Lecture Notes in Networks and Systems 647

Ajith Abraham · Tzung-Pei Hong · Ketan Kotecha · Kun Ma · Pooja Manghirmalani Mishra · Niketa Gandhi *Editors*

Hybrid Intelligent Systems

22nd International Conference on Hybrid Intelligent Systems (HIS 2022), December 13–15, 2022

Springer

Lecture Notes in Networks and Systems

Volume 647

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Ajith Abraham · Tzung-Pei Hong · Ketan Kotecha ·
Kun Ma · Pooja Manghirmalani Mishra ·
Niketa Gandhi
Editors

Hybrid Intelligent Systems

22nd International Conference on Hybrid Intelligent Systems (HIS 2022), December 13–15, 2022

Springer

Editors
Ajith Abraham
Faculty of Computing and Data Science
FLAME University
Pune, Maharashtra, India

Scientific Network for Innovation
and Research Excellence
Machine Intelligence Research Labs
Auburn, WA, USA

Ketan Kotecha
Symbiosis International University
Pune, India

Pooja Manghirmalani Mishra
Scientific Network for Innovation
and Research Excellence
Machine Intelligence Research Labs
Mala, Kerala, India

Tzung-Pei Hong
National University of Kaohsiung
Kaohsiung, Taiwan

Kun Ma
University of Jinan
Jinan, China

Niketa Gandhi
Scientific Network for Innovation
and Research Excellence
Machine Intelligence Research Labs
Auburn, WA, USA

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-27408-4 ISBN 978-3-031-27409-1 (eBook)
https://doi.org/10.1007/978-3-031-27409-1

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

HIS - IAS Organization

General Chairs

Ajith Abraham, Machine Intelligence Research Labs, USA
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Artūras Kaklauskas, Vilnius Gediminas Technical University, Lithuania

Program Chairs

Ketan Kotecha, Symbiosis International University, India
Ganeshsree Selvachandran, UCSI University, Malaysia

Publication Chairs

Niketa Gandhi, Machine Intelligence Research Labs, USA
Kun Ma, University of Jinan, China

Special Session Chair

Gabriella Casalino, University of Bari, Italy

Publicity Chairs

Pooja Manghirmalani Mishra, University of Mumbai, India
Anu Bajaj, Machine Intelligence Research Labs, USA

Publicity Team

Peeyush Singhal, SIT-Pune, India
Aswathy SU, Jyothi Engineering College, India
Shreya Biswas, Jadavpur University, India

International Program Committee

Aboli Marathe, Carnegie Mellon University, USA
Albert Alexander S., Vellore Institute of Technology, India
Alfonso Guarino, University of Foggia, Italy
Anu Bajaj, Thapar Institute of Engineering and Technology, India
Arthi Balakrishnan, SRM Institute of Science and Technology, India
Aswathy R. H., KPR Institute of Engineering and Technology, India
Aswathy S. U., Marian Engineering College, India
Cengiz Kahraman, Istanbul Technical University, Turkey
Devi Priya Rangasamy, Kongu Engineering College, India
Elİf Karakaya, Istanbul Medeniyet University, Turkey
Elizabeth Goldbarg, Federal University of Rio Grande do Norte, Brazil
Fariba Goodarzian, University of Seville, Spain
Gahangir Hossain, University of North Texas, USA
Gianluca Zaza, University of Bari “Aldo Moro”, Italy
Gowsic K., Mahendra Engineering College, India
Isabel S. Jesus, Institute of Engineering of Porto, Portugal
Islame Felipe Da Costa Fernandes, Federal University of Bahia (UFBA), Brazil
Jerry Chun-Wei Lin, Western Norway University of Applied Sciences, Bergen
José Everardo Bessa Maia, State University of Ceará, Brazil
Kun Ma, University of Jinan, China
Lalitha K., Kongu Engineering College, India
Lee Chang-Yong, Kongju National University, South Korea
M.Siva Sangari, KPR Institute of Engineering and Technology, India
Meera Ramadas, University College of Bahrain, Bahrain
Muhammet Raşit Cesur, Istanbul Medeniyet University, Turkey
Oscar Castillo, Tijuana Institute of Technology, Mexico
Padmashani R., PSG College of Technology, India

Paulo Henrique Asconavieta da Silva, Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense, Brazil
Pooja Manghirmalani Mishra, Machine Intelligence Research Labs, India
Prajoon P., Jyothi Engineering College, India
Radu-Emil Precup, Politehnica University of Timisoara, Romania
Sandeep Trivedi, Deloitte Consulting LLP, USA
Sandeep Verma, IIT Kharagpur, India
Sandhiya R., Kongu Engineering College, India
Sangeetha Shyam Kumar, PSG College of Technology, India
Sasikala K., Vinayaka Mission's Kirupananda Variyar Engineering College, India
Shalli Rani, Chitkara University, India
Sindhu P. M., Nagindas Khandwala College, India
Sruthi Kanakachalam, Kongu Engineering College, India
Suresh P., KPR Institute of Engineering and Technology, India
Suresh S., KPR Institute of Engineering and Technology, India
Thatiana C. N. Souza, Federal Rural University of the Semi-Arid, Brazil
Thiago Soares Marques, Federal University of Rio Grande do Norte, Brazil
Wen-Yang Lin, National University of Kaohsiung, Taiwan

Preface

Welcome to the 22nd International Conference on Hybrid Intelligent Systems (HIS 2022) and the 18th International Conference on Information Assurance and Security (IAS 2022) held in the World Wide Web during December 13–15, 2022. Due to the ongoing pandemic situation, both events were held online.

Hybridization of intelligent systems is a promising research field of modern artificial/computational intelligence concerned with the development of the next generation of intelligent systems. A fundamental stimulus to the investigations of Hybrid Intelligent Systems (HIS) is the awareness in the academic communities that combined approaches will be necessary if the remaining tough problems in computational intelligence are to be solved. Recently, hybrid intelligent systems are getting popular due to their capabilities in handling several real-world complexities involving imprecision, uncertainty, and vagueness. HIS 2022 received submissions from 28 countries, and each paper was reviewed by at least five reviewers in a standard peer-review process. Based on the recommendation by five independent referees, finally 97 papers were presented during the conference (acceptance rate of 34%).

Information assurance and security has become an important research issue in the networked and distributed information sharing environments. Finding effective ways to protect information systems, networks, and sensitive data within the critical information infrastructure is challenging even with the most advanced technology and trained professionals. The 16th International Conference on Information Assurance and Security (IAS) aims to bring together researchers, practitioners, developers, and policy-makers involved in multiple disciplines of information security and assurance to exchange ideas and to learn the latest development in this important field. IAS 2022 received submissions from 14 countries, and each paper was reviewed by at least five reviewers in a standard peer-review process. Based on the recommendation by five independent referees, finally 26 papers were presented during the conference (acceptance rate of 38%).

Many people have collaborated and worked hard to produce this year successful HIS–IAS conferences. First and foremost, we would like to thank all the authors for submitting their papers to the conference, for their presentations and discussions

during the conference. Our thanks to program committee members and reviewers, who carried out the most difficult work by carefully evaluating the submitted papers. Our special thanks to the following plenary speakers, for their exciting plenary talks:

- Kaisa Miettinen, University of Jyvaskyla, Finland
- Joanna Kolodziej, NASK-National Research Institute, Poland
- Katherine Malan, University of South Africa, South Africa
- Maki Sakamoto, The University of Electro-Communications, Japan
- Catarina Silva, University of Coimbra, Portugal
- Kaspar Riesen, University of Bern, Switzerland
- Mário Antunes, Polytechnic Institute of Leiria, Portugal
- Yifei Pu, College of Computer Science, Sichuan University, China
- Patrik Christen, FHNW, Institute for Information Systems, Olten, Switzerland
- Patricia Melin, Tijuana Institute of Technology, Mexico.

Our special thanks to the Springer Publication team for the wonderful support for the publication of these proceedings. We express our sincere thanks to the session chairs and organizing committee chairs for helping us to formulate a rich technical program. We express our sincere thanks to the organizing committee chairs for helping us to formulate a rich technical program. Enjoy reading the articles!

Maharashtra, India — Ajith Abraham
Kaohsiung, Taiwan — Tzung-Pei Hong
Pune, India — Ketan Kotecha
Jinan, China — Kun Ma
Mala, India — Pooja Manghirmalani Mishra
Auburn, USA — Niketa Gandhi

Contents

Hybrid Intelligent Systems

Bibliometric Analysis of Studies on Lexical Simplification 3
Gayatri Venugopal and Dhanya Pramod

Convolutional Neural Networks for Face Detection and Face Mask Multiclass Classification 13
Alexis Campos, Patricia Melin, and Daniela Sánchez

A Robust Self-generating Training ANFIS Algorithm for Time Series and Non-time Series Intended for Non-linear Optimization 21
A. Stanley Raj and H. Mary Henrietta

An IoT System Design for Industrial Zone Environmental Monitoring Systems 32
Ha Duyen Trung

A Comparison of YOLO Networks for Ship Detection and Classification from Optical Remote-Sensing Images 43
Ha Duyen Trung

Design and Implementation of Transceiver Module for Inter FPGA Routing 53
C. Hemanth, R. G. Sangeetha, and R. Ragamathana

Intelligent Multi-level Analytics Approach to Predict Water Quality Index 63
Samaher Al-Janabi and Zahraa Al-Barmani

Hybridized Deep Learning Model with Optimization Algorithm: A Novel Methodology for Prediction of Natural Gas 79
Hadeer Majed, Samaher Al-Janabi, and Saif Mahmood

PMFRO: Personalized Men's Fashion Recommendation Using Dynamic Ontological Models 96
S. Arunkumar, Gerard Deepak, J. Sheeba Priyadarshini, and A. Santhanavijayan

Hybrid Diet Recommender System Using Machine Learning Technique 106
N. Vignesh, S. Bhuvaneswari, Ketan Kotecha, and V. Subramaniyaswamy

QG-SKI: Question Classification and MCQ Question Generation Using Sequential Knowledge Induction 116
R. Dhanvardini, Gerard Deepak, and A. Santhanavijayan

A Transfer Learning Approach to the Development of an Automation System for Recognizing Guava Disease Using CNN Models for Feasible Fruit Production 127
Rashiduzzaman Shakil, Bonna Akter, Aditya Rajbongshi, Umme Sara, Mala Rani Barman, and Aditi Dhali

Using Intention of Online Food Delivery Services in Industry 4.0: Evidence from Vietnam 142
Nguyen Thi Ngan and Bui Huy Khoi

A Comprehensive Study and Understanding—A Neurocomputing Prediction Techniques in Renewable Energies 152
Ghada S. Mohammed, Samaher Al-Janabi, and Thekra Haider

Predicting Participants' Performance in Programming Contests Using Deep Learning Techniques 166
Md. Mahbubur Rahman, Badhan Chandra Das, Al Amin Biswas, and Md. Musfique Anwar

Fuzzy Kernel Weighted Random Projection Ensemble Clustering For High Dimensional Data 177
Ines Lahmar, Aida Zaier, Mohamed Yahia, and Ridha Boaullegue

A Novel Lightweight Lung Cancer Classifier Through Hybridization of DNN and Comparative Feature Optimizer 188
Sandeep Trivedi, Nikhil Patel, and Nuruzzaman Faruqui

A Smart Eye Detection System Using Digital Certification to Combat the Spread of COVID-19 (SEDDC) 198
Murad Al-Rajab, Ibrahim Alqatawneh, Ahmad Jasim Jasmy, and Syed Muhammad Noman

Hyperspectral Image Classification Using Denoised Stacked Auto Encoder-Based Restricted Boltzmann Machine Classifier 213
N. Yuvaraj, K. Praghash, R. Arshath Raja, S. Chidambaram, and D. Shreecharan

Prediction Type of Codon Effect in Each Disease Based on Intelligent Data Analysis Techniques 222
Zena A. Kadhuim and Samaher Al-Janabi

A Machine Learning-Based Traditional and Ensemble Technique for Predicting Breast Cancer 237
Aunik Hasan Mridul, Md. Jahidul Islam, Asifuzzaman Asif, Mushfiqur Rahman, and Mohammad Jahangir Alam

Recommender System for Scholarly Articles to Monitor COVID-19 Trends in Social Media Based on Low-Cost Topic Modeling 249
Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha

Statistical and Deep Machine Learning Techniques to Forecast Cryptocurrency Volatility 260
Ángeles Cebrián-Hernández, Enrique Jiménez-Rodríguez, and Antonio J. Tallón-Ballesteros

I-DLMI: Web Image Recommendation Using Deep Learning and Machine Intelligence 270
Beulah Divya Kannan and Gerard Deepak

Uncertain Configurable IoT Composition With QoT Properties 281
Soura Boulaares, Salma Sassi, Djamal Benslimane, and Sami Faiz

SR-Net: A Super-Resolution Image Based on DWT and DCNN 291
Nesrine Chaibi, Asma Eladel, and Mourad Zaied

Performance of Sine Cosine Algorithm for ANN Tuning and Training for IoT Security 302
Nebojsa Bacanin, Miodrag Zivkovic, Zlatko Hajdarevic, Stefana Janicijevic, Anni Dasho, Marina Marjanovic, and Luka Jovanovic

A Review of Deep Learning Techniques for Human Activity Recognition 313
Aayush Dhattarwal and Saroj Ratnoo

Selection of Replicas with Predictions of Resources Consumption 328
José Monteiro, Óscar Oliveira, and Davide Carneiro

VGATS-JSSP: Variant Genetic Algorithm and Tabu Search Applied to the Job Shop Scheduling Problem 337
Khadija Assafra, Bechir Alaya, Salah Zidi, and Mounir Zrigui

Socio-fashion Dataset: A Fashion Attribute Data Generated Using Fashion-Related Social Images 350
Seema Wazarkar, Bettahally N. Keshavamurthy, and Evander Darius Sequeira

Epileptic MEG Networks Connectivity Obtained by MNE, sLORETA, cMEM and dsPM 357
Ichrak ElBehy, Abir Hadriche, Ridha Jarray, and Nawel Jmail

Human Interaction and Classification Via K-ary Tree Hashing Over Body Pose Attributes Using Sports Data 366
Sandeep Trivedi, Nikhil Patel, Nuruzzaman Faruqui, and Sheikh Badar ud din Tahir

Bi-objective Grouping and Tabu Search 379
M. Beatriz Bernábe Loranca, M. Marleni Reyes, Carmen Cerón Garnica, and Alberto Carrillo Canán

Evacuation Centers Choice by Intuitionistic Fuzzy Graph 391
Alexander Bozhenyuk, Evgeniya Gerasimenko, and Sergey Rodzin

Movie Sentiment Analysis Based on Machine Learning Algorithms: Comparative Study 401
Nouha Arfaoui

Fish School Search Algorithm for Constrained Optimization 412
J. P. M. Alcântara, J. B. Monteiro-Filho, I. M. C. Albuquerque, J. L. Villar-Dias, M. G. P. Lacerda, and F. B. Lima-Neto

Text Mining-Based Author Profiling: Literature Review, Trends and Challenges 423
Fethi Fkih and Delel Rhouma

Prioritizing Management Action of Stricto Sensu Course: Data Analysis Supported by the k-means Algorithm 432
Luciano Azevedo de Souza, Wesley do Canto Souza, Welesson Flávio da Silva, Hudson Hübner de Souza, João Carlos Correia Baptista Soares de Mello, and Helder Gomes Costa

Prediction of Dementia Using SMOTE Based Oversampling and Stacking Classifier 441
Ferdib-Al-Islam, Mostofa Shariar Sanim, Md. Rahatul Islam, Shahid Rahman, Rafi Afzal, and Khan Mehedi Hasan

Sentiment Analysis of Real-Time Health Care Twitter Data Using Hadoop Ecosystem 453
Shaik Asif Hussain and Sana Al Ghawi

A Review on Applications of Computer Vision 464
Gaurav Singh, Parth Pidadi, and Dnyaneshwar S. Malwad

Analyzing and Augmenting the Linear Classification Models 480
Pooja Manghirmalani Mishra and Sushil Kulkarni

Literature Review on Recommender Systems: Techniques, Trends and Challenges 493
Fethi Fkih and Delel Rhouma

Detection of Heart Diseases Using CNN-LSTM 501
Hend Karoui, Sihem Hamza, and Yassine Ben Ayed

Incremental Cluster Interpretation with Fuzzy ART in Web Analytics 510
Wui-Lee Chang, Sing-Ling Ong, and Jill Ling

TURBaN: A Theory-Guided Model for Unemployment Rate Prediction Using Bayesian Network in Pandemic Scenario 521
Monidipa Das, Aysha Basheer, and Sanghamitra Bandyopadhyay

Pre-training Meets Clustering: A Hybrid Extractive Multi-document Summarization Model 532
Akanksha Karotia and Seba Susan

GAN Based Restyling of Arabic Handwritten Historical Documents ... 543
Mohamed Ali Erromh, Haïfa Nakouri, and Imen Boukhris

A New Filter Feature Selection Method Based on a Game Theoretic Decision Tree 556
Mihai Suciu and Rodica Ioana Lung

Erasable-Itemset Mining for Sequential Product Databases 566
Tzung-Pei Hong, Yi-Li Chen, Wei-Ming Huang, and Yu-Chuan Tsai

A Model for Making Dynamic Collective Decisions in Emergency Evacuation Tasks in Fuzzy Conditions 575
Vladislav I. Danilchenko and Viktor M. Kureychik

Conversion Operation: From Semi-structured Collection of Documents to Column-Oriented Structure 585
Hana Mallek, Faiza Ghozzi, and Faiez Gargouri

Mobile Image Compression Using Singular Value Decomposition and Deep Learning 595
Madhav Avasthi, Gayatri Venugopal, and Sachin Naik

Optimization of Traffic Light Cycles Using Genetic Algorithms and Surrogate Models 607
Andrés Leandro and Gabriel Luque

The Algorithm of the Unified Mechanism for Encoding and Decoding Solutions When Placing VLSI Components in Conditions of Different Orientation of Different-Sized Components 618
Vladislav I. Danilchenko, Eugenia V. Danilchenko, and Viktor M. Kureychik

Machine Learning-Based Social Media Text Analysis: Impact of the Rising Fuel Prices on Electric Vehicles 625
Kamal H. Jihad, Mohammed Rashad Baker, Mariem Farhat, and Mondher Frikha

MobileNet-Based Model for Histopathological Breast Cancer Image Classification 636
Imen Mohamed ben ahmed, Rania Maalej, and Monji Kherallah

Investigating the Use of a Distance-Weighted Criterion in Wrapper-Based Semi-supervised Methods 644
João C. Xavier Júnior, Cephas A. da S. Barreto, Arthur C. Gorgônio, Anne Magály de P. Canuto, Mateus F. Barros, and Victor V. Targino

Elections in Twitter Era: Predicting Winning Party in US Elections 2020 Using Deep Learning 655
Soham Chari, Rashmi T, Hitesh Mohan Kumain, and Hemant Rathore

Intuitionistic Multi-criteria Group Decision-Making for Evacuation Modelling with Storage at Nodes 668
Evgeniya Gerasimenko and Alexander Bozhenyuk

Task-Cloud Resource Mapping Heuristic Based on EET Value for Scheduling Tasks in Cloud Environment 680
Pazhanisamy Vanitha, Gobichettipalayam Krishnaswamy Kamalam, and V. P. Gayathri

BTSAH: Batch Task Scheduling Algorithm Based on Hungarian Algorithm in Cloud Computing Environment 690
Gobichettipalayam Krishnaswamy Kamalam, Sandhiya Raja, and Sruthi Kanakachalam

IoT Data Ness: From Streaming to Added Value 703
Ricardo Correia, Cristovão Sousa, and Davide Carneiro

Machine Learning-Based Social Media News Popularity Prediction 714
Rafsun Jani, Md. Shariful Islam Shanto, Badhan Chandra Das, and Khan Md. Hasib

Hand Gesture Control of Video Player 726
R. G. Sangeetha, C. Hemanth, Karthika S. Nair, Akhil R. Nair, and K. Nithin Shine

Comparative Analysis of Intrusion Detection System using ML and DL Techniques 736
C. K. Sunil, Sujan Reddy, Shashikantha G. Kanber, V. R. Sandeep, and Nagamma Patil

A Bee Colony Optimization Algorithm to Tuning Membership Functions in a Type-1 Fuzzy Logic System Applied in the Stabilization of a D.C. Motor Speed Controller 746
Leticia Amador-Angulo and Oscar Castillo

Binary Classification with Genetic Algorithms. A Study on Fitness Functions 756
Noémi Gaskó

SA-K2PC: Optimizing K2PC with Simulated Annealing for Bayesian Structure Learning 762
Samar Bouazizi, Emna Benmohamed, and Hela Ltifi

A Gaussian Mixture Clustering Approach Based on Extremal Optimization 776
Rodica Ioana Lung

Assessing the Performance of Hospital Waste Management in Tunisia Using a Fuzzy-Based Approach OWA and TOPSIS During COVID-19 Pandemic 786
Zaineb Abdellaoui, Mouna Derbel, and Ahmed Ghorbel

Applying ELECTRE TRI to Sort States According the Performance of Their Alumni in Brazilian National High School Exam (ENEM) 804
Helder Gomes Costa, Luciano Azevedo de Souza, and Marcos Costa Roboredo

Consumer Acceptance of Artificial Intelligence Constructs on Brand Loyalty in Online Shopping: Evidence from India 814
Shivani Malhan and Shikha Agnihotri

Performance Analysis of Turbo Codes for Wireless OFDM-based FSO Communication System 824
Ritu Gupta

Optimal Sizing and Placement of Distributed Generation in Eastern Grid of Bhutan Using Genetic Algorithm 831
Rajesh Rai, Roshan Dahal, Kinley Wangchuk, Sonam Dorji, K. Praghash, and S. Chidambaram

ANN Based MPPT Using Boost Converter for Solar Water Pumping Using DC Motor 841
Tshewang Jurme, Thinley Phelgay, Pema Gyeltshen, Sonam Dorji, Thinley Tobgay, K. Praghash, and S. Chidambaram

Sentiment Analysis from TWITTER Using NLTK 852
Nagendra Panini Challa, K. Reddy Madhavi, B. Naseeba, B. Balaji Bhanu, and Chandragiri Naresh

Cardiac Anomaly Detection Using Machine Learning 862
B. Naseeba, A. Prem Sai Haranath, Sasi Preetham Pamarthi, S. Farook, B. Balaji Bhanu, and B. Narendra Kumar Rao

Toxic Comment Classification 872
B. Naseeba, Pothuri Hemanth Raga Sai, B. Venkata Phani Karthik, Chengamma Chitteti, Katari Sai, and J. Avanija

Topic Modeling Approaches—A Comparative Analysis 881
D. Lakshminarayana Reddy and C. Shoba Bindu

Survey on Different ML Algorithms Applied on Neuroimaging for Brain Tumor Analysis (Detection, Features Selection, Segmentation and Classification) 893
K. R. Lavanya and C. Shoba Bindu

Visual OutDecK: A Web APP for Supporting Multicriteria Decision Modelling of Outranking Choice Problems 907
Helder Gomes Costa

Concepts for Energy Management in the Evolution of Smart Grids 917
Ritu Ritu

Optimized Load Balancing and Routing Using Machine Learning Approach in Intelligent Transportation Systems: A Survey 929
M. Saravanan, R. Devipriya, K. Sakthivel, J. G. Sujith, A. Saminathan, and S. Vijesh

Outlier Detection from Mixed Attribute Space Using Hybrid Model 940
Lingam Sunitha, M. Bal Raju, Shanthi Makka, and Shravya Ramasahayam

An ERP Implementation Case Study in the South African Retail Sector 948
Oluwasegun Julius Aroba, Kameshni K. Chinsamy, and Tsepo G. Makwakwa

Analysis of SARIMA-BiLSTM-BiGRU in Furniture Time Series Forecasting 959
K. Mouthami, N. Yuvaraj, and R. I. Pooja

VANET Handoff from IEEE 80.11p to Cellular Network Based on Discharging with Handover Pronouncement Based on Software Defined Network (DHP-SDN) 971
M. Sarvavnan, R. Lakshmi Narayanan, and K. Kavitha

An Automatic Detection of Heart Block from ECG Images Using YOLOv4 981
Samar Das, Omlan Hasan, Anupam Chowdhury, Sultan Md Aslam, and Syed Md. Minhaz Hossain

Attendance Automation System with Facial Authorization and Body Temperature Using Cloud Based Viola-Jones Face Recognition Algorithm 991
R. Devi Priya, P. Kirupa, S. Manoj Kumar, and K. Mouthami

Accident Prediction in Smart Vehicle Urban City Communication Using Machine Learning Algorithm 1002
M. Saravanan, K. Sakthivel, J. G. Sujith, A. Saminathan, and S. Vijesh

Analytical Study of Starbucks Using Clustering 1013
Surya Nandan Panwar, Saliya Goyal, and Prafulla Bafna

Analytical Study of Effects on Business Sectors During Pandemic-Data Mining Approach 1022
Samruddhi Pawar, Shubham Agarwal, and Prafulla Bafna

Financial Big Data Analysis Using Anti-tampering Blockchain-Based Deep Learning 1031
K. Praghash, N. Yuvaraj, Geno Peter, Albert Alexander Stonier, and R. Devi Priya

A Handy Diagnostic Tool for Early Congestive Heart Failure Prediction Using Catboost Classifier 1041
S. Mythili, S. Pousia, M. Kalamani, V. Hindhuja, C. Nimisha, and C. Jayabharathi

Hybrid Convolutional Multilayer Perceptron for Cyber Physical Systems (HCMP-CPS) 1053
S. Pousia, S. Mythili, M. Kalamani, R. Manjith, J. P. Shri Tharanyaa, and C. Jayabharathi

Information Assurance and Security

Deployment of Co-operative Farming Ecosystems Using Blockchain ... 1067
Aishwarya Mahapatra, Pranav Gupta, Latika Swarnkar, Deeya Gupta, and Jayaprakash Kar

Bayesian Consideration for Influencing a Consumer's Intention to Purchase a COVID-19 Test Stick 1082
Nguyen Thi Ngan and Bui Huy Khoi

Analysis and Risk Consideration of Worldwide Cyber Incidents Related to Cryptoassets 1093
Kazumasa Omote, Yuto Tsuzuki, Keisho Ito, Ryohei Kishibuchi, Cao Yan, and Shohei Yada

Authenticated Encryption Engine for IoT Application 1102
Heera Wali, B. H. Shraddha, and Nalini C. Iyer

Multi-layer Intrusion Detection on the USB-IDS-1 Dataset 1114
Quang-Vinh Dang

Predictive Anomaly Detection 1122
Wassim Berriche and Francoise Sailhan

Quantum-Defended Lattice-Based Anonymous Mutual Authentication and Key-Exchange Scheme for the Smart-Grid System 1132
Hema Shekhawat and Daya Sagar Gupta

Intelligent Cybersecurity Awareness and Assessment System (ICAAS) 1143
Sumitra Biswal

A Study on Written Communication About Client-Side Web Security 1154
Sampsa Rauti, Samuli Laato, and Ali Farooq

It's All Connected: Detecting Phishing Transaction Records on Ethereum Using Link Prediction 1167
Chidimma Opara, Yingke Chen, and Bo Wei

An Efficient Deep Learning Framework FPR Detecting and Classifying Depression Using Electroencephalogram Signals 1179
S. U. Aswathy, Bibin Vincent, Pramod Mathew Jacob, Nisha Aniyan, Doney Daniel, and Jyothi Thomas

Comparative Study of Compact Descriptors for Vector Map Protection 1189
A. S. Asanov, Y. D. Vybornova, and V. A. Fedoseev

DDoS Detection Approach Based on Continual Learning in the SDN Environment 1199
Ameni Chetouane and Kamel Karoui

Secure e-Voting System—A Review 1209
Urmila Devi and Shweta Bansal

Securing East-West Communication in a Distributed SDN 1225
Hamdi Eltaief, Kawther Thabet, and El Kamel Ali

Implementing Autoencoder Compression to Intrusion Detection System 1235
I Gede Agung Krisna Pamungkas, Tohari Ahmad, Royyana Muslim Ijtihadie, and Ary Mazharuddin Shiddiqi

Secure East-West Communication to Authenticate Mobile Devices in a Distributed and Hierarchical SDN 1244
Maroua Moatemri, Hamdi Eltaief, Ali El Kamel, and Habib Youssef

Cyber Security Issues: Web Attack Investigation 1254
Sabrina Tarannum, Syed Md. Minhaz Hossain, and Taufique Sayeed

Encrypting the Colored Image by Diagonalizing 3D Non-linear Chaotic Map 1270
Rahul, Tanya Singhal, Saloni Sharma, and Smarth Chand

Study of Third-Party Analytics Services on University Websites 1284
Timi Heino, Sampsa Rauti, Robin Carlsson, and Ville Leppänen

A Systematic Literature Review on Security Aspects of Virtualization 1293
Jehan Hasneen, Vishnupriya Narayanan, and Kazi Masum Sadique

Detection of Presentation Attacks on Facial Authentication Systems Using Intel RealSense Depth Cameras 1303
A. A. Tarasov, A. Y. Denisova, and V. A. Fedoseev

Big Data Between Quality and Security 1315
Hiba El Balbali, Anas Abou El Kalam, and Mohamed Talha

Learning Discriminative Representations for Malware Family Classification 1327
Ayman El Aassal and Shou-Hsuan Stephen Huang

Host-Based Intrusion Detection: A Behavioral Approach Using Graph Model 1337
Zechun Cao and Shou-Hsuan Stephen Huang

Isolation Forest Based Anomaly Detection Approach for Wireless Body Area Networks 1347
Murad A. Rassam

Author Index 1359

Hybrid Intelligent Systems



Bibliometric Analysis of Studies on Lexical Simplification

Gayatri Venugopal[1(✉)] and Dhanya Pramod[2]

[1] Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India
gayatri.venugopal@sicsr.ac.in

[2] Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India

Abstract. Text simplification is the process of improving the accessibility of text by modifying the text in such a way that it becomes easy for the reader to understand, while at the same time retaining the meaning of the text. Lexical simplification is a subpart of text simplification wherein the words in the text are replaced with their simpler synonyms. Our study aimed to examine the work done in the area of lexical simplification in various languages around the world. We conducted this study to ascertain the progress of the field over the years. We included articles from journals indexed in Scopus, Web of Science and the Association for Computational Linguistics (ACL) anthology. We analysed various attributes of the articles and observed that journal publications received a significantly larger number of citations as compared to conference publications. The need for simplification studies in languages besides English was one of the other major findings. Although we saw an increase in collaboration among authors, there is a need for more collaboration among authors from different countries, which presents an opportunity for conducting cross-lingual studies in this area. The observations reported in this paper indicate the growth of this specialised area of natural language processing, and also direct researchers' attention to the fact that there is a wide scope for conducting more diverse research in this area. The data used for this study is available on https://github.com/gayatrivenugopal/bibliometric_lexical_simplification.

Keywords: bibliometric study · lexical simplification · natural language processing

1 Introduction

Natural language processing is a rapidly evolving field involving a multitude of tasks such as sentiment analysis, opinion mining, machine translation and named entity recognition to name a few. One such task is text simplification, which refers to the modification of text in such a way that it becomes more comprehensible for the reader without loss of information. Text simplification promotes the use of plain language in texts belonging to various domains such as legal, education, business etc. Text simplification, in turn can be categorised as syntactic simplification and lexical simplification based on the methods

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 3–12, 2023.
https://doi.org/10.1007/978-3-031-27409-1_1

used to simplify the text. Syntactic simplification refers the process of modifying the syntax of a sentence in a given text in order to make it simpler to understand, whereas lexical simplification refers the process of replacing one or more complex words in a sentence with its simpler synonym keeping the context of the complex word in mind. The current study aims to examine the work done in the area of lexical simplification in various languages around the world. It has proven to be useful for readers who are new to the language readers with reading disabilities such as dyslexia [1], aphasia [2], and also readers with a poor level of literacy, and children [3]. Lexical simplification is composed of various steps, i.e., complex word identification, substitution generation, word sense disambiguation and synonym ranking [4]. Each sub-task of lexical simplification in itself acts as a focused area of research. Hence we lay importance not just on lexical simplification, but also on the sub-tasks involved in lexical simplification, while retrieving studies for this analysis.

2 Related Work

Bibliometrics refers to the quantitative study of publications and their authors [5]. Such studies have been conducted in various fields including natural language processing, in order to discover patterns in existing studies and to identify the areas for potential research. Keramatfar and Amirkhani [6] conducted a bibliometric study on sentiment analysis and opinion mining using articles from Web of Science and Scopus databases. They used tools such as BibExcel [7], VOSviewer [8] and Microsoft Excel [9]. They observed that English was the dominant language in this field, occupying roughly 99% of the 3225 articles analysed by them. They also found that the papers with more number of authors had a higher citation count, indicating that collaborative research may be one of the factors leading to good quality papers. Another study on sentiment analysis [10] performed an extensive bibliometric analysis of the work done in this field. They analysed the trends in this research area, used structural topic modeling to identify the key research topics and performed collaboration analysis among many other analyses to determine the popularity of the field and explore future directions. They analysed results based on not just the quantity of publications but also the quality, by using H-index values of authors. Yu et al. [11] conducted a bibliometric study of the use of support vector machines in research. They used the web of science database for their research and visualised the results using VOSviewer. They analysed the papers published by researchers in China, including their collaboration with international researchers. They also used co-occurrence analysis to identify the keywords that commonly appear together in order to determine the terms that are most focused upon. Wang et al. [12] conducted a similar study covering research conducted from 1999 to 2018. They used 5 Microsoft Excel and VOSviewer to analyse the trends in publications, collaborations, affiliations, keywords etc. Radev et al. [13] studied papers published in Association for Computational Linguistics (ACL) and created networks that indicated paper citations, author citations and author collaborations.

The broad objective of our study was to conduct a bibliometric analysis of the publications in the area of lexical simplification, in order to discover patterns and gaps in the existing structure of work which could lead to future studies that would help advance

the field. Hence we analysed the papers that reported studies on lexical simplification, complex word identification and lexical complexity. The subsequent section covers the details of the analyses.

3 Methodology

The study included papers published in three databases – Scopus, Web of Science and Association for Computational Linguistics (ACL) Anthology. These sources were chosen as these databases and library are prominent in the field of natural language processing and computational linguistics. Scopus and Web of Science contain high quality publications in other fields as well. We extracted details of primary documents from Scopus, that is, documents whose information is readily available in the database, as opposed to secondary documents that are present in the reference lists of primary documents and are not present in the database. We searched for publications using the keywords lexical simplification, complex word identification, lexical complexity prediction, lexical complexity, complex word, text simplification, and lexical simplification. We thus obtained 770 relevant results from Scopus using the following search string, as on April 12 2021, and 543 results from Web of Science on April 25 2021. At the time of writing this paper, there existed no API to extract information about papers from the ACL Anthology, which is a database of all the papers published in ACL conference proceedings. However, the data is available on GitHub (https://github.com/acl-org/acl-anthology) in an XML format. We retrieved publications for every year from 1991 till 2020 as 1991 was the earliest year for which metadata was available, and obtained 139 results as on April 13 2021.

We observed that a few papers in Scopus showed 0 citations. However these papers have over 100 citations in other sources, such as the proceedings of the conference in which the paper was presented. Most of these conference proceedings were available in the ACL anthology. Hence we did not analyse citations based on a specific source. The ACL anthology dataset does not consist of information related to affiliation of authors and number of citations. Therefore the results reported in this paper with regard to these attributes have been generated from the data retrieved from Scopus and Web of Science. We used the scholarly Python package (https://pypi.org/project/scholarly/) to extract the number of citations for all the publications from Google Scholar.

There were 392 records that were present in more than one source (Scopus, Web of Science and ACL). We performed filtering in such a way that the duplicate records that were present in ACL were removed. Among the records that were present in both Web of Science and Scopus, the records from Web of Science were removed. The resultant set consisted of 875 records. The next step was to identify the inconsistent columns. Among these fields, the values for a few fields such as language and source were missing. We used the information from other fields such as 7 abstract and publication type or the name of the conference, to extract the missing values. Certain values such as publisher information were not available for a few of the records, hence we manually extracted the data for these fields. If the value of a field could not be found in any of the three databases, the value for the field of the record under consideration was left blank. The language field was populated by searching for a language in the abstract. Therefore,

if the author/s did not mention the language in the abstract, the language field for that record was left blank. The list of languages was retrieved from https://www.searchify.ca/list-of-languagesworld/. The problem of non-uniform headings in all three databases was dealt with by taking a union of all the headings for creating the final dataset.

The PRISMA statement that explains the process we adopted to filter the records can be seen in Fig. 1.

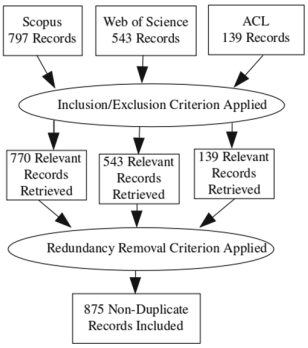


Fig. 1. PRISMA Statement

The data used for this study is available on https://github.com/gayatrivenugopal/bibliometric_lexical_simplification. We wrote scripts in Python to include the various analyses in our study, which have been reported in the subsequent section reports.

4 Results and Discussion

We observed that out of the 875 records, 415 publications were presented in conferences/workshops, which received a total of 7,540 citations, whereas 460 records were journal publications, which received 31,298 citations, signifying the relevance of journal publications and their reliability.

The percentage of publications with a single author has increased significantly over time and peaked in the year 2020 with over 17 publications being authored by a single researcher. We performed the Mann-Kendall trend test [14] (Hussain & Mahmud, 2019) and observed an increasing trend ($p = 4.738139658400087e{-}09$). However, if we compare these numbers with the total number of publications in each year, we can see that the proportion of publications with single author to the total publications in the year has reduced over the years, as is clearly visible in Fig. 2.

The field gained popularity from around the year 2011. There has been an increase in this count in the years 2016, 2018 and 2019. The SemEval tasks on complex word identification were held in the years 2016 and 2018 [15, 16]. The number of publications have peaked in the year 2020 indicating a possibility for further growth in the coming years.

We analysed the trend in the number of authors publishing in these areas in each year and obtained the graph shown in Fig. 3.

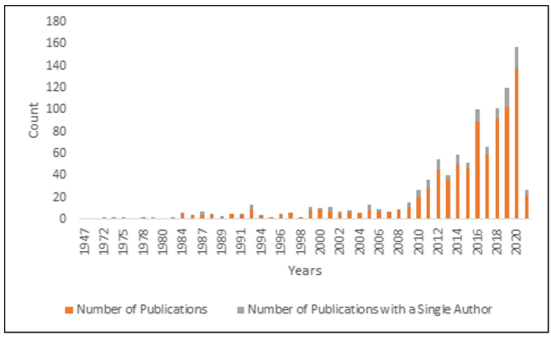


Fig. 2. Number of publications in each year

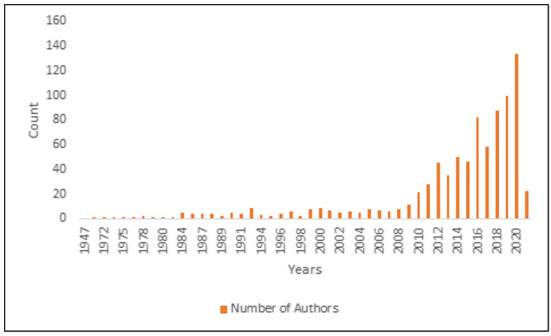


Fig. 3. Count of authors in each year

As can be seen, the field has grown significantly over the past few years and peaked in 2020.

Out of the 50 records that were available for extracting the language in focus, we observed the results depicted in Fig. 4 with respect to the popularity of each language.

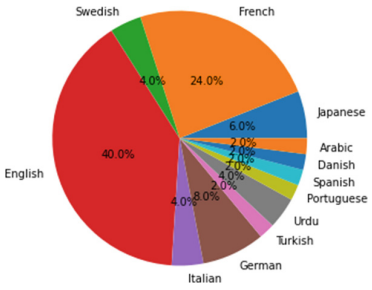


Fig. 4. Count of languages

We observed that English was a prominent language of choice of the researchers (that have mentioned the language in the abstract of the paper), followed by French.

The other languages had comparable counts of publications, indicating that more work needs to be done in languages other than English. We used the collaborative coefficient in order to study the collaboration among authors. The collaborative coefficient was devised by Ajiferuke et al. [17], which was later modified by Savanur and Srikanth [18] as collaborative coefficient does not produce 1 for maximum collaboration, unlike Modified Collaborative Coefficient (MCC). The MCC values for each year can be seen in Fig. 5.

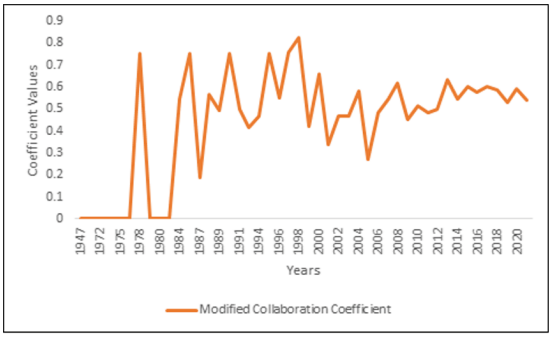


Fig. 5. Modified Collaborative Coefficient of Publications for each Year

As can be seen from the figure, more collaboration is required in this field as the coefficient decreases around the year 2000 and has not increased to a significant level since. However it should be noted that these values have been derived entirely from the data we acquired from the three databases alone. It is also observed that although the collaboration coefficient values are high in certain years such as 1978, the number of publications in 1978 were only 2, with one publication being written by 4 authors and the other publication being written by two authors. Therefore we analysed years with at least 10 publications and observed the results depicted in Fig. 6.

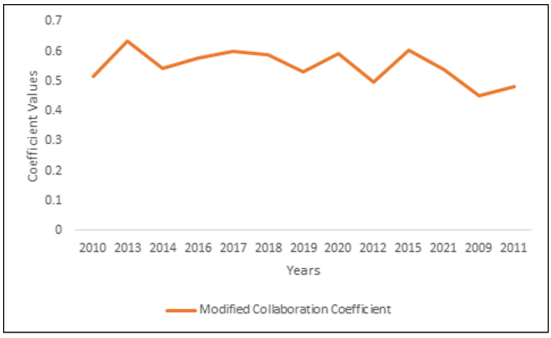


Fig. 6. Modified Collaborative Coefficient of Publications for each Year with Minimum 10 Publications

We can see that the collaboration is increasing although there are slumps in certain years such as 2019, a year in which the number of publications was high.

We calculated the correlation between the h-index of a country for a year and the number of publications from the country in the given year (in the specific research areas under consideration). We obtained 314 records, which were further normalised using min-max normalisation. We then calculated the Pearson's correlation coefficient and observed the value to be 0.4691. This indicates that there is a moderate correlation between the h-index of a country for a year and the number of publications contributed in this field during that year, which implies the significance of the field.

Readers working in this area would be familiar with the names mentioned in the figure. Figure 7 depicts the collaboration density among authors based on the data obtained from Scopus. We received a similar result for the data obtained from Web of Science. The darker the yellow colour, the more collaborative work has been done among the authors.

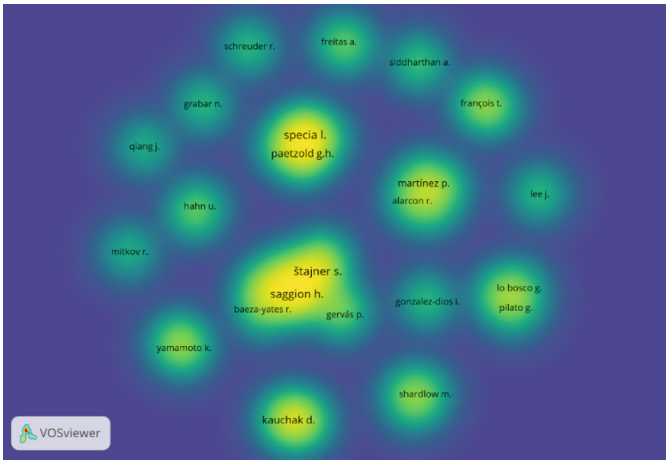


Fig. 7. Density visualisation of authors obtained from Scopus

In order to analyse the collaboration among countries, we determined the number of collaborations among the countries for which the data was available. We observed that there was not more than one instance of collaboration between any two countries. We believe that there could be more instances, however the data related to this was not readily available. As can be seen from Fig. 8, authors from United Kingdom have collaborated with authors from various other countries in this field. However, the collaboration among authors from other countries could be an aspect that could be focused on by researchers working in this area.

We determined the number of publications per author and plotted the top ten authors on a graph as shown in Fig. 9.

As can be seen, researcher Horacio Saggion has been very active in this field, closely followed by Lucia Specia and Sanja Stajner.

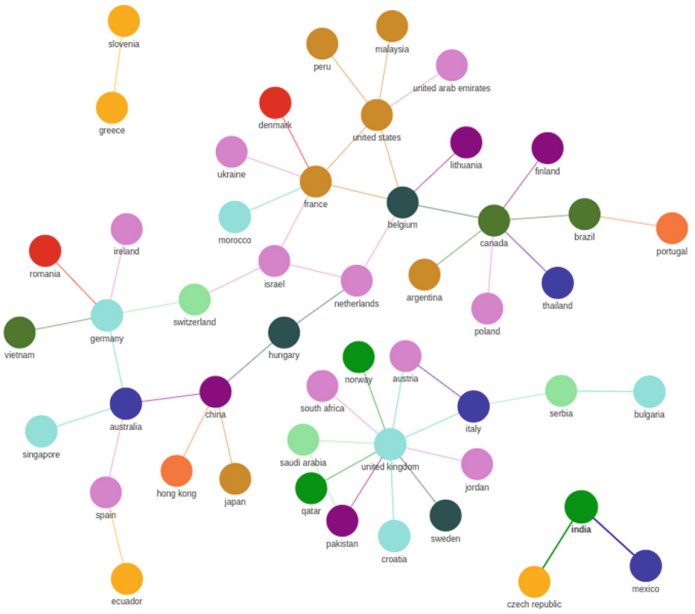


Fig. 8. Collaboration among countries

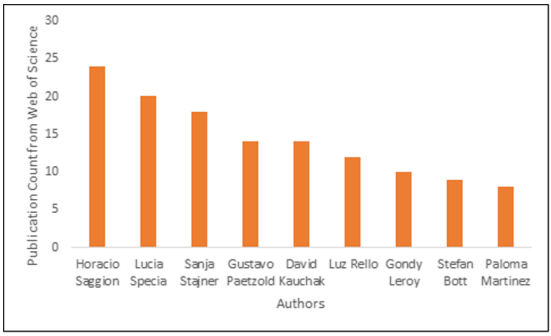


Fig. 9. The publication count of the top ten authors across the world

Finally, we analysed the citation data for the records. Figure 10 consists of the citation count for each country for which the data was available.

We observed that publications from Spain, United States and United Kingdom received the maximum number of citations.

The objective of our study was to gain insights into the existing body of work in the area of lexical complexity and simplification, regardless of the language. We observed that although there is only a 10.5% difference in the number of publications in journals and conferences, there is an approximately 300% difference in the citations received by journal publications and the citations received by publications in conference proceedings. This indicates the significance of publishing in journals, although conferences are good

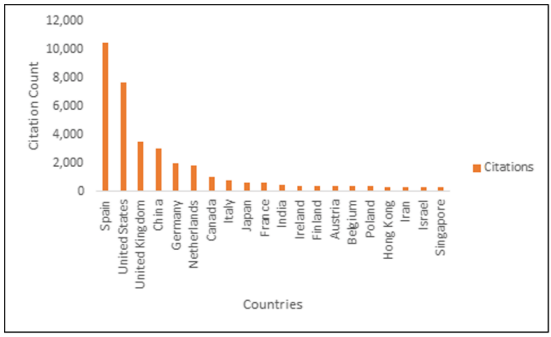


Fig. 10. Citation count for each country

venues for gaining feedback for the work from a diverse audience. The number of publications with single authors has reduced over the years, although the number of publications as well as the number of authors per year have increased especially post 2010, thus indicating more collaborative work in this field. With regard to language, the popularity of English has been established as most of the publications (for which the language related data was available) focused on English. We cannot deduce an inference entirely based on this observation, as only 50 articles, i.e., approximately 6% of the total number of articles contained information related to the language used, in their abstract. However, as compared to the observation made for other languages, we believe that there is a huge scope for work in languages other than English in this field. The modified collaboration coefficient values depicted in Fig. 10 depict an increase in collaboration among authors over the years, which reinstates our earlier claim that the field has evolved over the years. We can see a negative peak in 2019, though it does not indicate a significant decrease in collaboration.

The citation count graph displayed in Fig. 10 indicates the involvement of researchers from Spain, United States and United Kingdom. These countries also have the maximum number of publications in this field, and hence the large number of citations.

5 Conclusion

Through this study, we attempted to present the evolution of the field of lexical simplification and presented the observations and patterns we found. A major limitation was the absence of certain attributes, such as language, for the articles that were part of this study. A senior researcher in the area, Professor Dr. Emily M. Bender, emphasised on the importance of reporting the language under study, in research papers. This came to be known as the 'Bender Rule'. Along the same lines, we suggest that repositories that store papers related to natural language processing could add an additional section where the language/s associated with the paper can be mentioned. The growing number of papers and increasing collaboration indicates the growth of the field. We believe that cross-lingual lexical simplification research would encourage collaboration among authors from different countries. This study could be extended by including an analysis of the methods used for lexical simplification and the stages of lexical simplification, such as

complex word identification, word sense disambiguation etc. More work could also be done in studying the target users who were involved in these studies. For instance, certain studies involved their target readers in the annotation process, whereas other studies involved experts to annotate complex words. Another area that could be explored is the identification of similarities and/or patterns in the challenges and limitations reported by researchers in this area.

References

1. Rello, L., Baeza-Yates, R., Bott, S., Saggion, H.: Simplify or help? Text simplification strategies for people with dyslexia. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, pp. 1–10 (May, 2013)
2. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In: Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, pp. 7–10 (1998)
3. De Belder, J., Moens, M.F.: Text simplification for children. In: Proceedings of the SIGIR Workshop on Accessible Search Systems, pp. 19–26. ACM, New York (2010)
4. Shardlow, M.: A survey of automated text simplification. Int. J. Adv. Comput. Sci. Appl. **4**(1), 58–70 (2014)
5. Potter, W.G.: Introduction to bibliometrics. Library Trends **30**(5) (1981)
6. Keramatfar, A., Amirkhani, H.: Bibliometrics of sentiment analysis literature. J. Inf. Sci. **45**(1), 3–15 (2019)
7. Persson, O., Danell, R., Wiborg Schneider, J.: How to use Bibexcel for various types of bibliometric analysis. In: Åström, F., Danell, R., Larsen, B., Schneider, J. (eds.) Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday, pp. 9–24. International Society for Scientometrics and Informetrics, Leuven, Belgium (2009)
8. Van Eck, N.J., Waltman, L.: VOSviewer manual. Leiden: Univeristeit Leiden **1**(1), 1–53 (2013)
9. Microsoft Corporation: Microsoft Excel (2010). https://office.microsoft.com/excel
10. Chen, X., Xie, H.: A structural topic modeling-based bibliometric study of sentiment analysis literature. Cognit. Comput. **12**(6), 1097–1129 (2020)
11. Yu, D., Xu, Z., Wang, X.: Bibliometric analysis of support vector machines research trend: a case study in China. Int. J. Mach. Learn. Cybern. **11**(3), 715–728 (2020). https://doi.org/10.1007/s13042-019-01028-y
12. Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., Fan, L., Lei, J., et al.: Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed. J. Med. Internet Res. **22**(1), e16816 (2020)
13. Radev, D.R., Joseph, M.T., Gibson, B., Muthukrishnan, P.: A bibliometric and network analysis of the field of computational linguistics. J. Am. Soc. Inf. Sci. **67**(3), 683–706 (2016)
14. Mann, H.B.: Nonparametric tests against trend. Econometrica **13**, 245–259 (1945). https://doi.org/10.2307/1907187
15. Paetzold, G., Specia, L.: Semeval 2016 task 11: complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 560–569 (June 2016)
16. Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G.H., Specia, L., Štajner, S., Zampieri, M., et al.: A report on the complex word identification shared task (2018). arXiv:1804.09132
17. Ajiferuke, I., Burell, Q., Tague, J.: Collaborative coefficient: a single measure of the degree of collaboration in research. Scientometrics **14**(5–6), 421–433 (1988)
18. Savanur, K., Srikanth, R.: Modified collaborative coefficient: a new measure for quantifying the degree of research collaboration. Scientometrics **84**(2), 365–371 (2010)

Convolutional Neural Networks for Face Detection and Face Mask Multiclass Classification

Alexis Campos, Patricia Melin(✉), and Daniela Sánchez

Tijuana Institute of Technology, Tijuana, BC, Mexico
pmelin@tectijuana.mx

Abstract. In recent years, due to the COVID-19 pandemic, there have been a large number of infections among humans, causing the virus to spread around the world. According to recent studies, the use of masks has helped to prevent the spread of the virus, so it is very important to use them correctly. Using masks in public places has become a common practice these days and if it is not used correctly the virus will continue to be transmitted. The contribution of this work is the development of a convolutional neural network model to detect and classify the correct use of face masks. Deep learning methods are the most effective method to detect whether a person is using a mask properly. The proposed model was trained using the MaskedFace-Net dataset and evaluated with different images of it. The Caffe model is used for face detection, after which the image is preprocessed to extract features. These images are the input of the new convolutional neural network model, where it is classified among incorrect mask, non-mask, and mask. The proposed model achieves an accuracy rate of 99.69% in the test percentage, which is higher compared to other authors.

Keywords: Face mask · Convolutional neural network · Face detection

1 Introduction

Due to the new human coronavirus (COVID-19), there have been new respiratory symptoms and infections [1], and some of its symptoms are tiredness, dry cough, sore throat, fever, etc. This event has halted many activities worldwide due to the various effects it causes on humans.

The use of masks has worked as a strategy to decrease the spread of the COVID-19 virus, which has infected more than 430 million people worldwide according to the World Health Organization (until February 2022) [2]. One of the basic indications of the correct placement of the mask is that it should be placed covering the nose, mouth, and chin. Correctly performing this action will support the safety of oneself and the safety of others. Failure to follow these instructions could result in the spread of the virus to the people around.

During the COVID-19 pandemic in most countries, it became an obligation to use face masks [3], in February 2022 there were an estimated 50,595,554 new confirmed cases in the world [2], so it is necessary to identify people who correctly use face masks.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 13–20, 2023.
https://doi.org/10.1007/978-3-031-27409-1_2

Masks have different functions including preventing airborne viral particles from being transmitted between people and also allowing volatile particles to be filtered out of the air. The Centers for Disease Control and Prevention (CDC) recommendations indicate the use of surgical masks while exhaling air from the mouth and nose.

The development of computational methods employing machine learning makes it possible to automate the identification process through systems. Different studies use different deep learning models, such as YOLO [4–6], MobileNet [7, 8], Resnet [9–11] and Inception [12], these deep learning methods are sometimes preferred among the authors because they already have some recognition for the training of the Convolutional Neural Network (CNN) so they already have a good recognition rate.

In this research, a multiclass classification system is proposed for the recognition of the use of face mask, classifying them into three different classes, if the face mask is being used, the face mask is being used incorrectly, or if is not wearing a face mask. We design a convolutional neural model to be used by the system to detect and classify the correct use of face masks. The dataset provided by Cabani, named MaskedFaceNet [13], was used. The first 15,000 images of the dataset were preprocessed, identifying the region of interest using the Caffe model for face detection, then a feature extraction of the image was performed, which included resizing and RGB subtraction. Achieving a 99.69% accuracy, improving the percentage of accuracy compared to other authors. Therefore, the results of our initial experiments on the proposed model are presented.

The remainder of this article is organized into sections as follows. Section 2 mentions some related works in the field of classification and the use of the face mask. The proposed methodology is introduced in Sect. 3. Section 4 evaluates the model through various experiments. Finally, conclusions with possible future work are outlined in Sect. 5.

2 Related Works

Due to the COVID-19 virus, different techniques have been adapted through artificial intelligence to detect and classify people using face masks. This section will discuss some of the most relevant work on the classification of multi-class face masks.

In [14] the authors propose a CNN to detect people with face coverings used correctly, incorrectly, or without face coverings, using MaskedFaceNet and Flickr-Faces-HQ dataset [15] achieving an accuracy rate of 98.5%. In [7], the authors present a system to identify cover mask protocol violations, using the Haar Cascades to obtain the region of interest and the MobileNetV2 architecture as a model, achieving 99% accuracy with a MaskedFaceNet and Real Facemask Dataset. Similarly in [9], the author presents a graphical user interface (GUI) to identify the use of face covers by classifying it into the three classes of the previous authors, using the ResNet50 architecture with an accuracy of 91%, in the same way [10] used ResNet50 with 4 different datasets including MAFA, MaskedFace-Net and two from Kaggle. The author [5] proposed the standard use of coverslip recognition and detection algorithm based on YOLO-v4, achieving 98.3% accuracy. Other studies [16–18] present CNN models for the detection of coverslip usage using machine learning packages, such as Keras, OpenCV, TensorFlow, and Scikit-Learn.

The main differences found in the published articles are the architecture of their models, the data set, preprocessing and software libraries used to train their models.

Table 1 shows a comparison between the proposals of some authors who perform work similar to that proposed in this article. Our proposal is shown in the last row, with its respective characteristics.

Table 1. Authors' proposals for the classification of the use of face mask

1st Autor	Detection type	Classification model	Dataset	Software	Accuracy
Sethi [18]	Binary	CNN	MAFA	PyTorch	98.2%
Deshmukh [7]	Triple	MobileNetV2	RFMD, MaskedFace-Net	–	99%
Bhattarai [9]	Triple	ResNet50	Kaggle [19], MaskedFace-Net	OpenCV, Tensorflow, Keras	91%
Pham-Hoang-Nam [10]	Triple	ResNet50	Kaggle [19, 20], MaskedFace-NET, MAFA,	Tensorflow, Keras	94.59%
Yu [5]	Triple	YOLO-v4 Mejorado	RFMD, MaskedFace-Net	–	98.3%
Aydemir [21]	Triple	CNN	Manual, MaskedFace-Net	MATLAB	99.75
Soto-Paredes [11]	Triple	ResNet-18	MaskedFace-Net, Kaggle	PyTorch	99.05
Wang [12]	Triple	InceptionV2	RMFRD, MAFA, WIDER FACE, MaskedFace-Net	OpenCV, MATLAB	91.1%
Rudraraju [8]	Triple	MobileNet	RMFRD	OpenCV, Keras	90%
Jones [14]	Triple	CNN	MaskedFace-Net	Tensorflow, Keras	98.5%
Proposed Method in this Paper	Triple	CNN + Preprocessing	MaskedFace-Net	Tensorflow, Keras, OpenCV	99.69%

3 Proposed Method

This paper proposes a model combining deep learning, machine learning, and Python libraries. The proposal includes a CNN that allows the classification of the use of masks into three different classes (Mask, Incorrect Mask, and No Mask). The basic workflow of the proposed model is shown in Fig. 1.

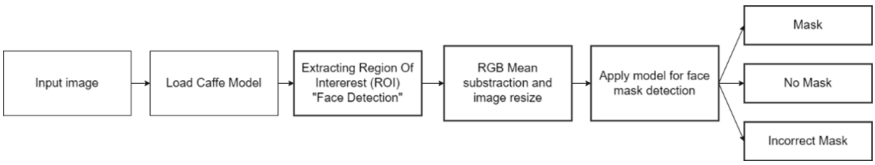


Fig. 1. Basic workflow.

3.1 General Architecture Description

Figure 2 shows the general architecture of the proposed convolutional neural network, including the learning and classification phase.

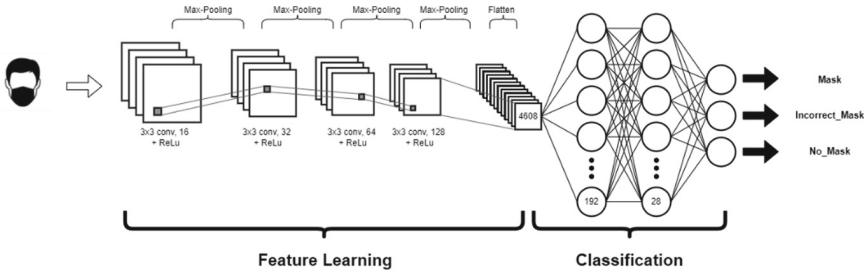


Fig. 2. The general architecture of the proposed method.

In order to classify the use of face masks, a convolutional neural network was designed and organized as follows. In the first part of the learning characteristics, four convolutional layers were considered, applying max-pooling techniques between each of them, as well as the ReLu activation function to improve the accuracy of the model, and the same padding was added to the convolutional layers.

3.2 Database

To train and validate the proposed convolutional neural network model, the MaskedFace-Net dataset for the Incorrect Mask and Mask classes, as well as, the Flickr-Faces-HQ Dataset (FFHQ) [15] face dataset was used. Therefore, in total 15000 images were used where each class containing the first 5000 images of the dataset. Some examples of classes are shown in Fig. 3. The images were separated into training, testing, and validation, where 70% was used to train the model, 20% for testing, and the remaining 10% for validation.

The dataset used is provided by Cabani [13], the author created a dataset called MaskedFace-Net, which contains 137,016 images with a resolution of 1024x1024 pixels, as the author mentions, this dataset is based on the Flickr-Faces-HQ (FFHQ) dataset, it was classified into two groups called Correctly Masked Face Dataset (CMFD) and Incorrectly Masked Face Dataset (IMFD).



Fig. 3. Examples of the database.

3.3 Creating a Model for Classification of Face Mask

The images used for training were classified into three different classes, “No_Mask” (using no mask), “Mask” (using a mask correct-ly), and “Incorrect_Mask” (using the mask incorrectly).

The model is based on an input image of $100 \times 100 \times 3$ pixels. Therefore, the input image is resized to these measurements. For each image in the dataset, to find the region of interest the Caffe model [22] was applied which automatically detects the face region, observe Fig. 4. The model was trained for 30 epochs with a batch of size 30.



Fig. 4. Sample of face detection.

In order to assist the convolutional neural network, the RGB subtraction technique was applied to the region of interest to help counteract the slight variations in the image, as shown in Fig. 5.

4 Experimental Results

For a fair comparison, the same dataset mentioned in [14] was used, taking the available features of the images used. The present model was trained with the dataset presented in Sect. 3, using the Python programming language with libraries such as Tensorflow and Keras. The model was evaluated with 30 trainings using the proposed model.

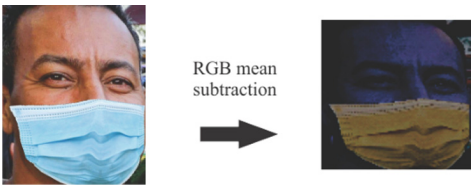


Fig. 5. Sample of pre-processing.

Table 2. Results of the proposed model

Training	Accuracy	Training	Accuracy
1	0.9958	16	0.9958
2	0.9958	17	0.9958
3	0.9958	18	0.9958
4	0.9958	19	0.9958
5	0.9958	20	0.9958
6	0.9958	21	0.9958
7	0.9969	22	0.9958
8	0.9969	23	0.9958
9	0.9969	24	0.9958
10	0.9958	25	0.9958
11	0.9958	26	0.9969
12	0.9958	27	0.9958
13	0.9958	28	0.9958
14	0.9958	29	0.9958
15	0.9958	30	0.9958

The results obtained are shown in Table 2, where it can be seen that the best result was obtained with training 7 with 99.69% accuracy.

Similarly, we can observe the confusion matrix evaluated with the test percentage, the confusion matrix is shown in Table 3.

In this paper to evaluate the effectiveness of the model, we consider 3 different parts of the MaskedFace-Net dataset, taking 15,000 different images for each part evaluating the model.

From the results in Table 4, we can observe the precision and loss values for each evaluated part. The model that had the best result with 99.69% was used to evaluate the different parts of the dataset, evaluating the first part 99.75% of accuracy was achieved, while for part 2 99.80% was achieved, part 3 being the last images of the dataset 99.63% of accuracy was obtained.

Table 3. Confusion matrix of the proposed model

Predicted / Actual	IncorrectMask	Mask	NoMask
IncorrectMask	1033	2	2
Mask	6	1005	2
NoMask	0	0	941

Table 4. Evaluating the model against different parts of the dataset

Training	Accuracy	Training	Accuracy
1	0.9958	16	0.9958
2	0.9958	17	0.9958
3	0.9958	18	0.9958

5 Conclusions

In this paper, a new CNN model was proposed to solve the problem of classifying and detecting the correct use of masks. The model can be classified into three different classes: Mask, NoMask, and IncorrectMask. In addition, to test the effectiveness of the model, the results were validated by evaluating it with different parts of the MaskedFace-Net dataset, the results showed that in general, this model allows to obtain a good classification percentage, achieving 99.69%. The model may be applied in real-time applications to help reduce the spread of the COVID-19 virus. In future work, other databases will be used for the classification and evaluation of the model in addition to testing in real-world applications.

References

1. Pedersen, S.F., Ho, Y.-C.: SARS-CoV-2: a storm is raging. J. Clin. Investig. **130**(5), 2202–2205 (2020)
2. World Health Organization, WHO Coronavirus (COVID-19) Dashboard, World Health Organization. https://covid19.who.int/. Accessed 25 Feb 2022
3. Erratum, MMWR. Morbidity and Mortality Weekly Report, vol. 70, no. 6, p. 293 (2021)
4. Singh, S., Ahuja, U., Kumar, M., Kumar, K., Sachdeva, M.: Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. Multimed. Tools Appl. **80**(13), 19753–19768 (2021). https://doi.org/10.1007/s11042-021-10711-8
5. Yu, J., Zhang, W.: Face mask wearing detection algorithm based on improved YOLO-v4. Sensors **21**(9), 3263 (2021)
6. Jiang, X., Gao, T., Zhu, Z., Zhao, Y.: Real-time face mask detection method based on YOLOv3. Electronics **10**(837), 1–17 (2021)

7. Deshmukh, M., Deshmukh, G., Pawar, P., Deore, P.: Covid-19 mask protocol violation detection using deep learning, computer vision. Int. Res. J. Eng. Technol. (IRJET) **8**(6), 3292–3295 (2021)
8. Rudraraju, S.R., Suryadevara, N.K., Negi, A.: Face mask detection at the fog computing gateway 2020. In: 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 521–524 (2020)
9. Bhattarai, B., Raj Pandeya, Y., Lee, J.: Deep learning-based face mask detection using automated GUI for COVID-19. In: 6th International Conference on Machine Learning Technologies, vol. 27, pp. 47–57 (2021)
10. Pham-Hoang-Nam, A., Le-Thi-Tuong, V., Phung-Khanh, L., Ly-Tu, N.: Densely populated regions face masks localization and classification using deep learning models. In: Proceedings of the Sixth International Conference on Research in Intelligent and Computing, pp. 71–76 (2022)
11. Soto-Paredes, C., Sulla-Torres, J.: Hybrid model of quantum transfer learning to classify face images with a COVID-19 mask. Int. J. Adv. Comput. Sci. Appl. **12**(10), 826–836 (2021)
12. Wang, B., Zhao, Y., Chen, P.: Hybrid transfer learning and broad learning system for wearing mask detection in the COVID-19 era. IEEE Trans. Instrum. Meas. **70**, 1–12 (2021)
13. Cabani, A., Hammoudi, K., Benhabiles, H., Melkemi, M.: MaskedFace-Net–a dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart Health **19**, 1–6 (2020)
14. Jones, D., Christoforou, C.: Mask recognition with computer vision in the age of a pandemic. In: The International FLAIRS Conference Proceedings, vol. 34(1), pp. 1–6 (2021)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **43**(12), 4217–4228 (2021)
16. Das, A., Wasif Ansari, M., Basak, R.: Covid-19 face mask detection using TensorFlow, Keras and OpenCV. In: 2020 IEEE 17th India Council International Conference (INDICON), pp. 1–5 (2020)
17. Kaur, G., et al.: Face mask recognition system using CNN model. Neurosci. Inf. **2**(3), 100035 (2022)
18. Sethi, S., Kathuria, M., Mamta, T.: A real-time integrated face mask detector to curtail spread of coronavirus. Comput. Model. Eng. Sci. **127**(2), 389–409 (2021)
19. Larxel: Face Mask Detection. https://www.kaggle.com/datasets/andrewmvd/face-mask-detection. Accessed 22 Mar 2022
20. Jangra, A.: Face Mask Detection 12K Images Dataset. https://www.kaggle.com/datasets/ashishjangra27/face-mask-12k-images-dataset/metadata. Accessed 22 Mar 2022
21. Aydemir, E., et al.: Hybrid deep feature generation for appropriate face mask use detection. Int. J. Environ. Res. Public Health **9**(4), 1–16 (2022)
22. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: MM 2014-Proceedings of the 2014 ACM Conference on Multimedia (2014)

A Robust Self-generating Training ANFIS Algorithm for Time Series and Non-time Series Intended for Non-linear Optimization

A. Stanley Raj[1] and H. Mary Henrietta[2(✉)]

[1] Loyola College, Chennai 600034, Tamil Nadu, India
[2] Saveetha Engineering College, Chennai 602105, Tamil Nadu, India
henrymaths123@gmail.com

Abstract. This paper provides an alternative to the conventional method of solving a complex problem using the artificial novel process. Algorithm testing is used to measure economic order (EOQ) quantities and groundwater depletion. This work incorporates both neurofuzzy and adaptive neurofuzzy to acquire the appropriate time series and non-time series tests. Effective asset management should be based on incorporating various decision variables like demand, setup costs, and order costs. The proposed self-training database algorithm announces an effective EOQ prediction model and water table level data graphically. Further, the data sets for both crisp and fuzzy models are examined and analyzed using the algorithm. The evaluation of the test results is suitable for use with any non-linear problem. This function proves that this algorithm works well for both time series and non-time series details.

Keywords: ANFIS · Economic order quantity (EOQ) · Groundwater level · Fuzzy logic

1 Introduction

Fuzzy sets referring to the ambiguity and uncertainty was first introduced by Zadeh [25]. This led to the EOQ model presented by Harris [6]. In asset management the strategy of EOQ is used for restoring and determining the total cost of the asset and is also controlled to reduce it. Previous demand was thought to be permanent which led to the collapse of the EOQ type. Therefore, there were models introduced with flexible requirements to deal with the volatile season in the business. To deal with the above problems, the management software can be suggested to customize the EOQ to derive a well-ordered order solution. Handling the constraints is important to come up with clever strategies to in solving real-time problems. Sinisa [22] studied a supply chain management and ANFIS was implemented to control the economic order quantity. Stanley et al. [23] incorporated ANFIS for examining the optimal order quantity.

Jang [11] initiated the study of artificial intelligence with asset management by combining a soft thinking system with flexible networks. This study, when paired with

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 21–31, 2023.
https://doi.org/10.1007/978-3-031-27409-1_3

artificial neural networks, (ANN) can make it easier in such unpredictable existing scenarios. A 'neural system' and a 'neurofuzzy' system are created when fuzzy and ANN interact. The well-known work of Aliev [2] has proposed two distinct structures, namely 'fuzzy-neural systems' and 'neuro fuzzy systems'. Fuzzy neural systems are used to obtain numerical data and practical knowledge data represented by fuzzy numbers, while neuro fuzzy systems have the important function of using mathematical relationships. Pedrycz [17] in the 1991, produced models for its behavior in relation to uncertainty and demonstrated connections about the theory in neural networks. Also, in 1992, Pedrycz [18] expanded his extensive study of neurons in pattern differentiation. In addition, NFN defined an unambiguous structure ordered by an algorithm by Jang's smart model (1993). Many disruptive neural networks are separated by communication between their neurons. In the year 1943, a mathematical model was developed by McCulloch with establishing a single neuron that checked the fidelity of the neuron function of the brain which was widely accepted for cognitive opportunities.

The asset management system known as inventory with neuro-fuzzy logic was launched by Balazs Lenart [3] in the year 2012. Besides, Aksoy [1] used ANFIS in the clothing trade to predicts its demand. Aengchuan and Pruksaphanrat [19] combined fuzzy inference system (FIS), Adaptive-Neuro-Fuzzy inference system (ANFIS) and ANN with different membership functions and solved asset problems. Apart from this, ANFIS and Gaussian membership activities have resulted in a full minimum cost. In 2015, another result that predicted ANFIS profits over ANN was made by Paul et al. [16] to maintain a good level of creativity in asset management problem.

Data-driven models are used in many fields of science including time series and non-time series. Time series data is translated using auto-regressive moving average (ARMA) and auto-regressive integrated average (ARIMA) and non-time series data are estimated using artificial neural networks (ANN), vector sustain equipment (SVM), adaptive neuro-fuzzy inference systems (ANFIS), and genetic engineering (GP) (Yoon et al. [24]; Fallah Mehdipour et al. [5]; Nourani et al. [15]). Shwetank [21] combined the ANFIS and Takagi-Sugeno fuzzy model to determine the quality of groundwater samples under five aspects. Dilip [4] had proposed multiple objective genetic algorithms using ANFIS to observe the groundwater level in wells. Hussam [9] applied the multiple objective genetic algorithms combined with ANFIS for studying the nitrate levels in groundwater. Rapantova [20] examined the groundwater contamination caused by uranium mines. Hass [7] inspected the quality of groundwater due to underground sewage in Berlin.

1.1 Materials and Methods

Many researchers have developed artificial intelligence hybrid approaches that have resulted in better results when employing ANFIS to solve challenging EOQ challenges. ANFIS is a commonly used learning method that uses complicated logic and a densely connected neural network to obtain the required output. A randomly generated asset is used to improve the asset management of the adaptive neuro-fuzzy inference system (ANFIS) in this model.

(1) Timeless series data analysis (EOQ model invention
(2) Data analysis on a timeline (groundwater fluctuations)

1.2 Inventory Model for EOQ

The ordering cost, O, and the constant demand rate coefficient, c, P stands for the coefficient of price-dependent demand. Selling price R, order size in Q, U is the unit purchase cost. The coefficient of constant holding cost is h are the input parameters for the corresponding model.

The total cost per cycle Kalaiarasi [12] and EOQ is derived

$$TC(Q) = \frac{O(c - pR)}{Q} + u(c - pP) + \frac{huQ}{2} \tag{1}$$

Differentiating partially w.r.t the order quantity parameter,

$$\frac{\partial T}{\partial Q} = -\frac{O(c - pR)}{Q^2} + \frac{hu}{2} \tag{2}$$

Equating $\frac{\partial T}{\partial Q} = 0$, the economic order quantity in crisp values is derived as

$$Q = \sqrt{\frac{2O(c - pR)}{uh}} \tag{3}$$

1.3 Groundwater Level Fluctuations

Groundwater exploration entails using geophysical tools to determine the subsurface hydrogeological surface area. According to studies, the geophysical approach not only creates subsurface characteristics but also detects signals from the environment in water, soil, and structures. As a result, any geophysical process's productive efficiency is determined by its ability to perceive by resolving discrepancies in surface water. Underground aquifers contain one of the nation's most valuable natural resources. The existence of human society is dependent on groundwater, which is a major supply of drinking water in India's urban and rural areas. The requirement for water has risen over time. India is dealing with issues such as water pollution and pollution as a result of poor management. As a result, millions of people do not have access to safe groundwater. Pollution of groundwater is caused by a combination of environmental and human factors. As a result, it's critical to be aware of issues like flooding, salt, agricultural toxicity, and industrial runoff, which are the primary causes of reduced groundwater levels. The application of neuro-fuzzy time series analysis has been widely employed. For stone filling ponds, Heydari et.al. [8] used the flow rate. Kisi [13] studied river flow using ANFIS. Mosleh [14] considered a hybrid model including ANFIS for the quality prediction of groundwater.

1.4 ANFIS Architecture

Initially, the information was entered into a certain level of membership grade, with the shooting power determining the following characteristics in each case.

Automated neural network pseudocode:

Calculate the workout's intensity.
Occasionally models are based on the number of repetitions

Repeat
Choose from a variety of free data to create your own sound reinforcement.
To choose the right data, use neural network activation functions.
Based on the data supplied, complete durability

Review the model's random weights for the best fit using error testing procedures.
The repetition will continue until the suspension requirements are reached.

The artificial network's architectural model
Begin now.

Algorithm Development

(1) Input set of training samples – x is utilized to increase the degree of membership, with the Gaussian membership function being applied here.

$$T_{1,i} = \mu_A(x), i = 1, 2, \ldots\ldots \text{n} \tag{4}$$

and

$$T_{1,i} = \mu_B(y), i = 1, 2, \ldots\ldots \text{n} \tag{5}$$

(2) Set the activation function

$$C^{x,1} \textit{ and perorm the activation function}$$

$$f(x, \sigma, c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \tag{6}$$

by {σ and c} are the membership function's parameters, also known as a parameter premise. The cluster bandwidth is denoted by b, and the cluster centre is denoted by c.

(3) Feed forward:

$$\textit{for each } i = 1,\ 2,\ 3 \textit{ and } l$$

Compute:

$$F^{x,l} = W^l C^{x,l-1} + B^l \textit{ and } C^{x,l} = \sigma\left(F^{x,l}\right) \tag{7}$$

Output Error:

$$\partial^{x,l} \textit{ compute the vector}$$

$$\partial^{x,l} = \nabla_c C_x \odot \sigma^1\left(F^{x,l}\right) \tag{8}$$

Error backpropagation and self-generating data

$$l = L - 1, L - 2, \ldots\ldots 2$$

$$\partial^{x,l} = ((W^{l+2})^T \partial^{x,l+1}) \odot \sigma^1\left(C^{x,l}\right) \tag{9}$$

Gradient descent

For each iteration $l = L, L - 2, \ldots .2$

By doubling each input signal, it serves to elicit firing-strength.

$$T_{2,i} = W_i = \mu_A(x)\mathrm{x}\mu_B(y), i = 1, 2. \tag{10}$$

Normalizes the *firing strength*

$$T_{3,i} = \overline{W_i} = \frac{W_i}{W + W\prime}, \quad i = 1, 2. \tag{11}$$

By adding all incoming signals, the ANFIS output signal is counted.

$$\Sigma_i = \overline{W_i} f_i = \frac{\Sigma_i W f_i}{\Sigma_i W_i}. \tag{12}$$

Update the weights

Calculates the output based on the rule's parameters as a result. $\{p_i, q_i \text{ and } r_i\}$

$$T_{4,i} = \overline{W_i} f_i = \overline{W_i}(p_i x + q_i y + r_i). \tag{13}$$

$$W^l \to W^l - \frac{\eta}{m} \sum_x \delta^{x,l}\left(C^{x,l-1}\right)^T \tag{14}$$

and

Biases

$$B^l \to B^l \to \frac{\eta}{m} \sum_x \delta^{x,l} \tag{15}$$

Model Creation and Implementation

Step 1:

The user can import data from the EOQ model. Text files, spreadsheets, CSV files, and other types of data must be imported.

Step 2:

ANFIS places a high value on training. As a result, ANFIS generates transaction data based on the number of duplicates and the amount of random data available. This artificial data automation will aid in the evaluation of sound and missing data. As a result, this is the most crucial phase in the algorithm.

Step 3:

The algorithm based on neuro fuzzy develops a synthetic database for each data link, which controls the sounds and faults in the data, and this is a crucial phase.

Step 4:

In this stage, the algorithm uses the coefficient of variation to calculate the percentage of data error.

Step 5:

A model is created depending on the number of refunds set by the user each time a program repeats a single transaction data. The algorithm's performance improves as the amount of transaction details increases.

Step 6:

Finally, the neural fuzzy algorithm predicts the Optimal order quantity model based on the fluctuating demand-rate.

This algorithm is built on a database of self-teaching. The system examines the definition parameters, standard deviation, high (high value), and low parameters in the automated model (low value). The technique generates performance data using random permissions after examining at all of the statistical variables from the input data. Each loop generates a new set of production data. The user must change the quantity of duplicate items to receive a big amount of transaction data. When the user increases the repeating value higher than the memory provided by the system to train the data, flexibility is lost. After a certain number of epochs, the system became unstable.

Error Estimation

The L1-norm error rating was used to reduce errors during the adjustment, and because it is more resilient than the L2-norm, it can be utilized in a variety of fields. Because the L2-norm squares are incorrect, the model performs substantially better when dealing with noisy data. By lowering the permissible error percentage to a minimum in this model, the most common issues that occur at ANFIS are avoided (10 percent in this study). The user repaired a genuine error by selecting the appropriate model parameters during iteration.

2 Results and Discussion

This research enables the ANFIS to test the data using an artificial training database generated by the given algorithm. The Gaussian membership function is given down in this EOQ model to forecast the outcome. Figure 1 displays the membership function used to train data. Figure 2 shows a self-made training database for EOQ. Figure 3 represents the ANFIS structure for the EOQ model. Figure 4 shows the tested data for EOQ. The bulk of the Economic Order is predicted by flexible demand using artificial intelligence databases using ANFIS.

ANFIS will put the data to the test using an algorithm-generated artificial training database. The membership function used to train data is shown in Fig. 1. Figure 4 depicts a training database for EOQ that was created by the author. Flexible demand, employing artificial intelligence databases and ANF, predicts the majority of the Economic Order. The business should be aware of the changing demand strategy as well as the Economic

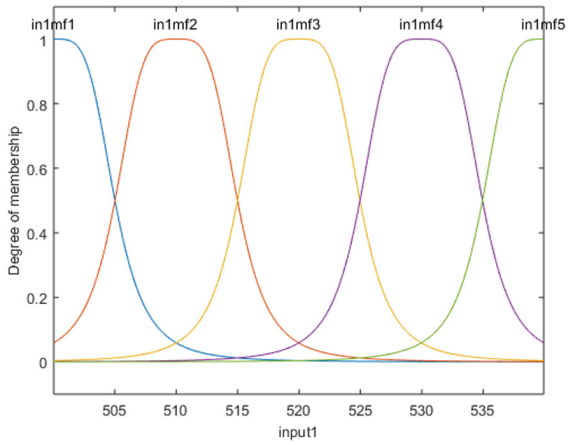


Fig. 1. The Gaussian membership function implemented to train the data

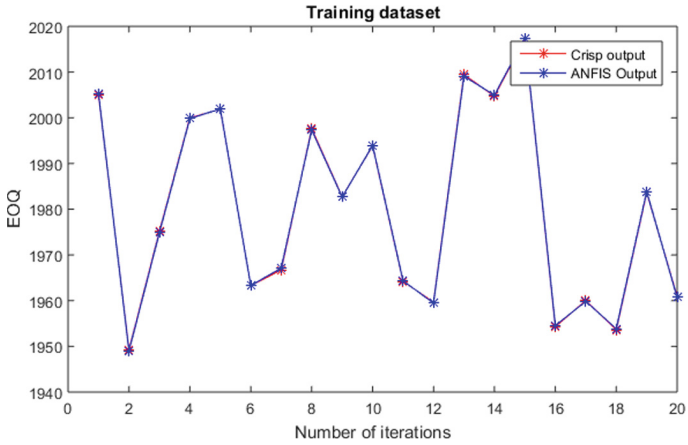


Fig. 2. The comparison between the self-generated training dataset and the original result

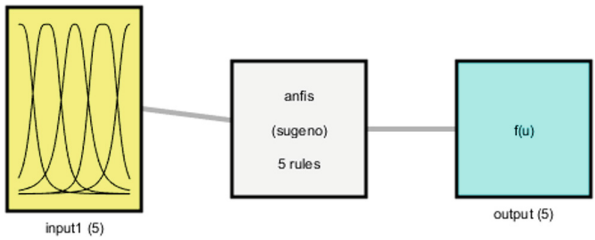


Fig. 3. ANFIS architecture for EOQ model

Order Quantity. As a result, we can readily estimate EOQ with fluctuating demand using

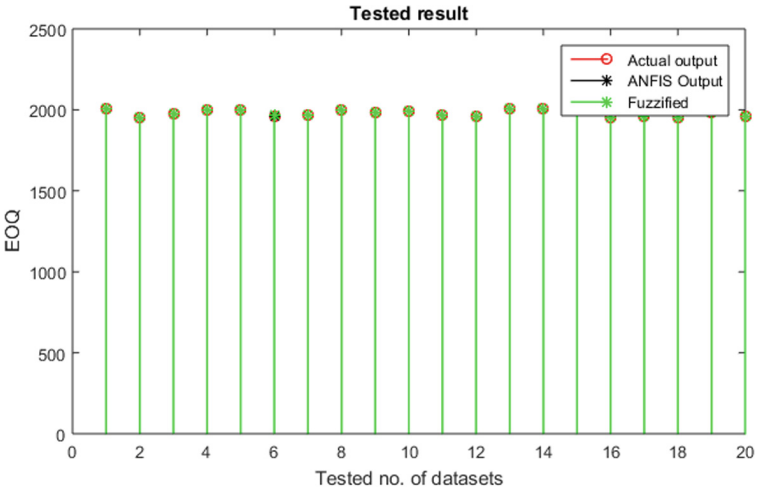


Fig. 4. Data examined using a synthetic training data-set generated by a self-generating system.

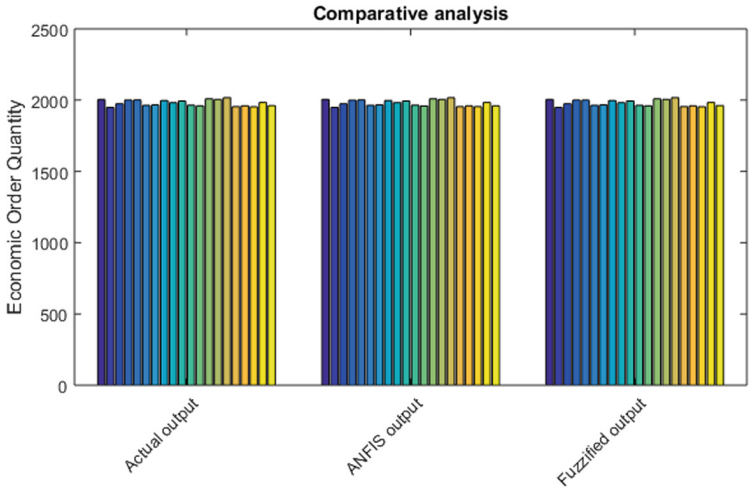


Fig. 5. A comparison of ANFIS output with crisp and fuzzified models in EOQ model.

this technique. When comparing the crisp and fuzzified models, this technique was successful (Fig. 5).

There is a clear correlation between performance and output data in ANFIS training. A water data training model is shown in Fig. 6. Figure 7 shows the results of the tests and a comparison of the three methods. ANFIS is one of the gentlest computer methods available for combining neural networks and sophisticated logic.

Personal data has a number of advantages,

1. 1.It includes the ability to erase errors or sounds in the data.

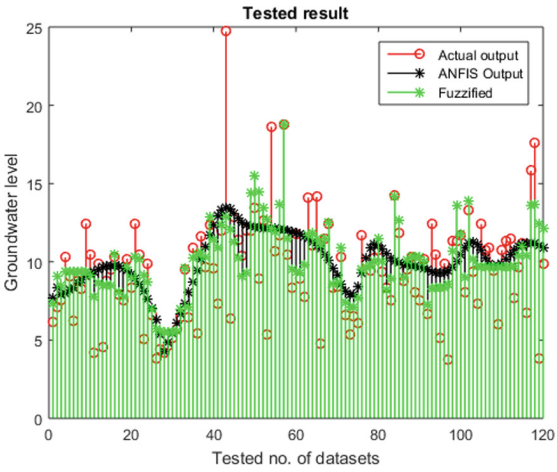


Fig. 6. Training data for groundwater fluctuation

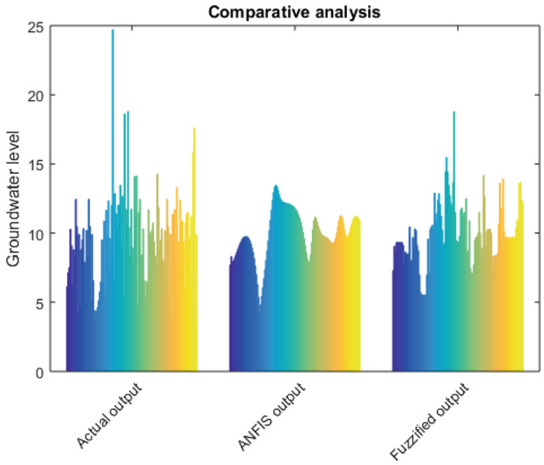


Fig. 7. A comparison of ANFIS output with crisp and fuzzified models for groundwater model

2. If data is lost between two data points, it may be included depending on the standard deviation and trend. The procedure will be followed.
3. The data for the training is expandable because data sets can move between the extrema of actual data and directly forecast the results, even if the data is out of line.
4. Using this approach to improve performance data makes it easier for the ANFIS system to determine output. Changing the data group's membership actions, which is time expensive, will forecast the immediate outcome after defuzzification.

An integrated platform for neural networks, fuzzy logic, and neuro-fuzzy networks can be used to create a variety of hybrid systems. The abstract notion, for example, can be used to mix findings from several neural networks; even if other hybrid systems are

developed, this current work has produced promising results when integrating abstract concepts with neural networks. Field validation shows that this approach has a promising prospect in measuring a wide range of off-line issues.

References

1. Aksoy, A., Ozturk, N., Sucky, E.: Demand forecasting for apparel manufacturers by using neuro-fuzzy techniques. J. Model. Manag. **9**(1), 18–35 (2014)
2. Aliev, R.A., Guirimov, B., Fazlohhahi, R., et al.: Evolutionary algorithm-based learning of fuzzy neural networks. Fuzzy Sets Syst. **160**(17), 2553–2566 (2009)
3. Lénárt, B., Grzybowska, K., Cimer, M.: Adaptive Inventory control in production systems. IN: International Conference on Hybrid Artificial Intelligence Systems, pp. 222–228 (2012)
4. Roy, D.K., Biswas, S.K., Mattar, M.A., et.al.: Groundwater level prediction using a multiple objective genetic algorithm-grey relational analysis based weighted ensemble of ANFIS Models. Water **13**(21), 3130 (2021)
5. Fallah-Mehdipour, E., Bozorg Haddad, O., Mariño, M.A.: Prediction and simulation of monthly groundwater levels by genetic programming. J. Hydro-Environ. Res. **7**, 253–260 (2013)
6. Harris, F.: Operations and Cost. AW Shaw Co., Chicago (1913)
7. Hass, U., Duünbier, U., Massmann, G.: Occurrence of psychoactive compounds and their metabolites in groundwater downgradient of a decommissioned sewage farm in Berlin (Germany). Environ. Sci. Pollut. Res. **19**, 2096–2106 (2012)
8. Heydari, M., Talaee, P.H.: Prediction of flow through rockfill dams using a neuro-fuzzy computing technique. Int. J. Appl. Math. Comput. Sci. **22**(3), 515–528 (2011)
9. Elzain, H.E., Chung, S.Y., Park, K.-H., et.al.: ANFIS-MOA models for the assessment of groundwater contamination vulnerability in a nitrate contaminated area. J. Environ. Manag. (2021)
10. Jang, J.R.: ANFIS: adaptive-network-based inference system. IEEE Trans. Syst. Man. Cybern (1993)
11. Jang, C., Chen, S.: Integrating indicator-based geostatistical estimation and aquifer Vulnerability of nitrate-N for establishing groundwater protection zones. J. Hydrol. **523**, 441–451 (2015)
12. Kalaiarasi, K., Sumathi, M., Mary Henrietta, H., Stanley, R.A.: Determining the efficiency of fuzzy logic EOQ inventory model with varying demand in comparison with Lagrangian and Kuhn-Tucker method through sensitivity analysis. J. Model Based Res. **1**(3), 1–12 (2020)
13. Kisi, O.: Discussion of application of neural network and adaptive neuro-fuzzy inference systems for river flow prediction. Hydrol. Sci. J. **55**(8), 1453–1454 (2010)
14. Al-adhaileh, M.H., Aldhyani, T.H., Alsaade, F.W., et.al.: Groundwater quality: the application of artificial intelligence. J. Environ. Pub. Health, 8425798 (2022)
15. Nourani, V., Alami, M.T., Vousoughi, F.D.: Wavelet-entropy data pre-processing approach for ANN-based groundwater level modeling. J. Hydrol. **524**, 255–269 (2015)
16. Paul, S.K., Azeem, A., Ghosh, A.K.: Application of adaptive neuro-fuzzy inference system and artificial neural network in inventory level forecasting. Int. J. Bus. Inf. Syst. **18**(3), 268–284 (2015)
17. Pedrycz, W.: Neurocomputations in relational systems. IEEE Trans. Pattern Anal. Mach. Intell. **13**(3), 289–297 (1991)
18. Pedrycz, W.: Fuzzy Neural Networks with reference neurons as pattern classifiers. IEEE Trans. Neural Netw. **3**(5), 770–775 (1992)

19. Aengchuan, P., Phruksaphanrat, B.: Comparison of fuzzy inference system (FIS), FIS with artificial neural networks (FIS +ANN) and FIS with adaptive neuro-fuzzy inference system (FIS+ANFIS) for inventory control. J. Intell. Manuf. **29**(4), 905–923 (2015)
20. Rapantova, N., Licbinska, M., Babka, O., et al.: Impact of uranium mines closure and abandonment on ground-water quality. Environ. Sci. Pollut. Res. **20**(11), 7590–7602 (2012)
21. Suhas, S., Chaudhary, J.K.: Hybridization of ANFIS and fuzzy logic for groundwater quality assessment. Groundw. Sustain. Dev. **18**, 100777 (2022)
22. Sremac, S., Zavadskas, E.K., Bojan, M., et.al.: Neuro-fuzzy inference systems approach to decision support system for economic order quantity. Econ. Res.-Ekonomska Istrazivanja **32**(1), 1114–1137 (2019)
23. Stanley Raj, A., Mary Henrietta, H., Kalaiarasi, K., Sumathi, M.: Rethinking the limits of optimization Economic Order Quantity (EOQ) using Self generating training model by Adaptive Neuro Fuzzy Inference System. In: Communications in Computer and Information Sciences, pp. 123–133. Springer (2021)
24. Yoon, H., Jun, S.C., Hyun, Y., et al.: A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. **396**, 128–138 (2011)
25. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)



An IoT System Design for Industrial Zone Environmental Monitoring Systems

Ha Duyen Trung(✉)

School of Electronical and Electronic Engineering (SEEE), Hanoi University of Science and Technology (HUST), Hanoi, Vietnam
trung.haduyen@hust.edu.vn

Abstract. This paper present the development of an Internet of Things (IoT) framework oriented to serve the management of industrial parks to continuously control and monitor the discharge of industrial park infrastructure investors, minimizing negative impacts on the surrounding living environment. In particular, we design and implement IoT end devices and IoT gateway based on open hardware platform for data collection and control of measuring and monitoring IoT devices. In addition, we build an open-source IoT cloud platform to support device management, data storage, processing, and analysis for applications of industrial parks and high-tech parks. The tested implementation has shown that the system design can be applied for the air and wastewater monitoring and management in industrial parks.

Keywords: IoT · open-source · gateway · devices · industial management

1 Introduction

In the recent years, the world is strongly transformed before the trend of "Internet of Things". According to the Ericsson Mobility Report, there are expected to be 28 billion connected devices, including 15 billion IoT-connected devices, including machine-to-machine (M2M) connections such as smart watches, street sensors, retail locations, consumer electronic devices such as televisions, automotive electronics, wearables, electronic musical instruments, digital cameras. The remaining 13 billion connections are from mobile phones, laptop PCs, and tablets [1]. According to McKinsey, IoT will contribute to the global economy of 11000 $ billion by 2025 [2].

The IoT has many different applications. One application that we currently hear about is "Smart City" with smart homes, all devices such as air conditioners, LED systems, health monitoring systems [3]. Intelligent sensor systems such as motion recognition, warning of air pollutants: NO_2 NO, SO_2, O_3, CO, PM10 and PM2.5 dust, and total suspended particles (TSP), are intelligently both connected and controlled via the Internet connections [4]. Moreover, in context

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 32–42, 2023.
https://doi.org/10.1007/978-3-031-27409-1_4

to the present standings of IoT, identification of the most prominent applications in the field of IoT have been highlighted and a comprehensive review has been done specifically in the field of precision agriculture [5].

Building an environmental monitoring network is one of the needs stemming from reality. Especially in the current era, when the continuous development of industries serving the country's modernization process. When the era of 4.0 Technology Revolution broke out, it caused certain negative impacts on the environment. Environmental protection has become a key topic that is concerned and focused on by society [5]. In this paper, we have implemented IoT devices to monitor environmental quality using different wireless connectivity. They are integrated on the same Gateway. The obtained data has been visualized in real-time on the dashboard web user interface and app platforms.

However, in addition to the positive contributions, the industrial development in general and the IZs system particularly in Vietnam are creating many challenges in environmental pollution due to solid waste, wastewater and gas industrial waste [6].

According to the World Bank, Vietnam can suffer losses due to environmental pollution up to 5.5% of annual GDP. Each year, Vietnam also loses 780 million USD in public health fields due to environmental pollution. Therefore, in this work, we develop the IoT framework for management applications in industrial and high-tech parks. Particularly, we design and implement IoT gateway devices based on open hardware platform for measuring, collecting, processing monitored environmental data from IoT end devices to platform via IoT gateways. In addition, we build an open-source IoT cloud platform to support device management, data storage, processing, and analysis for applications of industrial parks and high-tech parks.

The rest of this paper is organized as follows. Section 2 describes system architectures. Detail design descriptions of end devices, gateway, communication protocol and cloud, user applications are presented in this Section. Implementation results are presented in Sect. 2. Section 4 concludes this paper.

2 IoT System Design

There have been many research papers and application designs for monitoring of environmental parameters based on the IoT platform. However, each research method focuses on a certain radio protocol. In this article, we present a new method of multi-protocol wireless such as Bluetooth low energy, Z-wave, WiFi, Zigbee, Lora, and 4G. Such IoT wireless protocols are integrated on the Gateway. Figure 1 illustrates a horizontal architecture of the IoT network for industrial zone environmental monitoring system. In this architecture, air and wastewater sensors are embedded to IoT end devices for monitoring environmental parameters such as PM2.5, PM10, CO_2, temperature, humidity, EC conductivity, VOC, etc. End devices communicate with IoT gateway via various wireless protocols to send monitored parameters to servers for data aggregation. Industrial environment management can be supported by exploiting the bigdata analysis.

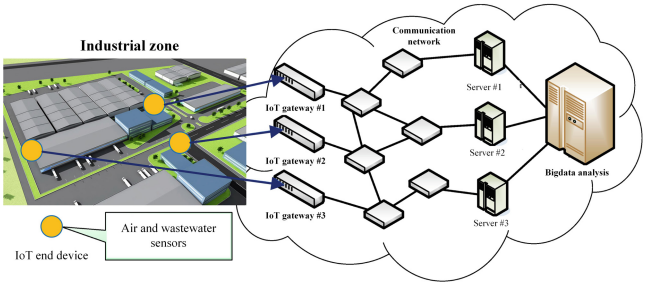


Fig. 1. A horizontal architecture of IoT network for industrial zone environmental monitoring system

Figure 2 shows the proposed system diagram of open IoT platform-based management applications for industrial zones. The system consists of many devices that are sensors that monitor environmental parameters and cameras. Each device uses a certain radio protocol. Data after measurement is sent back to Gateway. Here data will be pushed to the cloud via MQTT (Message Queuing Telemetry Transport) protocol [7–9]. Data in the cloud will be trained as well as used to display on user-friendly web and apps. The system is divided into each block to get a better understanding.

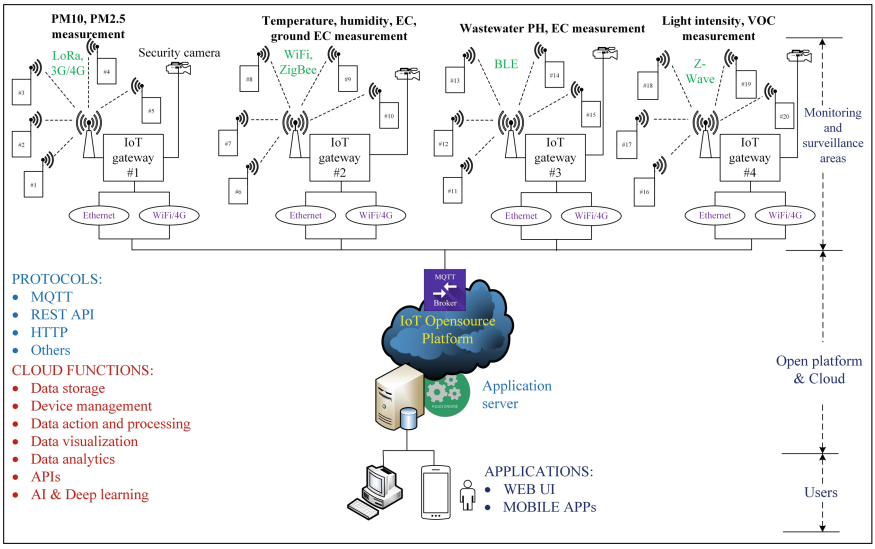


Fig. 2. A diagram of the proposed vertical open IoT platform for applications of industrial zone managements

2.1 Design of IoT Devices

As shown in the system overview, there are 20 devices with functions for monitoring different environments such as soil, water and air. There is also a security surveillance camera. We will go a little more deeply about the metrics the device tracking in each environment. Clearly, for the water environmental monitoring management, we use sensors of PH, EC conductivity, and temperature. For the air environmental monitoring management, we use sensors of temperature, humidity, light intensity, CO_2, VOC and dust PM2.5. Finally, we use sensors of temperature, humidity and EC conductivity in for monitoring soil environmental management. Cameras are used to monitor security, detect human movements and then send email notifications when it detects movement in the surveillance area. Each set of devices uses a certain wireless communication protocol. As a basis for making judgments and warnings. The design diagram of the hardware blocks is shown in the Fig. 3, including sensor, microprocessor, communication protocol and power block. First, we select the required sensors to measure the parameters for each of the environments shown in Table 1 as well as the corresponding measurement parameters. Figure 4 shows differences of the sensors used in the work.

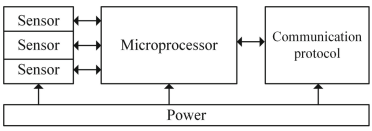


Fig. 3. The general block diagram of IoT devices

Table 1. Sensors used for the implementation

Sensor	Parameters
Analog electrical conductivity sensor (K = 1)	EC
Analog pH Sensor	pH
Digital temperature sensor DS18B20	Air temperature
Plantower PMS 7003	PM2.5 & PM10 dust
SHT31 temperature and humidity sensor	Temperature and humidity
MICS 6814 sensor	CO_2, VOC
Light intensity sensor Lux BH1750	Lux
MEC10-soil moisture, temperature sensor	Soil moisture & temperature

Choosing the controller unit for the hardware is important in electronic circuit fabrication. With the requirements of the given problem, we decided to



Fig. 4. Various air and wastewater sensors used for implementation of IoT devices

use the AT MEGA328P microprocessor. This is a microprocessor with a simple structure. The AT MEGA328P microprocessor includes 28 pins (23 I/O pins and 5 power pins), 32 registers, 3 programmable timers/counters, with internal and external interrupts, serial communication protocol USART, SPI, I2C. In addition, a 10-bit analog digital converter (ADC/DAC) that expands to 8 channels, operates with 5 power modes, can use up to 6 channels of pulse width modulation (PWM), supports the bootloader.

2.2 Design of IoT Gateway

At the gateway block, the authors use raspberry pi 3B embedded computer (2GB Ram version) which has integrated radio communication protocols corresponding to those on the device side. The special feature of this method is the integration of radio communication methods on the same Gateway, so that all can operate smoothly and without losing data packets. The block diagram in the Gate way is shown in Fig. 4. The Raspberry Pi 3B + uses a Broadcom BCM2837B0, quad-core A53 (ARMv8) processor, featuring a 64-bit quad-core chip clocked at 1.4 GHz. ARM calls the Cortex-A50 series "the world's best energy efficient 64-bit processors" thanks to being built on the ARMv8 instruction set architecture and bringing in new technical innovations. With a high degree of customization, ARM partners can tweak the Cortex-A50 generation core to apply it to SoC (System-on-Chip) chips for smartphones, tablets, PCs and even more. Lice are in servers. The A53 in the Cortex-A50 deliver an experience of half the power consumption of previous generations. The Cortex-A53 is also "the world's smallest 64-bit

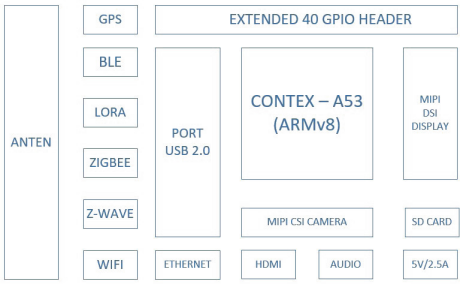


Fig. 5. The general block diagram of gateway devices

processor" to save the necessary space, so that manufacturers can create smaller, thinner smartphones and tablets. Thanks to the ARMv8 64-bit architecture, 64-bit computing will help the CPU calculate faster and manage a larger amount of the RAM memory, especially when performing heavy tasks.

With 40 extended GPIO pins in raspberry, connecting external modules is easy with full power and signal pins. There are also 4 USB 2.0 ports for connecting modules via usb port or connecting accessories such as keyboard, mouse, etc. The display of raspberry also has many ways. We can use an HDMI cable to connect to a big screen or we can use a small screen for raspberry pi through MIPI CSI. There is also one way that developers use it as remote access raspberry. This is done through an enthernet connection with a public IP range, to make this connection, make sure you have SSH enabled and use the same network to remote access the device. The archiving as well as the operating system for the raspberry pi is stored in the SD Card. The receiver modules of the radio protocols connect to the raspberry Pi through the GPIO pins or the USB jacks of the Raspberry Pi. The six wireless protocols used including Bluetooth Low Energy (BLE), WiFi, LoRa, ZigBee, Z-wave, and 4G cellular networks. With these wireless protocols, we can select the appropriate protocol for each environment location based on the transmission distance between the devices and the gateway.

2.3 IoT Cloud

After the gateway receives the data packets sent to each device, it send data to Cloud via MQTT protocol [10,11]. MQTT is a publish/subscribe protocol used for IoT devices with low bandwidth, high reliability and ability to be used in an unstable network. In a system using the MQTT protocol, multiple end device nodes (called the MQTT client - client) connect to an MQTT server (called the broker). Each client will subscribe to several channels (topic), for example "/ client1 / channel1", "/ client1 / channel2". This subscription process is called a "subscription". We subscribe to news on a Youtube channel. Each client will receive data when any other station sends registered data and channel. When a client sends data to that channel, it is called "publish". Cloud computing is

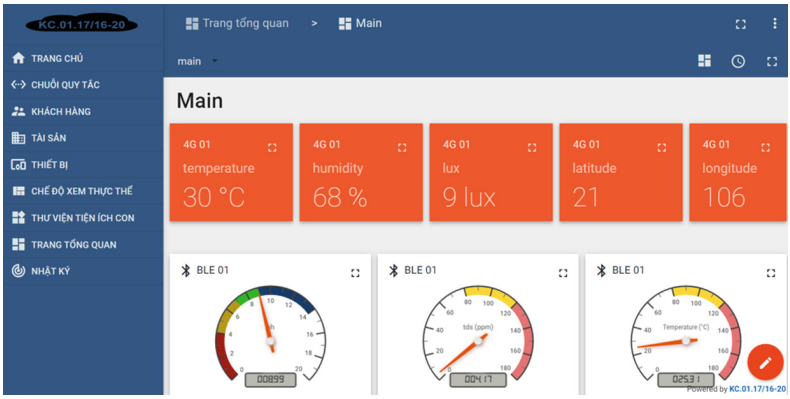


Fig. 6. Management interface of the open cloud platform

a solution that provides comprehensive information technology services in the Internet environment. There, resources will be provided, shared like the distribution line on the grid. Computers using this service run on a single system. That is, they will be configured to work together, different applications using the aggregated computing power. Cloud works in a completely different way from physical hardware. Cloud computing allows users to access servers, data, and Internet services. The cloud service provider owns, manages the hardware, and maintains the network connection. Meanwhile, users will be provided what they use through the web platform. Currently, there are four main cloud deployment models that are in common use including Public Cloud, Private Cloud, Hybrid Cloud and Community Cloud. There are many organizations and corporations that have been developing IoT standards, in which the OneM2M initiative aims to develop specifications that meet the needs of the general M2M Service layer [12]. Applications can be built using oneM2M-enabled devices sourced from multiple vendors. This allows the solution provider to build once and reuse it. This is a significant advantage in the lack of standards that restrict permutation between multiple technology and service providers, organizational boundaries, and IoT applications. The architecture standardized by oneM2M defines the Service layer IoT, i.e. the middleware between processing/communication hardware and IoT applications providing a rich set of functions needed for many IoT applications. OneM2M supports secure end-to-end data/control exchange between IoT devices and custom applications by providing functions for appropriate identification; authentication, encryption, remote licensing and activation, connection establishment, buffer, planning, device management.

In this work, we employ Thingsboard as cloud (Fig. 6). It is an open source IoT platform, allows for rapid development, management, and expansion of IoT projects. Thingsboard platform allows to collect, process, visualize and manage end devices. In addition, ThingsBoard allows the integration of end devices connected to legacy and third party systems using existing protocols. Connect to

OPC-UA (Open Platform Communication-Unified Architecture) server, MQTT broker by connecting via IoT Gateway. Thingsboar supports reliable remote data collection and storage. The collected data can be accessed using a custom web dashboard or server-side APIs [13].

2.4 User Applications

Web, apps help information get closer to users [14]. The information is displayed visually in the form of numbers and graphs in real time. Web applications, apps will use APIs to exchange data with the cloud and get parameters from the cloud, process those parameters and deliver them to users. With the measured data and stored in the cloud, the authors calculated the Air Quality Index (AQI) to warn about the air quality based on the international standard scale used.

3 Experimental Results

3.1 Experimental Setup

The designed system has been setup and tested at the Sai Dong B Industrial Park, Hanoi. Firstly, we drag, wire, install electrical equipment for computers, display screens, security cameras and WiFi sources. Then, dragging, wiring, installing wifi network system and setting up IP camera to have stable network source, security camera stream to other platforms smoothly, as shown in Fig. 7.



Fig. 7. The designed and implemented hardware prototypes of IoT gateway and end devices for the air and wastewater monitoring and surveillance of industrial zones

Next, we check the stability and safety of the electrical network as well as the Internet that has just been installed in the industrial park. We conduct division and survey of the places where the intended measuring equipment will be installed to divide the equipment to suit the measuring environment of those

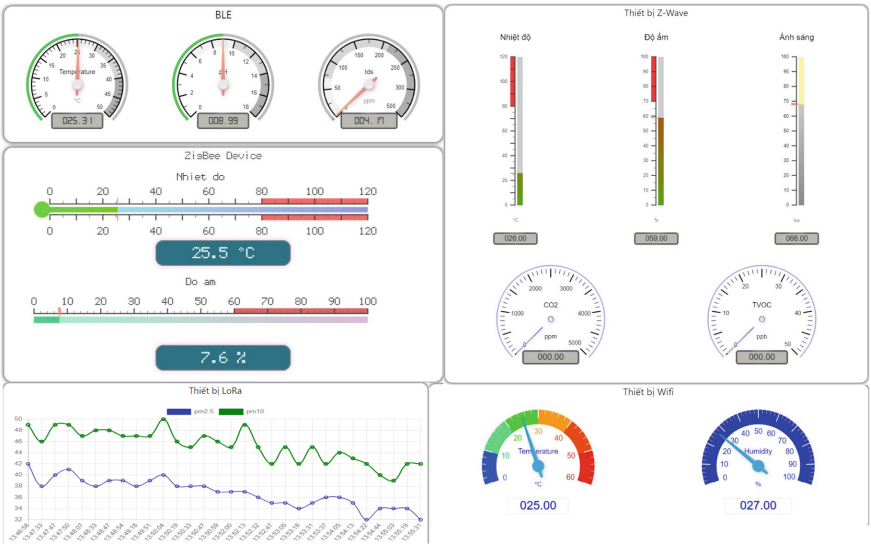


Fig. 8. The dashboards of air and wastewater parameters monitored in industrial zones

equipments. Then, install the Gateway at selected locations and establish wireless connections between the gateway and other IoT devices. Then running the gateway via SSH (Secure Shell), the gateway send parameters to the Server. Finally, we fix some errors that did not occur during the experiment at school. These errors are not very serious, so the repair time is not long.

It can be seen in Fig. 8 that when using one gateway and the PM 2.5 dust concentration sensor is located in an air-conditioned room, the PM2.5 dust concentration in the air is extremely low. On the other hand, Z-Wave device has the function of measuring parameters of CO_2, VOC (volatile organic compounds), Temperature, Humidity and Light for outdoor, so the indexes reflect different accurately the environmental situation. The school takes place in the industrial zone: high temperature, low humidity, high light intensity, CO_2 and VOC concentrations are very high compared to the normal threshold.

With the water environment, although according to theory, the test team will test at the wastewater environment. However, because this is a wastewater treatment plant with a closed process, along with that to ensure the safety of students as well as teachers, the management board does not allow the delegation to access the water area. Factory waste. The installation team can only take measurements at the treated water tank to the near final step. The results were very positive. Because the water tank is continuously pumped with fresh water from the pipeline, the water tank is less affected by the surrounding air. The result of a water temperature of 29.19^0C reflected this. The pH of the water is also close to the ideal of 6.38 and the amount of solids dissolved in the water is 0, possibly due to errors from the hardware, the results are not as expected.

4 Conclusion

We have presented in this paper the IoT framework system with the current situation of environmental pollution for air and wastewater management in industrial and high-tech zones. The system design focuses mostly on providing PaaS services to support the management of industrial zones companies. We have implemented the proposed system and shown that implementing dash-board surveillance for problem reporting in conjunction with the open platform system and dynamic routing models can give a significant increase of cost effectiveness. It is essential to apply in the management of industrial parks to continuously control and monitor the discharge of industrial park infrastructure investors, minimizing negative impacts on the surrounding living environment. Around the industrial park, saving energy, ensuring the lives of workers in the industrial park.

References

1. Ericsson, Ericsson Mobility Report (November 2016). https://www.ericsson.com/mobility-report
2. Akshay, L., Perkins, E., Contu, R., Middleton, P.: Gartner, Forecast: IoT Security, Worldwide (2016). Strategic Analysis Report No-G00302108. Gartner, Inc.
3. Rathore, M.M., Ahmad, A., Paul, A., Rho, S.: Urban planning and building smart cities based on the internet of things using big data analytics. Comput. Netw. **101**, 63–80 (2016)
4. Anagnostopoulos, T., Zaslavsky, A., Kolomvatsos, K., Medvedev, A., Amirian, P., Morley, J., et al.: Challenges and opportunities of waste management in IoT-enabled smart cities: a survey. IEEE Trans. Sustain. Comput. **2**, 275–289 (2017)
5. Khanna, A., Kaur, S.: Evolution of Internet of Things (IoT) and its significant impact in the field of precision agriculture. Comput. Electron. Agric. **157**, 218–231 (2019)
6. Qiu, X., Luo, H., Xu, G., Zhong, R., Huang, G.Q.: Physical assets and service sharing for IoT-enabled Supply Hub in Industrial Park (SHIP). J. Prod. Econ. **159**, 4–15 (2015)
7. Ngu, A.H., Gutierrez, M., Metsis, V., Nepal, S., Sheng, Q.Z.: Iot middleware: a survey on issues and enabling technologies. IEEE Internet Things J. **4**(1), 1–20 (2017)
8. Khanna, A., Kaur, S.: Internet of Things (IoT), applications and challenges: a comprehensive review. Wirel. Pers. Commun. **114**(2), 1687–1762 (2020). https://doi.org/10.1007/s11277-020-07446-4
9. Madakam, S., Ramaswamy, R., Tripathi, S.: Internet of Things (IoT): a literature review. J. Comput. Commun. **3**(05), 164 (2015)
10. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun. Surv. Tutor. **17**(4), 2347–2376 (2015)
11. Whitmore, A., Agarwal, A., Da, X.L.: The Internet of Things-a survey of topics and trends. Inf. Syst. Front. **17**(2), 261–274 (2015)
12. Trung, H.D., Hung, N.T., Trung, N.H.: Opensource based IoT platform and LoRa communications with edge device calibration for real-time monitoring systems. In: ICCSAMA, pp. 412–423 (2019)

13. Trung, H.D., Dung, N.X., Trung, N.H.: Building IoT analytics and machine learning with open source software for prediction of environmental data. In: HIS, pp. 134–143 (2020)
14. Abou-Zahra, S., Brewer, J., Cooper, M.: Web standards to enable an accessible and inclusive internet of things (IoT). In: 14th Web for All Conference on the Future of Accessible Work, vol. 9, pp. 1–9:4 (2017)



A Comparison of YOLO Networks for Ship Detection and Classification from Optical Remote-Sensing Images

Ha Duyen Trung(✉)

School of Electronical and Electronic Engineering (SEEE), Hanoi University of Science and Technology (HUST), No. 1, Dai Co Viet St, Hanoi, Vietnam
trung.haduyen@hust.edu.vn

Abstract. The waterway traffic is recently getting busier due to the strong development of the shipping industry. There are frequent collisions and other accidents between ships, it is necessary to detect these types of ships effectively to ensure waterway traffic safety. Ship detection technology based on computer vision employing optical remote sensing images has great significance to improve port management and maritime inspection. In recent years, convolutional neural networks (CNN) have achieved good results in ship target detection and recognition. In this paper, we train the YOLOv3 and the latest YOLOv4 model on the dataset. The experimental results show that YOLOv4 can be applied well in the field of ship detection and classification from optical remote sensing. Based on the obtained results, we compare the effectiveness of the models when applied to actual training on the same data set.

Keywords: YOLO Networks · Detection · Classification · Remote Sensing · Images Processing

1 Introduction

The science of remote sensing is growing, space agencies have deployed many satellites to orbit the earth. From there, it provides a large amount of information, remote sensing image data for research activities and applications to our lives. The need to apply artificial intelligence (AI) to remote sensing is also increasing, the development of automatic analysis models is the current and future trend and goal. The detection and classification of ships automatically based on satellite image data will partly help in the search and help cases at sea, as well as in the protection of national authority owners, navigating ships that enter the territory illegally.

The Automatic Identification System (AIS) was born in December 2004 by the International Maritime Organization (IMO) and the International Convention for the Safety of Life at Sea (SOLAS). All ships with a gross tonnage of 300 GT or more engaged in international transport, cargo ships with a gross tonnage of 500 GT or more engaged in inland and coastal transport, passenger ships must be equipped with AIS [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 43–52, 2023.
https://doi.org/10.1007/978-3-031-27409-1_5

Traditional ship detection methods are based on automatic identification system and ship features [1, 2]. Li et al. propose an improved dimensional spatial clustering algorithm to identify anomalous ship behavior [3]. Zhang et al. used AIS data to identify ships attempting to collide [4, 5]. Zhou et al. proposed a detection method to classify and identify the bow [6]. Zang et al. perform ship target detection from an unstable platform [7, 8]. Although these studies have achieved good results, there are generally problems such as low recognition accuracy and human interference. Therefore, it is difficult for the traditional ship detection method to achieve the ideal detection effect.

Recently, two-stage detection method and one-stage detection method are used to solve the problem of target detection through deep learning. The two-stage algorithm using region suggestion detection, mainly include AlexNet [9], VGG [10], ResNet [11], Fast-RCNN [12] and FasterRCNN [13]. Although the detection accuracy is better than the traditional method, the detection speed is slightly inadequate, the feature extraction process takes a long time, and it is difficult to achieve real-time detection efficiency. To ensure accuracy and improve detection speed, the proposed one-stage algorithm, one-stage detection does not use the idea of combining fine detection with coarse detection, but directly detects results in a single stage. The whole process does not need to detect the region hint and directly performs the end-to-end detection to import the image, the detection speed is greatly improved by the single-stage algorithm mainly consisting of SSD [14], YOLO [15], YOLOv2 [16], and YOLOv3 [17]. Most recently, Huang et al. propose an improved YOLOv3 network for intelligent detection and classification of ship images and videos [18, 19]. YOLOv5 based deep convolutional neural networks for vehicle recognition in smart university campus has been reported in [20]. However, the comparison of YOLO networks for ship detection and classification has not been reported in the literature.

The objective of the paper is determined to build and develop a model capable of detecting various kinds of ships through optical remote-sensing images employing artificial intelligence algorithms. More specifically, instead of having to use the usual visual inspection method, users can use this algorithmic model to accurately detect the coordinates of each type of ships with more accuracy, confidence, and high trust. From the above direction, we have planned the work to be done in this paper (i) Finding and processing data sets of remote sensing images of ships; (ii) Applying machine learning models to ship identification and classification; (iii) Compare the results obtained from different models to select the best model capable of detecting and classifying ships with high accuracy, within the scope of the study.

This paper is organized as follows. In Section II, we introduce the details of the YOLOv3 and YoLOv4 networks. In Section III, we conduct system implementation qualitatively and quantitatively. Experimental results are presented in Section III to evaluate the performance of the proposed method. Section IV concludes this paper.

2 Background

The structure of Yolov3 in Fig. 1 is an input image whose default size is $416 \times 416 \times 3$ and will be included in a backbone layer that is responsible for creating features from the input image to identify the features of the object. Then these feature classes will be

passed through the next processing layer to give the result the absolute coordinates of the object and the probability that the object is an object in the classes specified in data set.

YOLOv4 has many special enhancements that increase accuracy and speed over YOLOv3. According to [15], the structure of Yolov4 consists of 3 main parts: Backbone uses CSPDarknet53, neck uses SPP, PAN and head is the YOLOv3 (see Fig. 2).

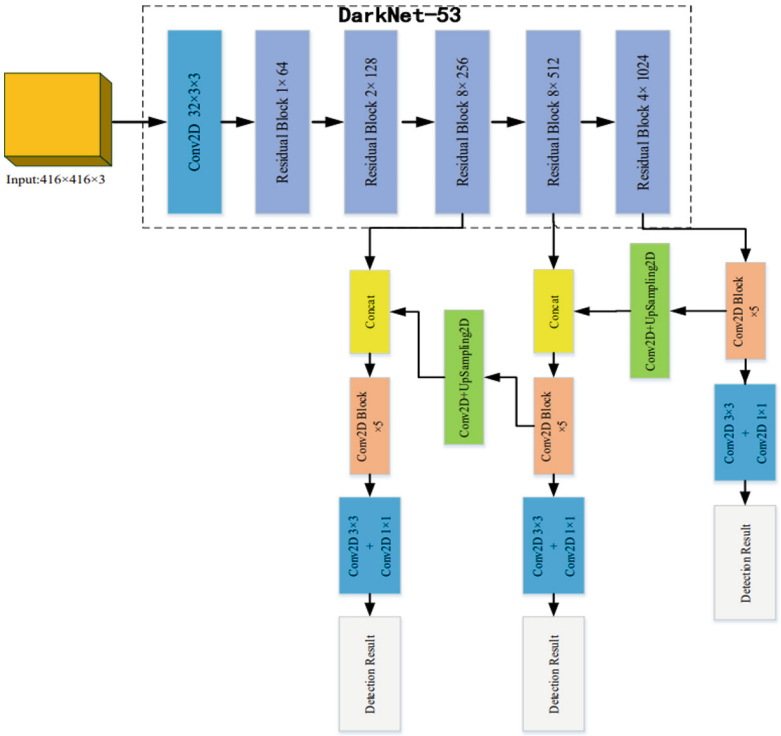


Fig. 1. Yolov3 architecture [7].

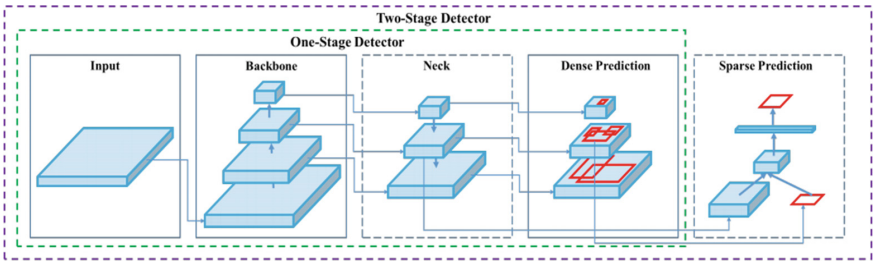


Fig. 2. YOLOv4 architecture.

3 System Implementation

3.1 Dataset

In this paper, the data used are optical images (RGB images) taken from above of boats at sea. In which, the dataset has more than 200,000 photos taken from above. Each image file is 768 × 768 × 3. All data is provided by kaggle – a site dedicated to organizing AI competitions and providing AI platforms and data. (https://www.kaggle.com/c/airbus-ship-detection/data). However, the data that the contest provides has many problems such as: too many photos without the boat; this is the data of the segmentation problem, not the detection problem, so the data will be different (not bounding box); and the data is labeled with the same label as the ship, so it cannot be distinguished. That's why we preprocessed the data to get a clean set of data to put into the train model.

For the dataset used for training, image processing without ships will reduce the training efficiency. To solve this problem, we use a script that uses the Pandas library to read the set of images used for training, categorize the images without objects (e.g., images with only sea surface) and remove them. Specifically, in the file train_ship_segmentations_v2.csv (this is the file containing information for masking the segmentation problem), images with boats will have column values "EncodedPixels" other than "NaN". Thereby we can filter images with ships for further processing.

The second problem is the data set from Kaggle used to serve the Image Segmentation problem. Although they are both aiming for the same goal with Object detection, which is to detect object containers, the output of the two articles is still very different. Therefore, we need to convert the data from segmentation to detection. Encoded Pixel is a type of data listed out in csv file used to replace masked image (masked image). This is a way to store labels for the Segmentation problem in a memory-optimized way (see Fig. 3).

	ImageId	EncodedPixels
0	00003e153.jpg	NaN
1	0001124c7.jpg	NaN
2	000155de5.jpg	264661 17 265429 33 266197 33 266965 33 267733...
3	000194a2d.jpg	360486 1 361252 4 362019 5 362785 8 363552 10 ...
4	000194a2d.jpg	51834 9 52602 9 53370 9 54138 9 54906 9 55674 ...

	ImageId	Ship	Boundingbox
0	00003e153.jpg	0	NaN
1	0001124c7.jpg	0	NaN
2	000155de5.jpg	1	[344, 465, 447, 501]
3	000194a2d.jpg	1	[469, 287, 491, 307]
4	000194a2d.jpg	1	[67, 377, 84, 386]
...	...	...	...
231718	fffedbb6b.jpg	0	NaN
231719	ffff2aa57.jpg	0	NaN
231720	ffff6e525.jpg	0	NaN
231721	ffffc50b4.jpg	0	NaN
231722	fffe97f3.jpg	0	NaN

Fig. 3. Data transformation to bounding box.

Assuming that an image is 768 × 768, the number of pixels in the image will be 589824 pixels. With a data of the form 1235 2 1459 1 5489 10 … that means taking 2 pixels starting from the 1235th pixel, taking one pixel starting from the 1459th pixel and taking 10 pixels starting from the 5489th pixel … and then it becomes 1235, 1236,

1459, 5489, 5490, 5491, 5492, 5493, 5494, 5495, 5496, 5497, 5498. From there we can determine the coordinates of the pixel along the x and y axes. And to get the data for the detection problem, we took the coordinates `(xmin, ymin) and (xmax, ymax)`.

Since the data set has not been labeled, we need to classify the types of ships for the classification problem along with the specific identification characteristics of each type, specifically as follows: (1) Cargo ship with identification of many squares, (2) tanker with identification of smooth and other ships. With the predefined bounding box coordinates [xmin, ymin and xmax, ymax], the result after hitting the bounding box will be as in the Fig. 4.

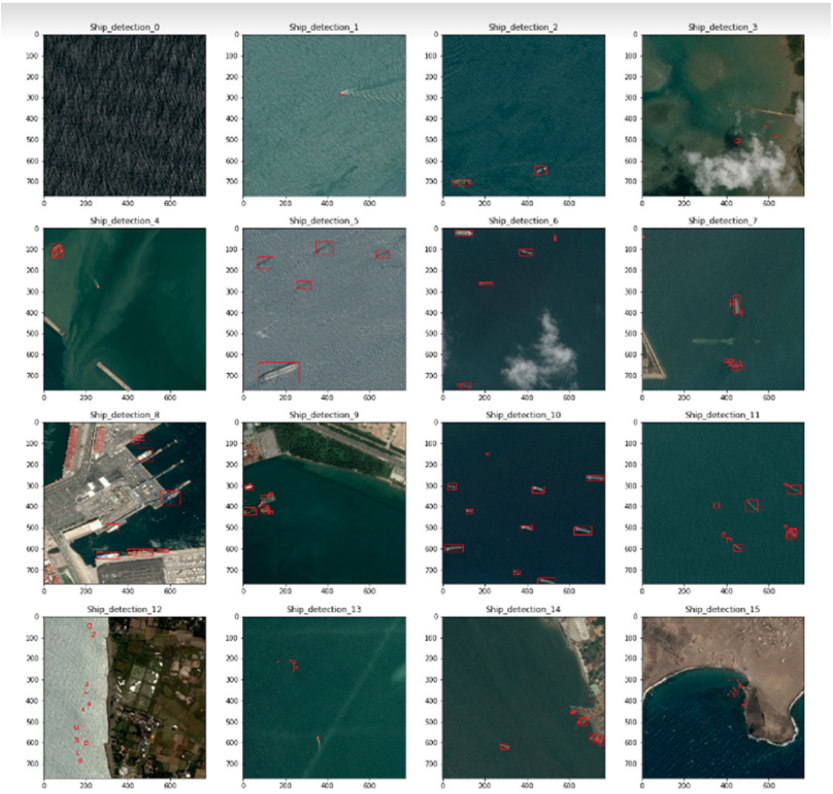


Fig. 4. Results of bounding box.

3.2 Implementation

In the implementation, three times with three image sizes of 256 × 256, 512 × 512 and 768 × 768 were used for training. Based on basic theory, the larger the image size, the more information the model will extract through the Convolution layers, so the accuracy will be higher with the larger the image size, but the processing speed of the image will be slower.

Loss function: The function calculates the difference between the output of the model (Prediction) and the actual result (Ground Truth). The smaller the difference, the better the model.

Optimal function: The function will calculate and optimize so that the difference of the model is minimal or in other words, optimize the loss function. One epoch will be one time the model is trained through all images in the image list with the path defined in the file train.txt, the more times the model is trained (the more accurate the prediction). When the model learned to a certain point will no longer learn, the loss function will give a constant value (saturation). Images will be included for batch learning with the number of images of a set defined as batch size.

Basic parameters of model training include the Epoch of 6000, the batch size of 64, the Optimal function is Adam, the Learning Rate of 0.001, the Momentum of Yolov3 - 0.9 [19].

3.3 Training Results

Results achieved for YOLOv3 after training:

The larger the model input image size, the longer the processing time (2 s → 10 s) (see Fig. 5).

With image size 512 × 512 gives much better results than 256 × 256 size.

The 768 × 768 image size results in only slightly better than 512 × 512. That proves that 512 × 512 is the most optimal image size that the Model can handle, i.e., if it is larger than 512 × 512, the results will not be improved.

With the lower the object prediction threshold (IoU Threshold) (0.75 → 0.5 → 0.25), the higher the mAP index. The reason is because with a low threshold the percentage of boxes appears more (there are more predictions about the position of the object), from which TP, FP and FN will increase, leading to an increase in mAP.

Because three image sizes of 256 × 256, 512 × 512 and 768 × 768 were trained using YOLOv3 and YOLOv4 network models for comparison purposes. We chose the size 512 × 512 because these two models provide the best results in the size. Some parameters were used such as compute-val-loss, batch-size 1, random-transform, epochs of 50 k, steps of 1000, and lr of 1e−3. Image result obtained as shown in Fig. 6.

Boats are detected but the accuracy rate is relatively low, sometimes there are cases where two boxes receive the same object or 1 box recognizes 2 objects close to each other. To improve this case, it is necessary to have more accurate data for the model to learn better.

3.4 Comparison Analysis

There are many model evaluation parameters, however the main parameter of mAP (mean Average Precision) is used in this paper, that can be expressed as [19, 20]

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall(TPR) = \frac{TP}{TP + FN} \tag{2}$$



Fig. 5. Ship detection using the YOLOv4.

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

$$AP@n = \frac{1}{GTP} \sum_{k}^{n} P@k \times rel@k. \tag{4}$$

In Eq. (4), GTP is the ground truth positive, P@k is the precision@k, and rel@k is the relevance function and selected by "0" and "1". Finally, mAP is defined as

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i. \tag{5}$$

Regarding the loss function of the models, Yolov3 and Yolov4, because they both use the Adam optimization function with the same Lr, the convergence is almost the same. Having the same starting point is because they both take pretrain from the previously trained data set. Finally, because Yolov4 has the part is more complex than Yolov3, the optimization takes longer than a few dozen epochs and Yolov3 will be easier to achieve convergence than Yolov4. The two methods seem to converge, but when looking at the last epoch, it is the difference(see Fig. 7 and Table 1).

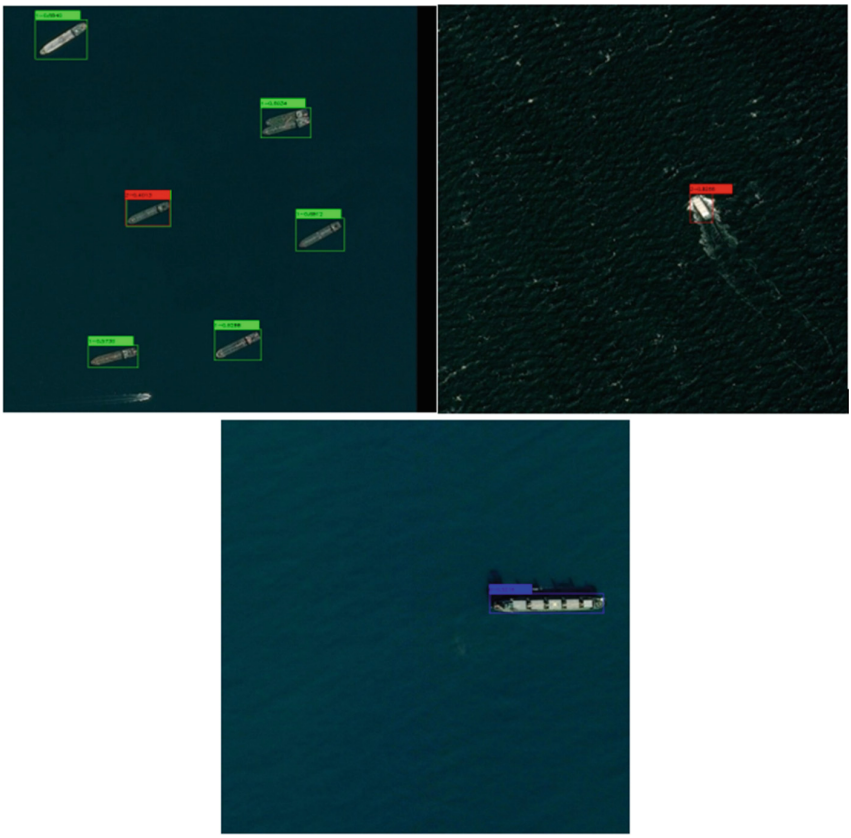


Fig. 6. EfficientDet result.

Table 1. Comparison of mAP for three models of YOLOv3, YOLOv4 and EfficientDet-D4.

	mAP25	mAP50	mAP75
YOLOv3	0.7636	0.74230	0.4654
YOLOv4	0.8344	0.8294	0.6032
EfficientDet-D4	0.5736	0.5635	0.4956

Yolov3 and Yolov4 Since they both use the Adam optimal function with the same Lr, the convergence is almost the same. The fact that they have the same starting point is because they all take pretrains from the previously trained data set. Finally, because Yolov4 is somewhat more complicated than Yolov3, the optimization takes longer than a few dozen epochs and Yolov3 will be easier to achieve convergence than Yolov4. The two lines seem to converge, but when you look at the last epoch, you will see the difference (see Fig. 8).

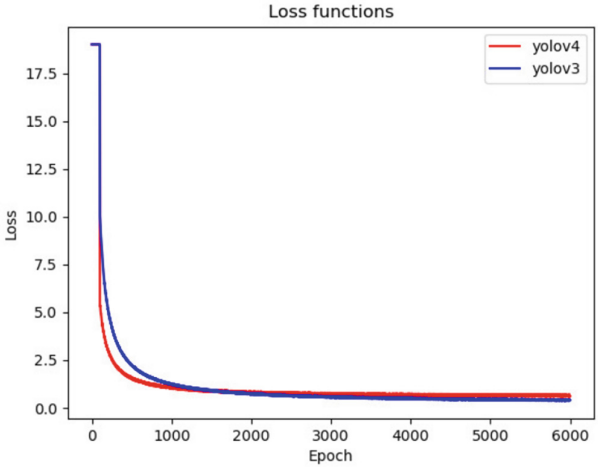


Fig. 7. Loss function of YOLOv3 and YOLOv4.

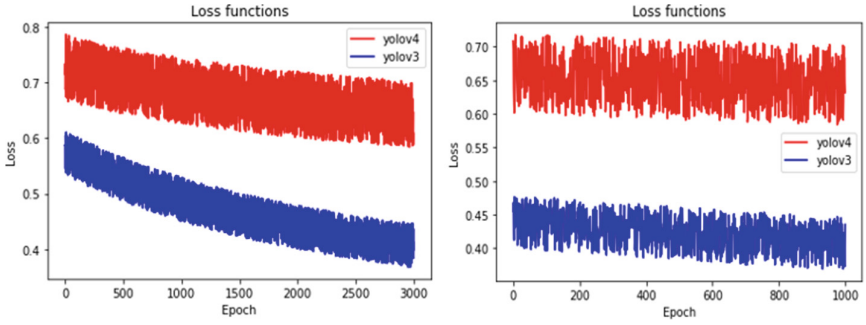


Fig. 8. Loss versus epoch for two models of YOLOv3 and YOLOv4.

4 Conclusions

This paper has applied a machine learning models to remote sensing image processing for the comparison of YOLO networks. More specifically, this paper provided general building models to help identify and classify ships on optical remote-sensing image data. Based on the experimental results, we compare the effectiveness of the two YOLO models when applying to actual training on the same data set.

References

1. Wang, J., Zhu, C., Zhou, Y., Zhang, W.: Vessel spatio-temporal knowledge discovery with AIS trajectories using coclustering. J. Navig. **70**(6), 1383–1400 (2017)
2. Bye, R.J., Aalberg, A.L.: Maritime navigation accidents and risk indicators: an exploratory statistical analysis using AIS data and accident reports. Reliab. Eng. Syst. Saf. **176**, 174–186 (2018)

3. Li, H., Liu, J., Wu, K., Yang, Z., Liu, R.W., Xiong, N.: Spatio-Temporal vessel trajectory clustering based on data mapping and density. IEEE Access **6**, 58939–58954 (2018)
4. Zhang, W., Goerlandt, F., Montewka, J., Kujala, P.: A method for detecting possible near miss ship collisions from AIS data. Ocean Eng. **107**, 60–69 (2015)
5. Luo, D., Zeng, S., Chen, J.: A probabilistic linguistic multiple attribute decision making based on a new correlation coefficient method and its application in hospital assessment. Mathematics **8**(3), 340 (2020)
6. Li, S., Zhou, Z., Wang, B., Wu, F.: A novel inshore ship detection via ship head classification and body boundary determination. IEEE Geosci. Remote Sens. Lett. **13**(12), 1920–1924 (2016)
7. Zhang, Y., Li, Q.-Z., Zang, F.-N.: Ship detection for visual maritime surveillance from non-stationary platforms. Ocean Eng. **141**, 53–63 (2017)
8. Zeng, S., Luo, D., Zhang, C., Li, X.: A correlation-based TOPSIS method for multiple attribute decision making with single-valued neutrosophic information. Int. J. Inf. Technol. Decis. Mak. **19**(1), 343–358 (2020)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: The International Conference on Neural Information Processing Systems, pp. 1097–1105 (2012)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015). http://arxiv.org/abs/1409.1556.
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision (2015)
13. Ren, S. He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2015)
14. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
15. Redmon, J., Divvala, S., Girshick, R., Arhadi, A.F.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
16. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517–6525 (2017)
17. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement (2018). http://arxiv.org/abs/1804.02767
18. Huang, Z.S., Wen, B., et al.: An intelligent ship image/video detection and classification method with improved regressive deep convolutional neural network. Complexity **1520872**, 11 (2020)
19. Hao, L., Deng, L., Yang, C., Liu, J., Gu, Z.: Enhanced YOLOv3 tiny network for real-time ship detection from visual image. IEEE Access **9**, 16692–16706 (2021)
20. Tra, H.T.H., Trung, H.D., Trung, N.H.: YOLOv5 based deep convolutional neural networks for vehicle recognition in smart university campus. In: Abraham, A., et al. (eds.) HIS 2021. LNNS, vol. 420, pp. 3–12. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96305-7_1



Design and Implementation of Transceiver Module for Inter FPGA Routing

C. Hemanth, R. G. Sangeetha(✉), and R. Ragamathana

Vellore Institute of Technology, Chennai, Tamil Nadu, India
{Hemanth.c,Sangeetha.rg}@vit.ac.in

Abstract. A Universal Asynchronous Receiver Transmitter (UART) is frequently used in conjunction with RS 232 standard, which sends parallel data through a serial line. The transmitter is essentially a special shift register that loads data in parallel and then shifts it out, bit by bit at a specific rate. The receiver, on the other hand, shifts in data bit by bit and then re-assembles the data. UART is implemented using FPGA by considering two Development and Education boards where each has a transceiver module. Bidirectional routing is established using RS 232 interface to communicate the two transceiver modules. This is designed and implemented using Quartus and Cyclone IV FPGA. The total power of the transceiver module using Cyclone IV is analyzed and compared with that of the transceiver implemented using different FPGAs.

Keywords: UART · Transceiver · Routing · Transmitter · Receiver · FPGA

1 Introduction

Transceiver is a combination of a transmitter and a receiver. It is a single package that transmits and receives analog or digital signals. Universal Asynchronous Receiver Transmitter (UART) is a hardware device for Asynchronous serial communication. UART is widely used, since it is one of the simplest serial communication techniques. It is used in many fields such as GPS receivers, GSM modems, Bluetooth modules, GPRS systems, wireless communication systems and RF applications.It is commonly used in conjunction with communication standards RS-485, RS-422 or RS-232.

UART converts both the incoming and outgoing signal into a serial binary stream. The transmitting UART converts the parallel data that is received from external devices such as CPU, into serial form by using parallel to serial converter. The receiving UART on the other hand converts the serial data back into parallel form by using serial to parallel converter. In UART communication, the data flows from the Tx pin of the transmitting UART to the Rx pin of the receiving UART. Similarly, the received data flows back from the Tx pin of the receiving UART to the Rx pin of the transmitting UART. In UART, there is no clock signal, which means the output bits from the transmitting UART are not synchronized with the sampling bits of the receiving UART. Since the communication is done asynchronously, instead of clock signal, start and stop bits are added to the transferred data packet by the transmitting UART. These start and stop bits define the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 53–62, 2023.
https://doi.org/10.1007/978-3-031-27409-1_6

starting and ending of the data packet, so that the receiving UART knows when it has to start reading the bits. The receiving UART starts to read the incoming bits, once it detects the start bit at a specific frequency known as baud rate. The working of UART is explained in [1, 2].

The measure of the speed of the data transfer is called the baud rate. It is expressed in bits per second (bps). Data is transferred to transmitting UART by the data bus from any external devices such as CPU in parallel form. Once the transmitting UART gets the parallel data from the data bus, it adds start bit, parity bit and stop bit to it, thus creating a data packet. The Tx pin of the transmitting UART transmits the data packet serially. The Rx pin of the receiving UART reads the data packet bit by bit. The receiving UART then converts the serial data back into parallel form and also segregates the start, parity and stop bits. The receiving UART then transfers the parallel data to the data bus on the receiving end.

Altera's Cyclone IV E is a FPGA that operates with a core voltage of 1.2 V and works under the ambient temperature of 50 °C. Cyclone IV E offers low power, high functionality and low cost. It works on a device EP4CE115 of package FBGA, with pin count 780 and a speed grade −7. It has 114480 LEs (Logic Elements), 529 user I/Os, 532 9-bit embedded multipliers, 4 PLLs and 20 Global clocks. The Parameters of Cyclone IV E is shown in Table 1

Table 1. Parameters of Cyclone IV E

Family	Cyclone IV E
Device	EP4CE115
Package	FBGA
Pin count	780
Speed grade	−7

In [3], researchers implemented the UART using different nanometer FPGA boards viz., Spartan-6, Spartan-3 and Virtex-4. In this paper, UART is implemented using Cyclone IV FPGA, a serial communication is established between two FPGAs using RS-232 interface and the obtained results are compared with the produced results of different nanometer FPGA boards.

2 UART Transmitter Module

The UART Transmitter is a special shift register which gets data from the external devices in parallel form. The parallel data is given to the data bus, which in turn gives the data to the transmitter. After the transmitter gets the parallel data from the data bus, it adds start bit, parity it and stop bit to the data. Start bit, also known as the synchronization bit, is placed before the data. The inactive data transmission line is generally controlled at a high voltage level. But in order to start the data transmission, the UART transmission drags the high voltage level to low voltage level. For data transmission, the voltage level

is controlled at low. The UART observes the drop of voltage from high to low, starts understanding the data and starts the data transmission. Generally, there will be only one start bit.

Parity bit is also known as fault checking bit, which ensures the receiver whether it has received the data correctly. It is of two ranges, odd parity and even parity. The parity bit can be assigned to either 0 or 1 accordingly which makes the number of 1's either even or odd depending on the type of the parity. Stop bit is usually placed at the end of the data packet. It is exactly opposite to the function of the start bit. Usually it has two bits, but only one bit is utilized frequently. It stops the data transmission, thereby changing the voltage from low to high level. The rise from low voltage level to high voltage level is observed by the UART and the data transmission is stopped. The data frame contains the data to be transmitted and is 8 bits long. The structure of data packet is shown in Fig. 1.



Fig. 1. Structure of Data Packet

The RTL design of UART transmitter is carried out using Verilog and synthesized using Quartus. The details on the number of logic elements, combinational functions, logic registers etc. are shown in the flow summary in Fig. 2. The RTL Schematic of UART transmitter is shown in Fig. 3

Revision Name	transceiver1
Top-level Entity Name	async_transmitter
Family	Cyclone IV E
Device	EP4CE115F29C7
Timing Models	Final
Total logic elements	42 / 114,480 (< 1 %)
Total combinational functions	41 / 114,480 (< 1 %)
Dedicated logic registers	28 / 114,480 (< 1 %)
Total registers	28
Total pins	12 / 529 (2 %)
Total virtual pins	0
Total memory bits	0 / 3,981,312 (0 %)
Embedded Multiplier 9-bit elements	0 / 532 (0 %)
Total PLLs	0 / 4 (0 %)

Fig. 2. Flow summary of UART Transmitter

3 UART Receiver Module

The Receiver examines each bit and receives the data. It determines whether the bit is 0 or 1 for a particular time period. For Example, If the transmitter takes 2 s to transmit the bit, then the receiver will take 1 s to examine the bit, whether it is 0 or 1 and wait for 2 s before examining the next bit. When the stop bit is sent by the transmitter, the receiver stops examining and the transmission line becomes idle.

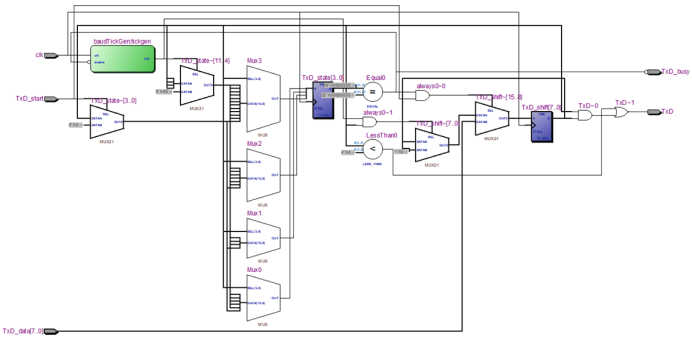


Fig. 3. RTL Schematic of UART Transmitter

Once the receiver receives all the bits it checks for parity bits. If no parity bit is available, the receiver encounters the stop bit. The missing stop bit may result in a garbage value which ultimately leads to framing error and it will be reported to the host processor. Framing error is due to the mismatches in the transmitter and receiver. The UART receiver discards the start, parity and stop bits automatically irrespective of the correctness of the received data. For the next transmission, the transmitter sends a new start bit after the stop bit for the previous transmission is sent.The UART Transmission and Reception is shown in Fig. 4.

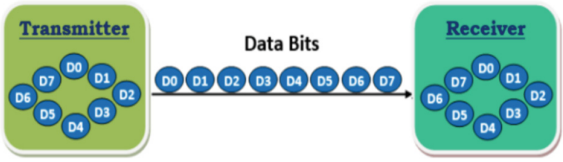


Fig. 4. UART Transmission and Reception

The RTL design of UART Receiver is carried out using Verilog and is synthesized using Quartus. The flow summary of UART Receiver is shown in Fig. 5. The RTL Schematic of UART Receiver is shown in Fig. 6.

Family	Cyclone IV E
Device	EP4CE115F29C7
Timing Models	Final
Total logic elements	50 / 114,480 (< 1 %)
Total combinational functions	50 / 114,480 (< 1 %)
Dedicated logic registers	25 / 114,480 (< 1 %)
Total registers	25
Total pins	11 / 529 (2 %)
Total virtual pins	0
Total memory bits	0 / 3,981,312 (0 %)
Embedded Multiplier 9-bit elements	0 / 532 (0 %)
Total PLLs	0 / 4 (0 %)

Fig. 5. Flow Summary of UART Receiver

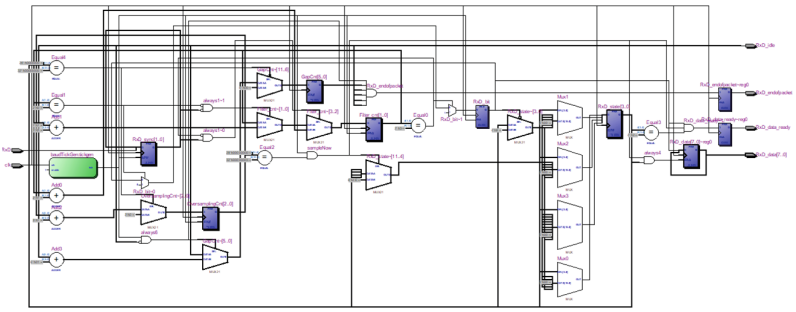


Fig. 6. RTL Schematic of UART Receiver

4 UART Transceiver Module

Transmitter and Receiver modules are combined together, in such a way that a single module transmits and receives the data simultaneously. The UART transceiver is a single package that has both transmitter and receiver modules [4, 5]. The block diagram of the UART transceiver module is shown in Fig. 7.

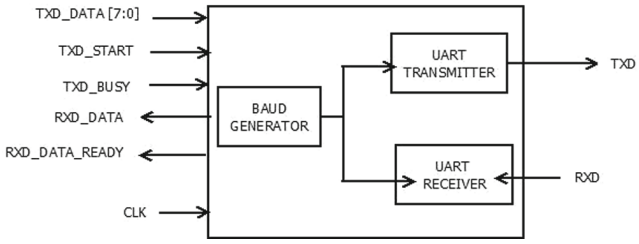


Fig. 7. UART Transceiver Module

In [6], the researchers have explained the working of transceivers in FPGA. The transceiver is synthesized using Quartus. The flow summary is shown in Fig. 8 and its RTL Schematic is shown in Fig. 9.

Family	Cyclone IV E
Device	EP4CE115F29C7
Timing Models	Final
Total logic elements	82 / 114,480 (< 1 %)
Total combinational functions	66 / 114,480 (< 1 %)
Dedicated logic registers	62 / 114,480 (< 1 %)
Total registers	62
Total pins	19 / 529 (4 %)
Total virtual pins	0
Total memory bits	0 / 3,981,312 (0 %)
Embedded Multiplier 9-bit elements	0 / 532 (0 %)
Total PLLs	0 / 4 (0 %)

Fig. 8. Flow Summary of UART Transceiver

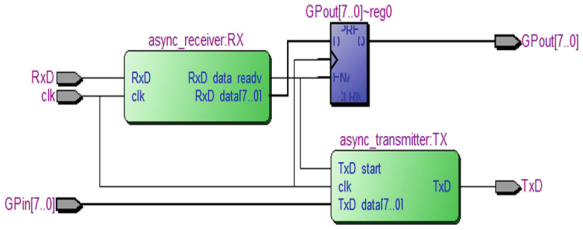


Fig. 9. RTL Schematic of UART Transceiver

The UART Transceiver is simulated using Modelsim and the Simulation Waveform is shown in Fig. 10.

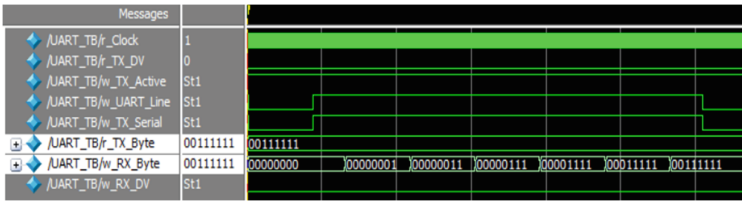


Fig. 10. UART Transceiver Simulation Waveform

The UART transceiver is designed with a clock frequency of 25 MHz and baud rate of 115200. Therefore, Clocks per bit is 217. The UART transmitter receives data, 00111111 from an external device and transmits it bit by bit through a serial data bus. The UART Receiver receives the data serially. When the receiver gets the stop bit, the transmission line becomes idle. The received data is rearranged by the receiver. The transmitted data, 00111111 is received at the receiver end.

5 Inter FPGA Routing

A Field Programmable Gate Array (FPGA) is an integrated circuit that has an array of programmable logic blocks and reconfigurable interconnects which makes the logic blocks to be wired together. Nowadays, FPGAs are widely used in the fields of Aerospace, Defense, Automotive, Electronics, IC Designs, Security, Video or Image Processing, wired and wireless communication etc.

Establishing a communication between two FPGAs is called the Inter FPGA Routing, shown in Fig. 11. It is widely used because it offers high execution speed, low cost and better testing experience. In [7] and [8], the researchers explained the need for Inter FPGA Routing. The different routing algorithms for Inter FPGA routing is in [9]. Communication between the FPGAs is done by Bidirectional Routing. Serial Interface RS 232 is used for routing the two FPGAs. RS 232 operates in full duplex mode.

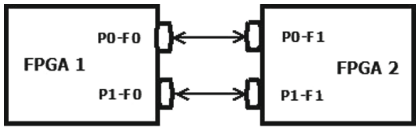


Fig. 11. Inter FPGA Routing

5.1 Working

Each FPGA is loaded with a transceiver module and a communication is established between the two transceiver modules. The transceiver module in the first FPGA is the transmitting UART and the one in the second FPGA is the receiving UART. Each UART has two pins, a Tx pin and a Rx pin. An 8-bit data is transmitted from the Tx pin of the transmitting UART to the Rx pin of the receiving UART. Similarly, the received data is transmitted back from the Tx pin of the receiving UART to the Rx pin of the transmitting UART as shown in Fig. 12.

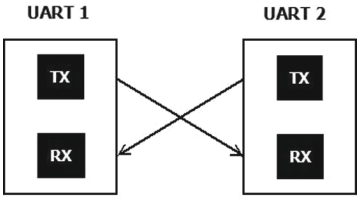


Fig. 12. UART Transceiver

RS 232 serial Interface is one of the simplest ways for the serial communication of data between two FPGAs. Two FPGAs are connected to each other through RS-232 Interface. On the transmitter side, it creates signal "TxD" by serializing the data to transmit and sends "busy" signal when the transmission is carried out, while on the receiver end it receives a signal "RxD" from outside the FPGA thereby de-serializing it for the easy use inside the FPGA. When the data is fully received, "data ready" is asserted.

This work is carried out by loading the SOF (system object file) of the transceiver module into two DE2 115 FPGA boards and these two FPGA boards are communicated with each other by RS 232 serial Interface.

6 Results and Discussion

The power Dissipation of the UART Transmitter, UART Receiver and UART Transceiver are shown in Figs. 13, 14 and 15. The power analyzer summary shows the dynamic,static,I/O and total power dissipation of the modules.

The total power dissipated by transmitter, receiver and transceiver are 0.133 W, 0.134 W and 0.144 W respectively under 1.2 V core voltage and 50 °C ambient temperature.The power analysis of transmitter, receiver and transceiver are shown in Table 2.

Family	Cyclone IV E
Device	EP4CE115F29C7
Power Models	Final
Total Thermal Power Dissipation	133.49 mW
Core Dynamic Thermal Power Dissipation	0.00 mW
Core Static Thermal Power Dissipation	98.47 mW
I/O Thermal Power Dissipation	35.02 mW

Fig. 13. Power Analysis of UART Transmitter

Family	Cyclone IV E
Device	EP4CE115F29C7
Power Models	Final
Total Thermal Power Dissipation	134.25 mW
Core Dynamic Thermal Power Dissipation	0.00 mW
Core Static Thermal Power Dissipation	98.47 mW
I/O Thermal Power Dissipation	35.77 mW

Fig. 14. Power Analysis of UART Receiver

Family	Cyclone IV E
Device	EP4CE115F29C8
Power Models	Final
Total Thermal Power Dissipation	144.72 mW
Core Dynamic Thermal Power Dissipation	0.00 mW
Core Static Thermal Power Dissipation	111.85 mW
I/O Thermal Power Dissipation	32.86 mW

Fig. 15. Power Analysis of UART Transceiver

The power dissipation comparison graph of UART transmitter, receiver and transceiver are shown in Fig. 16. From the graph, it can be clearly seen that the transceiver module dissipates more total power and static power when compared to transmitter and receiver. The transceiver module consumes less I/O power.

Table.2 Power Analysis

Cyclone IV E FPGA	Static power (Watt)	I/O power (Watt)	Total power (Watt)
UART transmitter	0.098	0.035	0.133
UART receiver	0.098	0.036	0.134
UART transceiver	0.111	0.032	0.144

It is observed that the Receiver dissipates 0.74% more power than that dissipated by the Transmitter and the Transceiver dissipates 7.4% more power than that dissipated by Receiver. The power consumed by the transceiver is more than that of transmitter and receiver since a communication is established between transmitter and receiver.

This work is carried out in Cyclone IV E at the ambient temperature of 50 °C. In [3], Keshav Kumar, et al. implemented UART using Virtex-4 and based on their results the total power dissipated by the transceiver using Virtex-4 is 0.177 W and the static power dissipated is 0.167 W under the same ambient temperature of 50 °C. The Power comparison of Virtex-4 and Cyclone IV is shown in Fig. 17.

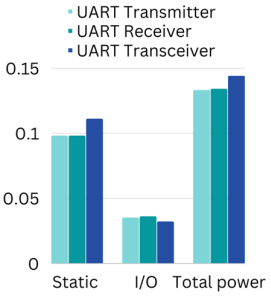


Fig. 16. Power Comparison of Cyclone IV

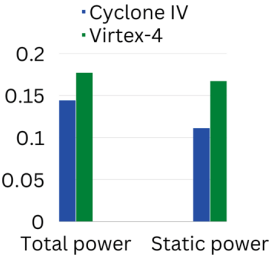


Fig. 17. Power Comparison of Virtex and Cyclone

The Power comparison between Cyclone IV E and Virtex-4, shows that Virtex-4 dissipates 22.9% more total power and 50.54% more static power than that dissipated by Cyclone IV E. From this comparison it can be seen that the UART transceiver implemented using Cyclone IV E consumes less power than that consumed by the transceiver implemented using Virtex-4. Therefore, Cyclone IV FPGA is a low power device.

7 Conclusion

The UART Transceiver module is designed in verilog and implemented using cyclone IV FPGA using Quartus. Serial communication is established between two FPGAs using RS-232 serial communication interface. From the simulation waveform of the transceiver in Fig. 10, it is seen that the data transmitted from the transmitting UART to the receiving UART and the data transmitted back from the receiving UART to the transmitting UART are the same. From the power comparison of Cyclone IV and Virtex-4 in Fig. 17, it is observed that Cyclone IV consumes 18.64% less power than that of Virtex-4, which shows that the UART transceiver implemented in Cyclone IV E dissipates less power when compared to Virtex-4 under the same ambient temperature of 50 °C according to the power analysis obtained from [3]. The UART transceiver module for Inter FPGA routing designed in this paper dissipates less power..

References

1. Nanda, U., Pattnaik, S.K.: Universal asynchronous receiver and transmitter (UART). In: 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, pp. 1–5 (2016)
2. Agrawal, R.K., Mishra, V.R.: The design of high speed UART. In: Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013) (2013). 978-1-4673-5758-6/13
3. Kumar, K., Kaur, A., Panda, S.N., Pandey, B.: Effect of different nano meter technology based FPGA on energy efficient UART design. In: 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, pp. 1–4 (2018)
4. Harutyunyan, S., Kaplanyan, T., Kirakosyan, A., Momjyan, A.: Design and verification of auto configurable UART controller. In: 2020 IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), pp. 347–350 (2020)
5. Gupta, A.K., Raman, A., Kumar, N., Ranjan, R.: Design and implementation of high-speed universal asynchronous receiver and transmitter (UART). In: 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 295–300 (2020)
6. Kumar, A., Pandey, B., Akbar Hussain, D.M., Atiqur Rahman, M., Jain, V., Bahanasse, A.: Frequency scaling and high speed transceiver logic based low power UART design on 45 nm FPGA. In: 2019 11th International Conference on Computational Intelligence and Communication Networks (CICN), Honolulu, HI, USA, pp. 88–92 (2019)
7. Farooq, U., Baig, I., Alzahrani, B.A.: An efficient inter-FPGA routing exploration environment for multi-FPGA systems. IEEE Access **6**, 56301–56310 (2018)
8. Farooq, U., Chotin-Avot, R., Azeem, M., Ravoson, M., Turki, M., Mehrez, H.: Inter-FPGA routing environment for performance exploration of multi-FPGA systems. In: 2016 International Symposium on Rapid System Prototyping (RSP), Pittsburgh, PA, pp. 1–7 (2016)
9. Tang, Q., Mehrez, H., Tuna, M.: Routing algorithm for multi-FPGA based systems using multi-point physical tracks. In: Proceeding of the International Symposium on Rapid System Prototyping (RSP), pp. 2–8 (Oct. 2013)

Intelligent Multi-level Analytics Approach to Predict Water Quality Index

Samaher Al-Janabi(✉) and Zahraa Al-Barmani

Faculty of Science for Women (SCIW), Department of Computer Science, University of Babylon, Babylon, Iraq
samaher@itnet.uobabylon.edu.iq

Abstract. In this paper will, building new miner called intelligent miner based on twelve concentrations to predict water quality called ($IM^{12}CP$-WQI). The main goal of that miner is to find water quality based on twelve types of concentrations that cause water pollution which is: Potential Hydrogen (PH), Total Dissolved Solids (TDS), Turbidity Unit NTU, Total Hardness (TH), Total Alkalinity, Calcium (Ca), Magnesium (Mg), Potassium (K), Sodium (Na), Chloride (Cl), Nitrogen Nitrate (NO3), and Sulfate (SO4). $IM^{12}CP$-WQI consists of four stages; the first stage related to data collection through two Seasons (i.e., summer & winter). The second stage, called pre-processing of data that include: (a) Normalization the dataset to make dataset in range (0, 1). (b) finding correlation between concentrations to know the direct or inverse correlation between those concentrations and their relationship with the water quality index WQI. The second stage involved building an optimization algorithm called DWM-Bat to find the optimum weight for each of the 12 compounds as well as the optimum number of M models for DMARS. The third phase involved building a mathematical model that combines these compounds, based on the development of MARS and drawing on the results of the previous stage, DWM-Bat. The last stage included the evaluation of the results obtained using three types of measures (R2, NSE, D) on the basis of which the value of WQI was determined based on that determined if the value of the WQI is less than 25, then it can be used for the purpose of drinking either between (26–50) it is used in fish lakes, as well as (51–75) it can be used in agriculture. Otherwise, it needs a refining process and reports are produced. Also, the results of the model ($IM^{12}CP$-WQI) were compared with the results of the models (MARS_Linear, MARS_poly, MARS sigmoid, MARS_RBF) under the same conditions and environment, finally; the results shown ($IM^{12}CP$-WQI) is pragmatic predictor of WQI.

Keywords: Deep learning · Multi-level analytics · $IM^{12}CP$-WQI · DWM-Bat · DMARS · Water Quality index

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 63–78, 2023.
https://doi.org/10.1007/978-3-031-27409-1_7

1 Introduction

Water is one of the most important resource for continuous life in the world. The source of water split into two types: "surface and groundwater water, in general, surface water is found in lakes, rivers, and reservoirs, while ground water lies under the surface of the land, it travels through and fills openings in the rocks". The water supply crisis is a harsh truth not only on a national level, but also on a global level. The recent Global Dangers report of the World Economic Forum lists the water supply crisis as one of the top five global risks to materialize over the next decade. On the basis of the current population trends and methods for water use, there is a strong indication that most African countries will exceed the limits of their usable water resources by 2025. The forecasted increases in temperature resulting from climate change will place additional demands on over-used water resources in the form of case dry's [1–6].

The major challenges of water are increasing water demand, water Scarcity, water pollution, inadequate access to safely, affordable water, sanitation, and climate change. That water pollution is the pollutant ion of water source such as oceans, rivers, seas, lakes, groundwater and aquifers by pollutant. Pollutants may end in the water by directly or indirectly application. This is the second most contamination type of the environmental after air pollution. The water quality depends on the eco-system and on human use, such as industrial pollution, wastewater and, more importantly, the overuse of water, which leads to reduce level of water. Water Quality is monitored by measurements taken at the original location and the assessment of water samples from the location achieving low costs and high efficiency in wastewater treatment is a popular challenge in developing states.

Prediction is one of the tasks achieve through data mining and artificial intelligent techniques; to find the discrete or continuous of facts based on the recent facts (i.e., the prediction techniques generated actual values if prediction build from real facet otherwise will generated the virtual values). Most prediction techniques based on the a statistical or probabilities tools for prediction of the future behaviors such as "Chi-squared Automatic Interaction Detection (CHAID), Exchange Chi-squared Automatic Interaction Detection (ECHAID), Random Forest Regression and Classification (RFRC), Multivariate Adaptive Regression Splines (MARS), and Boosted Tree Classifiers and Regression (BTCR)" [7].

Optimization is the process to finding of the best values dependent on the type of objective function for the problem identified. Generally speaking, the problem of maximizing or minimizing. There are many types of optimation namely *continuous optimization, bound constrained optimization, constrained optimization, derivative-free optimization, discrete optimization, global optimization, linear programming and non-differentiable optimization.* There are two types of objective function optimisation, a single objective function and a multiple objective function. In single-objective optimization, the decision to accept or decline solutions is based on the objective function value and there is only one search space. While one feature of multi-objective optimization involves potential conflicting objectives. There is therefore a trade-off between objectives, i.e. the improvement achieved for a single objective can only be achieved by

making concessions to a other objective. There is no optimal solution for all m objective functions at the same time. As a result, multiple-objective functions under a set of constrains specified [8].

The detection of Water Quality Index (WQI) is one of the most important challenges; therefore, this paper suggests a method to build an intelligent miner to predict of WQI through combination between one of optimation algorithm after developing called (DWM-Bat) with one of the prediction algorithms that based on mathematical principle called (DMARS).

2 Building IM^{12}CP-WQI

The model presents in this paper consist of two phases, the first including build the station as electrical circuit to collect the data related to 12 concentrations in real time and saved it on the master computer to preparing and processing in next phase. The second phase focuses on processing dataset after split it based on season identifier, the processing phase pass on many levels of learning to product forecaster can deal with different size of dataset. All the actives of this researcher summarization in Fig. 1 while the algorithm of *IM^{12}CP-WQI* model described in main algorithm. The main hypothesis used

- The file of water have the following: pH, TDS (mg/l), Hardness (as CaCO3) (mg/l), Alkalinity (as CaCO3) (mg/l), Nitrate (mg/l), Sulfate (mg/l), Chloride (mg/l), Turbidity (NTU), Calcium (mg/l), Magnesium (mg/l), Sodium(mg/l), finally Potassium (mg/l).
- Limitation\range for each parameters from Permissible Limit to Maximum Limit: pH [6.5–8.5] to No relaxation, TDS (mg/l) [500 to 2000], Hardness (as CaCO3) (mg/l) [200 to 600], Alkalinity (as CaCO3) (mg/l) [200 to 600], Nitrate (mg/l) [45 to No relaxation], Sulfate (mg/l) [200 to 400], Chloride (mg/l) [250 to 1000], Turbidity (NTU) [5–10 to 12], Calcium (mg/l) [50 to No relaxation], Magnesium (mg/l) [50 to No relaxation], Sodium(mg/l) [200 to No relaxation], finally Potassium (mg/l) [12 to No relaxation] (see Table 1).

2.1 Data Preprocess Stage

Dataset collection through two seasons in region of Iraq. To building the predictor as follow.

- Split the dataset for each season and save it in separated file hold the name of this season.
- apply the normalization for each column in dataset related to each season. Normalize used to all the datasets (*PH, TDS, NTU, TH, TA, Ca, Mg, K, Na, Cl, NO3, and SO4)* to make the value of that concentration in the range [0, 1].
- Finally, apply the correlation for column in dataset related to each season. Correlation Pearson used to correlation all the datasets (*PH, TDS, NTU, TH, TA, Ca, Mg, K, Na,*

Table 1. Main Chemical Parameters related to determined WQI [9]

Parameters	Unit	Recommended water quality standards (Sn)
PH		6.5–8.5
Turbidityx (NTU)	NTU	5
Totalxdissolved solid (TDS)	(mg/L)	500
Calciumx (Ca)	(mg/L)	75
Magnesiumx (Mg)	(mg/L)	50
Chloridex (Cl)	(mg/L)	250
Sodiumx (Na)	(mg/L)	200
Potassiumx (K)	(mg/L)	12
Sulfatex (SO4)	(mg/L)	250
Nitratex (NO3)	(mg/L)	50
Totalxalkalinity (CaCO3)	(mg/L)	200
Totalxhardness (CaCO3)	(mg/L)	500

Cl, NO3, and SO4) to know the correlation between the concentrations. Algorithm 1 explains the main steps of that stage.

Algorithm#1: Pre-Processing
Input: *Water dataset include two season and twelve concentration of (PH, TDS, NTU, TH, TA, Ca, Mg, K, Na, Cl, NO3, and SO4)*
Output: *Split season [id_season] have data in range [0, 1]*
// Split water pollution dataset according [id_season]
For *all samples in water dataset*
  IF *season_name = id_season*
    create a file called season_name
    Put all the concentrations related to season_name in that file
  End IF
End For
// Apply normalization
For *each row in water dataset*
  For *each column in water dataset*
    Compute Normalization
    $\text{Normalization} = \frac{x}{maxx}$
  End For
End For
// Apply correlation
For *each row in water dataset*
  For *each column in water dataset*
    Compute Correlation_Pearson
    $P_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$
  End For
End For
End pre-processing

Main Algorithm: IM[12]CP-WQI

Input: *water datasets of a two seasons include twelve concentrations* of water
(i.e., PH, TDS, NTU, TH, TA, Ca, Mg, K, Na, Cl, NO3, and SO4)

Output: *Predict the water quality index*

*// **Pre-Processing Stage***

```
1:    For each row in water dataset
2:        Call split Seasons
3:        For each column in water dataset
4:            Call Normalize
5:        End for
6:    End for
```

*// **Build DWM –Bat – DMARS Predictor***

```
7:    For each id_seasons
8:        For i in range (1: total number of samples [id_seasons])
9:            Split dataset according to 5- Cross-Validation into Training and
              Testing dataset
10:       End for
11:       For each Training part not used
12:           Call DWM –Bat     //determined the weights & number of the M
13:           Call DMARS        //predictive the value of twelve concentrations
14:       End for
15:       For each Testing part not used
16:           Test stopping conditions
17:               IF max error generation < Emax
18:                   Go to step 25
19:               Else
20:                   GO to step 8
21:               End IF
22:       End for
23:   End for
```

*// **Evaluation stage***

*24: **Call** Evaluation*

25: Generated report based on the result values of Water Quality Index (WQI)

End IM[12]CP-WQI

Fig. 1. Intelligent miner based on twelve concentrations to predict water quality

Where cov is the covariance between quantitative x and y, σ_x the standard deviation of x, σ_y the standard deviation of y, μ_x the average of x, μ_y the average of y, and E the expectation values.

2.2 Determine Weights of Concentrations and Number of Model (DWM-Bat)

In general, the BA is failing in satisfy the goal of it, when it arrives as max number of iterations without finding the goal, while it is a success in its' work when satisfy the following three steps (i.e., Evaluate the fitness of each Bat, Update individual and global bests, Update velocity and position of each Bat). These steps are repeated until some stopping condition is met. The goal of DWM-Bat is to determine the optimal (weight for each concentration, and number of base model of MARS "M"). Algorithm 2 shows the DWM-Bat step.

Algorithm#2: ***DWM –Bat***

Input: *Split seasons [id_seasons] have data in range [0, 1]*

Output: *Optimal of (weight for each concentration, number of base model "M")*

Initialization*: Pos: Position of population, NB :Population Size, W : weights of concentration, v_i :Velocity of population, Minimum (M),Maximum (M), Lower bound (Lb),Upper bound (Ub), f(pMBest), f(pwBest)=0.*

```
1:    For i=1 to max_iter          //max number of iteration
        // Determine number of base model(M)
2:      For number in base model (M) do
3:        Calculate fitness function
4:        IF fpM is better than f(pMBest)
5:          pMBest =fpM
6:        End IF
7:      End for
        // Determine weights
8:      For random weights w in W do
9:        Calculate fitness function
10:       IF fpw is better than f(pwBest)
11:         pwBest =fpw
12:       End IF
1         End for
3:
14:   End for
15:   gwBest = best w in W
16:   For random weights w in W do
17:     Calculate the velocity of weight      // as equation(1)
18:     Calculate the position of weight      // as equation(2)
19:   End for
20:     gmBest = best m in M
21:   For random weights m in M do
22:     Calculate the velocity of M       // as equation(3)
23:     Calculate the position of M       // as equation(4)
24:   End for
    End DWM –Bat
```

2.3 Develop MARS(DMARS)

Here we will train and predict concentrations movements for several epochs and see whether the predictions get better or worse over time. The Algorithm 3 shown how execution the DMARS.

Algorthim#3: DMARS

Input: *Training datasets, Split seasons [id_seasons] and the*
Optimal of (weight for each concentration, number of base model "M")

Output: *Prediction the value of WQI*

1: ***For*** *each id_seasons.*
2: *Specify Target variable Y.*
3: Building Model by following steps:
4: ***While*** complexity of Model<M, do following steps: //M determined through DWM-Bat
5: ***For*** each variable x, do following
6: ***For*** each Knot of variable x, Test each Knot according to equation:
7: $$(x-t)_+ = \begin{cases} x-t, & x > t \\ 0, & otherwise \end{cases}$$
8: Choose Knot for variable x, which decrease prediction error.
9: Add new basis function from variable x with knot to the Model.
10: ***IF*** complexity of Model>=M, then Stop building Model.
11: For each basis function in the Model make pruning as follow:
12: Calculate Generalized Cross Validation error according to equation:
13: $$GCV = \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\left(1-\frac{C}{N}\right)^2}$$
14: Where N: number of samples, C=1+cd // d is equal # of independent BFs, c is the penalty for adding a BF.
15: Remove function with high Generalized Cross Validation error.
16: After Model is completed, predict value of Y in parallel as follow:
// linear equation:
17: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ linear}$ // *as equation*$k(x_i, x_j) = x_i^T x_j$
// polynomial equation:
18: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ polynoimal}$ // *as equation*$(k(x_i, x_j) = (x_i^T x_j + \gamma)^d \quad d, \gamma > 0$
// sigmoid equation:
19: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ sigmoid}$ // *as equation* $\tanh(\gamma X_i^T X_j + r)$, γ>0
// RBF equation:
20: $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m * \boldsymbol{compute\ RBF}$ // *as equation*$(\exp(-\gamma \|X_i^T - X_j\|)^2$, γ>0
*// **DWM –Bat***
21: $y = f(x) = 100 + \sum_{m=1}^{M}$(best weight for all $concentration *$ $best\ value\ for\ all\ concentration$)
22: ***End*** *For*
End DMARS

2.4 Evaluation Stage

In this section, we will explain the evaluation of the predictor based on the compute three measures called (R^2, NSE and D), for each season to all Concentrations as shown in Algorithm 4.

3 Experiment and Results

Select the suitable parameters of any learning algorithm is considered one of the main challenges in the science, in general, MARS take a very long time in implementation

to give the result, therefore this section shows how DWM –Bat solves this problem and exceed this challenge.

In other words, the determination of weights and the number of model (M) are essential parameters that fundamentally affect DMARS performance. In general, the MARS based on the dynamic principle in selecting the parameters of it, the main parameters of DWM –Bat shown in Table 2.

Table 2. The Parameters Utilize in DWM –Bat

Parameter	Value
Number of bats (swarm size) (NB)	720
Minimum (M)	2
Maximum (M)	12
Determine frequency (pulse_frequency)	Pulse_frequency = 0*ones*(row_num,col_num)*
Loudness of pulse	1
Loudness decreasing factor(alpha)	0.995
Initial emission rate (init_emission_rate)	0.9
Emission rate increasing factor (gamma)	0.02
Bats initial velocity (init_vel)	0
Determine vector of initial velocity(velocity)	velocity = init_vel*ones*(row_num,col_num)*
Population Size (row_num)	*60*
(col_num)	*12*
Minimum value of observed matrix (min_val)	*0.0200*
Miaximum value of observed matrix (max_val)	*538*
Maximum number of iteration (max_iter)	*250*
Number of cells (n_var)	$n_var = row_num*col_num$
Lower bound (lb)	$lb = min_val*ones(row_num,col_num)$
Upper bound (ub)	$ub = max_val*ones(row_num,col_num)$
Position of bat (Pos)	$Pos = lb + rand(row_num,col_num)*(ub\text{-}lb)$
rand1, rand2	Random numbers that are in the range [0, 1]
Calculate velocity and position of weight of each concentrations	$vw = vw + rand1 * (pwBest - w) + rand2 * (gwBest - w)$ (1) $w = w + vw$ (2)
Calculate velocity and position of the # of M	$vm = vm + rand1 * (pmBest - m) + rand2 * (gmBest - m)$ (3) $m = m + vm$ (4)

By apply the DWM–Bat get the best weight of each the 12 contractions as follow: *PH = 0.247, NTU = 0.420, TDS = 0.004, Ca = 0.028, Mg = 0.042, Cl = 0.008, Na = 0.011, K =, 0.175, SO4 = 0.008, NO3 = 0.042, CaCO3(TA) = 0.011, and CaCO3(TH) = 0.004, while the optimal number of M related to winter and summer dataset is 9.*

DMARS is mainly based on the MARS algorithm, which is capable of handling the dynamic principle in selecting the parameters of it.

In this stage, forward the parameters result from DWM–Bat to DMARS that represents the weight of each material, number of model (M) with the dataset of that seasons generated from the best split of five cross-validations to represent training of DMARS the main parameters of that algorithm represent in Table 3. Then compute the prediction values based on the best split result from five cross-validations.

With respect to Eq. (5), the proposed approach found that TH, TDS, K, NO3, Na, PH, TA, Cl and Ca had a very important contribution in the prediction of the WQI in winter season from any of the remaining concentrations.

Example #1: Proof the accuracy of the proposed model through some of samples related to winter season, taking into account that the data is limited between 0 and 1 due to the normalization of it.

The use of the ideal (M) model number and the ideal weights that were determined from ***DWM-BA***, which are as follows: *M = 9; Weights = [PH = 0.247, NTU = 0.420, TDS = 0.004, Ca = 0.028, Mg = 0.042, Cl = 0.008, Na = 0.011, K = 0.175, SO4 = 0.008, NO3 = 0.042, CaCO3(TA) = 0.011, and CaCO3(TH) = 0.004]. In general, the* ranges of WQI based on the stander measures and possible use shown below (see Table 4).

Proof:

1-IF PH = 0.991; TDS = 0.675; Cl = 0.667; TA = 0.7939; Ca = 0.8634; TH = 0.825; NO3 = 0.194; Na = 0.300; K = 0.0012.

WQI (1) = 100*[0.991**0.247* + 0.675**0.004* + 0.667**0.008* + 0.794* *0.011* + 0.864* 0.*028* + 0.825* 0.*004* + 0.194**0.042* + 0.300**0.011* + 0.002**0.175]*

WQI (1) = 100* 0.300837 = 30.0837

Obviously, ***the WQI score is dependent on Case #2***

2-IF PH = 1.000; TDS = 0.729; Cl = 0.750; TA = 0.786;Ca = 0.0773; TH = 0.850; NO3 = 0.186;Na = 0.300;K = 0.002.

WQI (2) = 100*[1.000**0.247* + 0.729**0.004* **+** 0.750**0.008* **+** 0.786* *0.011* + 0.773* **0**.*028* + 0.850* 0.*004* + 0.186 **0.042* **+** 0.300**0.011* **+** 0.002**0.175]*

WQI (2) = 100*[0.301068] **=** 30.1068

Obviously, ***the WQI score is dependent on Case #2***

As for prediction values to WQI for two seasons winter and summer based on the best result of a split of five cross validations for $IM^{12}CP$-*WQI* model, where data for each season were divided into two parts, 80% samples training and 20% samples testing, and Ranging for all material from 0 to 1. We notice that the prediction values are very close to the real values and this indicates that the $IM^{12}CP$-*WQI* predictor is a good

Table 3. The Parameters Utilize in DMARS

Parameter	Description
Number of input variable(d)	d = 12
Datasets (x)	x = samples of winter season or samples of summer season
Number of columns (m)	m = 13
Number of row (n)	n = 60
Training data cases (Xtr, Ytr)	Xtr(i,:), Ytr(i), i = 1, …, n
Vector of maximums for input variables (x_max)	x_max(winter) = [0.06, 7.55, 538, 42.60, 381.66, 417.424, 88, 397.984, 15.32, 9.28, 457.20, 135.69, 94.27] x_max(sumer) = [0.060, 7.470, 539, 24.850, 325, 417.760, 92, 447.424, 6.700, 3.800, 427.760, 137.945, 87.707]
Vector of minimums for input variables (x_min)	x_min(winter) = [0.02, 7.240, 363, 21.300, 300, 28.800, 36, 2.35, 1.859, 1.780, 0.89, 20.146, 12.233] x_min(summer) = [0.0200, 6.900, 390, 14.200, 235, 24, 33.600, 2.355, 1, 0.920, 0.630, 64.857, 11.449]
Size of dataset (x_size)	*x_size(n,m) = x_size(60, 12)*

BF	Equation
BF_Z1	0.175*K // k = 0.985
BF_Z2	0.011*TH // TH = 0.86
BF_Z3	0.042*NO3 // NO3 = 0.761
BF_Z4	0.004*TDS // TDS = 0.55
BF_Z5	0.011*Na // Na = 0.415
BF_Z6	0.247*PH // PH = 0.371
BF_Z7	0.011*CaCo3(TA) // TA = 0.37
BF_Z8	0.008*Cl // Cl = 0.362
BF_Z9	0.028*Ca //Ca = 0.317

$$WQI = 100 * \sum_{(K=0)}^{M} (BF_ZK) = 100 * (\text{BF_Z1} + \text{BF_Z2} + \text{BF_Z3} + \text{BF_Z4} + \text{BF_Z5} + \text{BF_Z6} + \text{BF_Z7} + \text{BF_Z8} + \text{BF_Z9}) \quad (5)$$

predictor as it was able to predict the real values well, so it is a better predictor compare with MARS linear, MARS_Sig, MARS_RBF and MARS_Poly. As shown in Figs. 2, 3, 4, and 5.

Table 4. Generated report of WQI based on four cases

Case	WQI	Possible use
Case#1	Value in rang (0–25)	Drinkable
Case#2	Value in rang (26, 50)	Fit for aquarium and animal drinking
Case#3	Value in rang (51, 75)	Not suitable for drinking, but suitable for watering crops
Case#4	Value in rang (76, 100)	Unusable pollutant must go to recurrence

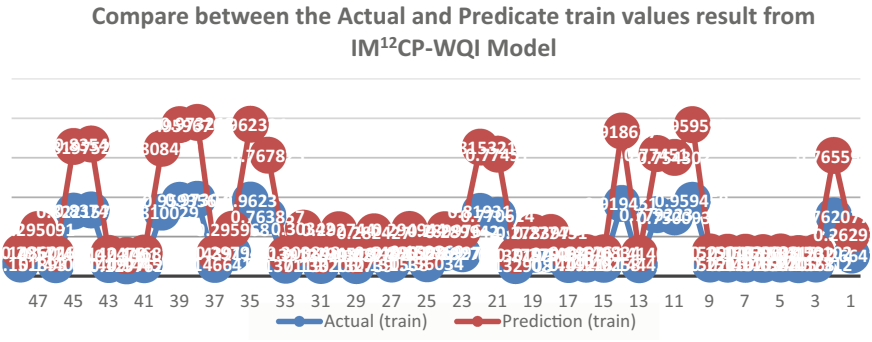


Fig. 2. Predictive Model *IM*^{12}*CP-WQI* for Training Dataset of Winter Season

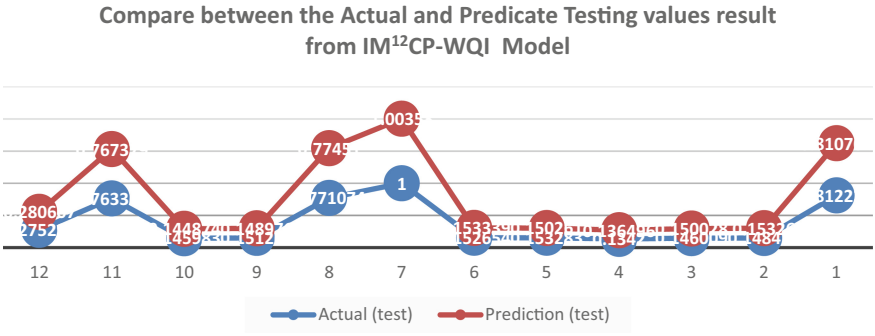


Fig. 3. Predictive Model *IM*^{12}*CP-WQI* for Testing dataset for Winter Season

The results shown ***IM*^{12}*CP-WQI*** model, were located closer to the reference point, indicating better performance compared to the other models. A comparison showed that the ***IM*^{12}*CP-WQI*** model generally converged faster and to a lower error value than the eithers model under same input combinations. The novel hybrid ***IM*^{12}*CP-WQI*** model showed more accurate WQI estimates with faster convergence rate than the other models.

The performances of the all models test in this study (i.e., MARS Linear, MARS_Poly, MARS_Sig, MARS_RBF, and MARS_DWM-BA) to predict the WQI

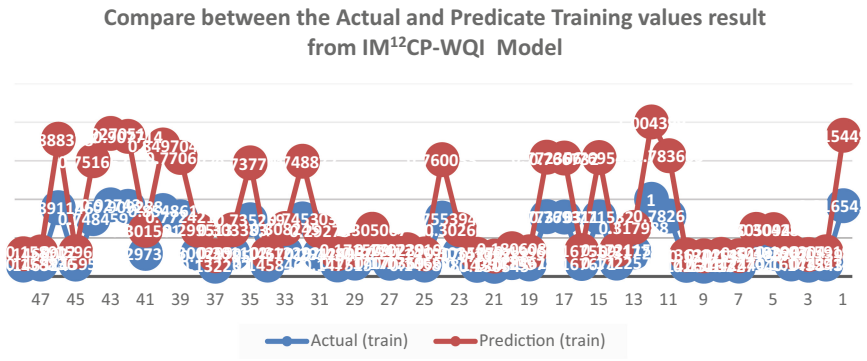


Fig. 4. Predictive Models $IM^{12}CP\text{-}WQI$ for Training Dataset to Summer Season

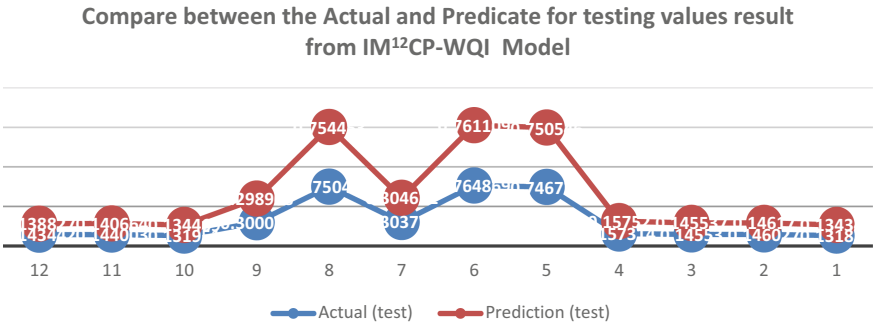


Fig. 5. Predictive Model $IM^{12}CP\text{-}WQI$ for Testing Dataset to Summer Season

were investigated for both training and testing stages for both seasons (winter and summer).

- In the training phase at winter season, for the prediction of WQI, $IM^{12}CP\text{-}WQI$ provided more accurate performance (R2 = 0.2202, NSE=0.9999, and D = 1) compare with other models, and MARS_RBF provided less accurate performance (R2 = −0.1148, NSE = −2.3411, and D = −16.6417) compare with other the models.
- While, in the testing phase at winter season for the prediction of WQI, IM12CP-WQI provided more accurate performance (R2 = 0.7919, NSE = 0.9999, and D = 1) compare with other models, in other side; MARS_RBF provided less accurate performance (R2 = −0.2034, NSE = −1.4032, and D = −2.5096).
- While, the evaluation of the summer season proves the training dataset of IM12CP-WQI give the best performance based on the three evaluation measures (R2 = 0.2331, NSE = 0.9999, and D = 1) compare with other models, and MARS_RBF provided less accurate performance (R2=0.751, NSE = −2.2284, and D = −12.0533) compare with the other models.
- Also, IM12CP-WQI provided more accurate performance for the three measures of testing dataset (R2 = 1.2688, NSE = 0.9999, and D = 1) compare with other models,

while; MARS_RBF provided less accurate performance (R2 = 2.7051, NSE = −2.185, and D = −2.6243).

4 Discussion

In this section, a quite few statistical measures are presented to evaluate the performance of the proposed models. Moreover, the results of the IM^{12}CP-WQI and MARS technologies compared with more than one core. The results proved that the IM^{12}CP-WQI model gives the best results according to the evaluation measures in two seasons related to the training and testing dataset, in general, this study answers the following questions [10–16]

- How Bat optimization algorithm can be useful in building an intelligent Miner?
- BOA works to modify the behavior of each in a particular environment gradually, depending on the behavior of their neighbors until they obtained the optimal solution.
- On the other hand, the MARS use the principle of the try and error in the selection of the basic parameters of their own and modified gradually to reach the values accepted for those parameters.
- Depending on the BOA and MARS of the above subject, we used the BOA principle to find the optimal weights for each concentration and the number of based models of the MARS.
- How to build a multi-level model with a combination of two technologies)MARS with BOA)?

Through, building new miner called IM^{12}CP-WQI that combining between the DWM –Bat and the DMARS. Where DWM –Bat used to find the best values of wights to each concentration with best number of M to DMARS while DMARS used to predict the water quality index (WQI).

- Is three evaluation measures enough to evaluate the results of suggested Miner?
- Yes, that measures are sufficient to evaluate the results of the miner to the both seasons.
- What is the benefit result from building miner by combination between DWM_Bat and DMARS?

By combining DWM_Bat and DMARS, reduce the execution time by defining MARS parameters but at the same time will increase the computational complexity.

5 Conclusions

We can summarize the main point performance in that paper as the follows: Water quality index dataset is a sensitive data need to accuracy techniques to extract a useful knowledge from it. Therefore; IM^{12}CP-WQI was able to solve this problem by giving results of high predictive accuracy, but on the other hand, it increased the mathematical complexities to obtain of that results. The main purpose of the normalization process is to convert data within a specified range of values to be handled more precisely at subsequent processing

stages. Especially since the concentrations are within different ranges and are measured in different units, so a normalization has been made to make them within a specific range to work on. Where the concentrations were placed between range (0, 1). This study proves the correlation between WQI and the important concentrations are k = 0.985, TH = 0.86, NO3 = 0.761, TDS = 0.55, Na = 0.415, PH = 0.371, TA = 0.37, Cl = 0.362, Ca = 0.317. This step focus on determined the important concentrations are Total Hardness (TH) that have negative relation with WQI and TDS. By apply the DWM-Bat get the best weight of each concentration as follow: W-PH = 0.247, W-NTU = 0.420, W-TDS = 0.004, W-Ca = 0.028, W-Mg = 0.042, W-Cl = 0.008, W-Na = 0.011, W-K = 0.175, W-SO4 = 0.008, W-NO3 = 0.042, W-CaCO3(TA) = 0.01l, and W-CaCO3(TH) = 0.004. While the optimal number of M related to both datasets are 9. This stage increases the accuracy of results and reduces the time required to training the MARS algorithm. Selection the best activation function to build the predictor based on mathematical concept, through build DMARS that replace the core of MARS by four types of functions (i.e., polynomial, sigmoid, RFB and linear). Results indicated that the MARS technique with linear and sigmoid kernel functions have stood at higher level of accuracy rather than the MARS approaches developed by other types of kernel functions. As the results of both training and testing indicated that MARS-linear and MARS-sig methods have provided relatively precise prediction for WQI, compared to the MARS_RBF and MARS_Poly. IM^{12}CP-WQI give pragmatic model of water quality index for different seasons indicates the water become high quality when the value of WQI is small value not exceed twenty-five will used to drink while other values highest than twenty-five to fifty. It is possible use to other uses, such as watering crops, fish lakes, and factories, except that requires a refining process to the water.

The following point give good idea for features works; Using other optimization algorithms based on search agent algorithm such as Whale Optimization Algorithm (WOA) or Particle Swarm Optimization (PSO) or Ant Lion Optimization (ALO). Investigation other prediction algorithm that adopts the mining principle such as Gradient Boosting Machine (GBM) or extreme gradient boosting (XGBoost). Verification from the prediction results based on other evaluation measures such as (Accuracy, Recall, Precision, F, and FB). Test the model on the new dataset that contain other concentrations rather than these used in this study.

Author Contributions. All authors contributed to the study conception and design. Data collection and analysis were performed by [Samaher Al-Janabi] and Zahra A. The first draft of the manuscript was written by [Samaher Al-Janabi] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations.

Conflict of Interest:. The authors declare that they have no conflict of interest.

Ethical Approval:. This article does not contain any studies with human participants or animals performed by any of the author.

References

1. Hudson, Z.: The applicability of advanced treatment processes in the management of deteriorating water quality in the Mid-Vaal river system. Environmental Sciences at the Potchefstroom Campus of the North-West University or Natural and Agricultural Sciences [1709] (2015). http://hdl.handle.net/10394/16075
2. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., García-Nieto, J.: Efficient Water Quality Prediction Using Supervised Machine Learning, vol. 11, p. 2210 (2019). https://doi.org/10.3390/w11112210
3. Aghalari, Z., Dahms, H.U., Sillanpää, M., et al.: Effectiveness of wastewater treatment systems in removing microbial agents: a systematic review. Global Health **16**, 13 (2020). https://doi.org/10.1186/s12992-020-0546-y
4. Singh, P., Kaur, P.D.: Review on data mining techniques for prediction of water quality. Int. J. Adv. Res. Comput. Sci. **8**(5), 396–401 (2017)
5. Qiu, Y., Li, J., Huang, X., Shi, H.: A feasible data-driven mining system to optimize wastewater treatment process design and operation. **10**, 1342 (2018). https://doi.org/10.3390/w10101342
6. Al-Janabi, S.: Smart system to create an optimal higher education environment using IDA and IOTs. Int. J. Comput. Appl. **42**(3), 244–259 (2020). https://doi.org/10.1080/1206212X.2018.1512460
7. Al-Janabi, S.: A novel agent-DKGBM predictor for business intelligence and analytics toward enterprise data discovery. J. Babylon Univ./Pure Appl. Sci. **23**(2) (2015)
8. Alkaim, A.F., Al-Janabi, S.: Multi objectives optimization to gas flaring reduction from oil production. In: Farhaoui, Y. (eds.) Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems, vol 81. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-23672-4_10
9. Ameen, H.A.: Spring water quality assessment using water quality index in villages of Barwari Bala, Duhok, Kurdistan Region, Iraq. Appl. Water Sci. **9**(8), 1–12 (2019). https://doi.org/10.1007/s13201-019-1080-z
10. Al-Janabi, S., Mahdi, M.A.: Evaluation prediction techniques to achievement an optimal biomedical analysis. Int. J. Grid and Utility Comput. **10**(5), 512–527 (2019).https://doi.org/10.1504/IJGUC.2019.102021.7
11. Al-Janabi, S., Patel, A., Fatlawi, H., Al-Shourbaji, I., Kalajdzic, K.: Empirical rapid and accurate prediction model for data mining tasks in cloud computing environments. In: 2014 International Congress on Technology, Communication and Knowledge (ICTCK), pp. 1–8 (2014). https://doi.org/10.1109/ICTCK.2014.7033495
12. Al_Janabi, S., Yaqoob, A., Mohammad, M.: Pragmatic method based on intelligent big data analytics to prediction air pollution. In: Big Data and Networks Technologies, BDNT 2019. Lecture Notes in Networks and Systems, pp. 84–109, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23672-4_8
13. Al-Janabi, S., Alkaim, A.F., Adel, Z.: An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft. Comput. **24**, 10943–10962 (2020). https://doi.org/10.1007/s00500-020-04905-9
14. Al-Janabi, S., Alkaim, A.F.: A comparative analysis of DNA protein synthesis for solving optimization problems: a novel nature-inspired algorithm. In: Abraham, A., Sasaki, H., Rios, R., Gandhi, N., Singh, U., Ma, K. (eds.) Proceedings of the 11th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2020) held during December 16–18. IBICA 2020. Advances in Intelligent Systems and Computing, vol. 1372, pp. 1–22. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73603-3_1
15. Al-Janabi, S., Kad, G.: Synthesis biometric materials based on cooperative among (DSA, WOA and gSpan-FBR) to water treatment. In: Abraham, A., et al. (eds.) Proceedings of

the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020). SoCPaR 2020. Advances in Intelligent Systems and Computing, vol. 1383, pp. 20–33. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73689-7_3

16. Al-Janabi, S ., Mohammad, M., Al-Sultan, A.: A new method for prediction of air pollution based on intelligent computation. Soft. Comput. **24**(1), 661–680 (2019).https://doi.org/10.1007/s00500-019-04495-1
17. Sharma, T.: Bat Algorithm: an Optimization Technique. Electrical & Instrumentation Engineering Department Thapar University, Patiala Declared as Deemed-to-be-University u/s 3 of the UGC Act., 1956 Post Bag No. 32, PATIALA–147004 Punjab (India) (2016). https://doi.org/10.13140/RG.2.2.13216.58884

Hybridized Deep Learning Model with Optimization Algorithm: A Novel Methodology for Prediction of Natural Gas

Hadeer Majed, Samaher Al-Janabi(✉), and Saif Mahmood

Faculty of Science for Women (SCIW), Department of Computer Science, University of Babylon, Hillah, Iraq
samaher@itnet.uobabylon.edu.iq

Abstract. This paper handles the main problem of natural gas through design the hybrid model based on developing one of predict data mining techniques. The model consists of four stages; The first stage collects data from a different source related to natural gas in real-time. The second stage, pre-processing is divided into multi steps including (a) Checking missing values. (b) Computing correlation among features and target. The third stage; building a predictive algorithm (DGSK-XGB). The fourth stage uses five evaluation measures in order to evaluate the results of the algorithm DGSK-XGB. As a results; we found DGSK-XGB give high accuracy reach to 93% compare with the tractional XGBoost; also, it reduces implementation time. And improving the performance.

Keywords: Natural Gas · XGboost · GSK · Optimization techniques

1 Introduction

The process of emission of gases in laboratories, or as a result of extracting some raw materials from the earth, or as a result of respiration of living organisms, is one of the most important processes for sustaining life. In general, these gases are divided into two types, some of them are poisonous and cause problems to the life of living organisms, and the other type is useful and necessary and used in many industries. Therefore, this paper attempts to build a model that classifies six basic types of those gases, which are (Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone, and Toluene) [1, 2].

The basic components of natural gas are (Methane (c1), Non-hydrocarbons (H2O, CO2, H2S), NGL (Ethane (c2), pentane (c5), and heavier fractions), LPG (propane (c3), Butane(c4)). To leave solely liquid natural gas, we shall eliminate both methane and non-hydrocarbons (water, carbon dioxide, hydrogen sulfide). That natural gas emits less CO2 than petroleum, which emits less CO2 than coal. The first choice is usually to save money and increase efficiency. One of the advantages of natural gas is that it burns entirely when used, and unlike other traditional energy sources, the carbon dioxide produced when burning is absolutely non-toxic [3, 4]. Natural gas is a pure gas by nature, and any contaminants that may be present in it may sometimes be simply and inexpensively

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 79–95, 2023.
https://doi.org/10.1007/978-3-031-27409-1_8

eliminated. Natural gas stations are not generally distributed and natural gas has a number of drawbacks, including the fact that extraction may be hazardous to the environment and necessitates the use of a pipeline, as well as the fact that methane leaks contribute to global warming. It asserts that increasing the pressure on gas at constant temperature reduces the volume of the gas [5]. In other words, Boyle's law asserts that volume is inversely proportional to pressure when the temperature and number of molecules stay constant. Natural gas is composed of hydrocarbon components such as methane, but also ethane, propane, butane, and pentane, all of which are referred to as natural gas liquids (NGLs), as well as impurities such as carbon dioxide (CO_2), hydrogen sulfide (H_2S), water, and nitrogen [6].

Intelligent Data Analysis (IDA) [7, 15, 26] is one of the pragmatic fields in computer science based on integration among the data Domain, Mathematical domain, and Algorithm domain; In general, to handle any problem through IDA must satisfy the following: (a) real problem: must found one of the real problems in one of the specific field of life, (b) design a new or a novel or hybrid model to solve it based on the integration among the above Three domains; (c) interpretation the result after analysis it to become understand & useful for any person not only for the person expert in the specific field of problem.

This paper will handle the main problem of Natural Gas that description in the above section by designing the hybrid model based on develop one of predict data mining technique through the optimization principle.

The problem of this work is divided into parts: The first part is related to programming challenges while; the second part is related to application challenges; In general; the prediction techniques are split into two fields; prediction techniques related to data mining and predictions related to neurocomputing; this work deal with the first type of prediction technique called XGboost; in general; XGboost is one of the data mining prediction techniques that characterized by many features that make it the best. These features (include XGboost give high accuracy results and work with huge data/stream data in real time but on other hand; the core of that algorithm is decision tree (DT) that have many limitations such as it requires choose the root of tree, determined the max number of levels of tree, also it have high computation and time of implementation. Therefore; the first challenge of this paper is how can avoid these limitations (i.e., high computation and time of implementation) of this algorithm and befit from their features. On other side; The problem of application can summarization by need of high efficiency prediction techniques; Therefore, the second challenge of this paper is how can avoid these limitations thought build an efficient technique to predict multi types of gas coming from different sensors.

2 Main Tools

Optimization [7, 15] is one of the main models in computer science based on find the best values such as max, min or benefit values through optimization function; In general; the optimization model split into single object function model or multi objective's function model also, some of these models based on constructions while the other not. There are many Techniques can used to find the optimal solution such as [8].

2.1 Optimization Techniques [9–11]

2.2 Particle Swarm Optimization Algorithm (PSO)

Eberhart and Kennedy devised one of the swarm intelligence methods, particle swarm optimization (PSO), in 1995. It's a population-based, stochastic algorithm inspired by social behaviors seen in confined birds. It is one of the approaches to evolutionary optimization.

2.3 Genetic Algorithm (GA)

Genetic algorithms were developed in 1960 by John Holland at the University of Michigan but did not become popular until the 1990s. Their main goal is to address issues when deterministic techniques are too expensive, And the genetic algorithm is a type of evolutionary algorithm that is inspired by biological evolution. It is the selection of parents, reproduction, and mutation of offspring.

2.4 Ant Lion Optimizer (ALO)

Mirjalili created ALO, a Metahorian swarm-based technique, in 2015 to imitate ant hunting behavior in nature. The lion-ant optimizer solves optimization issues by providing a heuristic after-factoring technique. It is an algorithm that is based on population. Antelopes and ants are the primary food sources for people.

2.5 Gaining-Sharing Knowledge-Based Algorithm (GSK) [12, 16]

Nature-inspired algorithms have been widely employed in several disciplines for tackling real-world optimization instances because they have a high ability to tackle non-linear, complicated, and challenging optimization issues. Algorithm for knowledge acquisition and sharing; It is a great example of a modern algorithm influenced by nature that uses real-life behavior as a source of inspiration for problem solutions (see Table 1).

Table 1. Analytic the Advantages and Disadvantages for Optimization Techniques.

O T	Advantage	Disadvantage
PSO	Simple to put into action There are a limited number of settings that must be adjusted It is possible to compute it in parallel The end consequence of it validation Locate the worldwide best solutions Convergent quick method Do not mutate and overlap Demonstrating a short implantation time	Selecting the initial values for its parameters using the concept of trial and error/at random It only works with scattering issues In a complicated issue, the solution will be locked in a local minimum
GA	It features a high number of parallel processors It is capable of optimizing a wide range of problems including discrete functions Continuous functions and multi-objective problems It delivers responses that improve with time There is no requirement for derivative information in a genetic algorithm	Implementing GA is still a work in progress GA necessitates less knowledge on the issue However, defining an objective function and ensuring that the representation and operators are correct may be tricky GA is computationally costly, which means it takes time
ALO	The search region is examined using this technique by selecting at random and walking at random as well The ALO algorithm has a high capacity to solve local optimization stagnation due to two factors: the first reason was the use of a roulette wheel, and the second component was the use of haphazard methods Relocates to a new location, and this site performs better throughout the optimization process, i.e. it retains search area areas It contains a few settings that you may change	The reduction in movement intensity is inversely related to the increase in repetitions Because of the random mobility, the population has a high degree of variety, which causes issues in the trapping process Because the method is not scaled, it is analogous to the black box problem
GSK	To resolve optimization issues GSK is a randomized, population-based algorithm that iterates the process of acquiring and sharing knowledge throughout a person's life Use the GSK method to tackle a series of realistic optimization problems that have been suggested In reality, it is simple to apply and a dependable approach for real-world parameter optimization	The algorithm is incapable of handling and solving multi-objective restricted optimization problems The method cannot address issues with enormous dimensions or on a wide scale Mixed-integer optimization issues cannot be solved

2.6 Prediction Techniques

Prediction is find event/value will occur in the future based on the recent facts, the prediction based on law say the predictor give the real values if it is build based on facts otherwise will give virtual values. In general; The prediction techniques split into two types technique based on data mining while the other based on neurocomputing techniques. This paper works with the first type of that technique. as explain below.

2.7 The Decision Tree (DT)

A decision tree is one of the simplest and most often used classification techniques. The Decision Tree method is part of the supervised learning algorithm family. The decision tree approach is also applicable to regression and classification issues [13].

2.8 Extra Trees Classifier (ETC)

Extra Trees Classifier is a decision tree-based ensemble learning approach. Extra Trees Classifier, like Random Forest, randomizes some decisions and data subsets to reduce over-learning and overfitting. Extra Trees Classifier. Trailing trees have a classifier [14].

2.9 Random Forest (RF)

Leo Breiman invented the random forest aggregation technique in 2001. According to Breman, "the generalization error of a forest of tree classifiers is dependent on the strength and interdependence of the individual trees in the forest" [17].

2.10 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting framework-based decision-tree-based ensemble Machine Learning approach. Artificial neural networks outperform all existing algorithms or frameworks in prediction problems involving unstructured data (images, text, etc.). Decision tree-based algorithms are the best [18] (see Table 2).

Table 2. Analytic the Advantages and Disadvantages for Prediction Techniques.

PT	Advantage	Disadvantage
DT [24]	Decision trees take less work for data preparation during pre-processing as compared to other methods Data normalization is not necessary for a decision tree Data scaling is not required for a decision tree Data missing values have no discernible impact on the decision tree generation process The decision tree technique is highly natural and simple to interact with technical teams as well as stakeholders	A tiny change in the data causes a significant change in the structure of the decision tree, resulting in instability When compare this approach to other algorithms, may see that the decision tree calculation become more complicated at times A decision tree is rehearsal time is frequently lengthy Because of the additional complexity and time required, decision tree training is more expensive For forecasting continuous values and performing regression, the Decision Tree approach is unsuccessful

(*continued*)

Table 2. (*continued*)

PT	Advantage	Disadvantage
ETC [25]	A sort of collective learning in which the outcomes of numerous non-correlated decision trees gathered in the forest are combined Increased predicting accuracy by using a meta-estimator DT should be generated using the original training sample Similar to the RF classifier, both ensemble learning models are used The manner trees are built differs from that of RF It chooses the optimum feature to partition the data based on the math Gini index criterion	Bad performance when Overfitting is a difficult problem to tackle A huge number of uncorrelated DTs are generated by the random sample
RF [26]	Both regression and classification are possible using RF The random forest generates accurate and understandable forecasts It can also successfully handle massive data categories In terms of accuracy in forecasting results, the random forest algorithm surpassed the decision tree method Noise has a less influence on Random Forest Missing values may be dealt with automatically using Random Forest Outliers are frequently tolerated by Random Forest and handled automatically	Model interpretability: Random Forest models are not easily understood because of the size of the trees, it can consume a large amount of memory Complexity: Unlike decision trees, Random Forest generates a large number of trees and aggregates their results Longer Training Period: Because Random Forest creates a large number of trees, it takes significantly longer to train than choice trees
XGBoost	The main benefit of XGB over gradient boosting machines is it has many hyperparameters that can be tweaked XGBoost has a feature for dealing with missing values It has several user-friendly features, including parallelization, distributed computing, cache optimization, and more The XGBoost outperforms the baseline systems in terms of performance It can benefit from out-of-core computation and scale seamlessly	XGBoost performs poorly on sparse and unstructured data Gradient Boosting is extremely sensitive to outliers since each classifier is compelled to correct the faults of the previous learners. Overall, the approach is not scalable

3 Proposed Method (HPM-STG)

This section presents the main stages of building the new predictor and shows the specific details for each stage. The hybrid Prediction Model for Six types of Natural Gas (HPM-STG) consist of four stages; The first stage collects data from a different source related to natural gas in real-time. The second stage, pre-processing is divided into multi steps including (a) Checking missing values. (b) Computing correlation among features and target. The third stage; building a predictive algorithm (DGSK-XGB). The fourth stage uses five evaluation measures in order to evaluate the results of the algorithm DGSK-XGB. The HPM-STG block diagram is shown in Fig. 1, and the steps of the model are shown in the algorithm (1). We can summarize the main stages of this research below:

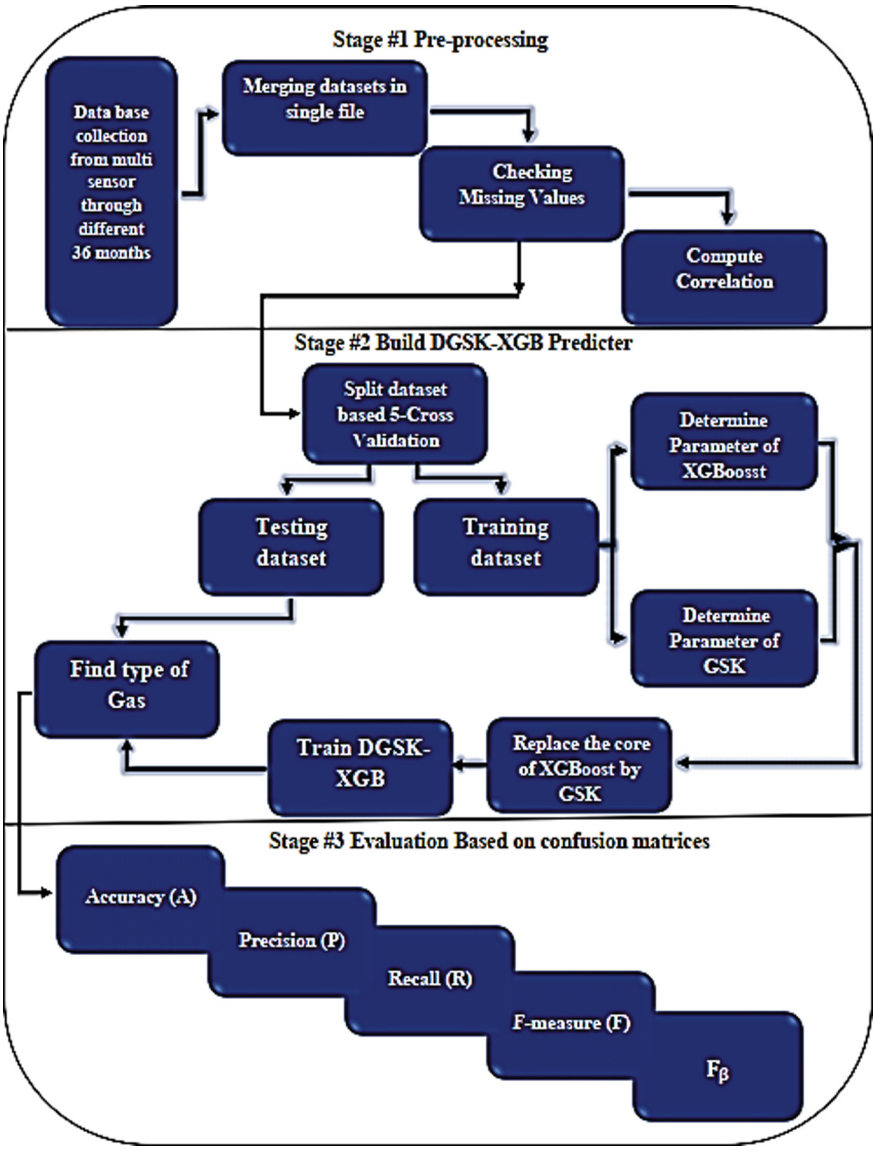


Fig. 1. Block diagram of DGSK-XGB Model

- Capture data from scientific location on internet where, these data collection from different sensors related to the natural gas.
- Through the pre-processing stage, check missing values and compute the correlation.
- Build a new predictor called (HPM-STG) by combining the benefits of GSK and XGBoost.
- Multi measures use to evaluate the predictor results include (accuracy, Precision, Recall, f-measurement, and Fb).

Algorithm# 1: Hybrid Prediction Model for Six Types of Gas (HPM-STG)

Input: Stream of real-time data capture from 16 sensors, each sensor, give 8 features; the total number of features 128 collect from 16 sensors

Output: *Predict the six types of Gas (Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone, and Toluene)*

*// **Pre-Processing Stage***

1: ***For*** *each row in dataset*

2: ***For*** *each column in dataset*

3: ***Call*** *Check Missing Values*

4: ***Call*** *Correlation*

5: ***End*** *for*

6: ***End*** *for*

*// **Build DGSK –XGB Predictor***

7: ***For*** *i in range (1: total number of samples in dataset)*

8: *Split dataset according to Five- Cross-Validation into Training and Testing dataset*

9: ***End*** *for*

10: ***For*** *each Training part*

11: ***Call*** *DGSK-XGB* ***//used Ackley Function as Function to test fitness function with GSK as kernel of XGboost***

12: ***End*** *for*

13: ***For*** *each Testing part*

14: *Test stopping conditions*

15: ***IF*** *max error generation < Emax*

16: ***Go to*** *step* ***21***

17: ***Else***

18: ***GO to*** *step* ***10***

19: ***End IF***

20: ***End*** *for*

*// **Evaluation stage***

21: ***Call*** *Evaluation*

End HPM-STG

4 Results

This section of the paper plain the main results; In addition, described the details of a database used to implement the DXGboost-GSk model.

4.1 Description of Dataset

The database has 16 sensors; each sensor gives 8 features therefore, the total number of features equal to 128. The data is affiliated to 36 months divided into 10 divisions. Each division is called a batch and the data belongs to 6 types of gases called Ammonia, Acetaldehyde, Acetone, Ethylene, Ethanol, and Toluene.

4.2 Result of Preprocessing

This stage begin form get the database from scientific internet sit, where these database aggregation from multi sensors through different periods of time include 36 months. Split into ten groups.

4.3 Checking Missing Value [21]

After Merging all datasets in single file; we checking if that file has missing values or not; if found drop the record from that dataset to satisfy the Law of prediction otherwise continuous. In general, in this step not dropping any record.

4.4 Correlation [19, 20]

The correlation is computed among all the features with the target to determine the main features effect in specific type of gas. In general, we Found three types of relationship among features and target; when the correlation forward in side (+1) this meaning the Positive relationship while If correlation value goes side (−1) this meaning the negative relationship between feature and target; otherwise, if correlation value is go side (0) this meaning not found any relationship between feature and target.

The effects and relationships among features. When the value of the adopted threshold is greater than or equal to 0.80.

4.5 Results of DXGBoost-GSk

This section of chapter will apply the main steps of predictor after spilt the dataset into training and testing parts through 5-cross validation Then grouping dataset by GSK after that; specific Label for each group through DXGboost; Final evaluation the results. The data is divided into training data test data as shown in Table 3. Through five cross validations, where; we build model based on certain percentage of the data, where this percentage of the data, where this percentage is for training and the rest for testing, and so on for the rest of the sections. Each time the error value is calculated, and who split gives the lowest error rate is depend on build the final model. In general; the total number of samples of these datasets are 13910.

Table 3. Number of samples of training and testing dataset based on five cross validations

Rate training dataset	# samples	Rate testing dataset	# samples
80%	11128	20%	2782
60%	8346	40%	5564
50%	6955	50%	6955
40%	5564	60%	8346
20%	2782	80%	11128

The Table 4 shows results of GSK based on three equations: junior, senior, and Ackley.

Table 4. The Result of GSK

It	Junior	Senior	Ackley
1	10.15019163870712	0.8498083612928795	22.753980010395882
2	9.35839324839964	1.64160675160036	22.725819840576897
3	8.621176953814654	2.378823046185346	22.627559663453333
4	7.935285368822167	3.064714631177833	22.739134598174868
5	7.297624744179685	3.702375255820315	22.63180468736198
6	6.705258323951894	4.294741676048106	22.736286138420425
7	6.155399906315438	4.844600093684562	22.73751724165138
8	5.64540760451318	5.35459239548682	22.678015204137193
9	5.1727778037666745	5.8272221962333255	22.7683895201492
10	4.735139310000001	6.264860689999999	22.732904122147605
11	4.33024768627229	6.66975231372771	22.730777667271113
12	3.9559797728608257	7.044020227139175	22.801723818612935
13	3.610328386980833	7.389671613019167	22.63505095191573
14	3.291397198172441	7.708602801827559	22.627375053302202
15	2.997395775429687	8.002604224570312	22.785544848141853
16	2.7266348021907447	8.273365197809255	22.77595747749035
17	2.477521455352944	8.522478544647056	22.7058029631555
18	2.248554944520475	8.751445055479525	22.687643377769465
19	2.038322207737026	8.961677792262973	22.701816441723256
20	1.845493760000001	9.15450624	22.763066773233398
21	1.6688196908972177	9.331180309102782	22.773781043057618

(*continued*)

Table 4. (*continued*)

It	Junior	Senior	Ackley
22	1.50712580775145	9.49287419224855	22.647336109599276
23	1.3593099207024493	9.640690079297551	22.682470735962827
24	1.2243382662004736	9.775661733799527	22.80833444732085
25	1.1012420654296875	9.898757934570312	22.65764842443089

The GSK algorithm is applied to the data and depends on three main parameters (Junior, Senior, Ackley) where each parameter depends on a certain law to be executed and indicates something where Junior means the amount of information to be obtained and Senior is the amount of information to be shared and they are the two principles The work of the GSK algorithm and the last parameter, Ackley [22, 23], which is its work to test the fitness function, is based on the optimization principle, So it is suitable for the working principle of the GSK algorithm.

While; the results of XGBoost after replacing their kernel with GSK are explained in Table 4.

In Table 5, the results of the developed method appeared, where it was found that the extent of convergence between Initial Residuals and New Residuals, as well as New Predictions, is the purpose of showing the value of the predictor to be closer to the real values, and whoever approaches the real values, the result is better, and each time the learning coefficient is added to expand the range It is useful to reach the real values by step by step, where if the jump is made quickly and the real values are reached, the results will be inaccurate, which is the reason for using the learning coefficient α and continuing until it approaches the real values.

Table 5. The result of HPM-STG

Iteration	Initial residuals	New predictions	New residuals
0	1.272424	8.187216	1.145182
1	−6.909718	6.223820	−6.218746
2	−6.913936	6.223398	−6.222542
3	−6.910175	6.223774	−6.219158
4	−6.772750	6.237517	−6.095475
5	−2.514800	6.663311	−2.263320
6	−6.914639	6.223328	−6.223175
7	−5.742731	6.340518	−5.168458
8	4.543536	7.369145	4.089182
9	−6.870089	6.227783	−6.183080

(*continued*)

Table 5. (*continued*)

Iteration	Initial residuals	New predictions	New residuals
10	−6.846299	6.230162	−6.161669
11	−3.359608	6.578831	−3.023647
12	2.267459	9.182251	2.040713
13	−6.912200	6.223572	−6.220980
14	−6.908514	6.223940	−6.217662
15	−6.914647	6.223327	−6.223182
16	−6.497434	6.265048	−5.847690
17	−6.683213	6.246470	−6.014892
18	−5.932537	6.321538	−5.339283
19	−6.914216	6.223370	−6.222794
20	−6.914572	6.223334	−6.223115
21	−6.893094	6.225482	−6.203784
22	−6.914734	6.223318	−6.223261
23	−5.683538	6.346438	−5.115184
24	−6.826928	6.232099	−6.144235
25	1.272424	8.187216	1.145182

In Table 6, the results of the Evaluation measures are shown, as it examines the efficiency of the model for each of the six types of gas, where each scale has a certain number that shows the accuracy of the system, and the best measure was found for each type of gas, as the results are shown in the above table.

Table 6. The result of Evaluation measures

Types of gas	Accuracy	Precision	Recall	F-measurement	F_β	Execution time (second)
Gas #1	0.4779	0.5032	0.7129	0.5900	0.5245	2.4878
Gas #2	0.5227	0.4982	1.5354	0.7523	0.5494	2.5358
Gas #3	1.2226	0.5455	2.5074	0.8961	0.6115	3.0889
Gas #4	0.6607	0.4798	1.4007	0.7148	0.5276	3.0782
Gas #5	0.4892	0.5023	0.4955	0.4989	0.5014	2.5627
Gas #6	0.4943	0.5004	1.5158	0.7524	0.5513	3.0828

In Table 7, the results were presented and it was a comparison between the developed method And the traditional method in terms of accuracy and execution time, where the accuracy appeared and the accuracy was 0.93, and it is considered a good accuracy as it can be relied upon in testing the model to know the extent of the model's reliability, and

the execution time took 4.70 It is an almost standard time in order to be useful in testing large models in a short time and useful in shortening the time when the data is large.

Table 7. The compare between the traditional XGBoost and DXGBoost-GSk

# Iteration	XGBoost		DXGBoost	
	Time	Accuracy	Time	Accuracy
1	2.9409520626068115	0.428063104	4.701775074005127ms	0.9368374562608915
2	2.956578493118286	0.387859209	4.702776193618774	0.9368374562608907
3	2.956578493118286	0.245783248	4.702776193618774	0.9368374562608898
4	2.956578493118286	1.452326905	4.702776193618774	0.9368374562608889
5	2.956578493118286	0.665733854	4.702776193618774	0.9368374562608881
6	2.956578493118286	0.59076485	4.702776193618774	0.9368374562608872
7	2.9658281803131104	0.562495346	4.702776193618774	0.9368374562608863
8	2.966827392578125	0.547653308	4.702776193618774	0.9368374562608854
9	2.9678261280059814	0.538508025	4.702776193618774	0.9368374562608847
10	2.9698259830474854	0.532307752	4.702776193618774	0.9368374562608838
11	2.970825433731079	0.527827222	4.702776193618774	0.9368374562608829
12	2.9728243350982666	0.52443808	4.702776193618774	0.936837456260882
13	2.973823070526123	0.521784852	4.702776193618774	0.9368374562608811
14	2.974822998046875	0.51965132	4.702776193618774	0.9368374562608803
15	2.975822925567627	0.517898412	4.702776193618774	0.9368374562608794
16	2.9768221378326416	0.516432615	4.702776193618774	0.9368374562608786
17	2.9778265953063965	0.515188728	4.702776193618774	0.9368374562608777
18	2.978820562362671	0.514119904	4.702776193618774	0.9368374562608769
19	2.9798214435577393	0.513191617	4.702776193618774	0.936837456260876
20	2.980821371078491	0.512377857	4.702776193618774	0.9368374562608751
21	2.981818675994873	0.511658661	4.702776193618774	0.9368374562608742
22	2.9828171730041504	0.511018452	4.702776193618774	0.9368374562608733
23	2.984816312789917	0.510444894	4.702776193618774	0.9368374562608726
24	2.9868156909942627	0.509928094	4.703782081604004	0.9368374562608717
25	2.9878153800964355	0.509460023	4.705773115158081	0.9368374562608708

As for the traditional method, where the best accuracy was 1.45 The worst accuracy was 0.24, which is ok, but its accuracy is less, it is basically unreliable, and the time it took to implement is 2.98. Although it took less implementation time than the developed method and also the accuracy was less than the proposed method, it is not useful, to be accurate.

Figure 2 shows the relationship between the developed method and the traditional method in terms of accuracy and was applied to the number of samples numbering 13910 and the number of columns 129 after applying the correlation to the data so that

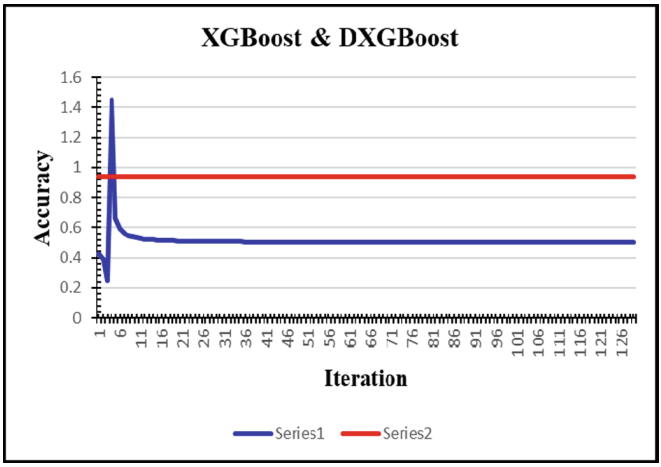


Fig. 2. Compare traditional XGBoost with DXGBoost from aspect accuracy

it becomes a matrix of 129 * 129. After applying the developed method to this matrix, the results shown in the above figure appear.

5 Conclusions

This section presents the most important conclusions reached through applying the HPM-STG into the dataset and focuses on how to avoid the both challenges (programming challenges and application challenges). In addition, we will suggest a set of recommendations for researchers to work on it in the future.

The process of emission of gases as a result of chemical reactions is one of the most important problems that cause air pollution and affect living organisms, although the process of analyzing these gases is a very complex issue and requires a lot of time. But HPM-STG is able to process a large flow of data in a small time.

The data used in this research characteristic as very huge and split into multi groups related to 10 months, therefore at the first; aggregation of all data in a single dataset, and find the data have high duplication therefore handle this problem by take only the different interval to work on it, this step reduces the computation.

The correlation used in that model to determine which features from the 128 related to sensors are more affect in determining the type of gases. In general, we found the following:

- The sensors more affect to determine the first gas is (FD1) in the first order and in the second-order (F23, FC1) while the not important sensors are (F05, F24, F25, F32) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the second gas (F63, FF3) in the first order and in the second-order are (F73, FA3, FE3) while the not important sensor is (F58) therefore to reduce the computation can be neglected.

- The sensors more affect to determine the second gas (FD3, FF3) in the first order and in the second-order is (FE3) while the not important sensors are (F06, F07, F08) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the second gas (FF3) in the first order and in the second-order is (FE3) while the not important sensors are (F06, F07, F08) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the fifth gas are (F31, F63) in the first order and in the second-order (FE3, FF3, FF7) while the not important sensor is (F12) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the fifth gas are (F21, F63, FE4) in the first order and in the second-order (F73, FB1, FF4) while the not important sensor is (F12) therefore to reduce the computation can be neglected.

GSK is one of the pragmatic tools to work with real data, where, GSK characteristic thorny working in parallel and give high accuracy. In general; it is based on three parameters (Ackley function, Junior Phase, Senior Phase). Therefore, replacing the kernel of XGBoost with GSK are get high accuracy results but on the other side, the computation is increased. To reduce implementation time.

This work avoids the main drawbacks of XGBoost; where the kernel of XGBoost is the Decision tree, this makes it need to determine the root; depth of the tree, In addition to high complexity. Through replace the kernel of it with GSK, enhance the performance of that algorithm from two points: reduce the implementation time and enhancement the performance. We can used the following idea for development this work in the futures

- It is possible to use another optimization algorithm that depends on the Agent principle as the kernel of the XGBoost algorithm, such as the Whale algorithm, the Lion algorithm, and the Practical swarm algorithm.
- The HPM-STG implementation on CPU as hardware while; we can implement on other hardware such as GPU or FPGA.
- It is also possible to use other types of sensors to study the effect of the emitted gas on the development of certain bacteria growth.
- It is possible to use another technology for the classification process such as the Deep learning algorithm represented by Long Short-Term Memory (LSTM).

References

1. Abad, A.R.B., et al.: Robust hybrid machine learning algorithms for gas flow rates prediction through wellhead chokes in gas condensate fields. Fuel **308**, 121872 (2022). https://doi.org/10.1016/j.fuel.2021.121872
2. Al-Janabi, S., Mahdi, M.A.: Evaluation prediction techniques to achievement an optimal biomedical analysis. Int. J. Grid Util. Comput. **10**(5), 512–527 (2019).https://doi.org/10.1504/ijguc.2019.102021
3. Alkaim, A.F., Al_Janabi, S.: Multi objectives optimization to gas flaring reduction from oil production. In: International Conference on Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems, pp. 117–139. Springer, Cham (April 2019). https://doi.org/10.1007/978-3-030-23672-4_10

4. Al-Janabi, S., Alkaim, A., Al-Janabi, E., et al.: (2021) Intelligent forecaster of concentrations (PM2.5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Comput. Appl. **33**, 14199–14229.https://doi.org/10.1007/s00521-021-06067-7
5. Al-Janabi, S., Alkaim, A.F.: A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft. Comput. **24**(1), 555–569 (2020)https://doi.org/10.1007/s00500-019-03972-x
6. Al-Janabi, S., Alkaim, A.F., Adel, Z.: An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft. Comput. **24**, 10943–10962 (2020)https://doi.org/10.1007/s00500-020-04905-9
7. Al_Janabi, S., Al_Shourbaji, I., Salman, M.A.: Assessing the suitability of soft computing approaches for forest fires prediction. Appl. Comput. Inf. **14**(2): 214–224 (2018). ISSN 2210-8327https://doi.org/10.1016/j.aci.2017.09.006
8. Chung, D.D.: Materials for electromagnetic interference shielding. Mater. Chem. Phys., 123587 (2020)https://doi.org/10.1016/j.matchemphys.2020.123587
9. Cotfas, L.A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D.S., Tajariol, F.: The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access **9**, 33203–33223 (2021).https://doi.org/10.1109/ACCESS.2021.3059821
10. da Veiga, A.P., Martins, I.O., Barcelos, J.G., Ferreira, M.V.D., Alves, E.B., da Silva, A.K., Barbosa Jr., J.R., et al.: Predicting thermal expansion pressure buildup in a deepwater oil well with an annulus partially filled with nitrogen. J. Petrol. Sci. Eng. **208**, 109275 (2022)https://doi.org/10.1016/j.petrol.2021.109275
11. Fernandez-Vidal, J., Gonzalez, R., Gasco, J., Llopis, J. (2022). Digitalization and corporate transformation: the case of European oil & gas firms. Technol. Forecast. Soc. Chang. **174**, 121293.https://doi.org/10.1016/j.techfore.2021.121293
12. Foroudi, S., Gharavi, A., Fatemi, M.: Assessment of two-phase relative permeability hysteresis models for oil/water, gas/water and gas/oil systems in mixed-wet porous media. Fuel **309**, 122150 (2022). https://doi.org/10.1016/j.fuel.2021.122150
13. Gao, Q., Xu, H., Li, A.: The analysis of commodity demand predication in supply chain network based on particle swarm optimization algorithm. J. Comput. Appl. Math. **400**, 113760 (2022). https://doi.org/10.1016/j.cam.2021.113760
14. Gonzalez, D.J., Francis, C.K., Shaw, G.M., Cullen, M.R., Baiocchi, M., Burke, M.: Upstream oil and gas production and ambient air pollution in California. Sci. Total Environ. **806**, 150298 (2022). https://doi.org/10.1016/j.scitotenv.2021.150298
15. Al-Janabi, S., Alkaim, A.: A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis. Egypt. Inf. J. (2022).https://doi.org/10.1016/j.eij.2022.01.004
16. Al-Janabi, S.: Overcoming the main challenges of knowledge discovery through tendency to the intelligent data analysis. In: 2021 International Conference on Data Analytics for Business and Industry (ICDABI), pp. 286–294 (2021)https://doi.org/10.1109/ICDABI53623.2021.9655916
17. Gupta, N., Nigam, S.: Crude oil price prediction using artificial neural network. Procedia Comput. Sci. **170**, 642–647 (2020). https://doi.org/10.1016/j.procs.2020.03.136
18. Hao, P., Di, L., Guo, L.: Estimation of crop evapotranspiration from MODIS data by combining random forest and trapezoidal models. Agric. Water Manag. **259**, 107249 (2022).https://doi.org/10.1016/j.agwat.2021.107249
19. Al-Janabi, S., Rawat, S., Patel, A., Al-Shourbaji, I.: Design and evaluation of a hybrid system for detection and prediction of faults in electrical transformers. Int. J. Electr. Power Energy Syst. **67**, 324–335 (2015)https://doi.org/10.1016/j.ijepes.2014.12.005
20. Houssein, E.H., Gad, A.G., Hussain, K., Suganthan, P.N.: Major advances in particle swarm optimization: theory, analysis, and application. Swarm Evol. Comput. **63**, 100868 (2021). https://doi.org/10.1016/j.swevo.2021.100868

21. Johny, J., Amos, S., Prabhu, R.: Optical fibre-based sensors for oil and gas applications. Sensors **21**(18), 6047 (2021). https://doi.org/10.3390/s21186047
22. Mahdi, M. A., & Al-Janabi, S.: A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers. In: International Conference on Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems, pp. 320–339. Springer, Cham (April 2019). https://doi.org/10.1007/978-3-030-23672-4_23
23. Kadhuim, Z.A., Al-Janabi, S.: Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis. Results Eng. **17** (2023). https://doi.org/10.1016/j.rineng.2022.100847
24. Mohammadpoor, M., Torabi, F.: Big Data analytics in oil and gas industry: an emerging trend. Petroleum **6**(4), 321–328 (2020). https://doi.org/10.1016/j.petlm.2018.11.001
25. Mohammed, G.S., Al-Janabi, S.: An innovative synthesis of optmization techniques (FDIRE-GSK) for generation electrical renewable energy from natural resources. Results Eng. **16** (2022). https://doi.org/10.1016/j.rineng.2022.100637
26. Ali, S.H.: A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining. In: IEEE,6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, pp. 951–961 (2012).https://doi.org/10.1007/978-90-313-8424-2_10



PMFRO: Personalized Men's Fashion Recommendation Using Dynamic Ontological Models

S. Arunkumar[1], Gerard Deepak[2(✉)], J. Sheeba Priyadarshini[3], and A. Santhanavijayan[4]

[1] Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

[2] Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India
gerard.deepak.christuni@gmail.com

[3] Deparment of Data Science, CHRIST (Deemed to Be University), Bengaluru, India

[4] National Institute of Technology, Tiruchirappalli, India

Abstract. There is a thriving need for an expert intelligent system for recommending fashion especially focusing on men's fashion. As it is an area which is neglected both in terms of fashion and modelling intelligent systems. So, in this paper the PMFRO framework for men's recommendation has been put forth which indicates the semantic similarity schemes with auxiliary knowledge and machine intelligence in a very systematic manner. The framework intelligently creates mapping of the preprocessed preferences and the user records and clicks with that of the items in the profile. So, this model aggregates community user profiles and also maps the men's fashion ontology using strategic semantic similarity schemes. Semantic similarity is evaluated using Lesk similarity and NPMI measures at several stages and instances with differential set thresholds and the dataset is classified using the feature control, machine learning bagging classifier which is an ensemble model in order to recommend the men's fashion. The PMFRO framework is an intelligent amalgamation and integration of auxiliary knowledge, strategic knowledge, user profile preferences as well as machine learning paradigms and semantic similarity models for recommending men's fashion and overall precision of 94.68% and FDR of 0.06 was achieved using the PMFRO model.

Keywords: Fashion Recommendation · Men's Fashion Recommendation · Ontology · Semantically Driven · User Click Records

1 Introduction

In today's digital world online shopping has set a huge foot in people's lifestyle. It eases the tiring process. E-commerce websites have used this to their advantage and have placed a very strong foothold in e-shopping especially in fashion industry. E-commerce websites are rated based on "How they present themselves to the user" i.e., recommendation system. For example, amazon's 'Item to Item collaborative filtering' is a forerunner

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 96–105, 2023.
https://doi.org/10.1007/978-3-031-27409-1_9

among recommendation systems as it secures a significant amount of user's preferences. It predicts a given customer's preferences on the basis of other customers i.e., collaborative process. These companies rigorously try to find "How good could you recommend fashionable entities?". This is important because a user would be satisfied mostly to his preferrable choice of fashion sense which pressurizes the need for an impeccable recommendation system. There should also be a consideration of the range of variety of preferences of the masses (From a typical conservative to a trendy neophile). So, the recommendation system should not be stereotyped to a particular way of suggestion, rather it should be inclusive to all kinds of people. Thus, the recommendation system needs to be tuned accordingly. So, the help was sought from leading fashion experts for the fashion ontology which is used in the classifier. The primary focus is on gender specific recommendation systems (men's fashion recommendation system in this paper). This recommendation system depends on the user's dynamic record clicks and past user preferences. These user record clicks approximately reflect the user's choice of interest(preference) which is the base of any recommendation model. The assumption is these user record clicks provide more accuracy on user's preference thus enhancing the recommendations. So, after consulting 146 fashion experts from various universities and organization to derive the ground truth about the contemporary fashion sense and fashion preferences to derive the ontology accordingly.

Motivation: Recommendation systems are the need of the hour because of raise in the entities over the internet, increase in data, exponential increase in digital transformation. Recommendation system of fashionable entities are scarce and underdeveloped despite the increase in demand and the surge in usage. These recommendation systems facilitate the user's choice in accordance to their preferences which can save time for the user. It also could be a driving factor that keeps the user engaged with the e-commerce website based on the satisfaction of user's previous usage. The world wide web reigns semantically inclined framework strategies which is knowledge centric is required to suit the needs of the web.

Contribution: The noble contribution of the framework includes classification of dataset using an ensemble bagging model with decision trees and random forest classifiers as independent classifiers. The ontology alignment is achieved using Lesk similarity and cosine similarity and ontology alignment happens between the terms obtained from dynamic user record clicks, past preferences and men's fashion ontology. The semantic similarity is evaluated using NPMI measure with differential threshold at several instances. The intelligent integration of community user profiles, user preference terms and items in the profile mapping of men's fashion ontology with the classified instances and computation of semantic similarity paradigms with differential thresholds is achieved in the model. Precision%, recall%, accuracy% and F-measure% is increased and False Discovery Rate (FDR) is decreased compared against the other baseline models.

The remaining part of the paper is presented under the following sections. The second section describes Related Work. The Proposed System Architecture is detailed in Sect. 3. The Results and Performance Analysis are shown in Sect. 4. Finally, Sect. 5 brings the paper to a conclusion.

2 Related Works

This paper has primarily referred and compared the proposed PMFRO model with VAFR model [1], FRVCR model [2] and DeepCDFR model [3]. VAFR model proposed by Kang et al., [1] put forth that performance of the recommendation can be considerably raised by directly studying fashion conscious image representations, i.e., by honing the representation of images and the system jointly thus they are able to show improvements over techniques such as Bayesian Personalized Ranking(BPR) and variants that utilize the pretrained visual features.FRVCR model proposed by Ruiping et al., [2] put forth a fashion compatibility knowledge learning method that integrates the visual compatibility relationships as well as style-based information. They also suggest a fashion recommendation method with domain adaptation strategy to relieve the distribution gap between items in target domain and items of external compatible outfits.DeepCDFR model proposed by Jaradat et al., [3] tries to solve the problem of complex recommendation possibilities that involve transfer of knowledge across multiple domains.The techniques used to accomplish this work encompass both architectural and algorithm design using deep learning technologies to scrutinize the effect of deep pixel-wise semantic segmentation and integration of text on recommendations quality.

Many researchers have proposed various types of recommendation systems. The approach differs drastically based on what you recommend. As there is discussion about Fashion recommendation system these are some works that this paper has referred to Hong et al., [8] have suggested a fabric suggestion algorithm based on perception which uses a computational model based on Fuzzy AHP and Fuzzy TOPSIS algorithms. This is integrated with a collaborative design process. Thus, the recommendation system uses a hierarchical interactive structure. Cosley et al., [9] have written about how recommendation system affects a user's opinion. The paper has a psychological approach on user's choice and the extent of recommendation system's influence and manipulation on user's choice. Thus, they could model their recommendation system accordingly This also proves the need for the recommendation system.Tu et al., [14] have proposed a novel Personalized intelligent fashion recommender. In this paper they have proposed three standalone models i) Recommendation models based on interaction ii) Apparel multimedia mining model with evolutionary hierarchies iii) Model for analyzing color tones. Zhou et al., [15] built mapping relations using the perceptual image of the user between design components of apparel, partial least squares as well as semantic differential to create a personalized online clothing shopping recommendation system. In [16–23] several models in support of the proposed literature have been depicted.

3 Proposed System Architecture

Figure 1 depicts the proposed system architecture for a framework to recommend men's fashion based on the user preferences and user record clicks. This is a user driven framework or a user driven model which is driven by user preferences and the recorded user clicks. The user record clicks are the previous history of user preferences (web usage data) of the current user profile. Previous web usage data is taken, his previous click through data (user record clicks) is taken as well as client's dynamic clicks are recorded. These

user clicks and the user preferences in the past history of the user profile are subjected to preprocessing (which involves Stop word removal, lemmatization, tokenization, and named entity recognition (NER)). So once the preprocessing of the user record clicks, dynamic user record clicks and the past user preferences are done, the individual terms T_n are obtained. Further the dataset is obtained.

Also, the community profiles refer to those profiles of several users who are participating in e-shopping and e-commerce recommendations on fashion websites such web usage data of 146 users who were experts in fashion were selected and their user profiles for men's fashion over the period of two weeks were collected. And those community profiles were again subjected to preprocessing which involves Stop word removal, lemmatization, tokenization, and named entity recognition (NER). And individual items in the user profiles along with the set fashion entities is stored in Hash set. And the items in the Hash set are further mapped with the terms preprocessed (i.e. T_n) using the cosine similarity with the threshold of 0.5. The reason for keeping the cosine similarity threshold as 0.5 is mainly due to the fact that large number of entities has to be aligned from the items in the profile and the terms also owing to the shallow number of items and the terms, the mapping is done liberally with a threshold of 0.5.The similarity between any documents or vectors is assessed using cosine similarity (1). It is the cosine of the angle formed by two vectors when they are projected in three dimensions.

$$\cos(x, y) = x \cdot y / ||x|| * ||y|| \tag{1}$$

Subsequently the mapped entities between the items in the community user profiles and the terms obtained from the current user click preferences is further mapped with men's fashion ontology which is modelled Men's fashion ontology is a perfect domain expert contributed ontology with consultation of several fashion experts who were 144 candidates who were first, second and final year undergrad as well as first year masters in fashion designing apparel technology courses specialized in men's fashion. From them the ground truth was collected on men's fashion on several occasions and several themes and a proper men's fashion ontology was formulated using web protégé. And this ontology is mapped with the entities resultant from mapping between the terms obtained (T_n) and the items in the profiles. This mapping is done again with the help of cosine similarity at a higher threshold of 0.75 in order to make sure that the entity entitlement takes place much more precisely. Finally, the features obtained from this resultant mapping is passed into the bagging classifier which uses decision trees and random forest classifiers.

Decision trees and random forest classifiers are used for bagging, features which are resultant from the second phase of mapping between the initially mapped entities and terms of the men's fashion ontology. Owing the shallow number of features, the features are passed randomly into the bagging classifier with decision trees and random forest classifier as the independent coherent classifiers. The dataset is classified and the classified instances are yielded for each of the classifier under each class we are computing the semantic similarity among the classified instances out of bagging classifier and entities aligned to in the initial obtained terms T_n and men's fashion ontology. Subsequently the term T_n obtained initially by preprocessing the user record clicks, dynamic user record clicks and the past user preferences and men's fashion ontology

are aligned by the Lesk similarity keeping the threshold as 0.75. The outcome of this alignment is further used to calculate the NPMI between this and that of the classified instances out of the bagging classifier.

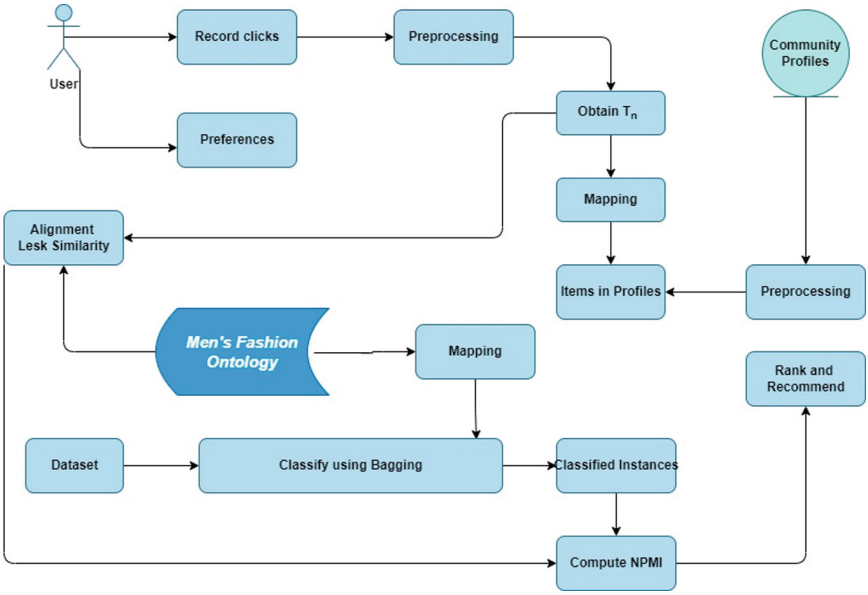


Fig. 1. PMFRO Architecture

The threshold for NPMI(3) is set as 0.5 because only positive values between NPMI(3) is taken (between 0 and 1), the threshold is set as mid (0.5) in order to increase the number of recommendations because already previous alignment using Lesk similarity has been done. The outcome of the NPMI(3) is ranked and further recommended to the user as the query facets and along with the men's fashion which is identified as the set for the recorded user click theme. Both the query facet as well as the expanded terms for the query and the respective attires together in terms of images are yielded to the user and that is further subjective for shopping or not handled by the e-commerce and business process UI.The Pointwise Mutual Information (PMI) is defined as the linear correlation between a characteristic and a class is measured by pointwise mutual information (PMI) and is depicted by Eq. (2). It is standardised between $[-1, +1]$, with -1 (within the limit) for never arising together, 0 for independence, and $+1$ for comprehensive co-occurrence. The Normalized Pointwise Mutual Information is depicted by Eq. (2).

$$pmi(X = x; Y = y) = \log\left[\frac{P(X = x, Y = y)}{P(X = x).P(Y = y)}\right] \tag{2}$$

$$npmi(x = x; Y = y) = \log\left[\frac{pmi(X = x; Y = y)}{h(X = x; Y = y)}\right] \tag{3}$$

where $h(X = x, Y = y)$ is the mutual self-information, which is calculated to be $-\log_2 p(X = x, Y = y)$.

Decision Tree is a Supervised machine learning method, that consists of nodes and branches. The internal nodes constitute the characteristic of the dataset, branches constitute rules for decisions and each and every leaf node constitutes a outcome. These Decision trees have 2 nodes (Decision Node and Leaf Node). Any choice is made using decision nodes and these decision nodes have numerous branches, where Leaf nodes are the result of those choices and they don't branch out any more. The tests are graded based on the characteristics of a given dataset. It is a schematic illustration of all possible outcomes to a problem or decision based on the given instances. Random Forest classifier which has a number of decision trees that operates as an ensemble on numerous subsets of the dataset usually trained with bagging. Random Forest classifier reduces overfitting of training data.

4 Implementation and Performance Evaluation

The implementation was done using i5 processor with 32GB RAM using google's colaboratory as the primary integrated development toolkit. The python's natural language toolkit (NLTK) was used for performing the preprocessing NLP tasks (Stop word removal, lemmatization, tokenization, and named entity recognition (NER)). The Ontology was manually modelled using Web Protégé and automatically generated with OntoCollab as a tool.The dataset used for implementation were standard datasets which were intelligently integrated and expanded by finding common annotations. If annotations were not common, they were integrated successively one after the other and ensured that all these documents yielded from these datasets were annotated, labeled with at least two annotations and labels i.e., the categories indicating these datasets were present in the final integrated dataset. The datasets included Myntra Men's Product Dataset [23], United States - Retail Sales: Men's Clothing Stores [24], Most popular fashion and clothing brands among men in Great Britain 2021 [25], Index of Factory Employment, Men's Clothing for United States [26].The put forth PMFRO framework was queried for 1782 queries whose ground truth has been collected from several fashion bloggers, fashion designing students and fashion experts who were aware of men's fashion. The number of consultant people where 942 people from several colleges and universities and the facts have been gathered and validated from them. And, in order to calculate & verify the performance of the suggested PMFRO model, the baseline models were also evaluated for the exact same dataset for the exact same no of queries as the proposed PMFRO framework.

Proposed PMFRO which is a personalized scheme for men's fashion recommendation is evaluated using precision%, recall%, accuracy%, F-measure% & false discovery rates (FDR) as potential metrics. From Table 1 it is clear that PMFRO yields 94.68% of overall average precision%, 97.45% average recall%, 96.06% average accuracy%, 96.04% average F-measure% with FDR of 0.06. Precision%, recall%, accuracy% with F-measure% yields the relevance of the recommendations and the False discovery rate (FDR) quantifies the no of false positives which are produced are furnished by this model. From Table 1 it is expressive that the proposed PMFRO is baselined

with VAFR [1], FRVCR [2] & DeepCDFR [3] models The VAFR [1] yields 90.23% of overall average precision%, 92.63% average recall%, 91.43% average accuracy%, 91.41% average F-measure% with FDR of 0.10. The FRVCR [2] yields 89.44% of overall average precision%, 93.63% average recall%, 91.53% average accuracy%, 91.48% average F-measure% with FDR of 0.11. The DeepCDFR [3] yields 88.12% of overall average precision%, 92.16% average recall%, 90.14% average accuracy%, 90.09% average F-measure% with FDR of 0.12.

Table 1. Performance Evaluation of PMFRO Model with the other baseline models

Model	Average of precision % P	Average of recall % R	Average of Accuracy % (P + R)/2	Average of F-Measure % (2*P*R)/(P + R)	FDR 1-Precision
VAFR [1]	90.23	92.63	91.43	91.41	0.10
FRVCR [2]	89.44	93.63	91.53	91.48	0.11
DeepCDFR [3]	88.12	92.16	90.14	90.09	0.12
Proposed PMFRO	94.68	97.45	96.06	96.04	0.06

The PMFRO has yielded the highest precision%, recall%, accuracy%, F-measure% & FDR's lowest value when evaluated against the baseline models. The reason why the PMFRO performs preferable than the baseline models because it is motivated by men's fashion ontology which is dynamically generated & mapping of ontology happens with the mapping of user preferences from past user record clicks. Apart from this, the community profiles of fashion stars & fashion experts along with the men's fashion ontology which is validated by fashion experts ensures that the right amount of auxiliary knowledge pertaining to men's fashion is prioritized and added to the model. Importantly the usage of the bagging classifier for the classification of dataset based upon the features obtained by the means of the ontology alignment, community profile contribution & user's past profile visits ensures that the feature bagging which is a strong ensemble classifier which is a feature control. Feature controlling (in machine learning model like bagging) makes sure the relevance is kept in track with user's relevance. The semantic similarities are computed using cosine similarity and Lesk similarity. The precision% vs no of recommendations distribution curve is depicted in Fig. 2 which is the line graph distribution for precision% vs number of recommendations for the proposed architecture and the baseline models and it is indicative whether the PMFRO occupies the highest hierarchy followed by other models. The second in the hierarchy is the VAFR model [1] The third in the hierarchy is the FRVCR model [2]. The lowermost in the hierarchy is the DeepCDFR model [3] in terms of precision %.

The Lesk similarity of ontology alignment and cosine similarity with various thresholds into the framework ensures that there is a strong relevance computational mechanism which is evidently present in the model. That is why the proposed PMFRO yields better

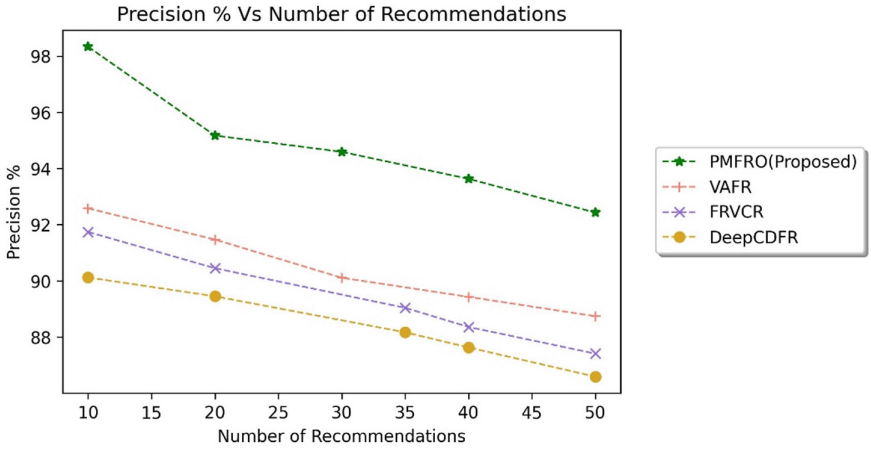


Fig. 2. Accuracy% vs No of Recommendations

results when compared to the baseline models. The reason why the VAFR model [1] doesn't perform as expected compared to the proposed model is because it is visually aware which takes into consideration about the visual features. Apart from this Siamese CNNs are used for classification. But the amount of moderated auxiliary knowledge into the model is minimalistic when compared to the proposed model and henceforth the VAFR model [1] doesn't perform as expected. The reason why the FRVCR model [2] doesn't perform as expected compared to the proposed model is because visual compatibility relationship is the key. Visual compatibility takes care of a high restrictive knowledge that is generated there which is shallow. Apart from this the relevance computational mechanisms in the FRVCR model [2] is not very strong which is why the FRVCR model [2] doesn't perform as expected compared to the proposed model. The reason why the DeepCDFR model [3] drastically lags compared to PMFRO model is because semantic segmentation of images was given more priority. So the entire model was made visual driven where the textual inputs were to be mapped with visual features. This feature mapping makes very complex. Instead, when it becomes an annotation driven model with expert opinion in terms of cognitive ontologies and user clicks would drive this better. Apart from this the laguna of textual knowledge with a deep learning model ensures that there is underfitting of textual content and henceforth the DeepCDFR model [3] also does not perform as expected when compared to the PMFRO model. Owing to all these reasons, and since the proposed model comprises of quality auxiliary knowledge with a strong relevance computation mechanism and feature control bagging classifier Thus the proposed PMFRO performs better than other baseline models.

5 Conclusion

This report contains successfully suggested a recommendation system which depends on user's dynamic record clicks and past preferences. This paper ensures that ensemble techniques and semantic similarity techniques yield better. And also have evaluated this recommendation model and compared this recommendation model against other baseline models and the outcomes reveal that the proposed model is comparatively better than other baseline models. The Dynamic generation and mapping of Ontology enhance the efficiency of the proposed model. This model is based on the ground truth of fashion sense from fashion experts. Thus, PMFRO is an annotation driven model with expert opinion in terms of cognitive ontologies. Better recommendations satisfy the customer's needs resulting in growth of business. Thus, a better model has been proposed and evaluated.

References

1. Kang, W., Fang, C., Wang, Z., McAuley, J.: Visually-aware fashion recommendation and design with generative image models. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 207–216 (2017). https://doi.org/10.1109/ICDM.2017.30
2. Yin, R., Li, K., Lu, J., Zhang, G.: Enhancing fashion recommendation with visual compatibility relationship. In: The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, pp. 3434–3440 (2019)
3. Jaradat, S.: Deep cross-domain fashion recommendation. In: Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17), pp. 407–410. Association for Computing Machinery, New York, NY, USA (2017)
4. Hwangbo, H., Kim, Y.S., Cha, K.J.: Recommendation system development for fashion retail e-commerce. Electron. Commer. Res. Appl. **28**, 94–101 (2018)
5. Stefani, M.A., Stefanis, V., Garofalakis, J.: CFRS: a trends-driven collaborative fashion recommendation system. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–4. IEEE (2019)
6. Shin, Y.-G., Yeo, Y.-J., Sagong, M.-C., Ji, S.-W., Ko, S.-J.: Deep fashion recommendation system with style feature decomposition. In: 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), pp. 301–305. IEEE (2019)
7. Liu, S., Liu, L., Yan, S.: Magic mirror: an intelligent fashion recommendation system. In: 2013 2nd IAPR Asian Conference on Pattern Recognition, pp. 11–15. IEEE (2013)
8. Hong, Y., Zeng, X., Bruniaux, P., Chen, Y., Zhang, X.: Development of a new knowledge-based fabric recommendation system by integrating the collaborative design process and multi-criteria decision support. Text. Res. J. **88**(23), 2682–2698 (2018)
9. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing? How recommender system interfaces affect users' opinions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 585–592 (2003)
10. Nakamura, M., Kenichiro, Y.: A study on the effects of consumer's personal difference on risk reduction behavior and internet shopping of clothes. Chukyo Bus. Rev. **10**, 133–164 (2014)
11. Wang, H., Wang, N.Y., Yeung, D.Y., Unger, M.: Collaborative deep learning for recommender systems. In: ACM KDD'15, pp. 1235–1244 (2015)
12. Yethindra, D.N., Deepak, G.: A semantic approach for fashion recommendation using logistic regression and ontologies. In: 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1–6. IEEE (2021)

13. Tian, M., Zhu, Z., Wang, C.: User-depth customized men's shirt design framework based on BI-LSTM. In: 2019 IEEE International Conference on Mechatronics and Automation (ICMA), pp. 988–992. IEEE (2019)
14. Tu, Q., Dong, L.: An intelligent personalized fashion recommendation system. In: 2010 International Conference on Communications, Circuits and Systems (ICCCAS), pp. 479–485. IEEE (2010)
15. Zhou, X., Dong, Z.: A personalized recommendation model for online apparel shopping based on Kansei engineering. Int. J. Cloth. Sci. Technol. (2017)
16. Surya, D., Deepak, G., Santhanavijayan, A.: KSTAR: a knowledge based approach for socially relevant term aggregation for web page recommendation. In: International Conference on Digital Technologies and Applications, pp. 555–564. Springer, Cham (January 2021)
17. Aditya, S., Muhil Aditya, P., Deepak, G., Santhanavijayan, A.: IIMDR: intelligence integration model for document retrieval. In: International Conference on Digital Technologies and Applications, pp. 707–717. Springer, Cham, (January 2021)
18. Varghese, L., Deepak, G., Santhanavijayan, A.: A fuzzy ontology driven integrated IoT approach for home automation. In: International Conference on Digital Technologies and Applications, pp. 271–277. Springer, Cham, (January 2021)
19. Surya, D., Deepak, G., Santhanavijayan, A.: Ontology-based knowledge description model for climate change. In: International Conference on Intelligent Systems Design and Applications, pp. 1124–1133. Springer, Cham (December 2020)
20. Manoj, N., Deepak, G.: ODFWR: an ontology driven framework for web service recommendation. In: Data Science and Security, pp. 150–158. Springer, Singapore (2021)
21. Singh, S., Deepak, G.: Towards a knowledge centric semantic approach for text summarization. In: Data Science and Security, pp. 1–9. Springer, Singapore (2021)
22. Roopak, N., Deepak, G., Santhanavijayan, A.: HCRDL: a hybridized approach for course recommendation using deep learning. In: Abraham, A., Piuri, V., Gandhi, N., Siarry, P., Kaklauskas, A., Madureira, A. (eds.) ISDA 2020. AISC, vol. 1351, pp. 1105–1113. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71187-0_102
23. Palvannan, S., Deepak, G.: TriboOnto: a strategic domain ontology model for conceptualization of tribology as a principal domain. In: International Conference on Electrical and Electronics Engineering, pp. 215–223. Springer, Singapore (2022)
24. Myntra Men's Product Dataset Men's Fashion Dataset
25. United States-Retail Sales: Men's Clothing Stores
26. Most popular fashion and clothing brands among men in Great Britain 2021
27. Index of Factory Employment, Men's Clothing for United States M08092USM331SNBR

Hybrid Diet Recommender System Using Machine Learning Technique

N. Vignesh[1], S. Bhuvaneswari[1], Ketan Kotecha[2], and V. Subramaniyaswamy[1](✉)

[1] School of Computing, SASTRA Deemed University, Thanjavur 613401, India
s.bhuvana@sastra.ac.in, vsubramaniyaswamy@gmail.com
[2] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India
head@scaai.siu.edu.in

Abstract. Obesity is a dangerous epidemic worldwide and is the root cause of many diseases. It is difficult for people to have the same diet with an optimized calorie intake as it becomes monotonous and boring. It will be much better if a dynamic diet can be generated depending upon the calories burnt by a person and their current Body Mass Index (BMI). The active diet planner could provide a person with some change regarding the food consumed and, at the same time, regulate the calorie intake depending upon the user's requirements. Previously proposed models are either focused only on one aspect of the nutritional information of food or on presenting a diet for a specific issue which is presently facing by user. The proposed system utilizes more balanced approach that focuses on most of the nutritional features of food, and can recommend different foods to a user depending on their BMI. The fat, carbohydrate, calorie, and protein content of food and the BMI of the user are considered while preparing the diet chart. K-means clustering is used to cluster food of similar nutritional content, and a random forest classifier is then used to build the model to recommend a diet for the user. The result of the system cannot be compared with a standard metric. Still, some of the factors that influence the performance of the diet recommender system include the truthfulness of the user while providing information to the design and the accuracy at which the parameters for the model had been set. The advantage of the system comes from the fact that the user has more options to choose from within their suitable range.

Keywords: Recommender system · BMI · Diet Chart · Machine learning · K-Means clustering · Random Forest classifier

1 Introduction

Obesity is a common, severe, and costly disease. Worldwide obesity has nearly tripled since 1975, and from data collected in 2016, it has seen that more than 1.9 billion adults were overweight, of which over 650 million were obese. Overweight and obesity are abnormal or excessive fat accumulation that may impair health. The fundamental cause is due to energy imbalance between calories consumed and calories expended. To handle

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 106–115, 2023.
https://doi.org/10.1007/978-3-031-27409-1_10

this situation, people who affected by obesity are depend fully on maintain diet to lead healthy lifestyle [1].

A diet can be maintained for weight loss, but it can lead to malnutrition if the diet is not planned correctly. Most of the available common diet plan generators only focus on providing static diet charts which may not account for dynamic diet plans according to user's behaviour [2]. As machine learning is applied across life science applications [axiom], the proposed work extended the idea of machine learning algorithm for dynamic diet chart generation. This dynamic diet chart can be suggested based on their daily calories expended by the users. The proposed system considers history of each user's preferences and stores it for future diet recommendations to provide different diet plans to diverse people.

As initial step, the input data is segregated according to time at which the users are able to consume food. Then, the collected data have been clustered based on the nutritional value of the various foods depending upon which are essential for weight loss, weight gain, or to maintain healthy diet. Afterwards, a popular classifier algorithm, random forest is applied to predict the closest food item with their nutritional value.

The rest of this paper is organized as follows. Section 2 compares previously known methods for diet recommender systems. The following section explains the proposed methodology and the techniques used. The next section covers the implementation and the experimental results of proposed diet recommender system. The final section contains conclusion of the work and the future analysis of the proposed work.

2 Related Works

This section presents existing research work to create a personalized food recommender system. Since the chosen field is widespread and active, only some of the most popular and recent ones are mentioned.

Rachel Year Toledo et al. proposed a system that incorporates a multi-criteria decision analysis tool used in the pre-filtering stage to filter out inappropriate foods for the current user characteristics. It also included an optimization-based step that generates a daily meal plan to recommend food that the user prefers, satisfies their daily requirements, and was not consumed recently [3]. Mohan Shrimal et al. proposed a recommender system that uses collaborative filtering and fuzzy logic. The proposed method can use the user's BMI to monitor their calorie targets and consider their background and preferences to provide food suggestions. Their plan also includes an Android-based pedometer that counts the number of steps taken during a particular workout [4].

CelstineEwendi et al. proposed a deep learning model that uses the user's characteristics like age, weight, calories, fibres, gender, cholesterol, fat, and sodium to detect the specific dish that can be served to an ill person who is suffering from a particular disease. Their model uses machine learning and deep learning algorithms like naïve Bayes, recurrent neural networks, and logistic regression for their implementation [5]. Prithvi Vasireddy proposed a system that implements an autonomous diet recommender bot and uses intelligent automation. The proposed method uses the macros and calories collected from a food database and the input from a user database to provide a specific diet recommended via e-mail. Their system is then scheduled to perform this task at particular time intervals and can be performed for many users with minimal effort [6].

Pallavi Chavan et al. proposed a hybrid recommender system using big data analytics and machine learning. Their research demonstrates the design, implementation, and evaluation of three types of recommender systems: collaborative filtering, content-based, and hybrid models. Their system provides health management by offering users food options based on their dietary needs, preferences in taste, and restrictions [7]. Nadia Tabassum et al. proposed a system to generate a diet plan to help diabetic patients calculate their daily calorie requirements and recommend the most suitable diet plan. The proposed recommender system uses fuzzy logic with macro and micro-level nutrients and individual dietary requirements to determine the optimal diet plan [8].

Samuel Manoharan et al. propose a system that considers the blood sugar level, blood pressure, fat, protein, cholesterol, and age and uses K-Clique embedded deep learning classifier recommendation system to suggest a diet for the patients. The newly proposed systems' accuracy and preciseness were compared with machine learning techniques like Naïve Bayes and logistic regression and deep learning techniques like MLP and RNN [9]. When compared with the models mentioned above, the proposed hybrid diet recommender system manages to increase the accuracy by which the model can optimize the diet plans and increase the range of foods that are available for the user to choose from.

Jong Hun-Kim et al. propose a system that considers user preference, personal information, amount of activity, disease history, and family history to recommend a customized diet [10]. This specific service consists of a single module that draws in nutrients and is adopted by users depending on the user-specified constraints; a separate module is then used here to determine the user's preference, and a scoring module is then generated that provides the score for the diet that was provided. The Soil test report comprises three major nutrients, namely Nitrogen (N), Phosphorus (P), and Potassium (K). We collected 2018–2019 soil reports and fertilizer recommendations as history data. The primary fertilizers recommended by most agricultural experts across various crops are Urea, Single Super Phosphate (SSP) and Unit of Potash (MOP).

3 Proposed Methodology

This section gives an overview of the proposed system—the basic diagram to recommend a diet using BMI is shown in Fig. 1. The food recommendation system generates a diet for the user to help them reduce, gain, or maintain their current Body Mass Index. The system considers the current BMI of the user and recommends a diet depending on the function needed [11].

3.1 Data Collection

Collecting the required datasets is one of the most critical tasks for the system. Most of the already present datasets did not contain all the required information. Thus, food information was scraped from various websites, and a dataset was created with only the required data. The data was collected in an unstructured format, then converted to Comma Separated Values (CSV) file and stored in the local database [13, 14].

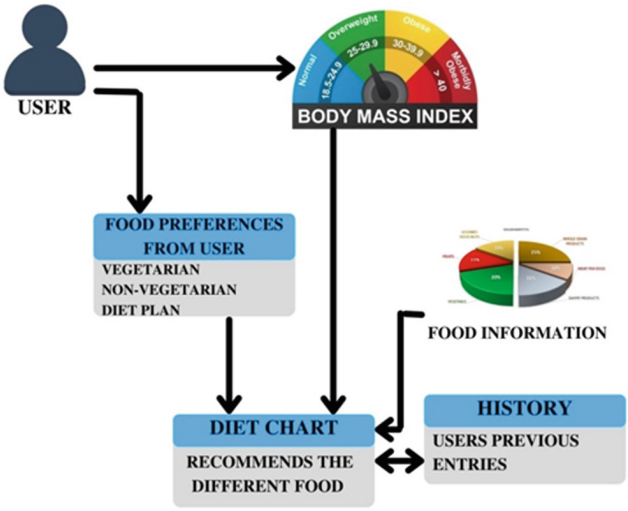


Fig. 1. Overview architecture of the system to recommend diet using BMI

3.2 Data Processing

The data collected in the previous step could be noisy and inconsistent. This leads to the building of a poor-quality model for the system. Hence, it is necessary to overcome this issue. First, data cleaning is required to handle the irrelevant and missing parts of the data. Any missing information can be retrieved from reputable food nutrition websites [15, 16]. Then, each food item in the dataset is assigned a specific six-digit binary number. Each digit represents the time of the day the food can be ingested, including Pre-breakfast, Breakfast, Pre-lunch, Lunch, Tea, and dinner. E.g., 100100.

After the data is pre-processed and accurate and high-quality data is obtained, then the data is clustered according to the timing at which food can be consumed. For this process, K-Means clustering is used. K-Means clustering is an unsupervised learning algorithm that can group unlabeled data into different clusters. It is a convenient way to discover the various categories of an unlabeled dataset. After the data is clustered, a classification algorithm is used to build the model according to the different available functions. The classifier used is a random forest classifier. A random forest classifier contains several decision trees on different subsets of the dataset. It averages the various trees formed to improve the model's predictive accuracy.

Once the dataset was created to be as accurate as possible, each food item was assigned a six-digit binary number denoting the specific food intake time. The six different timings are pre-breakfast, Breakfast, Pre-lunch, Lunch, tea, and Dinner. Any time a particular food can be consumed for a specific time, "1" is used in that spot. Eg.110100. The detailed workflow is illustrated in the Fig. 2 and step wise procedure is explained in procedure 1. The dataset was then clustered using K-Means clustering based on the different timings at which the food can be consumed. The silhouette coefficient was measured to determine the number of clusters that could be formed to provide the best results [17].

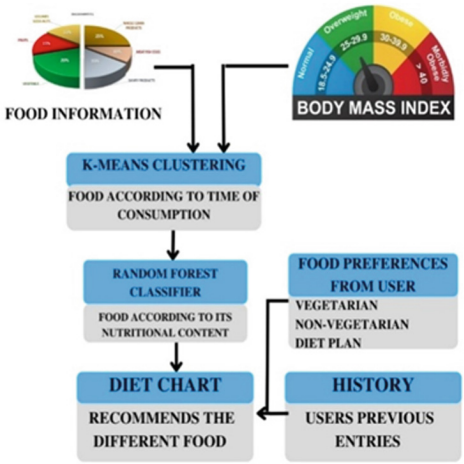


Fig. 2. Workflow of the Hybrid diet recommender system

Procedure 1: Diet Recommender System

Input: User BMI, food preferences, History
Output: Diet plan
Step 1: The food items are segregated depending on consumption.
Step 2: Apply K-Means clustering on the nutrients depending on whether they are helpful for weight loss, weight gain, or maintaining a healthy diet by following sub-steps.
- Step 2.1: Centers of clusters are selected
- Step 2.2: The distance between each data point and cluster is calculated.
- Step 2.3: The data point is assigned to the cluster center, which is minimum compared to all the available cluster centers
- Step 2.4: New cluster centers are then calculated, and the distance between each data point and the newly obtained cluster is obtained
- Step 2.5: If no other data point is reassigned, stop; else, repeat from step 2.3

Step 3: Apply Random Forests classifier to predict the nearest food items based on the diet using below sub-steps.
- Step 3.1: A dataset partition is considered from the whole dataset.
- Step 3.2: For each partition, a decision tree is constructed
- Step 3.3: An output is then obtained for each of the decision trees
- Step 3.4: The final result is then obtained using majority voting.

Step 4: The user receives the required input and runs throughout the model.
Step 5: The obtained output is displayed.

4 Experimental Results and Discussion

The experiments were performed using an Intel i-5 core processor with 8GB RAM. Python IDLE was used for the implementation of the program.

4.1 Dataset

The food information dataset was collected from multiple reputed sources and compiled into a single table. A six-digit binary number was assigned to each food item where each digit represents the time of the day at which the food can be ingested, which includes: Pre-breakfast, Breakfast, Pre-lunch, Lunch, Tea, and dinner. E.g., 100100. A sample of the dataset is shown in Table 1. The dataset was then separated into a training set and a testing set in the ratio of 70:30. A sample of the dataset used for the optimum nutrient constitution is given in Table 2.

Table 1. Sample of Dataset containing food information

Food_ID	Food	Measure	Grams	Calories	Calorie/grams	Protein	Fat	Sat. fat	Fibre	Carbs	Food time
1	Cows' milk	One qt	976	660	0.6762	32	40	36	0	48	110010
2	Milk skim	One qt	984	360	0.3659	36	0	0	0	52	110010
3	Buttermilk	1 cup	246	127	0.5163	9	5	4	0	13	110010
4	Evaporated, undiluted	1 cup	252	345	1.369	16	20	18	0	24	110010
5	Fortified milk	6 cups	1,419	1,373	0.9676	89	42	23	1.4	119	110010

Table 2. Sample of the dataset containing optimum nutritional quantities

Calories	Fats (gm)	Protein s(g)	Iron (mg)	Calcium (mg)	Sodium (mg)	Potassium (mg)	Carbohydrates (gm)
160	15	2	0.55	12	7	485	8.5
89	0.3	1.1	0.26	5	1	358	8.5
349	0.4	14	6.8	190	298	77	8.5

The silhouette coefficient was calculated to estimate the best quantity of clusters considered for K-Means clustering. The silhouette coefficient can be seen in Fig. 3.

The feature importance score was calculated to get the weightage given to a particular nutrient (in Fig. 4). For this model's purposes, the highest priority was given to carbohydrates and fats present in the food. The accuracy score of the proposed hybrid system when compared with other previously mentioned models using other machine learning methods is given in Fig. 5. The hybrid system is compared with MLP (Multilayer perceptron), RNN (Recurrent Neural Networks), and LSTM (Long Short-Term Memory). The comparison is given in Figs. 6 and 7.

The graphs show that the proposed hybrid system can slightly improve the previously present model's accuracy. This leads to more substantial improvements for the user using the system to retrieve a diet. The result of the system cannot be compared with a standard metric. Still, some factors that influence the diet recommender system's performance

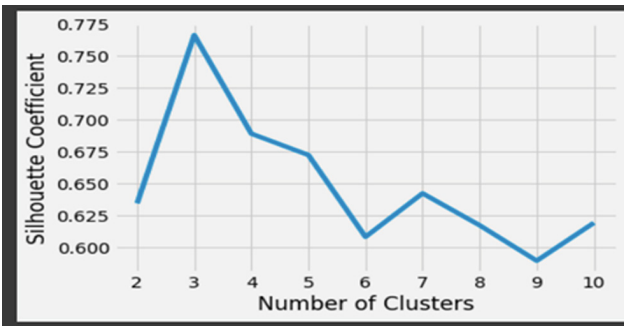


Fig. 3. Calculation of Silhouette coefficient

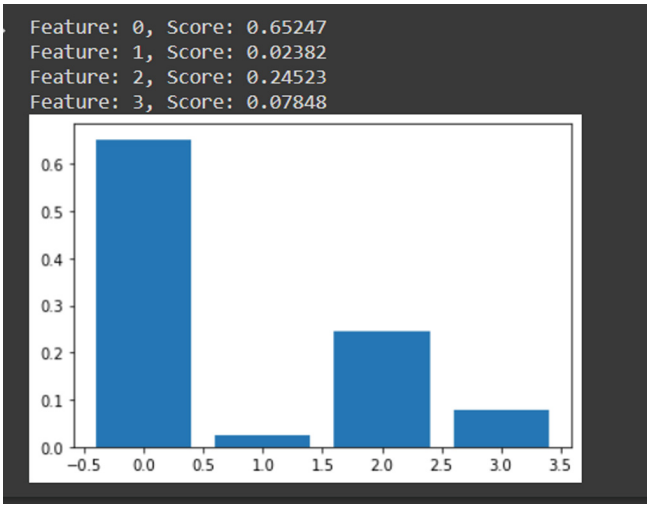


Fig. 4. Feature importance scores for the classifier

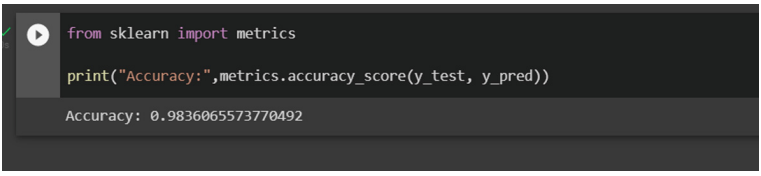


Fig. 5. Accuracy Score of the System

include the user's truthfulness while providing information to the design and the accuracy at which the parameters for the model had been set. A sample of the output obtained can be seen in Fig. 8.

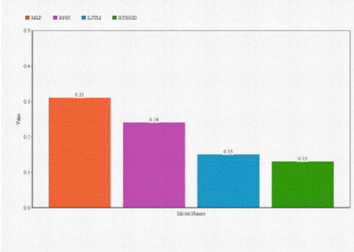


Fig. 6. Accuracy comparison of the hybrid system

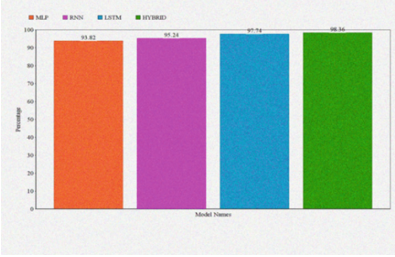


Fig. 7. Error comparison of the hybrid system

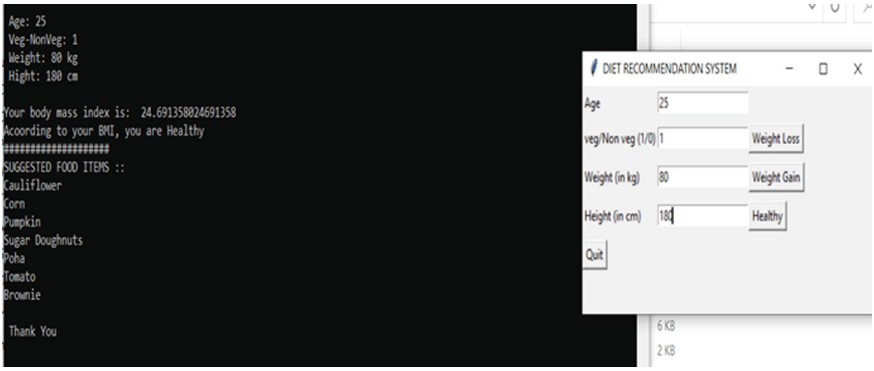


Fig. 8. Sample of the food items recommended by the system

5 Conclusion

The goal of this work is to develop an automatic diet recommender system to generate diet plan for user according to their BMI and food preferences. For this purpose, we have extended the idea of random forest classifier to recommend the final diet plan. Before this classification, the popular k-means clustering algorithm have been implemented for categories the food items based on its calories. It is worth to note that the proposed hybrid diet recommender system has performed well over the counterparts existing model. This system can be used as a tool for people to start toward a healthier lifestyle and improve

their nutritional necessities. This work can be extended by improvising the proposed system to make it available for cloud computing. It enables the users to share their diet plans to receive more varied and other recommendations. With the advent of cloud technology, it is also possible to recommend area-specific food items to the user so that it would be easier to acquire those items to follow the diet plan successfully. In addition, the recommendation fitness activities and the diet for the specific function that the user prefers.

Acknowledgments. The authors gratefully acknowledge the Science and Engineering Research Board (SERB), Department of Science & Technology, India, for the financial support through the Mathematical Research Impact Centric Support (MATRICS) scheme (MTR/2019/000542). The authors also acknowledge SASTRA Deemed University, Thanjavur, for extending infrastructural support to carry out this research.

References

1. World Health Organization (WHO): Fact Sheet, 312 (2011)
2. World Health Organization-Benefits of healthy diet https://www.who.int/initiatives/behealthy/healthy
3. Toledo, R.Y., Alzahrani, A.A., Martinez, L.: A food recommender system considers nutritional information and user preferences. IEEE Access **7**, 96695–96711 (2019)
4. Shrimal, M., Khavnekar, M., Thorat, S., Deone, J.: Nutriflow: a diet recommendation system (2021). SSRN 3866863
5. Iwendi, C., Khan, S., Anajemba, J.H., Bashir, A.K., Noor, F.: Realizing an efficient IoMT-assisted patient diet recommendation system through a machine learning model. IEEE Access **8**, 28462–28474 (2020)
6. Vasireddy, P.: An autonomous diet recommendation bot using intelligent automation. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 449–454. IEEE (May 2020)
7. Chavan, P., Thoms, B., Isaacs, J.: A recommender system for healthy food choices: building a hybrid model for recipe recommendations using big data sets. In: Proceedings of the 54th Hawaii International Conference on System Sciences, p. 3774 (January 2021)
8. Tabassum, N., Rehman, A., Hamid, M., Saleem, M., Malik, S., Alyas, T.: Intelligent nutrition diet recommender system for diabetic patients. Intell. Autom. Soft Comput. **30**(1), 319–335 (2021)
9. Manoharan, S.: Patient diet recommendation system using K clique and deep learning classifiers. J. Artif. Intell. **2**(02), 121–130 (2020)
10. Kim, J.-H., Lee, J.-H., Park, J.-S., Lee, Y.-H., Rim, K.-W.: Design of diet recommendation system for healthcare service based on user information. In: 2009Fourth International Conference on Computer Sciences and Convergence Information Technology, pp. 516–518 (2009). https://doi.org/10.1109/ICCIT.2009.293
11. Geetha, M., Saravanakumar, C., Ravikumar, K., &Muthulakshmi, V.: Human body analysis and diet recommendation system using machine learning techniques (2021)
12. Hsiao, J.H., Chang, H.: SmartDiet: a personal diet consultant for healthy meal planning. In: 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 421–425. IEEE (October 2010)
13. Princy, J., Senith, S., Kirubaraj, A.A., Vijaykumar, P.: A personalized food recommender system for women considering nutritional information. Int. J. Pharm. Res. **13**(2) (2021)

14. Agapito, G., Calabrese, B., Guzzi, P.H., Cannataro, M., Simeoni, M., Caré, I., Pujia, A., et al.: DIETOS: a recommender system for adaptive diet monitoring and personalized food suggestion. In 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–8. IEEE (October 2016)
15. Padmapritha, T., Subathra, B., Ozyetkin, M.M., Srinivasan, S., Bekirogulu, K., Kesavadev, J., Sanal, G., et al.: Smart artificial pancreas with diet recommender system for elderly diabetes. IFAC-PapersOnLine **53**(2), 16366–16371 (2020)
16. Ghosh, P., Bhattacharjee, D., Nasipuri, M.: Dynamic diet planner: a personal diet recommender system based on daily activity and physical condition. IRBM **42**(6), 442–456 (2021)
17. Chavan, S.V., Sambare, S.S., Joshi, A.: Diet recommendationsare based on prakriti and season using fuzzy ontology and type-2 fuzzy logic. In: 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6. IEEE (August 2016)
18. Pawar, R., Lardkhan, S., Jani, S., Lakhi, K.: NutriCure: a disease-based food recommender system. Int. J. Innov. Sci. Res. Technol. **6**, 2456–2165
19. Hernandez-Ocana, B., Chavez-Bosquez, O., Hernandez-Torruco, J., Canul-Reich, J., Pozos-Parra, P.: Bacterial foraging optimization algorithm for menu planning. IEEE Access **6**, 8619–8629 (2018)



QG-SKI: Question Classification and MCQ Question Generation Using Sequential Knowledge Induction

R. Dhanvardini[1], Gerard Deepak[2(✉)], and A. Santhanavijayan[3]

[1] Health Care Informatics Domain, Optum-United Health Groups, Hyderabad, India
[2] Department of Computer Science Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India
gerard.deepak.christuni@gmail.com
[3] Department of Computer Science Engineering, National Institute of Technology, Tiruchirappalli, India

Abstract. E-Learning has emerged as the most effective way of getting information in a range of sectors in the contemporary age. The utilisation of electronic content to provide education and development is referred to as e-learning. While the broadening internet has a plethora of e-learning tools, knowledge acquisition is not just an aspect that adds to an individual's enrichment. Assessment and evaluation are crucial parts of every learning system. Due to more complex assessments and quicker inspection, multiple choice questions are becoming extremely prevalent in current evaluations. However, establishing a diversified pool of MCQs relevant to a certain subject matter presents a hurdle. Manually creating high-quality MCQ exams is a time-consuming and arduous procedure that demands skill. As a result, research has concentrated on the automated construction of well-structured MCQ-based tests. This paper presents a paradigm using natural language processing based on semantic similarity and dynamic ontology. The proposed QG-SKI model uses LOD Cloud and Wiki Data to generate Ontologies dynamically and Knowledge reservoir is performed. The dataset is analysed using TF-IDF algorithm and the Semantic Similarity and Semantic Dissimilarity are computed using Shannon's entropy, Jaccard Similarity and Normalised Google Distance. These algorithms are executed for a multitude of degrees and levels to generate a similar of similar instances. The suggested model has a 98.15% accuracy and outperforms previous baseline models by dampening resilience.

Keywords: E-Assessment · E-learning · Dynamic Ontology · MCQ question generation

1 Introduction

Online learning resources uses computer or other digital devices to access the educational materials and to learn from it. The educational resources and the assessments of learners from those resources are both required for online learning. Learning tools are provided,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 116–126, 2023.
https://doi.org/10.1007/978-3-031-27409-1_11

and students may study from a variety of online sources. Automated questions and assessments from the learning materials, on the other hand, are necessary for the learner's evaluation. A succession of credible evaluations serves as indications of learners' depth of understanding and give a chance for friendly rivalry among peers, which helps the process escalate and become comprehensive.

Among the numerous forms of questions that are prevalent, multiple-choice questions are the most popular. These questions need vigilance, knowledge of the subject, and examination, as well as logic, which is frequently used during choice elimination. There is only a 20% chance of getting the correct answer out of the five possibilities offered. Consequently, the grading of these MCQs is extremely precise. In multiple choice questions, a "stem or question" is followed by a sequence of potential responses. Only one option is accurate, referred to as "key", while the others are referred to as "distractors". Rather than merely repeating lines from the corpus, the questions would have to be able to accurately detect the contextual.

Despite recent developments in NLP, creating high quality MCQ questions with complex attractors and distractors remains a time-consuming process. This work presents a unique technique based on dynamic ontologies for properly assessing and using it, depending on the semantic score. The produced distractors should have some distinguishing characteristics, such as meaning the same as the answer key, which gives the test participant a sense of uncertainty. This procedure must be followed precisely because it is the cornerstone question design phase.

Motivation: Due to a scholastic and cognitive transition in the administration of numerous online tests using Multiple Choice Questions, manually designing MCQ questions has become incredibly challenging. For subject compliant e-assessment, an efficient and automated method is required. In a world of ever-increasing knowledge and information, it's becoming progressively essential to adapt to the fast-paced virtual environment and have complex algorithms for producing appropriate questions for any given topic so that a student's actual inventiveness can be closely tracked. This serves as impetus for this research, as it highlights the necessity for a pertinent and systematic strategy to achieve at the development of Multiple-Choice Questions that will aid in a student's education.

Contribution: The framework proposed possess a novel approach for automatic MCQ question generation using QG-SKI model. The dataset is pre-processed with two phases. One, the dataset is inputted to LOD Cloud after obtaining keywords. Next, the dataset is computed using TF-IDF algorithm and categorical informative terms are obtained which is further passes through the Wiki Data API. The terminologies obtained from LOD Cloud and Wiki Data are combined to generate Ontologies. These dynamically generated ontologies help in establishing the Knowledge Reservoir. Then the keywords are computed for several degrees of Semantic Similarity and Semantic Dissimilarity using algorithms such as Shannon's entropy, Logistic regression and Decision trees, Jaccard Similarity and Normalised Google Distance (NGD). Experiments on the dataset resulted in a greater percentage of average precision, average recall, accuracy, F-Measure, as well as a very minimal False Discovery Rate by incorporating several techniques and methodologies into one. An overall F-measure of 98.141% and Accuracy of 98.15% is achieved.

Organisation: The remaining part of the paper is presented under the following sections. The second section describes Related Work. The Proposed System Architecture is detailed in Sect. 3. The Results and Performance Analysis are shown in Sect. 4. Finally, Sect. 5 brings the paper to a conclusion.

2 Related Work

Naresh Kumar et al. [1], develop OntoQuest, a system for generating multiple-choice questions depending on the user's preferred domain or topic. A summarization approach that relies on equivalence sets has been presented. WordNet combines dynamic information with static knowledge to improve overall accuracy. To produce the proper key, Jaccard similarity is considered. Rajesh Patra et al. [2], present a hybrid strategy for creating named entity distractors for MCQs. It illustrates how to automatically generate named entity distractors. The method employs a mix of statistical and semantic similarity. An approach based on predicate-argument extraction is used to calculate semantic similarity. Dhanya et al. [3], propose a Google T5 and Sense2Vec-based AI-assisted Online MCQ Generation Platform. They propose that all the NLP objectives be reconceptualized utilising T5 paradigm as a consistent text-to-text format with text strings as input and output. Sense2vec is a neural network model that includes extensive corpora to build vector space representations of words.

Rajat Agarwal et al. [4], present Automatic Multiple Choice Question Generation from Text leveraging Deep Learning and Linguistic Features. This paper describes an MCQ generation system that produces MCQs from a given text using linguistic characteristics and Deep Learning algorithms. The DL state-of-the-art model extracts significant data from a textual paragraph. Linguistic characteristics provide pairings of query (stem) and response (key). Using the key or the right answer, a distractor is developed. The MCQs dataset is supplemented with questions of same nature and level of difficulty using DL-based paraphrase models. I-Han Hsiao et al. [5], suggest a semantic PQG model to aid instructors in developing new programming problems and expanding evaluation items. The PQG model uses the Local Knowledge Graph (LKG) and Abstract Syntax Tree (AST) to transfer theoretical and technical programming skills from textbooks into a semantic network. For each query, the model searches the existing network for relevant code examples and uses the LKG/AST semantic structures to build a collection of questions. Neeti Vyas et al. [6], develops an Automated question and test-paper deployment tool that focuses on POS tagging, pronoun resolution, and summarisation. Questions are produced based on the text once it has been resolved and summarised. Kristiyan Vachev et al. [7], demonstrate Leaf, a method for building multiple-choice questions utilizing factual content.

Pranav et al. [8], proposes Automated Multiple-Choice Question Creation Using Synonymization and Factual Confirmation. This paper presents a technique for minimising the challenge's intensity by using abstractive LSTM series. Radovic et al. [9], presents an Ontology-Driven Learning Assessment Using the Script Concordance Test. The system is proposed using a unique automated SCT generating platform. The SCTonto ontology is used for knowledge representation in SCT question generation, with an emphasis on using electronic health records data for medical education. Pedro Lvarez

et al. [10], recommends using semantics and service technologies to create online MCQ tests automatically. The system comprises of a dynamic method for producing candidate distractors, a collection of heuristics for grading the adequacy of the distractors, and a distractors selection that considers the difficulty level of the tests.

Riken Shah et al. [11], introduces a technique for automatically generating MCQs from any given input text, as well as a collection of distractors. The algorithm is trained on a Wikipedia dataset that consists of Wikipedia article URLs. Keywords, which include both bigrams and unigrams, are retrieved and stored in a dictionary among many other knowledge base components. To produce distractors, we employed the Inverse Document Frequency (IDF) metric and the Context-Based Similarity method employing Paradigmatic Relation Discovery tools. In addition, to eliminate a question with inadequate information, the question creation process involves removing sentences that begin with Discourse Connectives. Baboucar Diatta et al. [12], discusses bilingual ontology to assist learners in question generation. Picking the most relevant linguistic assets and selecting the ontology label to be localised are two steps in the ontology localization process. Then Obtain and evaluate ontology label translation. To represent the two languages in their ontology, they use the paradigm that allows for the incorporation of multilingual information in the ontology using annotation features such as label and data property assertions. In [16–22] several models in support of the proposed literature have been depicted.

3 Proposed System Architecture

The principal objective of our proposed system is to use sequential knowledge induction to categorise and produce MCQ questions, key, and distractors for E-assessments. The entire architecture of the proposed framework is depicted in the Fig. 1. Text data is widely available and is utilised to assess and solve business and educational challenges. However, processing the data is necessary before using it for analysis or prediction. Tokenization, lemmatization, stop word removal, and named entity identification are all part of the pre-processing step. Tokenization is the method of dividing a text into manageable bits known as tokens. A large section of text is broken down into words or phrases. Then specified criteria is used to separate the input text into relevant tokens. It is a method of creating a huge dictionary in which each word is assigned a unique integer index. The sentences are subsequently converted from string sequences to integer sequences using this dictionary. The process of lemmatization is that the algorithm figures out the meaning of the word in the language it belongs to. It then determines how many letters must be removed to reduce it to its root word. The words are morphologically analysed during lemmatization. The aim of stop word removal is to eliminate terms that exist in all the articles in the corpora. Stop words encompass articles and pronouns in most cases. The purpose of named-entity recognition is to identify and categorise named items referenced in unstructured text into pre-defined categories.

On pre-processing the data, we obtain keywords which are used for the initial population by subjecting it to a Linked Open Data (LOD) cloud. LOD cloud is a Semantic Web of Linked Data that emerges as a Knowledge Graph. Later, SPARQL endpoints are

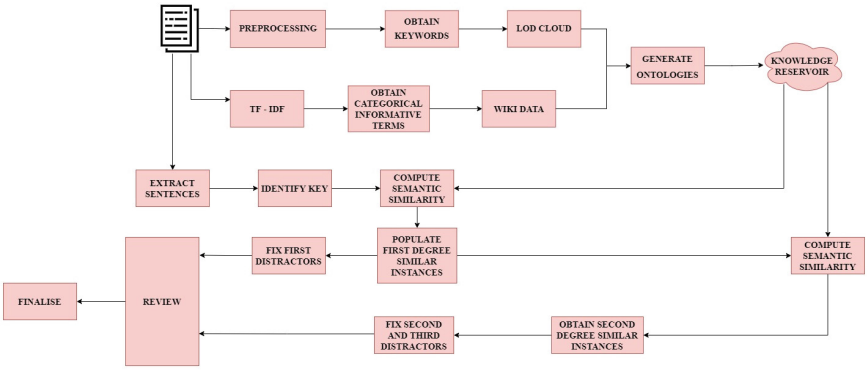


Fig. 1. Architecture of the proposed framework

used to query the LOD cloud by using the keywords obtained. The LOD Cloud Knowledge Graph and SPARQL Query Service Endpoints enable data access architecture in which hyperlinks function as data conductors across data sources.

In addition, the dataset is analysed in parallel using the TF-IDF model to obtain categorically informative terms, which has previously been pre-processed to produce keywords. Term Frequency-Inverse Document Frequency (TF-IDF) stop-words filtering is used in a variety of applications, including text summarization and categorisation. Because the TF-IDF weights words according to their significance, this approach may be used to discover which words are the most essential. This may be used to summarise articles more efficiently. It is composed of two terms: Term frequency (TF) and Inverse Document Frequency (IDF).

TF: Term Frequency (1) is a statistic that measures the number of times a terminology occurs in a text. Because the length of each document varies, it's possible that a term will appear more recurrently in longer documents than in shorter documents (2). As a result, it is expressed as follows as a means of normalisation:

$$\mathrm{TF}_t = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}} \tag{1}$$

$$\mathrm{tf}(\mathrm{t}, \mathrm{d}) = \frac{\mathrm{f}(\mathrm{t}, \mathrm{d})}{\sum t^{'} \in df\,(t^{'}, d)} \tag{2}$$

IDF: Inverse Document Frequency (3) is a statistic for measuring the importance of the frequency of a phrase. All elements are considered equal when calculating TF. However, it is common that terms such as "is," "of," and "that" appear frequently but have little significance (4). As a result, it is calculated as follows:

$$\mathrm{IDF}_t = \log_e\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right) \tag{3}$$

$$\mathrm{idf}(\mathrm{t}, \mathrm{D}) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \tag{4}$$

Term t is given a weight in document d via the TF-IDF$_{t;d}$ (5) weighting system:

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t \tag{5}$$

As a result, TF-IDF produces the most commonly recurring terms within a document corpus as well as the unusual term across all document corpora. This is subsequently sent to the Wikidata API, an open source linked database that serves as a central repository for structured data, which returns the appropriate terminology.Furthermore, the SPARQL queried data from the LOD cloud is combined with the terminologies from Wikidata to create Ontologies. Onto collab is a proposal for building knowledge bases using ontology modelling to improve the semantic properties of the World Wide Web. The keywords from the LOD cloud and Wikidata are fed into onto collab, which generates ontologies.The knowledge reservoir is built using ontologies, which contain some pre-existing domain information derived from web index terms retrieved directly from structural metadata. To formalise the knowledge reservoir, generated ontologies are pooled by generating at least one connection between each cluster of items. To acquire keys, the dataset is categorised by extracting features (sentences). The Shannon's entropy is utilised to compute semantic similarity between categorical terms on the dataset and created ontologies in this procedure.With the use of Shannon's entropy (6), semantic similarity assesses the distance between the semantic meanings of two words which is given as Eq. (6)

$$\text{Entropy} = -\sum_{i=1}^{k}[p(w_i) \times \log(p(w_i))] \tag{6}$$

To categorise the dataset using logistic regression and decision trees, ontologies are employed as special features. Logistic regression and decision tree are utilised as key classifiers to raise the heterogeneity of the relevant documents, as well as the significant subspace and the collection of documents in the classified set. The likelihood of a categorical dependent variable is predicted using logistic regression. It employs a sigmoid function to process the weighted combination of input information. A decision tree divides the input space into sections to classify inputs. It evaluates messages using a huge training dataset to learn a hierarchy of queries. Sentences are retrieved by comparing or recognising keywords in the document that are utilised in the key generation. To achieve the first degree of similar instances, the semantic similarity between key and knowledge reservoir is computed. The criterion for threshold of semantic similarity is estimated as 0.75.

Distractors are formed from first-degree comparable occurrences. Something that is significant to key is related to key but not identical to key. Rather of assessing dissimilarity and subsequently generating antonyms, the staged similar cases are computed. Minor dissimilarity can be produced based on transitivity or partial reliance by contrasting the similar of similar instances. By comparing the occurrences that are similar, the distinctly similar keyword is obtained called as distractor. The semantic similarity of the populated first-degree similar instances is computed once again to generate the second degree of similar instances. Then the second and third distractors are fixed. This entire computation of semantic dissimilarity to obtain the distractors is performed using Normalised Google Distance (NGD) and Jaccard Similarity.

Normalized Google Distance (7) is a semantic measure of similarity used by the search engine. By calculating the negative log of a term's probability, the Normalised Google Distance is utilised to generate a probability density function across search phrases that provides a Shannon-Fano code length. This method of calculating a code length from a search query is known as a Google compressor.

$$\mathrm{NGD(x, y)} = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}} \tag{7}$$

Jaccard similarity (8) is used to compute the correlation and heterogeneity of sample sets. It's calculated by dividing the size of the intersection by the size of the sample sets' union.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{8}$$

To acquire the first- and second-degree similar instances, semantic similarity is computed. The first, second, and third distractors are then computed using semantic dissimilarity. The threshold for semantic dissimilarity is determined as 0.45. The threshold is 0.45 instead of 0.25 since we are estimating dissimilarity among a subset of significantly similar keywords. Eventually, the question is presented with the key and three distractors and submitted for final evaluation. The complete system is finalised and formalised after the review output.

4 Implementation, Results and Performance Evaluation

The research is carried out using three distinct datasets that are combined into a single, huge integrated dataset namely Semantic Question Classification Datasets provided by FigShare [13], Kaggle's Questions vs Statements Classification based on SQuAD and SPAADIA dataset to distinguish between questions/statements [14] and Question Classification of CoQA – QcoC dataset by Kaggle [15]. The integration of the three distinct question categorization datasets into a single, large dataset is accomplished by manually annotating each dataset with a minimum of four to twelve annotations for each category of records. Latent Dirichlet Allocation and customised scrollers are used to dynamically annotate the data. Regardless, these three datasets were reordered using common category matching and similarity between these categories. Prioritization is created and placed at the end for all the unusual and mismatched category records. At the conclusion, all the matching records are combined. Therefore, by carefully merging each of these 3 datasets individually, a single huge dataset of proceedings is created.

The suggested QG-SKI is an automated question generating model based on question classification. The performance evaluation of the finalized MCQ generation with attractors and distractors will be assessed with the help of certain performance metrics. The percentage values of average precision, average recall, accuracy, F-measure, and False Discovery Rate metrics are used to evaluate the performance of the suggested knowledge centric question generation leveraging the framework QK-SKI. The significance of the findings is quantified by the evaluation metrics of average precision, average recall, accuracy, and F-measure. The number of false positives found in the provided

model is measured by the FDR. Table 1 and Fig. 2 shows the reliability of the proposed framework and the baseline models. Despite being an ontology-driven framework with certain unique features, the strength of the OntoQuest model can be increased by optimizing the density of auxiliary knowledge and including more distinctive and informative ontologies. The precision relevance computation method of OntoQuest, being a knowledge-centric semantically inclined framework, still has room for improvement. OMCQ employs a very basic matching technique and integrates a static domain ontology driven model that is part of the OWL framework. It employs wordnet linguistic resources and lexicons, resulting in knowledge that is based on linguistic structure rather than domain aggregation. Ontology has an extremely low density. The techniques for estimating relevancy are minimal and insignificant.

Table 1. Comparison of Performance of the proposed QG-SKI with other approaches

Search technique	Average precision %	Average recall %	Accuracy % (P + R)/2	F - Measure % (2*P*R)/(P + R)	FDR 1 - precision
OntoQuest [1]	95.82	97.32	96.57	96.564	0.05
OMCQ [2]	88.23	89.71	88.97	88.964	0.12
HyADisMCQ) [3]	89.71	91.82	90.76	90.753	0.11
Proposed QG-SKI	**97.23**	**99.07**	**98.15**	**98.141**	**0.03**

HyADisMCQ uses a hybrid technique for named entity distractors. The relevance computation algorithm is powerful since it employs both statistical and semantic similarity measurements. Existing real-world knowledge is included into the framework, and the knowledge density is kept to a minimum. The entity lacks richness, causing the distractors to deviate. Notwithstanding the shortcomings of the suggested baseline models outlined above, the proposed QG-SKI framework outperforms them with excellent precision. The LOD cloud and Wiki data add to the model's richness. The algorithm TF-IDF is used to obtain categorical informative terms. Ontologies are created dynamically. There are no static ontologies utilised. A knowledge reservoir is established. Semantic similarity is calculated using different threshold values at various levels and degrees. The Shannon' entropy is used to generate the attractor. For distractor synthesis, the Jaccard Similarity and Normalised Google Distance are utilised. Comparable of similar instances can be derived because of the continuous computation of semantic similarity. This improves accuracy and allows the architecture to exceed other models in performance. The precision percentage is shown on the line distribution curve Fig. 3.

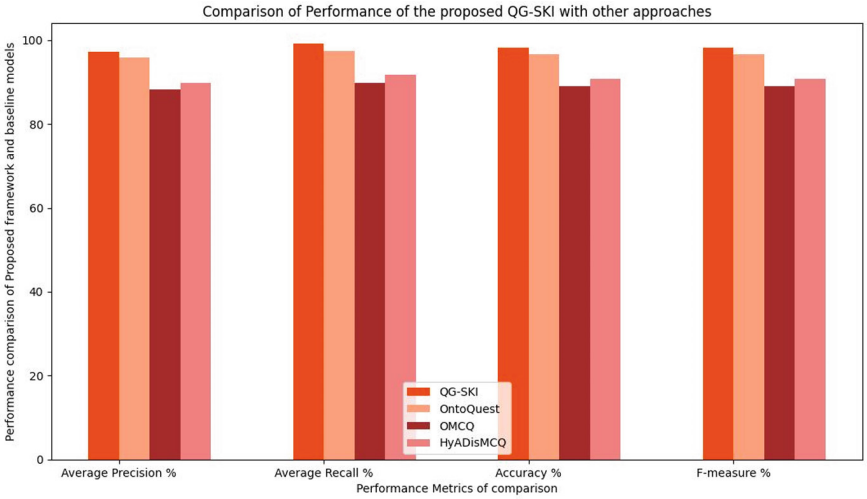


Fig. 2. Graph depicting Performance Comparison of the QG-SKI with other approaches

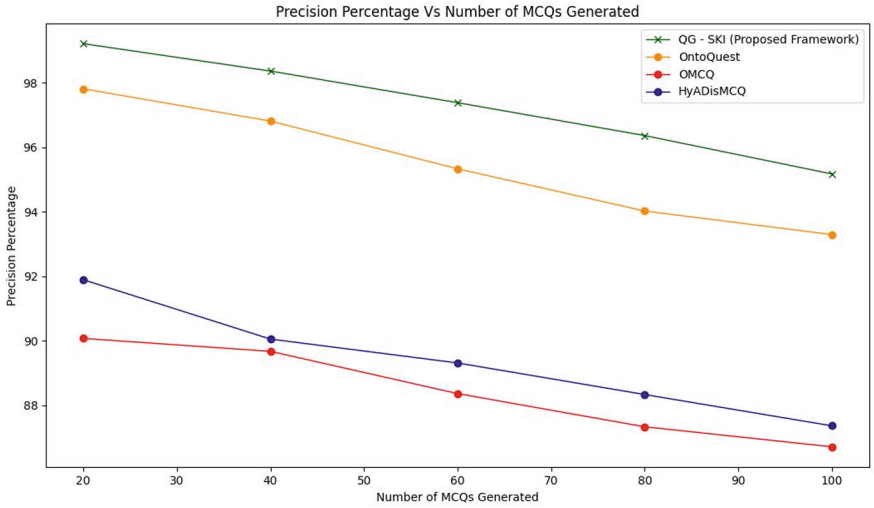


Fig. 3. Line distribution curve depicting Precision Percentage

5 Conclusion

A novel approach for automatically generating Multiple-Choice Questions for e-assessment from online corpora has been presented. In this research, a dynamic ontology for e-assessment systems is designed. The data is pre-processed with LOD Cloud and Wiki Data before being integrated to create ontologies. The Knowledge Repository has been built. The keywords are then evaluated using Shannon's entropy, Logistic regression and decision trees, Jaccard Similarity, and NGD for various degrees and levels of

semantic similarity and dissimilarity. The keywords that have been finely analysed are then examined and validated. The suggested algorithms' performance was compared to that of other existing algorithms. It may be determined from the experimental findings that the proposed algorithms enhanced the system's effectiveness. Average precision, average recall, accuracy, F-measure, and FDR are the performance measures used in the analysis, and the results are compared. When compared to previous research findings, QG-SKI is a highly robust approach with an overall accuracy of 98.15%, which is higher and more reliable. This work might be improved upon with yet more optimization improvements by including a more sophisticated semantic score that analyses the query sentence and assigns suitable weight to the relations encoded in the stem.

References

1. Deepak, G., Kumar, N., Bharadwaj, G.V.S.Y., Santhanavijayan, A.: OntoQuest: an ontological strategy for automatic question generation for e-assessment using static and dynamic knowledge. In: 2019 Fifteenth International Conference on Information Processing (ICINPRO), pp. 1–6. IEEE (December 2019)
2. Patra, R., Saha, S.K.: A hybrid approach for automatic generation of named entity distractors for multiple choice questions. Educ. Inf. Technol. **24**(2), 973–993 (2018). https://doi.org/10.1007/s10639-018-9814-3
3. Dhanya, N.M., Balaji, R.K., Akash, S.: AiXAM-AI assisted online MCQ generation platform using google T5 and Sense2Vec. In: 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 38–44. IEEE (February 2022)
4. Agarwal, R., Negi, V., Kalra, A., Mittal, A.: Deep learning and linguistic feature based automatic multiple choice question generation from text. In: International Conference on Distributed Computing and Internet Technology, pp. 260–264. Springer, Cham (January 2022)
5. Hsiao, I.H., Chung, C.Y.: AI-infused semantic model to enrich and expand programming question generation. J. Artif. Intell. Technol. **2**(2), 47–54 (2022)
6. Vyas, N., Kothari, H., Jain, A., Joshi, A.R.: Automated question and test-paper generation system. Int. J. Comput. Aided Eng. Technol. **16**(3), 362–378 (2022)
7. Vachev, K., Hardalov, M., Karadzhov, G., Georgiev, G., Koychev, I., Nakov, P.: Leaf: Multiple-Choice Question Generation (2022). arXiv:2201.09012
8. Pranav, M., Deepak, G., Santhanavijayan, A.: Automated multiple-choice question creation using synonymization and factual confirmation. In: Verma, P., Charan, C., Fernando, X., Ganesan, S. (eds.) Advances in Data Computing, Communication and Security. LNDECT, vol. 106, pp. 273–282. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-8403-6_24
9. Radovic, M., Petrovic, N., Tosic, M.: An ontology-driven learning assessment using the script concordance test. Appl. Sci. **12**(3), 1472 (2022)
10. Álvarez, P., Baldassarri, S.: Semantics and service technologies for the automatic generation of online MCQ tests. In: 2018 IEEE Global Engineering Education Conference (EDUCON), pp. 421–426. IEEE (April 2018)
11. Shah, R., Shah, D., Kurup, L.: Automatic question generation for intelligent tutoring systems. In: 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), pp. 127–132. IEEE (April 2017)
12. Diatta, B., Basse, A., Ouya, S.: Bilingual ontology-based automatic question generation. In: 2019 IEEE Global Engineering Education Conference (EDUCON), pp. 679–684. IEEE (April 2019)

13. Deepak, G., Pujari, R., Ekbal, A., Bhattacharyya, P.: Semantic Question Classification Datasets (2018). https://doi.org/10.6084/m9.figshare.6470726.v1
14. Khan, S.: Questions vs Statements Classification Based on SQuAD and SPAADIA dataset to distinguish between questions/statements (2021). https://www.kaggle.com/shahrukhkhan/questions-vs-statementsclassificationdataset
15. Question Classification of CoQA-QCoC. https://www.kaggle.com/saliimiabbas/question-classification-of-coqa-qcoc
16. Surya, D., Deepak, G., Santhanavijayan, A.: KSTAR: a knowledge-based approach for socially relevant term aggregation for web page recommendation. In: International Conference on Digital Technologies and Applications, pp. 555–564. Springer, Cham (January 2021)
17. Deepak, G., Priyadarshini, J.S., Babu, M.H.: A differential semantic algorithm for query relevant web page recommendation. In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 44–49. IEEE (October 2016)
18. Roopak, N., Deepak, G.: OntoKnowNHS: ontology driven knowledge centric novel hybridised semantic scheme for image recommendation using knowledge graph. In: Iberoamerican Knowledge Graphs and Semantic Web Conference, pp. 138–152. Springer, Cham (November 2021)
19. Ojha, R., Deepak, G.: Metadata driven semantically aware medical query expansion. In: Iberoamerican Knowledge Graphs and Semantic Web Conference, pp. 223–233. Springer, Cham (November 2021)
20. Rithish, H., Deepak, G., Santhanavijayan, A.: Automated assessment of question quality on online community forums. In: International Conference on Digital Technologies and Applications, pp. 791–800. Springer, Cham (January 2021)
21. Yethindra, D.N., Deepak, G.: A semantic approach for fashion recommendation using logistic regression and ontologies. In: 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1–6. IEEE (September 2021)
22. Deepak, G., Gulzar, Z., Leema, A.A.: An intelligent system for modeling and evaluation of domain ontologies for Crystallography as a prospective domain with a focus on their retrieval. Comput. Electr. Eng. **96**, 107604 (2021)

A Transfer Learning Approach to the Development of an Automation System for Recognizing Guava Disease Using CNN Models for Feasible Fruit Production

Rashiduzzaman Shakil[1(✉)], Bonna Akter[1], Aditya Rajbongshi[2], Umme Sara[2], Mala Rani Barman[3], and Aditi Dhali[4]

[1] Department of CSE, Daffodil International University, Dhaka, Bangladesh
{rashiduzzaman15-2655,bonna15-2585}@diu.edu.bd
[2] Department of CSE, National Institute of Textile Engineering and Research, Dhaka, Bangladesh
[3] Department of CSE, Sheikh Hasina University, Dhaka, Bangladesh
[4] Department of CSE, Jahangirnagar University, Dhaka, Bangladesh

Abstract. Guava (Psidium guava) is one of the most popular fruits which plays a vital role in the world economy. To increase guava production and sustain economic development, early detection and diagnosis of guava disease is important. As traditional recognition systems are time-consuming, expensive, and sometimes their predictions are also inaccurate, farmers are facing a lot of losses because of not getting the proper diagnosis and appropriate cure in time. In this study, an automatic system based on Convolution Neural Networks (CNN) models for recognizing guava disease has been proposed. To make the dataset more efficient, image processing techniques have been employed to boost the dataset which is collected from the local Guava Garden. For training and testing the applied models named InceptionResNetV2, ResNet50, and Xception with transfer learning technique, a total of 2,580 images in five categories such as Phytophthora, Red Rust, Scab, Stylar end rot, and Fresh leaf are utilized. To estimate the performance of each applied classifier, the six-performance evaluation metrics have been calculated where the Xception model conducted the highest accuracy of 98.88% which is good enough compared to other recent relevant works.

Keywords: Fruit's disease · Guava · InceptionResNetV2 · ResNet50 · Xception

1 Introduction

Humans can benefit greatly from the vitamins and minerals included in guava leaves and fruits. Due to Guavas' great nutritional and therapeutic properties, the fruit has gained widespread commercial success and is now grown in a variety of nations. However, various guava plant diseases are major issues that restrict output quantity and quality and dampen the economy.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 127–141, 2023.
https://doi.org/10.1007/978-3-031-27409-1_12

The guava sapling has been cultivated mainly by humans. The origin of guava seeds is sometimes obscured by the length of time they have been dispersed by birds and other four-legged creatures. It is still thought to be a region that reaches from southern Mexico into or through Central America. Since 1526, the West Indies, Bahamas, Bermuda, and southern Florida have grown guavas. It first appeared in 1847, and by 1886, it had spread over almost the whole state [1]. However, Guava also contributes significantly to the worldwide economy. According to a survey done in 2019, the annual global output of guavas was 55 million tons, with India accounting for approximately 45% of the total [2].

In the recent Era, the most critical fact limiting guava production is becoming a significant factor as well as it hampers the world economy. Cultivators face the most challenging difficulties in detecting and diagnosing guava fruit and leaf infections to overcome the barrier, which is quite impossible to do manually. The most common method for detecting and identifying diseases in plants and fruits is expert observation with the naked eye. Yet this requires constant expert monitoring, which may be cost-prohibitive for large farms. Not only are consultations with specialists costly, but in certain poor countries, farmers may need to travel long distances in order to get to them.

Cellphones and digital cameras make image-based automated systems more effective than traditional systems. The author collected the dataset used in the research from 2 hectares of field land.

This article addressed deep convolutional neural network (CNN) with transfer learning techniques to improve infinitesimal damage region learning and reduce computing complexity. CNN is one of the most convincing methods for pattern identification when working with a significant volume of data. CNN has promising results for detecting these diseases [3]. Plant disease detection and recognition based on Deep learning techniques can provide hints to identify the conditions and cure illnesses in the early stages. In addition, visual identification of plant diseases is costly, inefficient, and challenging and necessitates a trained botanist's assistance.

This research uses an image-based deep learning technique to develop an automation system by applying three CNN models named InceptionResNetV2, ResNet50, and Xception recognizing healthy leaves and four diseases, Phytophthora, Red Rust, Scab, and Stylar end rot, that affect guava fruit and leaf. The various image processing methods have been utilized to enhance the original dataset and make the system work well. The vital contribution is summarized as follows:

- An agro-based automation system to recognize guava diseases utilizing our original guava dataset that are available at Data in Brief [4].
- Proposed a fully connected and logistic layer-based architecture, Global Average Pooling2D, with a rectified linear unit function (ReLu).
- An authentic real-time dataset has been introduced in this paper, which the author collected.
- Discover the highest accuracy compared to existing relevant research.

2 Related Works

Currently, most machine learning and deep learning research focuses mainly on agriculture issues, as this sector contributes a lot to the world economy. But there is short research on fruits disease recognition such as guava, mango, jackfruits, etc.

Howlader et al. [5] created a D-CNN model to identify guava leaf disease. The model was created using 2705 images depicting four distinct illnesses. They achieved 98.74% and 99.43% accuracy during the training and testing phase, adopting 25 epochs.

Using a nine-layer convolutional neural network, Geetharamani and Pandia [6] developed a method to identify leaf fungus in plants. They worked on the Plant Village dataset, and the Kaggle dataset, which included 55448 images of 13 distinct plant leaves divided into 38 categories. SVM, logistic regression, decision tree, and KNN classifiers were also used to compare the proposed model, where the CNN model outperformed with remarkable prediction accuracy of 96.46%.

A multi-model pre-trained CNN model for identifying Apple and pest's disease was presented by Turkoglu et al. [7]. The AlexNet, GoogleNet, and DenseNet201 models utilizing 1192 images depicting four prevalent apple diseases. The DenseNet201 scored the highest accuracy among the applied models, with 96.10%.

Lakshmi [8] used an image classification system on an orange to test deep learning techniques' sweetness and quality detection. The goal of the study effort was applied to 5000 images, although the dataset's source was not revealed. SVM, AlexNet, SAE, and KSSAE were used to train the model, with KSSAE achieving the maximum accuracy of 92.1%. With a score of 96.1%, DenseNet201 seems to have the most excellent performance.

In order to diagnose mango disease, Trang et al. [9] suggested a deep residual network in combination with a contrast enhancement and transfer learning technique. The suggested algorithm correctly diagnosed three common illnesses based on 394 pictures, with an accuracy rate of 88.46%.

Nikhitha et al. [10] recommended employing the Inception V3 Model for fruit recognition and disease detection. They picked banana, apple, and cherry fruits as disease detection targets and solely used the Inception V3 model on them. This data was obtained from GitHub.

Ma et al. [11] proposed a deep convolutional neural network to identify four cucumber disorders diagnoses with a 93.4% recognition rate.

Prakash et al. [12] proposed an approach for diagnosing leaf diseases that relies on well-known image processing procedures such as preprocessing and classification. The provided technique is evaluated on a group of 60 photos, 35 of which are malignant and 25 of which are benign, with a 90% accuracy rate. K-means clustering is used to divide up the region impacted by the illness, and relevant features are extracted using GLCM. Subsequently, the SVM classifier is used to categorize the generated feature vector.

Buhaisi [13] used VGG16 model to detect the kind of pineapple utilizing 688 photos. The trained model got 100% accuracy and this dataset was most likely overfitting, or the accuracy would not have been possible.

Elleuch et al. [14] presented a deep learning diagnosis method. In this research, they used their newly created dataset containing five categories of plant data. They used transfer learning architecture with VGG-16 and Resnet to train their model. To compare

the validation of this model, they applied the proposed model to real and augmented data. VGG-16 with transfer learning gradually provided promising results in accuracy and reasonable accuracy of 99.02% and 98.35%.

Hafiz et al. [15] came up with a computer vision system that uses three convolutional neural network (CNN)-based models with different optimizers to find diseases in guavas. But they do not mention any reliable internet source for the collected data. The dropout value and third optimizer demonstrated promising accuracy when the dropout was 50%, which was 95.61%.

In order to detect guava disease, Meraj et al. [16] introduced a deep convolutional neural network-based technique using five different neural network structures. They used a locally collected dataset from Pakistan. The classification result proved that ResNet-101 was the best fit model for their work, achieving 97.74% accuracy.

Habib et al. [17] proposed a machine vision-based disease identification system for Guava, Papaya, and Jackfruit using nine important classifiers. Guava and jackfruit diseases were best identified by the Random Forest classifier obtaining 96.8% and 89.59% accuracy respectively.

3 Methodology

This section explains the step-by-step working procedure of guava disease recognition depicted in Fig. 1. Firstly, the guava image dataset was gathered at the field level. Then, the original images are augmented to boost the image dataset, a prerequisite for training and testing the CNN models. After completion of the augmentation, the new dataset is resized to the same size (224 * 224) and the same format (JPG). The dataset has also been separated into training and testing datasets for model generation. Finally, each classifier's performance is estimated to determine the best classifier to recognize the guava disease.

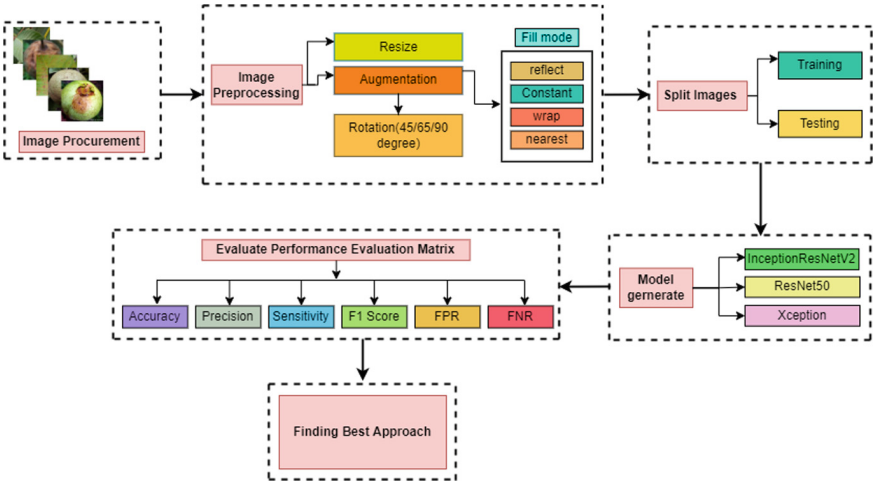


Fig. 1. Procedure of guava disease recognition

3.1 Image Acquisition

Evaluating the efficacy of deep learning-based models is greatly facilitated by the availability of a suitable dataset. More, image acquisition considers the crucial step in building a machine vision system. Diseased fruits are focused, and a particular distance is maintained while taking pictures. We collected the guava image dataset from the subtropical regions of Bangladesh, a guava garden with a camera to implement the model for guava disease recognition. The dataset consists of a total of 614 containing five classes where four classes have disease-affected image data and the rest one class with disease-free image data. The captured images were in RGB format. The detailed description of diseases is visualized in Table 1.

Table 1. Visualize disease with description

Disease Name	Description	Visualization
Fresh Leaf	• Guava's healthy leaves are green in color, and leaf veins are visible • The iron and vitamin C content of guava leaves is very high. That is an effective cure for a common cold and cough [18]	
Phytophthora	• Phytophthora are seen as brownish brown and grayish-black in guava fruits • Guava is seen soaked in water in the center of the affected area • The skin of affected fruits becomes soft • Infected guava stems become soft, and for this reason, the fruit falls off	
Red Rust	• The fungal infection Red rust deforms leaves and damages plants • Red Rust infected leaves turn brown, gradually dry out, and • spread to stems. Eventually, trees die	
Scab	• Scab is a fungal disease of guavas that is caused by the genus Pestalotiopsis • Infected surfaces become corky and ovoid • Scab affects the fruit's outer skin and lowers its quality and market value	
Stylar end rot	• Stylar end rot is believed to be caused by a fungal pathogen [19] • Circular to irregular discoloration of fruits starts from the stylar end side [20] • Infected fruits become soft, and this disease spreads until the whole fruit becomes brown and black	

3.2 Image Preprocessing

Image preprocessing is particularly first and most important for CNN models that employ images as input layers since it helps extract additional features and improves discrimination abilities. Building a CNN model is needed for a large data set. Data augmentation strategies improve performance since accuracy is increased due to the above process [21]. Due to the insufficiency of our data for CNN model construction, we have used augmentation methods to increase the size of the dataset. Besides, various preprocessing techniques have been adopted to make the image dataset the same size (224 * 224 pixels) and format before we train our model. The dataset distribution is presented in Table 2.

Table 2. Overall distribution of Guava Dataset

Disease Name	Original Data	Augmented Data	Utilized Data	Train
Fresh Leaf	140	407	547	428
Phytophthora	112	384	496	405
Red Rust	135	450	585	440
Scab	117	362	479	451
Stylar end rot	110	363	473	426
Total	614	1966	2580	2150

3.3 Model Description

Artificial neural networks known as convolutional neural networks (CNNs) are utilized in deep learning [22]. Its primary purpose is to assess visual data through the application of deep learning techniques [23]. A CNN model was constructed with the following layers: an input layer, a convolutional layer, a pooling layer, a fully connected layer, a hidden layer, and an activation function which is presented in Fig. 2.

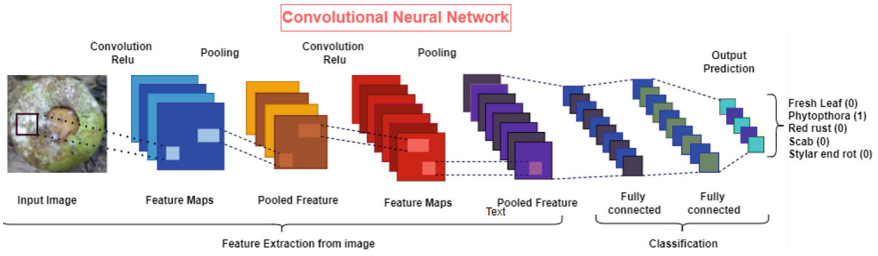


Fig. 2. The fundamental structure of CNN for guava disease recognition

When these layers are stacked, a CNN architecture will be formed. The most important aspects of the CNN architecture are the feature extraction and classification processes. We have employed three CNN models for recognizing the guava disease recognition.

3.3.1 InceptionResNetV2

Inception and ResNet are widely used deep convolutional neural networks, were combined to create InceptionResNetV2, which uses batch-normalization in place of summation for the conventional layers [24]. InceptionResNetV2 trained more than millions of images. Above a thousand filters, residual variances become too large, making it nearly impossible to train the model. So, the residuals are normalized to help stabilize the training of the network. InceptionResNetV2 was utilized in this research, and Fig. 3a provides a visual representation of its structured form.

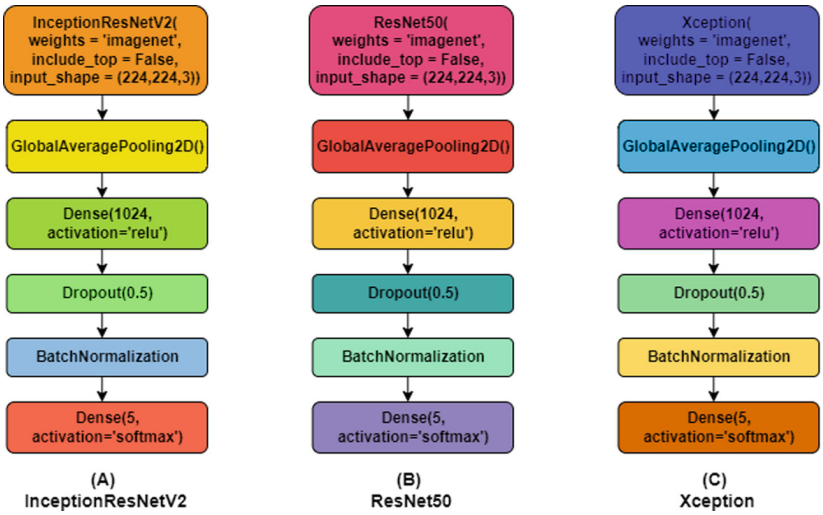


Fig. 3. Functional parameters of the applied models

3.3.2 ResNet50

Figure 3b visualizes the compressed form of ResNet50 as a convolutional neural network. It's also a deep residual network. It has around 50 layers for preprocessing [25]. After collecting data, it must be separated into two sets: training and testing. Each data instance in the training set has numerous characteristics, including a single target value.

3.3.3 Xception

The deep convolutional neural network architecture Xception uses depth-separable convolutions [26]. The Xception architecture-based feature extraction technique consists of

36 convolution layers. The 36 convolution layers were split into 14 modules for the first and last modules, each with its linear residual around them. The compressed format of Xeption employed in this work is shown in Fig. 3c.

4 Result and Discussion

4.1 Detailed of Environmental Setup

The proposed technique has been applied to Intel® Core™ i5-9600K processor, 480 GB SSD (Solid Disk Drive), 16 GB RAM (Random Access Memory), and GeForce GTX 1050 Ti D5 with 768 CUDA cores for both the training and validation phases. Poco X3 pro is used to capture field-level images, and it has a 48-megapixel camera and 8 GB ram. We completed the work in the Jupyter notebook with python version:3.8.5. For the recognition experiment, we had chosen CNN (Convolutional Neural Network), and inside CNN, three distinct models, InceptionResNetV2, ResNet50, and Xception, are utilized.

We had chosen 2150 images for training and 430 images for testing out of 2580 images. The training and testing percentage ratios were 80% and 20%, respectively. To determine the efficiency of the implemented models, we estimate six performance metrics: accuracy, sensitivity (TPR), precision, F1-Score, false positive rate (FPR), and false negative rate (FNR).

4.2 Experiment Result of Distinct Models

The models are trained and tested with 25 epochs for guava disease recognition. As three different models are applied, the epoch is affected differently to achieve accuracy. When the epoch was increased, the accuracy was raised, and the validation proportionally decreased. Figure 4, shows the visualization of Epoch vs. Accuracy after the completion of the adopted epoch, where the Xception model is well-performing label as Fig. 4(c). InceptionResNetV2 achieved the second-highest accuracy label Fig. 4(a) which is now cleared, and the ResNet50 model label Fig. 4(b) denotes the lowest accuracy among models.

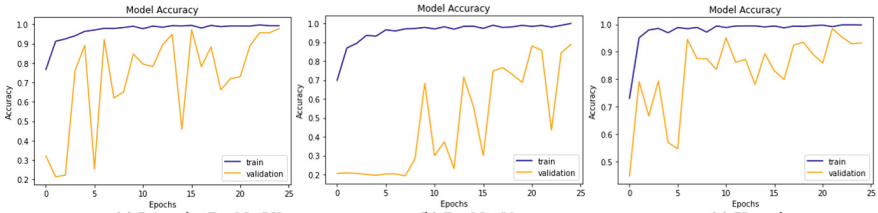


Fig. 4. Plotting of epoch vs. accuracy

The performance of a model depends on model loss that exists less. At the epoch's beginning, each model's accuracy is not up to mark where the loss is more. The epoch vs.

loss curve of the InceptionResNetV2, ResNet50, and Xception model is demonstrated in Fig. 5. When comparing the epoch vs. loss curves, the Xception model has minimum loss among all other model's label Fig. 5(c).

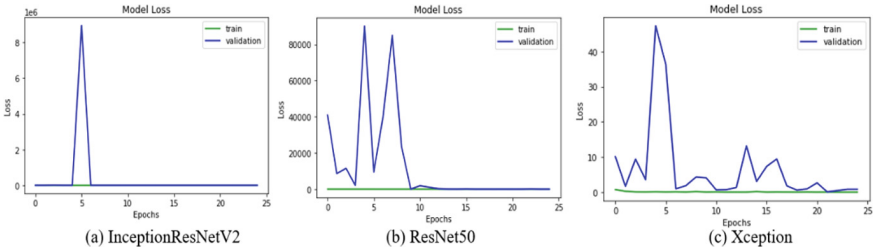


Fig. 5. Plotting of epochs vs. loss

ROC curves demonstrate the relationship between the true positive rate and the false positive rate for different threshold values. The needed individual values for distinct classes are added to compute the average in a micro-average ROC scheme. In contrast, a macro-average ROC curve estimates each class's required values individually and then take the average. The AUC (area under the curve) measures how effectively the model distinguishes between distinct classes; when evaluating test cases, an area of 1 is regarded as the best [27]. The micro-average and macro-average were graphically shown in Fig. 6, where InceptionResNetV2 and Xception both achieved the same micro-average and macro-average are 99% and 99%, respectively, and ResNet50 gained 97% micro-average and 98% micro-average.

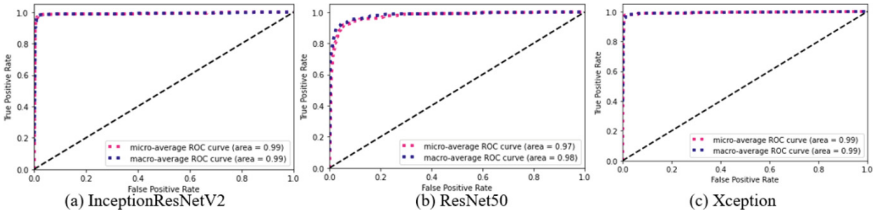
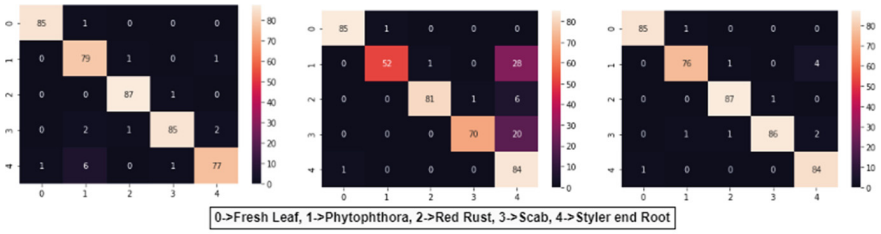


Fig. 6. Graphical representation of micro-average and macro-average

A confusion matrix is a visual representation of counts based on predicted and actual values [28]. Dimensionally, the confusion matrix for a multiclass scenario will be [n × n] [29, 30], where n > 2, and all matrices have the same number of rows and columns. As in our working procedure, we used five diseases for our proposed model so that a 5*5-dimensional confusion matrix will be generated for 430 testing images. Besides, for effectiveness and more realistic, we have constructed the plotting of those matrix which displays in Fig. 7. Then the 5*5 confusion matrix is converted to binary format.

Tables 3, 4 and 5 present the confusion matrix of InceptionResNetV2, ResNet50, and Xception model as binary format, respectively.



(a) InceptionResNetV2 (b) ResNet50 (c) Xception

Fig. 7. Graphical representation of confusion matrix

Table 3. Generated confusion matrix of InceptionResNetV2 model

Input	Five classes of guava disease					
	Fresh Leaf	Phytophthora	Red Rust	Scab	Stylar end rot	Total images
Fresh Leaf	85	1	0	0	0	86
Phytophthora	0	79	1	0	1	81
Red Rust	0	0	87	1	0	88
Scab	0	2	1	85	2	90
Stylar end rot	1	6	0	1	77	85
Total	86	88	89	87	80	430

Table 4. Generated Confusion Matrix for ResNet50 model

Input	Five classes of guava disease					
	Fresh Leaf	Phytophthora	Red Rust	Scab	Stylar end rot	Total images
Fresh Leaf	85	1	0	0	0	86
Phytophthora	0	52	1	0	28	81
Red Rust	0	0	81	1	6	88
Scab	0	0	0	70	20	90
Stylar end rot	1	0	0	0	84	85
Total	86	53	82	71	138	430

There are six distinct performance assessment metrics used to compare the quality of the various models, and their respective formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{2}$$

Table 5. Generated confusion matrix of Xception model

Input	Five classes of guava disease					
	Fresh Leaf	Phytophthora	Red Rust	Scab	Stylar end rot	Total images
Fresh Leaf	85	1	0	0	0	86
Phytophthora	0	76	1	0	4	81
Red Rust	0	0	87	1	0	88
Scab	0	1	1	86	2	90
Stylar end rot	1	0	0	0	84	85
Total	86	78	89	87	90	430

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{3}$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100\% \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \times 100\% \tag{5}$$

$$FNR = \frac{FN}{FN + TP} \times 100\% \tag{6}$$

Table 6 showed the result of InceptionResNetV2 models, constructed based on the following diseases: Fresh Leaf, Phytophthora, Red Rust, Scab, Stylar end rot. The highest precision, 98.84%, acquired Fresh Leaf, while the lowest precision, 89.77%, was gained by Phytophthora. Class wise accuracy are gradually 99.53%, 97.44%, 99.30%, 98.37%, 97.44% for selected diseases. Fresh Leaf gained the best result out of all of them.

Table 6. Class based performance evaluation metrics for InceptionResNetV2

Classifier	Disease Name	Accuracy (%)	Precision (%)	TPR (%)	F1-Score (%)	FPR (%)	FNR (%)
InceptionResNetV2	Fresh Leaf	99.53	98.84	98.83	98.84	0.29	1.16
	Phytophthora	97.44	89.77	97.53	93.49	2.58	2.47
	Red Rust	99.30	97.75	98.86	98.31	0.58	1.13
	Scab	98.37	97.70	94.44	96.05	0.59	5.56
	Stylar end rot	97.44	96.25	90.59	93.33	0.87	9.41

The accuracy obtained for Fresh Leaf, Phytophthora, Red Rust, Scab, Stylar end root is 99.53%, 93.02%, 98.14%, 95.12%, and 87.21%, respectively, in the ResNet50 model, as shown in Table 7. Class wise precision of ResNet50 is consistently 98.84%, 98.11%,

98.78%, 98.59%, and 60.78% for selected diseases. Fresh Leaf had the maximum sensitivity of 98.84%, while Phytophthora had the lowest sensitivity of 64.19%. The average accuracy of ResNet50 is 94.60%.

Table 7. Class based performance evaluation metrics for ResNet50

Classifier	Disease Name	Accuracy	Precision	TPR	F1-Score	FPR	FNR	Model Accuracy
ResNet50	Fresh Leaf	99.53%	98.84%	98.84%	98.84%	0.29%	1.16%	94.60%
	Phytophthora	93.02%	98.11%	64.19%	77.61%	0.29%	35.80%	
	Red Rust	98.14%	98.78%	92.05%	95.29%	0.29%	7.95%	
	Scab	95.12%	98.59%	77.78%	86.96%	0.29%	22.22%	
	Stylar end rot	87.21%	60.87%	98.82%	75.34%	15.65%	1.17%	

The result of the Xception model is shown in Table 8, where the maximum precision of 98.85% is attained by Scab disease. Among five classes, 99.53% was the highest accuracy, and 98.37% was the lowest accuracy achieved by Phytophthora and Stylar end rot. Sensitivity results are 98.84%, 93.83%, 98.86%, 95.56%, 98.82% corresponding of Fresh Leaf, phytophthora, Red Rust, Scab, Stylar end rot. Fresh leaf and Phytophthora seemed to have the greatest 98.84% and lowest 95.59% F1 scores, respectively.

Table 8. Class based performance evaluation metrics for Xception

Classifier		Accuracy	Precision	TPR	F1-Score	FPR	FNR	Model Accuracy
Xception	Fresh Leaf	99.53%	98.84%	98.84%	98.84%	0.29%	1.16%	98.88%
	Phytophthora	98.37%	97.44%	93.83%	95.59%	0.57%	6.17%	
	Red Rust	99.30%	97.75%	98.86%	98.31%	0.58%	1.13%	
	Scab	98.84%	98.85%	95.56%	97.18%	0.29%	4.44%	
	Stylar end rot	98.37%	93.33%	98.82%	96.00%	1.74%	1.18%	

4.3 Comparative Analysis with Other Existing Works

The consequence of research much depends on a comparison with the existing corresponding works. As we have worked with guava disease recognition by applying CNN models, comparing it with other research on guava recognition is required. M. R. Howlader et al. [5] worked with guava disease, where the highest accuracy was 98.74%. Another work was performed by Hafiz et al. [15] applying CNN models, and the accuracy was 95.61%. Some of the research on other fruits' disease recognition is also included. The comparative analysis of other existing work is presented in Table 9.

Table 9. Comparative study with other existing work

Completed work	Adopted Object	Repository	Measurement of Dataset	Applied Classifier/Model	Best Model	High and Low Accuracy
This work	Guava	Data in brief	2580	InceptionResNetV2, ResNet50, Xception	Xception	Xception: 98.88% ResNet50: 94.60%
Howlader et al. [5]	Guava	Publicly available known as BUGL2018	2705	SVM, LeNet-5, AlexNet, D-CNN	D-CNN	D-CNN: 98.74% SVM: 89.71%
Turkoglu et al. [7]	Apple	Malatya and Bingolcities of Turkey-[Field Level]	1192	AlexNet, GoogleNet, DenseNet201	DenseNet201	DenseNet201: 96.10% AlexNet: 94.7%
Lakshmi [8]	Orange	N/A	5000	SVM, AlexNet, SAE, KSSAE	KSSAE	KSSAE: 90.4% SVM: 75.10%
Trang et al. [9]	Mango	Plant Village dataset	394	Proposed Model, IncpetionV3, AlexNetV2, MobileNetV2	Proposed Method	Proposed Method: 88.46%
Nikhitha et al. [10]	Multiple Fruit	GitHub	539802	InceptionV3	InceptionV3	InceptionV3: 100%
Prakash et al. [12]	Citrus	Field Label	60	SVM	SVM	SVM: 90%
Hafiz et al. [15]	Guava	N/A	10000	CNN	CNN	CNN: 95.61%

5 Conclusion and Future Works

Food plant diseases cause a reduction in agricultural productivity in underdeveloped countries, which has repercussions for smallholder farmers. Consequently, it is critical to recognize ailments as soon as possible. Identification accuracy demonstrates the suggested CNN architecture with InceptionResNetV2, ResNet50, and Xception models is more effective and provides a superior solution for identifying guava disease. Inspired by our findings, we want to expand our dataset to include further classifications in the near future, and more leave diseases are planned to be added to make this model more accessible to users. We will also design to couple our model with a smartphone for a quick response, which could help farmers detect and prevent disease early on the spot.

References

1. Guava Details. https://hort.purdue.edu/newcrop/morton/guava.html. Accessed 22 June 2022
2. Guava. https://en.wikipedia.org/wiki/Guava. Accessed 25 June 2022
3. Mukti, I.Z., Biswas, D.: Transfer learning-based plant diseases detection using ResNet50. In: 2019 4th International Conference on Electrical Information and Communication Technology (EICT), pp. 1–6. IEEE (2019)
4. Rajbongshi, A., Sazzad, S., Shakil, R., Akter, B., Sara, U.: A comprehensive guava leaves and fruits dataset for guava disease recognition. Data Brief **42**, 108174 (2022)

5. Howlader, M.R., Habiba, U., Faisal, R.H., Rahman, M.M.: Automatic recognition of guava leaf diseases using deep convolution neural network. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–5. IEEE (2019)
6. Geetharamani, G., Pandian, A.: Identification of plant leaf diseases using a nine-layer deep convolutional neural network. Comput. Electr. Eng. **76**, 323–338 (2019)
7. Turkoglu, M., Hanbay, D., Sengur, A.: Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests. J. Ambient. Intell. Humaniz. Comput. , 1–11 (2019). https://doi.org/10.1007/s12652-019-01591-w
8. Lakshmi, J.V.N.: Image classification algorithm on oranges to perceive sweetness using deep learning techniques. In: AICTE Sponsored National Level E-Conference on Machine Learning as a Service for Industries MLSI (2020)
9. Trang, K., TonThat, L., Thao, N.G.M., Thi, N.T.T.: Mango diseases identification by a deep residual network with contrast enhancement and transfer learning. In: 2019 IEEE Conference on Sustainable Utilization and Development in Engineering and Technologies (CSUDET), pp. 138–142. IEEE (2019)
10. Nikhitha, M., Sri, S.R., Maheswari, B.U.: Fruit recognition and grade of disease detection using inception v3 model. In: 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1040–1043. IEEE (2019)
11. Ma, J., Du, K., Zheng, F., Zhang, L., Gong, Z., Sun, Z.: A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. Comput. Electron. Agric. **154**, 18–24 (2018)
12. Prakash, R.M., Saraswathy, G.P., Ramalakshmi, G., Mangaleswari, K.H., Kaviya, T.: Detection of leaf diseases and classification using digital image processing. In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–4. IEEE (2017)
13. Al Buhaisi, H.N.: Image-based pineapple type detection using deep learning. Int. J. Acad. Inf. Res. (IJAISR) **5**, 94–99 (2021)
14. Elleuch, M., Marzougui, F., Kherallah, M.: Diagnostic method based DL approach to detect the lack of elements from the leaves of diseased plants. Int. J. Hybr. Intell. Syst. 1–10 (2021)
15. Al Haque, A.F., Hafiz, R., Hakim, M.A. and Islam, G.R.: A computer vision system for guava disease detection and recommend curative solution using deep learning approach. In: 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1–6. IEEE (2019)
16. Mostafa, A.M., Kumar, S.A., Meraj, T., Rauf, H.T., Alnuaim, A.A., Alkhayyal, M.A.: Guava disease detection using deep convolutional neural networks. A case study of guava plants. Appl. Sci. **12**(1), 239 (2021)
17. Habib, M.T., Mia, M.J., Uddin, M.S., Ahmed, F.: An explorative analysis on the machine-vision-based disease recognition of three available fruits of Bangladesh. Viet. J. Comput. Sci. **9**(02), 115–134 (2022)
18. Benefit of Guava Leaf. https://food.ndtv.com/food-drinks/15-incredible-benefits-of-guava-leaf-tea-1445183/amp/1. Accessed 6 July 2022
19. Guava Disease Information. https://www.gardeningknowhow.com/edible/fruits/guava. Accessed 6 July 2022
20. Guava Crop Management. http://webapps.iihr.res.in:8086/cp-soilclimate1.html. Accessed 8 July 2022
21. Abbas, A., Jain, S., Gour, M., Vankudothu, S.: Tomato plant disease detection using transfer learning with C-GAN synthetic images. Comput. Electron. Agric. **187**, 106279 (2021)
22. Introduction of Convolutional neural network. https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/. Accessed 9 July 2022

23. Majumder, A., Rajbongshi, A., Rahman, M.M., Biswas, A.A.: Local freshwater fish recognition using different cnn architectures with transfer learning. Int. J. Adv. Sci. Eng. Inf. Technol. **11**(3), 1078–1083 (2021)
24. Hasan, M.K., Tanha, T., Amin, M.R., Faruk, O., Khan, M.M., Aljahdali, S., Masud, M.: Cataract disease detection by using transfer learning-based intelligent methods. Comput. Math. Meth. Med. (2021)
25. Ramkumar, M.O., Catharin, S.S., Ramachandran, V., Sakthikumar, A.: Cercospora identification in spinach leaves through resnet-50 based image processing. J. Phys. Conf. Ser. **1717**(1), 012046. IOP Publishing (2021)
26. Xception Model. https://maelfabien.github.io/deeplearning/xception/. Accessed 9 Aug 2022
27. Das, S., Aranya, O.R.R., Labiba, N.N.: Brain tumor classification using convolutional neural network. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1–5. IEEE (2019)
28. Jahan, S., et al.: Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. SN Appl. Sci. **3**(10), 1–17 (2021). https://doi.org/10.1007/s42452-021-04786-z
29. Rajbongshi, A., Biswas, A.A., Biswas, J., Shakil, R., Akter, B., Barman, M.R.: Sunflower diseases recognition using computer vision-based approach. In: 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), pp. 1–5. IEEE (2021)
30. Nawar, A., Sabuz, N.K., Siddiquee, S.M.T., Rabbani, M., Biswas, A.A., Majumder, A.: Skin disease recognition: a machine vision-based approach. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 1029–1034. IEEE (2021)



Using Intention of Online Food Delivery Services in Industry 4.0: Evidence from Vietnam

Nguyen Thi Ngan and Bui Huy Khoi(✉)

Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
buihuykhoi@iuh.edu.vn

Abstract. The online food ordering market in Vietnam is a potential and strongly developing market, which will be the industry attracting many domestic and foreign investors. The growing society, and increasing human needs, especially under the strong development of Industry 4.0, have made the online food delivery market in Vietnam hotter. With rapid growth, online food delivery service providers are also increasingly perfecting their services to attract customers and keep up with social trends. Online food delivery service in Vietnam in recent years has had significant development and is gradually replacing other traditional food delivery services. The outcomes of the AIC Algorithm for Using the intention of Online food delivery service (OFD) showed that 2 independent variables attitude (ATT) and social influence (SI) all have a favorable impact on the intention to use an online food delivery service (OFD). Previous research has shown that linear regression is effective. The AIC method is used in this study to make the best decision.

Keywords: Online Food Ordering Services · Perceived ease of use · Attitude · Time saving · And Social influence

1 Introduction

In new years, there have been quite a few studies in the world about food delivery services via the internet. Typically, following the COVID-19 outbreak in the Jabodetabek Area, the investigation examined factors influencing customers' intentions to use online food delivery services by Kartono and Tjahjadi [1], Prabowo and Nugroho [2] also published a study on Factors affecting Indonesian users' attitudes and behaviors Intent of Indonesian consumers towards OFD service using the Go-Food application and according to the research. Research by Ren et al. [3] on OFD in Cambodia: Research on elements affecting consumer using behavior intention. The outcomes of these studies display OFD has the following factors affecting intention to use e-mail services: perceived reliability affects Attitude, perceived relative advantage affects Attitude Perceived risk affects Attitudes, perceived reliability affects intention to use, and Perceived relative advantage affects intention to use and Attitude affects intention to use. Use, Hedonic motivation, online shopping experience first, save price, save time and ultimately convenience motivation, usefulness after use applies information technology innovation, perceive ease of use, performance expectations and the value of the price. The purpose of the chapter explores the AIC Algorithm for using the intention of online food delivery service (OFD).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 142–151, 2023.
https://doi.org/10.1007/978-3-031-27409-1_13

2 Literature Review

2.1 Using Intention of Online Food Delivery Service (OFD)

Using the intention of OFD service is a future-ready behavior of consumers [4], using the intention of OFD service will be affected by reasons of attitude, subjective norm, and perceived behavioral control [4]. In the research related to the application of information technology, according to Bauer [5], the use intention is also affected by the perceived risk factor.

In their study, Kartono and Tjahjadi [1], the using intention is expressed through the frequency of use, loyalty to the service will recommend and the intention to use this service will become a habit/lifestyle of consumers. Similar to the study of Prasetyo et al. [6] related to the use of e-mail services during the time of Covid-19 mentioned the intention to use is expressed by agreeing to use the service next time., plan to use and will try to use this service every day. And in the study of the elements influencing the using intention to the behavior of consumers in Cambodia, Ren et al. [3] suggested that the user intention is the use of mobile phone services instead of food ordering services, usually over the phone, is continued use and will recommend to others, this service will become my favorite service.

In summary, the using intention of the OFD service is the use of the OFD service instead of the usual ordering of food [3], the next use of the food ordering service [6] and the consumer will often, recommend to my friends this service [1].

2.2 Perceived Ease of Use (PEU)

People's predisposition to utilize or not use an application based on whether or not they believe it will help them perform their tasks better is known as perceived ease of use [7]. A study on the topic OFD research model in Cambodia: A research on elements influencing consumer using intention by Ren et al. [3] also mentioned factor PEU is one of the influences that directly impact the user intention. Specifically, Ren et al. [3] said that using the Internet phone service does not require mental effort, and ordering food from the Internet phone service is easy and understandable. PEU according to Prasetyo et al. [6] is that consumers can easily find what they need, the online food ordering application has a button that provides them with complete information and can complete the transaction easily, and the application has a good interface design. And according to the research results on topics related to OFD services, the relationship between PEU and using intention is positive [8, 9].

In summary, the perception of ease of use is that consumers find it easy to use the food ordering service, easy to understand, using this service does not require much brain [3], and the interface of a well-designed service [6].

2.3 Attitude (ATT)

Attitude to use first appeared in Ajzen's Theory of TRA [10]. According to Ajzen, the attitude variable has a direct impact on buyers' intentions. According to Ajzen and Fishbein [10], attitude is the belief in the attributes of a product and is a measure of trust

in the attributes of that product. Then, inheriting this theory of Ajzen and Fishbein [10], researchers have developed their theories and still agree with Ajzen and Fishbein [10], that attitude has a direct influence on behavioral intention. According to Kartono et al. [1], the features impacting the using intention of electronic communication services of people in the Jabodetabek zone include perceived risk, attitude and perception of relative advantage, and perception of reliability. Which, the author believes that the consumer's attitude influences the intention to use the positive feeling when using the service, they feel the online service is attractive to them when using the service, and they feel happy and satisfied.

In summary, the Attitudes toward online delivery services are satisfaction after using, having a pleasant experience when using them, and consumers feel that e-mail services are attractive to them [1].

From the above research results, it is shown that the connection between attitude and using intention for online food delivery services is positive [11, 12], in the same direction [13, 14].

2.4 Perceived Risk (PR)

The concept of perceived risk was first published in Bauer's Theory of Risk Perception (1960). According to Bauer [5], risk perception directly affects consumers' intention to use, including product-related perceived risks and online transaction-related perceived risks. Perceived risk in the shopping process is seen as an uncertain decision of the consumer when purchasing and must receive consequences from this decision.

In consumers, perceived risk is defined in different ways. Perceived risk from poor performance results, risk hazards and health risks, and costs. Perceived risks are divided into unsafe transactions, leakage of personal data, mishandling of orders, transportation risks, and other risks [1].

In summary, perceived risk is the risk perception of consumers about the product [5], perceived risk of unsafe transactions, leakage of personal data, mishandling of orders, and risks. Transportation and other risks [1].

In his study, Kartono et al. [1] also showed that the association between perceived risk and using intention is a negative relationship [15, 16], with different directions.

2.5 Time Saving (TS)

Saving time is using your time for things that give meaning to your life and work. Saving time is not wasting time on meaningless, unproductive tasks. And in today's faster and busier modern life so that work is not affected too much, many people have used services to save time and effort.

Timesaving orientation is the most important factor to influence customer motivation to use technology-based self-service. When a person is short on time owing to daily activities such as work and leisure activities, he or she looks for ways to save time. And in recent years, many people have had busy lifestyles they don't like the effort of foraging and waiting for food at restaurants. They want the food to come to them with little effort and be delivered as quickly as possible [2, 17]. According to Chai and Yat [17], a person who wants to save time will not choose to order food directly at a restaurant.

In summary, Time saving in using online services includes saving time in ordering, waiting, and transaction and payment [2, 17].

And in these studies, it has also been shown that the time saving factor has a positive link with the user's intention to online delivery services [18]. When consumers realize they save more time when using the service, they have a higher intention to use the food ordering service.

2.6 Social Influence (SI)

Cultural, social, personal, and psychological aspects all impact consumer purchasing behavior [19]. In which the social influence factor is understood as the influence of an individual or a group of people on the buying behavior of consumers. Every individual has people around them that influence their purchasing decisions [20]. These people can be reference groups, family, friends, colleagues, etc.

In their study, Prabowo and Nugroho [2] mentioned the social influence factor in the research model of Indonesian consumers' behavioral intentions towards OFD services for the Go-Food app. In it, Prabowo and Nugroho [2] said that social influence factors include: people who are important to me believe I should use food delivery apps, people who have control over my behavior who believe I should use delivery apps, and people whose opinions I respect believe I should use delivery apps that appreciate like me using food delivery apps. Chai and Yat [17] add a variable that that people who eat together affect the using intention and prefer to use the mobile phone service. Ren et al. [3] suggested that the social influencing factor that affects the intention to use e-mail services is that the surrounding people are important to consumers affect the intention to use, think consumers, recommend using, and have high ratings if used.

In summary, the Social Influence factor in the intention to use the Internet phone service is the surrounding people who are using it and who recommend me to use the service [3], who dine with me. I like to use online services [17], and those whose opinions are appreciated by me recommend me to use the service [2].

The above studies have concluded that the association between social influence factors and the user intention to services is a positive relationship [21, 22].

3 Method

After surveying the Google form for three weeks, we got 260 survey samples, however, of which only 241 were valid and usable for data analysis. We synthesize survey data using R software and give a set of 245 survey forms to analyze the elements impacting the using intention of OFD service for consumers in Vietnam. Table 1 describes the statistics of the sample characteristics.

The 5-point Likert scale is used to determine the degree to which the relevant variables are approved of. In order to evaluate the degree of agreement for all variables that were observed, this paper additionally employs a 5-point Likert scale, with 1 denoting disagreement., 5, and concur with Table 2.

All members of the research team and participants were blinded during the whole trial. The study participants had no touch with anyone from the outside world. The mean of factors is from 1.3527 to 4.4689.

Table 1. Statistics of Sample

Characteristics		Amount	Percent (%)
Sex	Male	46	19.1
	Female	195	80.9
Age	Below 18	12	5.0
	18–30	191	79.3
	31–40	30	12.4
	Above 40	8	3.3
Job	Student	99	41.1
	Officer	108	44.8
	Freelance	30	12.4
	Government staff	4	1.7
Monthly Income	Below 5 million VND	98	40.7
	5–10 million VND	131	54.4
	11–15 million VND	12	5.0
	Over 15 million VND	98	40.7

4 Results

4.1 Akaike Information Criterion (AIC)

The AIC was used by the R program to choose the best model. The AIC has been used in the theoretical environment for model selection [23]. When multicollinearity arises, the AIC approach may also handle a large number of independent variables. To estimate one or more dependent variables from one or more independent variables, AIC can be used as a regression model. The AIC is a significant and practical criterion for choosing a complete and simple model. A model with a lower AIC is chosen on the basis of the AIC information standard. The best model will terminate when the minimum AIC value is reached [24, 25].

R reports detail each phase of the search for the best model. The initial step is to use AIC $= -433.12$ for OFD $=$ f (PEU $+$ ATT $+$ PR $+$ TS $+$ SI) to analyze all 05 independent variables and stop with 02 independent variables with AIC $= -437.71$ for OFD $=$ f (ATT $+$ SI) in Table 3.

Two variables have a P-value lower than 0.05 [26], so they are correlated with Using the Intention of Online Food Delivery Service (OFD), which is in Table 4. Attitude (ATT), and Social influence (SI) impact Using the Intention of an Online Food Delivery Service (OFD).

4.2 Discussion

The results of the AIC Algorithm for Using the Intention of OFD showed that two independent variables, Social Influence (SI) and Attitude (ATT), have a positive and

Table 2. Factor and item

Factor	Mean	Item
Perceived ease of use (PEU)	**4.6846**	Ordering food and drinks from online food delivery services is easy
		The working of my food delivery app is clear and easy to understand
		Using a food delivery app won't require much brainpower
		I feel the online food delivery app has a good interface design that is easy to use
		Ordering food and drinks from online food delivery services is easy
Attitude (ATT)	**4.4689**	I realize satisfaction when using OFD services
		I find online food delivery services attractive
		I realize happiness when using OFD services
Perceived risk (PR)	**1.3527**	Risk of processing orders, not according to my requirements
		During transportation, the appearance and quality of the dish decreased
		Transaction risks
		Risk of leakage of personal information
Time saving (TS)	**4.4440**	Save time ordering and waiting
		Save transaction and payment time
Social influence (SI)	**4.3786**	People around me are using OFD services
		People around me advised me to use OFD services

(*continued*)

Table 2. (*continued*)

Factor	Mean	Item
		People who dine with me like to use OFD services
		Those whose opinions are appreciated by me advise me to use OFD services
Using Intention of Online Food Delivery Service (OFD)	**4.5332**	I will use the call service instead of the usual food ordering
		Next time, I will use online food delivery services
		I will recommend online food delivery services to friends, and colleagues…
		I will use online food delivery services regularly

Table 3. AIC Selection

Model	AIC
OFD = f (PEU + ATT + PR + TS + SI)	−433.12
OFD = f (ATT + PR + TS + SI)	−434.94
OFD = f (ATT + TS + SI)	−436.73
OFD = f (ATT + SI)	−437.71

Table 4. The coefficients

OFD	Estimate	SD	T	P-value	Decision
Intercept	4.10636				
ATT	−0.13893	0.06239	−2.227	0.026898	Accepted
SI	0.23927	0.06674	3.585	0.000409	Accepted

negative impact on the intention to use an online food delivery service, respectively. This is because their p-values are less than 0.05. In descending order, compare the level of influence of these 2 factors on the intention to shop online (OSI): social influence (0.23927), and attitude (−0.13893). The 95% confidence level accepts two associations as a result.

From the AIC Algorithm result, it is shown that the Social Influence factor has the best influence ($\beta = 0.23927$) on the intention to use online food delivery to consumers in Ho Chi Minh City, Vietnam. Therefore, businesses need to pay attention to and improve this factor to improve their intention to use internet telephony services and their delivery capabilities. It shows that attitude has a second influence ($\beta = 0.13893$) on using the intention of online food delivery services in Industry 4.0 for consumers in Vietnam. Therefore, businesses need to pay attention to and improve this factor to improve their intention to use online food delivery services in Industry 4.0 and their delivery capabilities.

5 Conclusion

The results of the AIC Algorithm showed that Using Intention of Online Food Delivery Service (OFD) was influenced by Attitude (ATT), Social influence (SI) and was not impacted by PEU, PR, and Time saving (TS) in integrating with the global trend, as well as the widespread adoption and growth of Industry 4.0. Many online sales systems are growing in Vietnam because of technological advancements, as well as altering payment and delivery methods. This is true of the online meal ordering service industry as well.

Limitations and Future work

The research results of this topic have certain contributions to academics and practical applications to the Online Food Delivery service industry in Vietnam. There are still many limitations in terms of time and money.

First, research topics in the same field in the world and Vietnam have been studied angles. These studies have studied, used different models, and presented a series of factors that affect the intention to use internet telephony services. This paper is also based on that idea and references those studies. However, to match the research context, the author only selects some factors to perform the analysis. Therefore, the author proposes that in future studies; it is possible to choose other models, and factors or combine theoretical models with factors to research and expand and develop for the Vietnamese market.

Next, this study, because of funding issues and regional cultural differences, the study was only conducted for consumers in HCMC. But there has not been an opportunity to expand the research to other provinces, especially in big cities with many consumers using mobile phone services such as Da Nang, Hanoi, Hai Phong, Can Tho, etc. If further studies are carried out widely in more provinces, the benefits to investors will be greater, helping businesses expand and develop in these localities.

Third, this study is only conducted with consumers who have had experience in using e-mail services without mentioning Now's other customers who are partners providing e-mail services (restaurants, bars, etc.) food, and the drivers who deliver the food. Online Food Delivery companies should not only research measures to attract and keep consumers using their e-mail service, but also have a strong connection with their partners. The close, strong relationship between consumers-Online Food Delivery Company-partners will help Online Food Delivery companies survive and succeed in today's fiercely competitive market.

Finally, this study did not use other methods, such as the structural equation model (SEM), to identify the hypotheses and theories and examine the cause-and-effect relationships between research concepts. It only performed data analysis and regression testing of the theoretical model. The aforementioned restrictions have allowed academics to move in a new path when studying online meal delivery and other internet services.

References

1. Kartono, R., Tjahjadi, J.K.: 'Factors Affecting Consumers' intentions to use online food delivery services during COVID-19 outbreak in Jabodetabek area. The Winners **22**(1) (2021)
2. Prabowo, G.T., Nugroho, A.: Factors that influence the attitude and behavioral intention of Indonesian users toward online food delivery service by the go-food application, pp. 204–210. Atlantis Press (2019)
3. Ren, S., Kwon, S.-D., and Cho, W.-S.: Online Food Delivery (OFD) services in Cambodia: A study of the factors influencing consumers' behavioral intentions to use (2021)
4. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**(2), 179–211 (1991)
5. Bauer, R.A.: Consumer behavior as risk taking. American Marketing Association (1960)
6. Prasetyo, Y.T., Tanto, H., Mariyanto, M., Hanjaya, C., Young, M.N., Persada, S.F., Miraja, B.A., Redi, A.A.N.P.: Factors affecting customer satisfaction and loyalty in online food delivery service during the covid-19 pandemic: its relation with open innovation. J. Open Innov. Technol. Market Complex. **7**(1), 76 (2021)

7. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. 319–340 (1989)
8. Farmani, M., Kimiaee, A., Fatollahzadeh, F.: Investigation of Relationship between ease of use, innovation tendency, perceived usefulness and intention to use technology: an empirical study. Indian J. Sci. Technol. **5**(11), 3678–3682 (2012)
9. Aprilivianto, A., Sugandini, D., Effendi, M.I.: Trust, Risk, Perceived Usefulness, and Ease of Use on Intention to Online, Shopping Behavior (2020)
10. Ajzen, I., Fishbein, M.: Belief, Attitude, Intention, and Behaviour: An Introduction to Theory and Research. Addison-Wesley, Reading (1975)
11. Pinto, P., Hawaldar, I.T., Pinto, S.: Antecedents of Behavioral Intention to Use Online Food Delivery Services: An Empirical Investigation', 2021
12. Yeo, V.C.S., Goh, S.-K., Rezaei, S.: Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services. J. Retail. Consum. Serv. **35**, 150–162 (2017)
13. Mensah, I.K.: Impact of government capacity and E-government performance on the adoption of E-Government services. Int. J. Publ. Admin. (2019)
14. Ray, A., Bala, P.K.: User generated content for exploring factors affecting intention to use travel and food delivery services. Int. J. Hosp. Manag. **92**, 102730 (2021)
15. Marafon, D.L., Basso, K., Espartel, L.B., de Barcellos, M.D., and Rech, E.: 'Perceived risk and intention to use internet banking', International Journal of Bank Marketing, 2018
16. Parry, M.E., Sarma, S., Yang, X.: The relationships among dimensions of perceived risk and the switching intentions of pioneer adopters in Japan. J. Int. Consum. Mark. **33**(1), 38–57 (2021)
17. Chai, L.T., Yat, D.N.C.: Online food delivery services: making food delivery the new normal. J. Market. Adv. Pract. **1**(1), 62–77 (2019)
18. Hwang, J., Kim, H.: The effects of expected benefits on image, desire, and behavioral intentions in the field of drone food delivery services after the outbreak of COVID-19. Sustainability **13**(1), 117 (2021)
19. Stet, M., Rosu, A.: PSPC (Personal, social, psychological, cultural) factors and effects on travel consumer behaviour. Econ. Manage. **17**(4), 1491–1496 (2012)
20. Gouwtama, T., Tambunan, D.B.: Factors that influence reseller purchasing decisions. KnE Soc. Sci. 239–245–239–245 (2021)
21. Yousuf, T.: Factors influencing intention to use online messaging services in Bangladesh. SSRN 2826472 (2016)
22. Chen, C.-J., Tsai, P.-H., Tang, J.-W.: How informational-based readiness and social influence affect usage intentions of self-service stores through different routes: an elaboration likelihood model perspective. Asia Pac. Bus. Rev. 1–30 (2021)
23. Mai, D.S., Hai, P.H., Khoi, B.H.: Optimal model choice using AIC Method and Naive Bayes Classification. Proc. IOP Conf. Ser. Mater. Sci. Eng. (2021)
24. Burnham, K.P., Anderson, D.R.: Multimodel inference: understanding AIC and BIC in model selection. Sociol. Meth. Res. **33**(2), 261–304 (2004)
25. Khoi, B.H.: Factors Influencing on University Reputation: Model Selection by AIC: Data Science for Financial Econometrics, pp. 177–188. Springer (2021)
26. Hill, R.C., Griffiths, W.E., Lim, G.C.: Principles of Econometrics. John Wiley & Sons (2018)

A Comprehensive Study and Understanding—A Neurocomputing Prediction Techniques in Renewable Energies

Ghada S. Mohammed[1], Samaher Al-Janabi[2(✉)], and Thekra Haider[1]

[1] Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq

[2] Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Hillah, Iraq

samaher@itnet.uobabylon.edu.iq

Abstract. Today's Renewable energy become the best solution to keep the environment from pollution and provide another source of generation energy. Data scientists are expected to be polyglots who understand math, code and can speak the language of generation energy from natural resources. This paper aims to display the main neurocomputing techniques for prediction in huge and complex renewable database to generation the energy form solar plant. Results clearly show that the LSTM improves the predictive accuracy, speed and cost of prediction. In addition, the results prove that LSTM can serve as a promising choice for current prediction techniques.

Keywords: Information Gain · LSTM · GRU · BLSTM · Alexnet · ZFNet · Renewable Energy

1 Introduction

As a result of the development in the world of technology and information, the digital revolution in different fields, this leads to a significant and noticeable increase in the need for energy, which has become an integral part of our lives. Most energy resources have a lot of limitations and drawbacks so shift towards the depended on the sources of renewable energy in solving problems of meeting the increasing demand for energy, reducing environmental impacts is considered one of the most critical challenge facing the world. Forecasting the amount of energy expected to be produced in the near future will help decision-makers to deal with the increasing demand for this energy and work to achieve a balance between energy production and consumption based on various forecasting techniques, but the prediction of the expected energy with high accuracy is considered a critical challenge, so our work aims to make a comparison among different Neurocomputing prediction techniques to find the most efficient techniques.

Intelligent Data Analysis (IDA) is one of the basic and influential tools in the process of decision making due to its importance in identifying new visions and ideas, it combines different strategies and techniques to collect data from multiple sources and use it

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 152–165, 2023.
https://doi.org/10.1007/978-3-031-27409-1_14

to discover knowledge and interpret it to be accurate and understandable to all. The process of intelligent data analysis begins with defining the problem, determining its data, then defining and using techniques such as artificial intelligence techniques, pattern recognition, and statistics techniques to obtain the required results, and then evaluating, interpreting, and explaining these results and their impact on the decision-making process.

Renewable Energy sources are environmental energy (alternative energy) that reduces the harmful impacts on the environment; This concept of energy is linked to the energy that is obtained from natural sources that produce enormous amounts of energy and is capable of regenerating naturally; the resources of friendly environment energy are driven by the wind, hydropower, ocean waves, biomass from photosynthesis, and direct solar energy; friendly environment energy has many advantages where it considers non-polluting, sustainable, (one-time installation), economic, ubiquitous, safe and it offers a wide variety of options and also there are drawbacks to some of the sources (Maybe more costly or its effects by some environmental influences) [1].

There are many challenges that have curbed renewable energy and didn't help it to expand; the most important of these challenges is: The high-cost challenge compared to the cost of traditional systems for power generation and it constitutes an obstacle to the expansion of this energy. Also, the reliability of environmental and industrial matters that can affect the efficiency of the source used to generate renewable energy is considered an important challenge, It must be taken into account when preparing any feasibility study for a future system for generating renewable energy; the technical innovations, the efficiency development of the methods and techniques that used in renewable energy generation systems are considered an important challenges in this field, these methods have a significant impact on turning some of the challenges into strengths point such as increasing the accuracy will increase the efficiency and reduce the time. Another important challenge in the field of Renewable Energy (RE) is the manpower where the energy generation systems from environmentally friendly sources need more manpower to operate power plants compared to traditional energy generation systems that rely heavily on technology in operation. Our work will try to deal with these challenges from two sides programmable side and the application side to overcome these challenges.

Prediction techniques can be classified into techniques that are related to data mining such as (Random Forest Regression and Classification (RFRC), Boosted Tree Classifiers and Regression (BTCR), Chi-squared Automatic Interaction Detection (CHAID), Bayesian Neural Networks Classifier (BNNC), Decision Tree (DT), Exchange Chi-squared Automatic Interaction Detection (ECHAID), and Multivariate Adaptive Regression Splines (MARS)) [2], and techniques that are related to Neurocomputing such as (Convolutional Neural Network (CNN), Recurrent Neural Network(RNN), Gated Recurrent Units (GRU), and many others algorithms) [3].

AI is a wide area of the researches that refer to the ability of machines to simulate the human intelligence, the most it is popular branches are ML and DL techniques; The algorithms of ML are trained on a different variety of data, and can improve the algorithm accuracy with more data [4], ML is very popular for making the prediction process due to its high performance in dealing with the data heterogeneity (the data that come from different resources, numerous types and complex characteristics) more

than the statistical methods, and handling complex prediction problems; Now with the development of DL techniques, these techniques considered extension of ML techniques and take advantage of the AI capabilities in the predict models) DL consisted of a large number of layers that were capable of learning the characteristics with an excellent level of abstraction [3], These Algorithms operate automatically with elimination of the need for manual operations [4].

2 Related Work

Many researchers have tried to developing prediction models based on deep learning techniques to solve the problem of the increasing demand and the urgent need for electrical energy due to the growing use of electronic devices. There are many different techniques that have been introduced to deal with this problem, through the review of previous works, it has been found that there are number limitation such as the time and the computation complexity and the accuracy problem as shown below.

The authors in [5] proposes an model combine (BLSTM) and the extended scope of wavelet transform to 24-h forecast the solar global horizontal irradiance for the Gujarat, Ahmedabad locations in India, to improve the forecasting accuracy, the input time series statistical features are extracted and decomposes the input into number of finite model functions then reduces it to trained the BLSTM networks. The author based on one year dataset to execution of the proposed model and using different matrices in the evaluation process; the model outperforms in compare with others models but there are some challenges with the design of this model such as: hyper parameters selection and the complexity of simulation time.

The authors in [6] proposed model for forecasting a wind speed based on using deep learning techniques (ConvGRU and 3D CNN)with variational Bayesian inference; historical information for two real-world case studies in the United States is used to apply the model. The results of the model performance evaluation show it outperforms other point forecast models (the Persistence model, Lasso Regression, artificial neural network, LSTM,CNN, GPR and Hidden Markov Model) due to the combination between techniques and using not too wide forecast intervals so the, model need to experiment to wider regions and using advance probabilistic methods to evaluate its performance.

The authors in [7] introduced a two-step wind power forecasting model, the first step is Variational Mode Decomposition (VMD), second step an improved residual-based deep Convolutional Neural Network (CNN). The used dataset was procured from a wind farm in Turkey. The results of the proposed method were compared with the results obtained from deep learning architectures (Squeeze Net, Google Net, ResNet-18, AlexNet, and VGG-16) as well as physical models based on available meteorological forecast data. The proposed method outperformed the other architectures and demonstrated promising results for very short-term wind power forecasting due to its competitive performance.

In [8] the authors Present a model to determine the strategy of real time dynamic energy management of Hybrid Energy Systems (HES) by using a deep reinforcement learning algorithm and training it on numerous data such as water demand, Wind Turbine(WT) output, photovoltaic(PV) output, electricity price, one year of load demand

data to obtain optimal energy management policy, the theory of information entropy is used to compute the Weight Factor (WF) and determine the best between different targets. Simulation results of this study show the optimal policy for control and the cost reduce by up to 14.17%. But the model have many limitation in his structure.

The authors in [9] proposed a method for trade-off multi-objective (practical swarm optimization (MOPSO) algorithm and Techniques for order of preference by similarity to ideal solution (TOPSIS)) that used to achieve a strategy for energy management in system optimal configuration; also examining the strategy on the real-world case; The results show that the TPC /COE/ EC set in (grid-connected, off-grid scheme) each one is optimal in different configurations. The method evaluate based on different perspectives (energy, economic, and environmental).

3 Theoretical Background

3.1 Multi Variant Analysis

The high dimensionality of the dataset that use to build the predictor model is a very important issue because the high dimension of the dataset can include input features that are irrelevant to the target feature so that, this will increase the time complexity of the model, also, the process of training will be slow and the required system memory will be a large amount, all which will reduce the model performance and overall effectiveness; so must select the only important features that have an impact and it useful in prediction the target feature and removing the excessive and non-informative feature [10]; the feature selection technique contribute in cost reduction and performance(accuracy)improvement; in this work information gain, entropy and correlation methods used to perform feature selection.

Information Gain (IG) is a popular filter (entropy-based) technique proposed by Quinlantat, it can be applied to categorical features [11] it represents how much information would be gained by each attribute (The attribute with the highest information gain, is selected), the Entropy(H) is the average amount of information (change in uncertainty)that needed to identify the attribute [12], the interval of the Entropy is [0, 1]. The (IG) measure is biased toward the attributes with many outcomes(values).

$$H = -\sum_{i-1}^{D} p_i * \log_2 p_i \tag{1}$$

$$IG = H - \sum \frac{DO}{DS} * H_J \tag{2}$$

where DO, DS is the dataset and sub-dataset, H_J is entropy of sub-dataset.

3.2 Long Short-Term Memory Algorithm (LSTMA)

It is one of Recurrent Neural networks (RNNs) that demonstrated clear superiority [15], the default behaviour is remembering the context for long intervals so it is capable of facilitating detection the long term dependencies. In LSTM, the memory cell is operated

instead of the activation function of the hidden state in the RNN; LSTM consists of cell with three gates (Memory cell, input, forget and output gates), The three gates regulate the preceding information (the flow of information to next step) while the cell used to remember the values (maintain the state) over different intervals [13]. Each gate has its special parameters that need to be trained. Also, there are hyper-parameters that need to be selected and optimized (hidden neurons number and batch size). Due to its impact on the performance of LSTM architecture [14, 15]; The LSTM architecture was presented by Hochreiter and Jürgen [16–18], there are many of modifications that performed on the classical LSTM architecture to decrease the design complexity and time complexity.

Algorithm 1: LSTM Algorithm

Input: Dataset.
Output: Predicted of target
Initialization the net parameters(#epochs , batch_size,input_size ,LR, # units)
//Control net complexity and input the data blocks to the allocated layers

1. **While** not meeting the max _epoch do
2. **for** each batch in max _iterations
3. **for** each x(t) in the batch
4. Compute(c(t) ,i(t) ,fg(t),cs(t),o(t),y(t))

$$c(t) = tanh(Wc(x(t), y(t-1)) + bc)$$
$$i(t) = \sigma(Wi(x(t), y(t-1)) + bi)$$
$$fg(t) = \sigma(Wf(x(t), y(t-1)) + bf)$$

cs(t)=f(t) ∗c(t-1) *(i(t)*c(t))

$$o(t) = \sigma(Wo(x(t), y(t-1)) + bo)$$
$$y(t) = o(t) * tanh(c(t))$$

5. **End for**
6. **End for**
7. **End while**
8. **End**

Where c(t) is the Memory cell, i(t) is the Input Gate, fg(t)is the Forget Gate, cs(t) is the New cell state, and o(t) is the output Gate; Wc, Wi, Wf and, Wo are the metrics of weights respectively; bc, bi, bf, andbo are represent the biases; x(t) is the input, σ is the logistic Sigmoid Function; for each batch needed W, y, b to be trained and updated input of the model.

4 Methodology

The proposed model consists from multi step as shown in Fig. 1.

4.1 Description of Dataset

In this work, there are two datasets (solar and weather dataset) each one of these datasets has different features; the solar dataset consists of 68778 samples while the weather dataset consists of 3182 samples. The solar panel dataset contains seven features

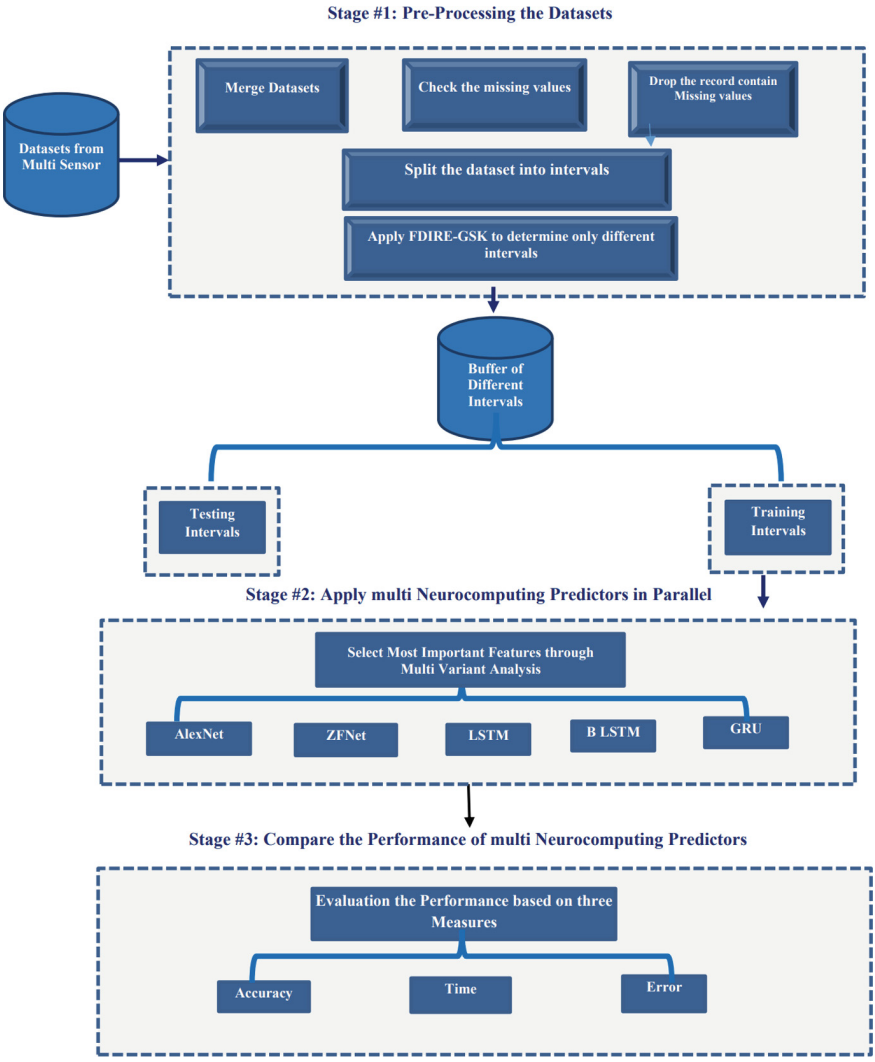


Fig. 1. The proposed model

(Date_time, plant_id, source_key, Dc_Power, Ac_Power, Daily_Yield, Total_Yield), whereas the weather dataset includes six features (Date_time, plant_id, source_key, Ambient_temperature, Module_temperature, irradiation).

4.2 Preprocessing the Data

This step involves handling the data sets, where:

1. The real time data capturing from multi sensors (solar plant sensor, weather sensor)
2. The datasets merging in one dataset based shared features (Source_key, and Date_time) this will causes to reduce the number of shared features and compressed the data in vertical manner.
3. Checking the new merged data set for the missing values, if there are any missing value the record will dropped this will causes to compressed the datasets in horizontal manner and this precise data compression will caused to reduce the time of computation.
4. Now the dataset will be cleaned, to increase the accuracy of predictor, the most important features in the dataset must determine, in our work using the, Information gain (that based on the computing the entropy) and the correlation methods to determine the importance of each features and its relation to the targets as shown in the Table 1.

Table 1. Information gain and correlation of the dataset features

Feature	Information Gain	Correlation
Dc_Power	1	1
Ac_Power	0.98746418	1
Daily_Yield	0.963139904	0.082
Total_Yield	0.996712596	0.039
Ambient_ Temperature	0.734895299	0.72
Module_ Temperature	0.734895299	0.95
Irradiation	0.734895299	0.99
Date	0.290856799	−0.037
Time	0.433488671	0.024
Hours	0.314651802	0.024
Minutes	0.110210969	0.0012
Minutes_ Pass	0.433488671	0.024
Date_Str	0.290856799	−0.037

From the Table 1 can show the Dc_Power feature has maximum information gain (1) and correlation to target feature (Dc_Power),also the Ac_Power has high correlation (1) to the target where the Date feature and data_str features have lowest correlation (−0.037). The Total_Yield features has highest information gain value () also the

Ac_Power has high information gain (0.98746418) where the Minutes feature has lowest information gain (0.110210969), these method determine the most related feature to target features and the most feature that have effect on the generation of Dc_Power

1. Now the datasets will contain the most important features only and based on the time and source key the data set will split into intervals and (each intervals for 15 min)
2. Based on the FDIRE-GSK algorithm, different intervals only will determine and saved in buffer to using them in the implementation of predictors; this determination to different intervals will increase the speed of performance of multi predictors model.
3. The data split into (Train_X (80%) of the data to train the model and Test_X(20%) to evaluate the model).

4.3 Built in Parallel Multi Predictor

In our work perform multi predictor in parallel in order to comparison between them and find the most accrues one, these predictors build based on some of Neurocomputing techniques (AlexNet, ZFNet, LSTM, BLSTM, GRU).

4.4 Performance Evaluation

The quality of the multi predictor model was evaluated using the error and accuracy measures. The results showed that a comparison between the predictors performance in terms of the error (mean square error) for each techniques and the accuracy. See algorithm 2 that show the main steps of the proposed model.

5 Results and Discussion

The aim of this work is predictions of maximum DC_Power that generated by solar plant, the prediction process based on different Neurocomputing techniques to compare between them and find the most efficient one., the merging process will reduce the number of processed features (reduce the number of columns from 13 feature to 11 features) then the process of cleaning the dataset from the missing values will reduce the number of processed samples (reduce the number of samples), this will increase the speed of predictor spatially the data collected in real time (Figs. 2, 3, 4 and 5).

Algorithm 2: Main Algorithm 9

Input: Data sets contain features capture from multi sensors

Output: Prediction maximum DC power

```
// Handel dataset based on FDIRE-GSK
for each i in NRK                                   // k:# datasets
  for each j in NCK                                 // NCK:#column
        Call Merging
  End for
End for
for  each i in NR
  for r each j in NC
  Checking the missing values
  Compute MIF                                       //MIF: Most important features
   End for
End for
for  each i in NR
  for each j in NMIF
   W[i]=stream o  data                              //split based on SOURCE _KE
                                                    &each 5 minutes
  End for
End for
for  each  l  in TNi                                //TNi: total number of intervals
  for r each j in NR
   Call FDIRE-GSK
  End for
End for
//Build in parallel multi predictor
for each i in NR
  for each j in NC
  Split Data (Train_X, Train_Y, Test_X, Test_Y)
  End for
End for
for each i in Train_X
  for each j in Train_X
   Call (AlexNet , ZFNet , LSTM, BLSTM  ,GRU)
  End for
End for
//Evaluation the performance
 Compute the error
 Compute the accuracy
End
```

The process of feature selection and determine the irrelevant feature in data set will effect on the accuracy of the predictor because it operate on most important features to the target, the selection based on the information gain value (that based on the entropy) and on the correlation between the datasets features and the target. In our work using the, Information gain (that based on the computing the entropy) and the correlation methods

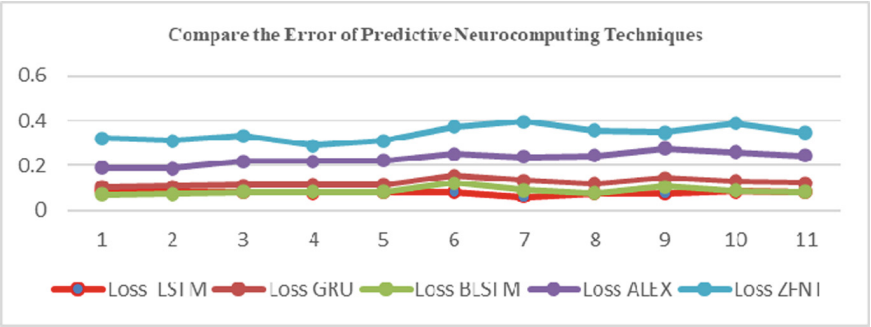


Fig. 2. Compare the Neurocomputing techniques based on Error

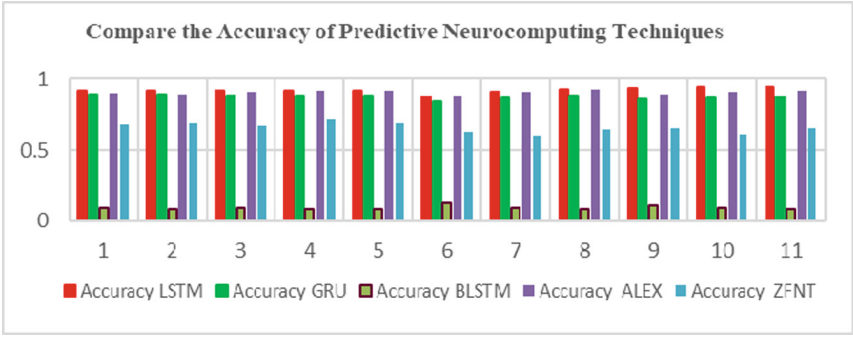


Fig. 3. Compare the Neurocomputing techniques based on Accuracy

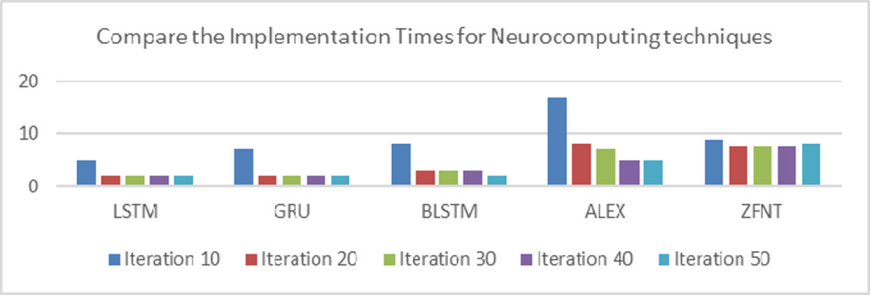


Fig. 4. Compare the Neurocomputing techniques based on Implementation Time

to determine the importance of each feature and its relation to the targets as shown in the Table 1. Building the multi predictor in parallel manner and compare between them based on the error of each predictor as shown in the Table 2, and the accuracy of the model as shown in the Table 3 while Table 4 shown the time required to execution each predictor. Finally; Total Time of each predictor shown in Table 5.

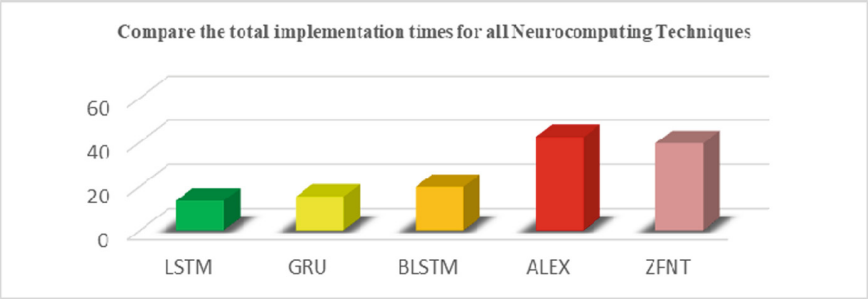


Fig. 5. Compare the Neurocomputing techniques based on Total Implementation Time

Table 2. Loss value of each predictor

LSTM	GRU	BLSTM	ALEX	ZFNT
0.091702	0.10561	0.073293	0.189727	0.320713
0.083836	0.10643	0.075757	0.188772	0.311013
0.081828	0.113829	0.0839	0.220824	0.33184
0.081664	0.114509	0.083522	0.220975	0.290288
0.081894	0.115322	0.083605	0.223171	0.312138
0.082543	0.156428	0.124139	0.251438	0.372194
0.062385	0.133367	0.092274	0.239495	0.395843
0.07846	0.119534	0.078328	0.243314	0.356411
0.076677	0.145451	0.107084	0.275609	0.349333
0.085221	0.130142	0.087893	0.260646	0.387477
0.083479	0.124513	0.083356	0.243735	0.346402

6 Conclusion and Future Works

This paper implemented neurocomputing techniques for predicting the DC_Power in renewable energy. In addition to that, it analyzed and compared some of the existing prediction neurocomputing techniques in an attempt to determine the main parameters that have the most important effects on their predictor. From the analysis, we found the techniques that are not dependent on randomization provided better results, while the ones using mathematical basis offered more powerful and faster solutions. In the light of this, mathematical basis is used in the proposed model.

The results show that the LSTM performs better than other prediction techniques in prediction in renewable domain. Also; it achieves an improvement in accuracy, speed of prediction and less cost. Therefore, the LSTM is promising choice compared to other prediction techniques. The experimental results also show that the LSTM employed in this work overcomes some of the shortcomings in other prediction techniques.

Table 3. Accuracy of each predictor

Accuracy LSTM	Accuracy GRU	Accuracy BLSTM	Accuracy ALEX	Accuracy ZFNT
0.918298	0.89439	0.094293	0.895707	0.679287
0.916164	0.89357	0.085757	0.89424	0.688987
0.918872	0.886171	0.0939	0.9061	0.66816
0.918936	0.885491	0.083522	0.916478	0.709712
0.919106	0.884678	0.083605	0.916395	0.687862
0.877957	0.843572	0.124139	0.875861	0.627806
0.907615	0.866633	0.092274	0.907726	0.604157
0.92154	0.880466	0.078328	0.921672	0.643589
0.933323	0.854549	0.107084	0.892916	0.650667
0.941779	0.869858	0.087893	0.912107	0.612523
0.946521	0.875487	0.083356	0.916644	0.653598

Table 4. The Time of each predictor model

Iteration	**LSTM** (s)	**GRU** (s)	**BLSTM** (s)	**ALEX** (s)	**ZFNT** (s)
10	5	7	8	17	9
20	2	2	3	8	7.5
30	2	2	3	7	7.5
40	2	2	3	5	7.5
50	2	2	2	5	8

Table 5. Total Time of each predictor model

Iteration	LSTM	GRU	BLSTM	ALEX	ZFNT
50	13.441	15.339	19.737	42.197	39.458

The results also show that some of predictors give very close results to each other such as (ALEX and ZFNT) while some of them are similar in both the work structure and the results such as (GRU and LSTM). As future work, we planning to develop LSTM by use optimization Algorithm (i.e., GSK). Using one of optimization algorithms such as swarm optimization, Ant Colony Optimization (ACO) and Genetic Algorithm (GA) to determine and select the most important features in order to reduce the time used in the predictor.

References

1. Al-Janabi, S., Alkaim, A.: A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis. Egypt. Informatics J. **23**(2), 271–290 (2022). https://doi.org/10.1016/j.eij.2022.01.004
2. Al-Janabi, S., Alkaim, A.F., Adel, Z.: An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft. Comput. **24**(14), 10943–10962 (2020). https://doi.org/10.1007/s00500-020-04905-9
3. Baydyk, T., Kussul, E., Wunsch II, D.C.: Intelligent Automation in Renewable Energy. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-02236-5
4. Al-Janabi, S., Mahdi, M.A.: Evaluation prediction techniques to achievement an optimal biomedical analysis. Int. J. Grid Util. Comput. **10**(5), 512–527 (2019)
5. Medina-Salgado, B., Sánchez-DelaCruz, E., Pozos-Parra, P., Sierra, J.E.: Urban traffic flow prediction techniques: a review. Sustain. Comput. Informatics Syst. **100739**,(2022). https://doi.org/10.1016/j.suscom.2022.100739
6. Sony, S., Dunphy, K., Sadhu, A., Capretz, M.: A systematic review of convolutional neural network-based structural condition assessment techniques. Eng. Struct. **226**, 111347 (2021). https://doi.org/10.1016/j.engstruct.2020.111347
7. Singla, P., Duhan, M., Saroha, S.: An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network. Earth Sci. Inf. 1–16 (2021). https://doi.org/10.1007/s12145-021-00723-1
8. Liu, Y., et al.: Probabilistic spatiotemporal wind speed forecasting based on a variational Bayesian deep learning model. Appl. Energy **260**, 114259 (2020). https://doi.org/10.1016/j.apenergy.2019.114259
9. Yildiz, C., Acikgoz, H., Korkmaz, D., Budak, U.: An improved residual-based convolutional neural network for very short-term wind power forecasting. Energy Convers. Manage. **228**, 113731 (2021). https://doi.org/10.1016/j.enconman.2020.113731
10. Zhang, G., et al.: Data-driven optimal energy management for a wind-solar-diesel-battery-reverse osmosis hybrid energy system using a deep reinforcement learning approach. Energy Convers. Manage. **227**, 113608 (2021). https://doi.org/10.1016/j.enconman.2020.113608
11. Zhao, P., Gou, F., Xu, W., Wang, J., Dai, Y.: Multi-objective optimization of a renewable power supply system with underwater compressed air energy storage for seawater reverse osmosis under two different operation schemes. Renew. Energy **181**, 71–90 (2022). https://doi.org/10.1016/j.renene.2021.09.041
12. Al-Janabi, S., Alkaim, A.F.: A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft. Comput. **24**(1), 555–569 (2019). https://doi.org/10.1007/s00500-019-03972-x
13. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. **53**(8), 5455–5516 (2020). https://doi.org/10.1007/s10462-020-09825-6
14. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Asari, V.K., et al.: The history began from alexnet: a comprehensive survey on deep learning approaches (2018). arXiv:1803.01164. https://doi.org/10.48550/arXiv.1803.01164
15. Mirzaei, S., Kang, J.L., Chu, K.Y.: A comparative study on long short-term memory and gated recurrent unit neural networks in fault diagnosis for chemical processes using visualization. J. Taiwan Inst. Chem. Eng. **130**, 104028 (2022). https://doi.org/10.1016/j.jtice.2021.08.016
16. Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F., Chandran, V.: Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. IEEE Access **6**, 49325–49338 (2018). https://doi.org/10.1109/ACCESS.2018.2868361

17. Darmawahyuni, A., Nurmaini, S., Caesarendra, W., Bhayyu, V., Rachmatullah, M.N.: Deep learning with a recurrent network structure in the sequence modeling of imbalanced data for ECG-rhythm classifier. Algorithms **12**(6), 118 (2019). https://doi.org/10.3390/a12060118
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)



Predicting Participants' Performance in Programming Contests Using Deep Learning Techniques

Md. Mahbubur Rahman[1], Badhan Chandra Das[2], Al Amin Biswas[3(✉)], and Md. Musfique Anwar[2]

[1] Iowa State University, Ames, Iowa, USA
mdrahman@iastate.edu
[2] Jahangirnagar University, Dhaka, Bangladesh
badhan0951@gmail.com, manwar@juniv.edu
[3] Daffodil International University, Dhaka, Bangladesh
alaminbiswas.cse@gmail.com

Abstract. In recent days, the number of technology enthusiasts is increasing day by day with the prevalence of technological products and easy access to the internet. Similarly, the amount of people working behind this rapid development is rising tremendously. Computer programmers consist of a large portion of those tech-savvy people. Codeforces, an online programming and contest hosting platform used by many competitive programmers worldwide. It is regarded as one of the most standardized platforms for practicing programming problems and participate in programming contests. In this research, we propose a framework that predicts the performance of any particular contestant in the upcoming competitions as well as predicts the rating after that contest based on their practice and the performance of their previous contests.

Keywords: Codeforces · Programming Contest · Performance Analysis and Prediction

1 Introduction

Codeforces is an online programming practice and contest hosting platform maintained by a group of competitive programmers from ITMO University, led by Mikhail Mirzayanov. According to Wikipedia, there were more than 600,000 registered users on this site. There are several certain features of Codeforces as follows. This site has been developed specially for competitive programmers while preparing for the programming contests. A registered user of this platform can use it in terms of practicing anytime and participating in the contests running at that time with the facility of the internet. There is a rating system commonly known as divisions of each contestant taking part in the contests based on their performance, i.e. capability to solve the problems according to their difficulty

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 166–176, 2023.
https://doi.org/10.1007/978-3-031-27409-1_15

level of that contest as well as the previous ones. The rating system, divisions and titles are shown in Table 1. The contestants can try to solve the unsolved problems of any contests, even after the contest, also known as upsolve. There are several types of contests that can be hosted in Codeforces. Among them, the most popular one is short contests held for two hours, which is also known as Codeforces Round. It can be conducted once a week. Another one is a team contest, where any registered user can invite any other registered users (at most two) for a contest. The users can also get connected (follow- following) with each other in order to watch updates of them. The trainers or institutions who organize the contests usually do this to track the progress of the trainees and students. One of the important and effective features of this widely used platform is, there is a community platform like Stack overflow, to get the solutions of the problems faced during the contest and in practice. However, this difference between this community platform and others is, it is dedicated for the competitive programmers trying to solve any programming problems while practicing independently or the problems after the contests. The users can also get a list of tagged problems, e.g. dynamic programming problems, greedy problems, etc. to practice and get experts or work on the weak parts of him or her on specific types of problems.

Table 1. Codeforces User Rating and Divisions

Rating Bounds	**Color**	**Division**	**Title**
>= 3000	Black & Red	1	Legendary Grandmaster
2600–2999	Red	1	International Grandmaster
2400–2599	Red	1	Grandmaster
2300–2399	Orange	1	International Master
2100–2299	Orange	1	Master
1900–2099	Violet	1/2	Candidate Master
1600–1899	Blue	2	Expert
1400–1399	Cyan	2/3	Specialist
1200–1399	Green	2/3	Pupil
<= 1199	Gray	2/3	Newbie

In this research, we propose a framework which predicts the performance of each individual programmer in upcoming contests based on his or her previous contests. The performance of the contestants is performed in two perspectives. First, we predict whether the rating of the contestant will increase or decrease, second, the rating itself of that corresponding contestant. The performance tracking of the contestants in order to recommend him to improve his performance in the impending ones. The main contributions of this paper is as follows.

1. We predict the performance of each contestant by analyzing his or her performances in the previous contests and practice problems.

2. The ratings of the programmers will also be predicted along with the percentage of increase or decrease of their ratings.
3. This experimental research is conducted on a real-world dataset obtained from Codeforces.

The remaining sections are organized as follows. The Sect. 2 covers relevant works in this topic. In Sects. 3 and 4, we explain the problem definition and proposed methodology respectively.The experimental outcomes are presented in Sect. 5. In Sect. 6, we conclude the paper.

2 Literature Review

To identify the gap in the available research, we have conducted extensive searches and investigations of numerous related studies. However, a very little amount of research work has been accomplished on this topic. Using the students' data of secondary school, Amra et al. [17] applied KNN and Naive Bayes classifiers to predict the students' performance. The obtained result showed that Naive Bayes outperformed KNN by attaining the accuracy of 93.6%. Babić et al. [15] tried to imbed the links between student academic motivation and their behaviour in the learning management system (LMS) course. Three different machine learning (ML) classifiers namely neural networks, support vector machines, and decision trees were applied to classify the students. Though the performance of all the classifiers were significant but the neural network was more promising than others applied models in detecting the student academic motivation based on the behaviour.

2.1 Academic Performance Prediction

Waheed et al. attempted to develop a system that can predict students' academic success based on clickstream data and assessment results in a virtual learning environment. They used the artificial neural network (ANN) to classify the student performance into different classes and compared the obtained result of ANN with two baseline methods namely support vector machines and logistic regression [14]. It is observed that ANN outperformed the baseline methods. Several works related to student performance prediction have also been accomplished [16,18–20].

2.2 Contest Performance Prediction

Sudha et al. [8] worked on the classification and recommendation of competitive programming problems using Convolution Neural Network (CNN). The goal of their proposed system is to determine the required approach for solving the problem. W. Looi analyzed single C++ source code submission on Codeforces and tried to predict a user's rank and country [9]. Among all the applied models, the neural network attained the highest accuracies of 77.2% in rank prediction

(within one rank) and 72.5% in the country's prediction. A. Alnahhas et al. investigated ML techniques to develop a system that can predict the contestant's future performance by dissecting their past rating record [10]. Here, they applied five different baseline machine learning approaches. Besides this, they proposed a new deep learning model for result comparison with baseline. To conduct this research, they collected public data from the Codeforces website. They found that most of the applied techniques attain an acceptable result but the deep learning model performed better than the baseline. Chowdhury et al. [11] trained a Kohonen Self organizing feature map (KSOFM) neural network on log data regarding programmers' performance. Here, programmers are grouped into three distinct clusters ie. 'at risk', 'intermediate', and 'expert'. The proportional rules made classification with an accuracy of 94.00%. Besides this, three more models namely multilayer neural networks, decision tree, and support vector machine were trained using the same dataset. Among them, feedforward multilayer neural networks and decision trees have achieved an accuracy of 97.00% and 96.00% respectively. The precision of the support vector machine was about 88.00%, but it attained the highest recall of 99.00% in terms of distinguishing 'at risk' students. By investigating ten years of TopCoder algorithm competitions, J. R. Garciaa et al. reported on the learning curves [12]. They also discussed how these learning curves are employed in university courses. Later, it can aid them to explain the impact of competitive programming in a class.
Ishizue et al. [13] employed machine learning models to try to simplify the process of predicting placement outcomes outside of the conventional, time-consuming placement examination and the level of programming competence outside of a programming contest. The explanatory variables consist of psychological assessments, programming tasks, and student-completed surveys.

Ohashi et al. proposed a unique feature extraction technique and convolutional neural networks to classify the source code. To demonstrate the proposed algorithm, they have used data of an online judge system. It is shown that the model performed well in predicting the right category with high accuracy. Intisar and Watanobe [11] tried to classify the programming problems. For this, they made use of the two topic modeling techniques namely Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) for extracting the relevant features. Then, by utilizing these topic modeling features and Naive TF-IDF features, six classifiers were trained. It is found a series of beneficial trade-offs between the applied models in terms of dimensionality and accuracy.

3 Proposed System

The proposed system gets started with the collection of dataset from online programming practice platform Codeforces using its public Application Programmable Interface (API). Then some pre-processing tasks had been performed on the collected data to convert them into sequences. Then some state-of-the art sequence to sequence models had been trained and tested on the collected data.

3.1 Dataset Collection

Our proposed framework includes two phases. At first, we collected the data of 100 contestants from Codeforces using codeforces public API. The data[1] includes contestant's ratings, competition ranks, problem submissions, and submission verdicts. After collecting the data, we did some pre-processing. We considered each contest as a timestamp. Each timestamp has four types of features.

1. Rating: Each contest represents a timestamp. The rating is a metric to evaluate an user/contestant. The more the rating is, the better performer the contestant is. This changes after each contest based on the rank of the user in that contest. We are going to predict this feature.
2. Rank: This is the rank/position of the contestant in that contest. The rank is decided by the solve rating of the contestant in that contest. The more the solve rating is, the better rank the contestant gets.
3. Solve Rating: Each contest has several problems and each problem has a different point based on its difficulty. The point of each problems decreases with time. The quicker a contestant solves a problem, the better points he gets. A contestant's solve rating is calculated by adding the points of each problem he solved during that competition.
4. Practice Features: This is the information of a contestant about how much practices he did after the previous contest and before the current contest.
   a. Accepted(AC): It represents the number of problems that the contestant solved before the current contest and after the previous contest.
   b. Wrong Answer(WA): It denotes the number of problems that were attempted by the contestant but failed to solve it correctly before the current competitions and after the previous competitions.

Finally, we built a dataset of sequences where each continuous 16 timestamps of a user are considered as a sequence. Among the 16 timestamps, first 15 timestamps were used as the input and 16th timestamp was used as the target. 80% of the sequences were used to train the models and the rest of the 20% of the sequences were used to test/evaluate the performance of the trained models.

3.2 Frameworks

In the second phase of our proposed system, we apply several state-of-the-art neural network models to predict the performance of each contestant in the impending programming contests based on previous contests. First, we describe the concepts of Recurrent Neural Network (RNN), since Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) both are categorized into that one. Then we describe Bidirectional LSTM (Bi-LSTM), and a combination of LSTM with an attention layer (LSTM+AL).

a. Recurrent Neural Network RNN is a special class of ANN, which was originally proposed by Hopfield [4]. There is a basic difference between the conventional simple feed-forward neural network and RNN. Whereas in a feed-forward

[1] https://cutt.ly/nL120M9.

network, information flows in a single direction from the input nodes to the output nodes via the hidden nodes, RNN remembers the past sequences as well as being operated by the present node i.e. the system comes back to its previous node while running the current note. As a result, cycles or loops happen in the network. As it visits its previous nodes in every iteration, these RNN approaches perform well in sequence tasks and are widely used in prediction tasks e.g. stock market prediction, language translation, etc.

b. Long Short-Term Memory As mentioned earlier, RNN remembers the past sequences and puts on the proper context. Moreover, it remembers that information for a small duration of time. As a result, RNN falls short in terms of long sequences of data needed to process. Long Short Term Memory, which is commonly known as LSTM is a particular type of RNN, proposed by Hochreiter et al. in 1997 [5], which can mitigate this issue. While putting new information RNN transforms the existing information once applied a function. As a result, the entire information gets modified, on the whole, it fails to infer any such consideration for important or less important information. On the other hand, LSTM makes little modification to the information. In LSTM, this information flow is called cell states. In this way, LSTMs can selectively remember or forget things as per the context. The LSTM Architecture varies a little in terms of its internal components. Unlike RNN, it contains four internal cells inside a single LSTM block. In order to build the LSTM model, we used four LSTM layers with 256 neurons. After each LSTM layer, we used a dropout layer with a drop rate of 0.5. Then, we added a dense layer of 100 neurons with the activation function Relu. At last, a dense layer was used to output the features of the next timestamp of the sequence.

c. Bi-LSTM In RNN, Bidirectional LSTM is commonly known as Bi-LSTM where Bidirectional RNN are just putting two independent RNNs together. Similarly, Bidirectional LSTM is putting two independent LSTMs together so that the networks can have both backward and forward information about the sequence at every time stamp. Bi-LSTM processes inputs in both the past-to-future and future-to-past directions. The thing that differentiates this approach from the unidirectional one is LSTM runs backward it's preserved information from the future and uses the two hidden states combined which is able to preserve information from both the past and future at a given time. The simple building block of bidirectional LSTM has been shown in Fig. 1.

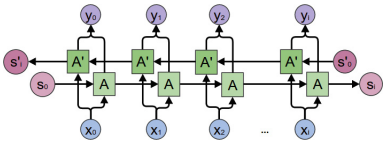


Fig. 1. Bi-Directional LSTM block

d. LSTM with Attention Mechanism The Attention Mechanism is one of the most widely used methods in the Deep Learning research area. It was first proposed by Bahdanau et al. in 2014 [7]. The main bottleneck of the earlier methods of Attention Mechanism such as encoder-decoder-based RNNs/LSTMs, fall apart to deal with long sequences. Moreover, those fail to emphasize any important sequences or patterns. Then, the idea of Bahdanau et al. was not only to keep track of long sequences but also to put more weight on the patterns which would be much needed to predict the outcome.

3.3 Models Configuration

In this paper, for all the models described above, we have configured four corresponding model layers with 256 neurons. After each layer, we used a dropout layer of drop rate 0.5. Then, we added a dense layer of 100 neurons with activation function Relu. Finally, a dense layer was used to output the features of the next timestamp of the sequence. In order to train the models, we used 'mae' and 'adam' as the loss function and optimizer respectfully. We trained all of the models for 1000 epochs with batch size 256. During training, weights of the best accuracy for each of the models were saved using the checkpoint. To check the efficacy of all the models with the test dataset, we used the saved weights of each of the trained models. The test dataset was sent through each of the models and accuracy was calculated to evaluate the models.

4 Experiment and Result Discussion

4.1 Evaluation Metrics

Metrics such as **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**, which are often used to determine correctness for continuous data, have been employed to evaluate our proposed framework. These two metrics have been increasingly utilized by researchers to demonstrate the efficacy of their method [1–3].

Mean Absolute Error *(MAE)*: It is a terminology that provides the measurement of flaws in the assessment compared with original values. It is also referred to **Absolute Accuracy Error (AAE)**. **MAE** represents the average of all **AE**s. The MAE is denoted as,

$$MAE(y, y') = \frac{1}{N} \sum_{k=1}^{N} \left| y_k - y'_k \right| \tag{1}$$

Here, y_k and y'_k denote the actual rating of k^th sample and the predicted rating of them from the models mentioned respectively. N refers the number of samples.

Root Mean Squared Error *(RMSE)*: It is a measurement of standard deviation that indicates how far the predicted value deviates from the actual value. Typically, this approach is suitable for finding the residuals' standard deviation. Residuals are the prediction errors, or the distance between the regression

line and the actual data points. Equation 2 shows how **RMSE(t)** is calculated. Where the notations are the same as described in Eq. 1.

$$RMSE(y, y') = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[y_k - y'_k\right]^2} \tag{2}$$

Mean Squared Error *(MSE)*: It is also known as Mean Squared Deviation (MSD), which is another well known evaluation metric userd for prediction tasks.

$$MSE(y, y') = \frac{1}{N}\sum_{k=1}^{N}\left[y_k - y'_k\right]^2 \tag{3}$$

R-squared (R^2): R-squared is a statistical metric that represents the proportion of the variance for an observed variable that's explained by a predicted variable or variables in a predictive model. The correlation describes the strength of the association between the observed and predicted values. It also describes the degree to which the variance of one variable explains the variation of the second variable. **R-squared** measure is ranged between 0 to 1 and usually mentioned as percentages. The more the value of this metric is, we consider the more precise the predictive model to be. Equation 4 shows how **R-squared** is calculated [21].

$$R^2(y, y') = 1 - \frac{\sum_{k=1}^{N}[y_k - y'_k]^2}{\sum_{k=1}^{N}[y_k - y']^2} \tag{4}$$

$$y' = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{5}$$

4.2 Experimental Results

Does considering contestants' practice as an input feature helpful for better accuracy during the contestants' performance prediction? To answer this question, at first, we ran the models without the contestant's practice feature. Then, we ran the models without the contestant's practice features. Result of the different models to predict the performance of the programmer on Codeforces is presented in Tables 2 and 3.

Table 2 presents the performance of different models for different evaluation metrics without considering the contestant's practice features. Among all the four models, the LSTM with attention achieved the lowest RMSE value (69.629) outperforming other three models: LSTM (73.133), GRU (85.968), and Bi-LSTM (75.024). The LSTM with attention model also outperformed other three models in terms of other three evaluation metrics: MSE, MAE and R^2. On the other hand, when the contestant's practice information is used as a feature, the performance of all models significantly improves (Table 3). The values of RMSE, MSE, and MAE decrease, and R^2 increase significantly for all models. From the above

Table 2. Experiments results of different models for different evaluation metrics without considering contestant's practice features

Performance Metrics	LSTM	LSTM + Attention	GRU	Bi-LSTM
RMSE	73.133	**69.629**	85.968	75.024
MSE	5348.436	**4848.302**	7390.642	5628.728
MAE	57.523	**54.243**	66.324	59.069
R^2	0.906	**0.930**	0.40	0.7942

Table 3. Experiments results of different models for different evaluation metrics considering the contestant's practice features

Performance Metric	LSTM	LSTM + Attention	GRU	Bi-LSTM
RMSE	59.287	**51.325**	72.342	62.244
MSE	3948.436	**3243.234**	6089.834	4467.907
MAE	42.67	**39.217**	53.219	45.989
R^2	0.946	**0.97**	0.884	0.928

two tables, we can see that the LSTM with Attention model provides better accuracy than others in both cases. By analyzing the four-evaluation metrics, it is observed that LSTM with the Attention model performed better than the other three applied models, and GRU performed worst among the applied four models. So, the sequence (from best to worst) of the performance of the models is LSTM with the Attention model, LSTM model, Bi-LSTM model, and GRU model.

5 Discussion

In this work, we showed how contestants' future performance could be predicted by employing deep learning models. We also found that using previous details practice information as input features improves the model's accuracy significantly. When the practice information is included as an input feature, LSTM with Attention performs best. The RMSE, MSE, MAE and R^2 of LSTM with the Attention model were 51.325, 3243.234, 39.217, and 0.97 respectively (Table 3). Then the second-best model is LSTM which got RMSE 59.287, MSE 3948.436, MAE 42.67, and R^2 0.946. We can see that LSTM with Attention provides at least 8% better performance in every metric than the second-best model. On the other hand, LSTM with Attention also achieved the highest efficacy excluding the practice information as the input feature as per our experiment. The RMSE, MSE, MAE, and R^2 of LSTM with the Attention model were 69.629, 4848.302, 54.243, and 0.93 respectively while the values of the second-best model LSTM are 73.133, 5348.436, 57.523 and 0.906 respectively (Table 2). LSTM with Attention shows at least 5% better performance in every metric than the second-best

model. We found LSTM with Attention as the best model in both cases. We can see that the performance of LSTM with the Attention model improves significantly when the practice information is used as the input features. The RMSE, MSE, MAE, and R^2 values of LSTM with Attention model improve about 26%, 33%, 24%, and 10% respectively. It proves that the more a contestant practices, the better he/she does well in the future. In other words, we can conclude that practices make a big difference in the upcoming competitions. Therefore, without practice information, the prediction is not as accurate as when the prediction is done with practice information. Here, we collected the practice information only from CodeForce's website. What if the contestant practices on other platforms or offline? In that case, our models will fall short to provide this performance. In the real world, these participants can get alert about their signs of progress by the predictions of our proposed method and they can improve their skills to perform well in their next contests. In the future, we are planning to include practice information from other platforms too. Besides, we calculated the practice information by summing the number of problems solved before the contest whereas we didn't consider the difficulty level of the problems. What if the contestant solves only the easier problems? It won't help him to perform better but the model will predict that he will do better.

6 Conclusion and Future Work

In this research, we provide a method for predicting participant ratings and analyzing their performance. We used a real-world Codeforces dataset to validate our methodology. The experiments had been conducted by considering the contestants' practice features as well as without considering them. In the future, we aim to consider each problem's (solved before the next competition) difficulty level as a feature and employ the data from other platforms also.

References

1. Jaidka, K., Ahmed, S., Skoric, M., Hilbert, M.: Predicting elections from social media: a three-country, three-method comparative study. Asian J. Commun. **29**(3), 252–273 (2019)
2. Bermingham, A., Smeaton, A.: On using Twitter to monitor political sentiment and predict election results. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), pp. 2–10
3. Das, B.C., Anwar, M.M., Sarker, I.H.: Reducing social media users' Biases to predict the outcome of Australian federal election 2019. In: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1–6. IEEE (2020)
4. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. **79**(8), 2554–2558 (1982)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling (2014). arXiv:1412.3555
7. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014). arXiv:1409.0473
8. Sudha, S., Arun Kumar, A., Muthu Nagappan, M., Suresh, R.: Classification and recommendation of competitive programming problems using cnn. In: International Conference on Intelligent Information Technologies, pp. 262–272 (2017). Springer, Singapore
9. Looi, W.: Analysis of code submissions in competitive programming contests. http://cs229.stanford.edu/proj2018/report/100.pdf
10. Alnahhas, A., Mourtada, N.: Predicting the performance of contestants in competitive programming using machine learning techniques. Olymp. Inform. **14**, 3–20 (2020). https://ioinformatics.org/journal/v14_2020/_20.pdf
11. Intisar, C.M., Watanobe, Y.: Classification of online judge programmers based on rule extraction from self organizing feature map. In: 2018 9th International Conference on Awareness Science and Technology (iCAST), pp. 313–318 (2018). IEEE
12. Garcia, J.R., Aguirre, V.E.: The learning curves of competitive programming. In: AIP Conference Proceedings, vol. 1618, No. 1, pp. 934–937 (2014). American Institute of Physics
13. Ishizue, R., Sakamoto, K., Washizaki, H., Fukazawa, Y.: Student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics. Res. Pract. Technol. Enhanc. Learn. **13**(1), 1–20 (2018). https://doi.org/10.1186/s41039-018-0075-y
14. Waheed, H., Hassan, S.U., Aljohani, N.R., Hardman, J., Alelyani, S., Nawaz, R.: Predicting academic performance of students from VLE big data using deep learning models. Comput. Hum. Behav. **104**, 106189 (2020)
15. Babić, I.: Machine learning methods in predicting the student academic motivation. Croat. Oper. Res. Rev. 443–461 (2017)
16. Xu, J., Moon, K.H., Van Der Schaar, M.: A machine learning approach for tracking and predicting student performance in degree programs. IEEE J. Sel. Top. Signal Process. **11**(5), 742–753 (2017)
17. Amra, I.A.A., Maghari, A.Y.: Students performance prediction using KNN and Naïve Bayesian. In: 2017 8th International Conference on Information Technology (ICIT), pp. 909–913 (2017). IEEE
18. Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., Radi, N.: Machine learning approaches to predict learning outcomes in Massive open online courses. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 713–720. IEEE (2017)
19. Zulfiker, M.S., Kabir, N., Biswas, A.A., Chakraborty, P., Rahman, M.M.: Predicting students' performance of the private universities of Bangladesh using machine learning approaches. Int. J. Adv. Comput. Sci. Appl. 11(3) (2020)
20. Ofori, F., Maina, E., Gitonga, R.: Using machine learning algorithms to predict students' performance and improve learning outcome: a literature based review. J. Inf. Technol. **4**(1) (2020)
21. Biswas, A.A., Basak, S.: Forecasting the trends and patterns of crime in Bangladesh using machine learning model. In: 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), pp. 114–118. IEEE (2019)



Fuzzy Kernel Weighted Random Projection Ensemble Clustering For High Dimensional Data

Ines Lahmar[1(✉)], Aida Zaier[2], Mohamed Yahia[3], and Ridha Boaullegue[2]

[1] MACS Laboratory, University of Gabes, Gabes, Tunisia
ines12lahmar@gmail.com
[2] Innov'Com Lab, University of Carthage Tunis, 1002 Tunis, Tunisia
[3] SYSCOM Laboratory ENIT, University Tunis El Manar, 1002 Tunis, Tunisia
ridha.bouallegue@ieee.org

Abstract. A clustering ensemble seeks to treat a consensus function by taking multiple base clustering. There are generally two main limitations: (1) High-dimensional data pose a challenge to current ensemble clustering techniques. (2) Generate all the fuzzy base clusterings regardless of their uncertainty and reliability, which makes them susceptible to low-quality. To address this issue, we developed a multi-kernel local weighted fuzzy random projection ensemble clustering. In particular, we develop a fuzzy cluster ensemble method based on a hybrid random projection with multiple KFCM, which can cope with high-dimensional data while guaranteeing the diversity of base clusterings. Based on the fuzzy similarity matrix derived from these local weight kernel fuzzy membership and random projection, an ensemble of diversified base clusterings can be constructed. Then, a fuzzy entropy-based on cluster reliability is used to estimate the fuzzy cluster-wise weighted diversity. Finally The fuzzy similarity matrix corresponding to each base clustering is weighted twice by the block coordinate descent method to find the best clustering result. Three types of consensus clustering are proposed. The experiment results on high-dimensional data demonstrate the efficiency of our method compared to state-of-the-art methods.

Keywords: Fuzzy ensemble clustering · High-dimensional data · Random projection · Kernel Fuzzy C-Means · Entropy

1 Introduction

Clustering is an unsupervised learning method. It can find hidden patterns and structures embedded in unlabeled data in the form of clusters. With the increasing advancement of data streaming from various data sources, we have witnessed the boosting growth of high-dimensional. It has brought a challenge to us. This can be realized by its reduction to a manageable data volume without significant loss of information [1]. This paper is about clustering methods that can be used with high-dimensional data.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 177–187, 2023.
https://doi.org/10.1007/978-3-031-27409-1_16

Ensemble clustering has gained more attention recently. It has emerged to aggregate multiple base clusterings in order to generate a stable consensus clustering, get the usual clusters and handle noise. Ensemble clustering has two phases: clustering generation and consensus clustering. An intuitive algorithm to treat low-quality base clusterings Some approaches have been made to validate each ensemble member and give it a weight to improve the consensus function [2]. Designed a clustering ensemble based on the assumption that all the clusters in the same base clustering have the same reliability [3]. Huang et al. proposed the local diversity strategy of clusters inside a base clustering [4]. Kernel-based methods, which depict various views as kernel matrices, integrate all the views using a weighted sum of the matrices [5].

High-dimensional data needs to be partitioned, which poses a new challenge to the fuzzy clustering ensemble. Many researchers have been presented for ensemble clustering of high-dimensional data, which made some feature models such as the random subspace [6], stratified feature sampling [7], and random projection [8]. For instance, a fuzzy cluster ensemble based on random projection and cumulative aggregation was introduced in [8]. Applied kernel to obtain more expressive features implicitly [9].

In this paper, we propose a new fuzzy ensemble clustering based on hybrid random projection with a KFCM strategy to aggregate fuzzy membership matrices for high-dimensional data. Moreover, to feat the cluster-wise weighted diversity in the multiple fuzzy base clusterings, a fuzzy entropy-based cluster validity measure is presented to local weight the cluster. On the basis of clustering uncertainty, the reliability of clustering is measured by a fuzzy ensemble-driven clustering index (FECI). Then, to get the consensus clustering, three types of consensus are presented. Multiple experiments are conducted on a high dimensional datasets, and the results illustrate the efficiency of of the proposed ensemble clustering method compared with the state-of-the-art methods.

The remainder of this paper is organized as follows. Related works are reviewed in Sect. 2. The proposed method is described in Sect. 3. The experimental results and comparisons are reported in Sect. 4. Finally, this paper is concluded in Sect. 5.

2 Proposed Approach

This part will introduce the prediction framework of multi-kernel local weighted fuzzy random projection ensemble clustering, named KFPEC. It generates fuzzy base clustering into a better consensus clustering to achieve the final clustering result. Brief overview is presented in Sect. 2.1. The Kernel Fuzzy C-means (KFCM) process is introduced in Sect. 2.2. The generation-based clustering is presented in Sect. 2.3. Finally, the consensus clustering is given in Sect. 2.4.

2.1 Brief Overview

In this paper, we propose a novel kernel weighted fuzzy random projection ensemble clustering framework. First, we create a fuzzy kernel using the fuzzy

similarity kernel and hybridize the kernel function with the random projection. Second, with each random kernel-projection pair, we build a fuzzy kernel similarity matrix for the data object. The multiple FCM is then integrated onto these similarity matrices to find an ensemble of base clusterings. Third, a fuzzy entropy-based measure on cluster reliability is used to estimate the fuzzy cluster-wise weighted diversity in fuzzy ensembles of multiple base clusterings. With the weighted vectors, we create the similarity matrix via a local weighting strategy regarding the reliability of each fuzzy cluster. Finally, three types of consensus clustering are proposed based on hierarchical clustering, fuzzy clustering, and bipartite graph strategy.

2.2 Kernel Fuzzy C-Means (KFCM)

A KFCM was presented to overcome noise sensitivity in FCM. The main idea was based on transforming the original space into an even-dimensional kernel space via a non-linear mapping (φ) so that the samples are linearly separable in the feature space. To resolve the problem of non-linear mapping for high-dimensional feature kernels, we use kernel functions. Consequently, it replaces Euclidean distance in fuzzy clustering with a kernel function, which is defined as the following:

Let $X = \{x1, x2, ..., xn\}$, which is classified into clusters C: $\{C1, C2, ..., Cc\}$. A fuzzy partition matrix $U = (U_{ij})$ characterizes fuzzy clustering, where U_{ij} is the degree of membership of x_j to its corresponding cluster in the range 0 to 1.

The objective function of FCM is assumed to be:

$$J_{min} = \sum_{i=1}^{c} \sum_{j=1}^{n} U_{ij}^{m} ||c_i - x_j||^2 \tag{1}$$

where J denotes the objective function, m denotes the fuzzy factor, c denotes the number of clusters, and n denotes the number of sample; U_{ij} is the membership degree of sample j to cluster i; c_i is the center of cluster i; x_j is the raw data of sample j.

The main idea of KFCM is to minimize the objective function that is converted into:

$$J_{min} = \sum_{i=1}^{c} \sum_{j=1}^{n} U_{ij}^{m} \left\| (\varphi(c_i), \varphi(x_j)) \right\|^2 \tag{2}$$

According to the conversion methodologies in kernel models, were have:

$$\left\| c_i - x_j \right\|^2 = (\varphi(c_i) - \varphi(x_j))^T (\varphi(c_i) - \varphi(x_j)) = \kappa(x_j, x_j) + \kappa(c_i, c_i) - 2\kappa(c_i, c_j) \tag{3}$$

In the case of the Gaussian kernel, according to $\kappa(x, x) = 1$, the objective function can be converted into:

$$J_{min} = 2 \sum_{i=1}^{c} \sum_{j=1}^{n} U_{ij}^{m} (1 - \kappa(c_i, x_j)) \tag{4}$$

Hybridizing with the fuzzy constraints and using the Lagrange multiplier approach to minimize the objective function, the clustering center c_i and membership matrix U_{ij} is updated by Eq.(3), which is defined as follows:

$$\kappa(x, x) = \exp^{-||x-y||^2/\sigma^2} \tag{5}$$

The following Eq. (7) can be defined.

$$U_{ij} = \frac{(1 - \kappa(c_i, x_j))^{1/(1-m)}}{\sum_{k=1}^{c}(1 - \kappa(c_i, x_j))^{1/(1-m)}} \tag{6}$$

2.3 Ensemble Generation Process

In ensemble generation, we hybridize multiple KFCM with random projection iterations that can be used to create a set of the fuzzy base clusters with local diversity. Then, we run the KFPEC M times to get fuzzy base clustering. To begin, a hybrid of KFCM and random projection is used to generate rich information in different subs-paces. As a result, given a feature in one of the datasets, $F = \{f1, , fn\}$, where f_i is the i-th feature. Specifically, we perform random projection M times to find M sub-spaces, denoted as F1. Let p be an n-dimensional random vector in which each attribute is randomly sampled from the normal distribution $\tau \in (0, 1)$. Obviously, if $\tau = 1$, that means every subspace is in fact the original feature vector with a given M kernels. Then, we construct M fuzzy local weight kernel similarity matrices and apply multiple KFCM to cluster them into p clusters. Thus, we obtain a fuzzy similarity matrix for every feature subset, which is formulated as:

$$\varsigma_{ij}^{m} = \sum_{d=1}^{M}(\pi^i(\boldsymbol{x_d}) * \pi^j(\boldsymbol{x_d})) \tag{7}$$

where $\pi(\boldsymbol{x_d})$ denotes the degree of membership data-object in each cluster $\in X$ that belongs to the corresponding fuzzy cluster. A similarity kernel measure s is defined as

$$s_{ij} = \exp \frac{-||x - y||^2}{2\sigma^2} \tag{8}$$

Thus, we can find M fuzzy kernel similarity matrix in terms of M projection as follows:

$$S_{ij}^{m} = \left[s_{ij}\varsigma_{ij}^{m}\right]_{ij} \tag{9}$$

Having obtained the M similarity matrix, we then exploit the KFCM to build the ensemble of base clusterings. Finally, based on the M kernel fuzzy similarity matrix, we can construct an ensemble of M base clusterings, defined as

$$\mathit{\Pi} = \left\{\pi^1, \pi^2, \pi^3, ..., \pi^M\right\} \tag{10}$$

where π^m is the m-th fuzzy base clustering in $\mathit{\Pi}$

2.4 Consensus Clustering

After getting M base clusterings, we need to explore the reliability and cluster-wise diversity of the ensembles and incorporate three types of consensus clusterings into the final clustering result.

Each cluster is a set of data objects. To estimate the reliability of each cluster, the cluster uncertainty estimation method based on fuzzy entropy measure calculates the uncertainty of the cluster by taking the cluster labels in the whole set [5]. The entropy of a fuzzy cluster $c_i \in c$ and a fuzzy base clustering $\pi^m \in \Pi$ can be calculated as

$$R(c_i, \pi^m) = -\sum_{j=1}^{h_m} p(c_i, c_j^m) \log_2 p(c_i, c_j^m) \tag{11}$$

With

$$p(c_i, c_j^m) = \frac{(c_i, c_j^m)}{\sum_{j=1}^{h_m} p(c_i, c_j^m)} \tag{12}$$

where h_m denotes the number of bases clusters. c_j^m is the jth cluster. As it is holds that $p(c_i, c_j^m)$ in the interval [0, 1] for any i, j, and m, which leads to $R(c_i, \pi^m) \in [0, +)$. When all the data objects in c_i belong to the same fuzzy cluster in π^m, the uncertainty of c_i with respect to π^m is zero.

Given the fuzzy base clustering, we can calculate the entropy of c_i and π^m for the entire ensemble Π as follows:

$$R^{\pi}(c_i) = \sum_{m=1}^{M} R(c_i, \pi^m) \tag{13}$$

After calculating the fuzzy entropy of each cluster, we consider the uncertainty of the cluster relative to the set through the notion of FECI and add weights to the data items within each cluster.

Given a cluster set of M fuzzy base clusters, the FECI of each cluster is computed as follows:

$$FECI(c_i) = \exp^{-\frac{R^{\Pi}(c_i)}{\theta . M}} \tag{14}$$

where $\theta > 0$ is the parameter to adjust the fuzzy cluster unreliability over the index.

The FECI metric is regarded as a reliability index for various fuzzy clusters in the ensemble. By using FECI as a cluster weighting strategy, we refine the similarity matrix by a locally weighted which is computed as follows:

$$B = \left\{B_{ij}\right\}_{N \times N} \tag{15}$$

$$B_{ij} = \frac{1}{M} \sum_{m=1}^{M} w_i^m S_{ij}^m \tag{16}$$

$$w_i^m = FECI(cl^m(x_i)) \tag{17}$$

where $cl^m(x_i)$ is the cluster in $\pi^m \in$ that object oi belongs to. Having generated cluster-wise diversity, we further create three types of consensus clustering to obtain final clustering, called HKFPEC, FKFPEC, and GBKFPEC.

In HKFPEC, a hierarchical agglomerative-based consensus clustering is presented in an iterative region merging iteratively to achieve a dendrogram uses the locally weighted similarity matrix as the initial regions set and the similarity matrix, defined as $Z^{(0)} = B$. The region merging is then developed iteratively. In each step, the two regions with the highest similarity are merged into a new and larger region.

Given $R^{(t)} = \{R_1^{(t)}, ..., R_{|R^{(t)}|}^{(t)}\}$ is the set of regions after the t-th step, whose fuzzy similarity matrix (see Eqs. (15) (16) and (17)) can be updated based on the average-link after region merging, resulting in:

$$Z^{(t)} = \left\{ Z_{ij}^{(t)} \right\}_{\left|R^{(t)}\right| \times \left|R^{(t)}\right|} \tag{18}$$

$$Z_{ij}^{(t)} = \frac{1}{\left|R^{(t)}\right| \left|R^{(t)}\right|} \sum_{x_u \in R_i^{(t)}, x_v \in R_j^{(t)}} a_{uv} \tag{19}$$

where $|R^{(t)}|$ is the number of regions and $R_i^{(t)}$ is the number of data samples.

In each iteration, the number of regions decreases by one. N denotes the number of the initial regions; it is obvious that all data objects will be combined into a root region and a dendrogram will be created after exactly $N-1$ iterations. Each level of the dendrogram represents a clustering result with a certain number of clusters and a certain number of clusters can be generated.

In FKFPEC, a fuzzy clustering method based on consensus clustering is presented. The optimal fuzzified membership matrix with a fuzzifier exponent m and the centers ς_0 can be presented to minimize the objective function $\varphi(\varsigma, \mu)$. The fuzzy consensus clustering is no longer represented by an integer vector π. We created a $n \times k$ membership matrix called μ. Thus, π_i should be replaced by a membership matrix of $n \times k_i$, where k_i is the number of clusters. The fuzzy consensus partition can be defined as:

$$\min_{(,\mu=} \sum_{i=1}^{l} w_i \varphi(\varsigma, \mu) \tag{20}$$

$$\varphi(\varsigma, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{k} u^m ij \left\| P_i, v_i \right\|^2 \tag{21}$$

where u is the degree of membership, $m \in (1, +\infty)$ is a user specified fuzzifier factor, and $w = (w_1, ..., w_l)$ is the vector of user-specified weights, with $||w||_1 = 1$. The v_i represents a set of unknown centers, and the P_i represents the sample.

In GBKFPEC, a fuzzy bipartite graph-based consensus clustering is presented. To construct a bipartite graph with both fuzzy clusters and data objects

treated as graph nodes. Then perform bipartite graph partitioning to find the clustering result. That is

$$\hat{G} = (U, V, \hat{E}) \tag{22}$$

where U, V denotes the node set and $\hat{E}$ denotes the edge set. A link between two nodes exists if and only if one of them is a data object and the other node is the fuzzy cluster that contains it. The link weight between a cluster's reliability and a membership degree is decided by the similarity between them. Let two nodes $u_i \in X$ and $v_j \in C$ have a link weight between them decided by two factors, i.e., their belonging-to relationship and the reliability of the connected cluster, which can be defined by the FECI metric.

$$\hat{E}_{ij} = \begin{cases} FECI(v_i) * F(x_j) & \text{if } u_i \in v_i \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

Where the FECI reflects the reliability of a fuzzy cluster, the entire ensemble of base clusterings. $F(x_j)$ is the membership degree of a data. Then, with the bipartite graph constructed, we proceed to partition the graph using the cut, which can efficiently partition the graph nodes into different node sets. The objects in each subset are treated as a cluster, and consensus clustering can be obtained.

3 Experiments

All of the experiments are developed in MATLAB R2017a on a 64-bit Microsoft Windows 10 computer with 8 GB of memory and an Intel Core i5-2410M CPU at 2.30 GHz processing speed. In our simulations, we compare the proposed methods with other methods, i.e., reliability-based graph partitioning fuzzy clustering ensemble (RGPFCE) [3], locally weighted ensemble clustering (LWEA, LWGP) [4], fuzzy consensus clustering (FCC) [5], probability trajectory based graph partitioning (PTGP) [10], K-means-based consensus clustering (KCC) [11], and entropy consensus clustering (ECC) [12].

Presentation results for the proposed algorithm of both measures NMI and ARI are run over 20 clustering iterations to investigate the effects of parameters. We choose the number of random projections to be set to 30. The weighting exponent m is 2. To produce the fuzzy base clusterings, the ensemble size M is set to 30. In each base clustering, the number of clusters is randomly selected in the range of $[2, \sqrt{N}]$.

3.1 Synthetic Datasets

In our experiments, 12 real-world datasets are used, namely, Multiple Features (MF), Image Segmentation (IS), MNIST, Optical Digit Recognition (ODR), Landsat Satellite (LS), UMist, USPS, Forest Covertype (FC), Texture, ISOLET, Breast Cancer (BC), and Flowers17 (as shown in Table 1).

Table 1. Datasets description

Datasets	Instances	Attributes	Classes	Source
MF	2.000	649	10	[13]
IS	2.310	19	7	[13]
MNIST	5.000	784	10	[14]
ODR	5.620	64	10	[13]
LS	6.435	36	6	[13]
UMist	575	10.304	20	[15]
USPS	11.000	2568	10	[14]
FC	11.340	54	10	[13]
Texture	5.500	40	11	[13]
ISOLET	7.797	617	26	[13]
BC	9	386	2	[13]
Flowers17	1.360	30.000	17	[16]

3.2 Evaluation Metrics

To access the quality of the clustering result, we used two evaluation metrics: normalized mutual information (NMI) [17] and adjusted Rand index (ARI) [18]. The NMI and ARI values are in the range of $[0, 1]$ and $[1, 1]$, respectively.

3.3 Comparison With The State of-Art Methods

The comparison results of the NMI score (see Table 2) indicate that our three proposed methods outperform both the other methods. The LWGP achieves the best NMI on 2 out of the 11 datasets, but the three proposed methods outperform the LWGP on most of the other datasets. In addition, they ranked in the top three with 19, 21, and 23 comparisons, respectively, while the baseline best method ranked in the top three with only 4 comparisons out of the total of given datasets. It can be concluded that the proposed methods achieve the overall best NMI values. For ARI scores (see Table 3), our proposed methods rank in the top three positions (17, 22, and 23 comparisons, respectively), while the baseline best method only ranks in the top three positions (3 comparisons).

To provide a summary view statistic for all datasets, we report the average score of various methods (see Tables 2 and 3). The average score is calculated by averaging the NMI (or ARI) values. As can be seen, our proposed methods achieve the highest average NMI of 68.2, 66.5, and 66.9, respectively, which is better than the fourth best average score of 58 (see Table 2). Similar advantages can be noted in terms of the average score of the proposed three methods with respect to ARI (see Table 3).

3.4 Execution Time

In this section, we evaluate the execution times of various ensemble clustering techniques. The time is calculated by the average output over 20 runs. Furthermore, larger sample sizes and larger dimensions lead to greater computational costs for the clustering approaches. As demonstrated in Table 4, the proposed three methods indicate comparable time efficiency with the other ensemble clustering methods.

Table 2. Average performance in terms of NMI of multiple approaches

Data set	RGPFCE	KCC	FCC	PTGP	ECC	LWEA	LWGP	HKFPEC	FKFPEC	BGKFPEC
MF	52.8	40.2	51.3	61.3	75.2	65.9	68.2	85.6	86.5	86.6
IS	27.2	39.5	40	61.1	61.1	62.1	62.9	63.7	63.2	69.2
MNIST	58.6	33.3	49.9	57.6	50	64.6	63.5	74.3	75.1	80.7
ODR	55.2	52.5	59.2	81.3	61.2	82.9	81.6	90.7	82.2	82.9
LS	48.9	30.4	45.6	62.5	39.2	61.6	64.4	65.7	68.4	77.2
UMist	63.9	60.8	61.1	62.6	61.3	62.9	62.5	79	80.2	77.8
USPS	61.8	27.8	30.2	56.5	52.7	63.3	61.4	77.6	74.9	75.4
FC	16	8.4	8.4	23.2	10.2	12.9	11.7	15.6	17.4	17.7
Texture	59.1	40.1	43.5	74.9	54	68.9	62.8	74.9	75.1	75.6
ISOLET	55.1	42	50.2	54.1	70	55.5	51.8	67.6	66.4	66.4
BC	71	76.5	68.2	76	79	65.5	66.2	81.2	80.3	79.4
Flowers17	22.5	24.9	24.7	24.9	24.1	21.8	21.6	27.5	28.9	29.5
average	49.34	39.7	44.35	58	47.92	57.32	56.55	66.95	66.55	68.2

Table 3. Average performance in terms of ARI of multiple approaches

Data set	RGPFCE	KCC	FCC	PTGP	ECC	LWEA	LWGP	HKFPEC	FKFPEC	BGKFPEC
MF	86.1	73	88.5	85.6	87.8	52.5	56.2	90.6	91.5	91
IS	72.9	59.5	51.2	62.9	50.6	52.2	52.9	83.1	81.7	89.5
MNIST	68.1	53.4	54.2	48.5	40.24	55	51.2	88.5	88.6	89
ODR	79.9	52.5	70	80.9	66.7	78.2	76.3	95.3	95.2	95.4
LS	62.6	48.8	54.7	52.6	44.2	56.8	58	80.1	82.5	82.7
UMist	63.1	60.8	64.2	33.4	31.2	56.8	58	72.4	71.3	72.7
USPS	63.9	51.1	55.2	43.9	45	63.3	61.4	86.3	88.3	87.3
FC	60	58.4	55.7	20	15.7	23.1	20.0	75.5	75.7	77.6
Texture	83.9	40.8	54.9	81.9	56.9	78.8	74.3	89.3	87.2	87.4
ISOLET	74.8	68.4	65.7	54.1	66.9	74.5	74.3	84.8	84.9	84.7
BC	88.1	76.1	89.6	85.7	87.6	85.7	86.2	94.4	94.5	94.4
Flowers17	19.2	24.1	15.7	9.2	9.7	20	19.5	27.9	33.8	35.5
average	68.55	55.57	59.96	54.89	53.89	58.07	57.35	80.68	81.26	82.26

Table 4. The execution times values of different clustering ensemble in seconds

Data set	RGPFCE	KCC	FCC	PTGP	ECC	LWEA	LWGP	HKFPEC	FKFPEC	BGKFPEC
MF	6.2	6.8	5.3	7.66	75.2	9.37	8.2	5.6	6.5	4.6
IS	27.2	39.5	40	61.1	61.1	62.1	62.9	63.7	63.2	69.2
MNIST	58.6	33.3	49.9	57.6	50	64.6	63.5	74.3	75.1	80.7
ODR	55.2	52.5	59.2	81.3	61.2	82.9	81.6	90.7	82.2	82.9
LS	48.9	30.4	45.6	62.5	39.2	61.6	64.4	33.7	31.4	31.2
UMist	113.9	115.8	99	87	86.3	101.2	105	79	78	77.8
USPS	6.8	7.2	3.9	5.5	7.7	8	8.8	7.6	8	7.1
FC	16	8.4	8.4	23.2	10.2	12.9	11.7	15.6	17.4	17.7
Texture	20	24.1	24.8	20.9	24	19.8	20.7	19.9	20.1	20
ISOLET	55.9	61.71	50.60	87.18	156.6	55.5	59.94	77.6	66.4	66
BC	71	76.5	68.2	76	79	65.5	66.2	81.2	80.3	79.4
Flowers17	222.5	204.9	200	204.9	206.4	206.9	21.6	177.9	189	188

4 Conclusion

In this paper, we present a model named multi-fuzzy kernel random projection ensemble clustering, which is capable of combining KFCM, random projection, and local weighted clusters. With the base clusterings defined, a fuzzy entropy-based metric is utilized to evaluate and weight the clusters with consideration to the distribution of the cluster labels in the entire ensemble. Finally, based on fuzzy kernel random projection, three ensemble clusterings are presented by incorporating three types of consensus results. The experiment results are validated on high-dimensional datasets, which have demonstrated the advantages of the proposed methods over other methods. Exploiting optimization in ensemble clustering should be an interesting future development.

References

1. Yang, M.S., Nataliani, Y.: A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy. IEEE Trans. Fuzzy Syst. **26**(2), 817–835 (2017)
2. Ilc, N.: Weighted cluster ensemble based on partition relevance analysis with reduction step. IEEE Access **8**, 113720–113736 (2020)
3. Bagherinia, A., Minaei-Bidgoli, B., Hosseinzadeh, M., Parvin, H.: Reliability-based fuzzy clustering ensemble. Fuzzy Sets Syst. **413**, 1–28 (2021)
4. Huang, D., Wang, C.D., Lai, J.H.: Locally weighted ensemble clustering. IEEE Trans. Cybern. **48**(5), 1460–1473 (2017)
5. Zhao, Y.P., Chen, L., Gan, M., Chen, C.P.: Multiple kernel fuzzy clustering with unsupervised random forests kernel and matrix-induced regularization. IEEE Access **7**, 3967–3979 (2018)
6. Gu, J., Jiao, L., Liu, F., Yang, S., Wang, R., Chen, P., Zhang, Y.: Random subspace based ensemble sparse representation. Pattern Recognit. **74**, 544–555 (2018)
7. Tian, J., Ren, Y., Cheng, X.: Stratified feature sampling for semi-supervised ensemble clustering. IEEE Access **7**, 128669–128675 (2019)
8. Rathore, P., Bezdek, J.C., Erfani, S.M., Rajasegarar, S., Palaniswami, M.: Ensemble fuzzy clustering using cumulative aggregation on random projections. IEEE Trans. Fuzzy Syst. **26**(3), 1510–1524 (2017)

9. Zeng, S., Wang, Z., Huang, R., Chen, L., Feng, D.: A study on multi-kernel intuitionistic fuzzy C-means clustering with multiple attributes. Neurocomputing **335**, 59–71 (2019)
10. Huang, D., Lai, J.H., Wang, C.D.: Robust ensemble clustering using probability trajectories. IEEE Trans. Knowl. Data Eng. **28**(5), 1312–1326 (2015)
11. Wu, J., Liu, H., Xiong, H., Cao, J., Chen, J.: K-means-based consensus clustering: a unified view. IEEE Trans. Knowl. Data Eng. **27**(1), 155–169 (2014)
12. Liu, H., Zhao, R., Fang, H., Cheng, F., Fu, Y., Liu, Y.Y.: Entropy-based consensus clustering for patient stratification. Bioinformatics **33**(17), 2691–2698 (2017)
13. Bache, K., Lichman, M.: UCI machine learning repository (2013)
14. Roweis, S.: http://www.cs.nyu.edu/
15. Graham, D.B., Allinson, N.M.: Characterising virtual eigensignatures for general purpose face recognition. In: Face Recognition, pp. 446–456. Springer, Berlin (1998)
16. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognit. (CVPR'06), vol. 2, pp. 1447–1454. IEEE (2006)
17. Strehl, A., Ghosh, J.: Cluster ensembles: a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
18. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)

A Novel Lightweight Lung Cancer Classifier Through Hybridization of DNN and Comparative Feature Optimizer

Sandeep Trivedi[1](✉), Nikhil Patel[2], and Nuruzzaman Faruqui[3]

[1] Deloitte Consulting LLP Texas, Houston, USA
sandeep.trived.ieee@gmail.com
[2] University of Dubuque, Iowa, USA
[3] Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

Abstract. The likelihood of successful early cancer nodule detection rises from 68% to 82% when a second radiologist aids in diagnosing lung cancer. Lung cancer nodules can be accurately classified by automatic diagnosis methods based on Convolutional Neural Networks (CNNs). However, complex calculations and high processing costs have emerged as significant obstacles to the smooth transfer of technology into commercially available products. This research presents the design, implementation, and evaluation of a unique lightweight deep learning-based hybrid classifier that obtains 97.09% accuracy while using an optimal architecture of four hidden layers and fifteen neurons. This classifier is straightforward, uses a novel self-comparative feature optimizer, and requires minimal computing resources, all of which open the way for creating a marketable solution to aid radiologists in diagnosing lung cancer.

Keywords: Lung Cancer · Deep Neural Network · Hybridization · Network Optimization · Feature Optimization

1 Introduction

Cancer develops when average cellular growth is disrupted due to mutations or aberrant gene alterations that usually do so [1, 2]. Since 2000, the number of people losing their lives to cancer has risen from 6.2 million to an estimated 10 million deaths annually by 2020 [3]. Lung tumor remains the leading cause of tumor death rate, with 1.80 million deaths (18%), and the global tumor burden is expected as around 28.40 million cases in 2040 [4]. The survival percentage for people with lung cancer can be increased to 90% by early identification [5]. X-ray, MRI, and CT scans diagnose lung cancer [6]. Radiologists must identify suspicious lung nodules to make radiography screening successful. This is especially important for tiny lung nodules. Literature shows that a single radiologist can properly diagnose 68% of lung nodules, and a second radiologist can enhance this to 82% [7]. This paper proposes a novel lightweight lung cancer classifier through hybridizing deep neural networks and comparative classifiers to assist radiologists in diagnosing lung cancer nodules more accurately.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 188–197, 2023.
https://doi.org/10.1007/978-3-031-27409-1_17

Convolutional Neural Networks (CNN), the state-of-art technology to automate lung cancer diagnosis from CT images, are computationally expensive [8]. Every new diagnosis helps machine learning models to be better at diagnosis. However, it is time-consuming and expensive to retrain a CNN every time new training data become available. A centralized server-based online learning approach is an efficient solution to this problem. Still, it imposes challenges on cloud computing resources. It demonstrates the necessity of a lightweight yet accurate lung cancer classifier proposed in this paper. In addition, technology acceptance is always challenging, which raises questions about the overall integrity and reliability of Computer Aided Diagnosis (CAD) systems. An innovative approach of self-comparative classifier has been developed, experimented with, hybridized with a Deep Neural Network (DNN), and presented in this paper.

This experiment aims to design a lightweight lung cancer classifier to assist radiologists in lung cancer diagnosis with reliable prediction through self-comparative classifiers. The core contributions of this paper are as follows:

- Lightweight hybrid lung cancer classifier with optimized network depth which classifies with 97.09% accuracy.
- The application of an innovative and effective self-comparative algorithm to identify the most relevant features.
- Exploration of the genetics algorithm-based feature optimization techniques in hybrid classifiers.

The rest of the paper has been organized into four different sections. The second section highlights recent research on lung cancer diagnosis using CAD systems and compares them with the proposed methodology. The third section explains the proposed methodology. The experimental results and performance evaluation have been analyzed and presented in the fourth section. Finally, the fifth section concludes the paper with a discussion.

2 Literature Review

A hybrid deep-CNN model named LungNet classifies lung cancer nodules into five classes with 96.81% accuracy. It has remarkable results from a research outcome perspective. However, the dependency on intensive image processing makes it computationally expensive [9]. The proposed methodology classifies malignant and benign classes with 97.09% accuracy. This approach is much less computationally expensive and requires minimal resources while retraining on new datasets.

A recent study demonstrates KNN-SVM hybridization, which achieved 97.6% accuracy. Although it achieved little better accuracy than the proposed methodology, the application of Grey Wolf Optimized (GWO), Whale Optimization Algorithm-Support Vector Machine (WOA-SVM), Advanced Clustering (AC), Advanced Surface Normal Overlap (ASNO) lung segmentation algorithm combined with another KNN-SVM hybrid classifier raise the question at the optimization complexity of the approach [10]. Each of these algorithms needs to be optimized. That means this approach generates acceptable

accuracy only under certain conditions that lack the generalizability of machine learning approaches. The proposed classifier gains similar performance with a much simpler architecture with better generalization.

Another KNN-SVM-Decision Tree hybrid classifier framework demonstrates promising performance. However, this method uses abstracted features from multiple sub-areas from enhanced images [11]. Sub-segmentation before feature extraction weakens the global correlation among features. Moreover, abstracted features question at the overall integrity of the methodology. The proposed methodology uses SURF features followed by Genetics Algorithm (GA) based optimizer to use actual but optimized features.

A contour-based Fuzzy c means centric hybrid method followed by a CNN shows 96.67% accuracy. After CT image binarization enriches the distinguishable features the CNN receives, the second-order statistical texture analysis method used in this paper. As a result, the approach achieved an accuracy of 96.67% [12]. The proposed methodology follows similar feature enhancement techniques. However, because of well-optimized DNN architecture, it gives better performance. A hybrid classifier by Ananya Bhattacharjee et al. achieves 92.14% accuracy [13], the CNN with residual connection and hybrid attention mechanism-based research conducted by Yanru Guo et al. shows 77.82% accuracy [14], and CNN applied on Thoracic radiography (chest X-ray) to detect lung cancer achieves 90% accuracy with AUC of 0.811 in M. Praveena et al. [15]. The accuracy of the proposed methodology outperforms other hybrid classifiers.

3 Methodology

The proposed methodology illustrated in Fig. 1 consists of four major parts—dataset and preprocessing, feature extraction, comparative feature optimizer, and network architecture and optimization.

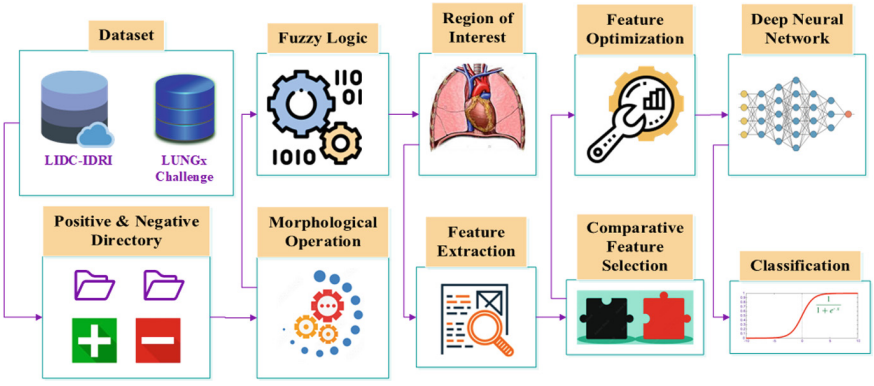


Fig. 1. Overview of the proposed methodology

The simple and lightweight design of the classifier, the application of an innovative comparative classifier, and an accuracy of 97.09% is the novel contributions of this methodology.

3.1 Dataset and Preprocessing

LIDC-IDRI and LUNGx datasets have been used in this paper. The LIDC-IDRI dataset contains more than 1000 cases with more than 244,000 CT scans. Four different experienced radiologists annotate the lung nodules of this dataset. They scale the degree of malignancy from 1 to 5 [16]. The LUNGx challenge dataset contains more than 22,000 CT images. The nodule location of this dataset is documented on CSV files available with the dataset [17]. The training images are stored in two directories representing positive and negative classes. The lung region is segmented by morphological operation assisted by Fuzzy Logic followed by Region of Interest (ROI) extraction [18].

3.2 Feature Extraction

The speeded-up robust features (SURF) have been used in this paper to extract the feature using a square filter with 9×9 pixels defined by Eq. 1.

$$I_f(x, y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I_s(i, j) \tag{1}$$

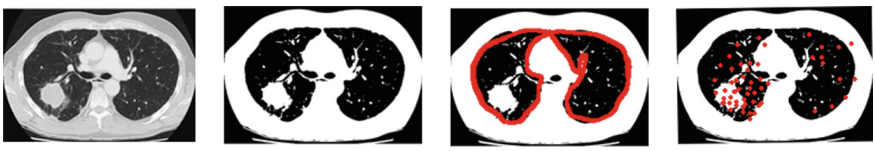
The point of interest is detected by the Hessian Matrix [20] defined by Eq. 2.

$$H(\rho, \sigma) = \begin{pmatrix} L_{xx}(\rho, \sigma)\ L_{xy}(\rho, \sigma) \\ L_{yx}(\rho, \sigma)\ L_{yy}(\rho, \sigma) \end{pmatrix} \tag{2}$$

Here in the above equation, the $L_{xx}(\rho, \sigma)$ represents the convolution of the second-order derivation of gaussian with the $I_s(i, j)$ image at point ρ. The location of the point of interest is calculated using Eq. 3.

$$\sigma_{approx} = c_f(x, y) \times \frac{f_{scale}}{f_{size}} \tag{3}$$

Here $c_f(x, y)$ is the current filter size, the f_{scale} is the scale of the base filter, and the f_{size} is the size of the base filter. An example of the original image, masked image, Region of Interest (ROI), and SURF features have been illustrated in Fig. 2.



(a) Original Image (b) Masked Image (c) ROI Image (d) SURF Features

Fig. 2. The images before and after processing with the surfing feature extracted from the ROI

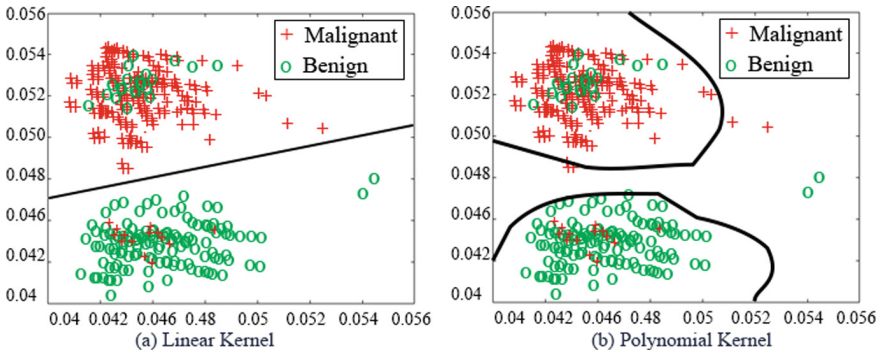


Fig. 3. Feature selection using linear and polynomial SVM kernel

3.3 Comparative Feature Optimizer

The comparative optimizer has been designed using Support Vector Machines [21] with two linear and polynomial kernels. The feature classification illustrated in Fig. 3 is compared using algorithm 1.

It selects the kernel that maximizes the distance between malignant and benign classes and passes the selected features to Genetics Algorithm (GA) based feature optimizer. Only the optimized features are used to train the Deep Neural Network (DNN).

Algorithm 1 Feature Selection and optimization

Input: $S_F[i,j] \leftarrow Surf\ Feature\ Extractor\ (I_P(x,y))$

1: $D_L, D_P \leftarrow dist(SVM_L(\mathrm{S[i,j]}),\ SVM_p(\mathrm{S[i,j]}))$

2: $F_{SVM} \leftarrow Max(D_L, D_P)$

3: $F_{Opt} \leftarrow Genetics\ Algorithm(F_{SVM})$

4: $C_o \leftarrow Deep\ Neural\ Network(F_{Opt})$

5: **if** $(p(C_o) > 0.62)$
$class \leftarrow Malignant$
else
$class \leftarrow Benign$

6: **end if**

3.4 Network Architecture and Optimization

A fully connected DNN with 6 hidden layers having 15 neurons in each layer has been designed with three input layers and one output layer. It is defined using Eq. 4.

$$Output = f_{sigmoid}\left(\sum_{i=1}^{4}\sum_{j=1}^{15} f_{sigmoid}((N_{ij} \times W_{ij}) + b_i)\right) \tag{4}$$

$$Here, i = hidden\ layer\ index,$$

$$j = hidden\ layer\ node\ index, N = node, W = weight,$$

$$b = bias, f_{HTST} = \text{sigmoid function}$$

The Levenberg-Marquardt backpropagation algorithm has been used as the learning rule of the proposed network where the weight update is governed by Eq. (5) where W_{j+1} is the updated weight, W_j is the current weight, $J^T J$ is the approximated Hessian matrix, μ transition constant, and e is the error vector.

$$W_{j+1} = W_j - \left[J^T J + \mu I\right]^{-1} J^T e \tag{5}$$

During the learning progress, the Mean Squared Error (MSE) defined by the following equation has been used as the performance measurement criteria where o_i is the actual value and $\widetilde{o_i}$ is the prediction from the network.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - \widetilde{o_i})^2$$

The network has been optimized from the learning curve illustrated in Fig. 4.

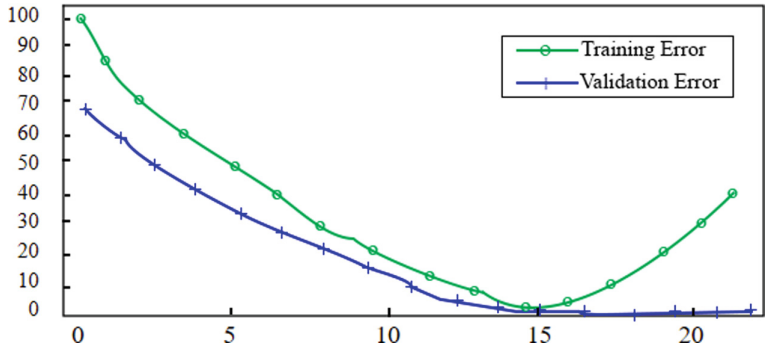


Fig. 4. Network optimization through the learning curve

The learning curve shows the incremental difference between training and validation errors for more than 15 neurons. It indicates that the network performs the best with 15 neurons in each hidden layer.

4 Experimental Results and Evaluation

The proposed classifier is implemented in a desktop computer with Microsoft Windows 10 Operating System, powered by Intel(R) Core (TM) i7-8700 processor, 16GB RAM, and GIGABYTE GeForce GT 730 2GB GDDR5 PCI EXPRESS Graphics Card. The

proposed methodology is coded in MATLAB 2021B. The performance of the proposed classifier has been measured using the evaluation matrices [22] listed in Table 1.

Table 2 shows the classification accuracy of the proposed classifier. It is significantly higher for the same dataset than for the polynomial kernel than the linear kernel.

One of the optimization strategies used in the proposed methodology is a 0.18% dropout rate which has been empirically calculated and later optimized through experimental results and comparison. The results of dropping out neurons at different layers have been listed in Table 3.

It has been observed that adding 0.18% dropout until the third hidden layer improves the network's performance. However, after that, the performance starts degrading. As a result, 0.18% dropout at layers 1, 2, and 3 have been used as the optimized dropout

Table 1. Evaluation Matrices

Evaluation Matrices	**Mathematical Definition**	**Performance Criteria**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Quality of prediction
Recall	$\frac{TP}{TP+FN}$	Correctness of true positive prediction
Specificity	$\frac{TN}{TN+FP}$	Correctness of true negative prediction
Error Rate	$\frac{FP+FN}{TP+FP+FN+TN}$	Incorrect prediction rate

Table 2. The classification accuracy score of both classifiers

Class	Dataset	Accuracy (%)	
		Linear Kernel	Polynomial Kernel
Malignant	LIDC-IDRI	82.04	97.09
Benign	LIDC-IDRI	79.45	96.20
Malignant	LUNGx	73.37	85.97
Benign	LUNGx	74.66	86.72

Table 3. Dropout analysis to optimize network performance

Hidden Layer	Dropout Rate (%)	Accuracy (%)	Recall (%)	Specificity (%)	Error Rate (%)
1	0.18	94.23	93.11	92.95	5.77
2	0.18	95.88	94.31	93.42	4.12
3	0.18	97.09	96.94	97.05	2.95
4	0.18	94.01	90.25	91.62	5.99

rate. The performance of the proposed network has been compared with another similar network and listed in Table 4.

Table 4. Performance comparison of the proposed classifier on the LIDC-IDRI dataset with recently published papers.

Models	Accuracy (%)	Recall (%)	Specificity (%)	AUC (%)
Texture CNN [23]	96.69	96.05	97.37	99.11
LungNet [9]	96.81	97.02	96.65	NA
Wei et al. [24]	87.65	89.30	86.00	94.20
MV-KBC [25]	91.60	86.52	94.00	96.70
Proposed Classifier	97.09	96.94	97.05	97.23

The performance comparison demonstrates that the proposed classifier performs better than other state-of-art hybrid classifiers.

5 Conclusion and Discussion

The experimenting classifier contains only four hidden layers, with 15 neurons in each hidden layer. With a 0.18 dropout rate, only 52 neurons participate in the classification process, which is much simpler than convolutional neural network-based classifiers published in recent literature. Moreover, the proposed classifies malignant and benign classes with 97.09% accuracy, which is higher than similar hybrid classifiers. Instead of using the features extracted by the feature extractor directly, an innovative comparative feature optimizer ensures that the network learns from the most relevant features. Moreover, the relevant features are further optimized before sending them to the deep neural network. The network proposed in this paper is well-optimized through the learning curve and empirical dropout analysis. As a result, even with simple and lightweight architecture, it performs better than similar classifiers. The self-comparative feature optimizer improves the reliability of the classification. The simple design, accurate prediction, and reliable classification make the proposed classifier an excellent fit to assist radiologists in lung cancer diagnosis.

References

1. Williams, R.R., Horm, J.W.: Association of cancer sites with tobacco and alcohol consumption and socioeconomic status of patients: interview study from the Third National Cancer Survey. J. Natl. Cancer Inst. **58**(3), 525–547 (1977)
2. Ravdin, P.M., Siminoff, I.A., Harvey, J.A.: Survey of breast cancer patients concerning their knowledge and expectations of adjuvant therapy. J. Clin. Oncol. **16**(2), 515–521 (1998)
3. Balaha, H.M., Saif, M., Tamer, A., Abdelhay, E.H.: Hybrid deep learning and genetic algorithms approach (HMB-DLGAHA) for the early ultrasound diagnoses of breast cancer. Neural Comput. Appl. 1–25 (2021). https://doi.org/10.1007/s00521-021-06851-5

4. Bicakci, M., Zaferaydin, O., Seyhankaracavus, A., Yilmaz, B.: Metabolic imaging based sub-classification of lung cancer. IEEE Access **8**, 218470–218476 (2020)
5. Liu, C., et al.: Blood-based liquid biopsy: insights into early detection and clinical management of lung cancer. Cancer Lett. **524**, 91–102 (2022)
6. Singh, G.A.P., Gupta, P.K.: Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Comput. Appl. **31**(10), 6863–6877 (2018). https://doi.org/10.1007/s00521-018-3518-x
7. Nasrullah, N., Sang, J., Alam, M.S., Mateen, M., Cai, B., Hu, H.: Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors **19**(17), 3722 (2019)
8. DeMille, K.J., Spear, A.D.: Convolutional neural networks for expediting the determination of minimum volume requirements for studies of microstructurally small cracks, Part I: Model implementation and predictions. Comput. Mater. Sci. **207**, 111290 (2022)
9. Faruqui, N., Yousuf, M.A., Whaiduzzaman, M., Azad, A.K.M., Barros, A., Moni, M.A.: LungNet: a hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data. Comput. Biol. Med. **139**, 104961 (2021)
10. Vijila Rani, K., Joseph Jawhar, S.: Lung lesion classification scheme using optimization techniques and hybrid (KNN-SVM) classifier. IETE J. Res. **68**(2), 1485–1499 (2022)
11. Kaur, J., Gupta, M.: Lung cancer detection using textural feature extraction and hybrid classification model. In: Proceedings of Third International Conference on Computing, Communications, and Cyber-Security, pp. 829–846. Springer, Singapore
12. Malathi, M., Sinthia, P., Madhanlal, U., Mahendrakan, K., Nalini, M.: Segmentation of CT lung images using FCM with active contour and CNN classifier. Asian Pac. J. Cancer Prevent. APJCP **23**(3), 905–910 (2022)
13. Bhattacharjee, A., Murugan, R., Goel, T.: A hybrid approach for lung cancer diagnosis using optimized random forest classification and K-means visualization algorithm. Health Technol. 1–14 (2022)
14. Guo, Y., et al.: Automated detection of lung cancer-caused metastasis by classifying scintigraphic images using convolutional neural network with residual connection and hybrid attention mechanism. Insights Imaging **13**(1), 1–13 (2022)
15. Praveena, M., Ravi, A., Srikanth, T., Praveen, B.H., Krishna, B.S., Mallik, A.S.: Lung cancer detection using deep learning approach CNN. In: 2022 7th International Conference on Communication and Electronics Systems (ICCES), pp. 1418–1423. IEEE
16. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A.V., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P.: Data from LIDC-IDRI (2015)
17. Kirby, J.S., et al.: LUNGx challenge for computerized lung nodule classification. J. Med. Imaging **3**(4), 044506 (2016)
18. Greaves, M., Hughes, W.: Cancer cell transmission via the placenta. Evol. Med. Public Health **2018**(1), 106–115 (2018)
19. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
20. Thacker, W.C.: The role of the Hessian matrix in fitting models to measurements. J. Geophys. Res. Oceans **94**(C5), 6177–6196 (1989)

21. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4), 18–28 (1998)
22. Handelman, G.S., et al.: Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. Am. J. Roentgenol. **212**(1), 38–43 (2019)
23. Ali, I., Muzammil, M., Haq, I.U., Khaliq, A.A., Abdullah, S.: Efficient lung nodule classification using transferable texture convolutional neural network. IEEE Access **8**, 175859–175870 (2020)
24. Wei, G., et al.: Lung nodule classification using local kernel regression models with out-of-sample extension. Biomed. Signal Process. Control **40**, 1–9 (2018)
25. Xie, Y., et al.: Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. IEEE Trans. Med. Imaging **38**(4), 991–1004 (2018)



A Smart Eye Detection System Using Digital Certification to Combat the Spread of COVID-19 (SEDDC)

Murad Al-Rajab[1](✉), Ibrahim Alqatawneh[2], Ahmad Jasim Jasmy[1], and Syed Muhammad Noman[1]

[1] College of Engineering, Abu Dhabi University, Abu Dhabi, UAE
murad.al-rajab@adu.ac.ae, {1075928,1076240}@students.adu.ac.ae
[2] School of Computing and Engineering, University of Huddersfield, Huddersfield, UK
i.alqatawneh2@hud.ac.uk

Abstract. The spread of the COVID-19 pandemic deeply affected the lifestyles of many billions of people. People had to change their ways of working, socializing, shopping and even studying. Governments all around the world made great efforts to combat the pandemic and promote a rapid return to normality. These governments issued policies, regulations, and other means to stop the spread of the disease. Many mobile applications were proposed and utilized to allow entrance to locations such as governmental premises, schools, universities, shopping malls, and a multitude of other locations. The applications most used being the monitoring of PCR (polymerase chain reaction) test results and vaccination status. The development of these applications is recent, and thus they have limitations which need to be overcome to provide an accurate and fast service for the public. This paper proposes a mobile application with an enhanced feature which can be used to speed the control process whereby public enter controlled locations. The proposed application can be used at entrances by security guards or designated personnel. The application relies on artificial intelligence techniques such as deep learning algorithms to read the iris and automatically recognize the COVID-19 status for the person in terms of their PCR test results and vaccination status. This proposed application is of promise because it would enhance safety while simultaneously facilitating a modern lifestyle by saving time compared to current applications used by the public.

Keywords: Mobile application · Eye detection · Artificial intelligence · Covid-19 combat

1 Introduction

The wide spread of COVID-19 has impacted most if not all areas of our lives, from employment and school to the smallest tasks in our daily lives. The coronavirus outbreak has spread to every nation on Earth. Governments have had to wrestle with new lockdown strategies to halt the spread of the coronavirus, with the result that national economies

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 198–212, 2023.
https://doi.org/10.1007/978-3-031-27409-1_18

and companies have experienced drastic changes [1]. As of mid-March 2022, the number of positive cases around the world had reached 476 million with 6 million deaths [32].

The pandemic has changed the way we interact and work. New professions have emerged, creating new opportunities while degrading many existing jobs. As a result of COVID-19, new technologies have emerged [2], data collection, detection of the presence of COVID-19, and evaluation of necessary actions are based heavily on emerging technologies such as artificial intelligence (AI), big data, digital health, 5G technology, real-time applications, and Internet of Medical Things (IoMT). To control the spread of COVID-19, it is essential to know the health status of each individual. To achieve this, governments have introduced several applications showing the required details. For example, the Al Hosn App in the UAE [3], Tawaklna in the KSA [4], the NZ Covid Tracer in New Zealand [5], Trace Together in Singapore [6], Sehet Misr in Egypt [7], COVID Alert in Canada [8] with others in many different countries [27].

Such applications often require the individual to carry the relevant documentation, in a suitable form, prove their COVID-19 test results and vaccination details. As per the law of each country, all citizens and residents who visit a government premises or visit malls, hotels, etc., need to present their details at the entrance to prove their negative results or vaccination details. However, these applications require access to the internet to verify their documentation, which may cause problems as some visitors might not have an internet connection or the data might not be available on their mobile phones. Moreover, there are those who may intentionally violate the procedures by using different ID accounts or provide a screenshot or a video recording of a past negative result to deceive the authorities and security personnel. Another common problem encountered is when visitors present their mobile phones to the security guards who then swipe the phone's screen to check the validity of the test results and the vaccination certificates (or to guide the visitors on how to open the application), however, this process is considered a violation to COVID-19 protocols.

The motivation of this paper is to propose a new mobile application which will overcome the aforementioned challenges in the existing mobile applications. The key features of the proposed mobile application are as follows: (i) Individuals are no longer required to have an internet connection to present their COVID-19 status on their mobile devices. (ii) The proposed application will be available on a controlled device by security personnel which verifies the validity of COVID-19 documents by scanning the face and iris of the individuals. (iii) The proposed application increases the COVID-19 safety protocol by not requiring security personnel to swipe or even touch an individual's devices. The proposed application will be linked with the government's centralized database in which the COVID-19 related details are stored. The contribution of this paper is to:

- Develop a modern mobile application that will assist in the combat of COVID-19.
- Propose an eye detection algorithm based on deep learning that can identify iris and automatically recognize the COVID-19 status.
- Propose mobile application that will support the government's efforts to return to normal life and facilitate the movement of citizens and residents after COVID-19.

The proposed application will be used to speed the control process of entering vital locations. Furthermore, the application will be available on all smartphone platforms with no cost. We intend to analyze the computational cost of the proposed system and we will suggest various techniques to optimize it.

The remainder of the paper is organized as follows. Section 2 discusses related work in the research domain and reviews related applications. Section 3 presents the proposed mobile application including the software architecture, application features, face and iris detection algorithm, and application scenario. Finally, Sect. 4 presents the conclusions.

2 Related Work

The main feature of the proposed application considers face and iris recognition, and this section reviews related work in the public domain. According to the authors in [27, 28], a mobile application has been developed in Singapore that enables the identification and tracking of individuals diagnosed with COVID-19 using Bluetooth signals to maintain records on the mobile phone and to identify when infected people are in close proximity to each other. This helps the Singapore government to identify and collect contact details of individuals infected with COVID-19 in order to mitigate the consequences. On the other hand, the authors in [24] explored technical challenges and ethical issues concerning digital vaccination certificates and some of the digital technologies that can enhance the performance of digital certificates. Even though digital certificates appear to be an important path ahead, they can pose significant questions concerning privacy issues, access without user consent, and improper standards when issuing certificates. To make sure the digital certificates are well maintained, the authors suggested that certificates contain QR codes, global protocols on handling digital certificates and user privacy. Also, there should be official rules and regulations set by the World Health Organization (WHO) on the use of digital certificates and other health tools.

The article [25] used a four-tier method survey to better understand the digital certificates available for COVID-19, examining prior scholarly studies that have suggested digital vaccination certificates and explored their benefits and challenges. The authors assessed all android smartphone applications provided for such certification both statically and dynamically to uncover any potential issues impacting the protection or privacy of the end-user. This was done by reviewing 54 country projects from across the world. It was noticed that there was a significant difference between Asia, America and Europe in terms of the level of privacy of the applications. According to the findings of the static analysis, 42 out of the 47 apps request authorization to utilize the camera, and around one-third of the apps require location permission and, additionally, ask for read/write authorization to the smartphone's external drives. These apps had to use a camera to read the digital certificates from the smartphone and read the QR code present in the certificate. Based on these results, it can be stated that European privacy laws guarantee a higher level of privacy than those from Asia and America, which frequently demand more delicate permits for location-based services and connection with external drives.

A review [26] sought to provide a thorough analysis of AI techniques that have been applied to identify face masks. One of the deep learning models, Inception-v3 showed an accuracy of 99.9% in detecting face masks. To assess the performance of deep learning

techniques it is important to use real life face mask images to ensure the obtained results are accurate and reliable. The research also indicated that in 2021 mask detection had been implemented using broader and deeper learning algorithms with an enhanced adaptive algorithm, such as Inception-v4, Mask R-CNN, Faster R-CNN, YOLOv3, or DenseNet. However, a number of AI techniques have been successfully applied in other areas of object recognition, but their application to identifying COVID-19 face masks remains untried. Due to the wide variety of mask styles, different camera pixel sizes, varying levels of obstacles, numerous variants of images, face mask identification has proven to be challenging.

The authors in [9] explored a number of variables that impact on the quality of the iris patterns. The size of the iris, picture quality, and capture wavelength were found to be important determinants. A case study of a present smartphone model is presented, along with a computation of the screen resolution achievable with an observable optical system. The system specifications for unsupervised acquisition in smartphones were outlined based on these assessments. Various layout methods were proposed, as well as important challenges and their solutions.

The authors in [11] proposed an adaptive approach for eye gaze monitoring in a cloud computing and mobile device context. The main issues found were involuntary head movement by the user and mobile devices that lacked sufficient computational capacity. To address these, an integrative cloud computing architecture was proposed with the help of neural networks which calculates the coordinates of the eye gaze and achieved high performance in a heterogeneous environment [11]. Chiara Galdi et al., proposed an algorithm called FIRE in [12]. FIRE is a novel multi-classifier for fast improper iris recognition, comprising a color descriptor, a texture descriptor, and a cluster descriptor. It was tested on a very challenging database, namely the MICHE-I DB composed of iris images collected by different mobile devices. A variety of different strategies were evaluated, and the best performers were chosen for fusion at the score level. The applicability of the suggested approach for iris detection under visible light is a significant feature, given that there are numerous application situations where near-infrared (NIR) illumination is neither accessible nor practical.

According to the results presented in [10], iris and periocular authentication is the most accurate biometric authentication for security measures. However, it requires the capture of a high-quality picture of the iris, which necessitates the use of a better image sensor and lens. Iris and periocular images are recorded at the same time, and periocular authentication is used to correct for any decrease of iris validation due to the use of a lower-quality camera. To obtain more accurate user authentication, the authors developed an authentication approach that used AdaBoost for the score fusion algorithm. AdaBoost is a common "boosting" technique that in most cases delivers greater discrimination reliability than individual discriminators [10].

The authors in [14] explored an eye recognition system using smartphones. According to the authors, rapidity of eye recognition, the variance between the user's gaze point and the center of the iris camera, the distance between the iris camera and the NIR LED illuminators are some of the main issues that affect the implementation of iris recognition systems. As a result, the suggested system was capable of detecting high-quality iris pictures and employed multiple image matching to improve identification rates. On

the other hand, the authors in [13] explored the feasibility of iris recognition on smartphone devices and suggested that the advanced spatial histogram could be helpful in iris recognition and matching features [13].

Capture circumstances have a great impact on the capabilities of iris recognition processes. However, the clarity of an imaging sensor does not always imply an increased detection performance. The method suggested in [13] was evaluated using the iris dataset, which included participants recorded indoors and outdoors using built-in cameras in monitored and unmonitored situations. The tested data gave fascinating information on the potential of developing mobile technology for iris detection [13].

Another study [15] introduced an iris identification system using self-organizing maps (SOM) to represent the iris of people in a low two-dimensional environment. Unmonitored approaches are used in SOM networks, making them ideal for mobile devices and personal detection techniques [15]. The suggested pixel-level technique combines RGB triples in each iris pixel's visible light spectrum with statistical classifiers generated by kurtosis and skewness on a surrounding window [15].

The authors in [16] developed and tested iris authentication systems for smartphones. The method relies on visible-wavelength eye pictures, which are captured by the smartphone's built-in camera. The system was introduced in four stages [16]. The first stage was iris classification, which included Haar Cascade Classifier training, pupil localization, and iris localization using a circular Hough Transform that can recognize the iris region from the captured image. In the second stage, the model employed a rubber sheet model to standardize the iris images, transforming them into a predefined sequence. In the third stage, a deep-sparse filtering technique extracted unique characteristics from the pattern. Finally, seven potential matching strategies were explored to determine which one of the seven systems would be used to validate the user.

The authors in [17] suggested a novel approach to authentication of eye tracking. As input, they used eye movement trajectories, which represent the direction of eye movement but not the precise look location on the screen. There was no need for an increased detector or a calibration procedure. The authors believe that this is the first such process ever implemented on smartphones.

Another contribution proposed a Light Version algorithm that helps to detect and capture iris images on smartphones [18]. The process includes three steps. First, they changed and redesigned the iris recognition algorithms that work in a smartphone environment. In the second step, the algorithm was extended to search for the best optimization solution for smartphone authentication and verification processes. Finally, they employed the updated CASIA-IrisV4 dataset. Their results demonstrated that the implementation of the LV recognition algorithm on smartphones yields better performance results in terms of CPU utilization, response time, and processing time [18].

The EyeVeri authentication system was proposed by [19]. This system is considered a new eye movement-based verification mechanism for protecting smartphone security. It uses the built-in front camera to capture human eye movement and uses signal processing and eye pattern matching techniques to investigate authentication processes [19]. The system performed well in the given tests and is considered a potential technique for identifying smartphone users.

The current system used by the citizens and residents of the UAE is called the "Al Hosn App" [13]. This is the official app for contact tracking and COVID-19 health status. All citizens and residents have this application on their mobile device in order to show their latest PCR result, vaccination, and travel history. There is a pass system, which shows green if the latest RT-PCR test result is negative and valid, grey if there is no valid PCR test, and red if the user tested positive. The application also helps check vaccination status, as well as vaccine information and records, travel tests, and a live QR code to download or update the app, which is required of all individuals with a stable internet or data connection.

According to [33], Interval type-2 fuzzy system evolved from Interval type-1 fuzzy system and performs well in a noisy environment. This paper presented a new method for fuzzy aggregation with a group of neural networks. The aggregator combines the outputs of the neural networks, and the overall output of the ensemble is higher than the outputs of an individual neural network. In the proposed method, the values given to the weighted average of the combined are estimated using fuzzy system. To represent the uncertainty of the aggregation Interval type-3 was used and the simulation data showed the ability of the Interval type-3 fuzzy aggregator to outperform both Interval type-2 and type-1 fuzzy aggregators.

For the purpose of forecasting COVID-19 data, the authors in [34] suggested the implementation of ensemble neural networks (ENNs) and type-3 fuzzy inference systems (FISs). Values from the type-3 FIS are used to integrate the results for each ENN component, with the ENN created using the firefly algorithm. The combination of the Ensemble Neural Network (ENNs), type-3 fuzzy logic, and firefly algorithm is referred to as ENNT3FL-FA. When ENNT3FL-FA was applied to COVID-19 data from 12 countries (including the USA, UK, Mexico, India and Germany), the authors claimed their system more accurately forecast proven COVID-19 cases than other integration methods which used type-1, and type-2 fuzzy weighted average.

The authors in [35] proposed a new unsupervised DL-based variational autoencoder (UDL-VAE) model for the recognition and classification of COVID-19. Inception-v4 with Adaptive Gradient Algorithm (Adagrad)-based extraction of features, unsupervised VAE-based classification, and Adaptive Wiener filtering (AWF)-based processing are all performed by the UDL-VAE model. The AWF methodology is the primary method for improving the quality of the medical images. The usable collection of features is extracted from the preprocessed image by the Adagrad model within Inception-v4. By using the Adagrad technique, it was possible to enhance the classification efficiency by adjusting the parameters of the Inception-v4 model. The correct class labels for the input medical photographs are then defined using an unsupervised VAE model. The experimentally determined values demonstrated the UDL-VAE model gave enhanced values of accuracy of 0.987 and 0.992 for prediction of binary and multiple classes, respectively. The authors explained that could be a very positive contribution to healthcare applications via the internet of things (IoT) and cloud based environments. Table 1 summarizes the features of the currently most common existing mobile applications.

Table 1. Most common existing COVID-19 mobile applications.

Application Name	Features and Advantages
Al Hosn App (United Arab Emirates) [3]	The Application helps check your status depending on your latest RT-PCR test, vaccination status/information/records, and any travelling tests
Tawakkalna (Kingdom of Saudi Arabia) [4]	Covid vaccine booking and doses info, covid test services, displays QR code. Positive and negative cases display. Shows permits during curfew. Color-coding system enables security status while also displaying health status
NZ Covid Tracer (New Zealand) [5]	Bluetooth tracing for positive users, QR code scanner, Quick test references, Updates contact information. The app can help medical teams contact positive individuals
Arogya Setu (India) [20]	Alerts for close positive contact cases, Self-assessment to analyze symptoms. For positive cases app turns red and start tracking the infected person and his/her contacts, Registration for vaccinations, Informs positive contact cases in last 14 days
COVID Alert (Canada) [8]	Sends and receives codes between phones through Bluetooth, every day app analyzes codes from positive cases. Positive individuals must notify others by entering one-time key in the app. Notifies close contacts of positive cases in last 14 days. Application can reduce the spread of the virus by proper detection and isolation
Trace Together (Singapore) [6]	Signals via Bluetooth trace positive covid route map. Mobile phones interchange anonymized IDs for positive cases. Guide, and support to isolate positive cases

(*continued*)

Table 1. (*continued*)

Application Name	Features and Advantages
COVID-19 Gov PK (Pakistan) [21]	Delivers awareness on covid, displays current status, constant alert for hand wash, covid recognition chatbots. A Radius Alert is under development to promote social distancing
Sehet Misr (Egypt) [7]	App enables covid positive individuals to connect with the medical team via WhatsApp, report suspected cases through the app, which then sends alerts to users when they are close to positive cases if their location is enabled. Provides tips and awareness on covid protection
PathCheck (United States of America) [22]	Notifies exposure to covid positive cases. Symptoms self-check, support for self-quarantine. Vaccination proof and dose reminders, record of health status
COVIDSafe (Australia) [23]	Functions by Bluetooth, displays new cases, confirmed cases, deaths etc. If a person tests positive his/her application scans for his/her close contact and calls them for tests and isolation. App also notifies of nearby positive cases

3 The Proposed Mobile Application

Since this study was conducted in UAE, we started our proposed mobile application by conducting a survey among 125 individuals in the UAE. The survey was of citizens and residents and included those who work as security personnel in different locations such as government buildings, malls, academic institutions, etc., who were in direct contact with the public and using the Al Hosn App at entries and reception areas for processing and checking. The survey results are illustrated in Fig. 1. It was found that 57.5% of the respondents agreed that they had internet problems while presenting the Al Hosn application. However, 45% of the participants did not feel safe when security guards swiped the screen and touched their mobile phone. On the other hand, 35% of those respondents presenting without the Al Hosn App on their mobile phones would recommend having an alternative solution.

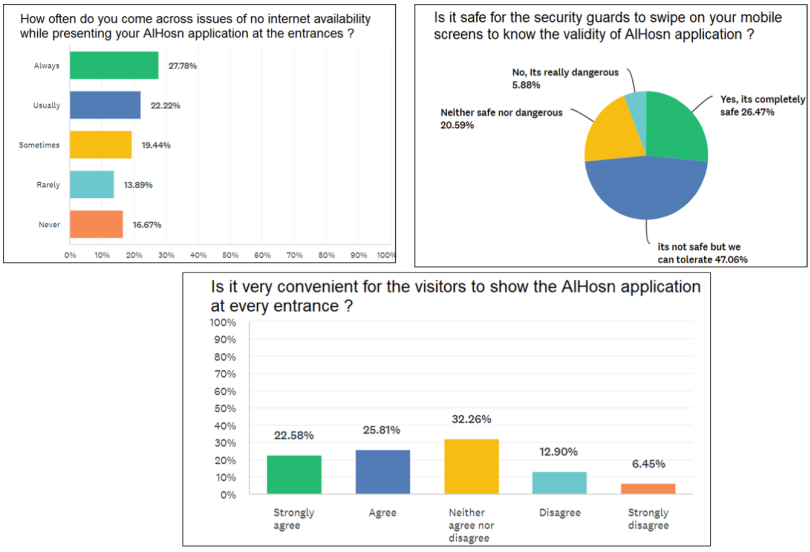


Fig. 1. Survey analysis results for the current mobile application

The common challenges faced were no internet availability, app takes time to load, updates very slow, generates a nervous state when the app does not work or malfunctions. The majority of respondents (71%) agreed that it is necessary for the security guard to know the COVID-19 status of all visitors. The most desirable feature favored by users was an application that works without the need of the Internet. Figure 1 presents the responses concerning issues that the public faces when using the current application.

3.1 Software Architecture

This section discusses the design of the software architecture in detail. Our proposed application adopts the well-established three-tier model, commonly used in software

development, and it consists of three tiers: presentation, application and database. Figure 2 shows the architecture of the proposed mobile application.

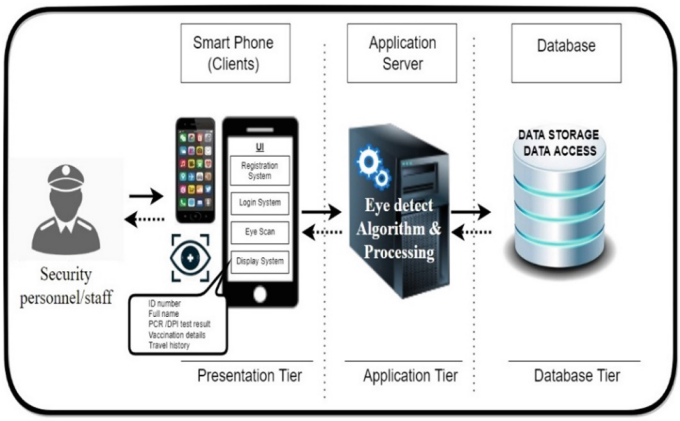


Fig. 2. Software architecture

1. **Presentation Tier**: the authorized user (security personnel or staff) activates the presentation tier which includes the following interface components: registration screen, login option, eye detection option and information display. First, the authorized user needs to register his/her details, which are verified by the health authority. Upon login, the registered user should be able to click on the scan button starting the application to scan the eyes. This enables the authorized user to request information from the health care database based on the iris of the visitor. Finally, the required information will be retrieved from the database and displayed on the smartphone being used by the authorized user. The information retrieved will include national ID number, full name, PCR test result, vaccination details and travel history of the person whose eye was scanned.
2. **Application Tier**: is the engine of the application. It is the middle layer and central to the entire application. It acts as the server for the data entered in the presentation layer. It receives the data from the recognized eye as an input from the presentation layer, and processes that data using a deep learning algorithm such as the Convolutional Neural Network (CNN). The function of the deep learning algorithm is to detect the features of the eye of the individual visitor, and then identify appropriate records from the database in order to retrieve full, detailed information of the individual visitor.
3. **Database Tier**: is the database layer where all the results obtained from the deep learning algorithm will be sent to compare and match the detected features of the recognized eye. The registration details provided by the authorized user will simultaneously be saved in the health authority database. Finally, the retrieved detailed information from the database is sent to the presentation layer and displayed on the mobile application.

3.2 Application Features

The proposed mobile application has the following features:

The information registered by the authorized user is saved in the health authority database. The application will also return a unique code for each organization when the sign-up process is completed successfully.
The application uses an eye detection algorithm based on deep learning to recognize the visitor's eye and retrieves appropriate details from the health care database.
All information is retrieved from the database of the health care organization, including COVID-19 vaccination details and travel history.
The application will display the following information: national ID number, full name, PCR test result (red/green), vaccination details (number and dates of doses), and travel history (latest arrival date to the country and dates of previous visits to the country).
The retrieved information will display two signs to the security personnel, red if the result is positive, displaying "STOP", or a green signifying "GO" if the result is negative.

3.3 Eye Detection Algorithm

A key feature of the proposed mobile application is an eye detection algorithm, as illustrated in Fig. 3. The smartphone camera captures the facial image of a visitor and focuses on the eyes in order to localize the inner (pupillary) and outer (limbic) boundaries [29]. The area surrounded by the internal and external boundaries of the iris may change due to pupil extension and contraction [31], so before comparing different iris images the effects of these variations are minimized. For this purpose, segmented iris regions are usually mapped to fixed-dimension regions [31]. Once the algorithm detects the relevant eye region, it crops the eye image into a rectangular shape, and proceeds to the next process where iris localization and normalization takes place. This can be done using Daugman's Rubber Sheet Model [30, 31]. Another advantage of normalization is that eye rotation (including head rotation) is reduced to a simple translation when matching [31]. The normalized iris image is then input into the CNN's functional extraction model. After the iris localization, the algorithm extracts the features and color of the iris in the form of a unique iris texture code. The CNN function vector is then input to the classification model to detect the features. Once the pre-processing steps have been completed, the obtained iris code is sent to the database to verify the owner. If the iris codes match, the database returns the required details relevant to the owner of the eye. One of the most important reasons CNN works so very well in computer vision tasks is that the CNN layers have millions of parameter-level deep networks that are able to capture and encrypt complex image features extremely well, achieving high performance [31].

3.4 Application Scenario

Figure 4a–d represents the user interface of the proposed mobile application. Figure 4a shows the signup screen, where authorized staff will enter the name of the organization, registration number and password to register themselves and their organization in the application. Figure 4b the eye detection process takes place when the "SCAN" button is

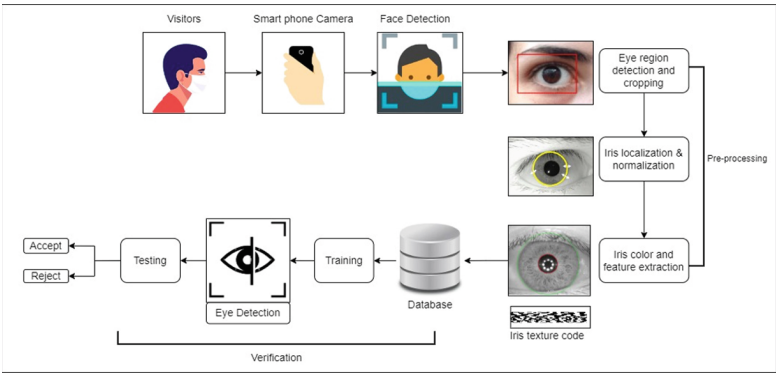
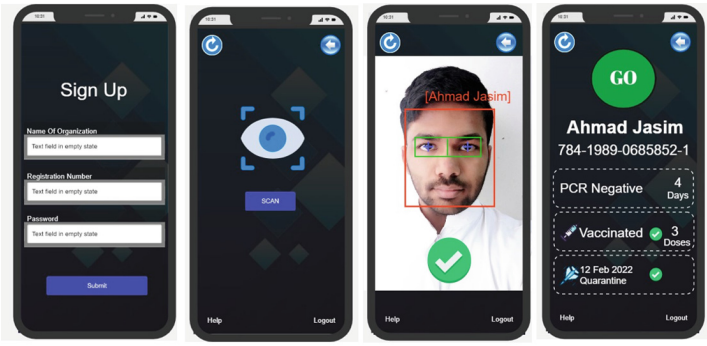


Fig. 3. Flowchart of the eye detection algorithm



(a) (b) (c) (d)

Fig. 4. Illustrative example of the proposed mobile application

triggered by an authorized member of staff, after which the device's camera will be given access to capture an image of the eye. Figure 4c shows the captured image of the visitor where both face and eyes are being detected. The name of the visitor will be displayed on the screen which also informs that the eye and facial features are those of the named individual. Figure 4d represents the obtained information of the visitor being detected, name and the ID number, PCR test results, vaccination details, and travel history. At the top of the screen, a green "GO" will be displayed for those meeting the necessary conditions, and a red "STOP" for those who do not. All the screens will have options to reload the retrieved information, return to home screen, help, and logout.

Government's consent for data transmission must be obtained by the application. A point-to-point data transfer network is created once the necessary approval has been received. The data shared on the network will be encrypted and digitally approved by the government. The shared information from the governmental database will not be saved in the application.

4 Limitation and Challenges

The limitations and challenges in the development of the proposed mobile application were:

Centralized database: the current stage of developing the mobile application will use a simulated centralized database. However, the integration governmental centralized databases are required for approval.

Upgrades and modifications: the proposed application will need specific upgrades and modifications to its system according to the changing environment. These issues will be addressed and considered when testing the application in real-life scenarios.

5 Conclusions

This paper proposes a convenient smart mobile application to help overcome the present difficulties of existing COVID-19 tracing applications. The proposed solution includes a smart eye detection algorithm based on a deep learning model that can identify an iris and automatically recognize the COVID-19 status of individuals. It is proposed the application is accessible through all smartphone platforms but can be used only by authorized staff of registered organizations. The application can help save time and resolve network access issues, as individuals will no longer need to carry out with them any written proof of their COVID-19 status. The proposed application will help governments combat and overcome the management of COVID-19 consequences and will support an easier return to normal life.

There are a number of interesting directions for a future work. Firstly, we will implement the proposed application and make it available on all smartphone platforms. Secondly, the proposed application will need specific upgrades and modifications to its system according to the changing environment. Finally, we intend to look for optimization techniques that can be used to justify the cost of the proposed mobile application.

References

1. WHO Coronavirus (COVID-19): World Health Organization, March 2022. https://covid19.who.int/
2. Mbunge, E., Akinnuwesi, B., Fashoto, S.G., Metfula, A.S., Mashwama, P.: A critical review of emerging technologies for tackling COVID-19 pandemic. Human Behav. Emerg. Technol. **3**(1), 25–39 (2021)
3. TDRA: The Al Hosn app. TDRA, 21 September 2021. https://u.ae/en/information-and-services/justice-safety-and-the-law/handling-the-covid-19-outbreak/smart-solutions-to-fight-covid-19/the-alhosn-uae-app
4. Tawakkalna: Kingdom of Saudi Arabia. https://ta.sdaia.gov.sa/en/index
5. Tracer, N.C.: Protect yourself, your Whānau, and your community. https://tracing.covid19.govt.nz/
6. Trace Together: Singapore Government Agency. https://www.tracetogether.gov.sg
7. El-Sabaa, R.: Egypt's health ministry launches coronavirus mobile application (2020)
8. COVID Alert: Government of Canada. https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html#a1

9. Thavalengal, S., Bigioi, P., Corcoran, P.: Iris authentication in handheld devices-considerations for constraint-free acquisition. IEEE Trans. Consum. Electron. **61**(2), 245–253 (2015)
10. Oishi, S., Ichino, M., Yoshiura, H.: Fusion of iris and periocular user authentication by adaboost for mobile devices. In: 2015 IEEE International Conference on Consumer Electronics (ICCE), pp. 428–429. IEEE (2015)
11. Kao, C.W., Yang, C.W., Fan, K.C., Hwang, B.J., Huang, C.P.: An adaptive eye gaze tracker system in the integrated cloud computing and mobile device. In: 2011 International Conference on Machine Learning and Cybernetics. IEEE (2011)
12. Galdi, C., Dugelay, J.L.: FIRE: fast iris recognition on mobile phones by combining colour and texture features. Pattern Recogn. Lett. **91**, 44–51 (2017)
13. Barra, S., Casanova, A., Narducci, F., Ricciardi, S.: Ubiquitous iris recognition by means of mobile devices. Pattern Recogn. Lett. **57**, 66–73 (2015)
14. Kim, D., Jung, Y., Toh, K.A., Son, B., Kim, J.: An empirical study on iris recognition in a mobile phone. Expert Syst. Appl. **54**, 328–339 (2016)
15. Abate, A.F., Barra, S., Gallo, L., Narducci, F.: Kurtosis and skewness at pixel level as input for SOM networks to iris recognition on mobile devices. Pattern Recogn. Lett. **91**, 37–43 (2017)
16. Elrefaei, L.A., Hamid, D.H., Bayazed, A.A., Bushnak, S.S., Maasher, S.Y.: Developing Iris recognition system for smartphone security. Multimedia Tools Appl. **77**(12), 14579–14603 (2018)
17. Liu, D., Dong, B., Gao, X., Wang, H.: Exploiting eye tracking for smartphone authentication. In: International Conference on Applied Cryptography and Network Security, pp. 457–477. Springer, Cham (2015)
18. Ali, S.A., Shah, M.A., Javed, T.A., Abdullah, S.M., Zafar, M.: Iris recognition system in smartphones using light version (lv) recognition algorithm. In: 2017 23rd International Conference on Automation and Computing (ICAC). IEEE (2017)
19. Song, C., Wang, A., Ren, K., Xu, W.: Eyeveri: a secure and usable approach for smartphone user authentication. In: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, pp. 1–9. IEEE (2016)
20. Arogya Setu: Government of India. https://www.aarogyasetu.gov.in/#why
21. COVID-19 Gov PK: Ministry of IT and Telecommunication, Pakisthan. https://moitt.gov.pk/NewsDetail/NjQ3NWQyMDMtYTBlYy00ZWE0LWI2YjctYmFmMjk4MTA1MWQ0
22. PathCheck: PathCheck Foundation. https://www.pathcheck.org/en/covid-19-exposure-notification-app?hsLang=en
23. COVIDSafe: Australian Government. https://www.covidsafe.gov.au/
24. Mbunge, E., Fashoto, S., Batani, J.: COVID-19 digital vaccination certificates and digital technologies: lessons from digital contact tracing apps (2021)
25. Karopoulos, G., Hernandez-Ramos, J.L., Kouliaridis, V., Kambourakis, G.: A survey on digital certificates approaches for the covid-19 pandemic. IEEE Access (2021)
26. Mbunge, E., Simelane, S., Fashoto, S.G., Akinnuwesi, B., Metfula, A.S.: Application of deep learning and machine learning models to detect COVID-19 face masks—a review. Sustain. Oper. Comput. **2**, 235–245 (2021)
27. Loucif, S., Al-Rajab, M., Salem, R., Akkila, N.: An overview of technologies deployed in GCC Countries to combat COVID-19. Period. Eng. Nat. Sci. (PEN) **10**(3), 102–121 (2022)
28. Whitelaw, S., Mamas, M.A., Topol, E., Van Spall, H.G.: Applications of digital technology in COVID-19 pandemic planning and response. Lancet Digital Health **2**(8), e435–e440 (2020)
29. Vyas, R., Kanumuri, T., Sheoran, G.: An approach for iris segmentation in constrained environments. In: Nature Inspired Computing. Springer, Singapore (2018)
30. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. (1993)

31. Nguyen, K., Fookes, C., Ross, A., Sridharan, S.: Iris recognition with off-the-shelf CNN features: a deep learning perspective. IEEE Access **6**, 18848–18855 (2017)
32. Coronavirus cases: Worldometer. https://www.worldometers.info/coronavirus. Accessed 07 Aug 2022
33. Castillo, O., Castro, J.R., Pulido, M., Melin, P.: Interval type-3 fuzzy aggregators for ensembles of neural networks in COVID-19 time series prediction. Eng. Appl. Artif. Intell. **114**, 105110 (2022)
34. Melin, P., Sánchez, D., Castro, J.R., Castillo, O.: Design of type-3 fuzzy systems and ensemble neural networks for COVID-19 time series prediction using a firefly algorithm. Axioms **11**(8), 410 (2022)
35. Mansour, R.F., Escorcia-Gutierrez, J., Gamarra, M., Gupta, D., Castillo, O., Kumar, S.: Unsupervised deep learning based variational autoencoder model for COVID-19 diagnosis and classification. Pattern Recogn. Lett. **151**, 267–274 (2021)



Hyperspectral Image Classification Using Denoised Stacked Auto Encoder-Based Restricted Boltzmann Machine Classifier

N. Yuvaraj[1], K. Praghash[2(✉)], R. Arshath Raja[1], S. Chidambaram[2], and D. Shreecharan[2]

[1] Research and Publications, ICT Academy, IIT Madras Research Park, Chennai, India
[2] Department of Electronics and Communication Engineering, CHRIST University, Bengaluru, India
prakashcospra@gmail.com

Abstract. This paper proposes a novel solution using an improved Stacked Auto Encoder (SAE) to deal with the problem of parametric instability associated with the classification of hyperspectral images from an extensive training set. The improved SAE reduces classification errors and discrepancies present within the individual classes. The data augmentation process resolves such constraints, where several images are produced during training by adding noises with various noise levels over an input HSI image. Further, this helps in increasing the difference between multiple classes of a training set. The improved SAE classifies HSI images using the principle of Denoising via Restricted Boltzmann Machine (RBM). This model ambiguously operates on selected bands through various band selection models. Such pre-processing, i.e., band selection, enables the classifier to eliminate noise from these bands to produce higher accuracy results. The simulation is conducted in PyTorch to validate the proposed deep DSAE-RBM under different noisy environments with various noise levels. The simulation results show that the proposed deep DSAE-RBM achieves a maximal classification rate of 92.62% without noise and 77.47% in the presence of noise.

Keywords: Stacked Auto Encoder · Hyperspectral images · Noise · Restricted Boltzmann Machine · PyTorch

1 Introduction

Hyperspectral imaging (HSI) has gained popularity in visual data processing for decades [1]. HSI has applications in biomedicine, food quality, agricultural legacy, and cultural heritage in Remote Sensing [2]. Since each pixel contains reflectance measurements over narrow-band spectral channels, it's possible to transmit more information about an image's spectral composition than RGB or multi-spectral data [3]. Current HSI acquisition methods [4] can provide high spectral resolution while providing sufficient throughput and spatial resolution [5].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 213–221, 2023.
https://doi.org/10.1007/978-3-031-27409-1_19

HSI's handling challenges limit the amount of data. A sparse data distribution causes the curse of dimensionality when multiple channels generate HSI data. HSI data processing is complex, and high-quality information isn't always possible. Due to redundancy, this study uses dimensionality reduction techniques to achieve high spatial resolution. Recent learning approaches [6] using Deep Learning (DL) architectures have solved the spatial resolution problem. These problems will always hamper traditional DL methods because they rely heavily on selected features.

This paper optimized several features for traditional DL for HSI data interpretation [7]. After using a simple linear classifier, the feature set and classifiers became more complex. DL solutions have a few drawbacks, but the most significant advantage is building models with higher and higher semantic layers until the data derives an appropriate representation for the task. Several methods work. Despite these benefits, the DL on hyperspectral data should be approached with caution. A large dataset is needed to avoid overfitting DL models' many parameters.

The study considers a dataset with hundreds of small samples. DL meets HSI lacks public datasets, which is its biggest flaw. When training data are scarce, it can lead to the so-called Hughes effect [8], models that cannot generalize due to dimensionality. Another issue hidden behind insufficient data for research is that the dataset has too much data, limiting possible solutions. The lack of labeled data forces us to use unsupervised algorithmic approaches. Data augmentation and DL design implementation improve many data-driven problems.

Stacked Auto Encoder (SAE) deals with parametric instability in hyperspectral image classification from an extensive training set.

The main contribution of the paper is:

- The authors improved the SAE to reduce classification errors and discrepancies in individual classes. The data augmentation process resolves such constraints, where several images are produced during training by adding noises with various noise levels over an input HSI image. Further, this helps increase the difference between multiple classes of a training set.
- The authors modified SAE to classify the HSI image using a Restricted Boltzmann Machine (RBM) principle of denoising.
- To operate on selected bands through band selection models, i.e., the classifier eliminates noise from these bands to produce higher accuracy results.

The paper's outline: Sect. 2 discusses the literature survey. In Sect. 3, we discuss the proposed deep DSAE-RBM classifier. Section 4 evaluates the entire model. Section 5 concludes the paper with directions for future scope.

2 Literature Survey

Bahraini et al. [8] avoided labeling errors in HSI classification using a modified mean shift (MMS). Machine learning algorithms classify denoised samples. classification errors are lower than before.

Shi et al. [9] proposed dual attention denoising for spectral and spatial HSI data. The attention module forms interdependencies between spatial feature maps, and channel attention simulates spectral branch correlations. This combination improves denoising models.

Xu et al. [10] proposed a dual-channel residual network (DCRN) for denoising the labels. Experiments show dual attention denoising, and other methods perform better.

Ghasrodashti and Sharma [11] used the spectral-spatial method to classify HSI images. WAE is used to extract spectral features. The fuzzy model improves auto-encoder-based classification. It improves classification accuracy over conventional models.

Miclea et al. [12] proposed a parallel approach (PA) for dimensionality-reduced feature extraction. The classifier uses controlled sampling and unbiased performance to classify spatial-spectral features.

The wavelet transform in the spectral domain and local binary patterns in the spatial domain extract features. An SVM-based supervised classifier combines spectral and spatial features. For the experimental validation, we propose a controlled sampling approach that ensures the independence of the selected samples for the training data set, respectively, the testing data set, offering unbiased performance results. Randomly selecting models for a hyperspectral dataset's learning and testing phases can cause overlapping, leading to biased classification results. The proposed approach, with controlled sampling, performs well on Indian pines, Salinas, and Pavia university datasets.

3 Proposed Method

In this section, we improve the stacked auto encoder (SAE) to deal with the problem of instability in parameters associated with the DL model while training the model with an extensive training set, as illustrated in Fig. 1. The SAE is modified to reduce the errors associated with the classification process and further reduces the discrepancy within the individual classes in the dataset. The difference is resolved by augmenting more datasets within the training via adding noises in various patterns in the input raw HSI image. Further, this helps increase the contrast between multiple classes of a training set.
In this section, we improve the stacked auto encoder (SAE) to deal with the problem of instability in parameters associated with the DL model while training the model with an extensive training set, as illustrated in Fig. 1. The SAE is modified to reduce the errors associated with the classification process and further reduces the discrepancy within the individual classes in the dataset. The difference is resolved by augmenting more datasets within the training via adding noises in various patterns in the input raw HSI image. Further, this helps increase the contrast between multiple classes of a training set.

3.1 Band Selection Segmentation

Additionally, band selection is employed to reduce data size in hyperspectral remote sensing, which acts as a band sensitivity analysis tool in selecting the bands specific to the region of interest. Various methods are found to operate on band selection on

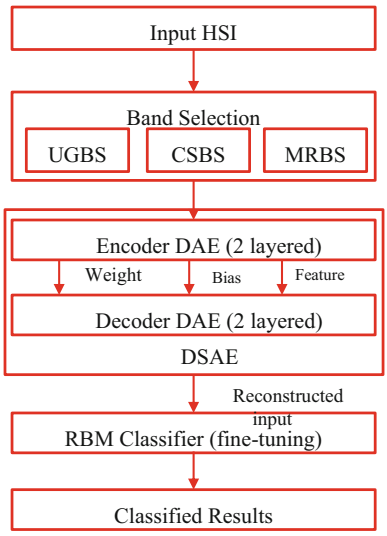


Fig. 1. Proposed HSI Classification Model

HSI images, and this includes (1) the Unsupervised Gradient Band Selection (UGBS) model [3] eliminates the redundant band using a gradient of volume. (2) Column Subset Band Selection (CSBS) [4] model is designed to maximize the volume of the selected band to reduce the size of the HSI image in noisy environments. (3) Manifold Ranking Salient Band Selection (MRBS) [5] method transforms the band vectors, especially in manifold space, and it tends to select a band-based ranking to tackle the unfitting data measurement present in the band difference. Once these pre-processing operations are performed, the study HSI features from the selected bands are classified in the following section.

3.2 Classification

The modified SAE is developed to classify the HSI image using the principle of denoising.

Pretraining by Denoised SAE

In an autoencoder, the input layer is larger than the encoding layer, making it smaller. Some applications require a more comprehensive encoding layer than the input layer to prevent the SAE from learning a mapping function. Limiting sparseness and distorting input data can avoid problems of mapping. The DSAE is stacked to create a deep learning network with hidden layers. Figure 1 shows SDAE's constant encoding and decoding layers. The first encoding output becomes the second input. With N hidden layers, layer (nactivation)'s function is given below in Eq. (1):

$$y(n+1) = f_e(W(n+1)y(n) + b(n+1)), n = 0, 1, ..., N-1 \quad (1)$$

where $y(0)$—input from the original data x, $y(N)$—output from the second encoding layer. Such output is regarded as the extracted high-level features from SDAE.

Similarly, the decoding output acts as a second decoding input. With N hidden layers, the activation function in the encoding of layer (n) is given in the below Eq. (2):

$$Z(n+1) = f_d\left(W(N-n)T \cdot z(n) + b^{'}(n+1)\right), n = 0, 1, ..., N-1 \tag{2}$$

where $z(0)$—input from the 1st decoding layer and $z(N)$—output from the 2nd decoding layer i.e., reconstructed original data x. The training of input HSI is given in Algorithm 1.

Algorithm 1: DSAE

- Randomly selection of HSI data
- Train encoding part (k) (1st encoding and last decoding layer)
- Attain the bias $b\,(1)$, weights $W\,(1)$, and features $y\,(1)$ (1st encoding outputs).
- Send the input $y(n)$ to the encoding layer $(n+1)$
- Train DAE $(n+1)$
- Attain the bias $b(n+1)$, weights $W(n+1)$ and features $y(n+1)$

The trained weights by DSAE are considered the initial weights for RBM to fine-tune the classification results during a training phase.

Fine-tuning by RBM

In RBM [6], there are two distinct forms of stochastic visible units: a layer of hidden units and a layer of hidden units. Both of these types of stochastic visible units are hidden units. RBMs have links between all of the visible units and the hidden units, but there are no connections between the remote units and the visible units. RBMs are capable of being modeled using bipartite graphs.

In RBM, $p(v, h; \theta)$, joint distribution over the visible v and hidden layers h with parameters θ, is defined as an energy function $E(v, h; \theta)$ and it is given in the below Eq. (3):

$$\mathrm{p}(\mathbf{v}, \mathbf{h}; \theta) = \frac{exp(-\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta))}{Z}, \tag{3}$$

where Z—partition function, the energy function is defined as in Eq. (4)

$$\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij} v_i h_j - \sum_{i=1}^{I} b_i v_i - \sum_{j=1}^{J} a_j h_j, \tag{4}$$

where wij—interaction between vi and hj, and I—visible units, J—hidden units, and aj and bi—bias terms.

This RBM classifier is a generative model that features the data distribution from input data via several hidden variables without any label information.

Deep DSAE-RBM includes pre-training from DSAE and fine-tuning by RBM. The network weights are trained via DSAE, and reconstructed learning helps find the relevant

features further. The consequences (learned) act as an initial weight for RBM. And RBM fine-tunes the overall classification process and obtains fine-tuned results. It is seen that the DSAE process is unsupervised, and RBM performs supervised operations with limited labeled data. Here, the initial features are produced from the outputs of the encoding part, and its production is given as input to the RBM layer. The sigmoid function is hence used as an activation function as in Eq. (5):

$$h(x) = \frac{1}{1 + exp(-Wx - b)} \tag{5}$$

Where x is the output of the last encoding layer.

The sigmoid output between the values 0 and 1 represents the classification results, where the classifier's feedback is used to fine-tune network weights. Such feedback fine-tuning uses cost function as in Eq. (6):

$$\text{Cost} = -\frac{1}{m}\left[\sum_{i=1}^{m} l(i)log(h(x(i))) + (1 - l(i))log(1 - h(x(i)))\right] \tag{6}$$

where $l(i)$—HSI sample label $x(i)$.

The reduction of cost function helps in updating the weights in the network, and this is solved by the minimal stochastic gradient descent (*Cost*) method. The steps relating to fine-tuning the outputs during network training are given below.

Algorithm 2: Training network with DSAE-RBM

- Train the initial weights using DSAE
- Set the initial weight of RBM randomly
- Obtain classification results from the initial weight
- Tune the weights iteratively using cost function minimization
- End the process

4 Results and Discussions

This section presents classification accuracy results using the proposed deep learning classifier on various reduction techniques such as UGBS, CSBS, and MRBS. The models are evaluated in the presence of noise (noise augmented) and on original images (raw images). Validation of the proposed model in such a testing environment is conducted to test the robustness of the model. The addition of noises on raw images increased the number of samples required to train the classifier with noises and, similarly, test the classifier.

The simulation is conducted in the PyTorch package [7], which offers high-level features for HSI classification. The simulation runs on a high-end computing engine with 32GB of GPU and RAM on an i7 core processor. The proposed deep learning model is tested on different datasets of varying classes: the Indian Pines Dataset, the Pavia University Dataset, and the Kennedy Space Center (KSC).

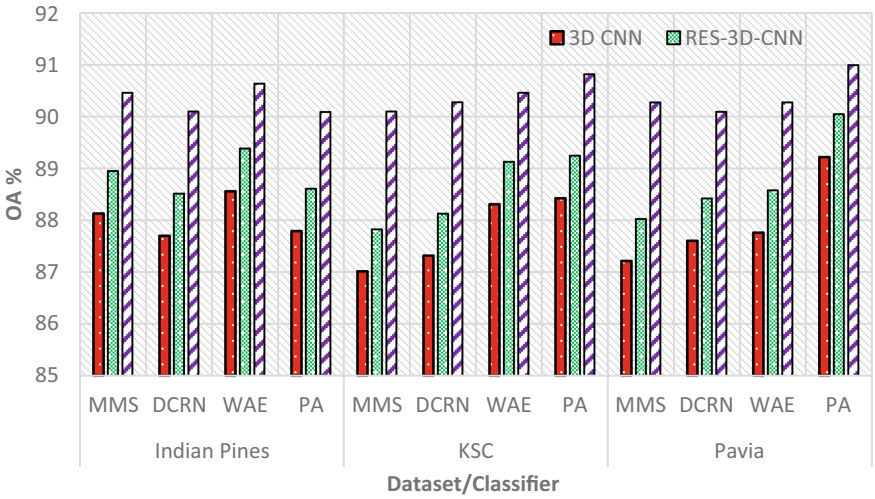


Fig. 2. Overall Accuracy with Conventional HSI Classifiers

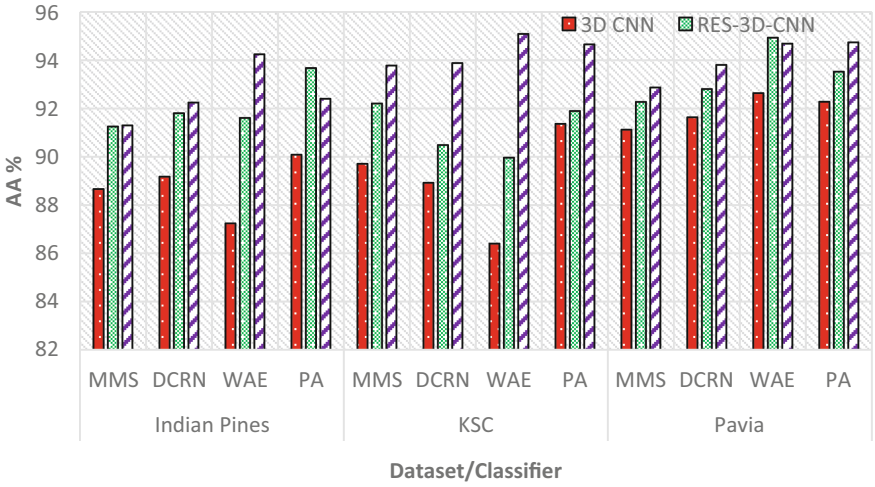


Fig. 3. Overall Accuracy with Conventional HSI Classifiers

Additionally, the proposed method is compared with various other existing methods that includes MMS, DCRN, WAE and PA as illustrated in Figs. 2, 3 and 4. Thus, it is seen that the proposed deep DSAE-RBM shows an increased rate of average accuracy, overall accuracy, and nominal kappa coefficient value than the other methods. Rather than fusing the spatial-spectral features, the reduction of dimensionality poses a greater adaptability for the deep learning models to attain an increased accuracy rates than the conventional models.

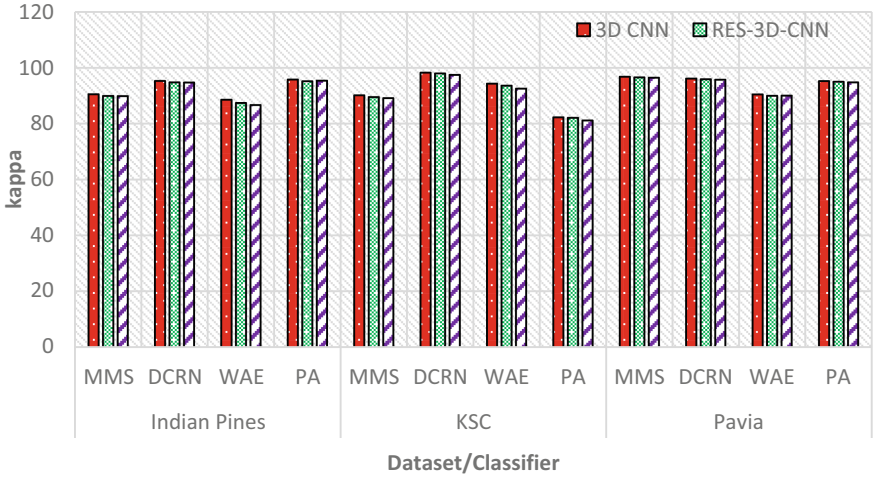


Fig. 4. Overall Accuracy with Conventional HSI Classifiers

5 Conclusions

In this paper, we developed a deep DSAE-RBM for classifying HIS from a large, augmented dataset. The data is supplemented by multiplying the dataset images using various noise addition levels. The deep DSAE-RBM is developed to classify the HSI image using the principle of denoising by RBM. The pre-processing model uses different band selection techniques like UGBS, CSBS, and MRBS that help select bands and perform classification in the presence of noise. The robust simulation to evaluate the model's efficacy shows that the proposed deep DSAE-RBM achieves an OA rate of 92.62% in the fact of noise and 77.47% in the absence of noise. The increasing noise level in dB shows that in the presence of Local shift noise, Multiplicative noise, and AWGN, the accuracy rates are higher in the proposed classifier than in the other classifiers. Among these three noises, the robustness of deep DSAE-RBM is better than other noise variants and levels in the presence of Local shift noise.

References

1. Li, W., Wu, G., Zhang, F., Du, Q.: Hyperspectral image classification using deep pixel-pair features. IEEE Trans. Geosci. Remote Sens. **55**(2), 844–853 (2016)
2. Ran, L., Zhang, Y., Wei, W., Zhang, Q.: A hyperspectral image classification framework with spatial pixel pair features. Sensors **17**(10), 2421 (2017)
3. Zhong, Z., Li, J., Luo, Z., Chapman, M.: Spectral–spatial residual network for hyperspectral image classification: a 3-D deep learning framework. IEEE Trans. Geosci. Remote Sens. **56**(2), 847–858 (2017)
4. Liu, X., Sun, Q., Meng, Y., Fu, M., Bourennane, S.: Hyperspectral image classification based on parameter-optimized 3D-CNNs combined with transfer learning and virtual samples. Remote Sens. **10**(9), 1425 (2018)
5. Ouyang, N., Zhu, T., Lin, L.: A convolutional neural network trained by joint loss for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. **16**(3), 457–461 (2018)

6. Demertzis, K., Iliadis, L., Pimenidis, E., Kikiras, P.: Variational restricted Boltzmann machines to automated anomaly detection. Neural Comput. Appl. 1–14 (2022). https://doi.org/10.1007/s00521-022-07060-4
7. Zhang, Y., Xia, J., Jiang, B.: REANN: A PyTorch-based end-to-end multi-functional deep neural network package for molecular, reactive, and periodic systems. J. Chem. Phys. **156**(11), 114801 (2022)
8. Bahraini, T., Azimpour, P., Yazdi, H.S.: Modified-mean-shift-based noisy label detection for hyperspectral image classification. Comput. Geosci. **155**, 104843 (2021)
9. Shi, Q., Tang, X., Yang, T., Liu, R., Zhang, L.: Hyperspectral image denoising using a 3-D attention denoising network. IEEE Trans. Geosci. Remote Sens. **59**(12), 10348–10363 (2021)
10. Xu, Y., et al.: Dual-channel residual network for hyperspectral image classification with noisy labels. IEEE Trans. Geosci. Remote Sens. **60**, 1–11 (2021)
11. Ghasrodashti, E.K., Sharma, N.: Hyperspectral image classification using an extended auto-encoder method. Signal Process. Image Commun. **92**, 116111 (2021)
12. Miclea, A.V., Terebes, R.M., Meza, S., Cislariu, M.: On spectral-spatial classification of hyperspectral images using image denoising and enhancement techniques, wavelet transforms and controlled data set partitioning. Remote Sensing **14**(6), 1475 (2022)

Prediction Type of Codon Effect in Each Disease Based on Intelligent Data Analysis Techniques

Zena A. Kadhuim[1] and Samaher Al-Janabi[2(✉)]

[1] Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

[2] Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq

Samaher@itnet.uobabylon.edu.iq

Abstract. To determine the codon usage effect on protein expression genome-wide, we performed whole-proteome quantitative analyses of FFST and LSTM whole-cell extract by mass spectrometry experiments. These analyses led to the identification and quantification proteins. Five human diseases are due to an excessive number of "cytosine (C), adenine (A), guanine(G)" as (i.e., CAG)repeats in the coding regions of five different genes. We have analyzed the repeat regions in four of these genes from nonhuman primates, which are not known to suffer from the diseases. These primates have CAG repeats at the same sites as in human alleles, and there is similar polymorphism of repeat number, but this number is smaller than in the human genes. In some of the genes, the segment of poly (CAG) has expanded in nonhuman primates, but the process has advanced further in the human lineage than in other primate lineages, thereby predisposing to diseases of CAG reiteration. Adjacent to stretches of homogeneous present-day codon repeats, previously existing codons of the same kind have undergone nucleotide substitutions with high frequency. Where these lead to amino acid substitutions, the effect will be to reduce the length of the original homopolymer stretch in the protein. In addition, RNA-sequencing (seq) analysis of the mRNA was performed to determine correlations between mRNA levels with codon usage biases. To determine the codon usage bias of genes, the codon bias index (CBI) for every protein-coding gene in the genome was calculated. CBI ranges from −1, indicating that all codons within a gene are nonpreferred, to +1, indicating that all codons are the most preferred, with a value of 0 indicative of random use. Because CBI estimates the codon bias for each gene rather than for individual codons, the relative codon biases of different genes can be compared.

Keywords: Codon DNA · Protein · Information Gain · LSTM

1 Introduction

A Disease is an extraneous condition that affects the body's organs with sporadic damage, and its functions stop working either temporarily or for a long time [1]. Recently many disease present like (COVID19, hemorrhagic fever ([2]. It is even estimated that

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 222–236, 2023.
https://doi.org/10.1007/978-3-031-27409-1_20

the number of patients with the aforementioned diseases is three million patient daily. The reason behind the tendency of effected disease is people not apply protective equipment and also not eat an healthy food [3]. The human need to protect their body against this deathly disease. All human has mRNA (Ribonucleic Acid) It is a complex compound with a high molecular weight that is involved in the protein synthesis process inside the cell [4]. Each mRNA sequence contain number of limited codon that effect directly on indirectly to disease caused to human [5]. Intelligent Data Analysis (IDA) is an interdisciplinary of study that focuses on extracting meaningful knowledge from data using techniques from artificial intelligence, high-performance computing, pattern recognition, and statistics [6]. The IDA process include three primary step, first must work on problem from real word and understand both problem and parameter of these problem second build model for this problem (clustering, classification, prediction, optimization, etc.) and evaluate the result. Finally Interprets the results so that they are understandable by all specialists and no specialists [7]. Data analysis is divided into four types Descriptive Analysis, Diagnostic Analysis, Predictive Analysis and Prescriptive Analysis. Descriptive analysis is a sort of data analysis that helps to explain, show, or summarize data points in a constructive way so, that patterns can develop that satisfy all of data conditions. While Diagnostic Analysis is the technique of using data to determine the origins of trends and correlations between variables is known as diagnostic analytics. After using descriptive analytics to find trends, it might be seen as a logical next step. Predictive analytics is a group of statistical approaches that evaluate current and historical data to, generate predictions about future or otherwise unknown events. It includes data mining, predictive modelling, and machine learning. Finally Prescriptive analytics is after data analytics maturity stage, for better decision in appropriate time. Different type of data analysis has been introduced with their used in different field and advantage of it. Main aim of intelligent Data analysis is to extract knowledge from data. Prediction is a data analysis task for predicting a value not known of target feature, we know the prediction techniques split based on the scientific field into two fields; prediction techniques related to data mining and prediction related to deep learning (i.e., Neuron computing Techniques) [8]. Different type of data analysis has been introduced with their used in different field and advantage of it. Aim of prediction is the process of making estimates for the future on the basis of past and present data and its impact on analysing trends. Bioinformatics is a sub discipline of biology and computer science, concerned with the extract, storage, analysis, and dissemination of biological data, for manage data in such a way that it, allows easy access to the existing information and to, submit new entries as they are produced and developing technological tools that help analyse data biology. Bioinformatics encompasses a wide range of disciplines, including Drug Designing, Genomics, Proteomics, System Biology, Machine Learning, Advanced Algorithms for Bioinformatics, Structural Biology, Computational Biology, and many others. Bioinformatics is consisted of complex DNA and amino acid sequences called protein after extracting from DNA. Bioinformatics is itself a great area of research at present [9].

The reset of paper summarized as: related works are explained in Sect. 2. In Sect. 3 present the theoretical background on techniques use, Sect. 4 show the methodology

of this work. Section 5 shows the results and discussion. Finally, Sect. 6 states the conclusions and future work of this model.

2 Related Work

Protein prediction is one of the most important concerns that directly affect people's lives and the continuance of a healthy lifestyle in general. The goal of this method is to establish a prediction of a mount of disease method for dealing with multiple type of disease and stop it later. Therefore, many researchers work on this filed summarized below.

In [10], Ahmed et al., implement a new model based on artificial intelligence to perform genome sequence analysis of human that infected by COVID-19 and other viruses that like COVID-19 example SARS and MERS and Ebola and middle east respiratory syndrome. The system helps to get important information from the genome sequences of different viruses. This done by extracting information of COVID-19 and perform comparative data analysis to original RNA sequences to detect gene continue virus and their frequency by count of amino acids.at end of method, classifier-based machine learning called support vector machine used to classify different genome sequences. The proposed work uses (accuracy) for measuring performance of algorithm. This study implements high accuracy level (97%) for COVID-19.

In [11], Narmadha and Pravin introduce a method called graph coloring-deep neural network to predict influence protein in infectious diseases. The method starts by coloring the protein that have more interaction in the network represent disease. The main aim of this method is to development of drug and early diagnosis and treatment of the disease. They used various datasets for different diseases (cancer, diabetes, asthma and HPV viral infection). The result show that for predicting cancer 92.328% accuracy, 93.121% precision, 92.874% recall and f measurement 91.102%.

In [12], Asad Khan et al. proposed a new method to predict the existence of m6A in RNA sequences this method used statistical and chemical properties of nucleotides and called (m6A-pred predictor) and uses random forest classifier to predict m6A by identify features that was discriminative. The proposed work uses (accuracy) and (Mathew correlation coefficient values) for measuring performance of algorithm. This study show high accuracy level (78.58%) with Mathew correlation coefficient values (79.65%) of 0.5717. Our work similarity with this work in evaluation measurement but differ from method used to discover protein based on intelligent data analysis and techniques used.

The authors in [13] predict protein position of S-sulfenylation by a new method called SulSite-GTB. This protein involved in a different biological processes important for life like (signaling of cell, increasing stress). The methods summarize by four steps: combine amino acid composition, dipeptide composition, grouped weight encoding, K nearest neighbors, position-specific amino acid propensity, position-weighted amino acid composition, and pseudo-position specific score matrix feature extraction. Secondly, to process the data on class imbalance, To remove the redundant and unnecessary features, the least absolute shrinkage and selection operator (LASSO) is used. Finally, to predict sulfenylation sites, the best feature subset is fed into a gradient tree boosting classifier. Prediction accuracy is 92.86%.

As for the work in [14], Athilakshmi et al. design a method using deep learning to discover anomaly causing genes in mRNA sequences cause brain disorders such as Alzheimer's disease and Parkinson's disease (Table 1).

3 Theoretical Background

The following section show the main concepts used in this paper.

3.1 Fast Frequency Sub-graph Mining Algorithm (FFSMA)

Frequency Sub-graph mining algorithm is an algorithm-based pattern growth for extracting all frequent sub-graph from data and then accepting the most frequent sub-graph according to some minimum support [15]. Many algorithms of FSM Work with graph, FFSM is a Fast Frequent Sub-graph mining algorithm its outperform all FSM algorithms include (gSpan, CloGraMi and Hybrid tree miner) because of two reason, first its contain incidence matrix normalize that compute each node and its connected edge second for each sub-matrix add all possible edge that have not found in it [16, 25].

3.2 Feature Selection

The act of selecting a subset of pertinent features, or variables, from a larger data collection in order to build models is known as feature selection. Other names for feature selection include variable selection, attribute selection, and variable subset selection [17]. It makes the machine learning algorithm less complex, allowing faster training, and is simpler to understand. If the proper subset is selected, a model's accuracy is increased. Finally, feature selection minimizes over fitting [17] 18. Entropy is a metric for a data-generating function's diversity or randomness. Full entropy data is utterly random, and no discernible patterns can be discovered. Data with low entropy offers the ability or potential to forecast newly created values [19]. On the other side Information gain, which is the decrease in entropy or surprise caused by altering a dataset. By comparing the entropy of the dataset before and after a transformation, information gain is computed, Entropy can be used to determine how a change to the dataset, such as the distribution of classes, affects the dataset's purity using information gain. A lower entropy indicates greater purity or decreased surprise [20, 21].

The connection between two variables is known as a correlation. Using the features of the input data, we can forecast our target variable using these variables. Based on their association, many variables are put together in this metric [22].

3.3 Deep Prediction Neuro Computing Techniques

Prediction is a method that used to predict some value or features according to founded once [17] prediction techniques, either related to data mining (SVM [23], LR [24], RF [25]) or related to Deep prediction neuro computing techniques (LSTM [26], BiLSTM [27], MLSTM [28], RNN [29], GRU [30]. Prediction in deep neuro techniques is outperform data mining techniques in term of accuracy but on the other hand take long time for predict an accurate result [30].

Table 1. Summarized on literate survey

Author	Data set/Database	Preprocessing	Method	Evaluation
Wang [13]	Independent test set (protein sequence) https://github.com/QUST-AIBBDRC/SulSite-GTB/.	Feature en-coding	SulSite-GTB	Accuracy
Athilakshmi et al. [14]	Gene Sets of Alzheimer's and Parkinson's http://www.genecards.org	Feature en-coding	DL based Anomaly Detection	MSE
Khan et. al. [12]	RNA sequences	Feature ex-traction	m6A-predection	Accuracy and Mathew correlation coeffi-cient
Narmada and Pravin [11]	Protein sequence Collection of PPI (string DB, IntAct, DIP) database	Segmentation	graph coloring-deep neu-ral network	Confusion matrix
Imran Ahmed et al. [10]	DNA sequences	Feature ex-traction	ML	Accuracy

4 Proposed Method

The data set used in our work is Codon usage Data Set published in machine learning respiratory at https://archive.ics.uci.edu/ml/datasets/Codon+usage#. There is 64 codon of Amino Acid related to 13028 disease. Each one of these Amino Acid have a percentage of bias in each disease [31]. In this paper we show how grouping codons effect positively in term of reduce computing and time for entering to farther stages.

4.1 Step of Proposed Method

Amino acids produce the taste of food and keep us healthy. For example, they are used for sports nutrition, medicine, beauty products and to reduce calorie intake. In this proposed method. We implement some of intelligent data analysis techniques for reduce the computation of working on dataset in [31]. All work summarized under the following main points:

- For all CTUG dataset, data pre-processing is performed to group feature by feature selection. We calculate information gain for all feature after convert all description feature to numeric value feature.
- Calculate Minkonisky distance for result from one for creating group between features.
- Grouping features into sixteen group four nodes for each sub-group.
- Enter all group to FFSMA for delete duplication subgroup for each group.
- The confidence of relation is computed to see how correct rule of result extracted.
- Finally, the results enter to long short-term memory to see how result valid

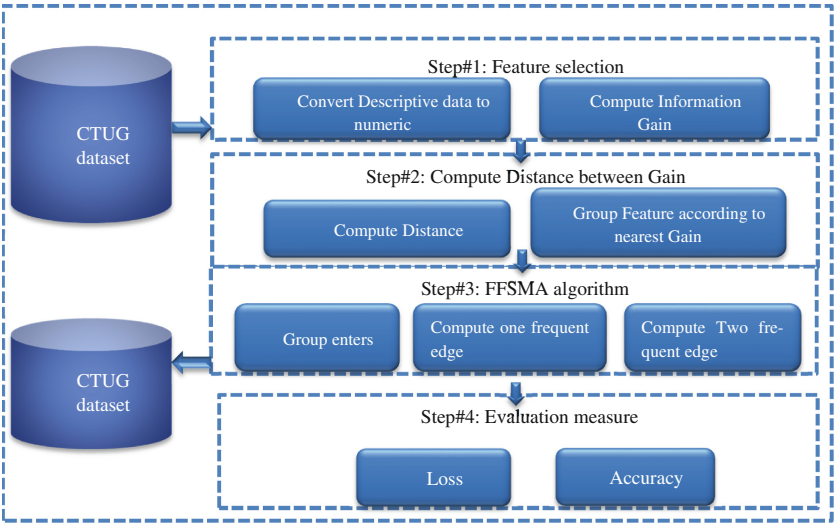


Fig. 1. Proposed New MVA-FFSM Method

- Our step of work is implemented via several stages to finally effect of each codon the details of proposed algorithm explain in algorithm #1 and Fig. 1.

Algorithm #1

Input: CUTG dataset //set of all codon related to each disease of kingdom entitled in dataset.

Output: // non-frequent records

Begin

For all element in CUTG dataset

Perform a feature selection

Compute information Gain(feature, Target)

Remove features have less information gain

End for

For all features in CUTG dataset

Compute minkonisky distance for between inf.G

Group CUTG base on minkonisky distance (nearest Gain)

End for

For all features in group of dataset perform FFSMA

While F(k) > 0 \\ for candidate generation set to← Φ

For i= 1 to n

TFG(Gi,K+1)= Φ

for each k-edge frequent sub graph gn(Gi,)∈FG(Gi,)

N← list of all edge (e) not connected to GI

For each e in N do

gn(G,k+1) ← gn(G,k+1) U e

If gn(G,k+1) not belong (Gi,K+1)

TFG(Gi,K+1)= TFG(Gi,K+1) U gn(G,k+1)

End

End

End

***For* each gn(G,k+1) in TFG(Gi,k+1)**

If gn(G,k+1) not belong TFGC(G,k+1) then

TFGC(G,k+1)= TFGC(G,k+1)U n(G,k+1)

Freq_gn(G,k+1)=1

Else

Freq_gn(G,k+1)= Freq_gn(G,k+1)+1

End

F(K+1) =0

For each gn(G, k+1) in TFGC(k+1)

if Freq_gn(G, k+1)> minisup then

FGS(GDB)= FGS(GDB) U {gn(G, k+1)}

F(k+1)= F(k+1)+1

End

End

For i= 1 to length of N

FG(G, k+1)=Φ

For each gn(Gi, k+1) in TFGC(k+1) do

if gn(Gi, k+1)∈FGS(GDB) then

FG(G, k+1)= FG(G, k+1) U {gn(G, k+1)}

End

End

K=k+1

End for

Delete non maximum frequent sub graph

4.2 Data Preprocessing (Feature Selection)

All CTUG dataset is entered to the system. For our work we need only specific filed for work on it, this field is name of disease and all 64 amino acid to compute there effect to disease. So, for this we show the 64 codon feature by compute information gain for each feature with target that represent Disease (SpieseName).

First, we calculate entropy for each 64 features:

$$\textbf{Entropy} = -\sum_{n=1}^{l} \boldsymbol{pi} * \boldsymbol{log_2 pi} \tag{1}$$

Then compute information gain for each features with Target (features, Disease):

$$\boldsymbol{Information\ Gain} = \boldsymbol{Target\ entropy} - \boldsymbol{Entropy\ child} \tag{2}$$

$$\boldsymbol{Entropy\ child} = \sum E\ splits * W \tag{3}$$

where:

Pi	probability of element in column
E splits	calculate Entropy for splits depend on feature selected
W	how many element found in splits
Target Entropy	Entropy of target feature

4.3 Selecting Different Records

After preparing data, we have 13028 record and each one of them have 64 codon effect to it. In this work we work on more frequent disease to see the effect of 64 codon. Before FFSMA work we group features according to nearest value of information gain by minkonisky distance.

$$\mathrm{M(D)} = \left(\sum \mathrm{i} = 1\mathrm{n}|\mathrm{Xi} - \mathrm{Yi}|\mathrm{p}\right)1/\mathrm{p} \tag{4}$$

where:

xi	vector one of value
y1	vector one of value
p	default value of squirt.

Then group 64 feature into 16 group of 4 nodes according to value of information gain from lowest to highest:

[['UAG', 'UAA', 'UGA', 'CGG'], ['CGA', 'UGC', 'AGG', 'UGU'], ['UCG', 'CGU', 'ACG', 'CCG'], ['AGU', 'AGC', 'CAU', 'GGG'], ['UGG', 'CGC', 'AGA', 'CAC'], ['CCU', 'UAC', 'UCC', 'GCG'], ['CCC', 'UCU', 'GUA', 'AUG'], ['ACU', 'UUG', 'CUA', 'UCA'], ['CCA', 'GUC', 'GCA', 'GGA'], ['GGU', 'CUU', 'AAC', 'GUU'], ['GCU', 'CAA', 'CAG', 'UAU'], ['GUG', 'UUC', 'ACA', 'GGC'], ['AUA',

'CUC', 'ACC', 'CUG'], ['UUA', 'GAC', 'AUC', 'AAG'], ['GAU', 'UUU', 'AAU', 'GAG'], ['GAA', 'GCC', 'AAA', 'AUU']]

Then after group it, each group is a sub graph enter to FFSMA and dealing with each column as value of node and compute frequent edge as follow:

Enter **Sub-graph 1** that have 4 node to FFSMA algorithm:

- N = AUU, GCC, AAA, GAA, whereas:

First compute One frequent edge:

- E1 = {AUU, GCC} at Attrition T1,
- E2 = {GCC, AAA} at Attrition T1,
- E3 = {AAA, GAA} at Attrition T1,

Second compute Two frequent edge:

- E1E2 = {AUU, GCC, AAA} at Attrition T2,
- E2E3 = {GCC, AAA, GAA} at Attrition T2,

Third compute Three frequent edge:

- E1E2E3 = {AUU, GCC, AAA, GAA} at Attrition T3,

Results of Remove Duplication Sup-graph FFSMA.

Origin Sub-graph—Results Sub-graph = Deleted Duplication Sub-graph.

5 Results and Discussion

The result of our work represents how each feature selection can reduce a computation of our work according to Gain of each feature and scale it, Table 2 shows the result of our features with normalized disease.

In Table 2, there are five columns: in the first column are the main characteristics in the dataset, which represent the codons of each disease, which are 64 codons found in all creatures that are associated with 13,028 diseases, in the second column the entropy values for each codon from among the 64 codons, and the third column It represents the value of the information gain in relation to the codon and its association with each disease, the next column is the conversion of the entropy values with the scaling function between (1 and −1) and the last column is the normalization of the information gain value to be between (1 and 0). Figure 2 represent important codon related to each disease that in range (1, 0).

Table 2. Illustrate the relation between feature and disease in term of Gain, Correlation

Feature	Entropy	Gain	Correlation	GN	GS
UUU	11.86638	5.14777	0.148125	0.748361	0.496721
UUC	11.65497	4.94355	0.292521	0.718672	0.437344
UUA	11.65547	5.02011	0.194589	0.729802	0.459604
UUG	11.17456	4.64629	−0.11831	0.675458	0.350915
CUU	11.46138	4.75966	0.255333	0.691939	0.383878
CUC	11.65372	4.97407	0.226646	0.723109	0.446218
CUA	11.3578	4.6904	0.410942	0.68187	0.36374
CUG	11.63117	4.9805	−0.17578	0.724044	0.448087
AUU	12.04675	5.32029	0.221926	0.773441	0.546881
AUC	11.83587	5.11225	0.243239	0.743197	0.486394
AUA	11.64205	4.97307	0.325031	0.722963	0.445927
AUG	11.26315	4.60613	−0.29516	0.669619	0.339238
GUU	11.46679	4.79023	−0.14321	0.696383	0.392766
GUC	11.37616	4.7118	−0.11657	0.684981	0.369962
GUA	11.27364	4.60142	0.167697	0.668935	0.337869
GUG	11.45692	4.92349	−0.28198	0.715756	0.431511
GCU	11.49003	4.79992	−0.08557	0.697792	0.395583
GCC	11.94552	5.24025	0.021014	0.761805	0.52361
GCA	11.40672	4.71649	0.067029	0.685663	0.371326
GCG	10.83358	4.50338	−0.3017	0.654682	0.309364
CCU	11.04018	4.37547	0.029213	0.636087	0.272174
CCC	11.24163	4.57031	0.185779	0.664412	0.328824
CCA	11.37799	4.69991	0.241677	0.683253	0.366505
CCG	10.56902	4.21152	−0.27587	0.612253	0.224505
UGG	10.77833	4.2783	−0.26737	0.621961	0.243921
GGU	11.40249	4.75339	−0.22887	0.691027	0.382055
GGC	11.4311	4.97036	−0.17221	0.722569	0.445139
GGA	11.62821	4.73219	0.231896	0.687945	0.375891
GGG	10.88869	4.26292	−0.05329	0.619725	0.23945
UCU	11.27058	4.5941	0.107305	0.66787	0.335741
UCC	11.17281	4.48813	0.265926	0.652465	0.30493
UCA	11.372	4.69187	0.274799	0.682084	0.364168

(*continued*)

Table 2. (*continued*)

Feature	Entropy	Gain	Correlation	GN	GS
UCG	10.36185	4.02861	−0.20983	0.585662	0.171324
AGU	10.68159	4.23582	−0.15994	0.615785	0.23157
AGC	10.8407	4.25646	−0.12671	0.618786	0.237571
ACU	11.29789	4.61877	0.031038	0.671457	0.342914
ACC	11.68092	4.97473	0.132544	0.723205	0.446409
ACA	11.66511	4.96873	0.226249	0.722332	0.444665
ACG	10.57329	4.1723	−0.29004	0.606551	0.213102
UAU	11.57677	4.90166	−0.02509	0.712582	0.425164
UAC	11.15769	4.48105	0.029102	0.651436	0.302871
CAA	11.48509	4.81417	0.015902	0.699863	0.399726
CAG	11.34212	4.84099	−0.30663	0.703762	0.407524
AAU	11.86419	5.16431	−0.09713	0.750765	0.50153
AAC	11.47363	4.77758	0.038796	0.694544	0.389088
UGU	10.30101	3.96058	−0.09584	0.575772	0.151544
UGC	10.40001	3.9367	−0.04899	0.5723	0.144601
CAU	10.89653	4.2591	0.002728	0.61917	0.238339
CAC	11.01853	4.36198	0.186279	0.634126	0.268252
AAA	11.99542	5.31061	−0.08887	0.772034	0.544067
AAG	11.63019	5.11527	−0.23254	0.743636	0.487272
CGU	10.49915	4.09973	−0.25308	0.596001	0.192002
CGC	10.73903	4.29636	−0.27613	0.624586	0.249172
CGA	10.37923	3.82518	0.301776	0.556088	0.112176
CGG	9.7358	3.65667	−0.1815	0.531591	0.063182
AGA	10.12659	4.2995	−0.1104	0.625043	0.250085
AGG	9.52235	3.9369	−0.11753	0.572329	0.144659
GAU	11.75633	5.14641	−0.34814	0.748163	0.496326
GAC	11.7045	5.04355	−0.25787	0.733209	0.466419
GAA	11.87286	5.19904	−0.22795	0.755814	0.511628
GAG	11.69344	5.16498	−0.27195	0.750862	0.501725
UAA	8.00838	2.22431	0.111846	0.323361	−0.35328
UAG	5.96591	1.36074	0.045477	0.197818	−0.60436
UGA	8.51548	2.7087	0.442073	0.393779	−0.21244

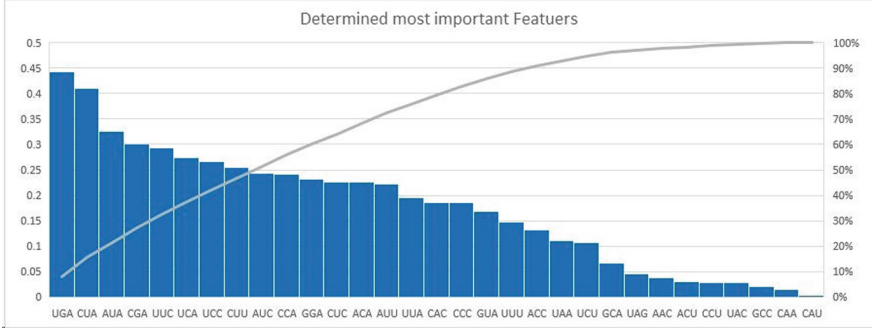


Fig. 2. Relation between codon to target disease

Then all feature must Grouped that represent a graph to enter to FFSMA algorithm, All 16 group of 64 codons from lowest to highs gain:

G1 ['UAG', 1.36074, 'UAA', 2.22431, 'UGA', 2.7087, 'CGG', 3.65667]
G2 ['CGA', 3.82518, 'UGC', 3.9367, 'AGG', 3.9369, 'UGU', 3.96058]
G3 ['UCG', 4.02861, 'CGU', 4.09973, 'ACG', 4.1723, 'CCG', 4.21152]
G4 ['AGU', 4.23582, 'AGC', 4.25646, 'CAU', 4.2591, 'GGG', 4.26292]
G5 ['UGG', 4.2783, 'CGC', 4.29636, 'AGA', 4.2995, 'CAC', 4.36198]
G6 ['CCU', 4.37547, 'UAC', 4.48105, 'UCC', 4.48813, 'GCG', 4.50338]
G7 ['CCC', 4.57031, 'UCU', 4.5941, 'GUA', 4.60142, 'AUG', 4.60613]
G8 ['ACU', 4.61877, 'UUG', 4.64629, 'CUA', 4.6904, 'UCA', 4.69187]
G9 ['CCA', 4.69991, 'GUC', 4.7118, 'GCA', 4.71649, 'GGA', 4.73219]
G10 ['GGU', 4.75339, 'CUU', 4.75966, 'AAC', 4.77758, 'GUU', 4.79023]
G11 ['GCU', 4.79992, 'CAA', 4.81417, 'CAG', 4.84099, 'UAU', 4.90166]
G12 ['GUG', 4.92349, 'UUC', 4.94355, 'ACA', 4.96873, 'GGC', 4.97036]
G13 ['AUA', 4.97307, 'CUC', 4.97407, 'ACC', 4.97473, 'CUG', 4.9805]
G14 ['UUA', 5.02011, 'GAC', 5.04355, 'AUC', 5.11225, 'AAG', 5.11527]
G15 ['GAU', 5.14641, 'UUU', 5.14777, 'AAU', 5.16431, 'GAG', 5.16498]
G16 ['GAA', 5.19904, 'GCC', 5.24025, 'AAA', 5.31061, 'AUU', 5.32029].

All sixteen-sub group enter to frequency sub graph mining algorithm (FFSMA) to remove duplication sub-graph. FFSMA results reduce the number of rows for whole dataset by remove frequent edge and save only different row that effect to each different disease and testing by association rule mining of dataset. Also the time of using techniques of preprocessing dataset is reduced compare with time of working to all dataset from 0.22 to 0.16 s.

Because of sensitive dataset, we select only rule that have high relation to feature. So in this case we select second rule and so forth.

The original dataset is [13028 rows × 65 columns of features and normalized Disease]. The rule for each group is:

G1 **entered** [13028 rows × 4 columns] **out is:** [12037 rows × 4 columns]
G2 **entered** [13028 rows × 4 columns] **out is:** [12452 rows × 4 columns]

G3 **entered** [13028 rows × 4 columns] **out is:** [12615 rows × 4 columns]
G4 **entered** [13028 rows × 4 columns] **out is:** [12769 rows × 4 columns]
G5 **entered** [13028 rows × 4 columns] **out is:** [12701 rows × 4 columns]
G6 **entered** [13028 rows × 4 columns] **out is:** [12811 rows × 4 columns]
G7 **entered** [13028 rows × 4 columns] **out is:** [12826 rows × 4 columns]
G8 **entered** [13028 rows × 4 columns] **out is:** [12852 rows × 4 columns]
G9 **entered** [13028 rows × 4 columns] **out is:** [12814 rows × 4 columns]
G10 **entered** [13028 rows × 4 columns] **out is:** [12839 rows × 4 columns]
G11 **entered** [13028 rows × 4 columns] **out is:** [12818 rows × 4 columns]
G12 **entered** [13028 rows × 4 columns] **out is:** [12837 rows × 4 columns]
G13 **entered** [13028 rows × 4 columns] **out is:** [12814 rows × 4 columns]
G14 **entered** [13028 rows × 4 columns] **out is:** [12849 rows × 4 columns]
G15 **entered** [13028 rows × 4 columns] **out is:** [12863 rows × 4 columns]
G16 **entered** [13028 rows × 4 columns] **out is:** [12866 rows × 4 columns]

Finally, results of FFSMA must be trained by Long Short-Term Memory (LSTM) that produce a result according to different splitting (Table 3).

Table 3. Measurements criteria

Rate of Training and Testing Dataset	MSE (%)	ACCURACY (%)
50 train, 50 test	0.003	94.2431
70 train, 30 test	0.0019	94.678
90 train, 10 test	0.0005	96.162

We see how accuracy is increase according to splitting between multiple value of training and testing.

6 Conclusion

To determine the codon usage bias of genes, the codon bias index (CBI) for every protein-coding gene in the genome was calculated. CBI ranges from −1, indicating that all codons within a gene are nonpreferred, to +1, indicating that all codons are the most preferred, with a value of 0 indicative of random use. Because CBI estimates the codon bias for each gene rather than for individual codons, the relative codon biases of different genes can be compared. The accuracy of proposed method is 96.162% while MSE is 0.0005.

References

1. Al-Janabi, S.: Overcoming the main challenges of knowledge discovery through tendency to the intelligent data analysis. Int. Conf. Data Anal. Bus. Ind. (ICDABI) **2021**, 286–294 (2021)
2. Kadhuim, Z.A., Al-Janabi, S.: Intelligent deep analysis of DNA sequences based on FFGM to enhancement the performance and reduce the computation. Egypt. Inform. J. **24**(2), 173–190 (2023). https://doi.org/10.1016/j.eij.2023.02.004
3. Vitiello, A., Ferrara, F.: Brief review of the mRNA vaccines COVID-19. Inflammopharmacology **29**(3), 645–649 (2021). https://doi.org/10.1007/s10787-021-00811-0
4. Toor, R., Chana, I.: Exploring diet associations with Covid-19 and other diseases: a network analysis–based approach. Med. Biol. Eng. Compu. **60**(4), 991–1013 (2022). https://doi.org/10.1007/s11517-022-02505-3
5. Kadhuim, Z.A., Al-Janabi, S.: Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis. Results Eng. **17**, 100847 (2023). https://doi.org/10.1016/j.rineng.2022.100847
6. Nambou, K., Anakpa, M., Tong, Y.S.: Human genes with codon usage bias similar to that of the nonstructural protein 1 gene of influenza A viruses are conjointly involved in the infectious pathogenesis of influenza A viruses. Genetica 1–19 (2022). https://doi.org/10.1007/s10709-022-00155-9
7. Al-Janabi, S., Al-Janabi, Z.: Development of deep learning method for predicting DC power based on renewable solar energy and multi-parameters function. Neural Comput. Appl. (2023). https://doi.org/10.1007/s00521-023-08480-6
8. Al-Janabi, S., Al-Barmani, Z.: Intelligent multi-level analytics of soft computing approach to predict water quality index (IM12CP-WQI). Soft Comput. (2023). https://doi.org/10.1007/s00500-023-07953-z
9. Li, Q., Zhang, L., Xu, L., et al.: Identification and classification of promoters using the attention mechanism based on long short-term memory. Front. Comput. Sci. **16**, 164348 (2022)
10. Ahmed, I., Jeon, G.: Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. Interdisc. Sci. Comput. Life Sci. 1–16 (2021). https://doi.org/10.1007/s12539-021-00465-0
11. Narmadha, D., Pravin, A.: An intelligent computer-aided approach for target protein prediction in infectious diseases. Soft. Comput. **24**(19), 14707–14720 (2020). https://doi.org/10.1007/s00500-020-04815-w
12. Khan, A., Rehman, H.U., Habib, U., Ijaz, U.: Detecting N6-methyladenosine sites from RNA transcriptomes using random forest. J. Comput. Sci. **4**,(2020). https://doi.org/10.1016/j.jocss.2020.101238
13. Wang, M., Song, L., Zhang, Y., Gao, H., Yan, L., Yu, B.: Malsite-deep: prediction of protein malonylation sites through deep learning and multi-information fusion based on NearMiss-2 strategy. Knowl. Based Syst. **240**, 108191 (2022)
14. Athilakshmi, R., Jacob, S.G., Rajavel, R.: Protein sequence based anomaly detection for neuro-degenerative disorders through deep learning techniques. In: Peter, J.D., Alavi, A.H., Javadi, B. (eds.) Advances in Big Data and Cloud Computing. AISC, vol. 750, pp. 547–554. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1882-5_48
15. Cheng, H., Yu, J.X.: Graph mining. In: Liu, L., Özsu, M.T. (Eds.) Encyclopedia of Database Systems. Springer, New York, (2018)
16. Mohammed, G.S., Al-Janabi, S.: An innovative synthesis of optmization techniques (FDIRE GSK) for generation electrical renewable energy from natural resources. Results Eng. **16**, 100637 (2022). https://doi.org/10.1016/j.rineng.2022.100637
17. Kadhim, A.I.: Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. In: 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 124–128

18. Wang, S., Tang, J., Liu, H.: Feature selection. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA (2017). https://doi.org/10.1007/978-1-4899-7687-1_101
19. Khan, M.A., Akram, T., Sharif, M., Javed, K., Raza, M., Saba, T.: An automated system for cucumber leaf diseased spot detection and classification using improved saliency method and deep features selection. Multimedia Tools Appl. **79**(25–26), 18627–18656 (2020). https://doi.org/10.1007/s11042-020-08726-8
20. Jia, W., Sun, M., Lian, J., Hou, S.: Feature dimensionality reduction: a review. Complex Intell. Syst. 1–31 (2022). https://doi.org/10.1007/s40747-021-00637-x
21. Rodriguez-Galiano, V., Luque-Espinar, J., Chica-Olmo, M., Mendes, M.P.: Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. Sci. Total Environ. **624**, 661–672 (2018)
22. Saqib, P., Qamar, U., Aslam, A., Ahmad, A.: Hybrid of filters and genetic algorithm-random forests based wrapper approach for feature selection and prediction. In: Intelligent Computing-Proceedings of the Computing Conference, vol. 998, pp. 190–199. Springer (2019)
23. Al-Janabi, S., Alkaim, A.: A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis. Egypt. Informatics J. **23**(2), 271–290 (2022). https://doi.org/10.1016/j.eij.2022.01.004
24. Liew, B.X.W., Kovacs, F.M., Rügamer, D., Royuela, A.: Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. Eur. Spine J. 1 (2022). https://doi.org/10.1007/s00586-022-07188-w
25. Hatwell, J., Gaber, M.M., Azad, R.M.A.: CHIRPS: Explaining random forest classification. Artif. Intell. Rev. **53**, 5747–5788 (2020)
26. Rodriguez-Galiano, V., Luque-Espinar, J., Chica-Olmo, M., Mendes, M.P.: Feature selection approaches for predictive modelling of foreseeing the principles of genome architecture. Nat. Rev. Genet. **23**, 2–3 (2022)
27. Liu, H., Zhou, M., Liu, Q.: An embedded feature selection method for imbalanced data classification. IEEE/CAA J. Autom. Sin. **6**, 703–715 (2019)
28. Lu, M.: Embedded feature selection accounting for unknown data heterogeneity. Expert Syst. Appl. **119** (2019)
29. Ansari, G., Ahmad, T., Doja, M.N.: Hybrid Filter-Wrapper feature selection method for sentiment classification. Arab. J. Sci. Eng. **44**, 9191–9208 (2019)
30. Jazayeri, A., Yang, C.: Frequent subgraph mining algorithms in static and temporal graph-transaction settings: a survey. IEEE Trans. Big Data (2021)
31. Khomtchouk, B.B.: Codon usage bias levels predict taxonomic identity and genetic composition (2020)



A Machine Learning-Based Traditional and Ensemble Technique for Predicting Breast Cancer

Aunik Hasan Mridul[1], Md. Jahidul Islam[1(✉)], Asifuzzaman Asif[2,2(✉)], Mushfiqur Rahman[1,1(✉)], and Mohammad Jahangir Alam[1,1(✉)]

[1] Daffodil International University, Dhaka, Bangladesh
{Aunik15-2732,Jahudul15-2753,Mushfiqur.cse,
Jahangir.cse}@diu.edu.bd
[2] Lovely Professional University, Phagwara, Panjab, India
asif.11900157@lpu.in

Abstract. Breast cancer is a physical disease and increasing in recent years. The topic is known widely in the recent world. Most women are suffering from problem of breast cancer. The disease is measured by the differences between normal and affected area ratio and the rate of uncontrolled increase of the tissue. Many studies have been conducted in the past to predict and recognize breast cancer. We have found some good opportunities to improve the technique. We propose predicting the risks and making early awareness using effective algorithm models. Our proposed method can be easily implemented in real life and is suitable for easy breast cancer predictions. The dataset was collected from Kaggle. In our model, we have implemented some different classifiers named Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), and K-Nearest Classifier algorithms. Logistic Regression and Random Forest Classifier were performed well with 98.245% testing accuracy. Other algorithms like Gradient Boosting 91.228%, and K-Nearest 92.105% testing accuracy. We also used some different ensemble models to justify the performances. We have used Bagging LRB 94.736%, RFB 94.736%, GBB 95.614%, and KNB 92.105% accuracy, Boosting LRBO 96.491%, RFBO 99.122%, and GBBO 98.218% accuracy, and Voting algorithm LRGK with 95.614% accuracy. We have used hyper-parameter tuning in each classifier to assign the best parameters. The experimental study indicates breast cancer predictions with a higher degree of accuracy and evaluated the findings of other current studies, RFBO with 99.122% accuracy being the best performance.

Keywords: Breast cancer · Prediction · Machine Learning · Algorithms · Ensemble Model

1 Introduction

In the recent era, several tissues are being damaged or growing uncontrolled, known as cancer. When uncontrolled tissue or damaged tissue creates cancer in a woman's breast

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 237–248, 2023.
https://doi.org/10.1007/978-3-031-27409-1_21

it is known as breast cancer. The rate of this patient is increasing at a significant rate. But the main problem is to identify or recognize the damaged area at the time of diagnosis. Machine learning can be the best part of a significant aspect in predicting the presence of breast cancer from responsive health datasets by exploring several features and patient diagnosis records. In our work, we have explored the patient's diagnosis reports and found some important parameters to determine the disease. The dataset was about the shape and size of tissues in a woman's body and identifying the presence of cancer in her breast or not. Many other researchers have collaborated to use machine learning algorithms to identify the cancer tissue in the body. But their accuracy and technique were not suitable or smooth to predict breast cancer. To improve the prediction of breast cancer in a woman's body, we propose our technique to improve the accuracy rate. Two types of machine learning approaches are present. One of them is supervised and another is unsupervised. Supervised learning works with the data which is labeled and gives an output from input based on the example input-output pairs. The working data is training data from the dataset. Unsupervised learning works with the unlabeled data and creates the model to work with its patterns and information which was not detected previously.

2 Related Works

Some Machine Learning classifiers we have implemented for our breast cancer classification and they are suitable for our proposed work. The meaning of tree structure is Machine Learning algorithms that are based on decision tree models to run decision models [1, 2]. Researchers Rani and Dhenakaran have proposed models focused on Modified Neural Network (MNN) as make predictions of cancer tissue growth rate. The proposed model resulted in an accuracy of 97.80% [3]. Researcher Li et al. also modified an SVM classifier to predict the cancer tissue. The proposed model performed with an accuracy of 84.12%, specificity of 78.80%, and sensitivity of 2.86% [4]. Gomez-Flores and Hernandez-Lopez proposed a model to detect cancer tissue with an 82.0% AUC score [5]. Liu et al. developed an SVC model to acquire the classification of breast cancer tissue with 67.31% accuracy, 47.62% sensitivity, and 80.65% specificity [6]. Irfan et al. also proposed CNN and SVM models to classify breast cancer with a precision rate of about 98.9% [7]. SVM, AdaBoost, Naive Bayesian, K-NN, Perceptron, and Extreme Learning Machine models were proposed by Lahoura et al. with 98.68% accuracy, 91.30% recall, 90.54% precision, and 81.29% F1-score [8].

3 Classifier and Ensemble Models

In our study, we used Machine Learning (ML) based classifiers like Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (KN).

Logistic Regression

Logistic Regression (LR) is one kind of machine learning classifier algorithm where the class label has two categories, there are yes or no like a binary (0/1). Logistic regression is useful in discrete variables but it allows the mixed value of continuous variables and

discrete predictors [11]. The concept is shown in Fig. 1. Logistic Regression accepts the method of supervised machine learning. The basic Eq. (1) is shown below [10].

$$h\Theta(x) = \frac{1}{1} + e - (\beta o + \beta 1X) \tag{1}$$

$h\Theta(x)$	*is known as the result of the function, here* $0 \leq h\Theta(x) \geq 1$
$\beta 1$	*is known as the slope*
βo	*is the y-grab*
X	*is the variable that is independent*
$(\beta o + \beta 1X)$	*derived from the equation of a line* $Y\,(predicted) = (\beta o + \beta 1X) + error.$

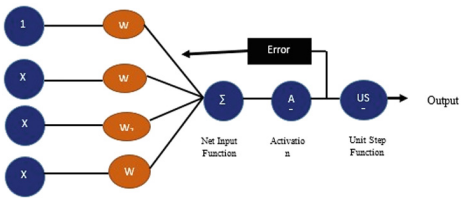


Fig. 1. Working Principle of Logistic Regression

3.1 Random Forest

Random Forest is one kind of machine learning-based classifier ensemble method that consists of different Decision Tree algorithms [12, 25, 28]. RF creates several multiple decision trees during the time of algorithm training to result in an optimal decision model which can result in the best accuracy than the single decision tree model. The concept is shown in Fig. 2.

But it is applicable in large datasets. The calculation of the mean of total decision tree algorithms is done with the Random Forest [13, 14]. The mean of two decision tree algorithms was calculated in the Random Forest algorithm (2).

$$j = \frac{1}{B} + \sum_{b=1}^{B} fb(X') \tag{2}$$

Concerning $X = \{x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n\}$ with respect to $Y = \{y_1, y_2, y_3, \ldots\ldots\ldots\ldots y_n\}$ with the lower to upper limit is 1 to B.

Sample $x' =$ mean of the sum of the prediction $\sum_{b=1}^{B} fb(X')$ for every summation.

3.2 Gradient Boosting

Gradient Boosting (GB) is one of the machine learning-based boosting algorithms that composes the loss function. The concept is shown in Fig. 3. It works with the combination

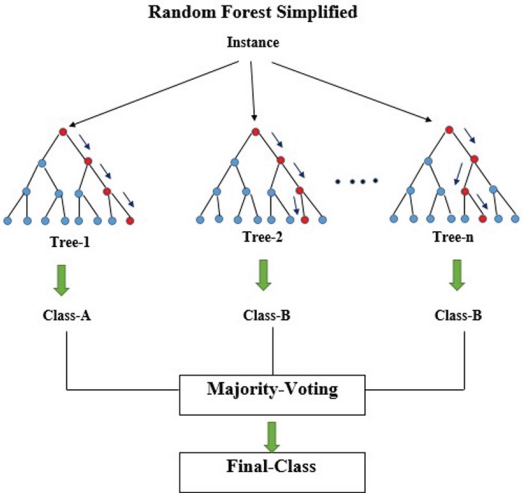


Fig. 2. Working Principle of Random Forest

and optimization of weak learners to decrease the loss function of a model. It removes overfitting to increase the performance of an algorithm [27]. Here $fi(x) =$ loss function with correlated negative gradients $(-\rho ixgm(X))$, $m =$ number of iterations.

Feature increment $(i) = 1, 2, 3, \ldots.., m$. Therefore, the optimal function $F(X)$ after m^{-th} iteration (3) is shown below [15].

$$F(X) = \sum_{i=0}^{m} fi(x) \tag{3}$$

Here, $gm =$ the path of loss function's fast decreasing $F(X) = Fn - 1(X)$ the decision tree's target is to solve the mistakes by previous learners [16, 17]. The negative iteration is shown below (4).

$$gm - \left(\frac{\partial L(y, F(X))}{\partial F(X)}\right) F(X) = Fm - 1(X) \tag{4}$$

K-Nearest

K-Nearest Neighbors is one of the machine learning algorithms that is mostly used in non-parametric classification methods as it allows the new data and existing data. The concept is shown in Fig. 4. It studies the Euclidean distance new (x_1, x_2) and existing (y_1, y_2) data (5) [18, 19, 26].

$$Euclidean\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

Ensemble Methods of Machine Learning

The ensemble method refers to the multiple classifiers that result in the best accuracy and effectiveness for the weak classifiers to create them as a strong classifier. It was applied

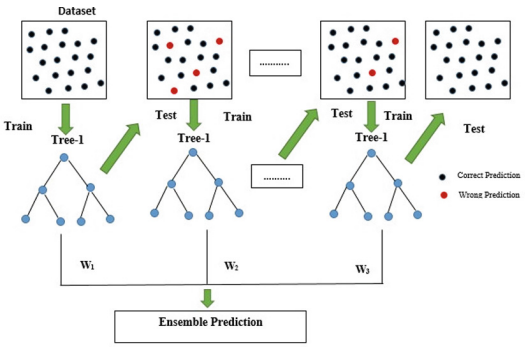


Fig. 3. Working Principle of Gradient Boosting

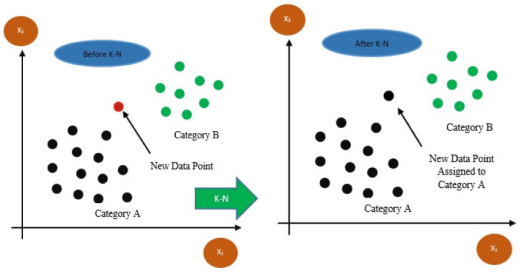


Fig. 4. Working Principle of K-Nearest

in our study because of variable handling, uncertainty, and bias, reduces variances, combines prediction of multiple models, and reduces the spread of predictions [20, 21]. Three ensemble methods were used in our study. We used Bagging, Boosting, and Voting ensemble, models.

Bagging

Bagging refers to the decrease of variance, diminishing handling, and missing variables. It enhances stability for different algorithms but is mainly applicable to decision tree algorithms. The concept is shown in Fig. 5. The formula of the Bagging model for classification is shown below (6) [17].

Here $f'(x)$ is the average of $fi(x)$ for i = 1, 2, 3, … T.

$$f'(x) = sign\left(\sum_{i=1}^{T} fi(x)\right) \tag{6}$$

Boosting

Boosting refers to the technique that uses a weighted average to work with several algorithms and makes the weak learners strong learners boost the accuracy of independent models creating the loss functions [23]. The concept is shown in Fig. 6. In our study,

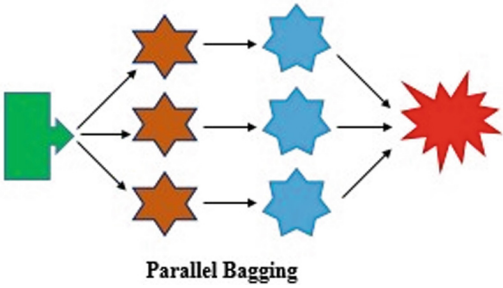


Fig. 5. Working Principle of Bagging

the boosting method is applied in the machine training and calculating the testing part to make the model hybrid. The proposed equation is shown below [22].

Here, $\Upsilon_t = ½\text{-}\epsilon_t$ (how much f_t is on the weighted sample) (7).

$$\frac{1}{n}\sum_{i=1}^{n} I(y_j g(x_i) < 0) \leq \prod_{t=1}^{T} \sqrt{1 - 4{\Upsilon_t}^2} \tag{7}$$

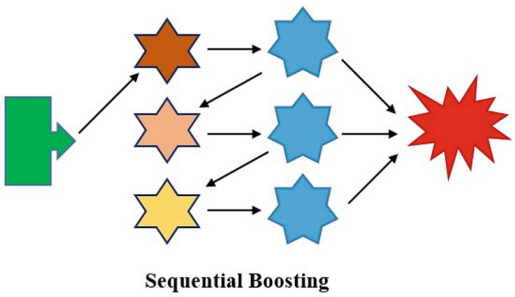


Fig. 6. Working Principle of Boosting

Voting

Voting classifiers refer to the combination of different classifiers to predict the class based on the best majority of voting. It means the model creates training by different models to predict results by combining the majority of voting. The concept is shown in Fig. 7. The equation we have used is shown below [23].

Here, w_j = weight that can be assigned to the jth classifier (8).

$$y' = argmax \sum_{j=1}^{m} w_j p_{ij}(\tag{8}$$

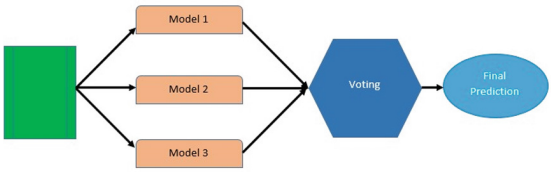


Fig. 7. Working Principle of Voting

4 Research Methodology

As the dataset was collected from Kaggle [9], the dataset was almost ready for implementation. The column and the row size are 32 and 569 respectively. The diagnosis column classifies the rate of breast cancer. All the attributes were important to predict breast cancer. Patients are separated into 2 conditions Malignant and Begin. Here Malignant was used as M and Begin was used as B. We have converted these values with nominal values. There 0 denotes 'B' and 1 denotes 'M'. We have calculated the rate of these two conditions. There were 357 patients in Begin stage and the rest 212 patients were in the Malignant stage. The ratio is shown in Fig. 8.

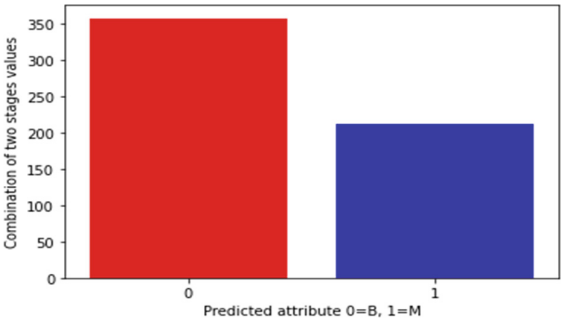


Fig. 8. Number of target values

The dataset contains nominal values and there were no missing or incorrect values. A comprehensive explanation of the dataset with its range is shown in Table 1.

Statistical Analysis

The analysis part is an important part of any kind of research work. This segment depends on developing and evaluating the algorithms I have used. As we have chosen comma separated valued (CSV) file to implement, we have to follow some steps to clean the dataset and make it usable. We have used several steps like data collection, pre-processing, etc.

In this study, we have used different four types of algorithms like Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), and K-Nearest (KN) Classifier algorithms. The best accuracy was LR and RF about 98.25%. Then Bagging, Boosting

Table 1. Details of the dataset

Attributes	Description	Value Range	Types of values
Diagnosis	Malignant or Begin	0 and 1	Integer
Radius_mean	Radius of Lobes	6.98 to 28.1	Float
Texture_mean	Mean of Surface Texture	9.71 to 39.28	Float
Perimeter_mean	Outer Perimeter of Lobes	43.8 to 188.5	Float
Area_mean	Mean Area of Lobes	143.5 to 2501	Float
Smoothness_mean	Mean of Smoothness Levels	0.05 to 0.163	Float
Compactness_mean	Mean of Compactness	0.02 to 0.345	Float
Concavity_mean	Mean of Concavity	0 to 0.426	Float
Concave points_mean	Mean of Concave Points	0 to 0.201	Float
Symmetry_mean	Mean of Symmetry	0.11 to 0.304	Float
Fractal_dimension_mean	Mean of Fractal Dimension	0.05 to 0.1	Float
Radius_se	SE of Radius	0.11 to 2.87	Float
Texture_mean	SE of Texture	0.36 to 4.88	Float
Perimeter_se	Perimeter of SE	0.76 to 22	Float
Area_se	Area of SE	6.8 to 542	Float
Smoothness_se	SE of Smoothness	0 to 0.03	Float
Compactness_se	SE of Compactness	0 to 0.14	Float
Concavity_se	SE of Concavity	0 to 0.4	Float
Concave points_se	SE of Concave Points	0 to 0.05	Float
Symmetry_se	SE of Symmetry	0.01 to 0.08	Float
Fractal_dimension_se	SE of Fractal Dimension	0 to 0.03	Float
Radius_worst	Worst Radius	7.93 to 36	Float
Texture_worst	Worst Texture	12 to 49.54	Float
Perimeter_worst	Worst Perimeter	50.4 to 251	Float
Area_worst	Worst Area	185 to 4254	Float
Smoothness_worst	Worst Smoothness	0.07 to 0.22	Float
Compactness_worst	Worst Compactness	0.03 to 1.06	Float
Concavity_worst	Worst Concavity	0 to 1.25	Float
Concave points_worst	Worst Concave Points	0 to 0.29	Float
Symmetry_worst	Worst Symmetry	0.16 to 0.66	Float
Fractal_dimension_worst	Worst Fractal Dimension	0.06 to 0.21	Float

and Voting algorithms were used and we got the best accuracy in RFBO was 99.122%. We have used 10-fold cross-validation, and hyperparameter tuning.

Flow Chart

We have used 80% in training and 20% in the testing part. Then we implemented the general classifier algorithms. The classifier evaluation was measured and then we used Bagging, Boosting, and then voting algorithms which are shown in Fig. 9.

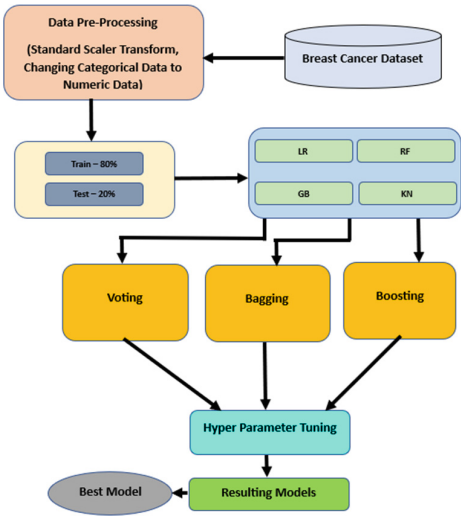


Fig. 9. Methodology

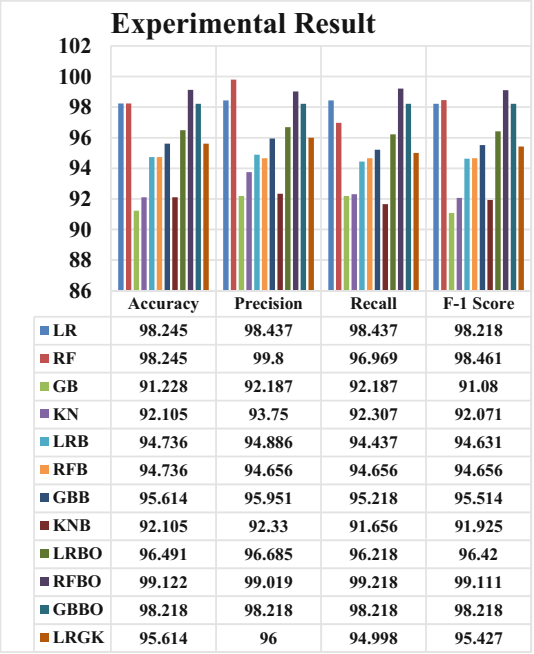
5 Experimental Results

We have calculated the outcome at the beginning and end of hybrid methods. We got the best 99.122% accuracy using RFBO. Then we got LR and RF with about 98.245% accuracy. Boosting model GBBO got 98.218% and LRBO got 96.491% testing accuracy. The Precision score was best in RF about 99.8% but RFBO got 99.019%, LR got 98.437%, and GBBO got 98.218%. The Recall score was best in RFBO about 99.218% but LR got 98.437%, GBBO got 98.218%, and RF got 96.696%. The F-1 score was best in RFBO about 99.111% but RF got 98.461%, LR and GBBO got 98.218%. All results are shown in Fig. 10.

In our study, we have calculated the run time for every different model shown in Fig. 11. We got the longest runtime of GBB for 492 ms where the lowest of KN was measured for 6.13 ms the runtime is shown below for every model.

6 Conclusion and Future Work

The present world is the modern world. Everything in the world is now technologically advanced and easy. Everyone in the world can familiarize themselves with the new



	Accuracy	Precision	Recall	F-1 Score
LR	98.245	98.437	98.437	98.218
RF	98.245	99.8	96.969	98.461
GB	91.228	92.187	92.187	91.08
KN	92.105	93.75	92.307	92.071
LRB	94.736	94.886	94.437	94.631
RFB	94.736	94.656	94.656	94.656
GBB	95.614	95.951	95.218	95.514
KNB	92.105	92.33	91.656	91.925
LRBO	96.491	96.685	96.218	96.42
RFBO	99.122	99.019	99.218	99.111
GBBO	98.218	98.218	98.218	98.218
LRGK	95.614	96	94.998	95.427

Fig. 10. Overall Outputs of models

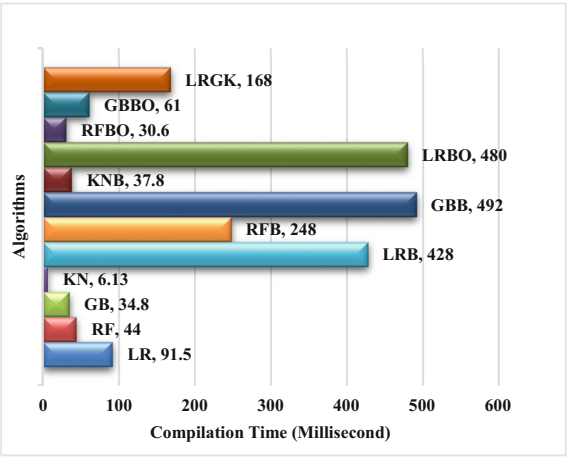


Fig. 11. Runtime Calculation

technology. With the help of technology, we have proposed is so much easy and low time-consuming. We have tried to reduce the complexity of breast cancer prediction among people. Our people can be benefited from our exciting models. We have to ensure the proposal is practical and we are promising to add many more features to our proposal

in the future and ensure we will work on more popular things. We are expressing this hope. We are human beings, we have mortality. We are affecting several diseases in our daily life. Some of us have recovery essentials but most of us suffer from cancers. As we are living in a developing world, the treatment and diagnosis technologies are more dynamic and accurate. New technologies have shortened the time and complexity of breast cancer disease identification. We have tried to do something new for our people. We hope our model will be accepted by the people. We have worked on some algorithms here and plan to add more in the future for better performance.

References

1. Yang, L., Shami, A.: On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing **415**, 295–316 (2020)
2. Khan, F., Kanwal, S., Alamri, S., Mumtaz, B.: Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction. IEEE Access **8**, 20954–20964 (2020)
3. Rani, V.M.K., Dhenakaran, S.S.: Classification of ultrasound breast cancer tumor images using neural learning and predicting the tumor growth rate. Multimedia Tools Appl. **79**(23–24), 16967–16985 (2019). https://doi.org/10.1007/s11042-019-7487-6
4. Li, Y., Liu, Y., Zhang, M., Zhang, G., Wang, Z., Luo, J.: Radiomics with attribute bagging for breast tumor classification using multimodal ultrasound images. J. Ultrasound Med. **39**(2), 361–371 (2020)
5. Gómez-Flores, W., Hernández-López, J.: Assessment of the invariance and discriminant power of morphological features under geometric transformations for breast tumor classification. Comput. Meth. Progr. Biomed. **185**, article 105173 (2020)
6. Liu, Y., Ren, L., Cao, X., Tong, Y.: Breast tumors recognition based on edge feature extraction using support vector machine. Biomed. Signal Process. Control **58**(101825), 1–8 (2020)
7. Irfan, R., Almazroi, A.A., Rauf, H.T., Damaševičius, R., Nasr, E.A., Abdelgawad, A.E.: Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion. Diagnostics **11**(7), 1212 (2021)
8. Lahoura, H., Singh, A., Aggarwal et al.: Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. Diagnostics **11**(2), 241 (2021)
9. Breast Cancer Dataset. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
10. What is Correlation in Machine Learning? https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47. Accessed: 6 Aug 2020
11. Mary Gladence, L., Karthi, M., Maria Anu, V.: A statistical comparison of logistic regression and different bayes classification methods for machine learning. ARPN J. Eng. Appl. Sci. **10**(14) (2015). ISSN 1819-6608
12. Logistic Regression for Machine Learning. https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/. Accessed 6 Aug 2021
13. Ghosh, P., Karim, A., Atik, S.T., Afrin, S., Saifuzzaman, M.: Expert cancer model using supervised algorithms with a LASSO selection approach. Int. J. Electr. Comput. Eng. (IJECE) **11**(3), 2631 (2021)
14. Nahar, N., Ara, F.: Liver disease prediction by using different decision tree techniques. Int. J. Data Mining Knowl. Manage. Process **8**(2), 01–09 (2018)
15. Aljahdali, S., Hussain, S.N.: Comparative prediction performance with support vector machine and random forest classification techniques. Int. J. Comput. Appl. **69**(11) (2013)

16. Bentéjac, C., Csörgő, A., Martínez-Muñoz, G.: A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. **54**(3), 1937–1967 (2020). https://doi.org/10.1007/s10462-020-09896-5
17. Drucker, H., Cortes, C., Jackel, L.D., LeCun, Y., Vapnik, V.: Boosting and other ensemble methods. Neural Comput. **6**(6), 1289–1301 (1994)
18. Pasha, M., Fatima, M.: Comparative analysis of meta learning algorithms for liver disease detection. J. Softw. **12**(12), 923–933 (2017)
19. Wang, Y., Jha, S., Chaudhuri, K.: Analyzing the robustness of nearest neighbors to adversarial examples. In: International Conference on Machine Learning, pp. 5133–5142. PMLR (2018)
20. Sharma, A., Suryawanshi, A.: A novel method for detecting spam email using KNN classification with spearman correlation as distance measure. Int. J. Comput. Appl. **136**(6), 28–35 (2016)
21. Hou, Z.-H.: Ensemble Methods: Foundations and Algorithms. CRC Press (2012)
22. Emmens, A., Croux, C.: Bagging and boosting classification trees to predict churn. J. Market. Res. **43**(2), 276–286 (2006)
23. Islam, R., Beeravolu, A.R., Islam, M.A.H., Karim, A., Azam, S., Mukti, S.A.: a performance based study on deep learning algorithms in the efficient prediction of heart disease. In: 2021 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1–6. IEEE (2021)
24. Tajmen, S., Karim, A., Mridul, A.H., Azam, S., Ghosh, P., Dhaly, A., Hossain, M.N.: A machine learning based proposition for automated and methodical prediction of liver disease. In: The 10th International Conference on Computer and Communications Management in Japan (2022)
25. Molla, S., et al.: A predictive analysis framework of heart disease using machine learning approaches. Bull. Electr. Eng. Informatics **11**(5), 2705–2716 (2022)
26. Afrin, S., et al.: Supervised machine learning based liver disease prediction approach with LASSO feature selection. Bull. Electr. Eng. Informatics **10**(6), 3369–4337 (2021)
27. Ghosh, P., et al.: Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access **9**, 19304–19326 (2021)
28. Jubier Ali, M., Chandra Das, B., Saha, S., Biswas, A.A., Chakraborty, P.: A comparative study of machine learning algorithms to detect cardiovascular disease with feature selection method. In: Skala, V., Singh, T.P., Choudhury, T., Tomar, R., Abul Bashar, M. (Eds.) Machine Intelligence and Data Science Applications. Lecture Notes on Data Engineering and Communications Technologies, vol. 132. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-2347-0_45

Recommender System for Scholarly Articles to Monitor COVID-19 Trends in Social Media Based on Low-Cost Topic Modeling

Houcemeddine Turki(✉), Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha

Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
turkiabdelwaheb@hotmail.fr,
{mohamedali.hajtaieb,mohamed.benaouicha}@fss.usf.tn

Abstract. During the last years, many computer systems have been developed to track and monitor COVID-19 social network interactions. However, these systems have been mainly based on robust probabilistic approaches like Latent Dirichlet Allocation (LDA). In another context, health recommender systems have always been personalized to the needs of single users instead of regional communities. Such applications will not be useful in the context of a public health emergency such as COVID-19 where general insights about local populations are needed by health policy makers to solve critical issues in a timely basis. In this research paper, we propose to modify LDA by letting it be driven by knowledge resources and we demonstrate how we can apply our topic modeling method to local social network interactions about COVID-19 to generate precise topic clusters reflecting the social trends about the pandemic at a low cost. Then, we outline how terms in every topic cluster can be converted into a search query to generate scholarly publications from PubMed Central for adjusting COVID-19 trendy thoughts in a population.

Keywords: Recommender System · Scholarly Publications · Social Network Analysis · Topic Modeling · Latent Dirichlet Allocation

1 Introduction

The analysis of social media interactions related to a disease outbreak like *COVID-19* can be very useful to assess the general perception of the concerned disease by a local population, identify rumors and conspiracy theories about the widespread medical condition, and track the spread and effect of official information, news and guidelines about the disease outbreak among a specific community [4]. Data provided by social networking sites are characterized by their volume, variety, veracity,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 249–259, 2023.
https://doi.org/10.1007/978-3-031-27409-1_22

velocity, and value and can consequently provide a huge real-time and ever-growing amount of information reflecting various aspects of the current social response to the COVID-19 pandemic and facilitating rapid data-driven decision-making to face any encountered societal problem [10].

However, most of the systems allowing social network analysis related to COVID-19 mostly depend on purely probabilistic approaches that do neither consider the semantic features of the assessed texts nor have a transparent way for identifying how results are returned [12,25]. These methods range from Latent Dirichlet Allocation and Latent Semantic Analysis to Word Embeddings and Neural Networks.

In this research paper, we investigate the creation of a novel approach that integrates free knowledge resources and open-source algorithms in the Latent Dirichlet Allocation of social network interactions related to the COVID-19 pandemic in *Facebook*[1] for generating a precise topic modeling of the topics of interest related to the ongoing disease outbreak for a local population at a low cost. Besides, to enable decision-making for monitoring the real-time social impact of the COVID-19 pandemic, we propose to use the returned topic clusters to recommend scholarly publications that can be used by health professionals and authorities to fight widespread misinformation and provide interesting accurate guidelines for their communities concerned by COVID-19 through the data mining of *PubMed Central*,[2] a database of open access biomedical research publications available online. We begin by providing an overview of social network analysis for crisis management as well as scholarly publication recommender systems (Sect. 2). Then, we outline our knowledge-based approach for the LDA-based recommendation of scholarly publications for COVID-19 societal responses based on the social network interactions of a given population related to the COVID-19 pandemic (Sect. 3). Finally, we give conclusions about our system and we draw future directions for our research work (Sect. 4).

2 Overview

2.1 Social Network Analysis for Crisis Management

Since their creation, social network sites have served as tools for online communication between individuals all over the world allowing them to effectively share their opinions, their habits, their statuses, and their thoughts in real-time with a wide audience [9]. The ability of these online platforms (e.g., *Facebook*) to establish virtual connections between individuals has permitted these websites to have billions of users within a few years of work [9]. Nowadays, thanks to their growth, social networks provide real-time big data about human con-

[1] https://www.facebook.com.
[2] https://www.ncbi.nlm.nih.gov/pmc/.

cerns including political and health crises. This resulted in the emergence of a significant research trend of using social network interactions to track crisis responses.

Social network analysis permits to parse textual posts issued by the users using common *natural language processing* techniques [13], *topic modeling* [4] and advanced *machine learning* techniques [12,25] and to analyze the graphs of non-textual interactions around posts (i.e., *shares*, *likes*, and *dislikes*) using a range of techniques from the perspective of network science and knowledge engineering [15]. The application of computer methods to analyze social network data can reliably reflect the sentiments and thoughts of a given community about the crisis and help identify and predict the geographical and socio-economic evolution of the phenomenon [4]. Social network analysis can be a valuable tool for detecting and eliminating inconsistent posts spreading misinformation and rumors across social networking sites leaving room for accurate posts and knowledge about the considered topic to get more disseminated [1,16]. The sum of all this inferred information will be efficient for aiding the recommendation of actions and resources to solve the considered crisis.

2.2 Scholarly Paper Recommendation

For centuries, scholarly publications have been considered a medium for documenting and disseminating scientific breakthroughs and advanced knowledge in multiple research fields ranging from medicine and biology to arts and humanities [19]. That is why they provide a snapshot of the latest specialized information that can be used to analyze and study a topic (e.g., *crisis*) and troubleshoot all the faced real-life matters related to it. Such knowledge can be explored through the analysis of the full texts of scholarly publications or the mining of the bibliographic metadata of these papers in bibliographic databases like *PubMed* and *Web of Science* using a variety of techniques including *Natural Language Processing*, *Machine Learning*, *Embeddings*, and *Semantic Technologies* [20]. With the rise of digital libraries in the computer age, new types of information about the timely online social interest in scholarly publications have emerged such as usage statistics, shares in social networks, and queries in search engines [7]. The combination of both data types coupled with social network analysis enables the development of knowledge-based systems to identify the main trendy topics for users as well as to measure the similarity between scholarly publications and user interests [5]. The outcomes of such intelligent systems will allow the generation of accurate recommendations of scholarly articles to meet the user needs [5].

Most recommender systems try to generate a *user interest ontology* based on the full texts of its scholarly readings and social network posts and then compare the generated user interest profile with unread research publications using semantic similarity measures to find the best papers to be recommended [18]. As well, there are some collaborative filtering approaches for the recommendation of scholarly publications for a given user based on the readings and behaviors of other users [5]. Several directions can be followed to develop this social network-based approach and combine it with content-based and social

interest-based approaches for achieving a better accuracy of scholarly recommendations. Despite the variety of scholarly publication recommender systems, quite all of them propose further readings based on the interests of a particular user and not of a global community. This can be not relevant in the context of population health where global measures are required. Previous efforts for the social recommendation of scholarly publications using LDA have mainly been based on characterizing the user interests through the topic modeling of their scholarly publications [3], of their social interactions and profiles [24], or of the research papers they interacted with online [2,23]. Several initiatives also considered the computation of user similarity based on LDA outputs to recommend publications for a given user (so-called *collaborative filtering*) [23]. Despite the value of these methods that recommend scholarly publications for single users of social networks, these approaches cannot be efficient in the situation of a broad crisis like COVID-19 when specialized information is requested on a large scale. In this research paper, we propose to use *Latent Dirichlet Allocation* (LDA) for modeling the interests of a whole regional community based on their social media interactions and we use the generated outputs for recommending further scholarly readings for this population based on content-based filtering. The approach we are proposing envisions supporting the multilingualism and variety of the social interactions regarding COVID-19 at the scale of a large community and accordingly formulate search queries to find research publications in the *PubMed Central* database to solve misinformation and support key facts and concerns about the outbreak.

3 Proposed Approach

Figure 1 illustrates the different components of the architecture conceived and implemented for recommending scholarly publications based on the social data analysis for tracking and monitoring the COVID-19 pandemic in the Tunisian Context. We mainly focus on *Facebook* and *Twitter* COVID-19-related posts in particular Tunisian Arabic posts. In this regard, a keyword-based search approach is performed by scraping Facebook public pages and using the *Twitter4J* API for the *Twitter* microblogging website. The pages to be scraped are chosen through human verification of their restricted coverage of Tunisia-related topics. After being anonymized, the posts are filtered according to a built vocabulary for COVID-19 based on an open knowledge graph and machine translation. The collected posts are ingested through Apache Flume connectors and Apache Kafka cluster to be analyzed.

However, the received Facebook posts and tweets are characterized by their heterogeneity in terms of schema, syntax, and semantics raising different challenges mainly related to data pre-processing relative to each social network. Indeed, this heterogeneity requires specific treatment for each social network for identifying COVID-19-related social entities. In this regard, and to overcome the challenges related to data pre-processing, we resort to the use of the Social Network OWL (*SNOWL*) ontology [17]. This ontology is used to uniformly model

posts and tweets independently of the source in which they reside. *SNOWL* presents a shared vocabulary between different online social networks. It models different social data entities namely, users, content (e.g., posts, comments, videos, etc.), and user-content interactions. *Author* concept is used for presenting users across online social networks and their related metadata such as name, age, interest, etc. *Publication* concept is used, also, for modeling posts and tweets. Furthermore, this ontology models a new concept namely popularity. Indeed, *SNOWL* defines user popularity-related concepts (e.g., number of friends, number of followers, etc.) and content popularity-related concepts (e.g., number of shares, number of comments, etc.). The popularity concept plays an important role in identifying content's reputation (e.g., the most shared COVID-19 posts) and identifying the most influencers' profiles. In addition, *SNOWL* includes also concepts serving for modeling user's opinion through the reuse of the *MARL*[3] ontology this is helpful to identify the polarity (i.e., positive, negative, neutral) of each collected post. It is worth mentioning that through the use of *SNOWL* ontology we can select posts according to their publication data indeed this ontology reuses also the *TIME*[4] ontology. Therefore, the posts and tweets are transformed into RDF triples according to the *SNOWL* ontology *TBox*. In addition, the resulting RDF triples are stored based on a distributed RDF storage system. The triples are queried based on SPARQL queries by the Latent Dirichlet Allocation (LDA) algorithm to detect COVID-19-related trends. When local COVID-19-related topic clusters are identified, the ten most relevant terms for every cluster are combined to create a *PubMed Central* query that finds the most relevant research publications that correspond to every topic.

3.1 Multilingual Topic Modelling

As a consequence of the multilingualism of social network interactions [14], classical topic modeling methods need to be largely revised and improved for better efficiency in characterizing social interests [22]. To solve this problem, multilingual topic modeling algorithms have been developed based on language identification followed by named entity extraction, entity disambiguation and linking, and finally, the application of the topic models on a mono-lingual or language-neutral representation of documents [6,22]. More precisely, LDA is a probabilistic generative model with latent variables. The exploited implementations are Mallet[5] and Gensim.[6] The parameters of this model are:

- The number k of subjects to extract.
- The two hyper-parameters α and β; α acts on the distribution of documents D (social posts) between the topics and β acts on the distribution of words between themes.

[3] http://www.gsi.upm.es:9080/ontologies/marl/.
[4] https://www.w3.org/TR/owl-time/.
[5] https://mimno.github.io/Mallet/.
[6] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/.

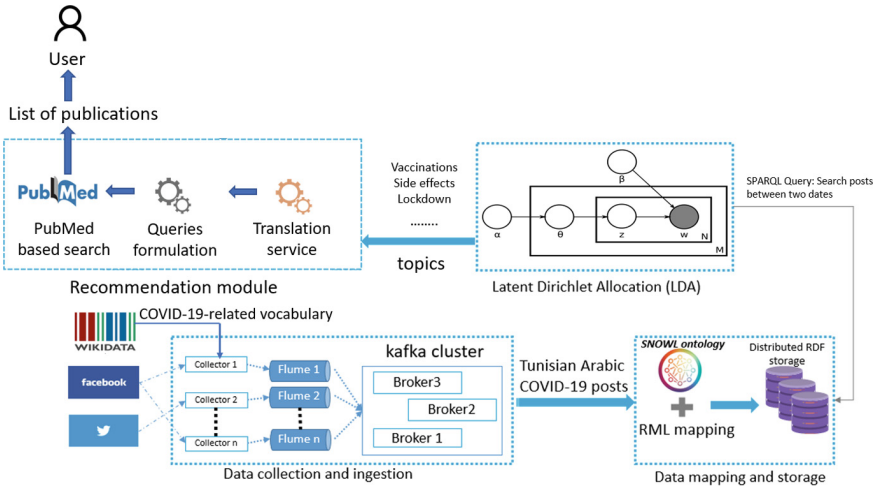


Fig. 1. Architecture of the scholarly paper recommender system to monitor the COVID-19 pandemic based on social data analysis

The LDA is a 3-level hierarchical model. Let W be the set of words in a post noted d and Z be the vector of the topics corresponding to all words in all posts, the document generation process by the LDA model works as follows:

- Choose the k number of subjects to extract.
- For each document $d \in D$, choose a distribution law θ among the subjects.
- For each word $w \in W$ of d, choose a subject $z \in Z$ respecting the law θ.

In the context of our approach, we built a COVID-19-related vocabulary through the extraction of labels, descriptions, and aliases of the Wikidata[7] items related to COVID-19. As an open and multilingual knowledge graph, Wikidata provides a wide range of data about the outbreak in a variety of languages, including Arabic, French, and English [21]. The vocabulary is enriched using machine translation outputs to avoid gaps in the language representation of the COVID-19 knowledge. For this purpose, *MyMemory*[8] is used as a public API for machine translation coupled to the use of the *Optimaize Language Detector*[9] Java Libary for the identification of the source languages of posts. Later, a set of users is automatically extracted from the official *Facebook* and *Twitter* pages tracking the pandemic status in Tunisia and providing the daily update of statistics. After, users are explored to extract the posts and identify those talking about COVID-19 through the built vocabulary. Selected posts are then ingested into big data architecture for integrating posts coming from different social networks using the *SNOWL* ontology. The mapping capability serves to represent

[7] A freely available multilingual knowledge graph (https://www.wikidata.org).
[8] https://mymemory.translated.net/.
[9] https://github.com/optimaize/language-detector.

the Arabic textual data in the post as *RDF* triplets according to the common concepts defined in the ontology. To inquire about the *RDF* database, *SPARQL* services are implemented for handling access to the data. So, the returned data as a response to a query fixing the time window will be the input for the topic modeling module exploiting the *LDA* method as a statistical and language-independent approach. The topics are provided according to a personalized configuration fixing the number of topics and words in each topic and exploited as input for the recommendation module.

3.2 Search Engine-Based Recommendation

As a large-scale bibliographic database, *PubMed Central* needs to be parsed using a search engine to enable medical practitioners and the general audience to find proper evidence about a fact. The sum of the provided contributions resulted in the creation of the "Best Match" new relevance search algorithm for *PubMed Central* [11]. This algorithm processes the search results using the *BM25* term-weighting function and then re-ranks them using *LambdaMART*, a high-performance pre-trained model that classifies publications using multiple characteristics extracted from queries and documents [11].

Table 1. Behavior of PubMed Central Search Engine for *AND* queries

Assessed Feature	PMC Query	@10	@100	Runtime (sec.)
Baseline	"Cough" AND "Symptom" AND "COVID-19"	–	–	1.000
Duplicate Keyword	"Cough" AND "Cough" AND "Symptom" AND "COVID-19"	0.8	0.73	0.955
Duplicate Keyword	"Cough" AND "Symptom" AND "COVID-19" AND "COVID-19"	0.4	0.61	1.035
Duplicate Keyword	"Cough" AND "Symptom" AND "Symptom" AND "COVID-19"	0.4	0.65	0.965
Not Exact Match	Cough AND Symptom AND COVID-19	0	0.09	1.005
Keyword Order	"Cough" AND "COVID-19" AND "Symptom"	1	1	0.935
Keyword Order	"COVID-19" AND "Cough" AND "Symptom"	1	1	0.895
Keyword Order	"COVID-19" AND "Symptom" AND "Cough"	1	1	0.93
Keyword Order	"Symptom" AND "Cough" AND "COVID-19"	1	1	0.955
Keyword Order	"Symptom" AND "COVID-19" AND "Cough"	1	1	0.975

To see the practical behavior of this novel algorithm, we apply several user queries to it trying to find publications where cough is featured as a symptom of COVID-19. These queries assess multiple characteristics, particularly the duplication of keywords, the use of exact match, the order of keywords, and the use of

logical operators. The evaluation of the user queries will be through a comparison with the baseline user query *"Cough" AND "Symptoms" AND "COVID-19"* revealing the scholarly publication where there is a certain mention of cough as a symptom of COVID-19. The evaluation will be based on three metrics: the agreement between the ten first results of a query with the ones of the baseline (@10), the agreement between the hundred first results of a query with the ones of the baseline (@100), and the runtime of the query in seconds. The source code implemented in Python 3.9 and used for retrieving the metrics can be found at https://shorturl.at/flLM2. When performing this evaluation, we found out that keyword order does not influence the search results of the query when using *AND* as a logical operator as shown in Table 1. This is significantly confirmed in Table 2 for queries using *OR* as a logical operator. However, it is revealed in the two tables that the query runtime tends to be largely shortened when the most specific keyword is put first in the query (i.e., COVID-19 in our situation).

Furthermore, when assessing whether the queries using quotation marks to find exact matches of keywords provide similar search results to the ones not using quotation marks, we found a very large difference in the returned scholarly evidence between the two types of user queries (Tables 1 and 2). This verifies that the use of quotation marks can cause the missing of several relevant papers from the search results although this user behavior can be useful to return specific publications on the topic of the query. Moreover, when the keyword is mentioned twice in a user query, it significantly influences the order of returned results. This demonstrates that keyword duplication can be practically used to emphasize one keyword in the query over another one, allowing to have more customized search results. The queries that do not use quotation marks or that include duplicate keywords tend to be only slightly slower if the used logical operator is *OR* as shown in the two tables, proving that such practices are not expensive from a computational point of view. Besides, the comparison of the use of *OR* vs. the use of *AND* as a logical operator between query keywords (Table 2) reveals that the papers that include all keywords tend to be ranked first by the *PubMed Central* search engine even when OR is used in the user query. These patterns are important to find the best way to find relevant research papers related to a set of terms. Subsequently, we will benefit from them to find the best way to retrieve relevant research papers related to the output of the topic modeling of COVID-19 trends in social networks. We use *OR* as a logical operator between the terms of the LDA cluster and we link the created search query to *COVID-19* using the *AND* operator to ensure that the *PubMed Central* results corresponding to the cluster are contextualized to the COVID-19 pandemic. Let S be the main topic of the collected posts (*COVID-19* in our context), w_i be the i^{th} most relevant word for the topic cluster, and N be the number of words that are considered to represent every topic cluster ($N \in \mathbb{N}$), the query that should be used to extract the most relevant scholarly publications for a given cluster is reflected by the following equation:

$$S \wedge (\vee_{i \leq N} w_i)$$

Such a method can be customized by emphasizing more relevant terms by including them multiple times in the query as shown in Table 1. However, we did not use this feature to save runtime in the *PubMed Central* queries. The result of our method will be the *PubMed Central*-indexed scholarly publications including most of the main words of the considered topic cluster according to the *Best Match* sorting method. This goes in line with previous efforts of using search engines as tools to drive knowledge-based systems in healthcare [8].

Table 2. Behavior of PubMed Central Search Engine for *OR* queries

Assessed Feature	PMC Query	@10	@100	Runtime (sec.)
OR vs. AND	"Cough" OR "Symptom" OR "COVID-19"	1	0.94	1.123
Keyword Order	"Cough" OR "COVID-19" OR "Symptom"	1	0.94	1.088
Duplicate Keyword	"Cough" OR "Symptom" OR "COVID-19" OR "COVID-19"	0.4	0.54	1.2
Not Exact Match	Cough OR Symptom OR COVID-19	0	0.06	1.245
Not Exact Match	Cough OR Symptom OR COVID-19	0	0.06	1.245

4 Conclusion and Future Works

This concurrent research is considered a recommender system to provide scholarly publications to monitor and track the COVID-19 pandemic which depends on the analysis of the data of the social platforms Facebook and Twitter. This study focuses on the Tunisian context but the process can be generalized to cover other languages. It exploits an ontology-based integration solution based on Big Data frameworks and lower-cost topic modeling. This proposed approach also remains valid for exploring other events and gives the possibility for an in-depth analysis of well-selected topics in a recursive way. The LDA output is considered a fuzzy classification affecting the posts to the extracted topics. In future works, we plan to broaden our work to cover other languages and to go deeper in the analysis by developing a recursive process able to zoom in on the topics by extracting the sub-topics and building predictive models which are favored by the probabilistic generative of LDA.

Acknowledgments. This paper is supported by the Ministry of Higher Education and Scientific Research in Tunisia (MoHESR) in the framework of Project PRFCOV19-D1-P1. This work is a part of the initiative entitled *Semantic Applications for Biomedical Data Science* and managed by *SisonkeBiotik*, a community for machine learning and healthcare in Africa.

References

1. Ahmed, W., Vidal-Alaball, J., Downing, J., López Seguí, F.: Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. J. Med. Internet Res. **22**(5), e19458 (2020)
2. Amami, M., Faiz, R., Stella, F., Pasi, G.: A graph based approach to scientific paper recommendation. In: Proceedings of the International Conference on Web Intelligence, pp. 777–782. WI '17, Association for Computing Machinery, New York, NY, USA (2017)
3. Amami, M., Pasi, G., Stella, F., Faiz, R.: An LDA-based approach to scientific paper recommendation. In: Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) Natural Language Processing and Information Systems, pp. 200–210. Springer International Publishing, Cham (2016)
4. Amara, A., Hadj Taieb, M.A., Ben Aouicha, M.: Multilingual topic modeling for tracking covid-19 trends based on facebook data analysis. Appl. Intell. **51**(5), 3052–3073 (2021)
5. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**(4), 305–338 (2015)
6. Bhargava, P., Spasojevic, N., Ellinger, S., Rao, A., Menon, A., Fuhrmann, S., Hu, G.: Learning to map wikidata entities to predefined topics. In: Companion Proceedings of The 2019 World Wide Web Conference, pp. 1194–1202. WWW '19, Association for Computing Machinery, New York, NY, USA (2019)
7. Bornmann, L.: Validity of altimetrics data for measuring societal impact: a study using data from altimetric and f1000prime. J. Inf. **8**(4), 935–950 (2014)
8. Celi, L.A., Zimolzak, A.J., Stone, D.J.: Dynamic clinical data mining: search engine-based decision support. JMIR Med. Inform. **2**(1), e13 (2014). Jun
9. Clark, J.L., Algoe, S.B., Green, M.C.: Social network sites and well-being: the role of social connection. Curr. Dir. Psychol. Sci. **27**(1), 32–37 (2017)
10. Demchenko, Y., Ngo, C., de Laat, C., Membrey, P., Gordijenko, D.: Big security for big data: addressing security challenges for the big data infrastructure. In: Jonker, W., Petković, M. (eds.) Secure Data Management, pp. 76–94. Springer International Publishing, Cham (2014)
11. Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., Osipov, M., Kholodov, M., Ismagilov, R., Mohan, S., et al.: Best match: new relevance search for PubMed. PLOS Biol. **16**(8), e2005343 (2018)
12. Hossain, T., Logan IV, R.L., Ugarte, A., Matsubara, Y., Young, S., Singh, S.: COVIDLies: detecting COVID-19 misinformation on social media. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics, Online (2020)
13. Kanakaraj, M., Guddeti, R.M.R.: Performance analysis of ensemble methods on twitter sentiment analysis using NLP techniques. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), pp. 169–170 (2015)
14. Kashina, A.: Case study of language preferences in social media of Tunisia. In: Proceedings of the International Conference Digital Age: Traditions, Modernity and Innovations (ICDATMI 2020), pp. 111–115. Atlantis Press (2020)
15. Kim, J., Hastak, M.: Social network analysis: characteristics of online social networks after a disaster. Int. J. Inf. Manag. **38**(1), 86–96 (2018)
16. Lanius, C., Weber, R., MacKenzie, W.I.: Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey. Soc. Netw. Anal. Min. **11**(1), 32:1–32:15 (2021)

17. Sebei, H., Hadj Taieb, M.A., Ben Aouicha, M.: SNOWL model: social networks unification-based semantic data integration. Knowl. Inf. Syst. **62**(11), 4297–4336 (2020)
18. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests, pp. 29–38. JCDL '10, Association for Computing Machinery, New York, NY, USA (2010)
19. Townsend, R.B.: History and the future of scholarly publishing. Perspect. Hist. **41**(3), 34–41 (2003)
20. Turki, H., Hadj Taieb, M.A., Ben Aouicha, M., Fraumann, G., Hauschke, C., Heller, L.: Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata. Front. Res. Metr. Anal. **6**, 694307 (2021)
21. Turki, H., Hadj Taieb, M.A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., Labra Gayo, J.E., Youngstrom, E.A., Banat, M., Das, D., et al.: Representing covid-19 information in collaborative knowledge graphs: The case of wikidata. Semant. Web **13**(2), 233–264 (2022)
22. Vulić, I., De Smet, W., Tang, J., Moens, M.F.: Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. Inf. Process. Manag. **51**(1), 111–147 (2015)
23. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 448–456. KDD '11, Association for Computing Machinery, New York, NY, USA (2011)
24. Younus, A., Qureshi, M.A., Manchanda, P., O'Riordan, C., Pasi, G.: Utilizing Microblog Data in a Topic Modelling Framework for Scientific Articles' Recommendation, pp. 384–395. Springer International Publishing, Cham (2014)
25. Zamani, M., Schwartz, H.A., Eichstaedt, J., Guntuku, S.C., Virinchipuram Ganesan, A., Clouston, S., Giorgi, S.: Understanding weekly COVID-19 concerns through dynamic content-specific LDA topic modeling. In: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pp. 193–198. Association for Computational Linguistics, Online (2020)

Statistical and Deep Machine Learning Techniques to Forecast Cryptocurrency Volatility

Ángeles Cebrián-Hernández[1](✉), Enrique Jiménez-Rodríguez[2], and Antonio J. Tallón-Ballesteros[3]

[1] Department of Applied Economics, Seville University, Seville, Spain
mcebrian1@us.es
[2] Department of Financial Economics, Pablo de Olavide University, Seville, Spain
[3] Department of Electronic, Computer Systems and Automatic Engineering, University of Huelva, Huelva, Spain

Abstract. This paper studies cryptocurrency volatility forecast covering a state-of-the-art review as well empirical comparison through supervised learning. This research has two main objectives. The first objective is the usage of artificial intelligence in the field of predicting the volatility of cryptocurrencies, in particular bitcoin. In this work, supervised machine learning algorithms from two different perspectives, a statistical one and a deep learning one, are compared to predict bitcoin volatility using economic-financial variables as additional information in the models. The second objective is to compare the fit of artificial intelligence models with traditional econometric models such as Multivariate GARCH (traditional generalized heteroscedasticity conditional autoregressive model (M-GARCH).

Keywords: Bitcoin · Volatility · Machine Learning · Random Forest · Neural Networks · DCC M-GARCH

1 Introduction

The concept of cryptocurrencies began in 2008 with the publication of the Bitcoin project by Satoshi Nakamoto [1], which described a digital currency based on a sophisticated peer-to-peer (p2p) protocol that allowed online payments to be sent directly to a recipient without going through a financial institution. At the time, a potential non-sovereign asset that was fully decentralized and isolated from the uncertainties of a country or market was presented as a great value proposition [2]. All cryptographically controlled transactions make them secure, validated and stored in blockchain by a decentralized network [3]. Many authors have sought relationships between Bitcoin and other assets of various kinds. Vassiliadis et al. [4] note that there is a strong correlation between Bitcoin price and trading volume and transaction cost, and there is some relationship with gold, crude oil and stock index. Statistics has always offered techniques and models to make predictions as accurate as possible. Models such as the GARCH family have

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 260–269, 2023.
https://doi.org/10.1007/978-3-031-27409-1_23

been pioneers in this type of time series forecasting. Authors as Katsiampa [5] focus on comparing these models for volatility prediction. In recent decades, new phenomena have emerged that have pushed traditional forecasting to store and process large amounts of data. In [6] it shown that the volatility of Bitcoin does not behave like that of exchange rates (EUR/USD) or commodities such as oil. For the results of correlation, they use the DCC-MGARCH model, after demonstrating that it performs better than other variants of M-GARCH.

The Machine Learning (ML) methodology and, in particular, the concept of Artificial Neural Networks (ANN), both belonging to the AI field, are the most widely used. However, it is important to note that there is still no well-defined boundary between traditional statistical prediction models and ML procedures. See, for example, the discussions by Barker [7], Januschowski et al. [8] and Israel et al. [9] for an excellent description of the differences between "traditional" and ML procedures. Both AI techniques have had unprecedented popularity in the field of price prediction of all types of financial assets, including cryptoassets.

Within classical ML algorithms, we find several studies, such as Panagiotidis et al. [10], where authors use the LASSO (Least Absolute Shrinkage and Selection Operator) algorithm [11] to analyze a dataset with several predictors of stock, commodity, bond and exchange rate markets to investigate the determinants of bitcoin. Derbentsev et al. [12] apply two of the most powerful ensemble methods, Random Forests and Stochastic Gradient Boosting Machine, to three of the most capitalized coins: Bitcoin, Ethereum and Ripple. Oviedo-Gómez et al. [13] use AI to evaluate different cryptocurrency market variables through a quantile regression model to identify the best predictors for Bitcoin price prediction using machine learning models. Within the field of neural networks, already in 1981, White [14] conducted research illustrating the use of artificial neural networks in the prediction of financial variables. Since then, the study and application of Artificial Neural Networks in the field of finance and economics has increased. In the 1990s, Franses et al. [15] proposed an ANN-based graphical method to investigate seasonal patterns in time series. In more recent studies, Zhengyang et al. [16] use multiple experiments predicting Bitcoin prices separately using ANN-LSTM, where the authors use a hybrid of convolutional neural networks (CNN) with LSTM to predict the prices of the three cryptocurrencies with the largest market capitalization: Bitcoin, Ethereum and Ripple.

Žunić and Dželihodžić [17] make use of recurrent neural networks (RNN) in the prediction model of cryptocurrency values; real-world data for three cryptocurrencies (Bitcoin, Ethereum and Litecoin) were used in the experiments. Another application that AI has been given is making predictions based on the analysis of cryptocurrency investor sentiment; this is the case of Madan et al. [18], which propose a Bitcoin prediction approach based on machine learning algorithms to examine Bitcoin price behavior by comparing its variations with those of tweet volume and Google Trends data.

The objective of this research is to focus on predicting Bitcoin volatility using economic-financial variables that correlate well with the cryptocurrency. For this purpose, we use artificial intelligence models, namely machine learning (ML) and neural networks (NR), in order to compare the results between them and between traditional

statistical models such as Multivariate GARCH. The financial potential of cryptocurrencies as an investment asset is indisputable. Although also the debate between academics and financial professionals in relation to its nature. Therefore, Hazlett & Luther [19] and Yermack [20] question whether it is really a coin. Either way, it is clear that Bitcoin or Ethereum are investable assets with a high degree of diversification and return potential, and this motivates the interest of investors. Thus, the analysis of cryptocurrencies goes beyond answering the question of what type of asset it is, the main objective is to limit its characteristics as an asset: liquidity, risk and profitability. To do this, this research aims to contribute to the existing discussion, developing models that, supported by artificial intelligence, improve the volatility forecast of traditional GARCH models.

This paper aims at comparing the cryptocurrency volatility forecasting via classical and statistical machine learning algorithms.

To present this research, the paper is divided into three parts. First, an empirical comparison of volatility predictions provided by Machine Learning, Ridge, Lasso, Elastic-net, k-NN, Random Forest, Gradient Boosting and XGBoost models is performed. From the best prediction model obtained (Random Forest Regression), an optimization of its hyperparameters is performed to achieve the lowest possible prediction error. In a second part, the RNN is implemented and compared with the optimized Random Forest model, analyzing the indicators (MAE, RMSE, MAPE). The last part consists of an empirical comparison of volatility forecasts generated by M GARCH and artificial intelligence models. Machine learning methods in time series forecasting are expected to be superior to traditional econometric models.

2 Problem Description

The data used to perform the analysis have been the daily closing price of Bitcoin (BTC) and the price closes of the financial variables NVDA, RIOT, KBR, WTI, GOLD and EURUSD (see [6]) that have been selected to build the different models, both Machine Learning and Neural Networks. All data have been extracted from the Datastream® database. The sample focuses on the time window from December 2016 to May 2022. The dataset contains 2008 instances. Table 1 presents the variables used in the research and Fig. 1 shows the correlation matrix.

For the treatment of the data, first, the returns of the variables, l_t, have been calculated as a logarithmic rate $l_t = ln\left(\frac{p_t}{p_{t-1}}\right)$ where p_t are the daily prices at market close of the variable in period t. Next, we need to divide the dataset into test and training data. The first 1605 (80%) days are employed for training and the remaining 401 (20%) days for testing.

3 Methodology and Experimentation

The main contribution of this paper is a new approach to forecast the cryptocurrency volatility using on the one hand the classical machine learning approaches and on the other hand to provide a comparison against statistical machine learning methods such traditional generalized heteroscedasticity conditional autoregressive models. The data

	BTC	RIOT	NVDA	KBR	WTI	GOLD	EUR/USD	VISA
BTC	1.000000	0.870776	0.881361	0.866943	0.543102	0.710036	0.265414	0.742410
RIOT	0.870776	1.000000	0.648200	0.600050	0.397286	0.463237	0.439937	0.510409
NVDA	0.881361	0.648200	1.000000	0.893355	0.556472	0.790909	0.164781	0.753588
KBR	0.866943	0.600050	0.893355	1.000000	0.674924	0.765572	-0.002908	0.814001
WTI	0.543102	0.397286	0.556472	0.674924	1.000000	0.184845	0.010647	0.304060
GOLD	0.710036	0.463237	0.790909	0.765572	0.184845	1.000000	0.168709	0.878019
EUR/USD	0.265414	0.439937	0.164781	-0.002908	0.010647	0.168709	1.000000	0.179453
VISA	0.742410	0.510409	0.753588	0.814001	0.304060	0.878019	0.179453	1.000000

Fig. 1. Correlation matrix

Table 1. Variables

Variable	Definition
Technological variables	
NVDA	Multinational company specialized in the development of graphics processing units (GPU) and integrated circuit technologies
RIOT	Bitcoin mining company that supports the blockchain
KBR	American engineering and construction company
Commodities	
WTI	Crude oil futures
GOLD	Gold futures
Payment methods	
EURUSD	Exchange rate
VISA	Multinational financial services company

pipeline for the experimentation starts, firstly, with the data partition in training and testing sets, whose percentages have been mentioned a few lines before and, secondly, feature selection is applied only in the training set; thirdly, the projection operator allow to get the reduced testing set; fourthly, the regressor is trained using the training set; and finally, the regression model performance is assessed using the reduced testing set.

The models considered within the ML are presented below. Ridge Regression (RR) is a method for estimating the coefficients of multiple regression models in scenarios where the independent variables are highly correlated. Lasso (Least Absolute Shrinkage and Selection Operator) (LR) regression model is a method that combines regression with a procedure of shrinking some parameters towards zero and variable selection, imposing a restriction or a penalty on the regression coefficients. Elastic.net is a regularized regression method that linearly combines the penalties of the RR and LR methods. Random Forest Regression (RF) consists of a set of individual decision trees, each trained

with a slightly different sample of the training data (generated by bootstrapping). The *k*NN algorithm uses "feature similarity" to predict new data values. This means that the new point is assigned a value based on its similarity to the points in the training set, as measured by the Euclidean distance. Gradient Boosting (GB) is a generalization of the AdaBoost algorithm that allows the use of any cost function, as long as it is differentiable. It consists of a set of individual decision trees, trained sequentially and using the gradient descent loss function. XGBoost (XGB) is a set of GB-based decision trees designed to be highly scalable. Like GB, XGB builds an additive expansion of the objective function by minimizing a loss function. Neural networks are computational models composed of a large number of procedural elements (neurons) organized in layers and interconnected with each other. For time series analysis and forecasting, the single layer feed-forward network is the most commonly used model structure. See Zhang et al. [21]. The statistical methodology used is a multivariate generalization of the GARCH (p, q) model. Engle [22] proposes the Dynamic Conditional Correlation Multivariate GARCH Model (DCC-MGARCH). The choice of this model is due to its good behaviour in predicting the volatility of Bitcoin [6]. Generally speaking, to compare the prediction accuracy of each of the models, the mean absolute error (MAE), root mean square error (RMSE), and mean absolute percent error (MAPE) and R^2 are used as metrics. The best forecasts are obtained by minimizing these forecast evaluation statistics.

4 Results

This section reports the test results to forecast the cryptocurrency volatility through classical machine learning algorithms and via statistical machine learning regressors Table 2 shows the different error metrics obtained evaluating testing data for each of the models considered. They are all similar, although Random Forest is the algorithm that obtained the best results in almost all of them, except in MAPE (2.847128) which is outperformed by Lasso and Elastic-net. The last column of Table 2 shows the execution time in seconds for each of the algorithms. All of them are very similar, not reaching the second of compilation, except Gradient Boosting with 1.583719 s.

Table 2. Prediction error metrics of ML techniques

	MAE	RMSE	MAPE	Time[a]
Ridge	0.713806	1.037230	2.205566	0.015961
Lasso	0.750058	1.084110	1.037448	0.013307
Elastic-net	0.750058	1.084110	1.037448	0.005981
k-NN	0.799023	1.115097	3.158620	0.003956
Random Forest	0.726618	1.046791	2.847128	0.011244
Gradient Boosting	0.745677	1.106103	2.286150	1.583719
XGBoost	0.805525	1.177339	4.031861	0.804226

[a]Computer: Intel(R) Core (TM) i7-1185G7 and installed RAM 16.0 GB

Next, in view of the fact that Random Forest performs better than the other models for predicting Bitcoin volatility, an optimization of the RF model ORF (Optimized Random Forest) is performed. A hyperparameter adjustment is performed for a total of 3 settings. 300 hyperparameter combinations are studied to see which one produces better validation metrics. By compiling the model with the best hyperparameter combination we obtain a very significant improvement over the previous model, making it a model that fits our data almost perfectly: $\boldsymbol{R^2 = 0.996100}$**, MAE (0.0393), RMSE (0.0619) and MAPE (0.2898)**. The hyperparameters used for model optimization are shown in Table 3.

Table 3. Optimized Random Forest model Feature

Parameter name	Description	Best value
n-estimators	The number of trees in RFR	400
max-features	The largest number of features to consider when branching	sqrt
max-depth	The maximum depth of a single tree	10
min-samples-split	The minimum number of samples requided to split an internal node	2
min-samples-leaf	The minimum number of samples requided to be at a leaf node	4

Figure 2 shows graphically the good fit provided by the optimized Random Forest model for both the training set and the test set. Will Neural Networks outperform Random Forest predictive fitting?

Figure 3 plots the feature importance (variable importance); i.e., it describes which features are relevant within the RF model. Its purpose is to help better understand the solved problem and, sometimes, to improve the model by feature selection. In our case, feature importance refers to techniques that assign a score to the input variables (exogenous financial variables) based on their usefulness in predicting the target variable (Bitcoin volatility). There are different types of importance scores but, in our case, permutation importance scores have been chosen, as they are the most widely used in the literature related to RF Regression models; it shows the importance of the variables within the model. RIOT, VISA, KBR and NVDA are the features that contribute most to the model. This is anticipated as it has been shown in previous research that Bitcoin behaves more like technology variables than commodities or fiat currency.

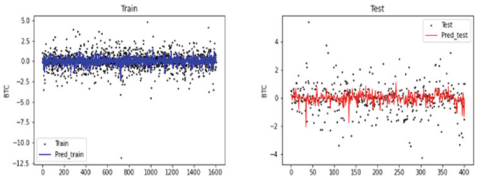


Fig. 2. ORF model adjustment.

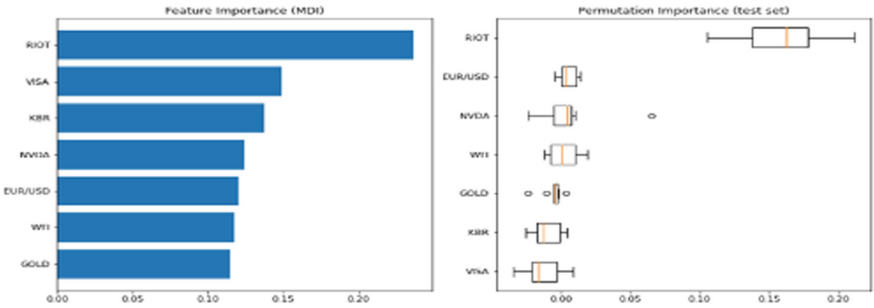


Fig. 3. Feature importance and Permutation importance (ORF model).

Deep Learning models equipped with ANN architecture are used with the objective of effectively acquiring the Bitcoin volatility movement pattern based on the same financial variables used in the ML models. Three neural networks with different number of parameters have been created (see Table 4). All of them trained for 10000 epochs, with cost function (**MSE**), optimization algorithm (**Adam**, $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate $= 0.01$) and validation metrics (Network 1: **MAE** $=$ 0.284802, **RMSE** $=$ 0.665205, **MAPE** $=$ 1.931694), (Network 2: **MAE** $=$ 0.748799, **RMSE** $=$ 1.083263, **MAPE** $=$ 1.026771), (Network 3: **MAE** $=$ 0.167559, **RMSE** $=$ 0.459763, **MAPE** $=$ 0.720910). It can be seen that their results are much worse than those obtained with the optimized Random Forest model, and in general with all ML algorithms. This is due to the aforementioned overfitting problem. The line of research is still open to obtain more data, having the same time horizon but with a higher data frequency of seconds or minutes –this is how the neural network can work. As mentioned above, daily data frequency has been considered to compare the prediction results between GARCH-M models and those provided by the AI. DCC-Model results are shown in Table 6.

Table 4. Neural networks parameters and associated time for the training.

	Parameters	Time (s)
Network 1	4737	511.826952
Network 2	14977	554.423578
Network 3	19073	594.990106

Table 5 shows the comparison of the results obtained by the statistical model and by the ML and ANN models. The last column added is the computational run time of the models. First, it is observed that the neural network models behave inefficiently compared to the ML models, in this case the optimized Random Forest model. The settings and runtimes are very high, so the networks do not seem to be a good alternative for our prediction.

The difference between the fits between DCC model and Random Forest is tiny, being improved by DCC model. This difference should not be taken into account in

Table 5. Test results: M-GARCH vs. AI

Model	MAE	RMSE	MAPE	Time (sec.)
DCC-MGARCH	0.043619	0.058076	−1.83975	398.2411
ORF	0.039304	0.061914	0.289898	0.74931
Network 1	0.284802	0.665205	1.931694	511.82695
Network 2	0.748799	1.083263	1.026771	554.42358
Network 3	0.167559	0.459763	0.720910	594.99011

view of the run time of the models. Random Forest obtains almost the same fit in a much shorter time than DCC model. While Random Forest predicts BTC volatility in less than a second, DCC model predicts it in almost 7 min. In our study, Random Forest model is definitely considered the best model for predicting BTC volatility.

5 Conclusions

Several conclusions are drawn from this research. When comparing the statistical measures of fit (MAE, RMSE, MAPE) of the ML models considered (Ridge, Lasso, Elasticnet, k-NN, Random Forest, Gradient Boosting and XGBoost), the RF model was found to be the best. However, although it is considered the best, due to the small difference between the values of the other models, its optimization is proposed (RFO). If we compare this optimal model with the M-GARCH DCC model, it appears that there is no significant difference between them to predict Bitcoin volatility. On the other hand, there is a significant difference between the runtime of the models, with the time being significantly shorter for the ORF (0.74931 sc.) versus 398.24 s for the DCC). Due to the small difference between the fit of the two models and the large difference between the execution time, the ORF machine learning model is taken as the best. Chen et al. [23] show that statistical methods perform better for low-frequency data with high-dimensional features, while machine learning models outperform statistical methods for high-frequency data. Within AI, if we compare the ORF model with ANNs, there is a big difference between the fit measures of the models. ANNs do not perform well in predicting Bitcoin volatility as there is a large overfitting problem. This may be due to the small amount of data available for the network to learn. ANNs are techniques and algorithms created for classical machine learning although bear in mind the trendy deep learning may be also use for big data; currently the highest magnitude of data storage is yottabyte. Note that for comparison purposes, this study uses the same data frequency as [6] and these results in fewer observations being available than a higher temporal data frequency scenario.

Table 6. DCC Multivariate M-GARCH model.

		Coeff	**Std. Err**	**z**
ARCH_BTC	arch L1	0.1674697	0.0240597	6.96**
	garch L1	0.8031184	0.0217433	36.94**
	_cons	0.0001231	0.0000211	5.84**
ARCH_RIOT	arch L1	0.3431427	0.0499861	6.86**
	garch L1	0.6388863	0.0447449	14.28**
	_cons	0.0003157	0.0000659	4.79**
ARCH_VISA	arch L1	0.2964985	0.0401611	7.38**
	garch L1	0.6674416	0.0340958	19.58**
	_cons	0.0000108	1.92e-06	5.60**
ARCH_NVDA	arch L1	0.2772477	0.0365308	7.59**
	garch L1	0.6096899	0.0421564	14.46**
	_cons	0.0000862	0.000014	6.16**
ARCH_KBR	arch L1	0.0919411	0.0107591	8.55**
	garch L1	0.8633705	0.0150463	57.38**
	_cons	0.0000213	4.26e-06	5.00**
ARCH_WTI	arch L1	0.1797799	0.0222289	8.09**
	garch L1	0.8049584	0.0184035	43.74**
	_cons	0.0000162	3.16e-06	5.13**
ARCH_GOLD	arch L1	0.0430722	0.0068363	6.30**
	garch L1	0.9514239	0.0088663	107.31**
	_cons	4.90e-07	2.18e-07	2.24**
ARCH_EURUSD	arch L1	0.2078883	0.1126673	1.85*
	garch L1	0.3966805	.4376379	0.91
	_cons	6.06e-06	5.16e-06	1.17
Adjustment	λ_1	.0657344	.0091189	7.21**
	λ_2	.6378354	.0722788	8.82**

*Significance level $\alpha = 0.1$; **significance level $\alpha = 0.5$.

References

1. Nakamato, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. Bitcoin, pp. 1–9 (2009)
2. Weber, B.: Bitcoin and the legitimacy crisis of money. Camb. J. Econ. **40**, 17–41 (2015)
3. Neil, G., Halaburda, H.: Can we predict the winner in a market with network effects? Competition in cryptocurrency market. Games **7**(3), 16 (2016)
4. Vassiliadis, S., Papadopoulos, P., Rangoussi, M., Konieczny, T., Gralewski, J.: Bitcoin value analysis based on cross-correlations. J. Internet Bank. Commerce **S7**(22) (2017)

5. Katsiampa, P.: Volatility estimation for Bitcoin: a comparison of GARCH models. Econ. Lett. **158**, 3–6 (2017)
6. Cebrián-Hernández, Á., Jiménez-Rodríguez, E.: Modeling of the bitcoin volatility through key financial environment variables: an application of conditional correlation MGARCH models. Mathematics **3**(9), 267 (2021)
7. Barker, J.: Machine learning in M4: what makes a good unstructured model? Int. J. Forecast. **1**(36) (2019)
8. Januschowski, T., Gasthaus, J., Wang, Y. Salinas, D., Flunkert, V., Bohlke-Scheider, M. y Lallot, C.: Criteria for classifying forecasting methods. Int. J. Forecast. **36**, 167–177 (2020)
9. Israel, R., Kelly, B.T., Moskowitz, T.J.: Can machines' learn' finance? J. Invest. Manage. (2020)
10. Panagiotidis, T., Stengos, T., Vravosinos, O.: On the determinants of bitcoin returns: A LASSO approach. Fin. Res. Lett. **27**, 235–240 (2018)
11. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodological) **58**(1), 267–288 (1996)
12. Derbentsev, V., Babenko, V., Khrustalev, K.I.R.I.L.L., Obruch, H., Khrustalova: Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices. Int. J. Eng. **1**(34), 140–148 (2021)
13. Oviedo-Gómez, A., Candelo-Viáfara, J.M., Manotas-Duque, D.F.: Bitcoin price forecasting through crypto market variables: quantile regression and machine learning approaches. Handbook on Decision Making, pp. 253–271. Springer, Cham (2023)
14. White, H.: Economic prediction using neural networks: the case of IBM daily stock returns. Neural Networks in Finance and Investing, pp. II459-II482 (1988)
15. Franses, P.H., Draisma, G.: Recognizing changing seasonal patterns using artificial neural networks. J. Econometr. **81**(1), 273–280 (1997)
16. Zhengyang, W., Xingzhou, L., Jinjin, R., Jiaqing, K.: Prediction of cryptocurrency price dynamics with multiple machine learning techniques. In: Proceedings of the 2019 4th International Conference, New York, NY, USA
17. Žunić, A., Dželihodžić, A.: Predicting the value of cryptocurrencies using machine learning algorithms. In: International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies. Springer, Cham (2023)
18. Madan, I., Saluja, S., Zhao, A.: Comercio automatizado de bitcoin a través de algoritmo de aprendizaje automático, vol. 20 (2015)
19. Hazlett, P.K., Luther, W.J.: Is bitcoin money? And what that means. Rev. Econ. Financ. **77**, 144–149 (2020)
20. Yermack, D.: Is bitcoin a real currency? An economic appraisal. In: Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data, pp. 31–43. Elsevier, Amsterdam, The Netherlands (2015)
21. Zhang, G., Patuwo, B.E., Hu, M.Y.: Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. **1**(14), 35–62 (1998)
22. Engle, R.: Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. J. Bus. Econ. Stat. **20**, 339–350 (2002)
23. Chen, Z., Li, C., Sun, W.: Bitcoin price prediction using machine learning: An approach to sample dimension engineering. J. Comput. Appl. Math. **635**, 112395 (2020)



I-DLMI: Web Image Recommendation Using Deep Learning and Machine Intelligence

Beulah Divya Kannan[1] and Gerard Deepak[1,2(✉)]

[1] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India
gerard.deepak.christuni@gmail.com

[2] Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

Abstract. Web Image Recommendation is the need of the hour because of the increasing exponential contents especially the multimedia content in the World Wide Web. The IDLMI framework has been proposed which is a query centric knowledge driven approach for Web Image recommendation. This model hybridizes Wiki Data and YAGO for entity enrichment and generation of Meta Data takes place which is further subjected to semantic similarity computation and mapped reduce algorithm for mapping and reducing the complexity within the Pearson's correlation coefficient. LSTM (Deep Learning Intrinsic Classifier) is used for automatic classification of data set. The model also classifies the data for upper ontology which enhances auxiliary knowledge. Performance of the proposed I-DLMI framework approach is calculated by utilizing F-measure, Precision, Recall, Accuracy percentages and False Discovery Rate (FDR) as the potential metrics. The proposed model furnishes the largest average precision of 96.02%, largest average recall percentage of 98.19%, largest average accuracy percentage of 97.10% and the largest F-measure percentage of 97.09%, while the lowest FDR is 0.17.

Keywords: Cosine Similarity · Image Recommendation · LSTM · Semantics · Shanon's Entropy

1 Introduction

Digitization has increased in unprecedented rates in the modern times. It has also proved helpful for various industries starting from Medicine, Trade, Public services and finance. What digitization actually means is that it converts the information received into digital format and it is used in various business models and provide an intuitive way of dealing with revenue. The information of the World Wide Web has increased, the end-users have increased because of internet availability. Today everything is connected, users are connected with the internet and henceforth data is increasing exponentially higher and the web is the most dynamic entity present today. As Data is highly increasing at an exponential rate, the web is to be configured into Web 3.0 according to Sir Tim

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 270–280, 2023.
https://doi.org/10.1007/978-3-031-27409-1_24

Berner Lee. The Web 3.0 is a semantic structure of the web where the density of the web data is quite high and every entity of the web is linked. The use of Multimedia has increased unprecedently due to the present day You-tubing culture and various social media platforms like Instagram, Flickr, Twitter etc. Every image found on the World Wide Web must be annotated, tagged or labelled. Only an annotated image will be retrieved rightly but in the present-day scenario, the number of images uploaded is excessively large, high and exhaustive. Tagging is just a mere optional phenomenon which should not be the case, the recommendation of images is essential especially on Web 3.0 of images.

Motivation: Owing to the structural density of the Web 3.0 which is exponentially increasing, multimedia contents and web image contents on the internet is a requisite necessity and special strategies, paradigms and models are used for retrieving images from Web 3.0 which is the semantic standard of the web that is of utmost importance.

Contribution: The known contribution includes the encompassment of hybridization of Wiki Data and YAGO knowledge sources stores for entity enrichment and Meta Data Generation. The approach of Meta Data generation is to yield meta data pool of entities with computation of semantic similarity and Shanon's entropy subjecting it to Map Reduce with Pearson Correlation Coefficient which are the key contributors. The employment of LSTM (deep learning intrinsic classifier) is used to classify data set and upper ontologies of the proposed model.

Organization: The rest of the paper is organized as follows. Section 2 depicts the Related Works. Section 3 depicts the Proposed System Architecture. Section 4 depicts the Implementation. Section 5 depicts Performance Evaluation. Paper is concluded in Sect. 6.

2 Related Works

Rachagolla et al. [4] has come with a strategy for recommending events with the help of Machine Learning. This system proposes a framework by examining data about events and by recommendation of good events to the ones who are not aware of where they are, so this system recommends accurate events for them. Meng et al. [5] proposed a model which supports Propagation which is cross modal for recommending images. The paper deals with the process. Of cross-modal manifold propagation (CMP) for the recommendation of images. CMP supports visual dissemination to report visual records of users by depending on semantic visual manifold. Chen et al. [6] proposed a recommendation model by integrating knowledge graph and Image features. A multimodal and a recommendation model has been used that incorporates Knowledge Graph with Image (KG-I) features. This model also uses visual embedding, knowledge embedding and structure embedding. Deepak et al. [7] proposed an intelligent based model for social relevance term accumulation for the recommendation of pages from the web. The data is extracted using WordNet. Classification algorithms like Random Forest is employed and Optimization of Ant Colony is used to find the shortest distance with the help of graphs. Yung et al. [8] dealt with a recommendation model for the web browser inbuilt with Augmented reality. This system creates a web browser encompassed with AR and recommends a web browser called as A2W browser that provides continuous

user-driven web browser occurrences influenced by headsets inbuilt with AR. Depeng et al. [9] proposed a Deep knowledge-aware framework to recommend web services. The framework proposes a graph based on knowledge and represents web services recommendation and an attention module. A neural network that uses Deep Learning has been dealt with to create the big-level attributes of user-service attributes. Le et al. [10] deals with an Attention Model with hierarchy to recommend images. Matrix factorization is used in this paper, this system finds important aspects that influences user's untapped preferences. Wan et al. [11] proposed a customized Image Recommending prototype paramount to Photo and buyer-thing collective process. To incorporate customized recommendation this model uses Bayesian Personalized Ranking. An attention mechanism is also introduced in this model to indicate users' different predilection on their concerned images. Viken et al. [12] dealt with a recommendation system, recommending tourist places using convolutional neural network. This system deals with a phone application that takes the user's preference and recommends hotels, restaurants and attractions accordingly. This system uses K-modes clustering model that is used for training the dataset. Xianfen et al. [13] proposed a Web page recommendation model using twofold clustering method i.e. behavior of user and topic relation. This system combines density-based clustering and k-means clustering. Amey et al. [14] proposed a Face emotion-based music recommendation system. This paper proposes a smart agent that sorts music according to the emotions expressed in each song and then recommends a song album that depends on the user's emotion. In [14–20] several models in support of the proposed literature have been depicted.

3 Proposed System Architecture

Figure 1 illustrates the suggested strategy of the system architecture of semantically inclined map reduced based web image recommendation framework in which the query of the user is considered as input and the query of the input is put through pre-processing. Pre-processing pertains to removal of stop words, lemmatization, tokenization and named entity identification. The input query is enriched and yielded as user query word. Query words are then further sent into Wiki Data and Yago (knowledge stores or knowledge bases). The query words are sent into WikiData through the WikiData API and the matching relevant entities from WikiData are knowledge based/knowledge stored are then harvested.The entities which are outcome of WikiData are sent into the YAGO knowledge base for further yielding of entities which are relevant to the Query words. Finally, after the Query Preprocessing Phase, the entity enrichment takes place by leveraging and harvesting the entities from WikiData and Yago knowledge bases. These entities from the Query words, Wiki Data and YAGO knowledge bases are used to generate the Metadata. The Metadata is generated using the Dspace meta-tag harvester as shown in Fig. 2.

Dspace is an digital dynamic opensource repository. Web based interfacing makes it easy for the user to create item that get archived by depositing files. Dspace is created to deal with any format from simple text files to complex datasets. An archival item consists of related, grouped content and **metadata**. The metadata of the item is indexed respectively for browsing purposes. Dspace provides **functional preservation**. When

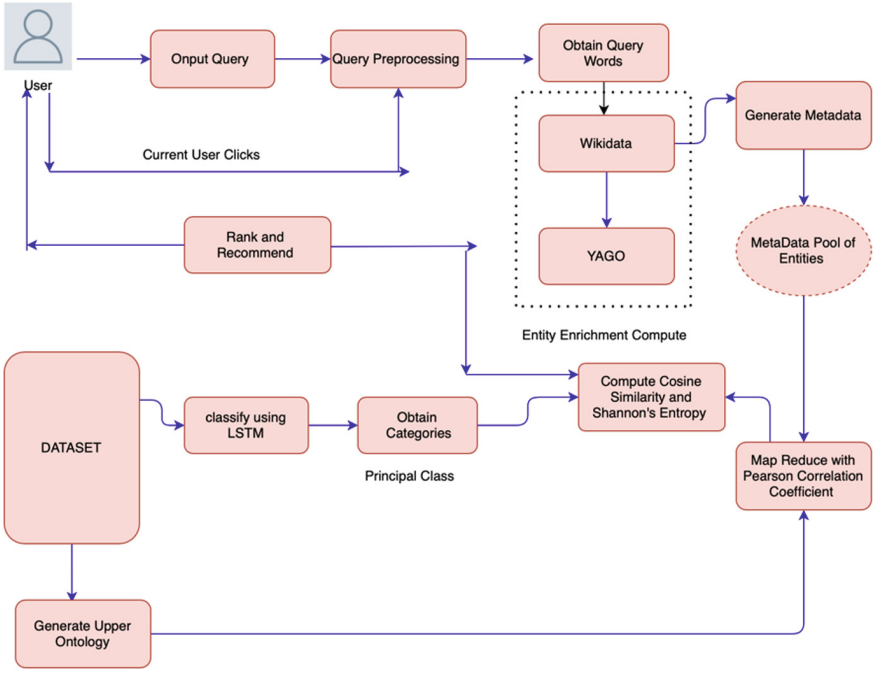


Fig. 1. Proposed System Architecture

the item is found, the Web-native formatted files is showed in the Web browser and other non-supportable formats are to be opened with other application programs respectively.

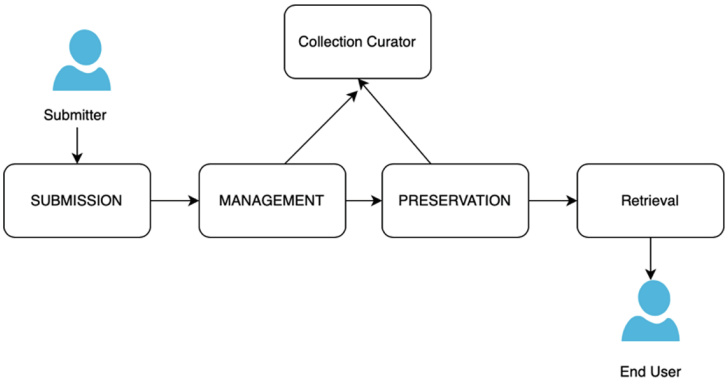


Fig. 2. Dspace Meta-Tag Harvester

The Meta Data Generation results in yielding a meta data pool of entries which is stored in a separate space that is further used for computation. The next stage necessitates the pre-processing of the data set. The pre-processed categorical web image data set is classified intrinsically by using the LSTM classifier. LSTM is a deep learning intrinsic

classifier. Long short-term memory (LSTMs) is a neural network artificially made and branches from Deep Learning. LSTMs are used to solve the problems faced by RNN's. RNN suffers from continual reliance problem. If more and more information piles up then, RNN becomes less effective in learning things. LSTM makes the neural network to retain the memory of the stuff that it needs to keep hold of context, but can also forget the stuff that is no longer necessary.

LSTM classifies the dataset by auto crafted feature selection. The classified instances from the LSTM are used to mark the principal classes i.e., the classes discovered by the LSTM classifier and subsequently the data set is used to generate the upper ontologies. The upper Ontologies are generated using OntoCollab and StarDog as tools, however in order to ensure Upper ontology is being used, the only three hierarchies of ontologies are retained by eliminating the fourth hierarchy from the root node thus eliminating other individuals.

The generated Upper Ontologies are further linked and mapped with the Meta Data pool of entities and only the entities which are relevant to the upper Ontologies from the meta data pool of entities are retained in the Upper Ontology Map in order to formulate subsequent process of knowledge subgraphs. This is done by using Map Reduced Algorithm along with Pearson's Correlation Coefficient with a threshold of 40% that is taken into consideration.

The MapReduce contains two important tasks, Map and Reduce. This programming model pushes your code into multiple servers and those servers process and run the code using MapReduce. The mapper class in the MapReduce algorithm takes in the input, tokenizes the input, maps, shuffles and sorts the input. The reducer class on the other hand searches and reduces the input and gives out the respective output. Pearson's Correlation Coefficient is depicted by Eq. (1).

$$\frac{\rho_{xy=} Cov(x, y)}{\sigma x \sigma y} \tag{1}$$

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{N} (x_i - \mu x)(y_i - \mu y) \tag{2}$$

Cov(x,y) defines the covariance of (x, y) where n is the number of data points and is depicted by Eq. (2).

X is the data values of x

Y is the data values of y

μ_X *is the mean of* x

μ_y *is the mean of* y

σ_x *is the standard deviation of* x

σ_y *is the standard deviation of* y

Pearson's Correlation Coefficient states the strength and direction of Relation between two variables we take into account.

The Semantics Invalidity is subsequently computed from the knowledge subgraphs and the principal categories which are outcome of the classified LSTM entities that are

used to compute the cosine similarity along with Shanon Entropy. The cosine similarity is set as 0.75 and the step deviation of Shanon's entropy is set as 0.25. It is because the Relevance is very high for cosine similarity and step deviation is moderately high in term of Computation of Shanon's Entropy.The Cosine Similarity states if two points are similar or not. It measures the similarity between two points in vector space. It is measured by taking the angle between the points P1 and P2

$$\text{Similarity} = \text{Cos}(\theta) = \frac{\text{A.B}}{||\text{A}||\text{B}||} \tag{3}$$

Equation (3) depicts the formula for cosine similarity. Shanon's Entropy on the other hand measures the uncertainty of probability distribution and is depicted in Eq. (4)

$$H(x) = -\in_x P(x)\log P(x) = \in_x P(x).\log\left(\frac{1}{P(x)}\right) \tag{4}$$

P(x) measures the probability of event X.

1/P(x) measures the amount of information.

Ultimately the coordinated entities are reranked in ascending form of the similarity in semantics and is suggested along with all the matched images comprising of these entities to the user. If the user has been satisfied with search, the recommendation stops, If the user is dissatisfied, the current user is then captured and sent further for preprocessing and the process goes on until there are no user clicks available i.e., when the user has come with consensus with the image recommended.

4 Implementation

The implementation of the paper was carried using Python, the recent version of Python that uses i7 processor under Intel Core inbuilt with a ram of 16 GB. Python's Natural Language Toolkit (NLTK) was made use of in order to carry out the language processing task. Ontology was semi-automatically modelled using OntoCollab and static ontologies using WebProtege. This paper uses three datasets. Experimentations are conducted using the first dataset i.e., Stylish Product Image Dataset which contains 65,000 Records of Fashion Product Image [21].The second dataset is Recommender Systems and Personalization Datasets which has been used [22]. The third data set used is Various Tagged Images and Labeled images suited for multi-label classifiers and recommendation systems [23]. A large dataset is synthesized by integrating and committing three distinct participant datasets in which the Stylish Product image Dataset contains 65,000 Records of Fashion Product image. Recommendation systems and personalization datasets is considered for each of these stats using customized image crawler, images are crawled and annotated and nearly 78,000 records of several art images relevant to the second dataset is available and is tagged by using the tags available in Recommendation Systems and Personalization Datasets i.e. UCSD CSE Research Project, Behance Community Art Data [Dataset]. The third dataset participating is Various Tagged Images from Kaggle by greg [24] In this dataset, labeled images are suited for multi-label classifiers and recommendation systems. These three datasets are further annotated using customized

annotators and they are used for implementation. Experimentations are conducted for the same datasets for both baseline models and the proposed model. The baseline model is evaluated for the same dataset as for the proposed model.

5 Results and Performance Evaluation

The carrying-out of the suggested I-DLMI framework is calculated using F-measure, Accuracy, Recall and Precision. Percentages and False Discovery Rate (FDR) as potential metrics. Accuracy, Recall, F-measure and Precision talks about the relevancy of results. The FDR takes account of the counts of faulty positives yielded by the framework. From Table 1, I-DLMI model's performance is computed for 5258 Queries where the ground truth is assimilated for over a period of 144 days from 912 users. The I-DLMI is baselined with NPR, AIRS, NWIR models in order to compare and benchmark the I-DLMI model. In order to ensure proper relevance of the results, the performance of NPRI, AIRS, NWIR models were evaluated for the same set of I-DLMI model and they are tabulated in Table 1.

Table 1. Comparison of Performance of the proposed SIRR with other approaches

Search Technique	**Average Precision %**	**Average Recall %**	**Accuracy %**	**F-Measure % (2*P*R/(P + R)**	**FDR (1-Precision)**
NPRI [1]	83.22	86.35	84.78	84.75	0.17
AIRS [2]	85.22	88.17	86.69	86.66	0.15
NWIR [3]	90.12	92.36	91.24	91.22	0.10
Proposed I-DLMI	**96.02**	**98.19**	**97.10**	**97.09**	**0.04**

Table 1 indicates the proposed I-DLMI structure which brings in the highest average precision of 96.02%, highest average recall percentage of 98.19%, highest average accuracy percentage of 97.10% and the highest F-measure percentage of 97.09%, while the lowest FDR is 0.17.The reason why NPRI model generated the minimal precision, F-measure, Accuracy and Recall with the lowest false discovery rate (FDR) is due to the reason that NPRI framework incorporates neural Bayesian personalized ranking. The incorporation of this neural network in the absence of auxiliary knowledge of inference makes the computational load very high. It only depends upon features and since the text feature in the dataset is definitely sparse, the neural network with personalized ranking does not work. The neural network becomes sparse and very indistinct. So, due to this reason, the NPRI model lags to its core.The AIRS model also does not perform as expected mainly due to the reason that it is a combination of both visual and semantic information. It is highly specific to a domain. But since two deep learning models are used and is completely driven by images, it matches the image features and text features. Since Image feature cannot directly be related with text features and also due to

the absence of auxiliary knowledge to promote the text, and absence of strong relevance computation mechanism makes this model definitely indistinct compared to the other models.

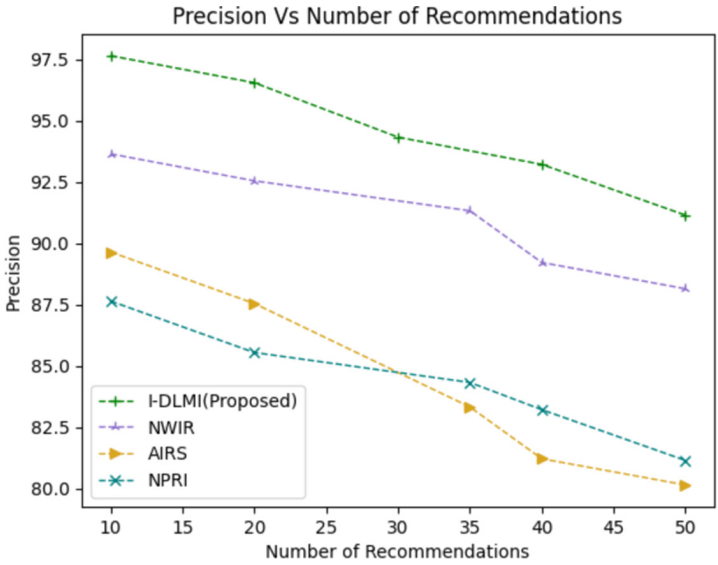


Fig. 3. Precision % versus Number of Recommendations Distribution Curve

NWIR model also does not perform well, although the performance is comparatively reliable than the other two baseline models, this model lags when compared to the proposed IDLMI model mainly for the reason it incorporates image retrieval model using bagging, weighted hashing, local structure information. Local structure information is extracted which ensures a small amount of auxiliary knowledge but relevance computation mechanism in this model is definitely highly insignificant and also the knowledge collected is significant. As a result, this model does not perform well and not the best fit.The I-DLMI model is the ideal model which is going to be used. The reason why the proposed I-DLMI model for web image recommendation is definitely better than the other baseline approaches because, it includes upper ontologies. Firstly, the upper ontologies generate a significant amount of knowledge and it performs better than the detailed ontologies because upper ontologies have significant concept distribution and relevancy is maintained because the detailed ontologies which becomes insignificant as the level increases. Secondly, the query classification is done using LSTM.The data set is classified using LSTM deep learning model where the features are automatically generated and the classification is highly accurate. The query is enriched by obtaining query words and passing it onto WikiData and YAGO model where entity enrichment takes place. Serial enrichment takes place using WikiData and YAGO of entities which in turn generates Meta Data.Entity Enrichment with heterogeneity takes place which increases knowledge. Exponential knowledge increases by generating Meta Data and relevant knowledge discovery of data set are done by using Upper Ontologies. Apart

from a very strong Cosine Similarity with Shannon's entropy, semantic similarity and computation model for relevance computation is present and Map Reduced based aggregation using Pearson's correlation coefficient that ensures the proposed I-DLMI model which performs much better than the base line models.

Figure 3 depicts the line graph of Number of Recommendations distribution Vs Precision curve for all the approaches. It is clear that the given I-DLMI model occupies the highest in the hierarchy. The NWIR model occupies the second in hierarchy. The AIRS occupies the third in hierarchy. The NPRI model occupies the fourth in hierarchy. The I-DLMI model occupies the first in the hierarchy because the model includes upper ontologies that consists of significant concept distribution and relevancy. The other models do not perform as great as the I-DLMI model. The disadvantage of NRPI model is that, it incorporates neural personalized ranking model, the incorporation of neural network in the absence of auxiliary knowledge for inference makes the computational very high. The data set in this model is highly sparse, so the neural network with personalized ranking does not work. The disadvantage of AIRS model is that, it is a combination of both visual and semantic information. Absence of auxiliary knowledge and absence of strong relevance computation makes this model definitely indistinct. On the other hand, The NWIR model lags when compared to the proposed I-DLMI model because this model incorporates retrieval of images using weighted hashing, bagging and local structure information. This model has a small amount of auxiliary knowledge and relevance computation in this mechanism is high which makes this model insignificant.

6 Conclusions

In this paper, Web image recommendation using Deep Learning and Machine Intelligence has been proposed. Due to the exponential increase in the information, Web Image recommendation has become of utmost importance. Every image found on the World Wide Webmust be annotated, tagged and labeled and hence an annotated image will be retrieved rightly. In this paper, the IDLMI framework that has been suggested is a query centric knowledge driven approach for Web image recommendation. This model hybridizes Wiki Data and YAGO for entity enrichment and enrichment of Meta Data. i.e., subjected to semantic similarity computation and Mapped Reduced Algorithm for Mapping and reducing the complexity. The next phase involves the pre-processing of the data set. The pre-processed categorical web image data set is classified intrinsically by using the LSTM classifier. The upper Ontologies are generated using OntoCollab and StarDog as tools. The execution of the proposed I-DLMI framework approach is calculated using F-measure, precision, Recall, Accuracy percentages and False Discovery Rate (FDR) which furnishes the highest average precision of 96.02%, highest average recall percentage of 98.19%, highest average accuracy percentage of 97.10% and the highest F-measure percentage of 97.09%, while the lowest FDR is 0.17.

References

1. Niu, W., Caverlee, J., Lu, H.: Neural personalized ranking for image recommendation. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 423–431 (2018

2. Hur, C., Hyun, C., Park, H.: Automatic image recommendation for economic topics using visual and semantic information. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pp. 182–184. IEEE (2020)
3. Li, H.: A novel web image retrieval method: bagging weighted hashing based on local structure information. Int. J. Grid Util. Comput. **11**(1), 10–20 (2020)
4. Varaprasad, R., Ramasubbareddy, S., Govinda, K.: Event recommendation system using machine learning techniques. In: Innovations in Computer Science and Engineering, pp. 627–634. Springer, Singapore (2022)
5. Jian, M., Guo, J., Fu, X., Wu, L., Jia, T.: Cross-modal manifold propagation for image recommendation. Appl. Sci. **12**(6), 3180 (2022)
6. Chen, Q., Guo, A., Du, Y., Zhang, Y., Zhu, Y.: Recommendation Model by Integrating Knowledge Graph and Image Features. **44**(5), 1723–1733 (2022)
7. Surya, D., Deepak, G., Santhanavijayan, A.: KSTAR: a knowledge-based approach for socially relevant term aggregation for web page recommendation. In: International Conference on Digital Technologies and Applications, pp. 555–564. Springer, Cham (2021)
8. Lam, K.Y., Lee, L.H., Hui, P.: A2w: Context-aware recommendation system for mobile augmented reality web browser. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2447–2455 (2021)
9. Dang, D., Chen, C., Li, H., Yan, R., Guo, Z., Wang, X.: Deep knowledge-aware framework for web service recommendation. J. Supercomput. **77**(12), 14280–14304 (2021). https://doi.org/10.1007/s11227-021-03832-2
10. Wu, L., Chen, L., Hong, R., Fu, Y., Xie, X., Wang, M.: A hierarchical attention model for social contextual image recommendation. IEEE Trans. Knowl. Data Eng. **32**(10), 1854–1867 (2019)
11. Zhang, W., Wang, Z., Chen, T.: Personalized image recommendation with photo importance and user-item interactive attention. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 501–506. IEEE (2019)
12. Parikh, V., Keskar, M., Dharia, D., Gotmare, P.: A tourist place recommendation and recognition system. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 218–222. IEEE (2018)
13. Xie, X., Wang, B.: Web page recommendation via twofold clustering: considering user behavior and topic relation. Neural Comput. Appl. **29**(1), 235–243 (2016). https://doi.org/10.1007/s00521-016-2444-z
14. Pawar, A., Kabade, T., Bandgar, P., Chirayil, R., Waykole, T.: Face emotion based music recommendation system. http://www.ijrpr.com. ISSN 2582, 7421
15. Surya, D., Deepak, G., Santhanavijayan, A.: KSTAR: a knowledge based approach for socially relevant term aggregation for web page recommendation. In: International Conference on Digital Technologies and Applications, pp. 555–564. Springer, Cham (2021)
16. Deepak, G., Priyadarshini, J.S., Babu, M.H.: A differential semantic algorithm for query relevant web page recommendation. In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 44–49. IEEE (2016)
17. Roopak, N., Deepak, G.: OntoKnowNHS: ontology driven knowledge centric novel hybridised semantic scheme for image recommendation using knowledge graph. In: Iberoamerican Knowledge Graphs and Semantic Web Conference, pp. 138–152. Springer, Cham (2021)
18. Ojha, R., Deepak, G.: Metadata driven semantically aware medical query expansion. In: Iberoamerican Knowledge Graphs and Semantic Web Conference, pp. 223–233. Springer, Cham (2021)
19. Rithish, H., Deepak, G., Santhanavijayan, A.: Automated assessment of question quality on online community forums. In: International Conference on Digital Technologies and Applications, pp. 791–800. Springer, Cham (2021)

20. Yethindra, D.N., Deepak, G.: A semantic approach for fashion recommendation using logistic regression and ontologies. In: 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1–6. IEEE (2021)
21. Deepak, G., Gulzar, Z., Leema, A.A.: An intelligent system for modeling and evaluation of domain ontologies for crystallography as a prospective domain with a focus on their retrieval. Comput. Electr. Eng. **96**, 107604 (2021)
22. Kumar, S.: Stylish Product Image Dataset (2022). https://www.kaggle.com/datasets/kuchhbhi/stylish-product-image-dataset
23. UCSD CSE Research Project, Behance Community Art Data. https://cseweb.ucsd.edu/~jmcauley/datasets.html
24. greg: Various Tagged Images (2020). https://www.kaggle.com/greg115/various-tagged-images

Uncertain Configurable IoT Composition With QoT Properties

Soura Boulaares[1(✉)], Salma Sassi[2], Djamal Benslimane[3], and Sami Faiz[4]

[1] National School for Computer Science, Manouba, Tunisia
boulaaressoura@gmail.com
[2] Faculty of law, Economic, and Management Sciences, Jendouba, Tunisia
[3] Claude Bernard Lyon 1 University, Lyon, France
[4] Higher Institute of Multimedia Arts, Manouba, Tunisia

Abstract. Concerns about **Quality-of-Service (QoS)** have arisen in an **Internet-of-Things (IoT)** environment due to the presence of a large number of heterogeneous devices that may be resource-constrained or dynamic. As a result, composing **IoT** services has become a challenging task. At different layers of the **IoT** architecture, quality approaches have been proposed that take a variety of QoS factors into account. Things are not implemented with QoS or exposed as services in the **IoT** context. Actually, the **Quality-of-Thing (QoT)** model of a thing is composed of duties that are each associated with a set of non-functional properties. It is difficult to evaluate the QoT as a non-functional parameter for heterogeneous thing composition. Uncertainty emerges as a consequence of the plethora of things as well as the variety of the composition paths. In this paper, we establish a standard method for aggregating Things with uncertainty awareness while taking **QoT** parameters into account.

Keywords: QoT · IoT Composition · Uncertainty · Configuration

1 Introduction

The **Internet-of-Things (IoT)** is a network of physical and logical components that are connected together for the purpose of exchanging information and serving the needs of an **IoT** service. In an open,dynamic and heterogeneous environment like the **IoT**, coming up with the right cost for a product or service is always "challenging." The availability of possible products and services, the characteristics of potential customers, and legislative act are just a few of the many factors that affect pricing. **IoT** is another **ICT** discipline that would benefit greatly from the ability to differentiate similar Things.

Therefore, how to select the "right" and "pertinent" things? To the best of our knowledge, Maamar et al. [15] developed the **Quality-of-Things (QoT)** model as a selection criterion with an **IoT** specificity (like **Quality-Of-Service (QoS)**). This model consists of a collection of nonfunctional attributes that

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 281–290, 2023.
https://doi.org/10.1007/978-3-031-27409-1_25

are specifically targeted at the peculiarities of things in terms of what they do, with whom they interact, and how they interact. In this essay, we refer to the functions that things carry out as their duties and categorize them into sensing, acting, and communicating.

Compared to existing works that handled **QoS** [5,13,16,17,19,20] , things are not exposed as services nor adopt **QoS**. In this paper we consider the duties of Thing. Hence, things are associated with their behaviour or duties having each a set of non-functional properties that constitute **QoT** model of a thing [12,15,18]. In the context of service composition, **QoS** was proposed for the composition process or the the workflow patterns [6]. **QoS** has been a major preoccupation in the fields of networking, real-time applications and middle-ware. However, few research groups have concentrated their efforts on enhancing workflow systems to support Quality of Service management[6].

Some work have been focused on handling the **QoS** over the composition process of the workflow patterns such as in [7] Thus, in the **IoT** context, the composition process is variable due the dynamic changes related to the environment and the the things relations and nature [4]. In fact , the **IoT** composition faces two main challenges, such as the configuration and the uncertainty. The variability is is handled through a configurable composition language (**Atlas+**) that we have presented in a previous work [4]. In the context of data composition, it consists of the best aggregation of composite services[1–3].Uncertainty generalizes incompleteness and imprecision of the composition process. It's related to which path of composition could be executed regarding a **QoT** of the same Thing duty. This challenge was modeled with the **QoS** non-Functional properties [8,10,11,14,21,22].

As a result, we aim to model the **QoT** through the configurable composition patterns formulated by **Atlas+** with uncertainty awareness. Our approach is based on the configurable composition patterns [4] and the classic composition patterns [9]. The main challenge is how to adapt the **QoT** to be represented by the new Framework and the handle the uncertainty of the configurable composition.

The rest of the paper is structured as follows.In Sect. 2 we present the related works, in Sect. 3. we represent our configurable composition based **QoT** with uncertainty awareness framework, in Sect. 4 we validate our approach by the proposed algorithms. And we conclude our work in the last section.

1.1 Motivation Scenario

Our illustrative scenario concerns the general **IoT** configurable composition where several composition plans could be selected to establish an **IoT** service composed from multiple composite services [4]. In Fig. 1, a composite Thing Service **(TS)** (a duty of a thing) has two execution paths with probabilities as p1 = 0.7 and p2 = 0.2 respectively.

There is one TS in each execution path. Taking the same Thing and with respect to the QoT model [15], multiple value could coexist depending on it's

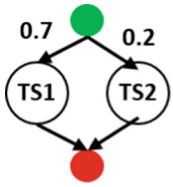


Fig. 1. Composition scenario

availability at a certain time or on the result of the previous composition execution. Let's consider response time and energy as QoT for assessing a certain TS. Each of which having different values for each possible **TS**. The result are depicted in Table 1.

Table 1. QoT for the available TS

Thing Service (TS)	Availabe TS	Response Time(T)	Energy(E)
Sensing TS1	TS11	10	10
	TS12	40	20
Sensing TS3	TS21	60	20
	TS22	100	10

The user requirements on **QoT** of the composite thing are: the response time should be less than 50 and the energy should be no more than 40°. In fact, when the energy is more then 40° we accord it the rate 20 else the rate is 10. The previous table depict that there are four possible aggregation, yet, we are unable of determining which is the most suitable path. The aggregation method is used to compute the final **QoT** and rank each of them according to the highest **QoT** [6,23]. As a result:

$$TimeofTST = p1 * T1 + p2 * T2 \tag{1}$$

$$EnergyofTSE = p1 * E1 + p2 * E2 \tag{2}$$

According to formulas 1, 2, and Table 1, there are four possible compositions where:

- composition 1: TS11 with TS21 with T=43 and E=19
- composition 2: TS11 with TS22 with T=67 and E=13
- composition 3: TS12 with TS12 with T=64 and E=26
- composition 4: TS12 with TS12 with T=88 and E=20

The above analysis shows that single values are not enough to represent the **QoT** of a composite TS and the lack of information about probability of each execution path prevents the effective selection for composite TS's component. There are ups and downs in the QoT of a composite service to execute different

paths. Adding QoT constraints on every execution path without considering the probability makes the user specified requirements too difficult to be satisfied.In fact, the thing composition should be built up by basic composition patterns, including parallel and conditional. Uncertainty should take into consideration all the patterns and fulfill the optimal aggregation. As a result we define the following challenges:

- Modeling method for the **IoT** composition configurable and classic patterns.
- Modeling method for the QoT of a component **IoT**.
- QoT estimation method for Thing composition patterns.

2 Background

2.1 Quality-of-Things Model

The **QoT**proposed in [12,15] define the non-functional properties of a Thing related to it's duties. This model would revolve around three duties : sensing (in the sense of collecting/capturing data), actuating (in the sense of processing/acting upon data), and communicating (in the sense of sharing/distributing data) as in Fig. 2.

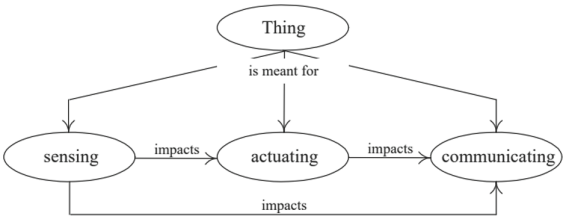


Fig. 2. Duties upon which a thing's QoT model is built [15]

As a result:

- A thing senses the surrounding environment in order to generates some outcomes.
- A thing actuates outcomes basic on the result from sensing.
- A thing communicates with environment based on the results of sensing and actuating.

2.2 Configurable IoT Composition and Workflow Patterns

A configurable composition plan reference model **(CCRM)** is a graph oriented with two essential components: nodes and links[4]. The nodes can be, for the atlas+ language, the primitives; Thing Service **(cTS/ TS)**, Thing Relationship **(cTR/ TR)**, Recipe **(cR/ R)** and operations or connectors such as **OR**,

exclusive Or **(XOR)** and **AND**. These connectors defines the configurable composition patterns. On the other hand, for the classic composition several patterns have been presented using the **QoS** based on the notion of the workflow such as sequence, parallel , conditional and loop [6,7,9,22].

3 State of the Art

Based on our reviews, limited are the analysis of **QoT/QoS**-sensitive **IoT** uncertain composition was produced. In this section we summarize the main and nearest works related to service composition **IoT** mainly Web Service Composition: In [16], the authors proposed a comparative study of some approaches of the composition of **IoT** services that are sensitive to quality of service. They made a comparison from the algorithms used, the majority of which are heuristics or meta-heuristics. In [12,15] a new model for addressing Quality-of-Things was proposed that consider the the non-functional properties related to thing duties. In [18], the authors presented an approach for the development of an ontological web language based on the OWL, called OWL-T (T means task). It could be used for users formally and semantically describing and specifying their needs to a high level of abstraction that can then be transformed into executable business processes by the underlying systems. The OWL-T aims to facilitate the modelling of complex applications or systems without considering the technical and low-level aspects of the underlying infrastructure. In the context of uncertainty,the authors [7] handled the **QoS**-oriented composition and defined an optimised composition plan with uncertain nonfunctional parameters for each Web Service. In [22], the authors proposed a probabilistic approach for handling service composition. The proposed approach handled any type of **QoS** probability distribution. In [14], modeled the problem of the uncertain **QoS**-aware web service composition with interval number and transform it into a multi-objective optimisation problem with global **QoS** constraints of user's preferences. In [6], the authors presented a predictive **QoS** model that makes it possible to compute the quality of service for workflows automatically based on atomic task **QoS** attributes.

Based on a review of the literature, we found that the most important quality of service attributes were response time, cost, availability and reliability. Energy consumption and location are two attributes that are important in the composition of **IoT** services. This can also be justified by the need for energy optimization of connected objects that are also closely related to the physical world. Thus, QoT modelling with knowledge of uncertainty in the **IoT** context with respect to composition patterns has not been addressed in any previous work.

4 Configurable Composition Based QoT with Uncertainty Awareness

4.1 Overview Of The QoT-Composition Architecture

The general architecture of our approach Uncertain configurable approach based on QoT (QoT-UCC) is depicted in Fig. 3. The first model consist of the QoT

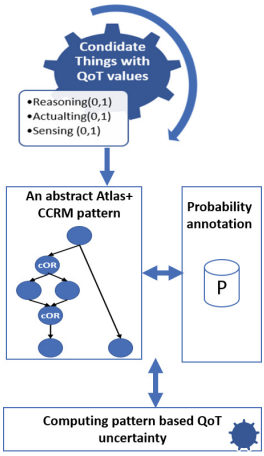


Fig. 3. The QoT-UCC architecture

definition in each Thing. Next the composition model is based on Atlas+ CCRM with probability annotation. Finally the uncertainty of the final composition is calculated through the patterns' formulas. Our approach will be detailed in the following sections.

4.2 The Composition Patterns Definition

Workflow control patterns in real-life business scenarios have been identified in several approaches,especially, in **IoT** context we defined a CCRM [4] that handles only the four composition patterns such as sequence, AND, OR and XOR. In our approach the loop patterns is not handled. These basic patterns include are the same as sequential pattern, parallel pattern, conditional pattern. Fig. 4 depicts the uncertain configurable composition patterns. In each composition patterns the probability of the path is depicted as p $p \in [0, 1]$ and each primitive of the composition patterns could be: a TS: Thing service, a TR:Thing Relationship or an R: thing recipe each of which is explained in [4]. In our example we show only the TS in each pattern. Each of which defines a specific composition of the possible primitive (TS/TR or R [4]).

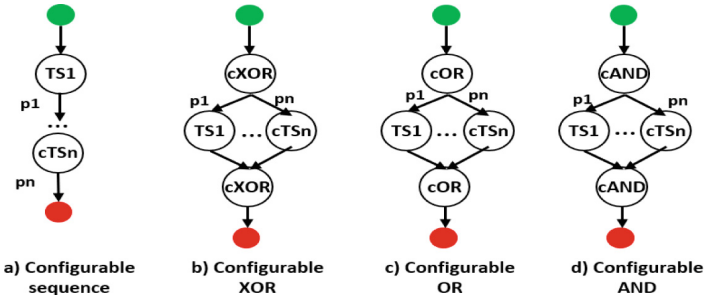


Fig. 4. Different Patterns of the Configurable Composition Model

- **(a) the configurable sequence:** is the set of primitives executed in sequence. A sequence may contain configurable and non-configurable primitives. A sequence could be active or blocked.
- **(b) configurable AND:** Configurable AND (cAND) is configured in classic AND. The AND connector consists of two or more parallel branches (in the case of a relationship or service block).
- **(c) configurable OR:** consists of sequences with n possibilities.And three possible connectors(OR,AND,XOR).
- **(d) configurable XOR:** it can be configured in sequence or in conventional XOR. This connector consists of two or more branches and among these multiple branches. Where, one and only one will be executed.

The composition is defined as the aggregation of all the possible customised configurable patterns. Each of the configurable composition pattern could be customised into a classic composition patterns as in Table 2.

Table 2. Configurable patterns customisation

	Classic Sequence	Classic AND	Classic OR	Classic XOR
Configurable Sequence (cSequence)	X			
Configurable AND (cAND)		X		
Configurable OR (cOR)	X	X	X	X
Configurable XOR (cXOR)				X

4.3 QoT Probability Aggregation Formulas for Composition Patterns

Based on **QoS** principle, the **QoT** metrics are classified into five categories according to their characteristics in different composition patterns, which are as follows: additive, multiplicative, concave (i.e., minimum), convex (i.e., max-

imum), and weighted additive. In each configurable composition we define the appropriate formula:

- **Sequential pattern:** let the probability of the incoming paths be $P_x = p_i \, where \, p \in [0,1] \; and \, i \in [1,n]$. The **QoT** of the energy consumption values are the possible VE_x and VE_y for the general E_x. As well, the **QoT** of the execution time are the possible VT_x and VT_x for the general T_x. Sequential are calculated as:

$$Energy: \; E_x = p_i \times VE_x + p_n \times VE_y \tag{3}$$

$$Time: \; T_x = p_i \times VT_x + p_n \times VT_y \tag{4}$$

- **AND pattern:** For AND patterns, let be P_x is the probability of the incoming paths. The **QoT** of each Energy and Time values is E_x and T_x respectively. The calculation pattern is as follows:

$$Energy: \; E_x = \sum_{i=1}^{n} p_i E_i \tag{5}$$

$$Time: \; T_x = \begin{cases} max(p_i T_y) \; For \, multiple \, possibilities \\ min(p_i T_y) \;\; For \, asignle \, possible \, path \end{cases} \tag{6}$$

- **XOR pattern:** let P_x the probability value for the incoming paths. And E_x and T_x the energy and time **QoT** values. The computation is as follows:

$$Energy: \; max\left(p_i VTE_x \cdots p_i VE_y\right) \tag{7}$$

$$Time: \; T_x = max\left(p_i VT_x \cdots p_i VT_y\right) \tag{8}$$

- **OR pattern:**let P_x the probability value for the incoming paths. And E_x and T_x the energy and time **QoT** values. The computation is as follows:

$$Energy: \;\; max\left(p_i \sum i = 1^n VE_i\right) \tag{9}$$

$$Time: \; max\left(p_i \prod_{i=1}^{n} VT_i\right) \tag{10}$$

The configurable composition with uncertain **QoT** is realised through the aggregation of all the paths. As a result, the computation of the final composition value correspond to the aggregation of all the patterns formulas after customisation. Hence the aggregation value is:

$$P(composition) = \sum_{i=1}^{n} pi \, pattern_i \tag{11}$$

Where pattern $\in \{cOR, cXOR, cAND, cSequence\}$

5 Conclusion

In this paper, we present a systematic **QoT** uncertain configurable composition approach that is able to provide comprehensive **QoT** information for a Things even with the existence of complex composition structures such as cAND,cOR and XOR. Regarding the space limitation the experimentation results are absent which will be detailed in the future.

References

1. Amdouni, S., Barhamgi, M., Benslimane, D., Faiz, R.: Handling uncertainty in data services composition. In: 2014 IEEE International Conference on Services Computing, pp. 653–660. IEEE (2014)
2. Boulaares, S., Omri, A., Sassi, S., Benslimane, D.: A probabilistic approach: a model for the uncertain representation and navigation of uncertain web resources. In: 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 24–31. IEEE (2018)
3. Boulaares, S., Sassi, S., BenSlimane, D., Faiz, S.: A probabilistic approach: uncertain navigation of the uncertain web. In: Concurrency and Computation: Practice and Experience, p. e7194 (2022)
4. Boulaares, S., Sassi, S., Benslimane, D., Maamar, Z., Faiz, S.: Toward a configurable thing composition language for the siot. In: International Conference on Intelligent Systems Design and Applications, pp. 488–497. Springer (2022)
5. Brogi, A., Forti, S.: QoS-aware deployment of iot applications through the fog. IEEE Internet Things J. **4**(5), 1185–1192 (2017)
6. Cardoso, J., Sheth, A., Miller, J., Arnold, J., Kochut, K.: Quality of service for workflows and web service processes. J. Web Semant. **1**(3), 281–308 (2004)
7. Falas, Ł., Stelmach, P.: Web service composition with uncertain non-functional parameters. In: Doctoral Conference on Computing, Electrical and Industrial Systems, pp. 45–52. Springer (2013)
8. Gao, H., Huang, W., Duan, Y., Yang, X., Zou, Q.: Research on cost-driven services composition in an uncertain environment. J. Internet Technol. **20**(3), 755–769 (2019)
9. Jaeger, M.C., Rojec-Goldmann, G., Muhl, G.: QoS aggregation for web service composition using workflow patterns. In: Proceedings. Eighth IEEE International Enterprise Distributed Object Computing Conference, 2004. EDOC 2004, pp. 149–159. IEEE (2004)
10. Jian, X., Zhu, Q., Xia, Y.: An interval-based fuzzy ranking approach for QoS uncertainty-aware service composition. Optik **127**(4), 2102–2110 (2016)
11. Li, L., Jin, Z., Li, G., Zheng, L., Wei, Q.: Modeling and analyzing the reliability and cost of service composition in the iot: A probabilistic approach. In: 2012 IEEE 19th International Conference on Web Services, pp. 584–591. IEEE (2012)
12. Maamar, Z., Faci, N., Kajan, E., Asim, M., Qamar, A.: Owl-t for a semantic description of iot. In: European Conference on Advances in Databases and Information Systems, pp. 108–117. Springer (2020)
13. Ming, Z., Yan, M.: QoS-aware computational method for iot composite service. J. China Univ. Posts Telecommun. **20**, 35–39 (2013)

14. Niu, S., Zou, G., Gan, Y., Xiang, Y., Zhang, B.: Towards the optimality of QoS-aware web service composition with uncertainty. Int. J. Web Grid Serv. **15**(1), 1–28 (2019)
15. Qamar, A., Asim, M., Maamar, Z., Saeed, S., Baker, T.: A quality-of-things model for assessing the internet-of-things' nonfunctional properties. Trans. Emerg. Telecommun. Technol. e3668 (2019)
16. Rabah, B., Mounine, H.S., Ouassila, H.: QoS-aware iot services composition: a survey. In: Distributed Sensing and Intelligent Systems, pp. 477–488. Springer (2022)
17. Sangaiah, A.K., Bian, G.B., Bozorgi, S.M., Suraki, M.Y., Hosseinabadi, A.A.R., Shareh, M.B.: A novel quality-of-service-aware web services composition using biogeography-based optimization algorithm. Soft Comput. **24**(11), 8125–8137 (2020)
18. Tran, V.X., Tsuji, H.: Owl-t: A task ontology language for automatic service composition. In: IEEE International Conference on Web Services (ICWS 2007), pp. 1164–1167. IEEE (2007)
19. White, G., Palade, A., Clarke, S.: QoS prediction for reliable service composition in iot. In: International Conference on Service-Oriented Computing, pp. 149–160. Springer (2017)
20. Zhang, M.W., Zhang, B., Liu, Y., Na, J., Zhu, Z.L.: Web service composition based on QoS rules. J. Comput. Sci. Technol. **25**(6), 1143–1156 (2010)
21. Zheng, H., Yang, J., Zhao, W.: Probabilistic QoS aggregations for service composition. ACM Trans. Web (TWEB) **10**(2), 1–36 (2016)
22. Zheng, H., Yang, J., Zhao, W., Bouguettaya, A.: QoS analysis for web service compositions based on probabilistic QoS. In: International Conference on Service-Oriented Computing, pp. 47–61. Springer (2011)
23. Zheng, H., Zhao, W., Yang, J., Bouguettaya, A.: QoS analysis for web service compositions with complex structures. IEEE Trans. Serv. Comput. **6**(3), 373–386 (2012)



SR-Net: A Super-Resolution Image Based on DWT and DCNN

Nesrine Chaibi[1,2(✉)], Asma Eladel[2,3], and Mourad Zaied[1,2]

[1] National Engineering School of Gabes, Gabes, Tunisia
nesrine.chaibi@enig.rnu.tn, mourad.zaied@ieee.org
[2] Research Team in Intelligent Machines (RTIM), Gabes, Tunisia
asma.eladel@ieee.org
[3] Higher Institute of Computing and Multimedia of Gabes, Gabes, Tunisia

Abstract. Recently, a surge of several research interests in deep learning has been sparked for image super-resolution. Basically, a deep convolutional neural network is trained to identify the correlation between low and high-resolution image patches. In other side, profiting from the power of wavelet transform to extract and predict the "missing de-tails" of the low-resolution images, we propose a new deep learning strategy to predict missing details of wavelet sub-bands in order to generate the high-resolution image which we called a super-resolution image based on discrete wavelet transform and deep convolutional neural network (SR-DWT-DCNN). By training various images such as Set5, Set14 and Urban100 datasets, good results are obtained proving the effectiveness and efficiency of our proposed method. The reconstructed image achieves high resolution value in less run time than existing methods based on based on the evaluation with PSNR and SSIM metrics.

Keywords: Deep Convolutional Neural Network · Discrete Wavelet Transform · High-Resolution Image · Low-Resolution Image · Single Image Super-Resolution

1 Introduction

The field of super-resolution has seen an enormous growth in interest over the last years. High-resolution images are decisive and incisive in several applications including medical imaging [1], satellite and astronomical imaging [2], and remote sensing [3]. Unfortunately, many factors such as technology, cost, size, weight, and quality prevent the use of sensors with the desired resolution in image capture devices. This problem is very challenging and many researchers have addressed the subject of image super-resolution. The process of super-resolution (SR), which is defined as reconstructing a high-resolution (HR) image from a low-resolution (LR) image, can be divided into two categories depending on the number of low resolution images entered: single image super-resolution (SISR) and multi-image super-resolution (MISR) [4]. The first category is single image super-resolution (SISR) which took one low-resolution image to reconstruct a high quality image. The second category is multi-image super-resolution

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 291–301, 2023.
https://doi.org/10.1007/978-3-031-27409-1_26

(MISR) which generates a high-resolution image from multiple low-resolution images that are captured from the same scene. Recently, SISR has outperformed other competing methods and has had a lot of success thanks to its robust feature extraction and representation capabilities [5]. For instance, examples from historical data are frequently used to create dictionaries of LR and HR image patches. Each low-resolution (LR) patch is then transformed to the high-resolution (HR) domain using these dictionaries. In this paper, we address the problem of single image super-resolution, and we propose to incorporate the challenge of image super resolution in the wavelet domain. The Discrete Wavelet Transform (DWT) has many advantages proved by their capability to extract details, depict the contextual and textual information of an image at different levels and to represent and store multi-resolution images [6]. And the prediction of wavelet coefficients for super-resolution has been successfully used to multi-frame of super-resolution. Due to the strong capacity of deep learning (DL), the main contribution of this research is to propose a method based on deep learning algorithms combined with second generation wavelets for image super-resolution with the capability of simultaneous noise reduction; that we called a super-resolution image based on discrete wavelet transform and deep convolutional neural network (SR-DWT-DCNN).

The rest of this paper is organized as follows: Section 2 presents relevant background concepts of SISR and DWT. Section 3 discusses the related works in the literature. The proposed method for single image super-resolution is detailed in Sect. 4. The experimental results are provided in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Background

2.1 Single Image Super-Resolution

Single image super-resolution (SISR) is a challenging unstable problem because the specific low-resolution (LR) input can coincide to a crop of possible high-resolution (HR) images, and the high-resolution space that we aim to map the low-resolution input to is usually unwilling [7].



Fig. 1. Sketch of the global framework of SISR [8]

In the typical single image super-resolution framework, as shown in Fig. 1, the LR image y is described as follows [8]:

$$y = (x \otimes k) \downarrow s + n \tag{1}$$

where $(x \otimes k)$ represents the convolution of the fuzzy kernel k and the unknown HR image x, $\downarrow s$ represents the down sampling operator with scale factor s, and n represents the independent noise component.

2.2 Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) plays an important role in many applications, such as JPEG-2000 images compression standard, computer graphics, numerical analysis, radar target distinguishing and so forth. Nowadays, research on the DWT is attracting a big deal of attention. As a result, different architectures have been proposed to process the DWT. The DWT is a multi-resolution technique capable of assessing different frequencies by different resolutions. The wavelet representation of a discrete signal x with n samples can be calculated by convolving x with the low-pass and high-pass filters and down-sampling the resulting signal by two, so each frequency band comprises n/2 samples. This technique decomposes the original image into two sub-bands: lower and higher bands [9]. In order to form multi-levels decomposition, the process is applied recursively to the average sub-band and can be extended from one dimension (1d) to multiple dimensions (2d or 3d) depending on input signal dimensions.

3 State of the Art

In the literature, a variety of SISR methods have been proposed that mainly have two drawbacks: one is the uncertain definition of the mapping that we seek to expand between the LR space and the HR space, and the other is the inefficiency of generating a complex high dimensional mapping given huge raw data [10]. Currently, mainstream SISR algorithms are mainly classified into three categories: interpolation-based methods, reconstruction-based methods and learning-based methods. Interpolation-based SISR methods, such as Bicubic interpolation [11] and Lanczos resampling [12], are very speedy and straightforward but show a lack of precision when the factor of interpolation is greater than two. Reconstruction-based SR methods [13, 14] often adopt sophisticated prior knowledge to restrict the possible solution space with an advantage of generating flexible and sharp details.

Nevertheless, the performance of many reconstruction-based methods degrades rapidly when the scale factor tends to increase, and these methods are usually time-consuming. So, learning-based SISR methods, also known as example-based methods, are widely investigated [15–17] because of their fast computation and outstanding performance. These methods often use machine learning algorithms to evaluate statistical correlations between the LR and its corresponding HR counterpart based on large amounts of training data. Meanwhile, many studies combined the strengths of reconstruction-based methods with learning-based approaches to further minimize artifacts affected by

various training examples [18, 19]. However, their super-resolved results are typically unsatisfying with large magnification factors.

Very recently, DL-based SISR algorithms have demonstrated great superiority to reconstruction-based and other learning-based methods for a variety of problems [20–22]. Generally, the family of deep learning-based SR algorithms differs in the following key ways: various types of network architectures, various types of activation functions, various types of learning principles and strategies, etc. While recovering lost high-frequency information in the frequency domain appears to be simpler, it has been overlooked in DL-based SISR methods. The wavelet transform (WT) is frequently used in signal processing due to its ability to extract features and perform multi-resolution analysis [23, 24]. Furthermore, the WT can depict the contextual and textual information of an image at several levels and has been shown to be an efficient and very intuitive technique for defining and maintaining multi-resolution images [25]. Consequently, many studies have been conducted on WT applications in the resolution field, such as a 2-D oriented WT method to compress remote sensing images [6, 26–28], an image classification method based on a combination of the WT and the neural network [27]. In [6], the discrete wavelet transform was combined with DCNN to predict the missing detail of approximation sub-band. Wen et al. [26] depicted a three-step super-resolution method for remote sensing images via the WT combined with the recursive Res-Net (WTCRR). Li et al. [28] reconstructed the infrared image sequences in the wavelet domain and obtained a significant increase of the spatial resolution. To the best of our knowledge, little research has concentrated on integrating the WT into DCNNs, which is expected to improve reconstruction accuracy further due to their respective merits.

4 Proposed Method

In this paper, to tackle super-resolution task, we propose a new approach of deep learning for single image super-resolution algorithms. Mainly, we focus on the combination of two areas: DCNN and DWT in the domain of super-resolution. In this section, we will explain and depict the overall architecture of the proposed method SR-DWT-DCNN (see Fig. 2).



Fig. 2. The architecture of the proposed approach "SR-DWT-DCNN"

The input of our network is a high-resolution image I (size: m*n) on which we apply YCbCr. YCbCr is a color space family that is utilized as part of the color image pipeline in video and digital photography systems. The luma component is represented by Y, whereas the blue-difference and red-difference chroma components are represented by

Cb and Cr, respectively. Then, we apply our method of super-resolution which is based on discrete wavelet transform and deep convolutional neural networks on each image Iy, Icb and Icr separately. As results, three reconstructed images SRy, SRcb and SRcr are generated and combined to generate the high-resolution image IHR (see Fig. 3). The main goal of our method is to minimize the noise and maximize the quality of the extracted features. To demonstrate our method in this paper, we will apply our method only in Iy image and we will give more details about our network architecture (see Fig. 3).

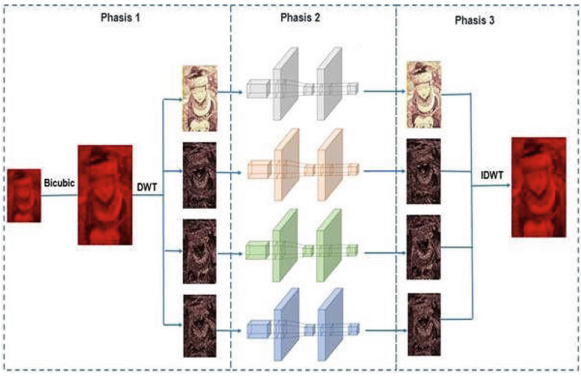


Fig. 3. The SR based DWT and DCNN network, which consists of three phases: the decomposition of *ILy*, the prediction of features from the four sub-bands, and the reconstruction of *SRy*.

As mention above, the input of our network is a high-resolution image Iy on which we will apply two transformations which are the down-sampling and the up-sampling in order to get a low-resolution image ILy. Using discrete wavelet transform, ILy was divided into four sub-bands LL, LH, HL and HH. Then, each sub-band was fed in his corresponding model which based on deep convolutional neural networks (DCNN). Finally, the inverse discrete transform was applied on the four generated sub-bands LL′, LH′, HL′ and HH′ to reconstruct the output SRy of our network. In the next sub-sections, we will detail the three phases.

Phasis 1: DWT for Sub-bands Extraction

Since 1980, the wavelet analysis was introduced and many wavelets have appeared such as Haar, Symlet, Coiets, Daubechies and so on. Recent researches proved the giant role of wavelets to solve the problem of super-resolution [6, 24, 25]. Therefore, profiting from its power of extracting effective high-level abstractions that bridge the LR and HR space, we applied the discrete wavelet transform to divide the input image into four sub-bands. In the first phasis, the Iy has been downed-in and zoomed-in using bicubic interpolation method with a scale value equal to S to achieve the low-resolution image ILy. Then, we used the discrete wavelet transform especially DB2 wavelet because it is more efficient for noise decreasing since the relevant features are those that persist across scales than other methods [25, 26]. After applying the DB2 wavelet transform, the

ILy has decomposed into LL, LH, HL and HH using single-level 2-D discrete wavelet transform (2d DWT). The three sub-bands LH, HL and HH contain edge information in different directions about the original image, which are used to improve our goal in the next step. A flowchart of 2d DWT using DB2 wavelet is represented in Fig. 4.

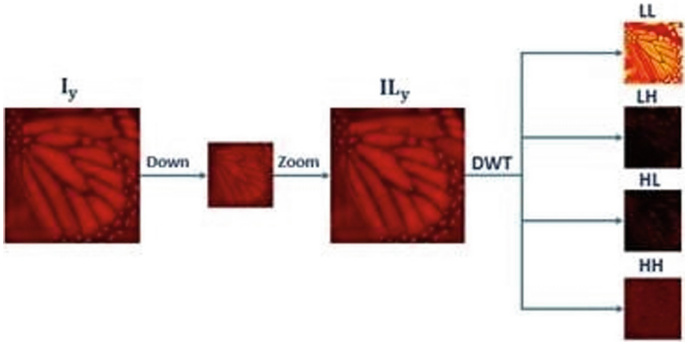


Fig. 4. A flowchart of 2d DWT in "Butterfly" image from Set5

Phasis 2: Enhance Resolution Using DCNN

The second phasis includes four deep convolutional network models. A DCNN for each sub-band. Thus, the first sub-band LL was fed to the DCNN trained on the approximation wavelet sub-band, the second sub-band LH was fed to the DCNN trained on the horizontal wavelet sub-band, the third sub-band HL was fed to the DCNN trained on the vertical wavelet sub-band and the last sub-band HH was fed to the DCNN trained on the diagonal wavelet sub-band. Each DCNN composed of three convolutional layers networks with $f1 = 9$, $f2 = 5$, $f3 = 5$, $n1 = 64$ and $n2 = 32$ trained on the ImageNet with up-scaling factor 2|3|4 (see Fig. 5).

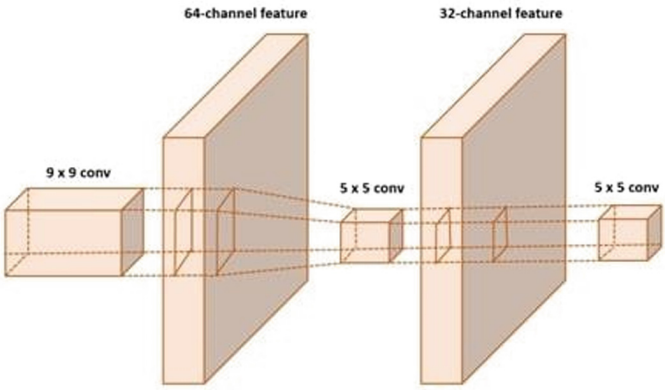


Fig. 5. DCNN network architecture

Feature extraction tries to capture the content of images. The first convolutional layer of each model (9*9 conv) extracts a set of feature maps. Thus, nonlinearly, $*$ these features are transformed to high-resolution patch representations. In this first operation, we convolve the image by a set of filters (n1 = 64), each of which is a basis. The output is composed of n1 feature maps whose each element is associated with a filter. After that, we map each of these n1-dimensional vectors into an n2-dimensional one. This $*$is equivalent to applying n2 filters (n2 = 32) which have a trivial spatial support 5*5. For the reconstruction, we use the output n2-dimensional vectors which are conceptually a representation of a high-resolution patch. The last layer aggregates the above high-resolution patch wise representations to generate the final high-resolution image. Our contribution in this phasis is that we proposed to implement four DCNN models; each one has as input one wavelet sub-band. The main goal of the first model is to predict the missing information for approximation wavelet sub-band while the three networks are implementing to predict the missing information for horizontal, vertical and diagonal wavelet sub-bands. The four models demand few training time and didn't increase the complexity of our method. As result of this step, four new wavelets sub-bands are generated LL′, LH′, HL′ and HH′ to reconstruct the high-resolution image.

Phasis 3: HR-Image Reconstruction

In this phasis, the 2nd inverse discrete wavelet transform (2d IDWT) can trace back the 2d DWT procedure by inverting the steps in Fig. 6. This allows the prediction and combination of wavelet coefficients to generate super-results. Consequently, the reconstructed high-resolution image S Ry was obtained via the inverse discrete wavelet transform (2d IDWT) of the new four wavelets sub-bands LL′, LH′, HL′ and HH′. Finally, we combined in RGB the three reconstructed images S Ry, S Rcb and S Rcr respectivly of Iy, Icb and Icr images to generate the high-resolution image IHR and we compare the reconstructed high-resolution image IHR and I using PSNR and SSIM metrics. Figure 8 shows the process of this phasis based in the inverse discrete wavelet transform IDWT.

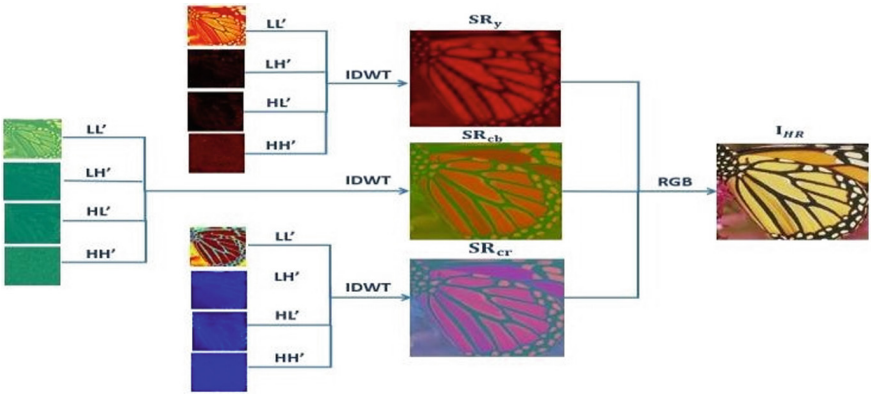


Fig. 6. A flowchart of IDWT in "Butterfly" image from Set5 and *IHR* reconstruction

5 Experimental Results

The proposed method's performance is evaluated in this section. To begin, we present the dataset that was deployed for the training and testing phases. The metrics used to evaluate the various methods were then provided. Finally, we compared our method to other super-resolution approaches. The 91 images from Yang et al. [16] are extensively used in the learning-based SR approach during the training stage. However, numerous studies demonstrate that the 91 photos are insufficient to push the network to its optimal performance for the super-resolution task. Set5 [29], Set14 [30] and Urban100 [31] datasets are employed in the testing stage. Huang et al. recently published a set of urban photos that is very interesting as it contains many challenging images that have been discarded by previous approaches.

In order to evaluate our approach, we used PSNR and SSIM [32] indices. These indices are widely used to evaluate super-resolution methods, because of their high correlation with the human perceptual scores [33]. We compare our SR-DWT-DCNN method with the state-of-the-art SR methods trained on different datasets, namely the deep convolutional neural network based on discrete wavelet transform for image super-resolution method (DCNNDWT) [6], SRCNN [20], and Bicubic Interpolation [11] which is used as the baseline. The quantitative results of PSNR and SSIM are shown in Table 1.

Table 1. The average results of PSNR (DB) and SSIM on the SET5 dataset

Eval	Scale	Bicubic	SRCNN	DCNNDWT	(Our)
PSNR	2	33.66	36.33	36.52	36.51
	3	30.39	32.75	33.43	**33.69**
	4	28.42	30.49	31.67	**31.98**
SSIM	2	0.9299	0.9542	0.972	**0.985**
	3	0.8682	0.9090	0.929	**0.946**
	4	0.8104	0.8628	0.884	**0.921**

As shown in Table 1, we observe that the Bicubic method get even lower scores than the SRCNN and DCNNDWT methods on PSNR and SSIM metrics. In the proposed method, we used the three details extracted from DWT which positively affect the obtained results by achieving the highest scores in most evaluation matrices in all experiments. When the up-scaling factor greater than 2, the average gains on PSNR and SSIM achieved by our SR-DWT-DCNN method are 0.98 and 0.154 dB. The average results are higher than the other approach on the three datasets. Also, the average gains on SSIM metric by our proposed method achieved the highest value. Also, comparing the SRCNN method with our method, we can observe clearly that the performance of SRCNN is far from converging. Moreover, our reached results can be improved by increasing the scale; and this due to the refinement of the extracted image details. However, obtained results by the other methods are decreased when reaching scale equal to 3

or 4. Furthermore, regardless of PSNR and SSIM metrics, SR-DWT-DCNN achieves the best performance and speed among all methods and specifically when the scaling factors superiors than 2. With moderate training, SR-DWT-DCNN outperforms existing state-of-the-art methods. Note that the running time of all algorithms using the same machine. Figures 7 and 8 show some reconstructed images on the 'Set5' dataset with an up-scaling factor respectively 3 and 4 using Bicubic, SRCNN, DCNNDWT and SR-DWT-DCNN methods.



Fig. 7. "Woman" image from Set5 with up-scaling 3.



Fig. 8. "Head" image from Set5 with up-scaling 4

6 Conclusion

In this paper, we presented a new method for Super Image Reconstruction based on DCNN and discrete wavelet transform. The main contributions of this paper is implementing four DCNN models with four inputs generated from discrete wavelet transform in order to predict the missing details. By this way, we guarantee the quality of the reconstructed image and the speed of running time. As a result, the effectiveness has improved. As future work, the proposed approach can be applied to solve the problem of Multi Image Super-Resolution and other low-level vision problems such as image denoising. Moreover, the effects of different wavelet basis can be examined in future works for super-resolution task.

References

1. Luján-García, J.E., et al.: A transfer learning method for pneumonia classification and visualization. Appl. Sci. **10**(8), 2908 (2020)
2. Puschmann, K.G., Kneer, F.: On super-resolution in astronomical imaging. Astron. Astrophys. **436**(1), 373–378 (2005)
3. Sabins, F.F.: Remote sensing for mineral exploration. Ore Geol. Rev. **14**(3–4), 157–183 (1999)
4. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Process. Mag. **20**(3), 21–36 (2003)
5. Mikaeli, E., Aghagolzadeh, A., Azghani, M.: Single-image super-resolution via patch-based and group-based local smoothness modeling. Vis. Comput. **36**(8), 1573–1589 (2019). https://doi.org/10.1007/s00371-019-01756-w
6. Chaibi, N., Eladel, A., Zaied, M.: Deep convolutional neural network based on wavelet transform for super image resolution. In: HIS Conference 2020, vol. 1375, pp. 114–123 (2020)
7. Yang, C.-Y., Ma, C., Yang, M.-H.: Single-image super-resolution: a benchmark. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, Proceedings, Part IV 13. Springer International Publishing, p. 386 (2014)
8. Yang, W., et al.: Deep learning for single image super-resolution: a brief review. IEEE Trans. Multim. **21**(12), 3106–3121 (2019)
9. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press (2008)
10. Xiong, Z., et al.: Single image super-resolution via image quality assessment-guided deep learning network. PloS one **15**(10), e0241313 (2020)
11. Keys, R.: Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. **29**(6), 1153–1160 (1981)
12. Duchon, C.E.: Lanczos filtering in one and two dimensions. J. Appl. Meteorol. Climatol. **18**(8), 1016–1022 (1979)
13. Dai, S., et al.: Softcuts: a soft ede smoothness prior for color image super-resolution. IEEE Trans. Image Process. **18**(5), 969–981 (2009)
14. Marquina, A., Osher, S.J.: Image super-resolution by tv regularization and bregman iteration. J. Sci. Comput. **37**, 367–382 (2008)
15. Cruz, C., et al.: Single image super-resolution based on Wiener filter in similarity domain. IEEE Trans. Image Process. **27**(3), 1376–1389 (2017)
16. Yang, J., et al.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
17. Luo, X., Yong, X., Yang, J.: Multi-resolution dictionary learning for face recognition. Pattern Recogn. **93**, 283–292 (2019)
18. Zhang, X.G.X.L.K., Tao, D., Li, J.: Coarse-to-fine learning for single-image super-resolution. IEEE Trans. Neural Netw. Learn. Syst. **28**, 1109–1122 (2017)
19. Yang, W., et al.: Consistent coding scheme for single-image super-resolution via independent dictionaries. IEEE Trans. Multim. **18**(3), 313–325 (2016)
20. Dong, C., et al.: "Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
21. Nguyen, K., et al. Super-resolution for biometrics: a comprehensive survey. Pattern Recogn. **78,** 23–42 (2018)
22. He, X., et al.: Ode-inspired network design for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
23. Aballe, A., et al.: Using wavelets transform in the analysis of electrochemical noise data. Electrochim. Acta **44**(26), 4805–4816 (1999)

24. Abbate, A., Frankel, J., Das, P.: Wavelet transform signal processing for dispersion analysis of ultrasonic signals. In: 1995 IEEE Ultrasonics Symposium. Proceedings. An International Symposium. Vol. 1. IEEE (1995)
25. Mallat, S.: Wavelets for a vision. Proc. IEEE **84**, 604–614 (1996)
26. Ma, W., et al.: Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net. IEEE Trans. Geosci. Remote Sens. **57**(6), 3512–3527 (2019)
27. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
28. Li, J., et al.: Wavelet domain superresolution reconstruction of infrared image sequences. In: Sensor Fusion: Architectures, Algorithms, and Applications V. Vol. 4385. SPIE (2001)
29. Bevilacqua, M., et al.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding, 135–1 (2012)
30. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24–30, 2010, Revised Selected Papers 7. Springer Berlin Heidelberg (2012)
31. Huang, J.-B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
32. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process **13**(4), 600–612 (2004)
33. Yang, C.-Y., Ma, C., Yang, M.-H.: Single-image super-resolution: a benchmark. In: European Conference on Computer Vision. Springer (2014)



Performance of Sine Cosine Algorithm for ANN Tuning and Training for IoT Security

Nebojsa Bacanin[1(✉)], Miodrag Zivkovic[1], Zlatko Hajdarevic[1], Stefana Janicijevic[1], Anni Dasho[2], Marina Marjanovic[1], and Luka Jovanovic[1]

[1] Singidunum University, Danijelova 32, 11000 Belgrade, Serbia
{nbacanin,mzivkovic,sjanicijevic,mmarjanovic}@singidunum.ac.rs,
{zlatko.hajdarevic.16,luka.jovanovic.191}@singimail.rs
[2] Luarasi University, Rruga e Elbasanit 59, Tirana 1000, Albania
anni.dasho@luarasi-univ.edu.al

Abstract. Recent advances in Internet technology ensured that the World Wide Web is now essential for millions of users, offering them a variety of services. As the number of online transactions grows, the number of hostile users who are trying to manipulate with sensitive data and steal user's private details, credit card data and money is also rising fast. To fight this threat, security companies developed a variety of security measures, aiming to protect both end user and business offering online services. Nowadays, machine learning methods are common part of the most of the contemporary security solutions. The research goal of this paper is proposal of the hybrid technique that uses multi-layer perceptron tuned by the well-known sine cosine algorithm. Sine cosine metaheuristics is utilized to determine the neural cell count within the hidden layer, and to obtain the weights and biases. The capabilities of the observed method were validated on a public web security benchmark dataset, and compared to the results obtained by other elite metaheuristics that have been tested under the same conditions. The simulation findings indicate that the introduced model surpassed other observed techniques, showing great deal of potential for practical use in this domain.

Keywords: ANN training · Sine cosine algorithm · IoT security · Industry 4.0

1 Introduction

The Industrial Revolution 4.0 has been driven by the recent significant development of the Internet of Things (IoT). The main goal of Industrial Revolution 4.0 supported is a transfer from traditional factories to smart factories. The IoT devices are now being installed and connected to equipment within the factory's

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 302–312, 2023.
https://doi.org/10.1007/978-3-031-27409-1_27

production chain and at clients' machines giving factory data that can improve quality and client satisfaction. One of the biggest problems in Industry 4.0 is failure detection and security, due that Industry 4.0 is in high dependence on IoT devices, and their secure and uninterrupted communication. The communication among these devices can be intercepted and overfloaded, and IoT devices can fail to provide service. To resolve these types of problems solutions for the real-time detection of the device's failures and attacks are in big demand and it has become the most important thing to consider when resolving IoT security. Some solutions for this type of problem can be provided by artificial intelligence (AI) and machine learning (ML).

AI constitutes a solid solution for problems that can arise in the domain of the network security, as machine learning models are capable of learning and adapting to the frequent changes in the environment. Although traditional security measures such as firewall and blacklists are still in use, they are not effective enough as they must be monitored and maintained all the time. Numerous scientist have recently investigated the possibility of improvement of the current approaches, and tried to leverage the network security through application of the AI methods. The most notable applications address intrusion detection, phishing attacks, IoT botnets discovery, and spam detection [3,12].

Multi-layer perceptron (MLP) is a most common AI model used today. It can achieve admirable level of accuracy on a variety of practical problems, however, it must be tuned for each individual problem, as the general solution that will attain the best performance in every domain isn't existing (no free lunch theorem). The MLP tuning task comprises of determining the count of units within the hidden layer, and input weight and bias merits, that is NP-hard tuning challenge by nature. Metaheuristics algorithms are considered as extremely effective in solving optimization problems in different domains, including NP-hard tasks that cannot be solved by applying conventional deterministic algorithms.

The main goal of this manuscript is to utilize a well-known sine cosine algorithm (SCA) [13], that is inspirited by the mathematical properties of sine and cosine functions, and apply it for tuning the number of hidden neurons, input weight and bias values of the MLP. The suggested approach has been validated on a well-known Web and network security dataset. To summarize, the most significant contributions of the proposed work are:

1. SCA metaheuristics is proposed for tuning hidden MLP hyperparameters.
2. The proposed model is adapted to tackle important challenge of Web and network security issues detection.
3. The proposed model has been validated on publicly obtainable network security benchmark dataset.
4. The findings of the introduced model have been evaluated and put into comparison with several other cutting-edge metaheuristics utilized to solve this particular problem on the same dataset.

The rest of this essay is organized as follows. Section 2 provides preliminaries on the neural networks and metaheuristics optimization. Section 3 shows the

utilized SCA approach. Section 4 brings the description of the simulation setup and displays the experimental findings. Finally, Sect. 5 brings conclusions and wraps up the paper.

2 Preliminaries and Related Works

2.1 Tuning the Parameters of an Artificial Neural Networks

Neural network (NN) training is an important task, with the main purpose to build a model with better capabilities. The function loss needs to be optimized during the learning process. One problem with the NN training process is overfitting. This problem occurs when there is a significant deviation in test and training accuracy, it indicates that the NN has been over-trained for specific data (training data), and it is not able to provide a good result when entering new data (test data). To solve this problem various approaches can be performed: dropout, drop connect, L1 and L2 regularization, early stopping, and so on [15].

The MLP training can be utilized with stochastic optimizers, that can break out from the local optima. If the goal is to tune both the weights and network architecture, MLP training becomes an extremely hard challenge. The MLP networks can be defined as a sort of feedforward neural networks (FFNN). The FFNNs consist of a set of neural cells. These neurons can be described as a sequences of completely interconnected layers. MLP contains three types of layers in this order: input, hidden and output. Neurons in MLP are one-directional and layers are bound with weights. Neurons are executing two operations: summation and activation. The summation operation is given by Eq. (1):

$$S_j = \sum_{i=1}^{n} \omega_{ij} I_i + \beta_j \tag{1}$$

where n represents the count of input values, I_i stands for the input value i, ω_{ij} stands for the connection weight, β_j denotes the bias term. The output of Eq.(1) executes the activation function. The best way to see the capabilities of any network is by measuring the loss function.

2.2 Swarm Intelligence

Swarm intelligence studies both natural and artificial systems where a big number of individuals have decentralized control and are self-organization to benefit their entire population. The inspiration for swarm intelligence came when observing the behavior of bird flocks when they are seeking food. Today we have many swarm intelligence approaches.

Survey of recent research indicates very successful combinations of a variety of neural network models with metaheuristic algorithms, together with a wide spectrum of other applications. Some cutting-edge applications of swarm intelligence optimization include predicting the nunmber of confirmed COVID-19 cases [18], COVID-19 MRI classifying task and sickness severity estimation

[7,8], computer-guided tumor MRI classifying process [5], feature selection challenge [11,20], cryptocurrencies fluctuations forecasting [14], network security and intrusion detection [2,19], cloud-edge computing task assignment [6], sensor networks optimization [4] and numerous other successful applications.

3 Proposed Method

3.1 Basic SCA

The inspiration for the sine cosine algorithm (SCA) was found in trigonometric functions from which the mathematical model is based on [13]. The position is updated by mathematical functions - trigonometric functions and because of that algorithm oscillation in the space of the optimum solution. The values that are returned in the ranges of $[-1, 1]$. At the initialization phase, it generates multiple solutions and every one of these solutions can be a candidate for the best solution in the area of search. Exploration and exploitation are controlled by randomized adaptive parameters. The position update is performed by two main equations [13]:

$$X_i^{t+1} = X_i^t + r_1 \cdot sin(r_2) \cdot |r_3 \cdot P_i^{*t} - X_i^t| \tag{2}$$

$$X_i^{t+1} = X_i^t + r_1 \cdot cos(r_2) \cdot |r_3 \cdot P_i^{*t} - X_i^t| \tag{3}$$

X_i^t and X_i^{t+1} represents the positioning of solution in dimension i-th and t-th and $i+1$-th round, in this order, created random pseudo numbers shown as r_{1-3}, P_i^* for the i-th dimension represents the location of the target, $||$ represents the absolute value. r_4 represents the control parameter and for this parameter, two equations are used:

$$X_i^{t+1} = \begin{cases} X_i^{t+1} = X_i^t + r_1 \cdot sin(r_2) \cdot |r_3 \cdot P_i^{*t} - X_i^t|, & r_4 < 0.5 \\ X_i^{t+1} = X_i^t + r_1 \cdot cos(r_2) \cdot |r_3 \cdot P_i^{*t} - X_i^t|, & r_4 \geq 0.5, \end{cases} \tag{4}$$

The search is controlled by four parameters, everyone is different and they are randomly generated. The main functions range of search is modified dynamically and this behavior balance to the global best solution. For repositioning near the solution, sine and cosine functions use cyclic sequences. This behavior guarantees exploitation. To increase randomness and its quality, the parameter values r_2 is changed to $[0, 2\Pi]$. The following equation is controlling diversification and exploitation balance:

$$r_1 = a - t\frac{a}{T}, \tag{5}$$

in which t represents the ongoing count of repetitions, T represents the maximum number of repetitions for every run, and a is a constant value. a constant is a hard-coded value, it can not be adjustable. Value for this parameter has been determined by previous experience, and because dropout regularization it is set to 2.0, this value is suggested in [13]. This type of dropout regularization also falls into NP-hard problems. Pseudo-code for SCA algorithm is next:

Algorithm 1 The SCA pseudocode

Generate collection of solutions (X)
while ($t < T$) **do**
  Evaluate agents using objective function
  Update the best agent(solution) determined until now ($P = X^*$)
  Update values $r1$, $r2$, $r3$, and $r4$
  Update the positions of individuals by applying Eq. (4)
end while
Return best obtained agent(solution)

3.2 Solution Encoding

All metaheuristics algorithms included in this research were used to first optimize the count of cells within the hidden layer, and then to tune the weight and bias merits. Lower bound for the number of neurons was set to $lb_nn = nf$, where nf denotes number of features, while the upper bound was set to $ub_nn = nf * 3$. Weight and bias values are set in range $[-1, 1]$. Each individual solution's vector length is given by $D = 1 + nf * ub_nn + ub_nn + ub_nn * no_classes + no_classes$.

As it can be seen, this problem is a mixed NP-hard challenge with both integer and real variables, where nn is integer, and weights and bias values are real. It makes this task very complex, as each individual in the population performs both optimization of the nn and network training, with significantly less training iteration than classic SGD method. However, since it is a large scale problem with a substantial amount of variables, it is very suitable to test the performance of the metaheuristics.

4 Experimental Findings and Discussion

4.1 Datasets

The dataset that we are using in this paper is generated by virtual machines with Windows 10 OS. The Windows 10 dataset has 125 features and attribute that represent attack type. There are seven types of attacks: DDoS, Injection, XSS, Password, Scanning, DoS, and MITM. Normal traffic has 4871 records in Windows 10 dataset, while DDoS has 4608, Injection has 612 records, XSS has 1268 records, Password has 3628 records, Scanning has 447 records, DoS has 525 records, MITM has 15 records. The dataset consists of 10,000 regular entries and 11,104 entries labeled as dangerous. This dataset could be utilized for both binary and multi-class classifying process, and the class distribution is shown in Fig. 1. In this paper, binary classification is utilized. Figure 2 shows the features heatmap.

4.2 Experimental Setup

The capabilities of the MLP optimized by the SCA method with respect to the convergence speed and general capabilities has been evaluated on the dataset

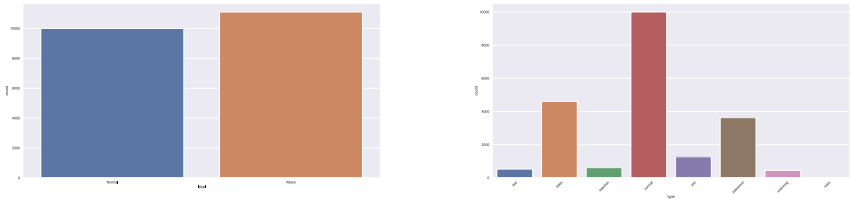


Fig. 1. Windows 10 dataset class distribution for binary and multi-class classification

given in the previous section. The experimental outcomes have been put into the comparison with the results attained by five other superior algorithms, employed in the same way, and used as a reference. The reference metaheuristics algorithms included AOA [1], ABC [10], FA [16], BA [17], and HHO [9]. Mentioned reference methods have been implemented independently for the sake of this manuscript, with the control parameters' setup as proposed in their respective publications. The experiments were executed as follows. The dataset was divided into train (80%) and test (20%) portions. All metaheuristics algorithms were used with 12 individuals in population ($N = 12$) and 10 independent run, with maximum of twelve iterations in a single run ($maxIter = 12$).

4.3 Experimental Results

Table 1 summarizes the overall metrics obtained by all algorithms on Win 10 dataset, for the objective function that is being minimized (error rate), and the best result in every category is bolded. It is possible to note that the MLP-SCA approach achieved superior level of performance for all observed metrics (best, worst, mean, median, standard deviation), and determined the network structure with 15 nodes in the hidden layer. Second-best value was obtained by MLP-HHO, while MLP-ABC finished at third place.

Table 2 brings forward the detailed metrics for the best solution for each observed algorithm. The best obtained accuracy on Win 10 dataset was again achieved by the MLP-SCA method, reaching the level 83.04%, and finishing infront of MLP-HHO that was behind by around 0.5%, with the accuracy of 82.54%. Other observed methods were left far behind, as the MLP-FA approach on the third position fell behind the observed method by almost 5%, with the highest accuracy of 78.8%. The suggested MLP-SCA method was superior in almost all other indicators as well, finishing in first place for eight out of ten indicators used.

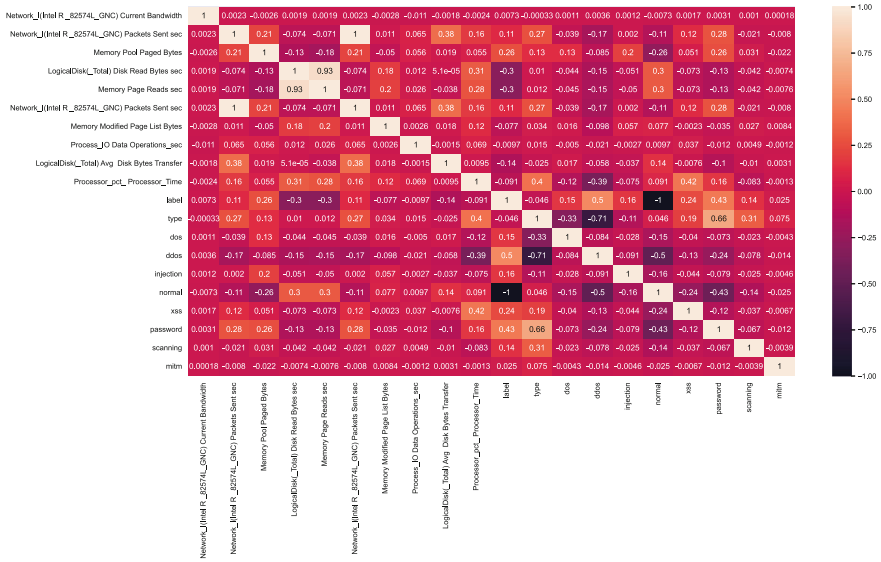


Fig. 2. Windows 10 dataset features' heatmap

Table 1. Overall metrics for all observed methods on Win 10 dataset

Method	MLP-SCA	MLP-AOA	MLP-ABC	MLP-FA	MLP-BA	MLP-HHO
Best	**0.169628**	0.226724	0.304430	0.212035	0.320540	0.174603
Worst	**0.225539**	0.389244	0.362000	0.453684	0.399905	0.327884
Mean	**0.188462**	0.306148	0.333333	0.358209	0.365612	0.249585
Median	**0.179341**	0.304312	0.333452	0.383558	0.371002	0.247927
Std	**0.021812**	0.071773	0.026308	0.089592	0.028590	0.070340
Var	**0.000476**	0.005151	0.000692	0.008027	0.000817	0.004948
Nn	15	23	10	30	28	10

Table 2. Detailed metrics for all observed methods on Win 10 dataset

	MLP-SCA	MLP-AOA	MLP-ABC	MLP-FA	MLP-BA	MLP-HHO
Accuracy (%)	**83.0372**	77.3276	69.557	78.7965	67.946	82.5397
Precision 0	0.911012	0.859407	0.805295	**0.921434**	0.666495	0.910331
Precision 1	**0.783001**	0.728159	0.653443	0.727835	0.690518	0.776659
M.Avg Precision	**0.843655**	0.790347	0.725393	0.819566	0.679135	0.839996
Recall 0	**0.711500**	0.623500	0.471500	0.604000	0.647500	0.700500
Recall 1	0.937416	0.908149	0.897344	**0.953624**	0.708240	0.937866
M.Avg. Recall	**0.830372**	0.773276	0.695570	0.787965	0.679460	0.825397
F1 score 0	**0.798989**	0.722689	0.594765	0.729689	0.656860	0.791749
F1 score 1	**0.853279**	0.808255	0.756213	0.825570	0.699267	0.849684
M.Avg. F1 score	**0.827555**	0.767712	0.679716	0.780140	0.679174	0.822233

In order to allow better visualization of the capabilities of the given model, the convergence graph of the objective function (error rate) and box plot diagrams for all observed algorithms are given in Fig. 3.

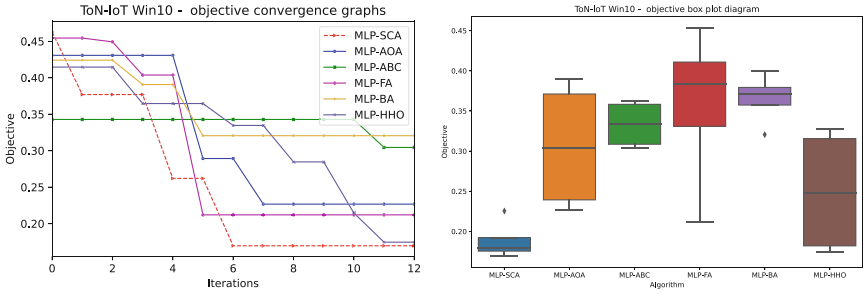


Fig. 3. Objective convergence and boxplot diagrams for all observed methods on Windows 10 dataset

The confusion matrices for all observed algorithms are shown in Fig. 4. It can be noted from the experimental outcomes that the proposed MLP-SCA is very well suited for tackling this problem, and it can be considered for practical implementation.

5 Conclusion

This manuscript proposed a hybrid ML-swarm intelligence approach to tackle the problem of web security. The well-known SCA metahueristics algorithm was used to establish the count of hidden neurons and weight and bias values for the MLP model. The proposed hybrid model was evaluated on a known benchmark Win 10 dataset, and the obtained results were collated to the outcomes achieved by five contending exceptional metaheuristics algorithms. The overall experimental outcomes clearly suggest that the proposed MLP-SCA method achieved superior level of performance, and has shown great deal of perspective to be practically implemented and used as the part of web security frameworks. The future examination in this domain should encompass additional verification of the suggested model, by utilizing additional real-world datasets, aiming to establish the confidence in the performance even further.

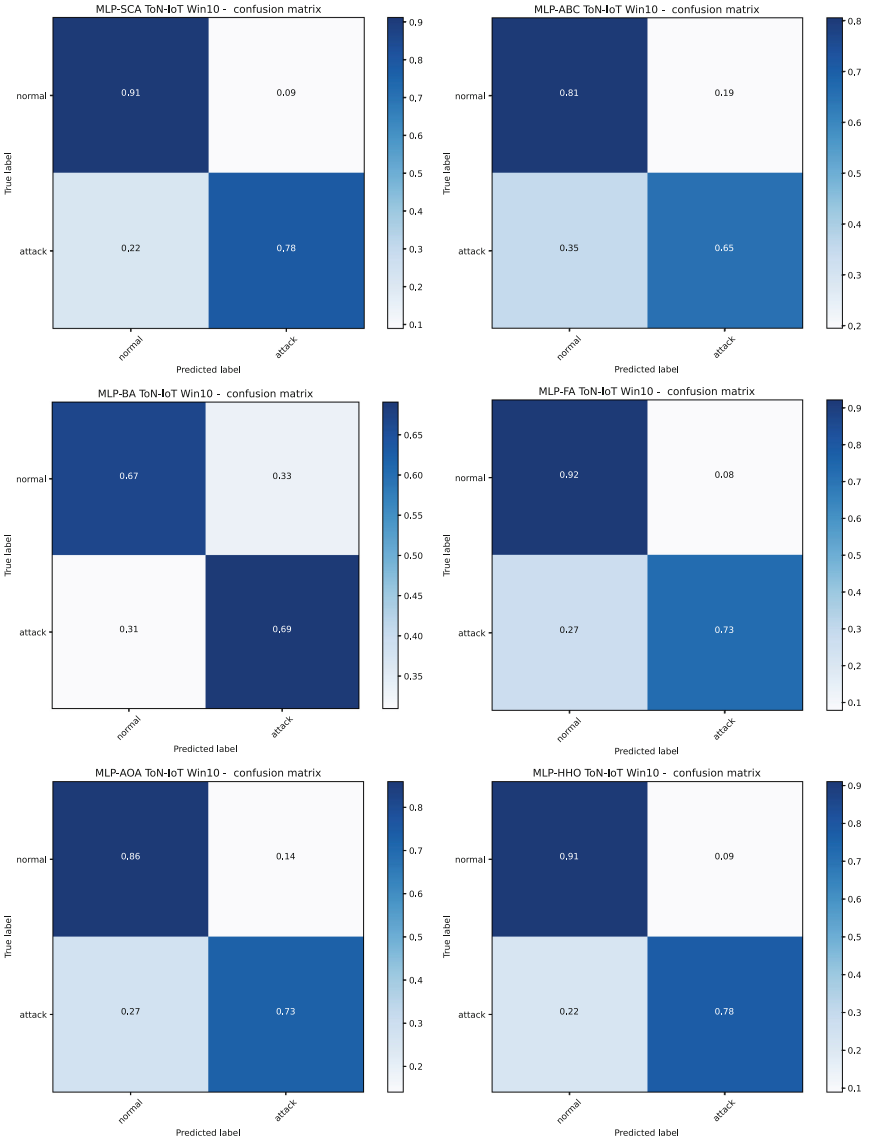


Fig. 4. Confusion matrices for all observed methods on Windows 10 dataset

References

1. Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., Gandomi, A.H.: The arithmetic optimization algorithm. Comput. Methods Appl. Mech. Eng. **376**, 113609 (2021)
2. AlHosni, N., Jovanovic, L., Antonijevic, M., Bukumira, M., Zivkovic, M., Strumberger, I., Mani, J.P., Bacanin, N.: The XgBoost model for network intrusion detection boosted by enhanced sine cosine algorithm. In: International Conference on Image Processing and Capsule Networks, pp. 213–228. Springer (2022)
3. Alqahtani, H., Sarker, I.H., Kalim, A., Hossain, M., Md, S., Ikhlaq, S., Hossain, S.: Cyber intrusion detection using machine learning classification techniques. In: International Conference on Computing Science, Communication and Security, pp. 121–131. Springer (2020)
4. Bacanin, N., Sarac, M., Budimirovic, N., Zivkovic, M., AlZubi, A.A., Bashir, A.K.: Smart wireless health care system using graph LSTM pollution prediction and dragonfly node localization. Sustain. Comput. Inf. Syst. **35**, 100711 (2022)
5. Bacanin, N., Zivkovic, M., Al-Turjman, F., Venkatachalam, K., Trojovskỳ, P., Strumberger, I., Bezdan, T.: Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. Sci. Rep. **12**(1), 1–20 (2022)
6. Bacanin, N., Zivkovic, M., Bezdan, T., Venkatachalam, K., Abouhawwash, M.: Modified firefly algorithm for workflow scheduling in cloud-edge environment. Neural Comput. Appl. **34**(11), 9043–9068 (2022)
7. Bezdan, T., Zivkovic, M., Bacanin, N., Chhabra, A., Suresh, M.: Feature selection by hybrid brain storm optimization algorithm for covid-19 classification. J. Comput. Biol. (2022)
8. Budimirovic, N., Prabhu, E., Antonijevic, M., Zivkovic, M., Bacanin, N., Strumberger, I., Venkatachalam, K.: Covid-19 severity prediction using enhanced whale with salp swarm feature classification. Comput. Mater. Contin., 1685–1698 (2022)
9. Heidari, A.A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., Chen, H.: Harris hawks optimization: algorithm and applications. Future Gener. Comput. Syst. **97**, 849–872 (2019)
10. Karaboga, D.: Artificial bee colony algorithm. Scholarpedia **5**(3), 6915 (2010)
11. Latha, R., Saravana Balaji, B., Bacanin, N., Strumberger, I., Zivkovic, M., Kabiljo, M.: Feature selection using grey wolf optimization with random differential grouping. Comput. Syst. Sci. Eng. **43**(1), 317–332 (2022)
12. Makkar, A., Garg, S., Kumar, N., Hossain, M.S., Ghoneim, A., Alrashoud, M.: An efficient spam detection technique for IoT devices using machine learning. IEEE Trans. Ind. Inf. **17**(2), 903–912 (2020)
13. Mirjalili, S.: SCA: a sine cosine algorithm for solving optimization problems. Knowl.-Based Syst. **96**, 120–133 (2016)
14. Salb, M., Zivkovic, M., Bacanin, N., Chhabra, A., Suresh, M.: Support vector machine performance improvements for cryptocurrency value forecasting by enhanced sine cosine algorithm. In: Computer Vision and Robotics, pp. 527–536. Springer (2022)
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
16. Yang, X.S.: Firefly algorithms for multimodal optimization. In: International Symposium on Stochastic Algorithms, pp. 169–178. Springer (2009)

17. Yang, X.S.: Bat algorithm for multi-objective optimisation. Int. J. Bio-Inspir. Comput. **3**(5), 267–274 (2011)
18. Zivkovic, M., Bacanin, N., Venkatachalam, K., Nayyar, A., Djordjevic, A., Strumberger, I., Al-Turjman, F.: Covid-19 cases prediction by using hybrid machine learning and beetle antennae search approach. Sustain. Cities Soc. **66**, 102669 (2021)
19. Zivkovic, M., Jovanovic, L., Ivanovic, M., Bacanin, N., Strumberger, I., Joseph, P.M.: XgBoost hyperparameters tuning by fitness-dependent optimizer for network intrusion detection. In: Communication and Intelligent Systems, pp. 947–962. Springer (2022)
20. Zivkovic, M., Stoean, C., Chhabra, A., Budimirovic, N., Petrovic, A., Bacanin, N.: Novel improved salp swarm algorithm: an application for feature selection. Sensors **22**(5), 1711 (2022)



A Review of Deep Learning Techniques for Human Activity Recognition

Aayush Dhattarwal(✉) and Saroj Ratnoo

Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar 125001, India
aayushdhattarwal@gmail.com

Abstract. In recent years, the research in Human Activity Recognition (HAR) has grown manifold due to the easy availability of data and its important role in many real-world applications. Since the performance of classical machine learning algorithms is not up to the mark, the focus is on applying deep learning algorithms for enhancing the efficacy of HAR systems. This review includes the research works carried out during the period of 2019–2022 in three recognition domains-human activity, surveillance systems and sign language. This review considers the methodologies applied, dataset used and the major findings and achievements of these recent HAR studies. Finally, the paper points out the various challenges in the field of Activity Recognition requiring further attention from researchers.

Keywords: Human Activity Recognition (HAR) · Deep Learning · Challenges in HAR · Computer Vision

1 Introduction

Human Activity Recognition (HAR) can be referred to as the process of identifying the physical actions of agents involved in performing the activities. The research in Human activity recognition (HAR) has grown manifolds because of its wide-ranging applications such as daily and sports related activity identification [3, 4, 11], surveillance systems [17, 19] and sign language recognition [21, 23, 24]. The availability of still image and video data featuring individuals engaged in a variety of activities has further sparked interest in the research for human activity recognition. Since performance of classical machine learning techniques depends on the efficacy of the feature extraction step, deep learning algorithms that auto extract features have become the primary focus for HAR these days [5, 6, 14].

In deep learning, the features are derived by applying some non-linear transformation operations on the raw data hierarchically, which in turn, determines the type of deep learning network. Some popular deep learning techniques incorporate Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Least Short-Term Memory (LSTM) networks. The literature has enough evidence to state that the performance of deep learning algorithms is high compared to the handcrafted feature extraction techniques [14]. However, deep learning is not without its challenges. Deep

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 313–327, 2023.
https://doi.org/10.1007/978-3-031-27409-1_28

learning algorithms require a very large amount of data in the training phase, and hence, the computational cost of these algorithms is significantly higher than the traditional machine learning methods. Moreover, the optimization of deep learning architectures is more complex than shallow learning methods.

This paper presents a review of the research work in applying deep learning algorithms for HAR from 2019 to 2022. There is a plethora of work in HAR and hence, due to space constraint, we have restricted our review to 25 research papers pertaining to three application domains, i.e., daily and sports activities, surveillance systems and sign language recognition. This will help the researchers to understand the state-of-the-art scenario for application of deep learning techniques for HAR to a large extent. The paper also highlights the challenges required to be addressed by the research community to enhance the performance of the HAR systems.

The rest of the paper is organized as follows. Section 2 describes the methodology for paper selection. Section 3 presents the literature review on the latest research trends on application of deep learning techniques in HAR. Section 4 lists the challenges faced by HAR systems. Section 5 concludes the paper.

2 Methodology and Paper Selection

This review focuses on research works investigating human activity recognition from image and video data. We have investigated the application of deep learning algorithms for recognizing daily and sports activities, suspicious activities for surveillance systems and sign language activities. Survey and review papers were omitted to prevent repetition.

The research papers are selected by carrying out a systematic literature search in IEEE Xplore, Springer, mdpi and ScienceDirect databases. The literature search consisted of three key concepts, (i) Human Activity Recognition, (ii) Computer Vision, and (iii) deep learning. The literature search was conducted using the following keywords: “HAR”, “human activity recognition”, “sensors” “vision”, “image”, “video”, “activity recognition”, “activity classification”, “optimization”, “deep learning”, “Weizmann”, “KTH”, “UCF sports”, “action recognition”, “HAR for surveillance”, “HAR for sports’ activities”, “HAR for old age houses”, “sign language recognition”. The shortlisted papers were carefully studied to check if the eligibility criteria were met. This way only 25 articles were found to be relevant.

The review is centered on the papers published from 2019 to 2022 inclusive. Figures 1 and 2 show the application area-wise and year-wise distributions of research papers respectively. Section 3 reviews the selected papers in detail in tabular form and gives a brief conclusion in the last column.

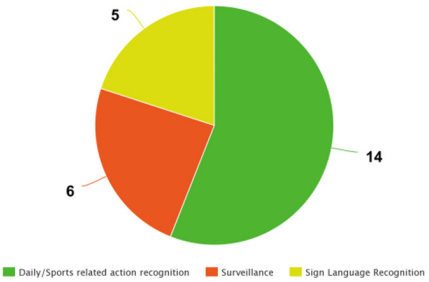


Fig. 1. Application Area-wise distribution of research papers

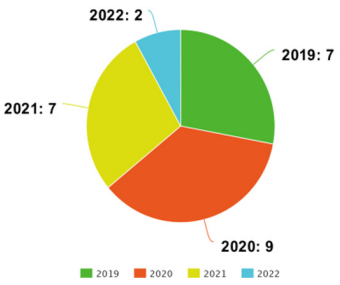


Fig. 2. Year-wise distribution of research papers

3 Review of Literature

3.1 Daily and Sports Related Action Recognition

A major part of Human Activity Recognition research is restricted to trivial daily activities like walking, sitting and waving hand but more recently it has also been expanded to include sports activities. This section considers daily and sports related action recognition as follows:

Noori et al. [1] uses Open-source Pose library to mine anatomical key points from RGB photos for human activity identification. The suggested technique achieves an overall accuracy of 92.4% on a publicly accessible activity dataset, which is superior to the greatest accuracy achieved using the traditional methodologies (78.5%) [1]. A Smartphone inertial accelerometer-based architecture for HAR is developed by Wan et al. [2]. The authors compare CNN, LSTM, BiLSTM, MLP and SVM models for real-time HAR with CNN achieving 91% accuracy on Pamap2 dataset and 92.27% on UCI Dataset [2]. Gnouma et al. [3] introduces a novel method for HAR based on "history of binary motion picture" (HBMI) combined with the Stacked Sparse Autoencoder framework. Excellent recognition rates are achieved without compromising the relevance of the method with the best recognition rate at 100% on the Weizmann dataset (5 actions) [3]. Vishwakarma, et al. [4] proposes a computationally effective and robust HAR scaffold by combining Spatial Distribution of Gradients (SDGs) and Difference of Gaussian (DoG)-based Spatio-Temporal Interest Points (STIP). The method outperforms previous works on the Ballet Dataset with SVM classifier achieving 95.62% accuracy [4]. To help

summarize a person's actions across a film, Chaudhary et al. [5] applies a video summarization method using dynamic images that proves to be cost-efficient with a significant improvement over the existing methods [5]. A deep learning model using residual blocks and BiLSTM is proposed by Li et al. [6]. Experimental results demonstrate that the suggested model improves the performance of previously published models while using fewer parameters [6].

Sargano et al. [7] envisages a new method based on 0-order fuzzy deep-rule based classifier with prototype nature. In this work, features extracted from UCF50 dataset by a pre-trained deep CNN are used for training and testing the model. The proposed classifier outperformed all existing algorithms by 3% achieving 99.50% accuracy while using a single feature descriptor in contrast to other methods which used multiple features [7]. Mazzia et al. [8] presents a short-term pose-based human action recognition using an action transformer, a self-attention model. The method is comprehensively compared to several state-of-the-art architectures [8]. Angelini et al. [9] formulates a novel human action recognition method using RGB and pose together for anomaly detection. The method is tested on UCF101 and MPOSE2019 datasets, significantly improving the recognition accuracy and processing time [9]. Osayamwen et al. [10] discusses probability-based class discrimination in deep learning for HAR with good results on both the KTH and Weizmann datasets comparable to related recent works [10].

Khan et al. [11] envisions a new 26-layered Convolutional Neural Network (CNN) architecture for accurate complex action recognition. The model achieves 81.4%, 98.3%, 98.7%, and 99.2% accuracy respectively on HMDB51, KTH, Weizmann, and UCF Sports datasets, which is an improvement over some of the existing works based on classical machine learning. The limitation of this method is choice of the final layer for feature extraction and the selection of active features [11]. Abdelbaky et al. [12] reviews understanding human motion in three dimensions using an unsupervised deep CNN with the accuracy of 92.67%, which outperforms the recent deep learning works on UCF sports dataset [12]. Sahoo et al. [13] uses sequential learning and depth estimated history images with data augmentation to avoid overfitting with the highest recognition rate of 97.67% on KTH dataset [13]. Tanberk et al. [14] focuses on human activity recognition using a hybrid deep model based on deep learning and dense optical flow which achieves the highest accuracy (96.2%) for MCDS dataset [14]. Some of the important papers in the area are tabulated in Table 1.

From the above table, deep learning (CNN) and its variants have shown significant performance improvement for daily and sports related action recognition across several publicly available datasets.

3.2 Surveillance

Effective and robust surveillance systems are important for maintaining the order at public places such as bus stands, railway stations and airports. Surveillance systems are also required for commercial markets, banks, government organizations and other similar institutions. It tries to detect or predict suspicious activities at public places with the help of an intelligent network of smart commercial off the shelf (COTS) video cameras [15]. Research related to Surveillance is summarized below:

Table 1. Deep Learning in HAR for Daily and Sports related Action Recognition.

References	Author/Year	Techniques Used	Dataset(s)	Performance Metrics Used	Summary
[3]	Gnouma/2019	Deep Recurrent Neural Network, LSTM	KTH, IXmas and Weizmann	Accuracy, Precision Recall, Memory Used	It introduces a novel method for HAR based on the history of binary motion image (HBMI) combined with the Stacked Sparse Auto-encoder framework achieving the best recognition rate at 100% for the Weizmann dataset
[4]	Vishwakarma/2019	Image processing, Algorithm for spatial distribution of gradients	Weizmann, KTH, Ballet Movements, Multi-view IXMAS	ARA, Accuracy	By combining Spatial Distribution of Gradients (SDGs) and Difference of Gaussian (DoG)-based Spatio-Temporal Interest Points (STIP), the method outperforms all other methods on Ballet Dataset with SVM classifier which achieves 95.62% accuracy

(*continued*)

Table 1. (*continued*)

References	Author/Year	Techniques Used	Dataset(s)	Performance Metrics Used	Summary
[5]	Chaudhary/2019	CNN	JHMDB and UCF-sports	ARR	It presents a dynamic image-based video summarization system that significantly outperforms state-of-art approaches, with ARR percentage of 94.5 for JHMDB and 92.6 for UCF-Sports Dataset
[6]	Li/2022	Residual Network and BiLSTM	WISDM and PAMAP2	Accuracy	The proposed method achieves a better performance than the existing models on WISDM and PAMAP2 datasets with the model accuracy at 97.32% and 97.15% respectively and requiring fewer parameters compared to existing models

(*continued*)

Table 1. (*continued*)

References	Author/Year	Techniques Used	Dataset(s)	Performance Metrics Used	Summary
[8]	Mazzia/2022	Multi Layered Perceptron (MLP), LSTM, Action Transformer (AcT) models	MPOSE2021	Accuracy	AcT is introduced as a basic, completely self-attentional architecture that regularly outperforms more complex networks providing a low latency solution. Authors also provide the dataset (MPOSE2021)
[11]	Khan/2021	Convolutional Neural Network (CNN)	HMDB51, UCF, KTH, and Weizmann	Accuracy, FNR, Testing Time	It uses a novel 26-layer CNN for HAR. The accuracy achieved on the four datasets are 81.4%, 99.2%, 98.3%, and 98.7% respectively which outperforms several earlier works
[14]	Tanberk/ 2020	3D-CNN, LSTM	(MCDS), and standard chess board video dataset (CDS)	Accuracy, Precision, Recall, F-Measure	It applies 3D-CNN with LSTM, the model successfully classifies forward human motion as a separate activity on MCDS. For MCDS, it has achieved the highest accuracy (96.2%)

Saba et al. [15] applies a novel CNN model named "L4-Branched-ActionNet" on CIFAR-100 dataset and attained 99.24% classification accuracy [15]. Ahmed et al. [16] presents Motion Classification Based on Image Disturbance which employs a CNN to extract information through convolutional layers and a Softmax classifier in a fully connected layer to categorize human motion. Experiments show high success rates of 98.75% with KTH, 92.24% with Ixmas, and 100% with the Weizmann datasets [16]. Human action recognition by combining DNNs is suggested by Khan et al. [17]. As a result, the suggested PDaUM-based method takes just the most reliable characteristics and feeds them into the Softmax for final recognition [17]. Li et al. [18] presents Deep Learning-Powered Feature Extraction and HAR Scheme. Extensive trials on a real dataset show that the PSDRNN is just as successful as the xyz-DRNN while requiring 56% less time on average for recognition and 80% less time for training [18].

Progga et al. [19] identifies children working as slaves using deep learning. The test accuracy of the CNN model was 90.625%, whereas that of the other two models, both based on transfer learning, was 95.312% and 96.875% [19]. Wu et al. [20] uses pre-trained CNN models for feature extraction and context mining. It utilizes a denoising auto-encoder of comparatively low complexity to deliver an efficient and accurate surveillance anomaly detection system that reduces the computational cost [20]. Some of the important papers in the area of Surveillance are given in Table 2.

The above discussion shows that relatively lesser variations of deep learning algorithms have been applied for surveillance applications. The area needs to be further explored.

3.3 Sign Language Recognition

Sign language is an altogether distinct style of human action where shapes and movements of hands with respect to the upper body are important for sign definition [21]. Research related to sign language recognition is reviewed in this section.

Ravi et al. [21] focuses on CNN that were trained to recognize signs in many languages achieving the accuracy of 89.69% on RGB spatial and optical flow input data [21]. Amor et al. [22] proposes the Arabic sign language alphabet recognition using a deep learning-based technique. CNN and LSTM is used in a pipeline achieving 97.5%accuracy [22]. Suneetha et al. [23] presents automatic sign language identification from video using a 8-stream convolutional neural network which achieves above 80% accuracy on various sign language datasets [23]. Kumar et al. [24] suggests joint distance and angular coded colour topographical descriptor for 3d sign language recognition using a 2-stream CNN which outperforms recent related works on CMU and NTU RGBD datasets [24]. Wadhawan et al. [25] presents a robust model for sign language recognition using deep learning-based CNN. The method achieves state-of-the-art recognition rates of 99.72% and 99.90% on colored and grayscale images [25]. Some of the most important works in the area are listed in Table 3.

The research in identifying sign language is scarce of all the three areas considered in this review. There is further scope of research in the area.

Table 2. Deep Learning in HAR for Surveillance

References	Author/Year	Techniques Used	Dataset(s)	Performance Metrics Used	Summary
[17]	Khan/2020	Deep Neural Network (DNN)	HMDB51, UCF Sports, YouTube, IXMAS, and KTH	Accuracy, FNR, Time	Using DNN-based high-level features on the HMDB51, UCF Sports, KTH, YouTube, and IXMAS datasets, the proposed algorithm achieves an accuracy of 93.7%, 98%, 97%, 99.4%, and 95.2%, respectively, surpassing all prior techniques
[16]	Ahmed/2020	CNN	KTH, IXmas, Weizmann	Accuracy, Time	Using CNN, recognition rates of 98.75% with KTH, 92.24% with Ixmas, and 100% with the Weizmann dataset are achieved
[18]	Li/2020	Feature Extraction, PSDRNN, TriPSDRNN	UniMiBSHAR dataset	Weighted F1-score, MAA	Power Spectral Density Recurrent Neural Network (PSDRNN) and tri-PSDRNN are used. TriPSDRNN achieves the best classification results outperforming the previous works

(*continued*)

Table 2. (*continued*)

References	Author/Year	Techniques Used	Dataset(s)	Performance Metrics Used	Summary
[19]	Progga/2020	CNN	Child labour dataset	Train Accuracy, Validation Accuracy, test Accuracy	It exploits CNN architecture to achieve 96.87% accuracy on Child Labour dataset
[20]	Wu/2020	Convolution Neural Network (CNN)	UCSD Ped1, UCSD Ped2	AUC, EER	Using contextual features with Deep CNN, model performance is improved, complexity and computational overhead is reduced, achieving a high AUC score of 92.4 on the Ped2 dataset

Table 3. Deep Learning in HAR for Sign Language Recognition

References	Author/Year	Techniques used	Dataset(s)	Performance Metrics Used	Summary
[21]	Ravi/2019	CNN, Sign language gesture recognition	RGB-D, BVCSL3D, MSR Daily Activity 3D, UT Kinect, G3D	Precision, recall, Accuracy	It uses four-stream CNN with a multi modal feature sharing method, the network performs better on all the datasets achieving 89.69% recognition rate on RGB spatial and optical flow input data
[22]	Amor/2021	Feature extraction, pattern recognition, Electromyography (EMG), CNN, LSTM	Arabic Sign Language Dataset	Accuracy	CNN with LSTM is used to process feature dependencies for identifying gestures from electromyographic (EMG) signals. This work achieves 97.5% accuracy
[23]	Suneetha/2021	Sign language recognition, M2DA-Net	MuHAVi, NUMA, NTU RGB D, Weizmann	Accuracy	An 8-stream convolutional neural network that models the multi-view motion deep attention network. It (M2DA-Net) achieves 85.12, 88.25, 89.98 and 82.25% accuracy for each of the datasets respectively

(*continued*)

Table 3. (*continued*)

References	Author/Year	Techniques used	Dataset(s)	Performance Metrics Used	Summary
[24]	Kumar/2019	Sign language recognition, CNN	3D ISL dataset (ISL3D), HDM05, CMU and NTU RGB - D (skeletal) action datasets	Accuracy	The proposed method outperforms all previous works on CMU, NTU RGBD datasets achieving a recognition rate of 92.67% and 94.42% respectively
[25]	Wadhawan/2020	CNN, Indian Sign Language (ISL)	Primary Collection	Precision, recall, F-score, Accuracy	The authors have tested the efficacy of the method by implementing 50 CNN models. The approach attains significantly higher rate of 99.90% and 99.72% on gray scale and colored images, respectively

4 Challenges in HAR and Research Directions

Although a lot of research has been carried out to enhance the performance of HAR, the domain is not without constraints and challenges. After reviewing the recent trends of research in HAR in the three application areas, some of the challenges are worth mentioning here. These challenges are applicable beyond the three domains studied for HAR in this paper.

Denoising. Any background noise in the data from ambient sensors affects the performance of HAR models. Moreover, the data collecting devices may also record data other than the main subject. Hence, denoising data obtained from images or videos is essential.

Dealing with Inter/Intra-subject variability. Inter or intra-subject variability in actions in presence of multiple users poses another challenge to HAR systems. Further, the positioning of sensors across the subjects must be uniform. The variability in sensor positions on human or other subjects may also increase the complexity of data being collected for activity recognition.

Availability of Large Labeled datasets. Deep learning algorithms always require a large repository of labeled data in the training phase. Non-availability of large amounts of labeled data particularly for newer domains is another difficulty faced by the researchers. Labeling data from sensors is a time-consuming process.

Skewed Class Distribution. The suspicious activities in surveillance or human or sports activity domains or in some related domains are rare. The skewed class distribution in favour of normal activities can significantly lower the performance of HAR systems. In such circumstances, the class imbalance must be addressed before applying any learning algorithm.

Space and Time Complexity. The major limitation of deep learning models for HAR is the exorbitant space and time complexity and setting the large number of parameters to reach an optimal performance. These activity recognition models trained in one domain cannot be deployed to other domains and researchers must start training the model all over again. Nowadays, the focus is on transfer learning where a model trained in one domain can be used for other related and similar domains with some least amount of training.

The future research can consider the challenges that are listed above and propose HAR systems that address these issues. Moreover, novel HAR approaches may develop scalable, cost-efficient activity recognition systems and consider activity recognition in unfavorable environments.

5 Conclusion

This study has reviewed different applications of deep learning algorithms for HAR in human and sports activities, surveillance systems and sign language recognition. It has considered 25 recent research works only from 2019 to 2022. After investigating research methodology, tools and techniques, dataset used in HAR systems, it is observed that the researchers have achieved quite some success for human activity recognition using deep

learning. However, the field of human activity recognition has a few challenges that also need to be addressed. This body of work may help in identifying the recent trends, and several difficulties associated with various approaches of human activity recognition using deep learning. It is evident from this review that the focus of research in HAR has largely been on daily and sports related action recognition which is gradually moving towards surveillance and sign language recognition systems. In future, scope of the research could be extended to include more domains of HAR and the techniques that can address the challenges identified in this paper.

References

1. Noori, F.M., Wallace, B., Uddin, Md.Z., Torresen, J.: A robust human activity recognition approach using OpenPose, motion features, and deep recurrent neural network. In: Felsberg, M., Forssén, P.-E., Sintorn, I.-M., Unger, J. (eds.) Image Analysis, pp. 299–310. Springer International Publishing, Cham (2019)
2. Wan, S., Qi, L., Xu, X., Tong, C., Gu, Z.: Deep learning models for real-time human activity recognition with smartphones. Mob. Netw. Appl. **25**(2), 743–755 (2019). https://doi.org/10.1007/s11036-019-01445-x
3. Gnouma, M., Ladjailia, A., Ejbali, R., Zaied, M.: Stacked sparse autoencoder and history of binary motion image for human activity recognition. Multim. Tools Appl. **78**(2), 2157–2179 (2018). https://doi.org/10.1007/s11042-018-6273-1
4. Vishwakarma, D.K., Dhiman, C.: A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel. Vis. Comput. **35**(11), 1595–1613 (2018). https://doi.org/10.1007/s00371-018-1560-4
5. Chaudhary, S., Dudhane, A., Patil, P., Murala, S.: Pose guided dynamic image network for human action recognition in Person centric videos. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8 (2019)
6. Li, Y., Wang, L.: Human activity recognition based on residual network and BiLSTM. Sensors 22 (2022)
7. Sargano, A.B., Gu, X., Angelov, P., Habib, Z.: Human action recognition using deep rule-based classifier. Multim. Tools Appl. **79**(41–42), 30653–30667 (2020). https://doi.org/10.1007/s11042-020-09381-9
8. Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., Chiaberge, M.: Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recogn. **124**, 108487 (2022)
9. Angelini, F., Naqvi, S.M.: Joint RGB-pose based human action recognition for anomaly detection applications. In: 2019 22th International Conference on Information Fusion (FUSION), pp. 1–7 (2019)
10. Osayamwen, F., Tapamo, J.-R.: Deep learning class discrimination based on prior probability for human activity recognition. IEEE Access **7**, 14747–14756 (2019)
11. Khan, M.A., Zhang, Y.-D., Khan, S.A., Attique, M., Rehman, A., Seo, S.: A resource conscious human action recognition framework using 26-layered deep convolutional neural network. Multim. Tools Appl. **80**(28–29), 35827–35849 (2020). https://doi.org/10.1007/s11042-020-09408-1
12. Abdelbaky, A., Aly, S.: Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network. Multim. Tools Appl. **80**(13), 20019–20043 (2021). https://doi.org/10.1007/s11042-021-10636-2

13. Sahoo, S.P., Ari, S., Mahapatra, K., Mohanty, S.P.: HAR-depth: a novel framework for human action recognition using sequential learning and depth estimated history images. IEEE Trans. Emerg. Topics Comput. Intell. **5**, 813–825 (2021)
14. Tanberk, S., Kilimci, Z.H., Tükel, D.B., Uysal, M., Akyokuş, S.: A Hybrid deep model using deep learning and dense optical flow approaches for human activity recognition. IEEE Access **8**, 19799–19809 (2020)
15. Saba, T., Rehman, A., Latif, R., Fati, S.M., Raza, M., Sharif, M.: Suspicious activity recognition using proposed deep L4-branched-actionnet with entropy coded ant colony system optimization. IEEE Access **9**, 89181–89197 (2021)
16. Ahmed, W.S., Karim, A.A.A.: Motion classification using CNN based on image difference. In: 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), pp. 1–6 (2020)
17. Khan, M.A., Javed, K., Khan, S.A., Saba, T., Habib, U., Khan, J.A., Abbasi, A.A.: Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multim. Tools Appl. (2020)
18. Li, X., Wang, Y., Zhang, B., Ma, J.: PSDRNN: an efficient and effective har scheme based on feature extraction and deep learning. IEEE Trans. Ind. Inf. **16**, 6703–6713 (2020)
19. Progga, F.T., Shahria, M.T., Arisha, A., Shanto, M.U.A.: A deep learning based approach to child labour detection. In: 2020 6th Information Technology International Seminar (ITIS), pp. 24–29 (2020)
20. Wu, C., Shao, S., Tunc, C., Hariri, S.: Video anomaly detection using pre-trained deep convolutional neural nets and context mining. In: 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), pp. 1–8 (2020)
21. Ravi, S., Suman, M., Kishore, P.V.V., Kumar, E.K., Kumar, M.T.K., Kumar, D.A.: Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition. J. Comput. Lang. **52**, 88–102 (2019)
22. Ben Hej Amor, A., El Ghoul, O., Jemni, M.: A deep learning based approach for Arabic Sign language alphabet recognition using electromyographic signals. In: 2021 8th International Conference on ICT & Accessibility (ICTA), pp. 1–4 (2021)
23. M. S., M.V.D., P. P.V.V. K.: Multi-view motion modelled deep attention networks (M2DA-Net) for video based sign language recognition. J. Vis. Commun. Image Represent. **78**, 103161 (2021)
24. Kumar, E.K., Kishore, P.V.V., Kiran Kumar, M.T., Kumar, D.A.: 3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2—stream CNN. Neurocomputing **372**, 40–54 (2020)
25. Wadhawan, A., Kumar, P.: Deep learning-based sign language recognition system for static signs. Neural Comput. Appl. **32**(12), 7957–7968 (2020). https://doi.org/10.1007/s00521-019-04691-y



Selection of Replicas with Predictions of Resources Consumption

José Monteiro, Óscar Oliveira, and Davide Carneiro(✉)

CIICESI, Escola Superior de Tecnologia e Gestão, Politécnico do Porto, Porto, Portugal
{8200793,oao,dcarneiro}@estg.ipp.pt

Abstract. The project Continuously Evolving Distributed Ensembles (CEDEs) aims to create a cost-effective environment for distributed training of Machine Learning models. In CEDEs, datasets are broken down into blocks, replicated and distributed through the cluster, so that Machine Learning tasks can take place in parallel. Models are thus a logical construct in CEDEs, made up of multiple base models. In this paper, we address the problem of distributing tasks across the cluster while adhering to the principle of data locality. The presented optimization problem assigns for each block a base model with the objective of minimizing the overall prevision of resources consumption. We present an instance generator and three datasets that will provide a means of comparison while analyzing solution methods to employ in this project. For testing the system architecture, we solved the datasets with an exact method and the computational results validate that to comply with the CEDEs requirements, the project needs for a more stable and less demanding solution method in terms of computational resources.

1 Introduction

The project Continuously Evolving Distributed Ensembles (CEDEs) aims to create a distributed environment for Machine Learning (ML) tasks (e.g. model training, scoring, predictions). One of its main goals is that models can evolve over time, as data changes, in a cost-effective manner. Several architectural aspects enable this.

A block-based distributed file system with replication is used e.g., Hadoop Distributed File System (HDFS; see [8]). This means that large datasets are split into relatively small fixed-size blocks. These blocks are then replicated, for increased availability and robustness, and distributed across the cluster. Thus, when a block is necessary, namely for training or predicting, there might be several available nodes to read from in the cluster. Moreover, each node will be in a different state in terms of available resources or job queues. There is thus the need, for each task, to select the most suitable replica.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 328–336, 2023.
https://doi.org/10.1007/978-3-031-27409-1_29

The problem exists in several tasks: when training a new model, when updating an existing model, and when making predictions. A new model is trained from a dataset selected by the user. However, since the dataset is split into blocks, several so-called base models are actually trained, one for each block. Therefore, the actual model is a logical construct, an ensemble [5], made up of multiple base models. The performance of this ensemble is obtained by averaging the performance of its base models.

Moreover, ensembles can be quickly and efficiently modified by adding or removing base models. This may happen as a requirement (the user may desire ensembles of different complexities) or as a way to deal with data streaming scenarios [11]: new base models can be trained for newly created blocks, which may eventually replace older or poorer ones. This allows the model to maintain its performance over time with minimal resource consumption.

Finally, the problem of selecting the best replicas also applies when making predictions, at two levels. On the one hand, predictions are often made on datasets that are stored in the file system (with replication). On the other hand, the base models themselves are stored in the file system and are also replicated. Therefore, this means that there will be several nodes with each necessary base model and with each necessary block. Thus, it is necessary to select the best ones at any moment.

One central principle that governs the entire system is the data locality principle [2], that is, rather than the data being transported across the cluster, the computation is moved to where the data is.

The architecture of CEDEs is depicted in Fig. 1. It is composed of several main components, namely:

- A front-end through which a human user can interact with the system. There is also an Application Programming Interface (API) for machine-to-machine calls.
- A storage layer (SL) implemented as an HDFS cluster were large datasets are split into blocks of fixed size.
- A metadata module (MM) estimates the cost of each individual task, i.e., base model training, based on meta-learning as described in [3,4].
- An optimization module (OM) that has, as main responsibility, to schedule the ensemble tasks considering the predictions given by the MM. A second responsibility, and the focus of this paper, is to assign to each of the dataset blocks (that is distributed and replicated through the cluster) a base model to minimize the overall resource consumption.
- A coordination module (CM) which interacts with the OM and MM.
- A blockchain module that records all the operations in the system.

This paper describes in more detail the secondary optimization problem solved by the OM for assigning ensemble base model to replicas to minimize the prevision of resources consumption, and the instance generator and solution method implemented to test the system architecture. As this problem, as far as we know, was never tackled in the literature, we needed to evaluate if the module could

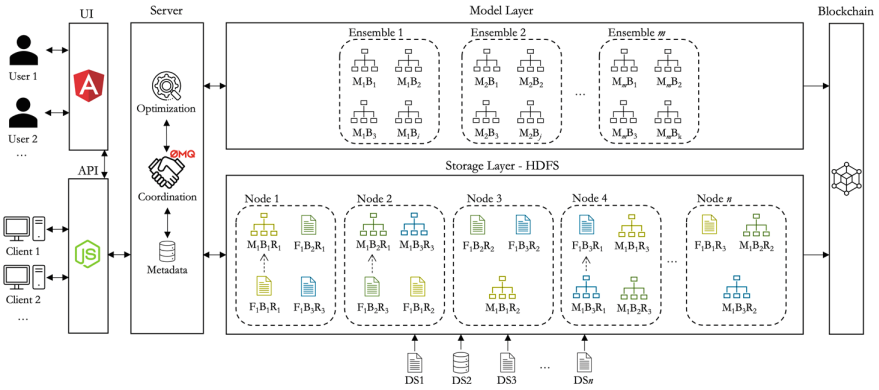


Fig. 1. Architecture of the CEDEs project

consider an exact solution method (through a solver) or if heuristic methods should be considered as, however without the guarantee of the obtention of the optimal solution, usually, they can provide good results with considerably less computational resources.

The remainder of this paper is structured as follows. Section 2 presents the optimization problem above mentioned. Section 3 presents the instance generator for the problem. In Sect. 4 computational experiments with the generated datasets and an exact solution method (using an optimization solver) are reported. Finally, conclusions and future work directions are given in Sect. 5.

2 Replica Selection

The problem considers a cluster with a set N of nodes in which datasets are stored to train machine learning ensembles (with a set M of base models). The considered file system (HDFS) creates replicas of the blocks of the datasets so that the same block ($b \in B$) is available simultaneously in multiple nodes ($n \in N$) to make predictions. Noteworthy, that although not yet considered by the optimization model, the emsemble base models ($m \in M$) are, also, replicated and stored by the file system. Therefore, multiple nodes will have the same base models available for making predictions. The set R represents the various resources to be considered (e.g., CPU, memory) by the optimization model. The values of the resources are represented by their percentage ($\in [0, 1]$) for current or predictions of usage.

let x_{nbm} be a binary variable which is equal to 1 if block $b \in B$ from node $n \in N$ will use model $m \in M$, 0 otherwise. Thus, the mathematical model can expressed as:

$$\min \sum_{r \in R} w_r \sum_{n \in N} \sum_{b \in B} \sum_{m \in M} x_{nbm} p_{nrm} \tag{1}$$

Subject to:

$$s_{nr} + \sum_{b \in B} \sum_{m \in M} x_{nbm} p_{nrm} \leq 1 \qquad \forall n \in N, r \in R \tag{2}$$

$$\sum_{n \in N} \sum_{m \in M} x_{nbm} = 1 \qquad \forall b \in B \tag{3}$$

$$\sum_{m \in M} x_{nbm} \leq a_{nb} \qquad \forall n \in N, b \in B \tag{4}$$

$$\sum_{n \in N} \sum_{b \in B} x_{nbm} \geq 1 \qquad \forall m \in M \tag{5}$$

$$x_{nbm} \in \{0, 1\} \qquad \forall n \in N, b \in B, m \in M \tag{6}$$

where:

- w_r represents the weight ($\in [0, 1]$) of the resource $r \in R$ on the calculation of the objective function. In addition, it is assumed that $\sum_{r \in R} w_r = 1$ is guaranteed.
- p_{nrm} represents, for the dataset under consideration, the prediction on the resource ($r \in R$) consumption on node $n \in N$ using model $m \in M$.
- s_{nr} represents the current resource usage $r \in R$ value on node $n \in N$.
- a_{nb} is a binary variable equal to 1 if block $b \in B$, from the dataset under consideration, has a replica on node $n \in N$, 0 otherwise.

Expression (1) denotes the objective to attain, namely the minimization of the resource consumption prediction considering the weights (w_r) on each resource. Constraints (2) ensure that the resources do not exceed their availability. Constraints (3) ensure that a replica of all blocks that constitute the dataset is chosen while (4) ensure that the replica exists at the node. Constraints (5) ensure that each model in M is used in the training at least one time (The number of blocks of a dataset determines the number of base models to be trained but the number of base models can be smaller than the number of blocks.). Constraints (6) ensure that all decision variables are binary.

As already stated, to the best of our knowledge, this problem was not approached in the literature. However, we refer to the following articles for the interested reader. In [10], the authors present a dynamic data locality-based replication for HDFS that considers a file popularity factor in the replication. in [7], the author proposes a best-fit approach to find best replica for the requesting users taking into account the limitation of their network or hardware capabilities. This algorithm matches the capabilities of grid users and the capabilities of replica providers. In [1] it is proposed a solution method that considers fairness among the users in the replica selection decisions in a Grid environment where the users are competing for the limited data resource.

3 Instance Generator

The OM receives the data in a JSON[1] object as represented in Listing 1. In the following listings, ... represents objects that were removed to facilitate the reading.

```
{ "nodes": [
    {
      "id": "23bccd93-98d9-4b47-b6d1-45b6d2af9fe0",
      "resources": { "cpu": 0.111, "mem": 0.122 },
      "previsions": [
        { "model": "regression", "resources": { "cpu": 0.111, "mem": 0.112 } },
        ...
      ]
    },
    ...
  ],
  "blocks": [
    {
      "id": "dc8beb40-378a-4ad8-aa61-16e4ff6e0a3a",
      "nodes": ["23bccd93-98d9-4b47-b6d1-45b6d2af9fe0", ...]
    },
    ...
  ]
}
```

Listing 1. Instance structure

In Listing 1 nodes (`1-9`) and blocks (`12-18`) are defined. Each node is defined by an identifier (`3`), the current resources consumption (`4`), and the previsions of resources consumption for training a block of the dataset using the corresponding model (`6`). Each dataset block is defined by an identifier (`14`) and a list of nodes in which a replica of this block exist (`15`).

To test the the system architecture and, specially, the OM, an instance generator was implemented to create datasets of instances (with the structure presented in Listing 1) using beta distribution [6] to generate the random values. This distribution was already used to create generators for other optimization problems, e.g., cutting and packing problems [9]. The reasoning for using this distribution is that it can assume a variety of different shapes (see Fig. 2), depending on the values of its parameters α and β, e.g., For α and β equal to 1 the distribution becomes a uniform distribution between 0 and 1 (represented with a straight line in Fig. 2). The probability density function for the beta distribution is given by:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 \mu^{\alpha-1}(1-u)^{\beta-1}d\mu}, 0 \leq x \leq 1, \alpha > 0 \text{ and } \beta > 0 \tag{7}$$

The following beta distributions (B), depicted in Fig. 2, are considered by the generator assuming the tuples (α, β):

$$B = [(1,3),(2,5),(2,2),(0.5,0.5),(1,1),(3,1),(5,2)]$$

Listing 2 presents the generator (JSON) configuration file structure.

[1] https://www.json.org/.

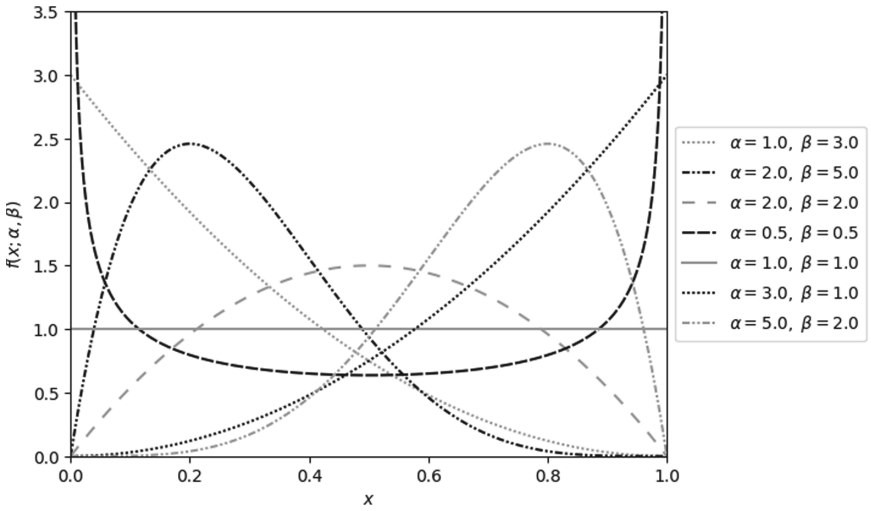


Fig. 2. Shapes of the beta distribution

```
1  { "folder": "./dataset",
2    "number_of_instances": 10,
3    "distribution": null,
4    "parameters": {
5      "number_of_nodes": [50, 100],
6      "number_of_blocks": [150, 200],
7      "replica_distribution": 0.25,
8      "resources": [
9        { "name": "cpu", "current-consumption": [0.2, 0.3] },
10       ...
11     ]
12   }
13 }
```

Listing 2. Generator configuration file

In this configuration file it can be specified: the output folder (`1`), the number of instances (n_{inst}) that will constitute the dataset (`2`), the beta distribution to be used (`3`) and the parameters that will define how each instances will be generated (`4-11`). If an integer value $\in [0, 6]$ is given for `distribution` (`3`) the corresponding beta distribution in B will be used. Otherwise, if a `null` value is given, n_{inst} instances for each distribution in B are generated, then n_{inst} instances are randomly selected between those $n_{inst} \times |B|$ generated instances.

To define how each instances will be generated, the following parameters can be used (`4-11`) with range of number of nodes (`5`), the range of number of blocks that constitute the dataset (`6`), and the percentage of nodes in which each replica must exist(`7`). The number of base models will be generated using the lower bound of the range of number of blocks (`6`) and the randomly generated number of blocks. The resources are defined by the name and the range to define the current node consumption of the corresponding resource (`9`).

The generator creates feasible solutions, as it defines the values constructing one solution (not guaranteed to be optimal) to generate the resource prevision values, in brief, first it is defined the cluster, with randomly generated nodes, blocks and models. Next, a solution is created (without previsions), i.e., for each block, a node in which it exist and a model is selected. Next, considering the generated solution, the models previsions using the available node free space for each resource are distributed randomly. Finally, the missing previsions are added to the solution using the minimum and maximum values of the previsions generated for each resource in the previous point as the range for the new models previsions.

4 Computational Experiments

The problem was modeled and implemented with the Google's mathematical optimization tools OR-Tools[2] for Python (using the SCIP mixed integer programming solver). The experimental tests were run on a computer with processor Intel(R) Core(TM) i7-8650U and 16Gb of RAM on Windows Subsystem for Linux version 2 of Windows 11 Pro.

The OM solver utilizes an JSON configuration file that allows to configure the solver behaviour. Most of the parameters are not reported in this paper as they only serve as input/output options and to limit the execution time for the exact solver. However it should be noted, that the configuration file contains an object that specifies which resources and limits to use for solving a particular instance.

We have generated three datasets for the purpose of testing the optimization method considering the data on Listing 3 varying the range of the number of nodes and blocks minimum models (5-7) accordingly to Table 1.

```
{ "folder": "../data/test1",
  "number_of_instances": 10,
  "distribution": null,
  "parameters": {
    "number_of_nodes": [15, 20],
    "number_of_blocks": [35,50],
    "replica_distribution": 0.50,
    "resources": [
      { "name": "cpu", "current-consumption": [0.2, 0.3] },
      { "name": "mem", "current-consumption": [0.1, 0.2] }
    ]
  }
}
```

Listing 3. Dataset 1 configuration file

Table 2 presents the computational times, in seconds, solving the datasets considering the weights on the resources consumption of 0.6 CPU and 0.4 for memory.

From Table 2 it can be stated that the exact method can, on some harder instances, requires an high computational time to solve the problem to optimality. As this solution method will serve an aid for decision making, the computational time should be more stable and predictable. The results obtained

[2] https://developers.google.com/optimization.

Table 1. Configuration for creating the datasets of instances

	Dataset 1	Dataset 2	Dataset 3
Number of nodes	[10, 20]	[10, 20]	[20, 40]
Number of blocks	[10, 20]	[20, 40]	[40, 60]

Table 2. Results solving the datasets with the exact method

Instance	Dataset 1	Dataset 2	Dataset 3
1	0.24	95.53	91.96
2	0.80	0.42	150.58
3	0.27	15.29	8.32
4	0.39	3.14	8.08
5	0.63	12.27	20.17
6	1.69	0.42	40.75
7	0.14	19.93	113.44
8	0.71	1.44	11.48
9	1.50	0.41	7.03
10	0.22	9.38	130.96
Average	0.66	15.82	58.28

justify the study of a more appropriate solution method such as an heuristic or metaheuristics. These approaches, although without the guarantee of finding the optimal solution, usually obtain good results with considerably less computational resources than the ones required by exact methods.

The generator, datasets and solver that support this paper are available from the corresponding author upon request.

5 Conclusions

Given the changing requirements of Machine Learning problems in recent years, particularly in terms of data volume, diversity, and speed, new techniques to deal with the accompanying challenges are required. CEDEs is a distributed learning system that works on top of a Hadoop cluster and takes advantage of blocks, replication, and balancing.

In this paper we presented the problem that the optimization module must solve assigning for each block dataset a base model with the objective of minimizing the overall prevision of resources consumption. Additionally, we present an instance generator and the results obtained solving to optimality three distinct datasets. These results demonstrated that the exact method requires on harder instances an high computational time, justifying the study of heuristics methods for solving this problem as a solution method that requires less computational resources is needed for satisfying usability requirements of the CEDEs project.

Extensions of this work will be done. Although the optimization module does not consider the problem presented in Sect. 2 in its isolated form, we expect to study the implementation of heuristic or metaheuristic solution methods using the results obtained by the exact method for comparison for this problem.

Acknowledgments. This work has been supported by national funds through FCT—Fundação para a Ciência e Tecnologia through projects UIDB/04728/2020 and EXPL/CCI-COM/0706/2021.

References

1. AL-Mistarihi, H.H.E., Yong, C.H.: On fairness, optimizing replica selection in data grids. IEEE Trans. Parallel Distrib. Syst. **20**(8), 1102–1111 (2009). https://doi.org/10.1109/TPDS.2008.264
2. Attiya, H.: Concurrency and the principle of data locality. IEEE Distrib. Syst. Online **8**(09), 3 (2007). https://doi.org/10.1109/MDSO.2007.53
3. Carneiro, D., Guimarães, M., Carvalho, M., Novais, P.: Using meta-learning to predict performance metrics in machine learning problems. Expert Syst. (2021). https://doi.org/10.1111/exsy.12900
4. Carneiro, D., Guimarães, M., Silva, F., Novais, P.: A predictive and user-centric approach to machine learning in data streaming scenarios. Neurocomputing **484**, 238–249 (2022). https://doi.org/10.1016/j.neucom.2021.07.100
5. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Front. Comput. Sci. **14**, 241–258 (2020). https://doi.org/10.1007/s11704-019-8208-z
6. Gupta, A.K., Nadarajah, S.: Handbook of Beta Distribution and Its Applications. CRC Press (2004)
7. Jaradat, A.: Replica selection algorithm in data grids: the best-fit approach. Adv. Sci. Technol. Res. J. **15**, 30–37 (2021). https://doi.org/10.12913/22998624/142214
8. Shvachko, K.V., Kuang, H., Radia, S.R., Chansler, R.J.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10 (2010)
9. Silva, E., Oliveira, J.F., Wäscher, G.: 2dcpackgen: a problem generator for two-dimensional rectangular cutting and packing problems. Eur. J. Oper. Res. **237**, 846–856 (2014). https://doi.org/10.1016/j.ejor.2014.02.059
10. Thu, M.P., Nwe, K.M., Aye, K.N.: Replication Based on Data Locality for Hadoop Distributed File System, pp. 663–667 (2019)
11. Zhou, L., Pan, S., Wang, J., Vasilakos, A.V.: Machine learning on big data: opportunities and challenges. Neurocomputing **237**, 350–361 (2017). http://orcid.org/10.1016/j.neucom.2017.01.026



VGATS-JSSP: Variant Genetic Algorithm and Tabu Search Applied to the Job Shop Scheduling Problem

Khadija Assafra[1(✉)], Bechir Alaya[2], Salah Zidi[2], and Mounir Zrigui[1]

[1] Research Laboratory in Algebrat, Numbers Theory and Intelligent Systems University of Monastir, Monastir, Tunisia
khadija.assafra@gmail.com

[2] Hatem Bettaher IResCoMath Research Unit, University of Gabes, Gabes, Tunisia

Abstract. In this article, we have studied the optimization problem of a JSSP production cell (Job-Shop Scheduling Problem) whose scheduling is very complex. Operational research and artificial intelligence-based heuristics and metaheuristics are only two of the many approaches and methodologies used to analyze this type of problem (neural network, genetic algorithms, fuzzy logic, tabu search, etc.). In this instance, we pick a technique based on the hybridization of TS (Tabu Search) and GA (Genetic Algorithm) to reduce the makespan (total time of all operations). We employed various benchmarks to compare our VGATS-JSSP (Variant Genetic Algorithm and Tabu Search applied to the Job Shop Scheduling Problem) with the literature to demonstrate the effectiveness of our solution.

Keywords: Optimization · Job Shop Scheduling Problem · Hybridization · Genetic Algorithm · Tabu Search

1 Introduction

To best meet the qualitative and/or quantitative needs of uncertain customers or managers in the industry, more complicated process management systems are deployed in the job shop setting [1]. This has enabled the development of new methods, especially in the Job Shop environment where demand quantities are unpredictable and large. The amount of requests automatically results in a large number of tasks that can lead to system overload [2].

This complexity is one of the reasons why the problems they pose are problems of optimization, planning, scheduling, and management which are generally recognized as very difficult to solve [3] . They must be studied methodically and rigorously to detect and quantify their impact on the quantitative and qualitative performance of the job shop [4].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 337–349, 2023.
https://doi.org/10.1007/978-3-031-27409-1_30

The task management problem consists of organizing and executing tasks in time, given time constraints and constraints related to the availability and use of the necessary resources.

Indeed, one of the challenging NP optimization issues explored for decades to identify optimal machine sequences is JSSP, which tries to schedule numerous jobs or operations on some machines where each operation has a unique machine route [5]. The primary goal of optimization was to reduce the maximum execution time (also known as Makespan) of all tasks [6].

The machine assignment problem and the operation sequence problem, which require assigning each operation to a machine and figuring out the start and end timings for each operation, are the two subproblems that must be resolved in order to complete the JSSP [7].

The JSSP issue is a significant problem nowadays. The majority of literature research focuses on speeding up utilization and decreasing completion times. As a result, the majority of studies concentrate on the use of heuristic and meta-heuristic techniques as SA (Simulated Annealing), PSO (Particle Swarm Optimization), FL (Fuzzy Logic), etc. The majority of methods are GA, TS, and ACO (Ant Colony Optimization) [8].

Academics are becoming more and more interested in the creation and use of hybrid meta-heuristics since these hybrid techniques integrate various ideas or elements of multiple meta-heuristics in an effort to combine their strengths and eradicate their shortcomings.

It is in this context that this article has been written. We propose a comparison between our results of VGATS-JSSP: Variant Genetic Algorithm and Tabu Search applied to the Job Shop Scheduling Problem, with the literature while using the same benchmarks and parameters.

The structure of this article is as follows: Section II provides examples of the JSSP-related tasks. The JSSP is described in Section III. GA and TS, their operators, and their parameters are discussed in Section IV. The dataset and results are described in Section V. Section VI offers a conclusion to the article.

2 Related Work

The use of GA in JSSP has been suggested by Davis et al. [9], in this study, 15 benchmarks were analyzed for scheduling operations and rescheduling new operations to reduce makespan. In [10], the authors designed the genetic algorithm to minimize manufacturing time, minimize total installation time, and minimize total transport time.

Ant Colony Optimization (ACO) in the JSSP, Jobs must recognize an appropriate machine to execute them. Just as ants search for the shortest distance to a source of food, activities should search for the shortest way to reach machines [11]. The ant house and food source are comparable at the beginning of the activity and at the end of JSSP. In [12], the authors proposed to improve the reach of the Flexible JSSP (FJSSP). The accompanying aspects are carried out on their improved ACO algorithm: select the machine rules problems, introduce

a uniform scattered component for the ants, modify the pheromone steering instrument, select the node strategy, and update the pheromone system.

The first implementation of Variable Neighborhood Search (VNS) to solve JSSP was introduced in 2006 by Sevkli and Aydin [13]. In [14], the authors offers a new VNS implementation, to minimize workshop planning time with installation times, based on different local search techniques. The VNS algorithm proposed by Ahmadian et al. in [8], consists of decomposing JIT-JSS into smaller subproblems, obtaining optimal or quasi-optimal sequences (to perform the operations) for the sub-problems, and generating a program, i.e. determining the time to complete operations.

Bożejko et al. in [15], presented a parallel solution of Tabu search for the Cyclic JSSP (CJSSP), It presents search by Tabu as a modification of the local search method.

Li and Gao in [16], suggested an effective hybrid approach that combines the tabu search (TS) and genetic algorithm (GA) for FJSSP to reduce makespan. The exploration is carried out using the GA, which has a powerful global search capability, and the mining is achieved using the TS, which has a strong local search capability.

Du et al. in [17] provided a timeline to reduce the amount of time needed to solve the Assembly Job Shop Scheduling Problem (AJSSP). A proposed and developed integrated hybrid particle swarm optimization (HPSO) technique PSO with Artificial Immune is used to solve the AJSSP because it is an NP-hard problem with high levels of complexity.

The solution presented in [18] is to apply VNS based on a GA to improve search capacity and balance intensification and diversification in the Job Shop environment. The VNS algorithm has shown excellent local search capability with structures for a thorough neighborhood search. Thus the genetic algorithm has a good capacity for global search.

3 Job Shop Scheduling Problem

Job Shop Scheduling Problem (JSSP), is a well-known insoluble combinatorial optimization problem; that was presented in [19]. It is one of the tough optimizations of NP problems studied for decades and that aims to schedule multiple operations on some machines. The optimization has mainly focused on minimizing the Makespan of entire operations.

The JSSP has been addressed by considering the availability of operations as well as the human resources and tools needed to execute an operation. The the objective here is to minimize the dwell time of the products in the workshop from the customer order until the end of product processing in the workshop [20].

On the other hand, in JSSP, operations are grouped into jobs; each job has its product range for which other constraints are introduced and assigned to machines [21]. The most basic version of JSSP is: The given n jobs $J_1, J_2, \ldots, J_n$, which must be scheduled on m machines with variable processing power, as

shown in Fig. 1, we present by the circles the jobs input and then the processing in the machines by rectangles.

However, most studies are interested in developing specific aspects of optimization for static or deterministic scenarios. Several propositions in the literature address different classes of manufacturing systems subjected to imponderable and unexpected events, such as cancellation of employment; machine failures; urgent orders modification; of the due date (advance or postponement) delay in the arrival of raw components, or materials; and changes in employment priority [22]. The factors to consider when describing the Job Shop problem are:

- Arrival model
- work order
- Performance evaluation criterion
- Number of machines (work stations),

There are two types of arrival patterns [23]:

- Static: n jobs come to an idle machine and want to be scheduled for work
- Dynamic: intermittent arrival

There are two types of work order:

- Fixed and repeated order- problem of flow shop
- Random order—All models are possible

Some performance evaluation criterion:
- Makespan (total completion time of all operations),
- average work time of jobs in the warehouse,
- Delay
- Average number of jobs in machines
- Use of machines.

The C_{max} parameter, the objective function that represents the minimum manufacturing time and indicates the performance measure used to minimize (the evaluation function). The value of C_{max} is equivalent to the production time it takes to complete all jobs, taking into account the restrictions imposed on the occupation of the machine [24].

- t_{ij} : represent the starting time of operation O_{ij}
- C_{ij}: represent the completion time.
- C_i : represent the completion time of job i

$$C_i = max\, C_{ij}, j = 1...n_i \tag{1}$$

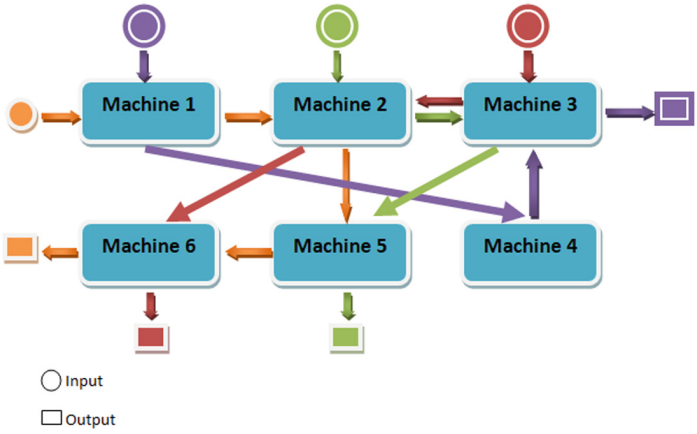


Fig. 1. Example of Job Shop scheduling

Let the following definitions:

- The first operation is an operation without predecessors: the operation O_{i1} is the first operation for work i.
- The end operation is an operation without successors: the operation $O_{in_i}$is the terminal operation for work i.
- A ready operation is an operation that has not yet been scheduled while all of its predecessors have been.
- A program no-idling schedule satisfies the no-idle constraint on each machine. In other words, if the operation O_{ij} is executed just before operation $O_{i'_{j'}}$ on the same machine, so:

$$C_{ij} = t_{ij} + P_{ij} = t_{i'_{j'}} \tag{2}$$

4 Genetic Algorithm and Tabu Search to Solve Job Shop Scheduling Problem

4.1 Genetic Algorithm

Genetic algorithms (GA) strive to reproduce the natural evolution of individuals respecting the law of survival declared by darwin. The basic principles of GA were originally developed by Holland to meet to specific needs in biology, were quickly applied to successfully solve problems combinatorial optimization in operations research and intelligence learning problems artificial.

As part of the application of GA in an optimization problem combinatorial, an analogy is developed between an individual in a population and a solution to a problem in global solution space.

The usage of genetic algorithms requires the following five fundamental components:

- A principle of coding the elements of the population, which consists in associating with each of the state-space points to a data structure, and the quality of that data encoding determines the success of genetic algorithms; although the binary encoding was originally widely used, actual encodings are now widely used, including in the fields of application for the optimization of problems with real variables.
- a method for creating the first population that must be able to create an uneven distribution of people to serve as a foundation for subsequent generations; the initial population's selection is crucial since it affects how quickly the world converges to its optimal state; It is crucial that the starting population is dispersed across the entire research region if there are few details available about the topic to be solved.
- a function to be optimized, called fitness or individual evaluation function,
- The crossover operator recomposes the genes of the individuals currently present in the population, while the mutation operator ensures state-space exploration. operators to explore state space and diversity the population across generations.
- the probability that the crossover and mutation operators will be applied, as well as the size of the population, the number of generations, or the stopping criterion.

4.2 Tabu Search

The idea of Tabu Search (TS) is defined as the exploration of the space of all possible solutions with sequential movement by Glover (1990)

TS is a local search method. it proceeds by exploring for a common solution all its neighborhood N(s). At each iteration, the best solution is this district is retained as a new solution even if its quality is lower than the current solutions. This strategy can lead to cycles, to avoid them we memorize the last k configurations visited in short-term memory, and we prohibit any movement which results in one of these configurations. This memory is called tabu memory or tabu list. It is one of the essentials of this method. She permits to avoid any cycle of length less than or equal to k.

By keeping the list tabu, the best solution may have a tabu status. In this case, we allow ourselves to accept this solution all the same in neglecting its tabu status, it is the application of the criterion suction.

4.3 Genetic Algorithms for Job Shop Scheduling Problem

A chromosomal representation of a solution is necessary for the application of GA to a specific issue (in our case, task planning). It is sufficient to present the sequencing of tasks on a single machine if the jobs move through the machines in the same order. Therefore, a calendar is considered a permutation defining the order in which the jobs pass through the machines. The position of a job in the defined chromosome is its order number in the sequence. The number of operations is measured from left to right in ascending order [25] .

In Fig. 2, example of 4 * 4 representation of the chromosome in JSSP, Oij consist of the operation with i: present the number of job ($i = 0 \ldots n-1$) and j: present the number of machine ($j = 0 \ldots m-1$), n and m are the total number of job and the total number of machine,For example, O00 refers to the first operation of Job Number 0 in Machine Number 0, while O11 represents the second operation of Job Number 1 in Machine Number 2. O32 denotes the third machine 2 and fourth job 3. The solutions to the JSSP are defined by the sequences of this number. An optimal solution is one that has the minimum makespan. In the crossover phase, we adopted the uniform crossover after population generation. Each gene is chosen at random from a set of related genes on the parent chromosomes. It's not always the case that combining two good solutions produces a better or equally good result. Given that the parents are excellent, there is a participated that the child will also be excellent. If the child is poor (a bad solution), it will be eliminated at "Selection" in the following iteration.

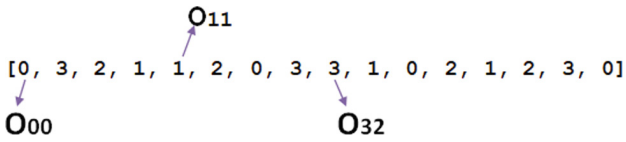


Fig. 2. 4*4 Chromosome representation

The Mutation operator consists of swap mutation, we select two genes from our chromosome and exchange their values.

For the selection phase, we employed the elitist technique, which entails maintaining a sort of population archive with the optimal non-dominated solutions discovered during the search. This population will take part in the reproduction and selection processes. In this instance, the selection pressure S, the probability of selecting a member of the present population of rank n, the probability of selecting a member of the Pareto population.

4.4 Tabu search for Job Shop Scheduling Problem

In this section, we demonstrate how the Tabu method's following components are implemented in JSSP:

- The generation of the initial solution
- The neighborhood generation function,
- Neighborhood assessment,
- The tabu list implementation.

The generation of the initial solution: In our application, we have generated a solution initial randomly, using coding by the priority rules. This type of coding allows us to have to each use a feasible solution thanks to the decoding algorithm used.

The neighborhood generation process: The quarter employed in [12] has a significant impact on the Tabu method's quality. In this section, we will present an improvement of the proposed neighborhood function by Gröflin and Klinkert for the job shop with blocking. For this, we will use the representation based on the alternative graphics.

Neighborhood assessment: The complexity of a resolution approach based on the local search is highly method dependent neighborhood assessment to determine the best neighbor. However, the full assessment, i.e. the calculation of the start dates of all operations, of each neighbor, takes a considerable time. It has been shown that nearly 90 the resolution time is taken by the evaluation of the neighborhoods [26].

The tabu list implementation: To avoid the trap of local optima in which the research process is in danger of collapsing. The research Tabu uses the tabu memory trick. The structure of the implanted memory is made using a circular list of size k. List management follows the First In First Out (FIFO) strategy (output order of list items is that of their insertion). This list is updated at each iteration of the search. At each iteration, a new element is introduced and an older one is published.

The items in the list must have enough information to accurately memorize the solution visited. In our case, the elements of the list are integers, each integer represents the number of the pair of alternating arcs concerned about the transition made on a solution to move to her neighbor.

5 Benchmarks and Results

The datasets utilized and the results are explained in great length in this section. The suggested algorithm is tested on a computer with the following specs after being implemented in the "Java" programming language: Microsoft Windows 10 Professional, an Intel Core i5-5200U processor clocked at 2.20 GHz, and 8 GB of RAM.

5.1 Benchmarks

Benchmarks are useful for knowing the performance of resolution methods.

In the literature, several benchmarks exist for operational research problems. Most of them are grouped on the site [27] of the OR Library where it is possible to download. There are also results and references from the authors of the benchmarks. Thereby as far as the JSSP is concerned, you can download a file containing 82 instances of different sizes grouping the main benchmarks in the literature and giving the source references for these instances. Instances have names made up of the letters that often represent the initials of the names of their authors and numbers to differentiate them between them. They are composed of n jobs and m machines and their size is given by n×m.

The most used instances for benchmarks in the literature are:

- abz5 to abz9 introduced by Adams et al. (1988),
- ft06 (6 × 6), ft10 (10 × 10), ft20 (20 × 5) introduced by Fisher and Thompson (1963)
- la01 to la40 constitute 40 instances of different sizes (10 × 5, 15 × 5, 20 × 5, 10 × 10,15 × 10, 20, 10, 30 × 10, and 15 × 15) are from Lawrence (1985);
- orb01 to orb10 come from Applegate and Cook (1991),
- swv01-swv20 introduced by Storer et al. (1992),
- yn1-yn4 introduced by Yamada and Nakano (1992). The Fisher and Thompson dataset ft06 example can be found in Fig. 3.

```
6  6
2  1  0  3  1  6  3  7  5  3  4  6
1  8  2  5  4  10 5  10 0  10 3  4
2  5  3  4  5  8  0  9  1  1  4  7
1  5  0  5  2  5  3  3  4  8  5  9
2  9  1  3  4  5  5  4  0  3  3  1
1  3  3  3  5  9  0  10 4  4  2  1
```

Fig. 3. Fisher and Thompson dataset ft06

5.2 Results

Sequential hybridization consists in applying several methods in such a way that the results of one method serve as initial solutions for the next.

In our case, we employed the TS to produce the population of GA because it has a worldwide knowledge base and explores the search space. Then, 10 solutions produced by the TS are translated into 10 chromosomes (population generation), which is the input of the GA before proceeding to the successive iterations (crossover, mutation, and selection).

The display of the tabu solution is in the form of sequences of operations, each sequence presents the job number $J_i (i = 0 \ldots n - 1$, n = number of jobs) , the machine number $M_j (j = 0 \ldots m - 1$, m = number of machines) and the execution time of the operation P_{ij}, accompanied by a cost (the makespan), as shown in Fig. 4.

The passage from a tabu-supplied ordering to a chromosome is shown in the Fig. 5, which presents an example of a result of the conversion into a chromosome.

This hybridization has given important results, while comparing with the GA and the TS of the literature.

```
solution1:****** Schedule: ******
Op. J0 M0 P3, Op. J3 M0 P5, Op. J2 M0 P9, Op. J5 M0 P10, Op. J1 M0 P10, Op. J4 M0 P3,
Op. J1 M1 P8, Op. J3 M1 P5, Op. J5 M1 P3, Op. J0 M1 P6, Op. J4 M1 P3, Op. J2 M1 P1,
Op. J0 M2 P1, Op. J2 M2 P5, Op. J1 M2 P5, Op. J4 M2 P9, Op. J3 M2 P5, Op. J5 M2 P1,
Op. J2 M3 P4, Op. J5 M3 P3, Op. J3 M3 P3, Op. J0 M3 P7, Op. J1 M3 P4, Op. J4 M3 P1,
Op. J1 M4 P10, Op. J4 M4 P5, Op. J2 M4 P7, Op. J3 M4 P8, Op. J5 M4 P4, Op. J0 M4 P6,
Op. J2 M5 P8, Op. J5 M5 P9, Op. J1 M5 P10, Op. J0 M5 P3, Op. J4 M5 P4, Op. J3 M5 P9,
cost: 55.0
```

Fig. 4. Example of a TS result for ft06

We tested different benchmarks, the table 1 shows our comparison. incorporated the benchmarks Abz5, La04, La10, FT06, and ORB01 into our test. In the majority of testing, VGATS-JSSP proved to be dependable, and it reduced the makespan given by the literature's research findings.

```
*chromosome1: [0, 3, 2, 5, 1, 4, 1, 3, 5, 0, 4, 2, 0,
2, 1, 4, 3, 5, 2, 5, 3, 0, 1, 4, 1, 4, 2, 3, 5, 0, 2,
5, 1, 0, 4, 3]
```

Fig. 5. The result of the TS schedule converted into a chromosome for ft06

The gantt chart in Fig. 5 for VGATS applied ft06 shows the scheduling of each operation's machine activity. The vertical line displays the machine numbers beginning with 0, while the horizontal line indicates the operation processing time unit. Each operation is represented by a different color, and the job number is shown by the operation's number on the job. The length of the bar represents the time required for that operation to finish on that machine.

Table 1. GA, TS, VGATS result for some dataset instances from (Lawrence, Adams et al., Fisher and Thompson and Applegate and Cook)

Benchmark	GA	TS	VGATS
abz5	1234	1234	963
ft06	55	55	45
la04	590	590	581
la10	958	958	958
orb01	1059	1059	1059

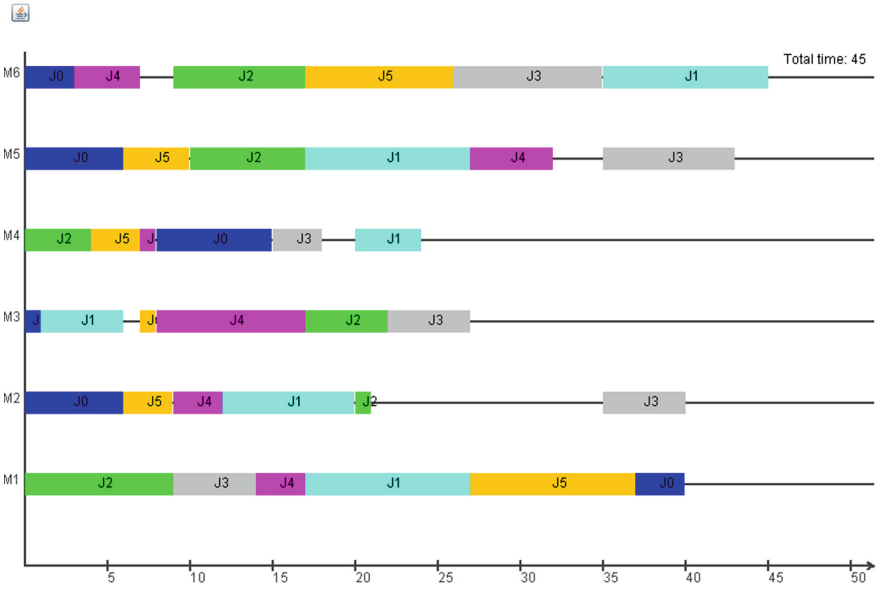


Fig. 6. Gantt chart of VGATS result for ft06

6 Conclusion

The scheduling issue for job shops is an NP problem. Various heuristic techniques are researched in the literature to tackle various iterations of the job shop scheduling problem. It is evident from the review of the various JSSP optimization strategies that the current methodologies cannot adapt to changing constraints and objectives. The job shop scheduling problem VGATS-JSSP is solved using genetic algorithms and Tabu Search, which are presented in this study. A partially workable solution is produced by the proposed one-dimensional solution representation and initialization technique. The results demonstrate quick convergence to the ideal solution.

Genetic algorithms and Tabu search are frequently used in future work to arrive at the better solutions. Combining the two meta-heuristics might result in improved performance.

References

1. Mohan, J., Lanka, K., Rao, A.N.: A review of dynamic job shop scheduling techniques. Procedia Manuf. **30**, 34–39 (2019)
2. Zhang, F., et al.: Evolving scheduling heuristicss via genetic programming with feature selection in dynamic flexible job-shop scheduling. iEEE Trans. Cybern. **51**(4), 1797–1811 (2020)

3. Bechir Alaya Alaya, B.: EE-(m,k)-Firm: a method to dynamic service level management in enterprise environment. In: Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017), vol 1, pp. 114–122 (2017). https://doi.org/10.5220/0006322401140122
4. Zhang, M., Tao, F., Nee, A.Y.C.: Digital twin enhanced dynamic job-shop scheduling. J. Manuf. Syst. **58**, 146–156 (2021)
5. Alaya, B.: EE-(m,k)-firm: operations management approach in enterprise environment. Ind. Eng. Manag. **05**(04) (2016). https://doi.org/10.4172/2169-0316.1000199.
6. Fang, Y., et al.: Digital-twin-based job shop scheduling toward smart manufacturing. IEEE Trans. Ind. Inform. **15**(12), 6425–6435 (2019)
7. Wang, L., et al.: Dynamic job-shop scheduling in smart manufacturing using deep reinforcement learning. Comput. Net. **190**, 107969 (2021)
8. Ahmadian, M.M., Salehipour, A., Cheng, T.C.E.: A meta-heuristic to solve the just-in-time job-shop scheduling problem. Eur. J. Oper. Res. **288**(1), 14–29 (2021)
9. Lin, L., Gen, M.: Hybrid evolutionary optimisation with learning for production scheduling: state-of-the-art survey on algorithms and applications. Int. J. Prod. Res. **56**(1–2), 193–223 (2018)
10. Zhang, G., et al.: An improved genetic algorithm for the flexible job shop scheduling problem with multiple time constraints. Swarm Evol. Comput. **54**, 100664 (2020)
11. Chaouch, I., Driss, O.B., Ghedira, K.: A novel dynamic assignment rule for the distributed job shop scheduling problem using a hybrid ant-based algorithm. Appl. Intell. 49(5), 1903–1924 (2019)
12. Hansen, P., et al.: Variable neighborhood search. In: Handbook of Metaheuristics, pp. 57–97. Springer, Cham (2019)
13. Abderrahim, M., Bekrar, A., Trentesaux, D., Aissani, N., Bouamrane, K.: Bi-local search based variable neighborhood search for job-shop scheduling problem with transport constraints. Optim. Lett. **16**(1), 255–280 (2020). https://doi.org/10.1007/s11590-020-01674-0
14. Tavakkoli-moghaddam, R., Azarkish, M., Sadeghnejad-Barkousaraie, A.: A new hybrid multi-objective Pareto archive PSO algorithm for a bi-objective job shop scheduling problem. Expert Syst. Appl. 38(9), 10812–10821 (2011)
15. Bożejko, Wojciech, et al. Parallel tabu search for the cyclic job shop scheduling problem. Comput. Ind. Eng. **113**, 512–524 (2017)
16. Li, X., Gao, L.: An effective hybrid genetic algorithm and tabu search for flexible job shop scheduling problem. Int. J. Prod. Econ. **174**, 93–110 (2016)
17. Du, H., Liu, D., Zhang, M.-H.: A hybrid algorithm based on particle swarm optimization and artificial immune for an assembly job shop scheduling problem. Math. Probl. Eng. (2016)
18. Zhang, G., et al.: A variable neighborhood search based genetic algorithm for flexible job shop scheduling problem. Cluster Comput. 22(5), 11561–11572 (2019)
19. Abukhader, R., Kakoore, S.: Artificial Intelligence for Vertical Farming-Controlling the Food Production (2021)
20. Zhou, B., Liao, X.: Particle filter and Levy flight-based decomposed multi-objective evolution hybridized particle swarm for flexible job shop greening scheduling with crane transportation. Appl. Soft Comput. **91**, 106217 (2020)
21. Cebi, C., Atac, E., Sahingoz, O.K.: Job shop scheduling problem and solution algorithms: a review. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), p. 1–7. IEEE (2020)

22. CUNHA, Bruno, MADUREIRA, Ana M., FONSECA, Benjamim, et al. Deep reinforcement learning as a job shop scheduling solver: A literature review. In : International Conference on Hybrid Intelligent Systems. Springer, Cham, 2018. p. 350-359
23. Semlali, S.C.B., Riffi, M.E., Chebihi, F.: Memetic chicken swarm algorithm for job shop scheduling problem. Int. J. Electr. Comput. Eng. **9**(3), 2075 (2019)
24. Kalshetty, Y.R., Adamuthe, A.C., Kumar, S.P.: Genetic algorithms with feasible operators for solving job shop scheduling problem. J. Sci. Res **64**, 310–321 (2020)
25. Gröflin, H., Klinkert, A.: A new neighborhood and tabu search for the blocking job shop. Discret. Appl. Math. 157(17), 3643–3655 (2009)
26. http://people.brunel.ac.uk/~mastjjb/jeb/info.html
27. http://people.brunel.ac.uk/~mastjjb/jeb/orlib/files/jobshop1.txt



Socio-fashion Dataset: A Fashion Attribute Data Generated Using Fashion-Related Social Images

Seema Wazarkar(✉), Bettahally N. Keshavamurthy, and Evander Darius Sequeira

National Institute of Technology Goa, Ponda, Goa, India
wazarkarseema@gmail.com

Abstract. Technological advancements helps different kinds of industries to gain maximum profit. Fashion or textile industries also trying to adopt the recent technical helplines to avoid risks and target optimal gain. In the recent years researchers have turned their focus towards fashion domain. In this paper, a dataset containing information/attribute values related to fashion is presented. This information is extracted from fashion-related images shared on social network i.e. Flickr which is a part of Fashion 10K dataset. Presented dataset contains information about 2053 fashion items related data. Along with the values for multiple attributes of fashion item/style, style labels are provided as class labels. The presented dataset is useful for fashion-related tasks like fashion analysis and forecasting and recommendation using small devices consuming less power.

Keywords: Fashion Analysis · Social Media · Style Prediction

1 Introduction

Nowadays, fashion become a part of day today life as most of the people like to use popular styles. Every fashion caries their own life cycle. Life cycle of the fashion indicates the popularity of a particular fashion at instance of time. There are three types of fashion life cycles i.e. short (fad), fashion and long (classic). Knowledge about life cycle of a particular fashion item is very useful for the business person in order to get maximum profit through managing the resources. Hence, fashion analysis plays an important role in industries of fashion and textile to accomplish fashion related different tasks like fashion trend analysis, recommendation, etc.

As rapid increase in number of users of social networks (e.g. Facebook, Twitter, etc.), huge amount of data is being uploaded daily on the social network from different locations. It contains diverse information which can be utilized to accomplish real world tasks in different fields. Initially, social data need to be analyzed and then utilized for the further use. As this data is available in huge size, it is very challenging to analyze it. Along with it, data possess characteristics like unstructured and heterogeneity. Social data contains two kinds of data, content and linkage data. Content data is existing in different forms like numeric (number of likes, tags), text (comments), images (profile picture, posts), audio, video, etc. Linkage data is about a relation between different users [1]. To accomplish real world tasks in the field of fashion, generally complex

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 350–356, 2023.
https://doi.org/10.1007/978-3-031-27409-1_31

multimedia data required to analyze. Out of above discussed data forms, image data is the most expressive and interesting data. It is very useful in the field of fashion as everyday large number of social users upload theirs photos. Those photos contain fashion related information as each person wears different styles of dresses as well as fashion accessories. Hence, it is important to analyze the social image data to extract fashion related information from it [2]. Dealing with the image data is not an easy task due to various aspects of it like size, complex in nature, needs high computational power. Hence, we have extracted information about fashion using social images in the form of numeric data which is one of the easily manageable data. Information about that fashion-related numeric dataset is presented in this paper is available at https://github.com/Seema224/Socio-Fashion-Dataset.

Organization of this paper is given as follows:

In Sect. 2, description of the presented dataset is provided. Applications of our dataset is given in Sect. 3. Then, our work is concluded in Sect. 4.

2 Data Description

The Socio-Fashion dataset has been created manually by analyzing the fashion-related images collected from social network. In this section, initially description about the data collection is provided. Then, statistics of the dataset is discussed.

2.1 Data Collection

Socio-fashion dataset has numerical data containing fashion related information. This data is obtained by analyzing fashion-related images uploaded on social network. Images to analyze are considered from Fashion 10000 dataset generated by Loni et al. [3]. Fashion 10000 dataset contains fashion and clothing related images from Flickr. It possess some irrelevant images also. Therefore, pre-processing is performed to remove those irrelevant images. Some classes in that dataset are merged to form new classes and also hierarchically arranged according to their relevance as shown in Fig. 1 and frequency of sub-categories from each class are provided in Fig. 2. Data is annotated manually with the help of students by giving knowledge of fashion attributes.

2.2 Data Statistics

In this dataset, information about total 2053 images is provided. Here, 287, 633, 918, 215 number of images from four classes i.e. bags, dresses, fashion accessories and footwear, respectively are analysed. Each category contains different number of sun-categories as mentioned in Table 1 which is provided as class labels in the proposed dataset.

Each category contains various kinds of attributes where only 3 attributes i.e. color, number of attributes and number of tags are common among all the categories. 4, 3, 1 and 3 are the category specific number of attributes for bags, dresses, fashion accessories and footwear, respectively. Details of the attributes are provided in Table 2. Here, wear_at indicates that "where to wear a given fashion accessory". In footwear category, with less and zip provides an information about whether less or zip is present or not for given

Table 1. General statistics about Socio-fashion dataset.

Category	Number of Sub-categories	Number of Images Considered
Bags	3	287
Dresses	5	633
Fashion Accessories	7	918
Footwear	7	215
Total	22	2053

style of footwear. Sole provides a value from range (0–5) which represents a size of sole i.e. 0 indicates flat footwear and 5 indicates footwear with very high hills. Closed foot provides an information about whether footwear is closed foot i.e. covering complete foot or not. For more details access metadata files (Fig. 3).

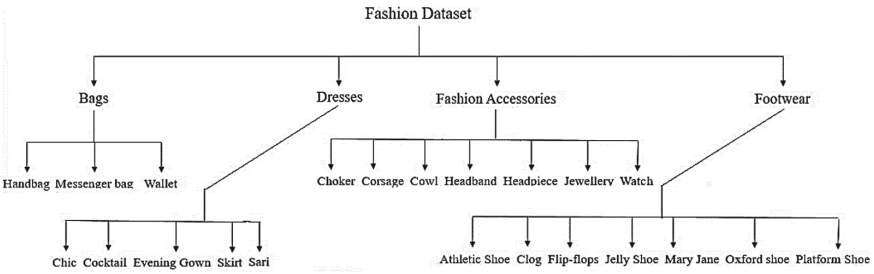


Fig. 1. The caption of the figure Hierarchical structure of classes in the dataset.

3 Data Analysis

Fashion related generated data is analyzed for style prediction using various machine learning algorithms like decision tree, random forest, Naïve Bayes classifier, linear discriminant analysis, multinomial logistic regression, decision tree regression and 3 best method approach (works based on 3 best methods mentioned earlier in the list of approaches used). Performance of these approaches are compared based on the evaluation metrics like accuracy, standard error and percentage error (Table 3).

4 Data Applications

Socio-fashion dataset is useful to test/validate the data mining algorithms used for the multimedia analysis. It includes techniques to accomplish tasks like classification, clustering, association, etc. For supervised tasks like classification which uses labelled data, dataset should be used directly as provided. But, for unsupervised tasks like clustering,

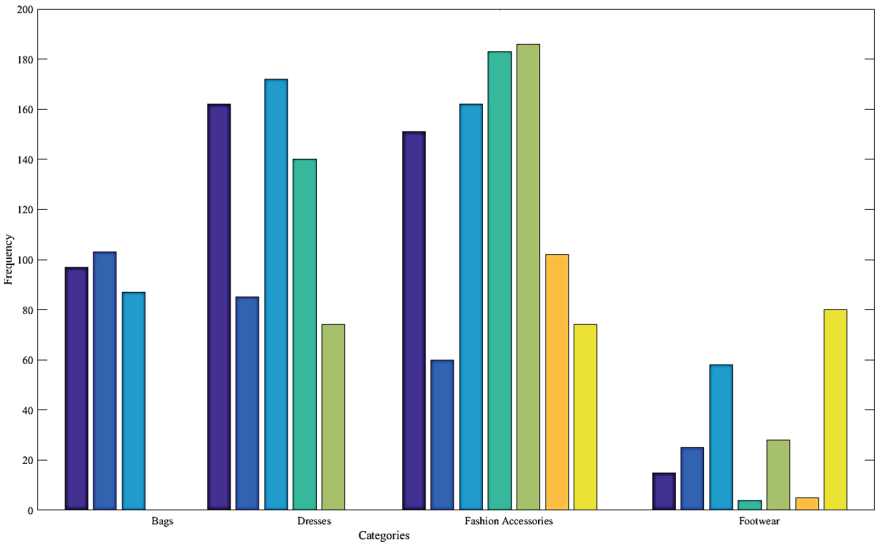


Fig. 2. Class wise frequency of each sub-category.

Table 2. Information about attributes.

Category	Category specific attributes
Bags	Fabric, Design, Gender, Shape
Dresses	Length, Neck, Design
Fashion Accessories	Wear_at
Footwear	With less, With zip, Sole, Closed foot

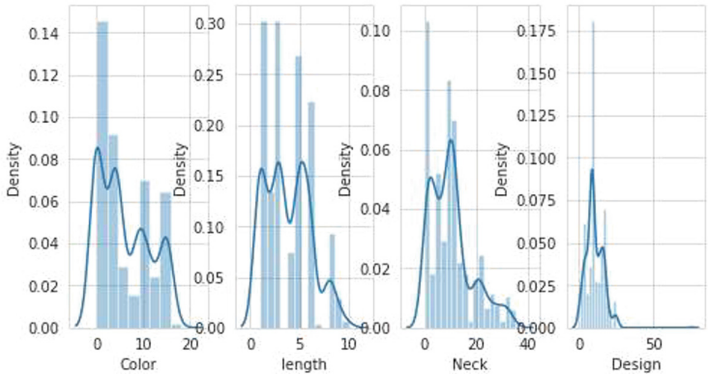


Fig. 3. Sample fashion attribute distribution visualization.

Table 3. Fashion data analysis using machine learning techniques.

Machine learning approach	Accuracy	Standard Error	Percentage Error
Decision Tree Classification	75.00	9.10	25.00
Random Forest Classification	76.02	10.30	23.98
Gaussian Naïve Bayes Classification	79.80	3.01	20.20
Linear Discriminant Analysis	67.10	9.22	32.90
Multinomial Logistic Regression	61.59	11.58	38.41
Decision Tree Regression	73.03	10.19	26.95
3-Best Methods Approach	93.79	3.04	6.21

style id's provided in dataset need to be removed as it is capable to work without class labels. As this dataset contains fashion related information, it can be also utilized for accomplishing the following fashion-related tasks:

Fashion trend analysis: Fashion trend analysis is a process of analyzing existing information about trends and its affecting factors. E.g. in fashion analysis, attributes like color, fabric, local environment, culture, etc. are the key driving elements for the change in fashion trends. Outcomes of this process can be utilized further for various important tasks like forecasting, recommendation, etc.

Fashion/style forecasting: Fashion/ style forecasting carried out to spot the upcoming trends/styles by analyzing the available data related to fashion. It is useful to take the important decisions in fashion and textile industries.

Fashion recommendation: Fashion recommendation provides a convenient way to the customers for identifying the favorite items. For fashion recommendation, fashion related data need to be analyzed using advanced machine learning techniques.

5 Background

According to the existing literature and resources, many researchers turned their focus towards fashion research after 2015 and provided fashion data in the form of images which are computationally expensive to analyze. Some of the popular fashion datasets are listed in the Table 4. These all datasets are in image and text format. These datasets are the motivation for the presented work. In Fig. 1, frequency of publications related fashion from Scopus is represented [4] (Fig. 4).

As social media is one of the live source it can be utilized to get current trends. Social data related to fashion is being considered for the present study. Fashion information mostly presented in the form of image and text but not in numeric form. Current dataset is providing fashion related information in numeric form and it will be available publically for researchers. Further it can be utilized for multi modal fashion studies by combining with existing other datasets and also for using transfer learning.

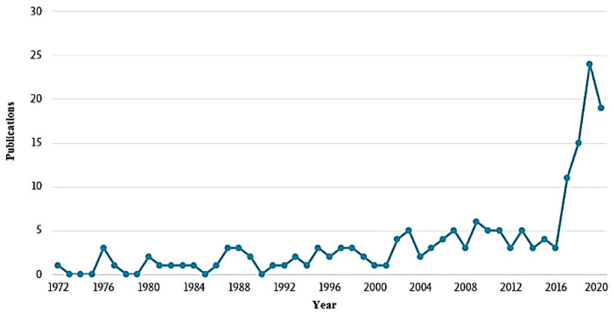


Fig. 4. Number of publications related to fashion over past many years.

Table 4. Existing fashion datasets.

Available Fashion Dataset	Year
Fashion-MNIST [5]	2017
Clothing Dataset [6]	2020
Large scale fashion (Deep-Fashion) Database [7]	2016
Fashion-Gen [8]	2018
iFashion [9]	2019
Fashionpedia [10]	2020

6 Conclusion

In this paper, numerical dataset on fashion is presented which is generated by using social fashion images. Through this dataset, we tried to present complex multimedia data in the simplest form i.e. numeric. As social network is being updated on daily basis, we have chosen social media images to extract fashion related information. This dataset is useful for the research in the field of fashion and machine learning. It can be utilized for the content data analysis, fashion forecasting, and fashion recommendation. As future work, new version of dataset will be created for men's fashion items as well updated version will consider more female fashion types with body shape information.

References

1. Aggarwal, C.: An Introduction to Social Network Data Analytics. Social Network Data Analytics, Springer, US (2011)
2. Kim, E., Fiore, A., Kim, H.: Fashion Trends: Analysis and Forecasting. Berg (2013)
3. Loni, B., Cheung, L., Riegler, M., Bozzon, A., Gottlieb, L., Larson, M.: Fashion 10000: an enriched social image dataset for fashion and clothing. In: Proceedings of the 5th ACM Multimedia Systems Conference, ACM, Singapore, pp. 41–46 (2014)
4. Scopus. Last accessed. https://www.scopus.com/.

5. Xiao, H., Rasul, K., Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms
6. Kaggle. https://www.kaggle.com/agrigorev/clothing-dataset-full
7. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Large-scale Fashion (DeepFashion) Database. The Chinese University of Hong Kong, Category and Attribute Prediction Benchmark, Xiaoou TangMultimedia Laboratory (2016)
8. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-gen: the generative fashion dataset and challenge (2018)
9. Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., ... Belongie, S.: The imaterialist fashion attribute dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
10. Jia, M., Shi, M., Sirotenko, M., Cui, Y., Cardie, C., Hariharan, B., Belongie, S.: Fashionpedia: ontology, segmentation, and an attribute localization dataset. In: European conference on computer vision, pp. 316–332. Springer, Cham (2020)

Epileptic MEG Networks Connectivity Obtained by MNE, sLORETA, cMEM and dsPM

Ichrak ElBehy[1(✉)], Abir Hadriche[1,2], Ridha Jarray[2], and Nawel Jmail[1,3]

[1] Digital Research Center of Sfax, Sfax University, Sfax, Tunisia
ichrakchouda@gmail.com, abir.hadriche.tn@ieee.org
[2] Regim Lab, ENIS, SfaxUniversity, Sfax, Tunisia
[3] Miracl Lab, Sfax University, Sfax, Tunisia

Abstract. The determination of relevant generators of excessive discharges in epilepsy is made possible by detecting electromagnetic sources within magnetoencephalography (MEG). MEG Neurologists employ indicator source localization as a diagnostic aid (presurgical investigation of epilepsy). Several ways to solving the forward and inverse source localisation challenges are discussed. The aim of this paper is to investigate four distributed inverse problem methods: estimation by minimum norm estimation MNE, standardized low resolution brain electromagnetic tomography sLORETA, Standard maximum entropy on the mean cMEM and Dynamic statistical parametric maps dSPM in defining networks connectivity of spiky epileptic events in Spatial resolution. We employed Jmail et al. Jamil et al. (Brain Topogr 29(5):752–765, 2016)'s pre-processing chain to estimate the rate of epileptic spike connection among MEG using sources spatial extend applied on pharmaco-resistant patient. We evaluated the cross correlation between extended active sources for each inverse approach. In fact, dsPM, MNE, and cMEM provide the highest amount of connectivity, all areas are connected, but they provide a low rate of correlation, while sLORETA provides the highest level of connection between all active sources. These findings demonstrate the consequences of these inverse problem approaches' basic assumptions, which entail direct cortical transmission. These results necessitate the employment of several localization approaches when analyzing interictal MEG spike location and epileptic zones.

Keywords: MNE · sLORETA · cMEM · dsPM · Cross correlation · Network connectivity · Spiky MEG events

1 Introduction

There are numerous tools to characterize brain functions or its pathologies, such as magneto- encephalography (MEG) and electro-encephalography (EEG) which are non-invasive ways used especially in neurological diseases as epilepsy. The main characteristic that may display advantages of EEG and MEG is that these techniques demand less detail about cortical tissue which aim in defining epileptic fit and sources. Correlation [1], directed transfer function [2], linear and non linear information based on correlation measures, dynamic causal modeling [3], and coherence [4], are several measures of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 357–365, 2023.
https://doi.org/10.1007/978-3-031-27409-1_32

connectivity [5, 6], used for cortical interaction between different brain regions. Many regions are implicated during the generation of paroxysmal discharges or as a propagation zone. Source localization is composed of both inverse and forward problem, when it is used to determine the responsible regions of excessive discharges called epileptogenic zones. For pharmaco resistant subject, Creating epileptic networks is a preoperative work that allows for the restriction of endless exact locations. So, examining and assessing the MEG biomarkers network connectivity (spikes or oscillatory events) [7–9, 11] is beginning with source localization then calculating connection using the forward and inverse problem. Four distributed inverse approaches will be used in this study to examine and evaluate connectivity measurements of spiky epileptic MEG events: Estimating the minimum norm MNE, Dynamic statistical parametric maps, dsPM standardized low resolution brain electromagnetic tomography sLORETA and Standard maximum entropy on the mean cMEM.

In reality, these inverse techniques (MNE, dsPM, cMEM, and sLORETA) are regarded as distributed methods [12, 16] that use same basic assumptions to generate active zones with different hypothesis. Where MNE normalizes the current density map while dsPM normalizes the using noise covariance. cMEM is featured by its capacity to recover the spatial extend of underlying sources. Our determined epileptic network connectedness of epileptic Spiky events includes a wide range of linkages between zones and rates of connection (links). As a result, it is required before epilepsy surgery to inquire the use of various inverse problem ways to improve the precision of precise generators and also how they indicate their neighbors.

In this work, we will first discuss our data base, and then we will use the preprocessing chain of Jmail et al. [7] to display connection metrics of the four inverse methods Lastly, we will demonstrate that SLORETA has the highest connectivity and correlation level but dsPM, MNE, cMEM show less correlation measures between active region of epileptic spiky events.

2 Materials and Methods

A. Materials

All analysis steps were carried out on “MATLAB” Mathwork, Natick, MA, with the aid of EEGLAB and Brainstorm toolbox (an accessible collaborative tool for analyzing brain recordings). The magnetoencephalography register of a pharmacoresistant patient from the clinical Neurophysiology Department of “La Timone” hospital was the source of our research data. inMarseille. A MEG recording with a sampling frequency of 1025 Hz demonstrates consistent and frequent intercritical activity as well as an important concordance of epileptic spikes. During MEG acquisition, no anti-epileptic medications or sleep deprivation were employed. Furthermore, our work was approved by the by the institutional Review committee of INSERM French Institution of Health. Table 1 displays the clinical information for our patient.

The patient’s MEG (MagnetoEncephalography) signal was captured on a system of 248 magnetometers, 4D Imaging, San Diego, California, located at the hospital Timone in Marseilles. The patient was supplied with head-mounted coils (for the 3-coil CTF

Table 1. Clinical information for our patient

The patient	Sex/Age	Construction MRI	Histological diagnostic	Pic MEG occurrence	Treatment at time of the record MEG	MEG: pre-op versus post op	Results surgical Class Engel (followed)
ZC	F29	Ordinary	Gliose	Abundant	phenytoin, + clobazam (20 mg/day), + Carbamazepine + Phenobarbital (50 mg/day)	Preoperative	Class 3 (5 years)

system) prior to data recording to identify the location of the head in reference to the sensor MEG. The spontaneous activity was recorded at a sampling rate of 2050 Hz using a 200 Hz anti-aliasing filter. A recording session is typically made up of 5 sessions of 3 min. The orientation of the head is recorded before and after all series by measuring the magnetic fields produced by the coils attached to the head. Series with head position changes of more than 5 mm were excluded from analysis (Table 2).

Table 2. MEG recording

Patient	Number of cluster	Total peak	Number of spikes in the selected cluster	%
ZC	3	28	12	42.85

B. Methods

Figure 1 displays the processes taken to identify the sources of epileptic spiky events, as described by Jmail et al. [7, 8]. To begin, spiky detection and selection were carried out by a stationary wavelet transform SWT filter rather than by an expert. Following that, we utilized the K-means algorithm to cluster spiky events, followed by FIR filtering to delineate low oscillatory components.

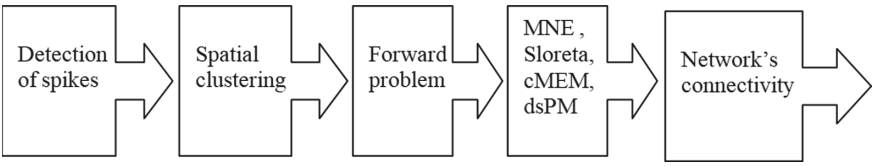


Fig. 1. Preprocessing steps of transient activity connectivity networks

The source was located by employing brainstorm to solve the direct and inverse problems. For the forward problem, we created a multiple-sphere head for the subjects, and after registering the MRI subjects, we imported cortical and scalp surfaces to brainstorm, and then we fitted a sphere for each sensor using three fiducial markers: nasion, left and right pre-auricular points [15]. For each subject, we used MNE, dsPM, cMEM and sLORETA to solve the inverse problem. Finally, a cross correlation and coherence was done to assess the connection and force between active cortical areas responsible for discharges initiation and propagation. Jmail et al. [7] provide more information on the preprocessing procedures of connection networks. We imported our spiky MEG events into Brainstorm, then utilized the four inverse problem methods: MNE, dsPM, cMEM, and sLORETA to identify active sources using the following parameters: a regularization parameter equal to the signal-to-noise ratio of three, sources limited to the cortex's normal direction, and a depth weighting of 0.5.

As a baseline activity prior to the discharges of our spiky events [7], we constructed our noise covariance matrix.

After obtaining an activation film for each inverse approach, we established visually active zones, As nodes of interest, local peaks with high amplitude called Scouts, after locating this scouts we employed 10 vertices for each scouts to have finally a spatial extend region, a thresholding process distinguishing spurious peaks.

Therefore, we had set a spatial extend for each scout that's to say all the scouts are surrounded by 10 vertices which had as result 5 regions at each hemisphere. Then, we disseminated a rotating dipole on each active region(spatial extend) obtained by our distributed techniques (MNE, dsPM, cMEM and sLORETA) in order to normalize our reconstructed time series [10]. Finally, we projected our data on dipoles for the five spatial extend zones, yielding a singletime course from which we calculated connection metrics using cross correlation [7].

C. Results of the Localization of MEG Spiky Events

Figure 2 shows active regions of our chosen spiky epileptic MEG data utilizing MNE, sLORETA, cMEM and dsPM.

MNE, sLORETA techniques yielded multiple active regions in common; additionally, dsPM and cMEM yielded significantly fewer active regions on hemispheres.

In Fig. 3, we have used our 4 distributed inverse problem technique to evaluate the rate of coupling between the active region of subject using spatial extend (scouts and vertices).

For each region of interest, we have reconstructed the time series, using the singular value decomposition following the projection on the regional dipoles. The following figures explain for the transients of the patient the time course at the level of the sources.

Then, for each region active during the discharges, we calculated a time course estimate. To assess the association between these active zones.

We calculated the cross-correlation between these time courses for each pair of signals, as show in the next section (Fig. 4).

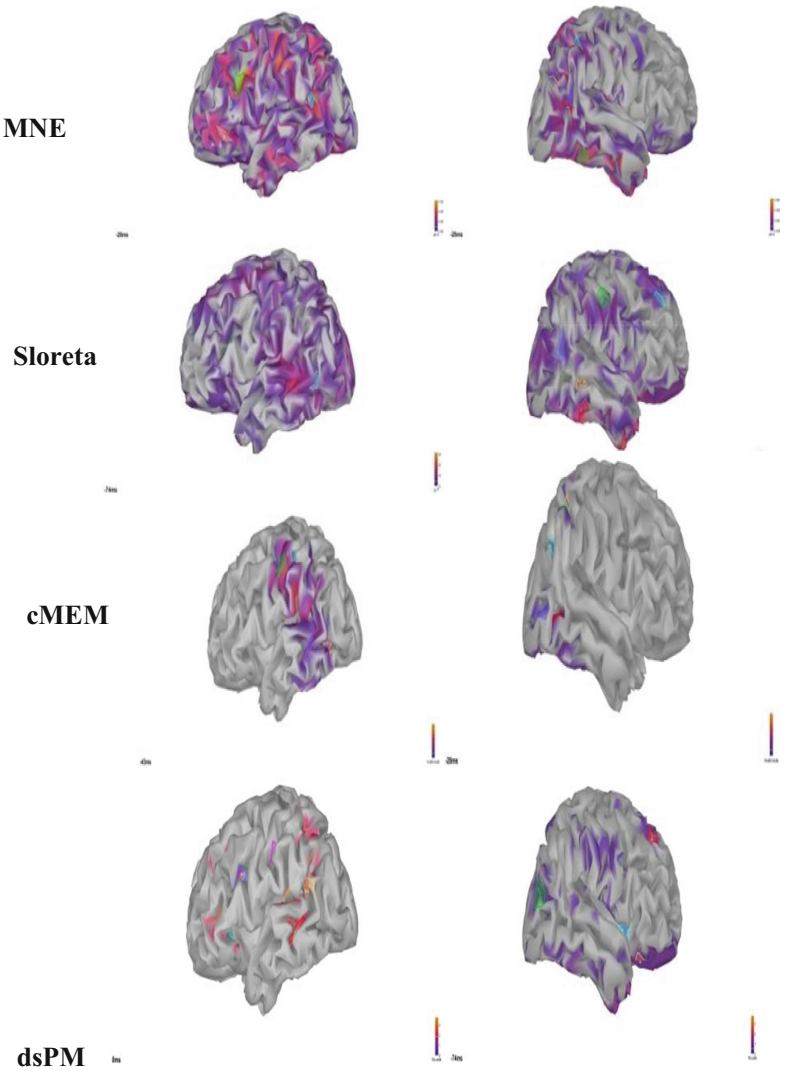


Fig. 2. Active regions using: MNE, sLORETA, CMEM and dsPM

dsPM, MNE, and cMEM provide the highest amount of connectivity, all areas are connected, but they provide a low rate of correlation, while sLORETA provides the highest level of connection between all active sources. These findings demonstrate the consequences of these inverse problem approaches' basic assumptions, which entail direct cortical transmission. These results necessitate the employment of several localization approaches when analyzing interictal MEG spike location and epileptic zones.

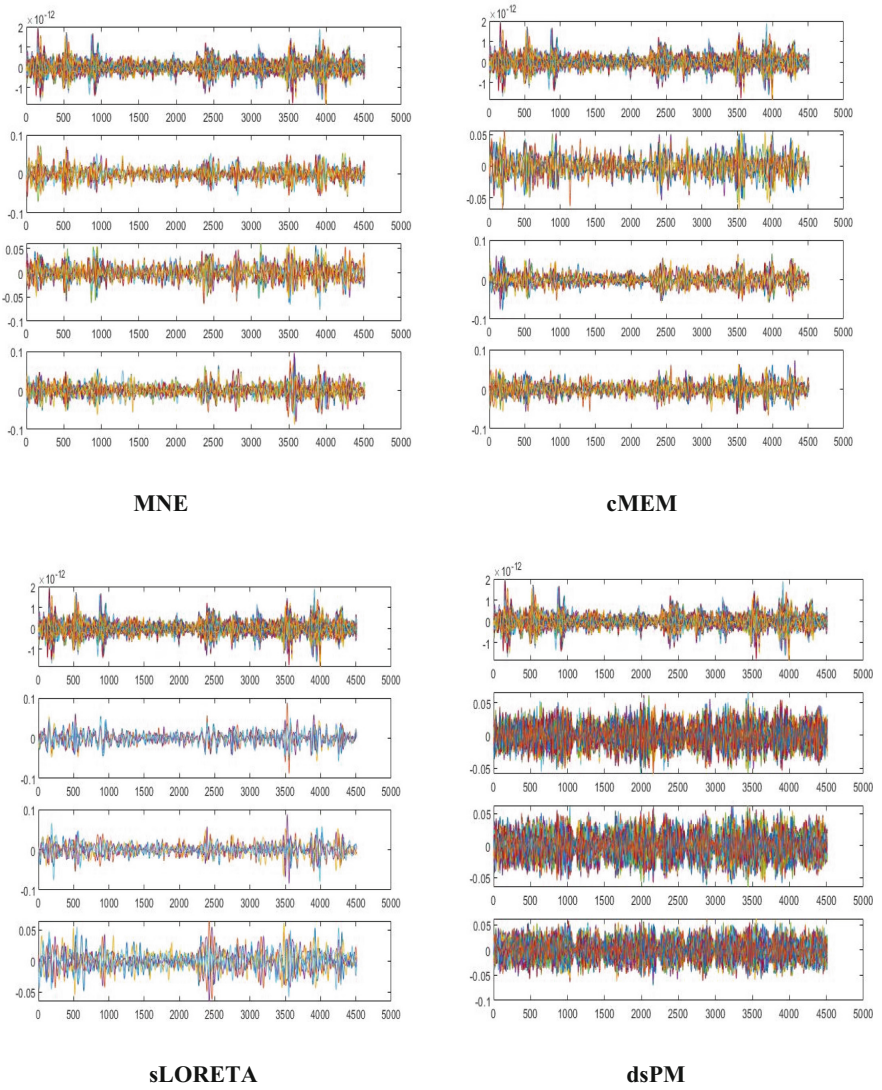


Fig. 3. MNE, SLORETA, cMEM and dsPM Scouts Time Series

3 Conclusion and Discussion

We begin in this research by establishing the network connection of spiky epileptic MEG events [7] which become extended spike (spatial extended), we emphasized four methods of resolving the inverse problem: MNE, dsPM, cMEM, and sLORETA, and we analyzed their impact on the average of coupling among cortical region responsible for excessive discharges. In fact, dsPM, MNE, and cMEM provide the highest amount of connectivity (all areas between scouts and vertices are connected), but they provide a low rate of

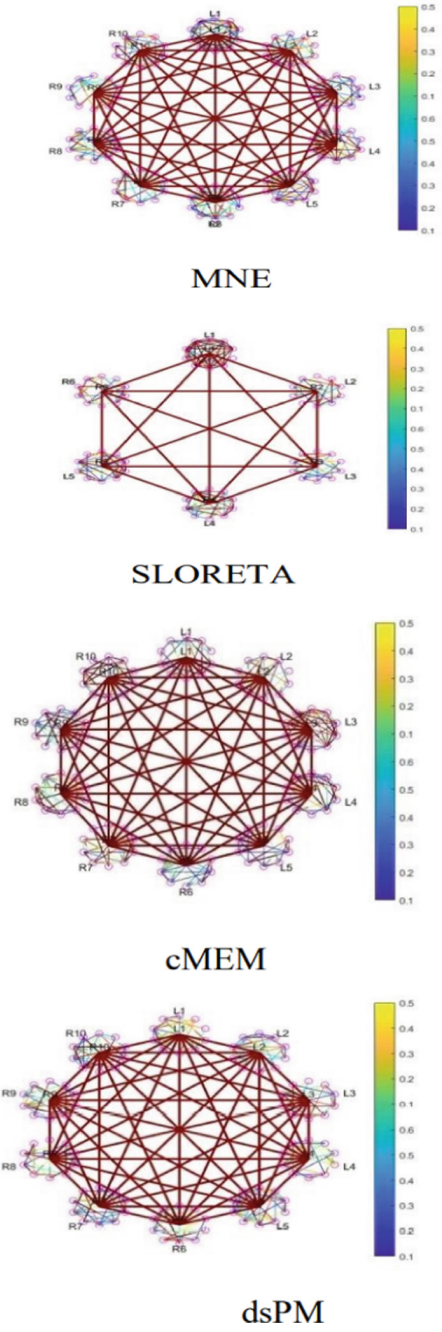


Fig. 4. Connectivity graph across regions, with a statistical threshold using: MNE, SLORETA, cMEM and dsPM, connection strength representing by Different colors

correlation, while sLORETA provides the highest level of connection between scouts and vertices and the highest rate of correlation. These findings demonstrate the consequences of these inverse problem approaches' basic assumptions, which entail direct cortical transmission. These findings necessitate the employment of several localization approaches when analyzing interictal MEG spike occurrences.

Each time sLORETA did localize with good precision, most active sources of epileptic spiky magnetoencephalography MEG events, with little or spurious activity in nearby or distant locations, sLORETA imply much more active region and much more propagation. These findings confirm the main concept of the four distributed inverse problem solutions and recommend the use of a variety of ways in handling the inverse problem during source localization and investigating accountable sources of excessive discharges. We recommend that in the future, we should evaluate these inverse problem solutions on oscillatory biomarkers in MEG and EEG to discover the best methodology that fits the best biomarkers accuracy in diagnosing epileptogenic zones and in predicting a buildup of a seizure [13, 14].

As future work, we recommend to test the four inverse problem methods on other patients to analyze and evaluate their effectiveness. Another track is to compare the results produced, by the alternative distributed approaches such as Eloreta (exact low resolution brain electromagnetic tomography), MCE (minimum current estimates), or ST-MAP (SpatioTemporal-Maximum A Posteriori). Meantime we propose to properly examine the relationship between these active zones, using other metrics such as, h2 and coherence.

Acknowledgment. This research was assisted by 20PJEC0613 "Hatem Ben Taher Tunisian Project".

References

1. Peled, A., Geva, A.B., Kremen, W.S., Blankfeld, H.M., Esfandiarfard, R., Nordahl, T.E.: Functional connectivity and working memory in schizophrenia: an EEG study. Int. J. Neurosci. **106**(1–2), 47–61 (2001). https://doi.org/10.3109/00207450109149737
2. Kaminski, M.J., Blinowska, K.J.: A new method of the description of the information flow in the brain structures. Biol. Cybern. **65**, 203–210 (1991). https://doi.org/10.1007/BF00198091
3. Friston, K.J., Harrison, L., Penny, W.: Dynamic causal modelling. Neuroimage J. **4**, 1273–1302 (2003). https://doi.org/10.1016/S1053-8119(03)00202-7
4. Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A.: Dynamic imaging of coherent sources: studying neural interactions in the human brain. Proc. Natl. Acad. Sci. **98**, 694–699 (2001). https://doi.org/10.1073/pnas.98.2.694
5. Horwitz, B.: The elusive concept of brain connectivity. J. Neuroimage **19**, 466–470 (2003). https://doi.org/10.1016/S1053-8119(03)00112-5
6. Darvas, F., Pantazis, D., Kucukaltun-Yildirim, E., Leahy, R.M.: Mapping human brain function with MEG and EEG: methods and validation. J. Neuroimage **23**(Suppl 1), S289–S299 (2004). https://doi.org/10.1016/j.neuroimage.2004.07.014
7. Jmail, N., Gavaret, M., Bartolomei, F., Chauvel, P., Badier, J.-M., Bénar, C.-G.: Comparison of brain networks during interictal oscillations and spikes on Magnetoencephalography and Intracerebral EEG. Brain Topogr. **29**(5), 752–765 (2016). https://doi.org/10.1007/s10548-016-0501-7

8. Jmail, N., Gavaret, M., Wendling, F., Badier, J.M., Bénar, C.G.: Despiking SEEG signals reveals dynamics of gamma band preictal activity. PhysiolMeas. **38**(2), N42–N56 (2017). https://doi.org/10.1088/1361-6579/38/2/N42
9. Jmail, N., Gavaret, M., Wendling, F., Badier, J.M., Bénar, C.G.: Despikifying SEEG signals using a temporal basis set. In: 15th International Intelligent Systems Design and Applications (ISDA), pp. 580–584. IEEE press, Marroc (2015). https://doi.org/10.1109/ISDA.2015.7489182
10. David, O., Garnero, L., Cosmelli, D., Varela, F.J.: Estimation of neural dynamics from MEG/EEG cortical current density maps:application to the reconstruction of large-scale cortical synchrony. IEEE Trans. Biomed. Eng. **49**, 975–987 (2002). https://doi.org/10.1109/TBME.2002.802013
11. Hadriche, A., Behy, I., Necibi, A., Kachouri, A., BenAmar, C., Jmail, N.: Assessment of effective network connectivity among MEG none contaminated epileptic transitory events.Comput. Math. Methods Med. (2021). https://doi.org/10.1155/2021/6406362
12. Jarray, R., Hadriche, A., Ben Amar, C., Jmail, N.: Comparison of inverse problem linear and non-linear methods for localization source: a combined TMS-EEG study (2021). arXiv preprint. arXiv:2112.00139. https://doi.org/10.48550/arXiv.2112.00139
13. Hadriche, A., ElBehy, I., Hajjej, A., Jmail, N.: Evaluation of techniques for predicting a build up of a seizure. In: International Conference on Intelligent Systems Design and Applications, pp. 816-827 (2021). https://doi.org/10.1007/978-3-030-96308-8_76
14. Jmail, N.: A build up of seizure prediction and detection software: a review. Ann. Clin. Med. Case Rep. **6** (14), 1–3 (2021)
15. Grova, C., Daunizea, J., Lina, J.M., Bénar, C.G., Benali, H., Gotman, J.B.: Evaluation of EEG localization methods using realistic simulations of interictal spikes. Neuroimage J. **29**,734–753 (2016)
16. Jmail, N., Hadriche, A., Behi, I., Necibi, A., Ben Amar, C.: A comparison of inverse problem methods For source localization of epileptic MEG spikes. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 867–870. https://doi.org/10.1109/BIBE.2019.00161

Human Interaction and Classification Via K-ary Tree Hashing Over Body Pose Attributes Using Sports Data

Sandeep Trivedi[1](✉), Nikhil Patel[2], Nuruzzaman Faruqui[3], and Sheikh Badar ud din Tahir[4]

[1] IEEE, Deloitte Consulting LLP Texas, Austin, USA
sandeep.trived.ieee@gmail.com
[2] University of Dubuque, Iowa, USA
[3] Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[4] Department of Software Engineering, Capital University of Science and Technology (CUST), Islamabad, Pakistan

Abstract. Human interaction has always been a critical aspect of social communication. Human action tracking and human behavior recognition are all indicators that assist in investigating human interaction and classification. Several features are considered to analyze human interaction classification in images and videos, including shape, the position of the human body parts, and their environmental effects. This paper approximated different human body key points to track their occurrence under challenging situations. Such tracking of critical body parts requires numerous features. Therefore, we first estimated human pose using key points and 2D human skeleton features to get full human body features. The extracted features are then served to t-DSNE in order to eliminate the redundant features. Finally, the optimized features are infused into the recognizer engine as a k-ary tree hashing algorithm. The experimental results have shown significant results on two benchmark datasets, including the UCF Sports Action dataset with an accuracy of 88.50% and an 89.45% mean recognition rate on the YouTube Action database. The results revealed that the proposed system had achieved better human body part tracking and classification when compared with other state-of-the-art techniques.

Keywords: Human-Human Interaction (HHI) · T-distributed Stochastic neighbor embedding (t-DSNE) · Human interaction classification (HIC) · Neural Network · K-ary tree hashing · Machine Learning

1 Introduction

Human-Human Interaction (HHI) classification requires detecting and analyzing interpersonal activities between two or multiple humans. These encounters can include commonplace actions such as conversing, passing objects, embracing, and waving. Similarly they can be supported by lifestyle actions, such as supporting a person standing up, sports

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 366–378, 2023.
https://doi.org/10.1007/978-3-031-27409-1_33

data, assisting another individual with walking, or attracting the attention of another individual. In addition, experts in this discipline are interested in suspicious behaviors such as touching a person's pocket, pushing someone, or fighting. Human interaction classification (HIC) has become a significant issue in the world of artificial intelligence due to its vast array of applications, which include sports, security, content-based video retrieval, medicine, and monitoring [1–3]. However, significant developments have been achieved in such areas, and numerous accurate human-to-human interaction systems have been created for a variety of applications; monitoring human interactions is still tricky for several reasons, including diverse views, clothing changes, poor lighting, distinct interactions with similar human movement, and the lack of vast and complex datasets [4, 5].While low-cost depth monitoring sensors, such as Microsoft Kinect, are nowadays extensively employed since they are less vulnerable to lighting environments than RGB cameras. In addition, many interactions sound related and are frequently misclassified. For instance, two people sharing a small object may resemble two individuals shaking hands. In contrast, the similar interaction becomes distinct when examined from multiple perspectives. Consequently, it is crucial to identify specific elements from photos that may easily distinguish among various actions that appear identical [6].

Unlike motion recognition, activity localization addresses the difficulty of determining the precise space-time region where an activity occurs. Compared to motion recognition, it presents a more extensive variety of issues, such as coping with background interference or the architectural diversity of the image, and has become the subject of many research articles [7]. Current effective activity localizes strategies and aims to split the movement using indications predicated on the action's appearance, motion, or a mix of the two [8]. UCF Sports is another dataset of the latest collections for action classification with actual actions in an uncontrolled setting [9]. The main problems of these systems are low accuracy rates, luminance effects, human silhouette extraction issues, and over-complex datasets. Various methods are based on traditional feature sets such as optical flow, distance, and histogram of the gradient. These systems have high error rates due to entire body features extraction techniques.

This scholarly study provides a novel methodology for the HIC system to deal with the research gap and effective video-based human interaction classification employing machine learning algorithms in a sports data environment. Human contours and frame conversion are derived from RGB sports data. Through a connected components-based technique, supplementary foreground recognition is applied. The next step is to find the human body's key points. We extract 15 key points of the human body, namely, the head, neck, right shoulder, right elbow, right hand, right hip, knee, and foot points. Similarly, the left side, we detect the left shoulder, left elbow, left hand, left hip, knee, and foot points. The 2D skeleton model is applied over these detected points. The bag of features extraction method is adopted as the next part in which we deal with crucial points and entire human body features. To deal with computational complexity, we applied a data optimization approach using the t-Distributed stochastic Neighbor embedding approach. Finally, for classification, we applied the K-ary tree hashing method. The primary contributions of this study include:

- Silhouette identification from RGB sports videos using a predicated approach to connected sections

- Human body key points extraction, 15 points are extracted.
- 2D skeleton and body key points features and full body features are extracted.
- Data optimization and mining of the repeated data through the t-Distributed stochastic Neighbor embedding approach and K-ary tree hashing is adopted as the classification method.

Segment II of the article details related research efforts, whereas Part III discusses the suggested system architecture. The development information and outcomes of the suggested method are presented in Sect. 3. Section 4 discusses several elements of the constructed system, while Section 5 offers the paper's conclusion and authors' suggestions for future research.

2 Related Work

Currently contributing to the creation of effective HHI systems are researchers. Previous methods are separated into several classes: marker-based and video-based. In the marker-based HHI framework, sensors such as reflecting spheres, light-emitting diodes, and thermal indicators are attached to the body of persons whose motions are being watched. These technologies are frequently applied in treatment [10]. For instance, [11] proposes a marker-based activity monitoring technology to assess the motions of several body components. The researchers contend that effective monitoring of the action of various body components can lead to improved medicinal recommendations.

The researchers of [12] added an IR monitor and an infrared transmitter to a remote hand skateboarding system for standard upper arm training. Eight individuals with inadequate upper arm motion were trained using the suggested apparatus. All individuals could operate the hand skateboarding across the assigned figure-of-eight pattern throughout four training sessions. However, the method was only examined on a limited sample of 10 actual patients. In video-based HIR approaches, personal interactions are captured using video cameras. In similar approaches, the primary procedure retrieves significant relevant characteristics or locations [13]. Based on these distinguishing characteristics, the activity executed in the movie is identified. Khan et al. [14] suggested an adaptive part-based simulation methodology to recognize and track human body components over successive frames. Their technology then monitored newborns' movements to discover a variety of movement abnormalities. They gathered the information utilizing Microsoft Kinect in a regional hospital, although it was simply RGB data. Khan et al. [15] presented a system for measuring the body kinematics of an individual undergoing Vojta therapy. They suggested using color characteristics and pixel placements to partition the human body in RGB video.

Then, researchers classified the right movements by applying multiclass SVM and a heterogeneous feature vector. Applying a graph analyzing neural network, Qi et al. [16] discovered and identified human–object connections in photos and videos. Their GPNN algorithm inferred, for a current scene, a parse graph consisting of the sports data networking structure defined by an adjacency matrix and the component labels. The suggested GPNN calculated the adjacency vectors and branch identifiers iteratively within architecture for message transmission inference. Liu et al. [17] adopted the few-shot learning (FSL) technique for HHI, which entails employing a small number of

examples to complete the job. However, this is challenging, and typical FSL approaches execute poorly in complicated sports scenarios.

Jaing et al. [18] Then, a late median fusion technique is applied to identify variable events. Liu et al. [19] introduced a unique hierarchical segmentation multifunctional learning methodology for the recognition and segmentation of human events simultaneously. They used the combining data approach with clustered simultaneous learning and a variable modeling strategy to enhance the features of human body joints. Abbasnejad et al. [20] created a novel approach that concurrently extracted spatiotemporal and contextual elements from video data. Then, a max-margin classification is trained, flexibly applying these features to identify activities with unknown starting and ending positions. Seemanthini et al. [21] devised a methodology for activity classification based on a convolution neural network and Fuzzy C-mean (FCM) for localization. Local Binary Pattern (LBP) and Hierarchical Centroid (HC) are utilized for feature extraction. Meng et al. [22] presented a novel methodology for feature vector presentation that collects stationary and kinetic mobility features to characterize the input event. In addition, various predictors were utilized to examine the behavior of events using a Support Vector Machine (SVM) and a Fisher vector.

3 Design and Method

In this section, we discuss our proposed method in detail; initially, video-based sports data is considered as input to the system, and preprocessing is performed to minimize the cost of the system. Human detection, body point detection, and machine learning-based feature extraction are performed; the next step is data optimization through t-DSNE and classification through the K-ary tree hashing algorithm. Figure 1 shows the detailed procedure of our proposed method.

3.1 Preprocessing

In this Subsection, we performed basic preprocessing steps to minimize the time cost and obtain more accurate results; Primarily we extract frame sequence from a given input. After that, we resize the images in format to avoid the extra computational costs. The objective of image normalization is to convert the data points of a vision to a routine basis to ensure that the image is formed more naturally in the human eye. A bilateral filter (BLF) enhances image resolution and eliminates disturbance. A BLF smooths the images while preserving the outlines of all the elements. Additionally, the bilateral filter modifies the actual image's luminosity pixels with a luminosity value derived from the adjacent pixels. Furthermore, the frequency kernel diminishes disparities in brightness, whereas the locational Gaussian diminishes disparities in dimensions. After implementing the bilateral filter, the generated allowed *Ifil* can be specified by utilizing.

$$Ig^{f} = 1/W_{pi} \sum_{x_i \in \partial} I(x_i f_r(||Ix_i - Ix||) g_s(||x_i - x||)) \tag{1}$$

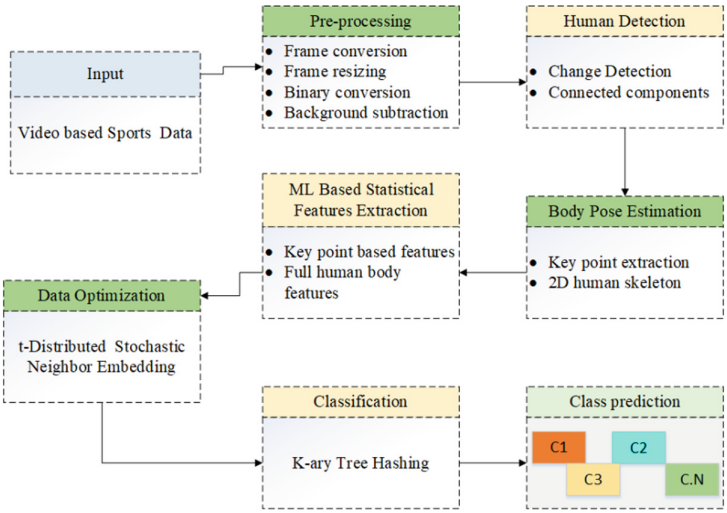


Fig. 1. The detailed overview of the proposed Human interaction classification procedure and data flow.

3.2 Human Detection

After preprocessing step, we need to extract a human silhouette from the given pre-processed data. In this research, we have various algorithms to perform this step. We utilized change detection and connected components-based approaches to achieve this. After extracting the human silhouette, we apply a bounding box to the human shape to identify that the outline is based on the human body and human body area. Figure 2 shows the results of human detection, subtraction.



(a) (b) (c)

Fig. 2. The example results of **a** background subtraction, **b** extracted human silhouette in binary format, **c** human detection with bounding box.

3.3 Body Pose Estimation

Following the extraction of the full-body contour, 12 essential body areas were chosen using a technique similar to that proposed by Dargazany et al. [23] in Algorithm 1.

Initially, the segmented outline was turned into a binary shape, and its outline was determined. Afterward, a geometric property was drawn around the body. Obtaining locations on the contour that was a component of the initial condition. Exactly five of these locations have been selected. Furthermore, an intermediate location was acquired by locating the contour's median. Through using six points achieved, six different essential points have been identified. Finding these extra points is straightforward: the median of any two main issues is determined, and the position on the shape nearest to the determined center is saved as an added main point. After that, we connect the key points to find the 2D skeleton. We link the head point to the neck, the neck to the right, the left shoulders, the shoulders to the elbow, and the hands. The neck is also associated with the mid, the mid is connected right/left hip, and the hip is connected with the right/left knee and feet. Figure 3 shows the overview of extracted key points and the human 2D skeleton model.

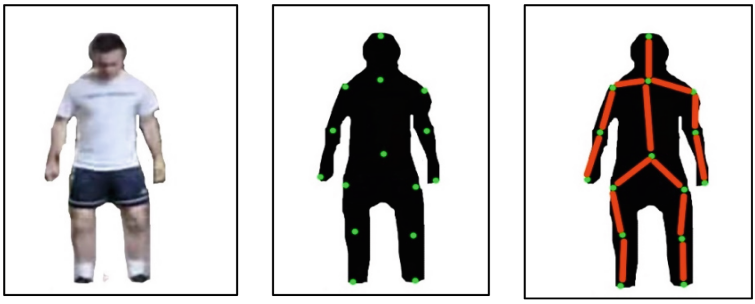


Fig. 3. The example results of **a** background subtraction, **b** extracted human body points in binary format, **c** human 2D skeleton.

Algorithm 1: Human body points extraction method

Input: Human silhouettes
Output: Human_BodyPoints (p1, p2, p3 ... p15)
contour← Getcontour(silhouette)
Convex hull← DrawPointl(silhouette)
%identifying first five human body points%
for point on convexhull:
if point in silhouette:
HumanBodyPoints.append(point)
%identifying 6th key point%
Center← GetContourcenter(silhouette)
Human_BodyPoints.append(Midpoint)
%identifying additional 9 points%
Nec ← Find15pixel(Head])
RE ← FindMidpoint(Rh_NC])
LE ← FindMidpoint(Lh_NC])
RH ← FindMidpoint(RH_RF])
RH ← FindMidpoint(RH_LF])
LK ← FindMidpoint((RF_RH])
RK ← FindMidpoint(LF_LH])
LS ← FindMidpoint(LE_NC])
RS ← FindMidpoint(RE_NC])
KeyBodyPoints.append(15 key points)
return KeyBodyPoints (p1, p2, p3 ... p15)

3.4 Machine Learning-Based Features

After completing the 2D stick model and human body points detection, we extract the machine learning-based features. There are two types of features: body points-based and full body features.

The Full body features: ORB

We extract full body-based features, in which the orientation of FAST and revolving BRIEF (ORB) is a quick and efficient analyzer of features. It detects critical points using the FAST (Features from Accelerated Segment Test) body point detection. In addition, it is a specialized version of the visual identifier BRIEF (Binary Robust Independent Elementary Features). ORB is resistant to scalability and translation. Sections of patching can be characterized by

$$M(p, q) = \sum w^p w^q Im(x, y) \tag{2}$$

Correspondingly, the visual components' intensities at the x and y coordinates are represented by p and q. These intervals allow for the identification of the centroid

$$C = \frac{mi}{mj}, \frac{mk}{mp} \tag{3}$$

The update's position is determined by

$$\alpha = an(mk.mi) \tag{4}$$

Figure 4 illustrates the outcomes of implementing an ORB feature to the obtained human figures.



(a) (b) (c)

Fig. 4. The example results of **a** background subtraction, **b** initial results of ORB features, **c** Final results of ORB features.

Human body points-based features: Distance features

In human body points features we target the main area of body joints and points. We find the distance of all the points and map them in a features vector. For this we consider head point is the starting region, we find the distance from head to neck ($Hp \rightarrow Nc$) neck to right shoulder ($Nc \rightarrow RS$) neck to left shoulder ($Nc \rightarrow LS$) right shoulder to right elbow ($Rs \rightarrow Re$), right elbow to right hand ($Re \rightarrow Rh$), left shoulder to left elbow ($LS \rightarrow Le$), left elbow to left hand ($Le \rightarrow Lh$), neck to mid ($Nc \rightarrow mid$), mid to right hip ($mid \rightarrow Rhi$), right hip to right knew ($Rhi \rightarrow Rk$), right knee to right foot ($Rk \rightarrow Rf$), mid to left hip ($mid \rightarrow Lhi$), left hip to left knee ($Lhi \rightarrow Lk$), left knee to left foot ($Lk \rightarrow Lf$).

$$Dis1 = \sum(d(Nc \rightarrow RS), d(Hp \rightarrow Nc), (Nc \rightarrow LS), d(Rs \rightarrow Re)) \tag{5}$$

$$Dis2 = \sum(d(Re \rightarrow Rh), d(LS \rightarrow Le), d(Le \rightarrow Lh)) \tag{6}$$

$$Dis3 = \sum(d(Nc \rightarrow mid), d(mid \rightarrow Rhi), d(Rhi \rightarrow Rk)) \tag{7}$$

$$Dis4 = \sum(d(Rk \rightarrow Rf), d(mid \rightarrow Lhi), d(Lhi \rightarrow Lk), d(Lk \rightarrow Lf)) \tag{8}$$

$$Distance_fe = \sum Dis1, Dis2, Dis3, Dis4 \tag{9}$$

Figure 5 shows the results and understanding of distance features.

3.5 Data Optimization and Classification

T-DSNE. After integrating all the features in the relevant field, it is necessary to apply particular tasks and approaches to the optimized number of features. We use a t-SNE-based data refinement technique to achieve this; this leads to an optimum collection of

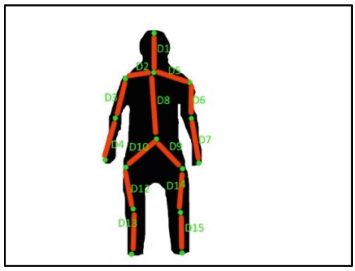


Fig. 5. The example procedure and layout of distance features we present 15 body points area.

data. Furthermore, we employed this array in our batch operations, which included estimation and classification. There are two primary methods to accomplishing this: Components held static consequently, the object changes and eliminates redundant material or converts the original functionality into a smaller subset of modified attributes with almost the same flexibility as the previous form. The t-distributed Stochastic Neighbor Embedding (t-SNE) technology of Maaten and Hinton [24] is used throughout the article. It is a relatively non process that separates and converts all classes with changing traits into optimal extra sections. As suggested by its name, this technique is based on statistical placement and is specifically designed to preserve the diversity of adjacent objects. The intensity of neighboring spots, also described as complexity, was adjusted to h, while the frequency response was adjusted to t. t-SNE is an effective technique for preserving the model's regional and global representation. While extracting the features using t-SNE, the assessed low-dimensional map has the same clustering approach as the current high-dimensional dataset. A Gaussian distribution must be generated throughout high-dimensional parameter combinations for the t-SNE method to be effective. There is a possibility that identical items are present. However, they are unlikely to be in the same position.

K-ary Hashing Algorithm. The K-ary tree hashing algorithm utilized the operations of the deeply embedded graph that specify the location with the most significant number of K successors. In addition, the most straightforward hashing methodology, considered the pre-step of K-ary tree hashing, was utilized for the recognition and classification method. This strategy is based on the dependent feature that employs the resemblance test over subgroups *Bi* and *Bj*.

$$k_b(j) = mod\left(\left(L_{dj} + M_d\right), N_d\right) \tag{10}$$

Nd, Ldj, and Md are the available randomized values extracted from the collection. The K-ary tree hashing methodology employs two ways to determine the optimized solution: a naive method for estimating the periodicity of adjacent routers and MinHashing for defining the quantity of any parameter. The naive technique is outlined in Algorithm 2.

Algorithm 2: Naïve Method for K-ary Tree Hashing

Require: L, K_i
Ensure: T(v)
% N is neighbor, L is Data, and T is size preservative approach%
1. $Tmp \leftarrow$sort ($L(K_i)$)
2. $r \leftarrow$min(r, $|K_i|$)
3. $t(i) \leftarrow [i,$ index($Tmp(1:r)$)]

4 Experimental Settings and Evaluation

This section provides extensive experimental detail of our proposed HIC system. Human interaction classification accuracy over key points detection was utilized to evaluate the performance of the proposed HIC model via two publicly accessible benchmark datasets, including UCF Sports Action and YouTube datasets. Additionally, we assessed the effectiveness of our system by determining the distance from the ground truth using optic flow, transportable body, and 180° intensity levels.

The UCF Sports Action database [25] comprises ten sports action classes, including *walking, diving, running, kicking, lifting, riding horse, swing-side, swing-bench, golf swing, and skateboarding*. While, the UCF YouTube action dataset [25] involves 11 action classes such as, *biking/cycling, walking with a dog, diving, horse back riding, volleyball spiking, basketball shooting, golf swinging, soccer juggling, trampoline jumping, swinging,* and *tennis swinging*. Figure 6 shows the sample of UCF sports action and the YouTube dataset.



Fig. 6. Sample images of UCF Sports Action dataset and YouTube dataset

Figure 7 exhibits the confusion matrix for the UCF Sports dataset for ten sports activity classes with an 88.50% recognition rate. Figure 8 shows the confusion matrix of the YouTube action dataset, attaining 89.45% accuracy over 11 sports activities.

Table 1 displays the evaluation results of the HIC proposed system compared with other state-of-the-art methods.

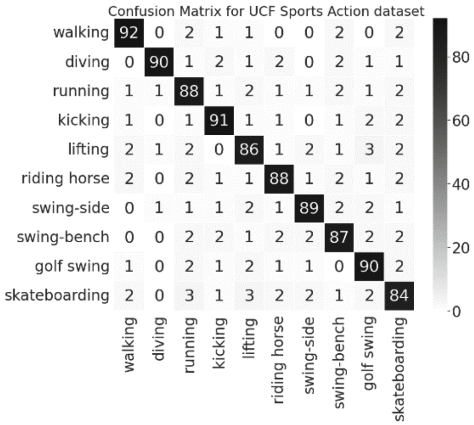


Fig. 7. Confusion Matrix of 10 sports activities on UCF Sports Action dataset

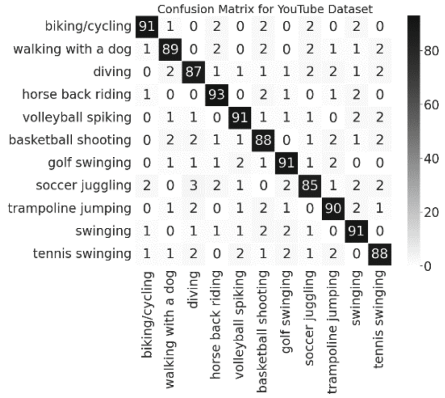


Fig. 8. Confusion Matrix of 11 action activities on UCF YouTube Action dataset

Table 1. HIC System Comparison with other State-of-the-Art Methods

Methods	UCF Sports Action Dataset (%)	Methods	UCF YouTube Action Dataset (%)
Multiple CNN [26]	78.46	PageRank [27]	71.2
Local trinary Patterns [28]	79.2	Dense trajectories [29]	84.2
Dense trajectories [29]	88	Kernelized Multiview Projection [30]	87.6
Proposed HIC	**88.50**	**Proposed HIC**	**89.45**

5 Conclusion

This paper introduced a robust 2D skeleton and key point features approach for tracking human body parts in gait event tracking and sports over action-based datasets. For feature minimization and optimization, we adopted the t-DSNE technique in order to select relevant features. Furthermore, a graph-based K-ary tree hashing algorithm is applied for sports and gate event tracking and classification. The experimental evaluation presented in our study demonstrates that our HIC-proposed system achieved a better recognition rate when compared with other state-of-the-art methods. Furthermore, this model significantly enhances human action tracking (including static and dynamic activities). In the future, we will deal with more complex interactions in indoor and outdoor settings. Furthermore, we will also focus on human-object interaction tracking in smart homes and healthcare.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. (2010). https://doi.org/10.1109/TPAMI.2008.284
2. Gholami, S., Noori, M.: You don't need labeled data for open-book question answering. Appl. Sci. **12**(1), 111 (2021)
3. Tahir, S.B.U.D., et al.: Stochastic recognition of human physical activities via augmented feature descriptors and random forest model. Sensors **22**(17), 6632 (2022)
4. Ghadi, Y.Y., Akhter, I., Aljuaid, H., Gochoo, M., Alsuhibany, S.A., Jalal, A., Park, J.: Extrinsic behavior prediction of pedestrians via maximum entropy Markov model and graph-based features mining. Appl. Sci. **12** (2022). https://doi.org/10.3390/app12125985
5. Bhargavi, D., Gholami, S., Pelaez Coyotl, E.: Jersey number detection using synthetic data in a low-data regime. Front. Artif. Intell. **221** (2022)
6. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: a review. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
7. Liu, M., Liu, H., Sun, Q., Zhang, T., Ding, R.: Salient pairwise spatio-temporal interest points for real-time activity recognition. CAAI Trans. Intell. Technol. (2016). https://doi.org/10.1016/j.trit.2016.03.001
8. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2010). https://doi.org/10.1007/978-3-642-15552-9_29
9. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. (2013). https://doi.org/10.1007/s00138-012-0450-4
10. Rado, D., Sankaran, A., Plasek, J., Nuckley, D., Keefe, D.F.: A real-time physical therapy visualization strategy to improve unsupervised patient rehabilitation. In: IEEE Visualization (2009)
11. Khan, M.H., Zöller, M., Farid, M.S., Grzegorzek, M.: Marker-based movement analysis of human body parts in therapeutic procedure. Sensors (Switzerland). (2020). https://doi.org/10.3390/s20113312
12. Chen, C.-C., Liu, C.-Y., Ciou, S.-H., Chen, S.-C., Chen, Y.-L.: Digitized hand skateboard based on IR-camera for upper limb rehabilitation. J. Med. Syst. **41**, 1–7 (2017)

13. Tian, Y., Cao, L., Liu, Z., Zhang, Z.: Hierarchical filtered motion for action recognition in crowded videos. IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev) **42**, 313–323 (2011)
14. Khan, M.H., Schneider, M., Farid, M.S., Grzegorzek, M.: Detection of infantile movement disorders in video data using deformable part-based model. Sensors **18**, 3202 (2018)
15. Khan, M.H., Helsper, J., Farid, M.S., Grzegorzek, M.: A computer vision-based system for monitoring Vojta therapy. Int. J. Med. Inform. **113**, 85–95 (2018)
16. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.-C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–417 (2018)
17. Liu, X., Ji, Z., Pang, Y., Han, J., Li, X.: Dgig-net: dynamic graph-in-graph networks for few-shot human-object interaction. IEEE Trans. Cybern (2021)
18. Jiang, Y.G., Dai, Q., Mei, T., Rui, Y., Chang, S.F.: Super fast event recognition in internet videos. IEEE Trans. Multimed. (2015). https://doi.org/10.1109/TMM.2015.2436813
19. Liu, A.-A., Su, Y.-T., Nie, W.-Z., Kankanhalli, M.: Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 102–114 (2016)
20. Abbasnejad, I., Sridharan, S., Denman, S., Fookes, C., Lucey, S.: Complex event detection using joint max margin and semantic features. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8 (2016)
21. Seemanthini, K., Manjunath, S.S., Srinivasa, G., Kiran, B., Sowmyasree, P.: A cognitive semantic-based approach for human event detection in videos. In: Smart Trends in Computing and Communications, pp. 243–253. Springer (2020)
22. Meng, Q., Zhu, H., Zhang, W., Piao, X., Zhang, A.: Action recognition using form and motion modalities. ACM Trans. Multimed. Comput. Commun. Appl. **16**, 1–16 (2020)
23. Dargazany, A., Nicolescu, M.: Human body parts tracking using torso tracking: applications to activity recognition. In: 2012 Ninth International Conference on Information Technology-New Generations, pp. 646–651 (2012)
24. der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9** (2008)
25. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. In: Computer Vision in Sports, pp. 181–208. Springer (2014)
26. de Oliveira Silva, V., de Barros Vidal, F., Soares Romariz, A.R.: Human action recognition based on a two-stream convolutional network classifier. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 774–778 (2017). https://doi.org/10.1109/ICMLA.2017.00-64
27. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild." In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003 (2009)
28. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 492–497 (2009). https://doi.org/10.1109/ICCV.2009.5459201
29. Wang, H., Kläser, A., Schmid, C., Liu, C.-L.: Action recognition by dense trajectories. In: CVPR 2011, pp. 3169–3176 (2011)
30. Shao, L., Liu, L., Yu, M.: Kernelized multiview projection for robust action recognition. Int. J. Comput. Vis. **118**, 115–129 (2016)

Bi-objective Grouping and Tabu Search

M. Beatriz Bernábe Loranca[1(✉)], M. Marleni Reyes[2], Carmen Cerón Garnica[3], and Alberto Carrillo Canán[3]

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla México, Puebla, México
beatriz.bernabe@gmail.com

[2] Escuela de Artes Plásticas y Audiovisuales, Benemérita Universidad Autónoma de Puebla México, Puebla, México
marleni.reyes@icloud.com

[3] Facultad de Filosofía y Letras, Benemérita Universidad Autónoma de Puebla México, Puebla, México
carmen.ceron@correo.buap.mx, acarrillo_mx@icloud.com

Abstract. When dealing with small sizes in zone design, the problem can have a polynomial cost solution by exact methods. Otherwise, the combinatorial nature of this problem is unavoidable and entails an increasing computational complexity that makes necessary the use of metaheuristics. Specifically, when using partitioned grouping as a tool to solve a territorial design problem, geometric compactness is indirectly satisfied, which is one of the compulsory restrictions in territorial design when optimizing for a single objective. However, the inclusion of additional cost functions such as homogeneity imply a greater difficulty since the objective function becomes multi-objective. In this case, partitioning is used to build compact groups over a territory and the partitions are adjusted to satisfy both, compactness and homogeneity, or balance the number of objects for each group. The work presented here gives answers to territorial design problems where the problem is presented as bi-objective and aims at striking a compromise between geometric compactness and homogeneity and the cardinality of the groups. The approximation method is Tabu Search.

Keywords: Clustering · Compactness · Homogeneity · Tabu Search

1 Introduction

Spatial data is important in Territorial Design (TD) problems or Zones to answer several issues of geographical nature. Zone design arises when small basic areas or geographical units must be merged into zones that are acceptable according to the requirements imposed by the problem under study. At this point, the merging is proposed as geographical and it is an implicit task in Territorial Design (TD). The problems of this kind have as fundamental principle the creation of groups of zones that are spatially compact, contiguous and/or connected.

The most common applications of TD include political districting, commerce, census, sampling classification (as the case presented here), etc.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 379–390, 2023.
https://doi.org/10.1007/978-3-031-27409-1_34

To incorporate TD to a grouping method, where data has a clear geographical component, it is inevitable to review the updated and classic literature on clustering methods. In this work, after the appropriate readings, classification by partitioning has been preferred just by its practical and satisfactory implicit result: the indirect developing of compact zones. Compact partitioning solves the problem of creating "polygonal" by using the Euclidian measure to expedite the compactness of each group's shape (distance minimization).

This type of problems, where the homogeneity in the cardinality of the groups is significant, is present in lots of logistic applications for the fair burden on vendors, the same number of voters in electoral problems, p-median in routing, etc., where also the compactness is essential.

This work centers on the above-mentioned homogeneity along with geometric compactness. The objectives of the presented model that solves the territorial partitioning are two: (1) homogeneity in the cardinality of the groups and (2) geometric compactness. Tabu Search (TS) is used to approximate the two conflicting functions.

The article is organized as follows: Sect. 1 as introduction. In Sect. 2 preliminaries and theoretical aspects are presented. Section 3 includes the description of the problem and its mathematical model. Section 4 shows the design of an algorithm with Tabu Search that we have called Partitioning with Penalized Tabu Search (PPTS). Lastly, the computing experience and conclusions are presented.

2 The Problem

Geographical partitioning is a useful tool in TD where spatial restrictions (adjacency, compactness) are demanded [15, 16]. The geographic clustering needs a combinatorial nature solution (NP-hard) such meta-heuristics one of the generated by non-exact optimization methods [2, 13]. Demographic criterion has been a restriction for demographic criterion [2, 7].

The geometrical irregularity restricts the use of adjacency methods and Delaunay triangulation and is a problem in the handling of several maps [7]. On real problems a single objective optimization is insufficient. Finding the optimal solutions for homogeneity in the number of objects and geometric compactness is a challenge. Bernábe [4], presented two approaches to achieve the homogeneity in the cardinality of the clusters with satisfactory results but implies the increase above 100 clusters.

On the other hand, using clustering and Tabu search, problems that need to find the optimum configuration of the crew and better routes to follow have been solved in order to satisfy some logistic actions in programming of medical attention as well as delivery of medical services from the hospital to patients, in a transport network. In this point, our proposal could be combined with the problem that the authors of medical logistic present [5].

Fog commuting has arisen as a new infrastructure of three layers: node levels, cloud services and businesses (clients). The levels of nodes give services to the layers of cloud and fog computing and also serve for the processes "in situ" in enterprises. Thus, the purpose of nodes layers is to give economic and high responsiveness services. Consequently, the layers of clouds are reserved for expensive processes and it is necessary to

solve the balance of optimum load between cloud and fog nodes. In the work, the author has attended the efficiency in the use of memory resources in these layers with simple Tabu search for an equilibrium of optimum load between cloud and fog nodes [6].

The contribution in this work is a multi-objective optimization model to find approximate optimal solutions for the two criteria: compactness and in-cluster homogeneity regarding several elements. The meta-heuristic chosen as approximation method was TS and it has been incorporated to the proposed model to deal with the complexity of the problem. The proposal has been applied to recent data from the districts of Colima and Puebla.

2.1 Algorithm Description

The algorithm proposed relates homogeneity in the cardinality of the clusters and the objectives of geometric compactness. This algorithm is partitioning around medoids including an approximation method with TS.

The algorithm takes an initial solution of k (possibly random) clusters. Exchanges between clusters elements or representatives produce a new configuration that constitutes a better solution. This new solutions, called "neighbors", are generated until a stopping condition is met. Two of the most widespread methods used to generate neighbor solutions so far are [11] Swap Method and Single Method.

In [9, 10] a classification of partitioning algorithms is described; the most noteworthy algorithms are K-Medians, PAM (Partitioning Around Medoids) and CLARA (Clustering Large Applications) [1, 10, 14]. The shortcomings of these algorithms are generally due to either the quality of the initial solution or the random adjustment of the neighboring solutions leading to local optima for the optimizing criteria.

3 Model

The optimization goals are compactness and homogeneity, and if homogeneity is considered as "hard" restriction, the process tends toward an ideal balancing. In this scenario, a pragmatic alternative is a "soft" restriction, and it means balancing as an additional goal that benefits the computation time penalizing the balance of the solution [8].

The optimization model uses the following definitions [2]:

Definition 1 Compactness

Let $Z = \{1, 2, \ldots, n\}$ be the set of n objects to classify, Z must be divided into k clusters in such a way that:

$$\bigcup_{i=1}^{k} G_i = Z, G_i \cap G_j = \emptyset, i \neq j, \ |G_i| \geq 1, i = 1, 2 \ldots k$$

A cluster G_m with $|G_m|>1$ is compact if each object $t \in G_m$ satisfies:

$$\min_{i \in G_m} d(t, i) > \min_{j \in Z - G_m} d(t, j), i \neq t \tag{1}$$

A cluster G_m with $|G_m| = 1$ is compact if its object $t \in G_m$ satisfies:

$$\min_{i \in Z - \{t\}} d(t, i) > \min_{j,l \in G_f} d(j, l), \forall f \neq m$$

$$\min_{i \in Z - \{t\}} d(t, i) > \min_{j,l \in G_f} d(j, l), \forall f \neq m$$

The neighborhood criterion between objects to achieve compactness is given by the pairs of distances described in 1.

Definition 2 Homogeneity (cardinality of the elements).

Let $Ti = |Gi| for\, i = 1, 2 \ldots k$ and given a homogeneity tolerance percentage $p \in [0, 1]$ that produces two bounds: an inferior bound $I = |n/k| - \lceil n/k * p \rceil$ and a superior bound $S = |n/k| + \lceil n/k * p \rceil$ where n is the number of geographical units and k, the number of clusters to form.

A solution is said to be non-homogeneous when

$$\exists Ti | Ti \notin [I, S], i = 1 \ldots k \tag{2}$$

3.1 Formulation

Let GU be the total number of geographical units and let the initial set of n objects $GU = \{x1, x2 \ldots xn\}$ where x_i is the i-th geographical unit (i is the index for GU and k is the number of zones (clusters).

To reference the formed clusters we define Z_I as the set of the GU that belong to zone l, c_t is the centroid and $d(i, j)$ is the Euclidian distance from node i to node j (i.e. from one GU to another). Then the following restrictions apply:

$Z_I \neq \varnothing$ for $l = 1, \ldots, k$ (the clusters are non-empty),

$Z_I \cap Z_m = \varnothing$ for $l \neq m$ (no GU appears in more than one cluster), and $\bigcup_{l=1}^{k} Z_l = GU$ (all GUs appear in at least one cluster).

Once the number k of centroids $c_t, t = 1, ..., k$ to use has been decided, they must be randomly selected and then assign the corresponding GU as follows: for each GU take

$$i = \min_{t=1,\ldots,k} \{d(i, c_t)\}$$

i.e. each GU is assigned to the nearest centroid c_t.

The homogeneity cost of the solution is defined by the following function:

$$h(T) = \begin{cases} T - S & if\ T > S \\ I - T & if\ T < I \\ 0 & otherwise \end{cases} \tag{3}$$

T is the size of a cluster, I is the inferior bound and S is the superior bound that delimit the ideal cluster size.

For each value of k the sum of the distance between the assigned *GUs* and the centroid is calculated as well as the sum of the remaining or missing elements of each cluster with respect to the given values of inferior and superior bounds. This values are weighted by w_1 and w_2 such that $w1 + w2 = 1$ and lastly the weighted values are summed. This value is minimized through n_{it} iterations. This can be expressed as (4):

$$\min_{l=1,\ldots,nit} \left\{ w1 \left(\min \left\{ \sum_{t=1}^{k} \sum_{i \in c_t} d(i, c_t) \right\} \right) + w2 \left(\sum_{j=1}^{k} h(T_j) \right) \right\} \tag{4}$$

3.2 Tabu Search Proposal

Tabu Search (TS) was introduced in 70s, Fred Glover and Manuel Laguna introduced the name and the methodology later in 1989 in their book Tabu Search [9].

TS guide a search process to negotiate regions that would be difficult to access otherwise. The restrictions are enacted or created by referencing memory structures that are designed for this specific purpose [9, 11].

Diverse applications make use of TS to achieve good quality in their solutions, our interest is centered on examining those works that reveal good results of TS in clustering problems. In [11] a modified tabu search is proposed that comprises two stages: the constructive stage, during which an initial solution is generated using the K-medians algorithm and an improvement stage, where a modified TS is used with the objective of improving on the solution of the constructive stage. The clustering algorithm extracts the main properties of k-medoids with a special emphasis on PAM [12, 14] and for the problem at hand achieved good quality solutions at a reasonable computational cost.

3.3 Data Structures

The first data structure is an array of initial size *k* (number of clusters to form) in which the centroids of each group are stored. To carry out TS it is necessary to define a list where the centroids will reside. The size of this list is dynamic, but the size of the centroid array plus the list of tabu centroids is equal to k throughout execution time, i.e. the centroids are divided in two classes, those that can be replaced and those than cannot (Fig. 1).

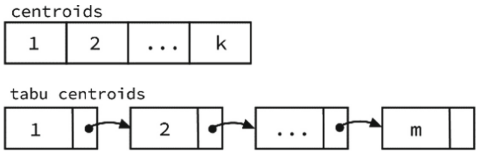


Fig. 1. Centroid structure

The geographical units (*GUs*), in this case Agebs, are stored as an array of initial size n (number of geographical units) and for Tabu Search a list is defined to store the

tabu GUs. The list of tabu *GUs* is of dynamic size and at any given moment its size plus the size of the *GUs* array is equal to n-k, since k *GUs* turn part of the centroid array.

At the homogeneity objective, control is demanded over the *GUs* assigned to each centroid; therefore, a matrix of size $n\text{-}k \times k$ has been included in the implementation. Each column represents a cluster or centroid and each cluster can have a maximum of n-k elements, not counting centroids.

Another array of size k is defined to store the size of each cluster. Initially each cluster has a size equal to 1, (a centroid is counted as an element of the cluster). This array is updated whenever the cost of each accepted solution is calculated. This update justifies the use of homogeneity, where it is necessary to assign the *GUs* to each centroid which guarantees having control over the size of each group in each accepted solution (Fig. 2).

1,1	1,2	...	1,k
2,1	2,2	...	2,k
...	...	...	...
n,1	n,2	...	n,k

Fig. 2. Clustering matrix

4 Algorithm

The following algorithm has been adapted with TS and the model introduced in Sect. 3. The function to optimize is given by Eq. 4. This algorithm is known as Partitioning with TS Penalty (PTSP) through this paper.

Algorithm 1. (PTSP) with Tabu Search

```
Input:
Number of clusters k
Number of iterations nit
Number of iterations for the second phase nit2
Number of iterations for perturbation ip

pc = 0
ic = 1
S = InitialSolution()
S* = S
While ic < nit do
    If UpperPenalty(S) == 0 then
        modeN = 1
    else
        modeN = 2
  end if
  costp = Cost(S)
  Movement(S, modeN)
  If Cost(S) > costp then
        pc = pc + 1
  end if
  If Cost(S) < Cost(S*) then
        S* = S
  end if
  If pc > ip then
        PerturbateSolution(S)
        pc = 0
  end if
  UpdateTabuLists()
  ic = ic + 1
end while
S = S*
ClearTabuStates()
For i = 0 to nit2 do
  Movement(S, modeN)
  If Cost(S) < Cost(S*) then
        S* = S
  end if
UpdateTabuLists()
ic = ic + 1
end for
Return S*
```

The documentation of the algorithm is as follows:

At line 1 the perturbation counter is set to 0. At lines 13 and 14 the perturbation counter is increased by 1 if the cost of the new solution is worse than that of the previous solution. When this counter reaches the maximum value given by *ip* the current solution is perturbed at lines 19–21. The perturbation consists in generating a new random solution and resetting the tabu lists, from such solution the search will be restarted.

At line 3 an initial solution is generated, choosing k objects at random as the cluster centroids, this solution is stored in S. Line 4 designates this solution as the best solution found so far and is represented by S*.

The first search iteration begins at line 5, and finishes when the penalty for the best solution found reaches 0 or when the iteration counter *ic* reaches the maximum number of iterations given by the user.

The "If" conditional inside the loop, at lines 6 to 10, modifies the way in which a centroid to be replaced is chosen in the neighborhood function. When the penalty is equal to 0, i.e. when there are no elements in the clusters that go over the upper bound of homogeneity (see the model in Sect. 3).

At line 11 the cost of the solution S is stored before performing the movement within it at line 12. When the movement is performed, the If conditional at lines 13–15 tests whether the cost of the new solution S is better than the previous cost; if it is not, the perturbation counter pd is increased. The "If" conditional at lines 16–18 takes care of updating the best solution found S* if the new solution just found S is even better.

Lastly, a second phase of searching over the best solution found is performed. This search consists in emptying the tabu lists and thus give place to movements over the best solution for a certain number of iterations (nit2) with the purpose of finding a better solution that can be near S* (a few movements away).

4.1 Neighborhood Function

In the proposed algorithm, the neighbors of the current solution are obtained in two ways: Select either the smallest or greatest cluster from the current solution considering its penalty cost (if there are elements below the lower bound or above the upper bound). If the cluster has size 1 then the cluster centroid is replaced by a randomly selected non-centroid geographical unit. The tabu tenure that was established is $k-1$ (number of groups -1) that through experimentation has shown the intensity level necessary to achieve acceleration with optimal costs.

4.2 Results

The tests were performed on hardware with the following characteristics:

- CPU: Dual Core AMD E-350 a 1.6 Ghz.
- RAM: 2GB DDR3.
- HDD: SATA-II 320GB 5400 RPM
- OS: Windows 7 Ultimate 32bits

For comparison with previous tests [4], the map of the Toluca valley, Mexico was considered. This map has 469 geographical units and tests were performed to create from 2 to 200 and 300 clusters. Table 1 summarizes the results from a prototype based on PAM (PP) but doesn't exclude the optimization model proposed in this paper. It is important to highlight that PP does not handle an approximation method, it was developed for this paper with the aim of experimenting the duality of the optimized functions, which remained in the final implementation in at least two features: (1) Excellent quality solutions were observed for the test performed, such that the greater part of the programming strategies at this point are kept for the next and final version. (2) For tests with more than 200 clusters the computational cost is high and it was not possible to record the cost of the solutions. This new proposal called Partitioning with TS Penalty (PTSP) is the main contribution of this paper. The results are shown in Table 2.

Table 1. Tests with Penalty PAM (PP) applying the model from Sect. 3.

		Penalty PAM	
	Compactness	Penalty	Time (seg)
2	36.5485	0	0.075
4	27.4569	0	0.222
6	31.4284	0	0.670
8	27.2486	0	1.051
10	19.9424	0	3.042
20	13.6542	0	24.041
40	8.6276	1	114.813
60	7.4026	9	261.200
80	5.3185	17	469.071
100	4.5248	6	860.164
120	3.9991	11	1391.139
140	2.8315	0	1894.436
160	2.8784	11	2249.452
180	2.0487	0	2325.100
200	1.6022	0	2482.393

We can see Partitioning test with TS Penalty (PTSP) in the next Table 2.

The algorithm 1 in Table 2 (PSD) corresponds to a previous proposal to solve homogeneity with modifications [4]. This algorithm minimizes the standard deviation of the cluster size compared to the ideal size. In the table the tests corresponding to PTSP can also be observed. An analysis of both tables shows that the model from Sect. 3 along with a new homogeneity measurement based on minimizing the remaining elements, is far superior with respect to PAM Penalty (Table 1) as PSD. Figure 3 shows a graphical

Table 2. Tests Test with PAM (Standard Deviation and PTSP. Compactness (Comp) and Penalty (Pen)

	PAM PSD			TS Penalty PSTP		
Group	Comp	Pen	Time	Comp	Pen	Time
2	37.2245	0	0.234	36.3865	0	990
4	30.9555	0	0.190	27.3666	0	2.010
6	29.4603	0	0.610	23.39{}53	0	3.360
8	24.4460	0	2.645	21.0318	0	4.293
10	17.1270	2	4.101	17.4750	0	5.531
20	13.0818	1	19.277	13.6844	0	24.261
40	7.5209	16	136.340	8.9531	0	89.232
60	5.1136	51	334.054	6.3707	26	93.405
80	3.8886	97	549.179	5.0451	28	202.395
100	3.0497	106	1042.438	3.9835	32	105.953
120	2.5883	103	1464.341	3.4795	39	93.612
140	2.1947	103	2080.419	2.9383	19	241.207
160	1.8786	92	2474.121	2.6919	49	89.398
180	1.6211	69	2566.385	2.5472	29	241.301
200	1.4206	49	2829.424	2.0897	19	79.366
220	1.2410	38	3249.674	1.9316	11	73.661
240	1.0895	100	3380.657	1.7650	58	70.450
260	0.9491	89	2508.559	1.5874	40	140.298
280	0.8077	77	2394.728	1.3897	31	140151
300	0.6807	60	2610.345	1.2878	21	127.297

result for 10 clusters that corresponds to the test of Table 2. This map was produced by the interface with a Geographic Information System (SIG) [3].

5 Conclusions

The computational results allow us to conclude with certainty that from the algorithms presented, PTSP reduce cost time. We must notice that PTSP accepts big sized instances that can't be tested with traditional algorithms due to the high computing time required.

We distinguish that our PTSP proposal surpasses the PAM Penalty algorithm (PP in Table 1) regarding time, because the execution time of PP increases quickly and exponentially with bigger instances, whereas PTSP maintains a performance up to 90% faster with the parameters used for this case (20,000 iterations). As we saw in Sect. 4, PTSP combines random and strategic neighbor selection operations. For this reason, its execution times can vary even with the same input parameters.

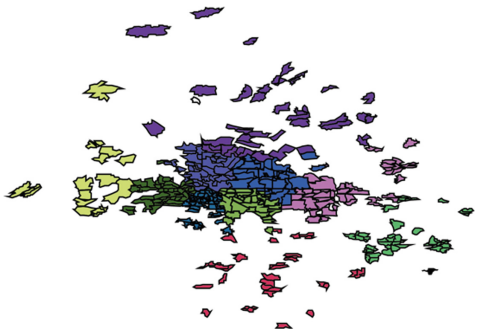


Fig. 3. Map of Toluca. Test obtained by PTSP in Table 2 for G = 10.

As a developing work, we are looking for a different algorithm from others authors and make comparison with the results of this work.

On the other hand, tabu search has been improved by incorporating by simulated annealing to obtain a hybridization to get better approximations.

Finally, it is estimated to incorporate our algorithm to clustering problems that requires a balance between their objects.

References

1. Anderberg, M.: Cluster Analysis for Applications. Academic Press (1973)
2. Bernábe, B., Espinosa, J., Ramiréz, J., Osorio, M.A.: Statistical comparative analysis of simulated annealing and variable neighborhood search for the geographical clustering problem. Computación y Sistemas **42**(3), 295–308 (2011)
3. Bernábe, B., González, R.: Integración de un sistema de información geográfica para algoritmos de particionamiento. Research in Computing Science, Avances en la Ingeniería del Lenguaje y Conocimiento **88**, 31–44 (2014)
4. Bernábe, B., Martínez, J.L., Olivares, E., et al.: Extensions to K-medoids with balance restrictions over the cardinality of the partitions. J. Appl. Res. Technol. **12**, 396–408 (2014)
5. Chaieb, M., Sassi, D.B.: Measuring and evaluating the home health care scheduling problem with simultaneous pick-up and delivery with time window using a Tabu search metaheuristic solution. Appl. Soft Comput. **113**, 107957 (2021)
6. Téllez, N., Jimeno, M., Salazar, A., Nino-Ruiz, E.: A tabu search method for load balancing in fog computing. Int. J. Artif. Intell **16**(2), 1–30 (2018)
7. Romero, D.: Formación de unidades primarias de muestreo. Forthcoming
8. García, J.P., Maheut, J.: Modelos de programación lineal: Definición de objetivos. In: Modelos y Métodos de Investigación de Operaciones. Procedimientos para Pensar, pp. 42–44 (2011). Available via DOCPLAYER. https://docplayer.es/3542781-Modelos-y-metodos-de-investigacion-de-operaciones-procedimientos-para-pensar.html. Accessed 22 Sept 2022
9. Glover, F., Laguna, M.: Tabu Search. Kluwer Academic Publishers (1997)
10. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons (1990)
11. Kharrousheh, A., Abdullah, S., Nazri, M.Z.A.: A Modified Tabu search approach for the clustering problem. J. Appl. Sci. **19**, 3447–3453 (2011)

12. Leiva, S.A., Torres, F.J.: Una revisión de los algoritmos de partición más comunes de conglomerados: un estudio comparativo. Revista Colombiana de Estadística **33**(2), 321–339 (2010)
13. Altman, M.: The computational complexity of automated redistricting: Is automation the answer? RutgersComput. Technol. Law J. **23**(1), 81–141 (1997)
14. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
15. Nickel, S., Schröder, M., Kalcsics, J.: Towards a unified territorial design approach—Applications, algorithms and GIS integration. Top J. Oper. Res. **13**, 1–74 (2005)
16. Salazar, M.A., González, J.L., Ríos, R.Z.: A Divide-and-conquer approach to commercial territory design. Computación y Sistemas **16**(3), 309–320 (2012)

Evacuation Centers Choice by Intuitionistic Fuzzy Graph

Alexander Bozhenyuk(✉), Evgeniya Gerasimenko, and Sergey Rodzin

Southern Federal University, Nekrasovsky 44, 347922 Taganrog, Russia
avb002@yandex.ru

Abstract. The problem of choosing places for evacuation centers is considered in this paper. We consider the case when the territory model is represented by an intuitionistic fuzzy graph. To solve this problem, the concept of a minimal antibase of such graph is introduced, and on its basis, the concept of an antibase set as an invariant of this graph is introduced too. A method and algorithm for calculating the minimal antibases are considered. The problem of finding all minimal antibases of the graph allows us to solve the task of determining the antibase set. The paper considers a numerical example of finding the antibase set of an intuitionistic fuzzy graph. The task of choosing the places of evacuation centers in an optimal way depends on their number. The calculation of the minimal antibase set allows us to directly solve this problem.

Keywords: Evacuation · Evacuation Centers · Intuitionistic Fuzzy Graph · Minimal Intuitionistic Antibase Vertex Subset · Antibase Set

1 Introduction

Evacuation, as a kind of response to a threatening situation caused by various situations of a natural or man-made nature, can mitigate the negative impact of a possible disaster on the population of a given territory. The evacuation of populations is a complex process. Therefore, evacuation planning plays an important role in ensuring its effectiveness.

To support planning and decision-making, approaches related to the optimization of evacuation routes are of great importance. At the same time, decision makers (DM), with a comprehensive assessment of the circumstances, face many factors and uncertainties.

In order to facilitate the management of evacuation operations, evacuation plans must be developed during the preparation phase. The work [1] presents various methods for support effective evacuation planning. However, both a general evacuation planning model and a general set of specific parameters that should be included in the plan as initial data are missing here. The work [2] considers various stages in the planning of flood evacuation. But there is no approach to assessing information about the current situation to justify the need for evacuation. The evacuation studies carried out in [3] identified the following tasks for the development of an evacuation plan at the preparation stage: determining the predicted parameters and disaster scenarios, characterizing the vulnerability, determining actions and data such as the capacity of the transport network,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 391–400, 2023.
https://doi.org/10.1007/978-3-031-27409-1_35

the number of evacuees, strategies and evacuation scenarios, their optimization, selection of an evacuation plan and its application in real time. The work [4] presents a program for modeling floods, traffic flows during evacuation, as well as optimization of possible strategies. According to their purpose, evacuation modeling tools can be divided into two types: models of specific disasters [5, 6] and models that provide evacuation [1, 7–9].

Existing evacuation traffic models can be classified as:

- flow models [10];
- agent-based models, in which individual vehicles are considered as agents with autonomous behavior interacting with other vehicles [11];
- scenario-based simulation models to identify evacuation bottlenecks [12].

The paper [13] presents a review of the literature on the methods of mathematical modeling of evacuation traffic. Time models taking into account critical paths are presented in [14].

The decision to initiate a mass evacuation plan based on a crisis assessment becomes a challenge for decision makers. Several issues related to this problem are considered in the literature: criteria for making decisions about evacuation, the decision-making process taking into account uncertain factors, as well as decision-making modeling [15, 16].

Accounting for forecast uncertainty is a complex part of the decision making. Several studies have been conducted to quantify the uncertainty of possible developments and to help decision makers determine what to plan for. Some studies emphasize the importance of interpreting uncertainty in predicting the level of danger and evacuation [17, 18].

Subjective uncertainty factors are not widely represented in the literature. They are difficult to model, so studies are required to consider subjective uncertainty in the process of planning to evacuate. At present, the need to model support for making decisions about evacuation is becoming increasingly important. Such tasks are difficult to formalize, characterized by incompleteness and fuzziness of the initial information, fuzziness of the goals set [19].

This paper considers one of the tasks that arises when supporting decision-making during evacuation, namely, the choice of locations for evacuation centers on the plan of a certain territory. At the same time, the territory model is represented by an intuitionistic fuzzy graph. In the graph under consideration, the vertices determine the locations of people and the possible locations of evacuation centers, and the intuitionistic degree assigned to the edges determines the degree of safety of movement along this edge. Concepts of the minimal antibase and the antibase set of intuitionistic fuzzy graph are introduced here. It is shown that the choice of the best placement of evacuation centers is equivalent to finding an intuitionistic fuzzy set of antibases for a given graph.

2 Preliminaries

The concept of a fuzzy set as a method of representing uncertainty was proposed and discussed in [20]. In the article [21], the fuzzy set was generalized as the concept of

an intuitionistic fuzzy set. In the latter, the degree of non-membership was added to the concept of the membership function of the fuzzy set.

The original definition of a fuzzy graph [22] was based on the concept of a fuzzy relationship between vertices [23]. The concepts of an intuitionistic fuzzy relation and an intuitionistic fuzzy graph were considered in the papers [24, 25]. The concepts of a dominating set, and a base set as invariant of intuitionistic fuzzy graph were introduced in the papers [26–28].

The intuitionistic fuzzy set on the set X is the set of triples [21] $\tilde{A} = \{\langle x, \mu_A(x), \nu_A(x)\rangle | x \in X\}$. Here $\mu(x) \in [0,1]$ is the membership function of x in $\tilde{A}$, and $\nu_A(x) \in [0, 1]$ is the non-membership function x in $\tilde{A}$. Moreover, for any $x \in X$ the values $\mu_A(x)$, and $\nu_A(x)$ must satisfy the condition $\mu_A(x) + \nu_A(x) \leq 1$.

The intuitionistic fuzzy relation $\tilde{R} = (\mu_R(x,y), \nu_R(x,y))$ on the set $X \times Y$ is the set $\tilde{R} = \{\langle (x,y), \mu_R(x,y), \nu_R(x,y)\rangle \mid (x,y) \in X \times Y\}$, where $\mu_R: X \times Y \to [0,1]$ and $\nu_R: X \times Y \to [0,1]$. In this case, the following condition is fulfilled $(\forall x,y \in X)[\mu_R(x,y) + \nu_R(x,y) \leq 1]$.

Let $p = (\mu(p),\nu(p))$ and $q = (\mu(q), \nu(q))$ be intuitionistic fuzzy variables, where $\mu(p) + \nu(p) \leq 1$ и $\mu(q) + \nu(q) \leq 1$. Then the operations "&" and "$\vee$" are defined as [15]:

$$p \,\&\, q = (\min(\mu(p), \mu(q)), \max(\nu(p), \nu(q))), \tag{1}$$

$$p \vee q = (\max(\mu(p), \mu(q)), \min(\nu(p), \nu(q))). \tag{2}$$

We will consider $p \leq q$ if $\mu(p) \leq \mu(q)$ and $\nu(p) \geq \nu(q)$. Otherwise, we will assume that p and q are incommensurable intuitionistic fuzzy variables.

An intuitionistic fuzzy graph [24, 25] is a pair $\tilde{G} = (\tilde{A}, \tilde{U})$, where $\tilde{A} = \langle V, \mu_A, \nu_A\rangle$ is an intuitionistic fuzzy set on the vertex set V, $\tilde{U} = \langle V \times V, \mu_U, \nu_U\rangle$ is an intuitionistic fuzzy set of edges, and the following inequalities hold:

$$\mu_U(xy) \leq \min(\mu_A(x), \mu_A(y)); \tag{3}$$

$$\nu_U(xy) \leq \max(\nu_A(x), \nu_A(y)); \tag{4}$$

$$(\forall x, y \in V)[0 \leq \mu_U(xy) + \nu_U(xy) \leq 1]. \tag{5}$$

3 Antibase Set

Let $\tilde{G} = (\tilde{A}, \tilde{U})$ be an intuitionistic fuzzy graph. Let $p(x,y) = (\mu(x,y),\nu(x,y))$ be an intuitionistic fuzzy variable that determines the degree of adjacency and degree of non-adjacency of vertex y from vertex x.

An intuitionistic fuzzy path $\tilde{L}(x_i, x_j)$ [29, 30] from a vertex x_i to a vertex x_j of a graph $\tilde{G} = (\tilde{A}, \tilde{U})$ is a directed sequence of vertices and edges in which the end vertex of any edge (except for x_j), is the starting vertex of the next arc.

The strength of the path $s(\tilde{L}(x_i, x_j))$ is determined by the smallest value of the degrees of vertices and edges included in this path. Taking into account expressions (3) and (4),

the strength $s(\tilde{L}(x_i, x_j))$ of the path $\tilde{L}(x_i, x_j)$ is determined only by the values of its edges: $s(\tilde{L}(x_i, x_j)) = \underset{(x_\alpha, x_\beta) \in \tilde{L}(x_i, x_j)}{\&} p(x_\beta, x_\beta)$. Here the operation & is defined according to expression (1).

Since the strength of the path depends on the intuitionistic degrees of the edges and does not depend on the degrees of the vertices, we will further consider intuitionistic fuzzy graphs with crisp vertices: $\tilde{G} = (V, \tilde{U})$.

The vertex x_j is *reachable* from the vertex x_i if there exists an intuitionistic fuzzy path $\tilde{L}(x_i, x_j)$ with degree $s\left(\tilde{L}(x_i, x_j)\right)$ different from (0,1). Each vertex x_i is considered to be reachable from itself with degree $s\left(\tilde{L}(x_i, x_i)\right) = (1, 0)$.

The degree of reachability of the vertex x_j from the vertex x_i is determined by the expression:

$$\gamma(x_i, x_j) = \underset{k \in \overline{1,t}}{\vee} \{s(\tilde{L}_k(x_i, x_j))\} = \underset{k \in \overline{1,t}}{\vee} \{s(\mu_k, \nu_k)\}\ . \tag{6}$$

Here t is the number of different paths from vertex x_i to vertex x_j. Here the operation $\vee$ is defined according to expression (2).

If among the paths there are paths with an incommensurable degree, then as the degree of reachability we will choose the value for which the membership degree (μ_k) is the largest.

Example 1 Consider the intuitionistic fuzzy graph $\tilde{G}_1$, shown in Fig. 1.

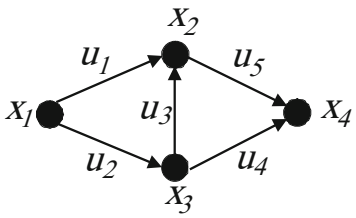


Fig. 1. Intuitionistic fuzzy graph $\tilde{G}_1$.

Table 1 gives an intuitionistic fuzzy set of edges:

Table 1. Intuitionistic fuzzy set edges of graph $\tilde{G}_1$.

u_1	u_2	u_3	u_4	u_5
(0.4,0.5)	(0.6,0.4)	(0.5,0.3)	(0.2,0.7)	(0.8,0.0)

Vertex x_1 is not reachable from the vertex x_4, but the vertex x_4 is reachable from the vertex x_1 by three ways:

- $\tilde{L}_1 = (x_1, u_1, x_2, u_2, x_4)$, with degree s_1 = (0.4,0.5) & (0.8,0) = (0.4,0.5);
- $\tilde{L}_2 = (x_1, u_2, x_3, u_4, x_4)$, with degree s_2 = (0.6,0.4) & (0.2,0.7) = (0.2,0.7);
- $\tilde{L}_3 = (x_1, u_2, x_3, u_3, x_2, u_5, x_4)$, with degree s_3 = (0.6,0.4) & (0.5,0.3) & (0.8,0) = (0.5,0.4).

In this case, the degree of reachability will be defined as: $\gamma(x_1, x_4) = (0.5, 0.4)$.

Example 2 Consider the intuitionistic fuzzy graph $\tilde{G}_2$, shown in Fig. 2. Table 2 gives an intuitionistic fuzzy set of graph $\tilde{G}_2$ edges.

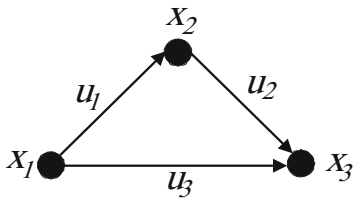


Fig. 2. Intuitionistic fuzzy graph $\tilde{G}_2$.

Table 2. Intuitionistic fuzzy set edges of graph $\tilde{G}_2$.

u_1	u_2	u_4
(0.8,0.1)	(0.3,0.2)	(0.5,0.3)

Vertex x_3 is reachable from the vertex x_1 by two ways with incommensurable degrees:

- $\tilde{L}_1 = (x_1, u_1, x_2, u_2, x_3)$, with degree s_1 = (0.8,0.1) & (0.3,0.2) = (0.3,0.2);
- $\tilde{L}_2 = (x_1, u_3, x_3)$, with degree $s_2 = (0.5, 0.3)$.

Therefore, the degree of reachability will be defined as: $\gamma(x_1, x_3) = (0.5, 0.3)$.
Let's the number of graph vertices $|V| = n$.

Definition 1 Intuitionistic fuzzy antibase of a graph $\tilde{G}$ is a subset of vertices $\overline{B}_\beta \subseteq V$, that have the property that at least one of these vertices is reachable from any other vertices $V \backslash \overline{B}_\beta$ with an intuitionistic reachability degree of at least $\beta = (\mu_\beta, \nu_\beta)$.

Definition 2 Intuitionistic fuzzy antibase will be called *minimal* if there is no other antibase $B' \subset \overline{B}_\beta$, with the same intuitionistic reachability degree β.

Minimal intuitionistic fuzzy antibase determines the best placement of evacuation centers in the territory modeled by graph $\tilde{G}$. In this case, evacuation centers number is determined by the vertices number of of the considered antibase.

The following property follows from the definition of an intuitionistic fuzzy antibase:

Property 1 Let $\overline{B}_\beta$ be an minimal intuitionistic fuzzy antibase. Then the following statement is true:

$(\forall x_i, x_j \in \overline{B}_\beta)[\gamma(x_i, x_j) < \beta]$.

In other words, the intuitionistic reachability degree between any two vertices belonging to the minimal intuitionistic fuzzy antibase $\overline{B}_\beta$ is less than the value β of this antibase.

Consider a family of subsets of minimal intuitionistic fuzzy antibases $\beta_i = \{\overline{B}_{i_1}, \overline{B}_{i_2}, ..., \overline{B}_{i_k}\}$, each of which consists of i vertices and has reachability degrees $\{\beta_{i_1}, \beta_{i_2}, ..., \beta_{i_k}\}$ respectively. Let β_i be the largest of these degrees. If the family $\beta_i = \varnothing$, then $\beta_i = \beta_{i-1}$.

Definition 2 We call the intuitionistic fuzzy set

$$\tilde{B} = \{< \beta_1/1 >, < \beta_2/2 >, ..., < \beta_n/n >\}$$

the *antibase set* of the graph $\tilde{G}$.

Thus, the antibase set determines the greatest possible reachability degree (β_i) for a given number of evacuation centers ($i = \overline{1,n}$).

Property 2 For antibase set, the following inequality holds true:

$$(0, 1) \le \beta_1 \le \beta_2 \le ... \le \beta_n = (1, 0).$$

4 Method for Finding Minimal Intuitionistic Fuzzy Antibases

We consider a method for finding the family of all minimal intuitionistic fuzzy antibases. This method is similar to the approach proposed in [31].

Let $\overline{B}_\beta$ be the minimal antibase with intuitionistic reachability degree $\beta = (\mu_\beta, \nu_\beta)$. Then the following expression is true:

$$(\forall x_i \in V)[x_i \in \overline{B}_\beta \vee (\exists x_j \in \overline{B}_\beta | \mu(x_i, x_j) \ge \mu_\beta \,\&\, \nu(x_i, x_j) \le \nu_\beta)]. \tag{7}$$

For each vertex $x_i \in V$ we introduce a variable p_i such that if $x_i \in \overline{B}_\beta$ then $p_i = 1$, and 0 otherwise. Let us associate the intuitionistic variable $\xi_{ij} = \beta = (\mu_\beta, \nu_\beta)$ for the expression $(\mu(x_i, x_j), \gamma(x_i, x_j)) \ge \beta$. Then, passing from the quantifier notation in expression (7) to logical operations, we obtain the truth:

$$\Phi_{\overline{B}} = \underset{i=\overline{1,n}}{\&} (p_i \vee \underset{j=\overline{1,n}}{\vee} (p_j \& \xi_{ij})).$$

Considering that $(\forall j = \overline{1,n})[\xi_{jj} = (1, 0)]$, and $(\forall i = \overline{1,n})\left[p_i \vee \underset{j}{\vee} p_j \& \xi_{ij} = \underset{j}{\vee} p_j \& \xi_{ij}\right]$, the last expression will be rewritten as:

$$\Phi_{\overline{B}} = \underset{i=\overline{1,n}}{\&} \left(\underset{j=\overline{1,n}}{\vee} (p_j \,\&\, \xi_{ij}) \right). \tag{8}$$

Let us open the brackets in expression (8) and reduce like terms, following the rules:

$$a \vee a \,\&\, b = a;\ \xi_1 \,\&\, a \vee \xi_2 \,\&\, a \,\&\, b = \xi_1 \,\&\, a \ \text{ if } \xi_1 \geq \xi_2. \tag{9}$$

Here, $a, b \in \{0, 1\}$, and $(0, 1) \leq \xi_1, \xi_2 \leq (1, 0)$.

Then the expression (8) can be rewritten as:

$$\Phi_{\overline{B}} = \underset{i=\overline{1,l}}{\vee} \left(p_{1_i} \,\&\, p_{2_i} \&\ldots \&\, p_{k_i} \,\&\, \beta_i\right). \tag{10}$$

The variables included in each parenthesis of expression (10) define the minimum antibase set with the intuitionistic reachability degree β_i. Having found all minimum antibase sets, we automatically determine the antibase set of the considered graph.

5 Example

Let's consider an example of the best placement of district evacuation centers, the model of which is represented by the intuitionistic fuzzy graph $\tilde{G}_3$, shown in Fig. 3. To do this, we will find all minimum antibases according to the considered approach.

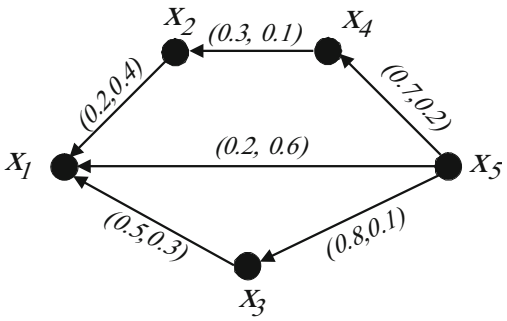


Fig. 3. Intuitionistic fuzzy graph $\tilde{G}_3$.

The adjacency matrix of the graph $\tilde{G}_3$ has the form:

$$R_X = \begin{array}{c|ccccc} & x_1 & x_2 & x_3 & x_4 & x_5 \\ x_1 & (1.0, 0.0) & (0.0, 1.0) & (0.0, 1.0) & (0.0, 1.0) & (0.0, 1.0) \\ x_2 & (0.2, 0.4) & (1.0, 0.0) & (0.0, 1.0) & (0.0, 1.0) & (0.0, 1.0) \\ x_3 & (0.5, 0.3) & (0.0, 1.0) & (1.0, 0.0) & (0.0, 1.0) & (0.0, 1.0) \\ x_4 & (0.0, 0.0) & (0.3, 0.1) & (0.0, 1.0) & (1.0, 0.0) & (0.0, 1.0) \\ x_5 & (0.2, 0.6) & (0.0, 0.0) & (0.8, 0.1) & (0.7, 0.2) & (1.0, 0.0) \end{array}$$

Based on the adjacency matrix, one can construct reachability matrices:

$$R_D = \begin{array}{c|ccccc} & x_1 & x_2 & x_3 & x_4 & x_5 \\ x_1 & (1.0, 0.0) & (0.0, 1.0) & (0.0, 1.0) & (0.0, 1.0) & (0.0, 1.0) \\ x_2 & (0.2, 0.4) & (1.0, 0.0) & (0.0, 1.0) & (0.0, 1.0) & (0.0, 1.0) \\ x_3 & (0.5, 0.3) & (0.0, 1.0) & (1.0, 0.0) & (0.0, 1.0) & (0.0, 1.0) \\ x_4 & (0.2, 0.4) & (0.3, 0.1) & (0.0, 1.0) & (1.0, 0.0) & (0.0, 1.0) \\ x_5 & (0.5, 0.3) & (0.3, 0.2) & (0.8, 0.1) & (0.7, 0.2) & (1.0, 0.0) \end{array}$$

Using the reachability matrix, we write the expression (8):

$$\Phi_{\overline{B}} = \big[(1.0,0.0)p_1\big]\&\big[(0.2, 0.4)p_1 \vee (1.0, 0.0)p_2\big] \& \big[(0.5, 0.3)p_1 \vee (1.0, 0.0)p_3\big]\& \\ \&\big[(0.2, 0.4)p_1 \vee (0.3, 0.1)p_2 \vee (1.0, 0.0)p_4\big]\& \\ \&\big[(0.5, 0.3)p_1 \vee (0.3, 0.2)p_2 \vee (0.8, 0.1)p_3 \vee (0.7, 0.2)p_4 \vee (1.0, 0.0)p_5\big].$$

Multiplying brackets 1 and 2, brackets 3 and 4, and using the absorption rules (8) we get:

$$\Phi_{\overline{B}} = \big[(0.2, 0.4)p_1 \vee (1.0, 0.0)p_1p_2\big]\& \\ \&\big[(0.2, 0.4)p_1 \vee (0.3, 0.3)p_1p_2 \vee (0.5, 0.3)p_1p_4 \vee (0.3, 0.1)p_2p_3 \vee (1.0, 0.0)p_3p_4\big]\& \\ \&\big[(0.5, 0.3)p_1 \vee (0.3, 0.2)p_2 \vee (0.8, 0.1)p_3 \vee (0.7, 0.2)p_4 \vee (1.0, 0.0)p_5\big].$$

Multiplying brackets 1 and 2, also using the absorption rules (8) we get:

$$\Phi_{\overline{B}} = [(0.2, 0.4)p_1 \vee (0.3, 0.3)p_1p_2 \vee (0.5, 0.3)p_1p_2p_4 \vee (0.3, 0.1)p_1p_2p_3 \vee \\ \vee (1.0, 0.0)p_1p_2p_3p_4]\& \\ \&\big[(0.5, 0.3)p_1 \vee (0.3, 0.2)p_2 \vee (0.8, 0.1)p_3 \vee (0.7, 0.2)p_4 \vee (1.0, 0.0)p_5\big].$$

Multiplying the brackets, we finally get:

$$\Phi_{\overline{B}} = (0.2, 0.4)p_1 \vee (0.3, 0.3)p_1p_2 \vee (0.5, 0.3)p_1p_2p_4 \vee (0.3, 0.1)p_1p_2p_3 \vee \\ \vee (0.8, 0.1)p_1p_2p_3p_4 \vee (1.0, 0.0)p_1p_2p_3p_4p_5.$$

Whence it follows that this graph has 6 minimum intuitionistic antibases. From here it follows that if we have 2 evacuation centers at our disposal, then the best places for their placement are the vertices x_1 and x_2.

The antibase set for the considered graph $\tilde{G}_3$ will look like:

$$\tilde{B} = \{\langle(0.2, 0.4)/1\rangle, \langle(0.3, 0.3)/2\rangle, \langle(0.5, 0.3)/3\rangle, \\ \langle(0.8, 0.1)/4\rangle, \langle(1, 0)/5\rangle\}.$$

This set, in particular, can help answer the question: does it make sense to use two evacuation centers, or can one be enough? In this case, the intuitionistic reachability degree will decrease from the value (0.3,0.3) to (0.2,0.4).

6 Conclusion and Future Scope

The problem of choosing places for evacuation centers when the territory model is represented by an intuitionistic fuzzy graph was considered. To solve this problem, the definitions of the minimal antibase and the antibase set of intuitionistic fuzzy graph were introduced. The method and algorithm for calculating all minimal antibases of graph have been considered. The numerical example of finding the antibase set has been reviewed. It is shown that the antibase set allows solving the problem of choosing the places of evacuation points in an optimal way, depending on the number of evacuation centers. In this paper we considered the case of optimal placement of evacuation centers at the vertices of the graph. In further studies, it is planned to consider cases of placing evacuation centers on the edges of the intuitionistic fuzzy graph. Which leads to the need to consider the problem of generating new graph vertices.

Acknowledgments. The research was funded by the Russian Science Foundation project No. 22–71-10121, https://rscf.ru/en/project/22-71-10121/ implemented by the Southern Federal University.

References

1. Shaw, D., et al.: Evacuation Responsiveness by Government Organisations (ERGO): Evacuation Preparedness Assessment Workbook. Technical report. Aston CRISIS Center. (2011)
2. Lumbroso, D., Vinet, F.: Tools to improve the production of emergency plans for floods: are they being used by the people that need them? J. Contingencies Crisis Manag. **20**, 149–165 (2012)
3. Hissel, M., François, H., Xiao, J.J.: Support for preventive mass evacuation planning in urban areas. IET Conf. Public. **582**, 159–165 (2011). https://doi.org/10.1049/cp.2011.0277
4. Chiu, Y., Liu, H.X.: Emergency Evacuation, Dynamique Transportation Model. Spring Street, NY 10013, USA: Springer Science Buisiness Media, LLC. (2008)
5. Bayram, V.: Optimization models for large scale network evacuation planning and management: a literature review. Surv. Oper. Res. Manag. Sci. **21**(2), 63–84 (2016)
6. Gao, Z., Qu, Y., Li, X., Long, J., Huang, H.-J.: Simulating the dynamic escape process in large public places. Oper. Res. **62**(6), 1344–1357 (2014)
7. Lazo, J.K., Waldman, D.M., Morrow, B.H., Thacher, J.A.: Household evacuation decision making and the benefits of improved hurricane forecasting: developing a framework for assessment. Weather Forecast. **25**(1), 207–219 (2010)
8. Simonovic, S.P., Ahmad, S.: Computer-based model for flood evacuation emergency planning. Nat. Hazards **34**(1), 25–51 (2005)
9. Dash, N., Gladwin, H.: Evacuation decision making and behavioral responses: individual and household. Nat. Hazard. Rev. **8**(3), 69–77 (2007)
10. Lessan, J., Kim, A.M.: Planning evacuation orders under evacuee compliance uncertainty. Saf. Sci. **156**, 105894 (2022)
11. Stepanov, A., Smith, M.J.: Multi-objective evacuation routing in transportation networks. Eur. J. Oper. Res. **198**(2), 435–446 (2009)
12. Lindell, M., Prater, C.: Critical behavioral assumptions in evacuation time estimate analysis for private vehicles: examples from hurricane research and planning. J. Urban Plann. Dev. **133**(1), 18–29 (2007)

13. Bretschneider, S.: Mathematical Models for Evacuation Planning in Urban Areas. Springer, Verlag Berlin Heidelberg (2013)
14. Hissel, F.: Methodology for the Implementation of Mass Evacuation Plans. CEMEF, France, Compiègne (2011)
15. Kailiponi, P.: Analyzing evacuation decision using Multi-Attribute Utility Theory (MAUT). Procedia Eng. **3**, 163–174 (2010)
16. Regnier, E.: Public evacuation decision and hurricane track uncertainty. Manag. Sci. **54**(1), 16–28 (2008)
17. Agumya, A., Hunter, G.J.: Responding to the consequences of uncertainty in geographical data. Int. J. Geogr. Inf. Sci. **16**(5), 405–417 (2002)
18. Kunz, M., Gret-Regamey, A., Hurni, L.: Visualization of uncertainty in natural hazards assessments using an interactive cartographic information system. Nat. Hazards **59**(3), 1735–1751 (2011)
19. Kacprzyk, J., Zadrozny, S., Nurmi, H., Bozhenyuk, A.: Towards innovation focused fuzzy decision making by consensus. In: Proceedings of IEEE International Conference on Fuzzy Systems. pp. 256–268 (2021)
20. Zadeh, L.A.: Fuzzy sets. Inf. Contr. **8**, 338–353 (1965)
21. Atanassov, K.T.: Intuitionistic fuzzy sets. In: Proceedings of VII ITKR's Session, Central Science and Technical Library, vol. 1697/84, pp. 6–24. Bulgarian Academy of Sciences, Sofia (1983)
22. Christofides, N.: Graph Theory. An Algorithmic Approach. Academic Press, London, UK (1976)
23. Zadeh, L.A.: Similarity relations and fuzzy orderings. Inf. Sci. **3**(2), 177–200 (1971)
24. Shannon, A., Atanassov K.T.: A first step to a theory of the intuitionistic fuzzy graphs. In: Lakov, D. (ed.) Proceeding of the FUBEST, pp. 59–61. Sofia, Bulgaria (1994)
25. Shannon, A., Atanassov, K.T.: Intuitionistic fuzzy graphs from α-, β- and (α, b)-levels. Notes on Intuitionistic Fuzzy Sets **1**(1), 32–35 (1995)
26. Karunambigai, M.G., Sivasankar, S., Palanivel, K.: Different types of domination in intuitionistic fuzzy graph. Ann. Pure Appl. Math. **14**(1), 87–101 (2017)
27. Shubatah, M.M., Tawfiq, L.N., AL-Abdli, A.A.-R.A.: Edge domination in intuitionistic fuzzy graphs. South East Asian J. Math. Math. Sci. **16**(3), 181–198 (2020)
28. Kahraman, C., Bozhenyuk, A., Knyazeva, M.: Internally stable set in intuitionistic fuzzy graph. Lecture Notes Netw. Syst. **504**, 566–572 (2022)
29. Bozhenyuk, A., Knyazeva, M., Rozenberg, I.: Algorithm for finding domination set in intuitionistic fuzzy graph. Atlantis Stud. Uncertainty Model. **1**, 72–76 (2019)
30. Bozhenyuk, A., Belyakov, S., Knyazeva, M., Rozenberg, I.: On computing domination set in intuitionistic fuzzy graph. Int. J. Comput. Intell. Syst. **14**(1), 617–624 (2021)
31. Bozhenyuk, A., Belyakov, S., Kacprzyk, J., Knyazeva, M.: The method of finding the base set of intuitionistic fuzzy graph. Adv. Intell. Syst. Comput. **1197**, 18–25 (2021)

Movie Sentiment Analysis Based on Machine Learning Algorithms: Comparative Study

Nouha Arfaoui(✉)

Research Team in Intelligent Machines, National Engineering School, Gabes, Tunisia
arfaoui.nouha@yahoo.fr

Abstract. The movie market products many movies each year. This makes hard to select the appropriate film to watch. The viewer, generally, has to read the feedback of the previous ones to make decision. He finds a problem because of the massive quantity of feedbacks. Hence, it is necessary to use Machine learning algorithms because they automatically analyze and classify the collected movies feedbacks into negative and positive reviews. Different works in the literature use machine learning for movies' sentiment analysis. They select some algorithms for the evaluation. And because it is always complicated to select the appropriate machine learning algorithms according to the requirements that get poor accuracy and performance, we propose, in this work, the implementation of 35 different algorithms for a good comparison. They are evaluated using the following metrics: Kappa and Accuracy.

1 Introduction

Movie of Film communicates ideas, tells stories, shares feelings, etc., through a set of moving images. It grows continuously. The statistics provided in [17] prove the importance of this market. Indeed, in the United States and Canada, Box Office revenue between 1980 and 2020 is 2.09 billion USD, the number of movies released between 2000 and 2020 is about 329, and the number of movie tickets sold between 1980 and 2020 is 224 million tickets.

During the COVID-19, the China exceeded the other country in term of movie revenue of three billion U.S. dollars. This amount includes online ticketing fees [14].

The continuous growth of this market requires offering a help to the viewers to decide if the movie is worth their time or not to save money and time. The selection of the movie is based on the feedback lefts by the previous viewers who already have watched that movies [27]. Therefore, according to [15], 34% of people are not influenced by the critic reviews, 60% are influenced and the rest are without opinion. This study proves the impact on the opinions on the final decision.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 401–411, 2023.
https://doi.org/10.1007/978-3-031-27409-1_36

Because of the huge quantity of the generated feedbacks, reading all of them to make the right decision seems to be a hard task. Hence, using Machine Learning (ML) algorithms to analyze the movie reviews is the appropriate solution. ML is a field of Computer Science in which machines or computers are able to learn without being programmed explicitly [26]. It can help in terms of effectiveness and efficiency. It will automatically classify and shorten the processing time [7].

Concerning the sentiment analysis, it is a concept related to Natural Language Processing (NLP). It is used to determine the feeling of a person in positive or negative comments by analyzing a large numbers of documents [24]. This process can be automated using the machine learns through training and testing the data [4].

In the literature, several works use ML algorithms to analyze the sentiments related to the movies, in order to automate the process of feeling extraction based on the different reviews and classify them into positive and negative. Compared to those works, we used over 35 different algorithms, we evaluated them and we compared them to determine the real best model. We used two different datasets applied in the most of the existing works as benchmarks. Concerning the metrics of the evaluation, we used, Kappa and Accuracy.

This work is organized as following. In Sect. 2, we will summarize some of the existing works that used ML algorithms for sentiment analysis related to movies. Section 3 describes our proposed methodology. Section4 defines the techniques used during the preprocessing step. Section 5 contains a description of the used data set as well as the specificities of the used algorithms and the metrics that we use for the evaluation. In Sect. 6, we compare the results of the evaluation using the two metrics Kappa and Accuracy. In the conclusion, we summarize our work and we give some perspectives as future work.

2 State of the Art

In this section, we summarize some of the existing works that used ML algorithms for movie sentiment analysis.

In [21], the authors use dataset Polarity v2.0 from Cornell movie review dataset. The latter is composed by 1000 documents of negative reviews and 1000 documents of positive reviews in English language. This dataset is used to test KNN (K-Nearest Neighbour) with Information gain features selection in order to determine the best K value. The used algorithms are compared to NB (Naive Bayes), SVM (Support Vector Machine) and RF (Random Forest) algorithms. In [23], the authors focus on the sentiment analysis of movie reviews written in Bangla language, since the latter is the language with the second-highest number of speakers. To achieve this goal, the authors use a dataset collected and labeled manually from publicly available comments and posts from social media websites. They use different ML algorithms for the comparison: SVM, MNB (Multinomial Naive Bayes) and LSTM (Long Short Term Memory). In [25], the authors propose using ML algorithms to classify the reviews related to

movies. They compare NB and RF in term of memory use. They collect data from different web sites like Times of India and Rotten Tomatos. They conclude that RF is better than NB in terms of time and memory to recommend the good movie to users. In [6], the authors study the sentiment analysis of movie reviews in Hindi language. They use, for this purpose, Hindi sentiwordnet which is a dictionary used for finding the polarity of the words. They compare the performance of two algorithms: RF and SVM according to several evaluation metrics. In [29], the authors implement a system that is able to classify sentiments from review documents into two classes: positive and negative. This system uses NB Classifier as ML algorithm and applies it to Movienthusiast. The latter is a movie reviews in Bahasa Indonesia website. The collected dataset is composed by 1201 movie reviews: 783 positive reviews and 418 negative. For the evaluation, the accuracy metric is used. In [27], the authors collect different movie review datasets with different sizes. Then, they apply a set of popular supervised ML algorithms. They compared the performance of the different algorithms using the accuracy. In [2], the authors apply several ML algorithms for sentiment analysis to a set of movie review datasets. They use, then, BNB (Bernoulli Naïve Bayes), DT (Decision Tree), SVM, ME (Maximum Entropy) and MNB. The different algorithms have been compared according to accuracy, recall, precision and F-score metrics. As a conclusion, MNB achieves better accuracy, precision and F-score while SVM has a higher recall. In [3], the authors propose the use of Bayesian Classifier for Sentiment Analysis of Indian Movie Review. They train the model using five feature selection algorithms which are: Chi-square, Info-gain, Gain-Ratio, One-R and relief attribute. For the evaluation, they applied two different metrics: F-Value and False Positive. In [22], the authors used three different ML algorithms to ensure the sentiment analysis. The algorithms are NB, KNN and RF. They are applied to data extracted from IMDb and evaluated using the accuracy metric. In [16], the authors provide an in-depth research of ML methods for sentiment analysis of Czech social media. They propose the use of ME and SVM as ML algorithms. The latter are applied to a dataset created from Facebook with 10,000 posts and labeled manually from two annotators. They use, also, two datasets extracted from online databases of movie and product reviews, whose sentiment labels were derived from the accompanying star ratings from users of the databases. In [20], the authors proposed a system to analyze the sentiment of movie reviews and visualize the result. The used data set is collected from IMDb movie reviews. Concerning the ML algorithm, the authors apply NB classifiers with two types of features: Bag of Words (BoW) and TF-IDF.

3 Methodology

In order to achieve our goal, in this section, we present our adapted methodology. The letter is composed by four steps.

- **Data Collection**: We collected reviewers from two datasets: Cornell Movie Review dataset and Large Movie Review Dataset. They are used in many works as benchmark.

- **Preprocessing**: It is a crucial step since the collected data is not clean. Hence, it is necessary to apply several techniques such as remove stop words, remove repeated characters, etc. The purpose is to get clean data to use it with different ML algorithms later. This step proved its efficiency since the accuracy increases when using preprocessed data.
- **Machine Learning algorithms**: This step is about using the proposed algorithms to classify the data. In this work, we are using over 35 different algorithms.
- **Evaluation**: This step helps us to determine the best model using several metrics which are: kappa and Accuracy.

4 Preprocessing Step

The preprocessing step is crucial to improve the quality of the classification and speed up the classification process. Several standards NLP techniques are applied. Based on many studies such as [1], choosing the appropriate combination may provide significant improvement on classification accuracy rather than enabling or disabling them all. Hence, in our work, we considered: Apply Tokenization, Convert the text to lower case, Remove the HTML tags, Remove special characters including hashtags and punctuations, Remove repeating characters from words, and Apply lemmatization using Part-of-Speech (PoS).

- **Tokenization**: It is one of the most common tasks when working with text. For a given input such a sentence, performing the tokenization is about chopping it up into pieces called tokens. Hence, a token is an instance of sequence of characters that are grouped together as a useful semantic unit for processing [8]. It can be performed by using: split( ) method, or NLTK library *from nltk.tokenize import word_tokenize*, or Keras library *from keras.preprocessing.text import text_to_word_sequence*, or genism *from gensim.utils import tokenize*, etc.
- **Remove repeating characters from words**: It is the case where a word is written by repeating some letters for example instead of "happy", they write "happyyyyyy" or "haaaapppyyyyyy". The different words imply the same thing, but since they are written differently, they are interpreted differently. Hence, the importance of this step to get the same word from different written.
- **Lemmatization**: It converts the word into its base form taking into consideration its context to get the meaningful base form of the word. It is more efficient than the stemming that it about deleting the last few characters leading to incorrect meanings and spelling errors [11].
  This step is performed by many libraries such as NLTK *from nltk.stem import WordNetLemmatizer, treetragger; import treetaggerwrapper, pattern; from pattern.en import lemma*; etc
- **PoS**: In the English language, there are eight parts of speech which are: noun, pronoun, verb, adjective, adverb, preposition, conjunction, and interjection. It is used to indicate how the word functions in meaning as well as grammatically within the sentence. An individual word can function as more than one part of

speech when used in different circumstances. Understanding parts of speech is essential for determining the correct definition of a word when using the dictionary [10].
Python offers many libraries to perform this task. As example, we can mention NLTK *from nltk.tag import pos_tag*.
Using PoS with lemmatization helps to improve the quality and the precession of the lemmatization results. For example, a lemmatiser should map *gone*, *going* and *went* into *go*. In order to achieve its purpose, lemmatisation requires to know about the context of a word, because the process relies on whether the word is a noun, a verb, etc., [9]. Hence, PoS is used as a part of the treatment

- **Remove special characters including hashtags and punctuations**: The punctuation and special characters are used to divide a text into sentences. They are frequent and they can affect the result of any text preprocessing especially the approaches that are based on the frequencies of the words. We can apply the regular expression in this step.
- **Stop words**: It is a list of words that carry little meaningful information, generally they do not add much meaning to the text like "a, an, or, the, etc". Performing this step helps to focus on the important words. In python there are different libraries, for example NLTK *from nltk.corpus import stopwords*, genism *from gensim.parsing.preprocessing import STOPWORDS*, etc. To the default list, it is possible to add new ones.

5 Machine Learning Algorithms

In this section, we will start by defining the structure of the used dataset, then, we will give the specificities of the used algorithms.

5.1 Data Set

In order to evaluate the different algorithms, we used in this work the following datasets (Table 2):

- **Cornell Movie Review dataset**: It is a collection of 1000 positive and 1000 negative processed reviews [5].
- **Large Movie Review Dataset**: It is a dataset for binary sentiment classification. It is composed of 25000 highly polar movie reviews for training, and 25000 for testing [18].

5.2 Metrics of evaluation

- **Accuracy (ACC)**: It is a metric used to evaluate the classification of models by measuring the ratio of correctly predicted instances over the total number of evaluated instances [12]. Its formula is as follow:

Table 1. Datasets description

Dataset name	Year of creation	Year of last update	Number of reviews	Number of classes
Cornell movie review	2002	2004	2000	2
Large movie review	2011	–	50000	2
Total	–	–	52000	2

$$ACC = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{1}$$

For binary classification, the formula of the accuracy is as follow:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- **Kappa (K)**: It is used to measure the degree of agreement between observed and predicted values for a dataset [19]. It can be calculated from the confusion matrix [13] as follow:

$$K = \frac{Total\ Accuracy - Random\ Accuracy}{1 - Random\ Accuracy} \tag{3}$$

where:

$$Total\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Random\ Accuracy = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{Total * Total} \tag{5}$$

where:

- TP: implies the actual value is positive and the model predicts a positive value.
- TN: implies the actual value is negative and the model predicts a negative value.
- FP: implies the actual value is negative but the model predicts a positive value.
- FN: implies the actual value is positive but the model predicts a negative value.

5.3 Machine Learning Algorithms

In this part, we present the specificities of 35 ML algorithms in term of: category, type and used parameters. We present, also, the values of Kappa and Accuracy for each algorithm as presented in the following table (Table 2).

Table 2. The specificities of the used Machine Learning Algorithms

Clf*i*	ML Algorithm	Category	Type	Parameters	Kappa	Accuracy
Clf1	LogisticRegression	Linear	Single	TF-IDF, L2 norm regularization	85.33%	92.67%
Clf2	SGD	Linear	Single	TF-IDF, L2 norm regularization, Hinge loss	82.66%	91.34%
Clf3	PassiveAggressive	Linear	Single	TF-IDF, max iteration = 50	95.70%	97.86%
Clf4	DecisionTree	Tree	Single	TF-IDF, Entropy index, max_depth = 20	62.21%	81.12%
Clf5	LinearSVC	SVM	Single	TF-IDF, L2 norm regularization	94.21%	97.11%
Clf6	NuSVC	SVM	Single	TF-IDF, nu = 0.5, RBFkernel	86.83%	93.42%
Clf7	MultinomialNB	Naïve bayes	Single	TF-IDF, alpha = 1.0	79.45%	89.72%
Clf8	MLP	Neural network	Single	TF-IDF, activation = relu, hidden_layer_sizes = 100	42.02%	71.06%
Clf9	RandomForest	Randomization	Ensemble	TF-IDF, Entropy index, n_estimators = 50	95.68%	97.84%
Clf10	LGBM	Boosting	Ensemble	TF-IDF	94.36%	97.18%
Clf11	GradientBoosting	Boosting	Ensemble	TF-IDF, loss function = deviance, n_estimators = 100	75.87%	87.94%
Clf12	XGB	Boosting	Ensemble	TF-IDF, booster = gbtree	63.01%	81.52%
Clf13	Bagging	Bagging	Ensemble	TF-IDF, Estimator = Decision Tree Classifier	90.10%	95.05%
Clf14	ExtraTrees	Randomization	Ensemble	TF-IDF, n_estimators = 100, criterion = 'gini'	95.60%	97.80%
Clf15	AdaBoost	Boosting	Ensemble	TF-IDF, Estimato = DecisionTree	61.60%	80.81%
Clf16	KNeighbors	Neighbors	Ensemble	TF-IDF, K = 5	69.62%	84.82%
Clf17	KNeighbors	Neighbors	Ensemble	TF-IDF, K = 2	75.8%	87.92%
Clf18	LogisticRegression and DecisionTree	–	Combined	TF-IDF, L2 norm regularization, Entropy index, max_depth = 20	80.12%	90.06%
Clf19	LogisticRegression and RandomForest	–	Combined	TF-IDF, L2 norm regularization, Entropy index, n_estimators = 50	90.36%	95.18%
Clf20	LogisticRegression and LinearSVC	–	Combined	TF-IDF, L2 norm regularization	90.13%	95.06%
Clf21	LogisticRegression and MultinomialNB	–	Combined	TF-IDF, L2 norm regularization, alpha = 1.0	81.81%	90.90%
Clf22	LogisticRegression and	–	Combined	TF-IDF, L2 norm regularization, Entropy	91.05%	95.52%
	PassiveAggressive			index, max_depth = 20		

continued

Table 2. continued

Clf*i*	ML Algorithm	Category	Type	Parameters	Kappa	Accuracy
Clf23	RandomForest and	–	Combined	TF-IDF, Entropy index,	94.59%	97.30%
	DecisionTree			n_estimators = 50, max_depth = 20		
Clf24	DecisionTree and	–	Combined	TF-IDF, Entropy index,	86.18%	93.09%
	LinearSVC			max_depth = 20, L2 norm regularization		
Clf25	DecisionTree and	–	Combined	TF-IDF, Entropy index,	75.79%	87.89%
	MultinomialNB			max_depth = 20, alpha = 1.0		
Clf26	DecisionTree and	–	Combined	TF-IDF, Entropy index,	87.23%	93.62%
	PassiveAggressive			max_depth = 20, max iteration = 50		
Clf27	LinearSVC and	–	Combined	TF-IDF, L2 norm regularization,	94.01%	97.00%
	RandomForest			Entropy index, n_estimators = 50		
Clf28	MultinomialNB and	–	Combined	TF-IDF, alpha = 1.0,	85.65%	92.82%
	RandomForest			Entropy index, n_estimators = 50 m		
Clf29	PassiveAggressive and	–	Combined	TF-IDF , max iteration = 50,	94.92%	97.46%
	RandomForest			Entropy index, n_estimators = 50		
Clf30	LinearSVC and	–	Combined	TF-IDF, L2 norm regularization, alpha = 1.0	85.60%	92.80%
	MultinomialNB					
Clf31	PassiveAggressive and	–	Combined	TF-IDF, max iteration = 50,	95.14%	97.57%
	LinearSVC			L2 norm regularization		
Clf32	PassiveAggressive and	–	Combined	TF-IDF, max iteration = 50, alpha = 1.0	86.13%	93.06%
	MultinomialNB					
Clf33	KNeighbors and	–	Combined	TF-IDF, K = 5, L2 norm regularization	80.97%	90.48%
	LogisticRegression					
Clf34	PassiveAggressive and	–	Combined	TF-IDF , max iteration = 50,	89.76%	94.88%
	MultinomialNB and			alpha = 1.0, L2 norm regularization		
	LogisticRegression					
Clf35	RandomForest and	–	Combined	TF-IDF , Entropy index,	93.98%	96.99%
	KNeighbors and			n_estimators = 50, K = 5,		
	LogisticRegression and LinearSVC			L2 norm regularization		

6 Comparison

In this section, we will compare the different values extracted previously to determine the convenient algorithm according to several metrics. For this purpose, we use the histogram for each metric.

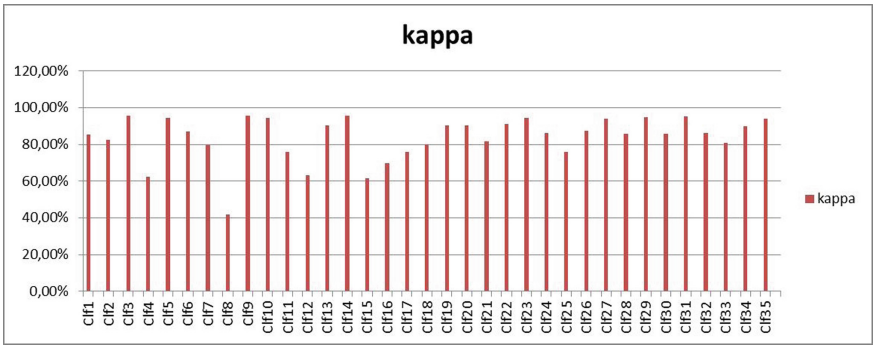


Fig. 1. The values of Kappa for the different machine learning algorithms

6.1 Kappa

Figure 1 presents the values of kappa as applied to the used ML algorithms. Based on the histogram, we can notice that the fourth higher values are for the Clf3 with 95.70%, then Clf9 with 95.68%, then Clf4 with 95.60% and finally Clf31 with 95.14%. For the lower values, they are 42.02% for Clf8, and 62.60% for Clf15.

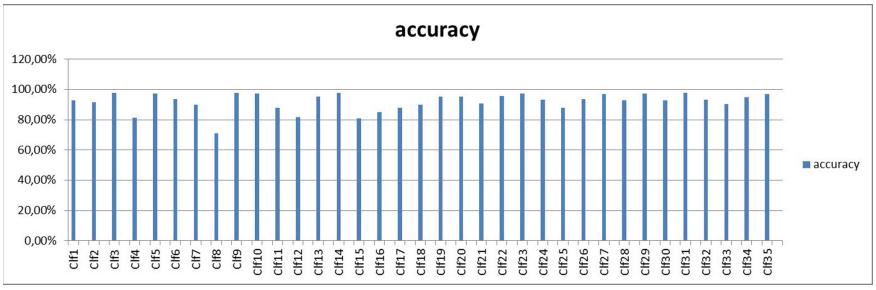


Fig. 2. The values of Accuracy for the different machine learning algorithms

6.2 Accuracy

Figure 2 presents the values of the Accuracy for the different ML algorithms that were evaluated. Based on the generated histogram, we can notice that the four higher values are for Clf3 with 97.86%, then Clf9 with 97.84%, then Clf14 with 97.80% and Clf31 with 97.57%. Concerning the lower values, they belong to Clf8 with 71.06%, and Clf15 with 80.81%.

7 Conclusion

This work focuses on the analysis of sentiment related to the movies. We used for this purpose different ML algorithms to automatically analyze and classify the collected movies feedbacks in order to facilitate the task to the viewer to make his choice.

Compared to the existing works we apply 35 different algorithms that are evaluated using different metrics which are: Kappa and Accuracy for good results.

To conclude, Passive Aggressive algorithm has the highest value according to Kappa and Accuracy with 95.70% and 97.86% respectively.

As perspective, we will extend this work to deal with unbalanced datasets and we will use our solution with other languages like Arabic.

References

1. KursatUysal, A., Gunal, S.: The impact of preprocessing on text classification. Inf. Process. Manag. **50**(1), 104–112 (2014)
2. Rahman , A., Hossen, M.S.: Sentiment analysis on movie review data using machine learning approach. In: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE (2019)
3. Tripathi , A., Trivedi, S.K.: Sentiment analyis of Indian movie review with various feature selection techniques. In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA). IEEE (2016)
4. Devi , B.L., Bai, V., Ramasubbareddy, S., Govinda, K.: Sentiment analysis on movie reviews. In: Emerging Research in Data Engineering Systems and Computer Communications, pp. 321–328. Springer, Singapore (2020)
5. Pang , B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124 (2005)
6. Nanda, C., Dua, M., Nanda, G.: Sentiment analysis of movie reviews in hindi language using machine learning. In: 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE (2018)
7. Sebastiani, F.: Machine learning in automated text categorization. CSUR **34**(1), 1–47 (2002)
8. Patil, G., Galande, V., Kekan, V., Dang, K.: Sentiment analysis using support vector machine. Int. J. Innov. Res. Comput. Commun. Eng. **2**(1), 2607–2612 (2014)
9. https://marcobonzanini.com/2015/01/26/stemming-lemmatisation-and-pos-tagging-with-python-and-nltk/. Cited 10 Sep 2021

10. http://www.butte.edu/departments/cas/tipsheets/grammar/parts_of_speech.html. Cited09 Sep 2021
11. https://www.machinelearningplus.com/nlp/lemmatization-examples-python/. Cited 10 Sep 2021
12. https://deepai.org/machine-learning-glossary-and-terms/gradient-boosting. Cited 10 Sep 2021
13. https://www.standardwisdom.com/2011/12/29/confusion-matrix-another-single-value-metric-kappa-statistic/. Cited 13 Dec 2021
14. https://www.statista.com/statistics/243180/leading-box-office-markets-workdwide-by-revenue/. Cited 19 Dec 2021
15. https://www.statista.com/statistics/682930/movie-critic-reviews-influence/. Cited19 Dec 2021
16. Habernal , I., Ptáček , T., Steinberger, J.: Sentiment analysis in czech social media using supervised machine learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (2013)
17. Navarro, J.G.: Film industry in the U.S.—statistics & fact. https://www.statista.com/topics/964/film/dossierKeyfigures. Cited 19 Dec 2021
18. Andrew, L., Maas, R. E., Daly , R. E., Pham , P. T., Huang, D., Ng , A. Y., Potts, C.: Learning word vectors for sentiment analysis. In: The 49th Annual Meeting of the Association for Computational Linguistics (2011)
19. Al-Rakhami, M.S., Al-Amri, A.M.: Lies kill, facts save: detecting COVID-19 misinformation in Twitter. IEEE Access **8**, 155961–155970 (2020)
20. Adam, N.L., Rosli, N.H., Soh, S.C.: Sentiment analysis on movie review using Naïve Bayes. In: 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS). IEEE (2021)
21. Daeli, N.O.F., Adiwijaya, A.: Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. J. Data Sci. Appl. **3**(1), 1–7 (2020)
22. Baid, P., Gupta, A., Chaplot, N.: Sentiment analysis of movie reviews using machine learning techniques. Int. J. Comput. Appl. **179**(7), 45–49 (2017)
23. Chowdhury, R.R., Hossain, M.S., Hossain, S., Andersson, K.: Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE (2019)
24. Mukherjee, S.: Sentiment Analysis, pp. 113–127. In ML. NET Revealed, Apress, Berkeley, CA (2021)
25. Untawale , T.M., Choudhari, G.: Implementation of sentiment classification of movie reviews by supervised machine learning approaches. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE (2019)
26. Ayodele, T.O.: Types of machine learning algorithms. New Adv. Mach. Learn. **3**, 19–48 (2010)
27. Bharathi , V., Upadhayaya, N.: Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis
28. Madani, Y., Erritali, M., Bouikhalene, B.: Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. Results Phys. 104266 (2021)
29. Nurdiansyah, Y., Bukhori, S., Hidayat, R.: Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method. J. Phys.: Conf. Ser. (2018). IOP Publishing

Fish School Search Algorithm for Constrained Optimization

J. P. M. Alcântara(✉), J. B. Monteiro-Filho, I. M. C. Albuquerque, J. L. Villar-Dias, M. G. P. Lacerda, and F. B. Lima-Neto

University of Pernambuco, Recife PE, Brazil
jpma@ecomp.poli.br

Abstract. In this work we investigate the effectiveness of the application of a niching metaheuristic of the Fish School Search family in solving constrained optimization problems. Sub-swarms are used to allow the achievement of many feasible regions to be exploited in terms of fitness function. The niching approach employed was wFSS, a version of the Fish School Search algorithm devised specifically to deal with multi-modal search spaces. A technique referred as rwFSS was conceived. Tests were performed in seven problems from CEC 2020 and a comparison with other approaches was carried out. Results show that rwFSS can handle some reasonable constrained search spaces and achieve results comparable to two of the CEC 2020 top ranked algorithms on constrained optimization. However, we also observed that the local search operator of wFSS and inherited by rwFSS makes it difficult to find and keep the individuals inside feasible regions when the search space presents a large number of equality constraints.

Keywords: Swarm Intelligence · Constrained Optimization · Weight-based FSS

1 Introduction

According to Koziel and Michalewicz [14], the General Nonlinear-Programming Problem (NLP) Consists in Finding $\boldsymbol{x}$ Such that:

$$Optimize\, f(x),\ x = (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

Where $x \in \mathcal{F} \subseteq S$. Objective function f is defined on the search space $S \subseteq \mathbb{R}^n$ and the set $\mathcal{F} \subseteq S$ defines the feasible region.

Search space S is defined as a sub-space of $\mathbb{R}^n$ and $m \geq 0$ constraints define the feasible space $\mathcal{F} \subseteq S$:

$$g_j \leq 0, for\, j = 1, \ldots, q,$$

$$and\ h_j(x) = 0, for\, j = q+1, \ldots, m = 0.$$

Equality constraints are commonly relaxed and transformed into inequality constraints [23] as: $|h_j(x) - \delta| \leq 0$, where δ is a very small tolerance value.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 412–422, 2023.
https://doi.org/10.1007/978-3-031-27409-1_37

Real-world optimization problems are usually constrained [13]. Hence, many metaheuristics capable of dealing with such problems has been proposed in the literature. Recent approaches include Genetic Algorithms [9, 14, 18], Differential Evolution [5, 11, 19, 28–30], Cultural Algorithm [15, 32], Particle Swarm Optimization [3, 6, 12, 13, 17, 27] and Artificial Bee Colony Optimization [1, 4, 16, 23].

Regarding the approaches applied to tackle constrained search, Mezura-Montes and Coello-Coello [22] present a simplified taxonomy of the common procedures in the literature:

1. Penalty functions—includes a penalization term in the objective function due to some constraint violation. This is a popular and easy-to-implement approach but has the drawback of requiring the adjustment of penalty weights.
2. Decoders—Consists of mapping the feasible region on search spaces where an unconstrained problem will be solved. The high computational cost required is the main disadvantage in its use.
3. Special operators—Mainly in evolutionary algorithms, operators can be designed in a way to prevent the creation of unfeasible individuals.
4. Separation of objective function and constraints—This approach, different from penalty functions, treat the feasible and the infeasible areas separately as two different objective functions (the infeasible area is usually transformed into a constraint violation function).

The Fish School Search (FSS) algorithm, presented originally in 2008 in the work of Bastos-Filho and Lima-Neto et al. [10], is a population-based continuous optimization technique inspired in the behavior of fish schools while looking for food. Each fish in the school represents a solution for a given optimization problem and the algorithm utilizes some key information of each fish to guide the search process to promising regions in the search space as well as avoiding early convergence to local optima.

Ever since the original version of FSS algorithm was developed, several modifications have been made to tackle different types of problems, such as multi-objective optimization [2], multi-solution optimization [20] and binary search [26]. Among those, a novel niching and multi-solution version known as wFSS was proposed [8].

Recently, Vilar-Dias et al. [32] proposed the cwFSS, an FSS family algorithm based on cultural algorithms to incorporate different kind of previous knowledge about the problem on the search process. Even though cwFSS can focus the search on specific areas it does not treat the prior knowledge as constraints, nor it is focused on finding feasible solutions.

To the best of the authors knowledge, the application of FSS in the solution of constrained optimization problems has never been reported before. Hence, in this work, a modification in niching weight based FSS (wFSS) was carried out. The separation of objective function and constraints was applied, and the niching feature was used for the population to find different feasible regions within the search space to be exploited in terms of fitness value.

This paper is organized as follows: Sect. 2 provides an overview of Fish School Search algorithm and its niching version, wFSS. Section 3 introduces the proposed

modifications to employ wFSS in constrained optimization problems. Section 4 presents the tests performed and results achieved.

2 Fish Schooling Inspired Search Procedures

2.1 Fish School Search Algorithm

FSS is a population-based search algorithm inspired in the behaviour of swimming fishes in a school that expands and contracts while looking for food. Each fish in an n-dimensional location in the search space represents a possible solution for the optimization problem. The success of the search process of a fish is measured by its weight, since well-succeeded fish are the ones that have been more successful in finding food.

FSS is composed by feeding and movement operators, the latter being divided into three subcomponents, which are:

Individual component of the movement: Every fish in the school performs a local search looking for promising regions in the search space. It is done as shown in Eq. (1):

$$x_i(t+1) = x_i(t) + rand(-1, 1)step_{ind} \tag{1}$$

where $x_i(t)$ and $x_i(t+1)$ represent the position of fish i before and after the individual movement operator, respectively. $rand(-1, 1)$ is a uniformly distributed random numbers array with the same dimension as $x_i(t)$ and values varying from -1 up to 1. $step_{ind}$ is a parameter that defines the maximum displacement for this movement. The new position $x_i(t+1)$ is only accepted if the fitness of fish i improves with the position change. If it is not the case, $x_i(t)$ remains the same and $x_i(t+1) = x_i(t)$.

Collective-instinctive component of the movement. A weighted average among the displacements of each fish in the school is computed according to Eq. (2):

$$I = \frac{\sum_{i=1}^{N} \Delta x_i \Delta f_i}{\sum_{i=1}^{N} \Delta f_i} \tag{2}$$

The weights of this weighted average are defined by the fitness improvement of teach fish. It means that the fish that experimented a higher improvement will have more influence in the decision of the direction of this collective movement.

After vector I computation, every fish will be encouraged to move according to (3):

$$x_i(t+1) = x_i(t) + I \tag{3}$$

Collective-volitive component of the movement. This operator is used to regulate exploration/exploitation abilities of the school during the search process. First, the barycenter B is calculated based on the position x_i and the weight W_i of each fish:

$$\frac{\sum_{i=1}^{N} x_i(t) W_i(t)}{\sum_{i=1}^{N} W_i(t)} \tag{4}$$

And then, if the total weight given by the sum of the weights of all N fishes in the school $\sum_{i=1}^{N} W_i$ has increased from the last to the current iteration, the fish are attracted to the barycenter according to Eq. (5). If the total school weight has not improved, the fish are spread away from the barycenter according to Eq. (6):

$$x_i(t+1) = x_i(t) - step_{vol} rand(0,1) \frac{x_i(t) - B(t)}{distance(x_i(t), B(t))} \tag{5}$$

$$x_i(t+1) = x_i(t) + step_{vol} rand(0,1) \frac{x_i(t) - B(t)}{distance(x_i(t), B(t))} \tag{6}$$

where $step_{vol}$ defines the maximum step performed in this operator, $distance(x_i(t), B(t))$ is the euclidean distance between fish i and the school's barycenter, $rand(0, 1)$ is an uniformly distributed random numbers array with the same dimension as B and its values varying between 0 and 1.

Besides the movement operators, it was also defined a feeding operator used to update the weights of every fish according to Eq. (7):

$$W_i(t+1) = W_i(t) + \frac{\Delta f_i}{\max(|\Delta f_i|)} \tag{7}$$

where $W_i(t)$ is the weight of a fish i, Δf_i is the fitness variation between the last and the new positions and $max(|\Delta f_i|)$ represents the maximum absolute value of the fitness variations among all the fish in the school.

W is only allowed to vary from 1 up to W_{scale}, which is a user defined parameter of the algorithm. Weights of all fishes are initially set to $W_{scale}/2$. The parameters $step_{ind}$ and $step_{vol}$ decay linearly throughout the search.

2.2 Weight-Based Fish School Search Algorithm

Introduced by Lima-Neto and Lacerda [8], wFSS is a weight-based niching version of FSS intended to provide multiple solutions for multi-modal optimization problems. The niching strategy is based on a new operator called Link Formator. This operator is responsible for defining a leader for each fish, what leads to the formation of sub-schools. This mechanism, performed by each fish individually. Works as follows: a fish a chooses randomly another fish b in the school. If b is heavier than a, then a now has a link with b and follows b (i.e. b leads a). Otherwise, nothing happens. However, if a already has a leader c and the sum of the weights of the followers of a is higher than the weight of b, then a stops following c and starts following b. In each iteration, if a becomes heavier than its leader, the link between them will be broken.

In In addition to Link Formator operator inclusion, some modifications were performed in the components of the movement operators to emphasize leaders influence on sub-swarms. Thus, the displacement vector I of the collective-instinctive component becomes:

$$I = \frac{\Delta x_i \Delta f_i + L \Delta x_l \Delta f_l}{\Delta f_i + \Delta f_l} \tag{8}$$

where L is 1 if fish i has a leader and 0 otherwise. Δx_l and Δf_l are the displacement and fitness variation of the leader of fish i. Furthermore, the influence of vector I in fishes' movements is increased along with iterations. This is represented by $x_i(t+1) = x_i(t)+\rho I$ with $\rho = \frac{current\ iteration}{It_{max}}$. The collective-volitive component of the movement was also modified in a sense that the barycenter is now calculated for each fish with relation to its leader. If the fish does not have a leader, its barycenter will be its current position. This means:

$$B(t) = \frac{x_i(t)w_i(t) + Lx_l(t)w_l}{w_i(t) + Lw_l(t)} \tag{9}$$

3 rwFSS

In this work, a few modifications to the wFSS to make the algorithm able to tackle constrained optimization problems are proposed. Basically, either fitness values or constraint violation are measured for every fish. In the beginning of each iteration, a decision must be done to define whether the fitness function or constraint violation will be used as the objective function.

The decision of which value to use as objective function is done according to the feasible individuals' proportion with relation to whole population. This means that, if the current feasible proportion of the population is higher than threshold σ, the search will be performed using the fitness function as objective function. If that is not the case, constraint violation will be then minimized. The threshold σ has a default value of 50%, but the user can adjust it according to the problem's needs.

The described procedure was applied to divide the search process in two different phases and to allow the algorithm to: phase 1—find many feasible regions; and, phase 2—optimize fitness within feasible regions. The niching feature of wFSS is useful in phase 1 once this feature will make the school able to find many different feasible regions. Moreover, every once the search changes from phase 1 to phase 2, an increase factor τ is applied in the steps of either Individual or Collective-volitive movement operators in order to augment the school mobility in the new phase. The algorithm described will be referred as rwFSS and its pseudocode is defined as follows:

```
1: Initialize user parameters
2: Initialize fish position randomly
3: while Stopping condition is not met do
4:    Calculate fitness for each fish
5:    Calculate constraint violation for each fish
6:    if Feasible proportion ≥ σ then
7:       Define fitness as objective function
8:    else
9:       Define constraint violation as objective function
10:   end if
11:   Run individual movement operator
12:   Run feeding operator
13:   Run collective-instinctive movement operator
14:   Run collective-vollitve movement operator
15: end while
```

The constraint violation measure applied in rwFSS was the same as in the work of Takahama and Sakai [30], as defined by Eq. (10).

$$\phi(x) = \sum_{j=1}^{q} \left| max\{0, g_j(x)\} \right|^p + \sum_{j=q+1}^{m} \left| \{0, h_j(x)\} \right|^p \tag{10}$$

Best fish selection was done using Deb's heuristic [9]:

1. Any feasible solution is preferred to any unfeasible solution.
2. Among two feasible solutions, the one having better fitness function value is preferred.
3. Among two unfeasible solutions, the one having smaller constraint violation is preferred.

Furthermore, the feeding operator version applied was the same as in the work of Monteiro et al. [24], where feeding becomes a normalization of both fitness and constraint violation values, as shown in Eq. (11).

$$W_i = W_{scale} + (1 - W_{scale}) \frac{f_i - \min(f)}{\max(f) - \min(f)} \tag{11}$$

In this equation, f will be the constraint violation values in phase 1 and fitness in phase 2. $min(f)$ And $max(f)$ are the minimum and maximum f values found in all the search process.

It is important to highlight that the normalization applied in Eq. (11) makes $max(f) \Rightarrow 1$ and $min(f) \Rightarrow W_{scale}$ once this equation is applied for minimization of both fitness function and constraint violation.

4 Experiments

To evaluate the proposed algorithms on search spaces with various constraints, a set of constrained optimization problems defined at CEC 2020 [21] have been solved.

The chosen CEC 2020 problems, as well as their features, are presented on Table 1. The problems selected to be included in the test set cover different feasible region's complexity, i.e., different combinations of equality and inequality constraints. The best feasible fitness indicates the best possible fitness result within a feasible region.

Table 1. Chosen CEC 2020's problems.

Problem	Dimension	Maximum Fitness Evaluations	Number of constraints		Best Feasible Fitness
			E	I	
RC01	9	100000	8	0	189.31
RC02	11	200000	9	0	7049
RC04	6	100000	4	1	0.38
RC08	2	100000	0	2	2
RC09	3	100000	1	1	2.55
RC12	7	100000	0	9	2.92
RC15	7	100000	0	7	2990

For the RC08, RC09, RC12 and RC15 the feasible threshold (σ) was set to 40%. Due to the very restricted feasible regions on functions RC01, RC02 and RC04 and the randomness of the rwFSS local search operator, a higher feasible proportion (σ) of 60% was chosen to focus the search on phase 1 and prevent feasible fishes to step out the feasible regions. rwFSS include the Stagnation Avoidance Routine [22] within the Individual movement operator, with α set to decay exponentially: $\alpha = 0.8e^{-0.007*t}$, where t is the current iteration.

Table 2 presents the results obtained in 25 runs of the rwFSS and two of the CEC 2020's top ranked algorithms on constrained optimization, enMODE [33] and BP-ϵMAg-ES [34] along with the p-value of the Wilcoxon Rank-Sum test. In all tests, the number of iterations has been set to the maximum number of fitness evaluations (max FEs) for each function.

Table 2 shows that the proposed algorithm managed to find feasible solutions in all runs for problems RC04, RC08, RC09, RC12 and RC15, which are those containing little or no equality constraints. On these functions, rwFSS found solutions comparable to the chosen CEC 2020's competitors. For RC01 and RC02, due to the presence of a considerable number of equality constraints, rwFSS got stuck in unfeasible regions. Despite not providing feasible solutions, on Fig. 1 shows that rwFSS can achieve regions with lower constraint violation values on fewer iterations compared to enMODE, making it suitable for problems with flexible constraints and that requires fewer fitness function calls.

The struggle of rwFSS to tackle some heavily constrained problems is related to the search mechanisms employed to the original FSS. The individual movement operator is based on a local search performed with a random jump. Therefore, in situations in which the feasible regions are very small, random jumps may neither guarantee that a fish can reach this region in phase 1 nor guarantee that a fish that has already reached it will remain there.

Table 2. CEC 2020's problems results.

		Feasible rate (%)	Mean Const. Violation	Best Const. Violation	Best Fitness	p-value
RC01	rwFSS	0	511.13	29.28	368.57	1E−09
	BP-ϵMAg-ES	100	0.00	0.00	189.32	
	EnMode	100	0.00	0.00	189.31	
RC02	rwFSS	0	120.91	31.67	16627.01	1E−09
	BP-ϵMAg-ES	100	0.00	0.00	7049.00	
	EnMode	100	0.00	0.00	7049.00	
RC04	rwFSS	100	0.00	0.00	0.38	1E−09
	BP-ϵMAg-ES	100	0.00	0.00	0.38	
	EnMode	100	0.00	0.00	0.38	
RC08	rwFSS	100	0.00	0.00	2,00	1
	BP-ϵMAg-ES	100	0,00	0.00	2.00	
	EnMode	100	0,00	0.00	2.00	
RC09	rwFSS	100	0.00	0.00	2.55	1
	BP-ϵMAg-ES	100	0.00	0.00	2.55	
	EnMode	100	0.00	0.00	2.55	
RC12	rwFSS	100	0.00	0.00	2.92	1
	BP-ϵMAg-ES	100	0.00	0.00	2.92	
	EnMode	100	0.00	0.00	2.92	
RC15	rwFSS	100	0.00	0,00	2998.35	1E−09
	BP-ϵMAg-ES	100	0.00	0.00	2994.40	
	EnMode	100	0.00	0.00	2990.00	

5 Conclusion

Several problems within Industry and Academia are constrained. Therefore, many approaches try to employ metaheuristic procedures to efficiently solve these problems. Different search strategies were developed and applied in both Evolutionary Computation and Swarm Intelligence techniques.

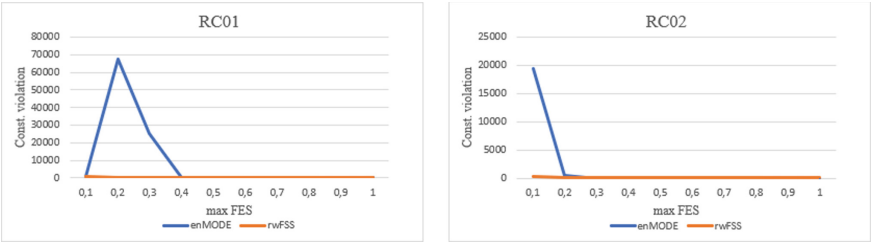


Fig. 1. Constraint violation comparison between enMODE and rwFSS over iterations for RC01 and RC02.

The first contribution in this work regards the proposal of a new approach to tackle constrained optimization tasks: the separation of objective function and constraint violation by the division of the search process in two phases. Phase 1 is intended to make the swarm to find many different feasible regions and, after that, phase 2 takes place to exploit the feasible regions in terms of fitness values.

This strategy, mainly in phase 1, requires a niching able algorithm. Thus, we selected wFSS, the multi-modal version of the Fish School Search algorithm, to be employed as base algorithm. Hence, we conceived a variation of wFSS named rwFSS embedding the division strategy.

To evaluate the proposed technique, seven problems from CEC 2020 have been solved. Results show that rwFSS can solve many hard constrained optimization problems. However, in some cases, specifically in problems containing feasible regions presenting geometric conditions in which the widths in some directions are much higher than in others, the algorithm's local search procedure brings difficulties for rwFSS to keep solutions feasible once phase 1 finishes. Such a known issue will be addressed in a future work. Even so, rwFSS also managed to achieve less unfeasible solutions within a significant smaller number of iterations compared to the CEC 2020 winner.

According to what has been found in the experiments presented in this work, the proposed strategy of dividing the search process in two different phases and apply a niching swarm optimization technique to find many feasible regions in phase 1 is an interesting approach to be explored. In future works, improvements in rwFSS could include adjustments on the sub-swarm's link formation to avoid that unfeasible fishes drag the subswarms to unfeasible regions and the implementation of a strategy to tackle equality constraints gradually.

References

1. Akay, B., Karaboga, D.: Artificial bee colony algorithm for large-scale problems and engineering design optimization. J. Intell. Manuf. **23**(4), 1001–1014 (2012)
2. Bastos-Filho, C.J.A., Guimarães, A.C.S.: Multi-objective fish school search. Int. J. Swarm Intell. Res. **6**(1), 23–40 (2015)
3. Bonyadi, M., Li, X., Michalewicz, Z.: A hybrid particle swarm with velocity mutation for constraint optimization problems. In: Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference—GECCO '13, p. 1 (2013)

4. Brajevic, I., Tuba, M.: An upgraded artificial bee colony (ABC) algorithm for constrained optimization problems. J. Intell. Manuf. **24**(4), 729–740 (2013)
5. Brest, J.: Constrained real-parameter optimization with ϵ-self-adaptive differential evolution. Stud. Comput. Intell. **198**, 73–93 (2009)
6. Campos, M., Krohling, R.A.: Hierarchical bare bones particle swarm for solving constrained optimization problems. In: 2013 IEEE Congress on Evolutionary Computation, CEC 2013, pp. 805–812 (2013)
7. Chootinan, P., Chen, A.: Constraint handling in genetic algorithms using a gradient-based repair method. Comput. Oper. Res. **33**(8), 2263–2281 (2006)
8. De Lima Neto, F.B., De Lacerda, M.G.P.: Multimodal fish school search algorithms based on local information for school splitting. In: Proceedings—1st BRICS Countries Congress on Computational Intelligence, BRICS-CCI 2013, pp. 158–165 (2013)
9. Deb, K.: An efficient constraint handling method for genetic algorithms. Comput. Methods Appl. Mech. Eng. **186**(2–4), 311–338 (2000)
10. Filho, C.J.a.B., Neto, F.B.D.L., Lins, A.J.C.C., Nascimento, A.I.S., Lima, M.P.: A novel search algorithm based on fish school behavior. In: Conference Proceedings—IEEE International Conference on Systems, Man and Cybernetics, pp. 2646–2651 (2008)
11. Hamza, N., Essam, D., Sarker, R.: Constraint consensus mutation based differential evolution for constrained optimization. IEEE Trans. Evol. Comput. (c):1–1 (2015)
12. Hu, X., Eberhart, R.: Solving constrained nonlinear optimization problems with particle swarm optimization. Optimization **2**(1), 1677–1681 (2002)
13. Jordehi, A.R.: A review on constraint handling strategies in particle swarm optimisation. Neural Comput. Appl. **26**(6), 1265–1275 (2015). https://doi.org/10.1007/s00521-014-1808-5
14. Koziel, S., Michalewicz, Z.: Evolutionary algorithms, homomorphous mappings, and constrained parameter optimization. Evol. Comput. **7**(1), 19–44 (1999)
15. Landa Becerra, R., Coello, C.A.C.: Cultured differential evolution for constrained optimization. Comput. Methods Appl. Mech. Eng. **195**(33–36), 4303–4322 (2006)
16. Li, X., Yin, M.: Self-adaptive constrained artificial bee colony for constrained numerical optimization. Neural Comput. Appl. **24**(3–4), 723–734 (2012). https://doi.org/10.1007/s00521-012-1285-7
17. Liang, J.J., Zhigang, S., Zhihui, L.: Coevolutionary comprehensive learning particle swarm optimizer. In: 2010 IEEE World Congress on Computational Intelligence, WCCI 2010—2010 IEEE Congress on Evolutionary Computation, CEC 2010, 450001(2):1–8 (2010)
18. Lin, C.-H.: A rough penalty genetic algorithm for constrained optimization. Inf. Sci. **241**, 119–137 (2013)
19. Liu, J., Teo, K.L., Wang, X., Wu, C.: An exact penalty function-based differential search algorithm for constrained global optimization. Soft. Comput. **20**(4), 1305–1313 (2015). https://doi.org/10.1007/s00500-015-1588-6
20. Madeiro, S.S., De Lima-Neto, F.B., Bastos-Filho, C.J.A., Do Nascimento Figueiredo, E.M.: Density as the segregation mechanism in fish school search for multimodal optimization problems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6729 LNCS(PART 2), 563–572 (2011)
21. Kumar, A., Wu, G., Ali, M., Mallipeddi, R., Suganthan, P.N., Das, S.: A test-suite of non-convex constrained optimization problems from the real-world and some baseline results. Swarm Evol. Comput. **56**, 100693 (2020)
22. Mezura-Montes, E., Coello Coello, C.A.: Constraint-handling in nature-inspired numerical optimization: Past, present and future. Swarm Evolut. Comput. **1**(4), 173–194 (2011)
23. Mezura-Montes, E., Velez-Koeppel, R.E.: Elitist artificial bee colony for constrained real-parameter optimization. In: 2010 IEEE World Congress on Computational Intelligence, WCCI 2010—2010 IEEE Congress on Evolutionary Computation, CEC 2010 (2010)

24. Monteiro, J.B., Albuquerque, I.M.C., Neto, F.B.L., Ferreira, F.V.S.: Comparison on novel fish school search approaches. In: 16th International Conference on Intelligent Systems Design and Applications (2016)
25. Monteiro, J.B., Albuquerque, I.M.C., Neto, F.B.L., Ferreira, F.V.S.: Optimizing multi-plateau functions with FSS-SAR (Stagnation Avoidance Routine). In: IEEE Symposium Series on Computational Intelligence (2016)
26. Sargo, J.A.G., Vieira, S.M., Sousa, J.M.C., Filho, C.J.A.B.: Binary Fish School Search applied to feature selection: application to ICU readmissions. In: IEEE International Conference on Fuzzy Systems, pp. 1366–1373 (2014)
27. Takahama, T., Sakai, S.: Contrained optimization by ϵ constrained swarm optimizer with ϵ-level control. In 4th IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology, pp. 1019–1029 (2005)
28. Takahama, T., Sakai, S.: Constrained optimization by the constrained differential evolution with gradient-based mutation and feasible elites. In: IEEE Congress on Evolution Computation, pp. 1–8 (2006)
29. Takahama, T., Sakai, S.: Solving difficult constrained optimization problems by the ϵ constrained differential evolution with gradient-based mutation. Stud. Comput. Intell. **198**, 51–72 (2009)
30. Takahama, T., Sakai, S.: Constrained optimization by the ϵ constrained differential evolution with an archive and gradient-based mutation. IEEE Congress Evol. Comput. **1**, 1–8 (2010)
31. Takahama, T., Sakai, S., Iwane, N.: Constrained optimization by the ϵ constrained hybrid algorithm of particle swarm optimization and genetic algorithm. Adv. Artif. Intell. **3809**(1), 389–400 (2005)
32. Vilar-Dias, J.L., Galindo, M.A.S., Lima-Neto, F.B.: Cultural weight-based fish school search: a flexible optimization algorithm for engineering. In: 2021 IEEE Congress on Evolutionary Computation (CEC), pp. 2370–2376 (2021). https://doi.org/10.1109/CEC45853.2021.9504779
33. Sallam, K.M., Elsayed, S.M., Chakrabortty, R.K., Ryan, M.J.: Multi-operator differential evolution algorithm for solving real-world constrained optimization problems. In: 2020 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2020). https://doi.org/10.1109/CEC48606.2020.9185722
34. Hellwig, M., Beyer, H. -G.: A modified matrix adaptation evolution strategy with restarts for constrained real-world problems. In: 2020 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2020). https://doi.org/10.1109/CEC48606.2020.9185566



Text Mining-Based Author Profiling: Literature Review, Trends and Challenges

Fethi Fkih[1,2(✉)] and Delel Rhouma[1,2]

[1] Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
fethi.fkih@gmail.com, f.fki@qu.edu.sa
[2] MARS Research Lab LR 17ES05, University of Sousse, Sousse, Tunisia

Abstract. Author profiling (AP) is a very interesting research field that can be involved in many application, such as, Information Retrieval, social network security, Recommender System, etc. This paper presents an in-depth literature review on Author Profiling (AP) techniques, concentrating on text mining approaches. Text Mining-based APs techniques can be categorized into three main classes: Linguistic-based AP, Statistical-based AP and a hybrid approach that combines both linguistic and statistic methods. Also, literature review shows the extensive use of classical Machine Learning and Deep Learning in this field. Besides, we perform in this paper a discussion of the presented models and the main challenges and trends in the AP domain.

Keywords: Author profiling · Text Mining · Machine Learning

1 Introduction

The rapid expansion of data on social media platforms (Facebook, Twitter, blogs… etc.) presents a big challenge for Author Profiling (AP) systems. In fact, it's a difficult task to know who write the posts in these platforms. AP aims to identify the demographic (age, gender, region, level of education) and psychological (personality, mental health) properties of a text's author, mainly users content produced in social media, by using specific techniques. However, we can describe the author profiling as the possibility to know the characteristics of people based on what they write. To infer the gender, age, native language, language variety of a user, or even when the user is lying, simply by analyzing her/his messages, opens up a wide range of security possibilities [1]. The AP techniques can be used in many applications in the fields of forensics, protection, marketing, fake profile recognition on online social networking sites, spam senders, etc. On other hand, the AP domain has to pass many challenges such as extracting features from text using text mining tools, the datasets availability, improving the performance of APs techniques, etc.

In this paper, we provide an in-depth literature review on main AP approaches. Besides, we present the most important challenges and trends in this field. The paper is organized as follows: in Sect. 2 we supply an overview on main AP approaches. In

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 423–431, 2023.
https://doi.org/10.1007/978-3-031-27409-1_38

Sect. 3, we provide a summary and a discussion. Finally, in Sect. 4 we conclude our paper.

2 Text Mining-Based Author Profiling Main Approaches

During the last decade, many researches have been launched in the author profiling field. AP evolution has coincided with the revolution of social media sites (Facebook, Twitter, blogs… etc.) then it is considered as an interesting topic for researchers in computer science. The method of turning unstructured text into relevant and actionable data is called text mining, also known as text analysis [2–5]. Via detecting topics, trends, and keywords, text mining enables you to gain valuable insights without having to manually go through all your data. They get the advantages of text mining techniques to perform the AP, as shown in Fig. 1. Text mining models for AP task that mentioned in previous research can be classified into three main approaches: statistical, linguistic and hybrid (as shown in Fig. 2).

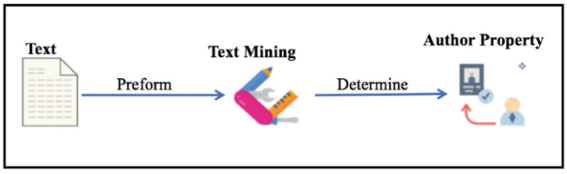


Fig. 1. Author profiling based on Text mining.

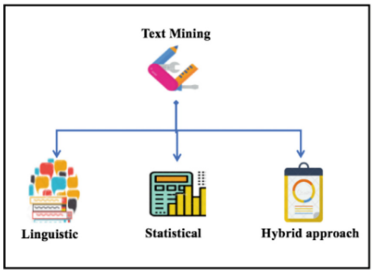


Fig. 2. Text mining main approaches.

2.1 Linguistic Approach

The linguistic approach aims to extract linguistic features using grammatical and syntactic knowledge. It is the knowledge of human languages grammar, syntax, semantics, rules and structure. There are two kind techniques that are common used to extract text features for AP, that are lexical-based and stylistic-based techniques.

Duong et al. [6] identified the age, gender and location of author in Vietnamese forum posts. They comparing the performance of the detection model depend on the stylometry features and content-based features. They applied Decision Tree, Bayes Networks

and Support Vector Machine learning methods. Their result showed that the features work well on short and free style text. Also, showed the content-based feature provide better result than stylometric feature. In [7], the authors showed their system working on Author Profiling task by PAN-2014 corpus. They identified the age and gender of author from (tweets, blogs, social media and hotel review) data sets. Their training data is provided by the PAN organizers. They extracted features from a text document by proceed different Natural Language Processing techniques and using the Random Forest classifier to determining the personal traits (age and gender) of author.

The authors of [8] used 60 textual meta-attributes to identify linguistic gender expression in tweets written in Portuguese. In order to identify the author's gender using three different machine-learning algorithms (BFTree, MNB, and SVM), short-length characters, grammar, words, structure, and morphology, multi-genre, content-free texts posted on Twitter are taken into account. The impact of the suggested meta-attributes on this process is also examined. To determine which of these traits performs best in the categorization of a corpus with neutral messages present, Chi-Square and information gain techniques are used in selection. Researchers in [9] built their system based on simple content-based features to identify authors age, gender and other personality traits. They used supervised machine learnings algorithm in PAN-2015 corpus. Several Machine Learning techniques (SVM, Random Forest and Naive Bayes) applied to train the models after, content-based features were extracted from the text. They showed the efficiency of content-based features approach in predicating the authors traits from anonymous text.

The work described in [10] focused on AP task in Urdu language with Facebook platform. They used the Urdu construction of sentences that written by English alphabets which transforms the language properties of the text. They looked at how existing AP approaches for multilingual texts that include English and Roman Urdu performed, primarily for the purposes of identifying gender and age. They created a multilingual corpus, created a bilingual dictionary by hand to translate Roman Urdu words into English, and model existing AP techniques using 64 different stylistic features for identifying gender and age on translated and multilingual corpora. Word and character N-grams, 11 lexical word-based features, 47 lexical character-based features, and 6 vocabulary richness measures are some of these features. They analyze and evaluate the behavior of their model. According to their analysis, content-based methods outperform stylistic methods for tasks like gender and age recognition as well as multilingual translation. Current author profiling techniques can be used for both multilingual and monolingual text (corpus obtained after translating multilingual corpus using bilingual dictionary). The authors in [11] presented novel approach for profiling the author of an anonymous text in English languages. Authors use machine learning approaches to obtain the best classification. They proposed a framework for age prediction based on advanced Bayesian networks to overcome the problem that mentioned in previously Bayesian networks work which is the Bayesian naïve classifier do not yield the best results. They relying their experiment on an English PAN2013 corpus. The results obtained are comparable to those obtained by the methods of the best state of the art. They found that the lexical classes do not enough for obtained good result for AP task. The authors in [12] addressed the task of user classification in social media especially in Twitter.

They inferred the values of user properties automatically. They used a machine learning technique that relies on a rich set of language attributes that are derived from such user data. On three tasks including the detection of political affiliation, ethnicity identification, and affinity qualities for a specific firm, they obtained an excellent experimental result. Miura et al. [13], prepared neural network models for the author profiling task of PAN2017. The NN have shown a good result in NLP task. The proposed system integrates character and word information with multiple Neural Network layers. They identified the gender in four languages corpora data sets (English, Spanish, Portuguese and Arabic).

2.2 Statistical Approach

Statistical approach considered a text as a bag of words. To extract relevant knowledge (called n-grams or co-occurrences) from textual data, the statistical approach is based on a frequency counting of words within the text [14, 15]. Castillo et al. [16] presented an approach to solve the author profiling especially determining age, gender and personality traits. The main focus of the approach is to build and enrich a co-occurrence graph using the theory of relation prediction to find the profile of an author using a technique of graph similarity. Given the identical training and testing resources, they used their method on the PAN2015 author profiling task's English language portion to obtain results that were competitive and not far off from the best results ever reported. After conducting tests, they came to the conclusion that adding additional edges to a graph representation based on the topological neighborhood of words can be a useful tool for identifying patterns in texts that originate from social media. Also, using a chart similarity provides a novel way to examine whether texts related to a particular group or personality characteristic are identical to an author's writing style.

Maharjan et al. [17] introduced a system used MapReduce (distributive computing techniques) programming paradigm for most parts of the training process, which makes their system fast. Their system used word n-grams including stop words, punctuations and emoticons as features and TF-IDF (term frequency inverse document frequency) as the weighing scheme. These are fed to the classification of logistic regression which predicts the authors' age and gender. The authors in [18], identified the gender and age of author from the SMS messages using ML approach. They used a technique of statistical feature selection to pick features that contribute significantly to the classifications of gender and age. They have performed a paired t-test to show that statistically significant improvement in performance. The evaluation done by using MAPonSMS@FIRE2018 shared task data set. Werlen [19] used SVM and Linear Discriminant Analysis (LDA) classifiers to present the AP approach. We examined the characteristics obtained from dictionaries of Linguistic Inquired and Word Count (LIWC). These are category-by-category frequencies of use of words that give an overview of how the author writes and what he/she is talking about. These are important features to differentiate gender, age-group, and personality, according to the experimental results.

The writers of [20] investigated an experiment including cross-genre analysis and author profiling of tweets in English and Spanish. They classified age and gender using the Support Vector Machine method (APtask). The genres evaluation originates from blogs, hotel reviews, earlier-collected tweets, and other social media platforms, while

their training set was compiled from tweets. The TF-IDF and word average were two feature extracting algorithms that were compared. The results show that in the majority of cross-genre problems for age and gender, employing average of word vectors surpasses TF-IDF. Sarra et al. in [21] proposed a purely statistic model for detecting bots in English and Spanish corpora. In fact, they used a Random Forest model with 17 stylometry-based features. The proposed model provided a good result. In the same context, the same authors in [22] applied their approach for the gender identification task for English and Spanish languages. Also, the model provided good finding when applied on PAN2019 corpus.

2.3 Hybrid Approach

Hybrid approach is a combination of the two previous approaches. It takes the advantage of statistical approach and linguistic approach. In this context, the authors in [23] proposed approach of solving the PAN2016 Author Profiling Task through social media posts, this includes classifying the gender and age of users. On the TF-IDF and verbosity features, they applied SVM classifiers and Neural Networks. Results indicated that while SVM classifiers perform well for English datasets, Neural Networks outperform them for Dutch and Spanish datasets. SVM classifiers perform better for English datasets, and Neural Networks perform better for Dutch and Spanish datasets, according to their findings. The task of automatically identifying authorship from anonymous data provided by the PAN2013 was presented in the work detailed in [24]. The authors' age and gender were determined by linguistic and stylistic features.

Different word lists are generated to determine each document's frequencies. We have created the list of stop terms, smiley list, lists of positive and negative words, etc. for creating the feature vector. A machine learning algorithm was used to classify the profile of the authors. Weka1 tool's Decision tree was used for the classification task. The authors in [25] performed gender identification from multimodal Twitter data which provided by the organizers of AP task2 at PAN2018. They interested on how everyday language reflects on social and personal choices. The organizers provided tweets and photos of users using Arabic, English, and Spanish languages. We established many significant textual features for the English dataset, including embedded words and stylistic features. To extract captions from images, an image captioning system was used and the textual features above were identified from the captions. On other side, Arabic and Spanish databases used a language-independent approach. After gathering term frequency-inverse document frequency (TF-IDF) of unigrams, singular value decomposition (SVD) was used to the term frequency-inverse document frequency (TF-IDF) vectors to reduce sparsity. To obtain the final feature vectors, latent semantic analysis (LSA) was used to the reduced vectors. For categorization, Support Vector Machine (SVM) was used.

The authors in [17], identified the gender of author in Russian languages text. They extracted the linguistic inquiry, word count, TF-IDF and n-grams in order to applied the conventional ML (SVM, decision tree, gradient boosting) and Conventional Neural Network (CNN). They used files from the RusProfiling and RusPersonality corpora as well as text from the gender limitation corpus to enrich their data (training and testing set). In the same context, the authors in [26], presented their working in author profiling task at PAN2017. They identified the gender of variety languages (English, Spanish,

Portuguese, and Arabic). They used character n-grams, word n-grams. Also, they use non-textual features which are binary, raw frequency, normalized frequency, log-entropy weighting, frequency threshold values and TF-ID. They experimented various ML algorithms that are lib-linear and lib SVM implementations of Support Vector Machines (SVM), multinomial naïve Bayes, ensemble classifier, meta-classifiers.

Poulston et al. [27], in conjunction with Support Vector Machines, they used two sets of text-based features, n-grams and topic models to predict gender, age and personality ratings. They applied their system on four different languages (Italian, English, Dutch and Spanish) corpora that provided by PAN2015. Every corpus was made up of sets of tweets from various Twitter users whose gender, age, and personality score had been determined. They demonstrate the usefulness of topic models and n-grams in a variety of languages.

Authors in [28] have proposed a Twitter user profiling classifier that takes advantage of deep learning techniques (Deep learning is a kind of machine learning algorithms that slowly extracts higher-level features from the input) to automatically produce user features that are suitable for AP tasks and that are able to combat covariance shift problems due to differences in training and test sets in data distribution. The designed system can achieve very interesting results of accuracy in both English and Spanish languages. Sarra et al. in [29] used a convolutional Neural Network (CNN) model for bot and gender identification in Tweeter. In fact, the extracted semantic (topics) and stylistic features from tweets content and they learned them to the CNN. The test of the proposed approach confirms its performance.

3 Summary and Discussion

As mentioned previously, the researchers prefer to use ML techniques and tools after text mining to provide the AP task with high performance. In particularly, they used supervised learning classification more than unsupervised learning clustering. Also, we can remark that SVMs were used mostly for linguistic (content-based, stylistic based) features that can be explained with its high performance for this kind of task. Many researchers in this field mainly concentrate on one form of approach whether linguistic or statistical. For the task of AP, age and gender property are the most identified. Furthermore, we can observe that researchers focus on studying some (English and Spanish languages, for instance) more than other languages. This observation can be explained with the wide existence of linguistic and semantic resources (ontologies, thesauri, dictionaries, semantic networks, etc..) for these languages. Whereas, this advantage is not available for many other languages, such as Arabic, where researchers still in the phase of preparing and building linguistic resources and tools.

In Table 1, we summarize the main characteristics of the models mentioned in the state of art section. We highlight in this table used approaches, the features extracted from the text, the learning type (supervised or not), the languages handled in the data set, and the proprieties identified by the author profiling.

Table1. Main approaches of the Author Profiling task

Model	Features type	Language	Proprieties of Author
Decision Tree, Bayes Networks and SVM [6]	Stylistic	Vietnamese	Age, Gender and Location
BFTree, MNB, SVM [8]	60 textual meta-attributes	Portuguese	Gender
Random Forest, SVM and Naive Bayes [7, 9]	Content-based	English	Gender, Age and other Personality Traits
Co-occurrence graph [16]	Graph similarity	English	Gender, Age and other Personality Traits
Random Forest [21]	Stylistic	English Spanish	Gender Bot/Human
Latent semantic analysis (LSA), SVM [25]	Stylistic	Arabic, English, and Spanish	Gender
Conventional Neural Network (CNN) [17]	Linguistic Inquiry and Word Count, TF-IDF and n-grams	Russian	Gender
MapRduce [17]	Word n-grams including stop words,	English	Age Gender
Deep learning techniques [28]	Automatically produce user features	English Spanish	Bot/Human Gender
Advanced Bayesian networks [11]	Lexical feature	English	Age
SVM classifiers and neural networks [23]	TF-IDF and verbosity features	Dutch, English and Spanish	Age Gender
Decision tree [24]	Linguistic and stylistic	English	Age Gender
Twitter user classification [12]	Linguistic	English	Political Affiliation, Ethnicity and Affinity for a particular business
Word embedding averages and SVMs [20]	Statistical	English Spanish	Age Gender
Support Vector Machines [27]	Statistical	Italian, English, Dutch and Spanish	Gender Age Personality Scores

(*continued*)

Table1. (*continued*)

Model	Features type	Language	Proprieties of Author
Convolutional Neural Network [29]	Statistical and semantic	English Spanish	Bot/Human Gender
Various ML algorithms [26]	Statistical	English, Spanish, Portuguese, and Arabic)	Gender
Neural Network Models [22]	NLP	English, Spanish, Portuguese and Arabic	Gender

4 Conclusion

In this work, we have provided an overview on the most important approaches in the Author Profiling field. In fact, the mentioned approaches were classified into two main categories: Text Mining-based approaches and Machine Learning-based approaches. For each approach, we have supplied its fundamental foundation and the targeted information (gender, age, etc.).

Moreover, we have presented the main challenges that should be overcome to improve the efficiency of future Author Profiling systems. Also, this work reveals that the most important factor for improving the performance of the AP systems is to improve linguistic and semantic resources and tools, accordingly.

References

1. HaCohen-Kerner, Y.: Survey on profiling age and gender of text authors. Expert Syst. Appl. **199** (2022)
2. Fkih, F., Nazih Omri, M.: Information retrieval from unstructured web text document based on automatic learning of the threshold. Int. J. Inf. Retr. Res. **2**(4), 12–30 (2012)
3. Fkih, F., Omri, M.N.: Hidden data states-based complex terminology extraction from textual web data model. Appl. Intell. **50**(6), 1813–1831 (2020). https://doi.org/10.1007/s10489-019-01568-4
4. Fkih, F., Nazih Omri, M.: Information retrieval from unstructured web text document based on automatic learning of the threshold. Int. J. Inf. Retr. Res. (IJIRR) **2**(4), (2012)
5. Fkih, F., Nazih Omri, M.: Hybridization of an index based on concept Lattice with a terminology extraction model for semantic information retrieval guided by WordNet. In: Abraham, A., Haqiq, A., Alimi, A., Mezzour, G., Rokbani, N., Muda, A. (eds.) Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016). HIS 2016. Advances in Intelligent Systems and Computing, Vol. 552. Springer, Cham (2017)
6. Duong, D.T., Pham, S.B., Tan, H.: Using content-based features for author profiling of Vietnamese forum posts. In: Recent Developments in Intelligent Information and Database Systems, pp. 287–296. Springer, Cham (2016)
7. Surendran, K., Gressel, G., Thara, S., Hrudya, P., Ashok, A., Poornachandran, P.: Ensemble learning approach for author profiling. In: Proceedings of CLEF (2014)

8. Filho, L., Ahirton Batista, J., Pasti, R., Nunes de Castro, L.: Gender classification of twitter data based on textual meta-attributes extraction. In: New Advances in Information Systems and Technologies. Springer, Cham, pp. 1025–1034 (2016)
9. Najib, F., Arshad Cheema, W., Adeel Nawab, R.M.: Author's Traits Prediction on Twitter Data using Content Based Approach. CLEF (Working Notes) (2015)
10. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on Facebook. Inf. Process. Manag. **53**(4), 886–904 (2017)
11. Mechti, S., Jaoua, M., Faiz, R., Bouhamed, H., Belguith, L.H.: Author Profiling: Age Prediction Based on Advanced Bayesian Networks. Res. Comput. Sci. **110,** 129–137 (2016)
12. Pennacchiotti, M., Popescu, A.-M.: A machine learning approach to twitter user classification. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
13. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T.: Author Profiling with Word+ Character Neural Attention Network. CLEF (Working Notes) (2017)
14. Fkih, F., Nazih Omri, M.: A statistical classifier based Markov chain for complex terms filtration. In: Proceedings of the International Conference on Web Informations and Technologies, ICWIT 2013, pp. 175–184, Hammamet, Tunisia, (2013)
15. Fkih, F., Nazih Omri, M.: Estimation of a priori decision threshold for collocations extraction: an empirical study. Int. J. Inf. Technol. Web Eng. (IJITWE) **8**(3) (2013)
16. Castillo, E., Cervantes, O., Vilariño, D.: Author profiling using a graph enrichment approach. J. Intell. Fuzzy Syst. **34**(5), 3003–3014 (2018)
17. Sboev, A., Moloshnikov, I., Gudovskikh, D., Selivanov, A., Rybka, R., Litvinova, T.: Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. Procedia Comput. Sci. **123**, 417–423 (2018)
18. Thenmozhi, D., Kalaivani, A., Aravindan, C.: Multi-lingual Author Profiling on SMS Messages using Machine Learning Approach with Statistical Feature Selection. FIRE (Working Notes) (2018)
19. Werlen, L.M.: Statistical learning methods for profiling analysis. Proceedings of CLEF (2015)
20. Bayot, R., Gonçalves, T.: Multilingual author profiling using word embedding averages and svms. In: 2016 10th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). IEEE (2016)
21. Ouni, S., Fkih, F., Omri, M.N.: Toward a new approach to author profiling based on the extraction of statistical features. Soc. Netw. Anal. Min. **11**(1), 1–16 (2021). https://doi.org/10.1007/s13278-021-00768-6
22. Ouni, S., Fkih, F., Omri, M.N.: Bots and gender detection on Twitter using stylistic features. In: Bădică, C., Treur, J., Benslimane, D., Hnatkowska, B., Krótkiewicz, M. (eds.) Advances in Computational Collective Intelligence. ICCCI 2022. Communications in Computer and Information Science, Vol. 1653. Springer, Cham (2022)
23. Dichiu, D., Rancea, I.: Using Machine Learning Algorithms for Author Profiling In Social Media. CLEF (Working Notes) (2016)
24. Gopal Patra, B., Banerjee, S., Das, D., Saikh, T., Bandyopadhyay, S.: Automatic author profiling based on linguistic and stylistic features. Notebook for PAN at CLEF 1179 (2013)
25. Patra, B.G., Gourav Das, K., Das, D.: Multimodal Author Profiling for Twitter. Notebook for PAN at CLEF (2018)
26. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language-and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling. CLEF (Working Notes) (2017)
27. Poulston, A., Stevenson, M., Bontcheva, K.: Topic models and n–gram language models for author profiling. In: Proceedings of CLEF (2015)
28. Fagni, T., Tesconi, M.: Profiling Twitter Users Using Autogenerated Features Invariant to Data Distribution (2019)
29. Ouni, S., Fkih, F., Omri, M.N.: Novel semantic and statistic features-based author profiling approach. J. Ambient Intell. Human Comput. (2022)

Prioritizing Management Action of Stricto Sensu Course: Data Analysis Supported by the k-means Algorithm

Luciano Azevedo de Souza[1](✉), Wesley do Canto Souza[1], Welesson Flávio da Silva[2], Hudson Hübner de Souza[1], João Carlos Correia Baptista Soares de Mello[1], and Helder Gomes Costa[1]

[1] Universidade Federal Fluminense, Niterói, RJ 24210-240, Brazil
{luciano,wesleycanto,jccbsmello,heldergc}@id.uff.br
[2] Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil
welesson.silva@ufv.br

Abstract. The challenge of balancing the benefits and pitfalls of adopting general or customized management strategies is always present. However, the planning effort must be sufficiently flexible because making judgments before complete and in-depth evaluation could mean making a decision too late. The management of a stricto sensu specialization course relies heavily on academic publication rates and citation counts, as is well known. In this study, we present and test a novel idea for cluster analysis using the k-means method and the creation of generic responses for the groups found. The SCOPUS platform of the 17 researchers who make up the faculty of a production engineering course in Brazil was examined for indicators over the previous ten years.

Keywords: Academic management · Clustering · k-means algorithm

1 Introduction

The expansion and consolidation of *stricto sensu* postgraduate courses were driven by the creation of the Coordination for the Improvement of Higher Education Personnel (CAPES), which has as one of its purposes the constant search for the improvement of its evaluation system (Martins et al. 2012).

The academic evaluation of a higher education course involves the application of different indicators and methods to measure the maturity in teaching and research of universities (Yu et al. 2022; Mingers et al. 2015). These assessments contribute to the development of educational institutions and the improvement of scientific research management (Meng 2022).

Attempting to balance the benefits of generic actions that could offer broad scope for producing the desired effects for the solution of a specific problem-situation with the lack of access to individual conditions that, in an individualized plan, can confer greater overall performance despite requiring excessive effort is a challenge that has been discussed in various fields of application in the existing literature.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 432–440, 2023.
https://doi.org/10.1007/978-3-031-27409-1_39

Particularly, the coordination of a *stricto sensu* specialization program is a challenging task, and the pursuit of a generic strategy to improve academic productivity has been deemed ineffective when considering the quantity of scientific work and its significance to the community.

On the other hand, it requires too much effort to establish individualized measures, which may be contested by an individual researcher's exposure and may cause conflict in the team.

Yu et al. (2022) apply K-means in a study of mapping and evaluation of data from the perspective of the clustering problem, in which parameters and local statistics were used to determine a special distribution parameter, to provide a cloud of points to be used from the parameters worked initially.

This study used the K-means algorithm to identify groups of researchers who are in similar situations, with the aim of creating support plans for each group of researchers by providing a middle ground between generic and tailored action.

2 Proposed Methodology

The Scopus database was searched for publications and citations between 2012 and 2021. Individual data were obtained from the 17 researchers that make up the *stricto sensu* faculty of production engineering at a public Brazilian institution. Figure 1 shows a step-by-step breakdown of the approaches used.

The methodological procedures used in this work are shown in Fig. 1.

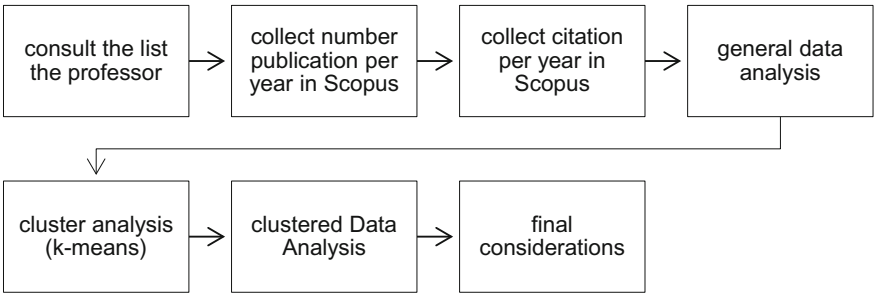


Fig. 1. Methodological procedures

It is important to note that as a criterion for counting individual products each year, it was not differentiated whether the researcher was the first author or not.

To gather the yearly citations per researcher, it was assumed that a single citation of a publication with multiple authors would get one citation each, thereby avoiding the fractioning of this indicator.

The operations were carried out on a PC running 64-bit Windows 10, with 8GB RAM and a 2.80GHz Intel(R) Core(TM) i5-8400 CPU.

The programs "ggplot2", "cluster", and "factoextra" were used in R (R 4.1.3), while the packages "openxlsx", "writexl" were used in MS Excel loading and saving. The interface development platform RStudio 2022.02.3 Build 492 was utilized.

The experimental results are shown in Sect. 3, and, in the Final considerations section, there is a discussion regarding the results, limitations, and future work proposals.

3 Experimental Results

The Fig. 2 depicts the annual publication and citation statistics per researcher in the Scopus database from 2012 to 2021.

Professor	Description	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Prof. 1	Publication	0	4	3	4	2	5	6	5	3	4	36
	Citation	16	33	35	41	45	77	129	170	222	222	990
Prof. 2	Publication	2	2	0	0	2	0	1	1	0	1	9
	Citation	1	4	1	0	3	0	2	0	6	2	19
Prof. 3	Publication	1	0	0	2	0	0	0	0	1	0	4
	Citation	0	0	0	0	0	0	0	0	0	1	1
Prof. 4	Publication	1	1	1	4	3	8	11	11	11	9	60
	Citation	0	5	11	5	9	3	13	37	58	66	207
Prof. 5	Publication	8	6	4	13	6	12	4	7	4	8	72
	Citation	37	69	80	91	88	126	91	104	135	164	985
Prof. 6	Publication	3	3	5	6	4	7	5	4	3	7	47
	Citation	39	49	59	66	69	87	77	90	123	135	794
Prof. 7	Publication	0	1	0	1	0	0	0	2	1	1	6
	Citation	0	0	2	2	0	3	1	1	4	8	21
Prof. 8	Publication	3	2	2	5	0	1	1	1	4	0	19
	Citation	13	18	19	14	21	18	26	18	20	25	192
Prof. 9	Publication	0	1	3	5	1	1	4	3	4	5	27
	Citation	6	6	10	19	18	9	12	6	13	30	129
Prof. 10	Publication	1	0	1	0	0	0	1	0	0	0	3
	Citation	0	0	0	1	2	1	4	2	10	8	28
Prof. 11	Publication	0	1	2	8	0	2	6	6	11	10	46
	Citation	4	6	7	8	14	20	33	38	77	116	323
Prof. 12	Publication	6	7	4	0	2	5	5	8	8	3	48
	Citation	78	109	123	132	138	153	215	250	291	293	1782
Prof. 13	Publication	0	1	0	0	0	1	1	0	0	1	4
	Citation	6	4	2	2	3	3	13	20	36	21	110
Prof. 14	Publication	7	10	3	14	5	6	5	7	4	7	68
	Citation	7	8	20	24	29	36	60	78	151	148	561
Prof. 15	Publication	4	3	6	1	1	0	1	3	2	0	21
	Citation	0	0	1	2	2	4	0	7	13	13	42
Prof. 16	Publication	0	1	0	1	1	0	3	2	1	2	11
	Citation	0	0	3	5	4	6	14	17	29	17	95
Prof. 17	Publication	8	8	3	4	8	12	19	14	16	28	120
	Citation	5	12	18	27	22	39	98	206	394	546	1367
Prof. 18	Publication	0	0	0	0	0	1	3	3	1	3	11
	Citation	0	3	1	0	1	5	2	8	15	26	61
Prof. 19	Publication	0	0	1	1	0	0	1	0	0	0	3
	Citation	0	0	0	0	0	3	0	3	10	7	23
Prof. 20	Publication	1	1	0	5	2	3	2	6	5	5	30
	Citation	0	0	0	1	9	6	22	16	52	44	150

Fig. 2. Individual researcher's publication and citation scores from 2012 to 2021

Due to the absence of mechanisms for such analysis on the Scopus database, the computed citations are not limited to publications published within the time. In this perspective, academics with older publications who continue to obtain citations on them acquire a comparative advantage, which should be recognized a method limitation. As a recommendation, these indications should be moderated by the researcher's age or the age of his first publication.

3.1 General Data

We recognized from the individual data table that there would be a greater proportion of publications in a small number of academics, but the concentration seemed to be smoother in the citations. Therefore, we compared the best five researchers (top 5) in each metric to the others and the results are shown in Table 1.

Table 1. Concentration of indicators

Indicator	Counting	Sharing
Publication		
Top 5	323	50.1%
Others	322	49.9%
Citation		
Top 5	5918	75.1%
Others	1962	24.9%

As we can see, the top-5 group generated in the years studied half of the papers of the group of 17 academics, confirming the worrisome discrepancies. The five most cited, on the other hand, account for a fifth of the 17 academics global indicators.

The top five academics in published works are not the same as the top five mentioned, indicating that such relationship cannot be captured in a single indicator without losing data distinction.

To accomplish our goal, we use year-by-year publishing and citation data to identify conglomerates.

3.2 K-means Clustering

We took as database for cluster identification the set of answers that indicated relevance. The purpose of the k-means method is to classify data by structuring the set into subsets whose features show intra-group similarities and differences with other groups [5, 6, 8, 17].

We utilized three ways to determine the number of groups needed to segregate the data. Elbow (Fig. 3a), Gap Stat (Fig. 3b), and Silhouette (Fig. 3c).

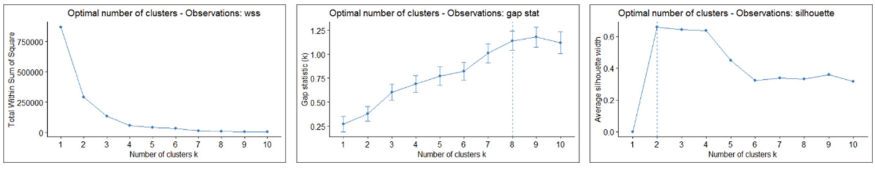


Fig. 3. Optimal number of Clusters

The Elbow method [5] indicated k = 3, Gap Stat methods [16] suggested k = 8 and the Silhouette method [9] indicates k = 2. To define the analysis, we did a visual comparison of the data with k ranging from 2 to 7 for, as shown in Fig. 4.

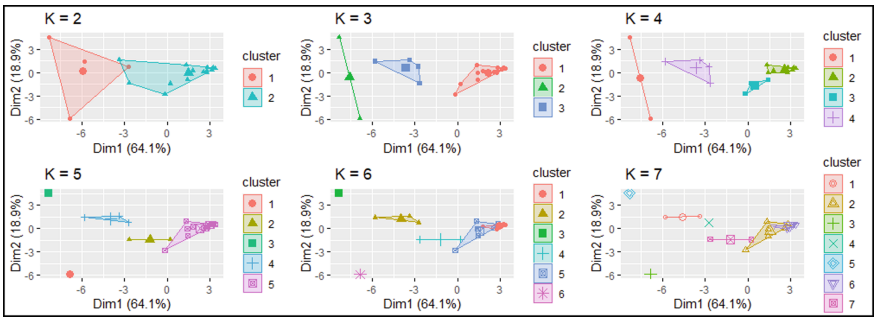


Fig. 4. Visual representations of clustering wit k = 2 to k = 5

We decided to split the data into four clusters since the comparative graph indicated that there is closeness within groups and isolation between clusters of observations. Figure 5 displays a depiction of the observation clustering.

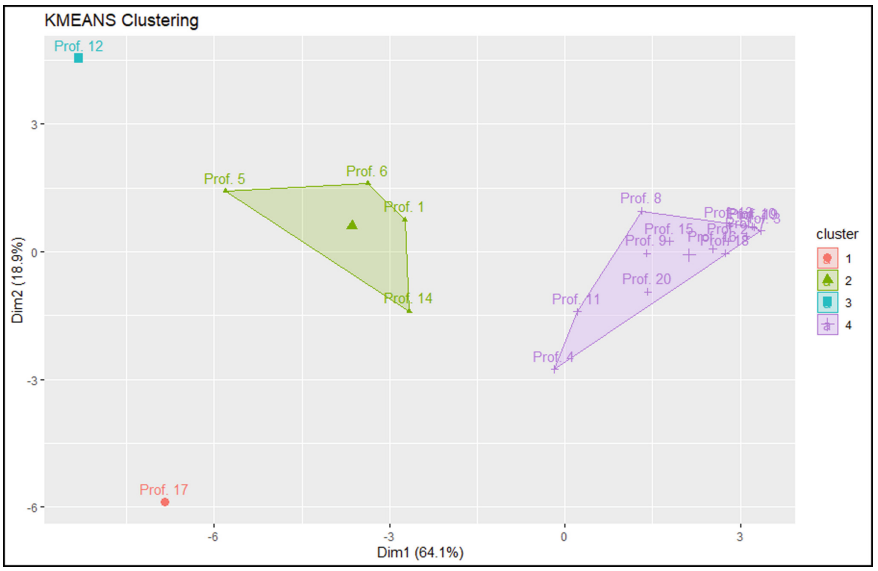


Fig. 5. Clustering data using k-means (k = 4)

It is possible to identify that Prof 12 and Prof 17 are isolated from the others and from each other, which justifies them being studied separately. The other researchers are classified into 2 groups. Researchers 1, 5, 6 and 14 make up Cluster 2 and the other 11 researchers are grouped in Cluster 4.

We compared the clusters using the average of publication and the result is shown in Fig. 6.

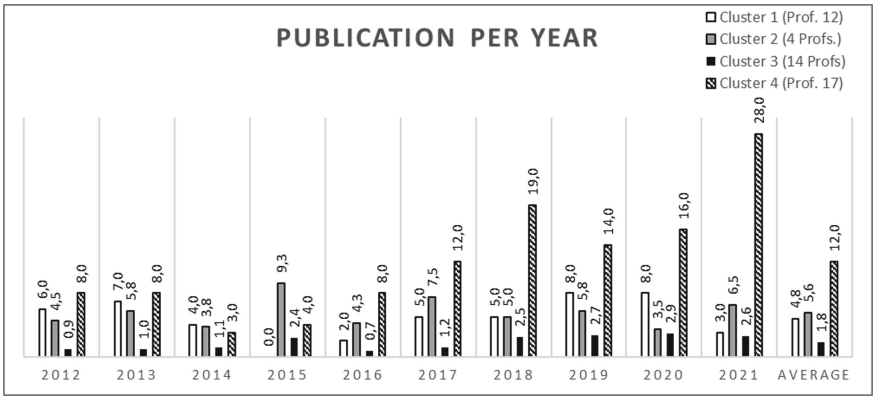


Fig. 6. Publication average per year by cluster

It's possible to learn that Prof 17 has a higher level of publication in comparison to the other clusters. It's also clear that cluster 3, with 14 components is the one where the 14 researchers had in last 10 years lower articles production rate.

Another evaluation was performed to compare the citation per cluster and the graphic is plotted in Fig. 7.

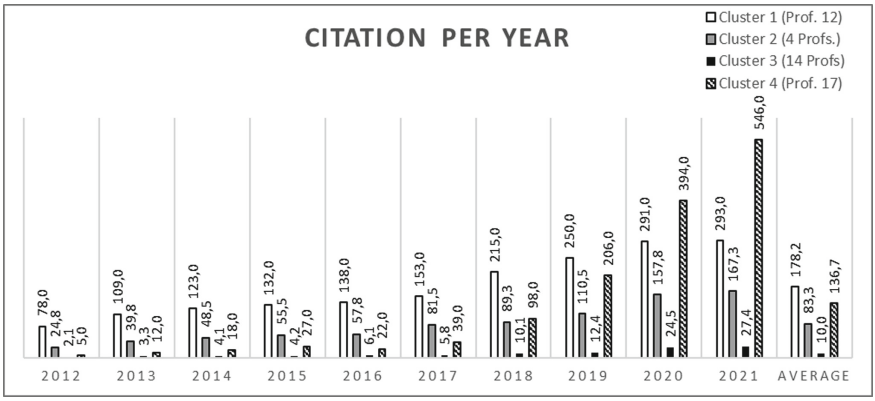


Fig. 7. Citation average per year by cluster

Researchers 12 and 17 have a high number of citations, as may be seen here. Researcher 17 has received a substantial number of citations in the previous three years, indicating that there is most likely a unique factor at work. Cluster 3 (14 Profs) has a significantly lower value than the rest. We also organizer the list researchers in Table 2 and added the respective h-index that is a metric that quantifies both publication productivity and citation impact [7].

Table 2. Researchers by cluster and respective h-index

	Researcher	H-index
Cluster 1	Prof. 12	24
Cluster 2	Prof. 1	18
	Prof. 5	18
	Prof. 6	16
	Prof. 14	14
Cluster 3	Prof. 2	2
	Prof. 3	1
	Prof. 4	7
	Prof. 7	3
	Prof. 8	6
	Prof. 9	7
	Prof. 10	2
	Prof. 11	11
	Prof. 13	4
	Prof. 15	4
	Prof. 16	5
	Prof. 18	5
	Prof. 19	3
	Prof. 20	8
Cluster 4	Prof. 17	17

Except for Researchers 12 and 17, who were recognized as unique clusters, it was feasible to find academics with comparable h-index ranges. Cluster 2 is made up of four researchers with higher h-indexes (14 to 18) who have been in the academic system for a longer period or have a high annual output of well-cited articles.

In general, we could suggest a prioritization of support action in group of Cluster 3, such as contracting translators, statistics professionals, copydesk reviewers.

4 Final Considerations

The goal of this work was supporting decision maker to establishing plans in adequate aggregation grade in which it could be possible offering support according to the differences and similarities of researchers in groups considering the main indicators of academic relevance (publication and citation). In this regard, the indicators of the previous ten years on the SCOPUS platform of the 17 researchers that comprise the staff of a production engineering course in Brazil were studied.

The k-means technique was employed for cluster analysis, and four groups were established for which actions might be designed.

The method demonstrated to be appropriated once the groups shown similar h-index (that wasn't used by k-means) and is a well-known and accepted synthetic indicator at the academy.

We intend to broaden the evaluation field to a field of knowledge as recommendations for future work, encompassing numerous researchers under comparable circumstances.

References

1. Azhari, B., Fajri, I.: Distance learning during the COVID-19 pandemic: School closure in Indonesia. Int. J. Math. Educ. Sci. Technol. (2021). https://doi.org/10.1080/0020739X.2021.1875072
2. Belle, L.J.: An evaluation of a key innovation: mobile learning. Acad. J. Interdiscip. Stud. **8**(2), 39–45 (2019). https://doi.org/10.2478/ajis-2019-0014
3. Bleustein-Blanchet, M.: Lead the change. Train. Ind. Mag. 16–41 (2016)
4. Criollo-C, S., Guerrero-Arias, A., Jaramillo-Alcázar, Á., Luján-Mora, S.: Mobile learning technologies for education: benefits and pending issues. Appl. Sci. (Switzerland) **11**(9) (2021). https://doi.org/10.3390/app11094111
5. Cuevas, A., Febrero, M., Fraiman, R. (2000). Estimating the number of clusters. Can. J. Stat. **28**(2)
6. de Souza, L.A., Costa, H.G.: Managing the conditions for project success: an approach using k-means clustering. In: Lecture Notes in Networks and Systems, Vol. 420. LNNS (2022). https://doi.org/10.1007/978-3-030-96305-7_37
7. Hirsch, J.E.: An index to quantify an individual's scientific research output (2005). https://www.pnas.org. https://doi.org/10.1073/pnas.0507655102
8. Jain, A.K.: Data clustering: 50 years beyond K-means q (2009). https://doi.org/10.1016/j.patrec.2009.09.011
9. Kaufman, L., Rousseeuw, P.J.: Finding groups in data : an introduction to cluster analysis 342 (2005)
10. Mierlus-Mazilu, I.: M-learning objects. In: ICEIE 2010—2010 International Conference on Electronics and Information Engineering, Proceedings, 1 (2010). https://doi.org/10.1109/ICEIE.2010.5559908
11. Noskova, T., Pavlova, T., Yakovleva, O.: A study of students' preferences in the information resources of the digital learning environment. J. Effic. Responsib. Educ. Sci. **14**(1), 53–65 (2021). https://doi.org/10.7160/eriesj.2021.140105
12. Pelletier, K., McCormack, M., Reeves, J., Robert, J., Arbino, N., Maha Al-Freih, with, Dickson-Deane, C., Guevara, C., Koster, L., Sánchez-Mendiola, M., Skallerup Bessette, L., Stine, J.: 2022 EDUCAUSE Horizon Report® Teaching and Learning Edition (2022). https://www.educause.edu/horizon-report-teaching-and-learning-2022
13. Ramos, M. M. L. C., Costa, H. G., da Azevedo, G.C.: Information and Communication Technologies in the Educational Process, pp. 329–363. IGI Global (2021). https://services.igi-global.com/resolvedoi/resolve.aspx?. https://doi.org/10.4018/978-1-7998-8816-1.ch016
14. Salinas-Sagbay, P., Sarango-Lapo, C.P., Barba, R.: Design of a mobile application for access to the remote laboratory. Commun. Computer and Inf. Sci. 1195 CCIS, 391–402 (2020). https://doi.org/10.1007/978-3-030-42531-9_31/COVER/
15. Shuja, A., Qureshi, I.A., Schaeffer, D.M., Zareen, M.: Effect of m-learning on students' academic performance mediated by facilitation discourse and flexibility. Knowl. Manag. E-Learning **11**(2), 158–200 (2019). https://doi.org/10.34105/J.KMEL.2019.11.009

16. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. Royal Stat. Soc. Series B: Stat. Methodol. **63**(2), 411–423 (2001). https://doi.org/10.1111/1467-9868.00293
17. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Qiang, ·, Hiroshi Motoda, Y., Mclachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D. J., Steinberg, D., Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**, 1–37 (2008). https://doi.org/10.1007/s10115-007-0114-2

Prediction of Dementia Using SMOTE Based Oversampling and Stacking Classifier

Ferdib-Al-Islam[1](✉), Mostofa Shariar Sanim[1], Md. Rahatul Islam[2], Shahid Rahman[3], Rafi Afzal[4], and Khan Mehedi Hasan[1]

[1] Northern University of Business and Technology Khulna, Khulna, Bangladesh
ferdib.bsmrstu@gmail.com
[2] Kyushu Institute of Technology, Kitakyushu, Japan
[3] Canadian University of Bangladesh, Dhaka, Bangladesh
[4] Bangladesh Advance Robotics Research Center, Dhaka, Bangladesh

Abstract. Dementia is an umbrella word that refers to the many symptoms of psychic degradation that manifest as forgetting. Typically, dementia and Alzheimer's disease are a little more challenging to examine in terms of symptoms since they begin with various views. There is no all in one test for determining if someone has dementia. Physicians identify Alzheimer's disease and other types of dementia using detailed health history, physical examination, laboratory tests, and the characteristic changes in thinking, everyday function, and behaviour associated with each kind. Clinical decision-making tools based on machine learning algorithms might improve clinical practice. In this paper, stacking based machine learning has been utilized to predict dementia from clinical information. At first, SMOTE was applied to remove data imbalance in dataset. Then, five base classifiers (LR, SVM, KNN, RF, and XGBoost) were used to form up stacking model. It achieved 91% of accuracy, 90% of precision and recall. The proposed work has shown better execution than the previous work.

Keywords: Dementia · Alzheimer's Disease · SMOTE · Machine Learning · Stacking Classifier

1 Introduction

Dementia is a neurodegenerative illness that causes nerve cells to die over time. It results in the loss of cognitive processes such as thinking, memory, and other mental capacities, which can occur due to trauma or natural aging. Dementia is a chronic, progressive, and irreversible disease. It affects approximately 44 million people worldwide, with one new case diagnosed every seven seconds. This figure is anticipated to quadruple every 20 years. Dementia has been characterized simply as a sickness (basically brain failure) that affects higher brain processes, and it is the most dreaded illness among adults over the age of 55. By 2050, it is anticipated that 131.5 million individuals globally will be living with dementia, with a global cost of $2 trillion by 2030 [1].

Alzheimer's disease is the most renowned cause of dementia. The cerebrum is composed of billions of intercommunicating nerve cells. Alzheimer's disease eradicates

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 441–452, 2023.
https://doi.org/10.1007/978-3-031-27409-1_40

the link between these cells. Proteins build and shape random structures referred to as tangles and tangles. Nerve cells eventually die, and cerebrum tissue is destroyed. The cerebrum also contains crucial synthetic chemicals that aid in transmitting signals between cells. Because people with Alzheimer's have fewer of these 'artificial couriers' in their brains, the indications do not spread as quickly [2]. Dementia is not a single disease. Instead, it describes various symptoms, including memory, reasoning, direction, language, learning limits, and relationship capacities. It is a progressive and continuing condition. Alzheimer's disease and dementia causes memory loss through expression in subjects. Alzheimer's disease is most commonly found in older adults [3]. It is a chronic neurological disease that usually develops gradually and manifests itself after some time. The most well-known early adverse effect is difficulty recalling constant instances. Because of the influence on the human cerebrum, Alzheimer's patients have headaches, mood swings, cognitive deterioration, and loss of judgment [4].

Numerous variables impact the recruitment of patients into clinical trials for Alzheimer's Disease and Related Dementia (ADRD). For instance, physician consciousness of clinical trial opportunities, the accessibility of study collaborators who can provide information about the research subject's functioning, the insensitivity of commonly used procedures in Alzheimer's trials, and considerations about labelling a patient with a serious dementia diagnosis. Accurate prediction of the beginning of ADRD in the future has numerous significant practical implications. It enables the identification of patients at high risk of developing ADRD, which aids in the clinical growth of innovative therapies. Patients are frequently found after developing symptoms and severe neurodegeneration [5]. To identify dementia in patients, several techniques using various datasets have been presented. When identifying dementia using clinical datasets, methods have flaws in accuracy, precision, and other performance measures [6, 7].

In this research, a dementia prediction system has been built using SMOTE based oversampling technique and stacking model. The base models used in this work were LR, SVM, KNN, RF, and XGBoost; LR was the meta classifier. The performances of individual models and voting model were also evaluated. This work has been removed the class imbalance problem existed in the previous work and also shown better performance.

The subsequent portion of the article is structured as follows: The "Literature Review" section covers recent studies in detecting and diagnosing dementia using machine learning and other methods. Several sub-sections of the "Methodology" section highlight the particulars of this study. The results are described in the "Result and Discussion" section. "Conclusion" is the concluding part of the article.

2 Literature Review

Akter and Ferdib-Al-Islam [2] divided dementia into three groups (AD Dementia, No Dementia, and Uncertain Dementia) in this study to diagnose Alzheimer's disease in its early stages using the XGBoost method, and they also showed the feature significance scores. The accuracy was 81% in that work, the precision was 85%, and the most important feature was "ageAtEntry." Class imbalance problem was not solved in that study. Hane et al. [5] used two years of data to forecast the fate of ADRD onset. Clinical notes with particular phrases and moods were presented in a de-identified format.

Clinical notes were integrated in a 100-dimensional feature space to identify common terms and abbreviations used by hospital systems and individual clinicians. When clinical notes were incorporated, the AUC increased from 85 to 94%, and the prognostic value (PPV) increased from 45.07 to 68.32% in the model at the onset of the disease. In years 3–6, when the quantity of notes was greatest, models containing clinical notes increased in both AUC and PPV; findings in years 7 and 8 with the lowest cohorts were mixed. Mar et al. [6] searched through 4,003 dementia patients' computerised health records using machine learning (random forest) approach. Neuropsychiatric symptoms were documented in 58% of electronic health records of patients. The psychotic cluster model's area under the curve was 0.80, whereas the depressed cluster model's area under the curve was 0.74. Additionally, the Kappa index and accuracy demonstrated enhanced discrimination in the psychotic model.

Zhu et al. [7] enlisted 5,272 people who completed a 37-item questionnaire. Three alternative feature selection strategies were evaluated to choose the most significant traits. The best attributes were then integrated with six classification algorithms to create the diagnostic models. Among the three feature selection approaches, Information Gain was the most successful. The Naive Bayes method performed the best (accuracy 81%, precision 82%, recall 81%, and F-measure 81%). So et al. [8] applied machine learning methods to develop a two-layer model, which was inspired by the method utilized in dementia support centres for primary dementia identification. When normal, mild cognitive impairment (MCI), and dementia F-measure values were assessed, the MLP had the maximum F-measure value of 97%, while MCI had the lowest.

Bennasar et al. [9] employed a study that includes 47 visual cues after thoroughly examining the available data and the most commonly published CDT rating methods in the medical literature. While comparing to a single-stage classifier, the findings revealed a substantial increase of 6.8% in discriminating between three stages of dementia (normal/functional, mild cognitive impairment/mild dementia, and moderate/severe dementia). When just distinguishing between normal and pathological circumstances, the results revealed a classification accuracy of more than 89% Mathkunti and Rangaswamy [10] explored the research by using ML techniques to improve the accuracy of identifying Parkinson's disease. The data collection in question is from the UCI's Online-based ML Repository. Accuracy, recall, and confusion matrix are computed using SVM, KNN, and LDA approaches. This implementation achieved a precision of 100% for SVM and KNN and 80% for LDA.

3 Methodology

The proposed work's approach has been divided into the subsequent steps:

- Preprocessing of Dataset
- EDA on Dataset
- Use of SMOTE
- ML Classifiers for Classification

3.1 Preprocessing of Dataset

The dataset utilized for this study is accessible on Kaggle [2]. This dataset comprises 1229 instances of 6 attributes and a target variable called "Dx1". The dataset details has been illustrated in [2].

Firstly, the irrelevant columns were dropped. Label encoding is a general strategy for achieving this goal. This research used label encoding to transform categorical data to numeric inputs. The purpose of feature scaling is to automatically resize every feature to the same size. In this investigation, each feature was treated to min-max scaling or normalization. It is a method of rescaling values between 0 and 1. The formula behind the use of normalization is described in (1):

$$Feat' = \frac{Feat - Feat_{\min}}{Feat_{\max} - Feat_{\min}} \tag{1}$$

where $Feat_{max}$ and $Feat_{min}$ denotes the peak and the bottom values of the feature.

3.2 EDA on Dataset

EDA is a technique for analyzing or understanding data and discovering perceptions or essential characteristics of the data. EDA has been performed in this dataset. The insights of EDA have been demonstrated in Figs. 1, 2, 3, and 4.

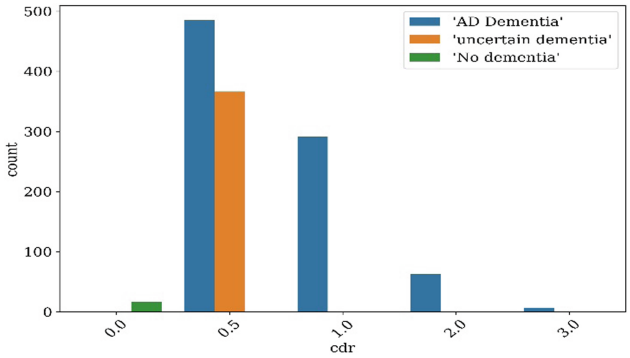


Fig. 1. 'cdr' data distribution as per the target variable

Figure 1 demonstrates the "cdr" data distribution as per the target variable. Clinical Dementia Rating (cdr) is mostly for "AD Dementia" and "uncertain dementia," with a rating of 0.5. Figure 2 illustrates the "mmse" data distribution as per the target variable. Mini-Mental State Examination (mmse) is mostly for values 28, 29, 24, 30, 23, 25, 27 and 26.

Figure 3 represents class-wise memory screening data distribution. "AD Dementia" is maximum at 1, and "Uncertain Dementia" is maximum at 0.5. Figure 4 shows the imbalance of the target variable. "AD Dementia" has the maximum instances, and the other two classes have low instances than "AD Dementia". A class balancing algorithm can remove this problem.

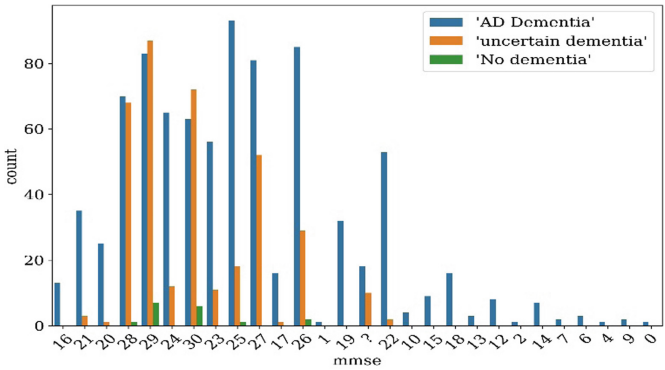


Fig. 2. 'mmse' data distribution as per the target variable

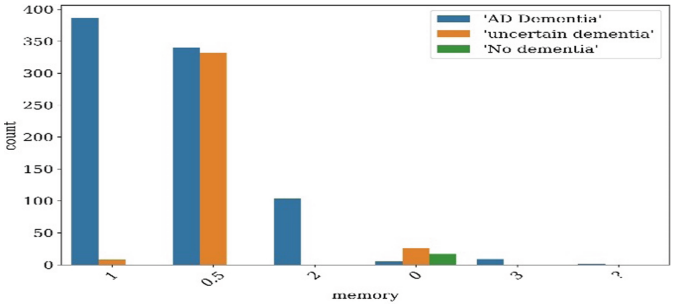


Fig. 3. 'memory' data distribution as per the target variable

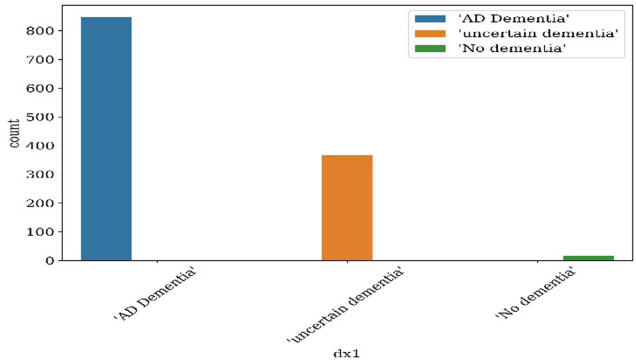


Fig. 4. Data distribution of the target variable

3.3 Use of SMOTE

Imbalanced classifications provide a challenge since the machine learning methods used to classify the data were developed on the assumption of an equal number of data examples for individual class. This consequence occurs in models that perform poorly,

most notably for the minority class. As the minority class is more substantial, and hence more disposed to classification errors than the majority class.

SMOTE is an approach for oversampling in which artificial data are generated for the minority class [11, 12]. This computation contributes to eradicating the over-fitting problem associated with random oversampling. To begin, N is initialised with the total number of oversampling perceptions. Typically, the binary class distribution strategy is selected with the objective of reaching a 1:1 ratio. In any instance, this might be modified based on the circumstances. At that point, the series begins by randomly picking a positive class occurrence. Then, the KNNs for that occurrence are collected. N of these K occurrences are then picked to generate further false occurrences. Using any distance metric, the variation in distance between the component vector and its neighbors is determined to accomplish this. This difference is now increased by any arbitrarily big incentive between 0 and 1 and added to the previous feature vector.

- Step 1: Putting the minority class set A, for each $x \in A$, the KNNs of x are acquired by finding the Euclidean distance among x and each other instance in set A.
- Step 2: The sampling rate N is set by the imbalanced extent. For each $x \in A$, N samples $(x_1, x_2, \ldots, x_n)$ are arbitrarily chosen from its KNN, and they build the set A'.
- Step 3: For each sample $x_k \in A$ $(k = 1, 2, 3, \ldots, N)$, the subsequent principle is applied to produce a fresh instance:

$$x' = x + rand(0, 1) \times |x - x_k| \tag{2}$$

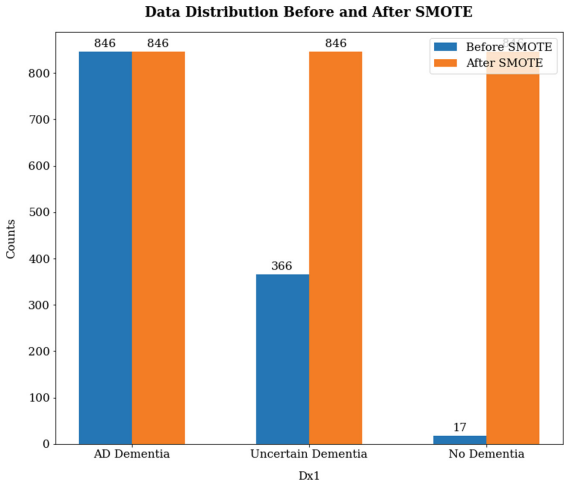


Fig. 5. Target variables class-wise data distribution earlier and subsequently use of SMOTE

Figure 5 demonstrates the class-wise data of the output variable earlier and subsequently using SMOTE. The corresponding number of instances of "AD Dementia," "Uncertain Dementia" and "No Dementia" was 846, 366, and 17 before using SMOTE.

The corresponding number of instances of "AD Dementia," "Uncertain Dementia," and "No Dementia" has become 846, 846, and 846 after using SMOTE.

3.4 ML Classifiers for Classification

The dataset has been splitted by 80:20 for preparing training and set, respectively. Firstly, LR, SVM, KNN, RF, and XGBoost have been applied to predict dementia. Voting classifier has been formed using these base classifiers and soft voting techniques. Stacking classifier has been made using the base mentioned above classifiers as level-0 model and logistic regression as meta-model. "GridSearchCV" technique has been utilized to find the optimal hyperparameters of the classifiers.

Logistic Regression. Logistic regression is a model with a set number of parameters dependent on the number of input characteristics and produces categorical results. It explains how perception is likely to take on one of those two values. LR is a very simple model that falls short of predictions compared to more complicated models. Table 1 describes the optimal parameters of the LR model.

Table 1. Optimal parameters of LR model

Parameter	Value
solver	"lbfgs"
penalty	"l2"
multi_class	"multinomial"

Support Vector Machine. SVM models represent several classes in multidimensional space through a hyperplane. The hyperplane will be constructed repeatedly using SVM to minimize errors. SVM seeks to classify datasets in order to discover the largest external hyperplane. Table 2 describes the optimal parameters of the SVM model.

Table 2. Optimal parameters of SVM model

Parameter	Value
kernel	"rbf"
C	1.0
gamma	"scale"

K-Nearest Neighbor. The k-nearest neighbour approach preserves all previously collected data and classifies new data points based on their similarity (e.g., distance functions). This refers to the occurrence of new data. It may then be easily classified using

the K-NN approach. K-NN is a straightforward classifier frequently used as a starting point for more complex classifiers like ANN. Table 3 describes the optimal parameters of the K-NN model.

Table 3. Optimal parameters of K-NN model

Parameter	Value
n_neighbors	11
metric	'minkowski'
p	2

Random Forest. It is a meta-algorithm that applies decision tree classifiers to several subsamples of the dataset and then utilizes averaging to enhance estimated accuracy. When the training set for the present tree is generated using sampling with replacement, about one-third of the occurrences are removed. When additional trees are introduced to the forest, this out-of-bag data is used to produce an approximate solution to the classification error. Table 4 describes the optimal parameters of the RF model.

Table 4. Optimal parameters of RF model

Parameter	Value
max_depth	2
n_estimators	200
random_state	0

XGBoost. XGBoost is a machine learning approach based on decision trees that use a gradient boosting system to solve problems. It is a kind of model in which new models are employed to calculate the residuals of prior models and are then incorporated in the final output prediction. This approach begins with a model to perform a forecast. Then, the model's loss is calculated. This is mitigated by training a new model. For classification, this model is being included to the ensemble. Table 5 describes the optimal hyperparameters of the XGBoost model.

Voting Classifier. Voting classifier has been made by combining the base classifiers—LR, SVM, KNN, RF, and XGBoost. Figure 6 shows the architecture of the voting model. Firstly, the base classifiers are trained on the training set. Then each model was tested on the test set, and then each was summed up. The soft voting technique has been chosen for picking up the final prediction.

Stacking Classifier. Stacked adaptation consists of stacking the output of a separate estimator and applying a classifier to obtain the final prediction. Stacking enables exploiting

Table 5. Optimal parameters of XGBoost model

Parameter	Value
objective	"reg:softmax"
max_depth	3
n_estimators	1000

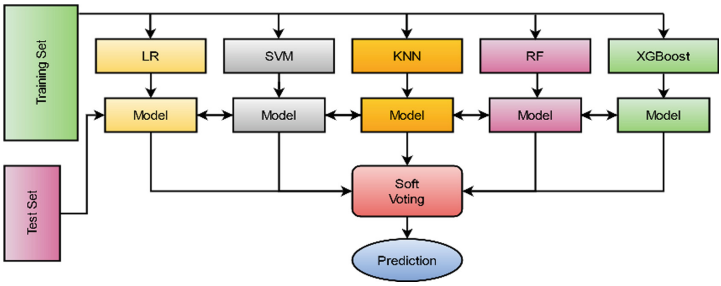


Fig. 6. Architecture of the voting model

the strength of each independent estimate by utilizing their conclusion as input to a final estimator. Figure 7 shows the architecture of the stacking model. Here, LR, SVM, KNN, RF, and XGBoost acted as level-0 model and then the LR model acted as the level-1 model.

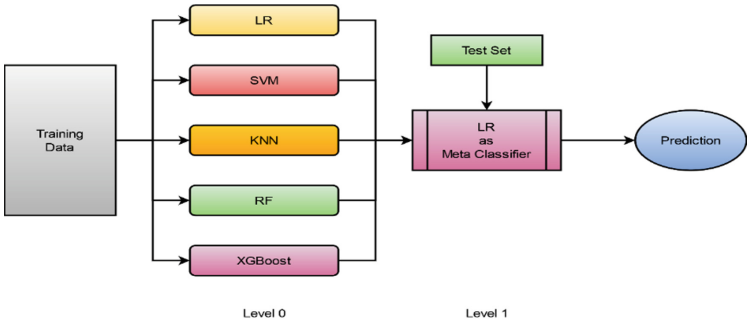


Fig. 7. Architecture of the stacking model

4 Result and Discussion

Patients' dementia has been predicted using classifiers such as LR, SVM, KNN, RF, XGBoost, Voting, and Stacking, as described earlier. The performance of the model was

measured by three separate performance metrics—accuracy, precision, and recall—in accordance with (3)–(5).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The comprehensive classification report of every ML model has been represented in Table 6. The performances of all models have been increased significantly after using SMOTE. Among the ML models, the stacking model performed better than the other models with 91% accuracy, 90% precision, and 90% recall after using SMOTE.

Table 6. Classification report of ML models

Model	Without SMOTE			Using SMOTE		
	Acc. (%)	Prec. (%)	Rec. (%)	Acc. (%)	Prec. (%)	Rec. (%)
LR	73	71	73	90	90	90
SVM	74	74	74	89	89	89
KNN	80	79	80	87	87	87
RF	77	78	77	89	90	89
XGBoost	81	85	83	88	89	88
Voting	82	82	82	91	90	89
Stacking	83	82	83	91	90	90

The confusion matrix of the stacking model has been illustrated in Fig. 8. 42 "AD Dementia" cases have been predicted as "Uncertain Dementia" and 15 "Uncertain Dementia" cases have been predicted as "AD Dementia". Total 57 errors were occurred in the prediction of dementia with the stacking model.

Figure 9 shows the performance comparison of the proposed work with the previous work. The proposed work has shown better performances in both accuracy and precision. This work reached 91% accuracy and 90% precision using a stacking model after applying SMOTE; Akter and Ferdib-Al-Islam [2] achieved 81% accuracy and 85% precision. From Table 6, it can be seen that the proposed stacking model also outperformed the previous work without using SMOTE. The work proposed in this paper eliminates the class imbalance issue that existed in the previous work and performed better in both metrics mentioned in Akter and Ferdib-Al-Islam [2].

5 Conclusion

Alzheimer's disease primarily causes dementia. Regardless, neither Alzheimer's disease nor Alzheimer's dementia is an unavoidable consequence of ageing. Dementia is not a

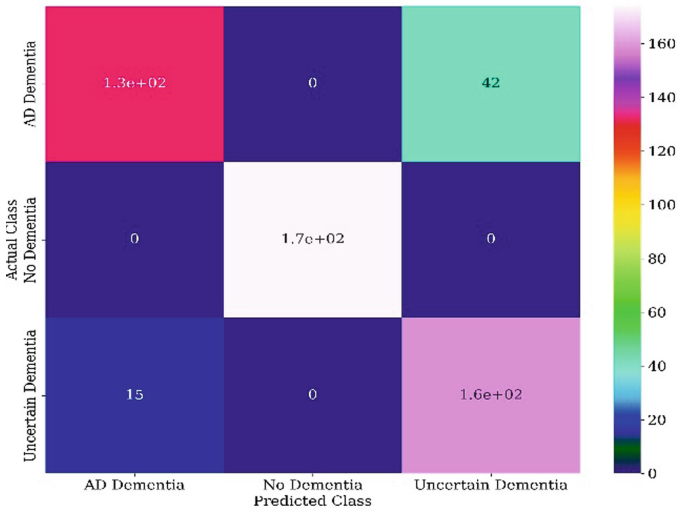


Fig. 8. Confusion matrix of stacking model

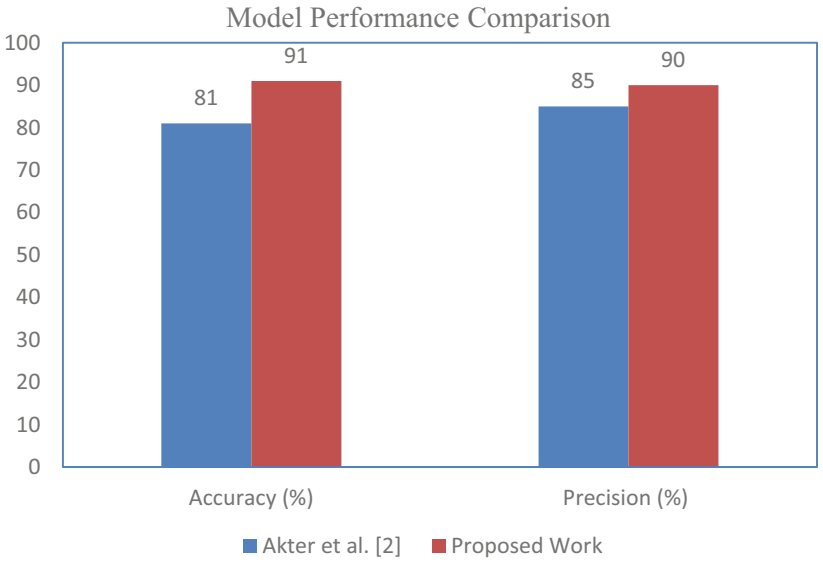


Fig. 9. Model performance comparison with previous work

normal part of ageing. It is caused by damage to synapses, impairing their ability to transmit information, which may affect one's thinking, behaviour, and emotions. This research predicts dementia using ensemble machine learning. This work also eliminates the data imbalance issue using SMOTE that existed in the previous work. The stacking classifier performs best with 91% accuracy, 90% precision, and recall compared to base

classifiers. Further analysis with other oversampling technique and other classification algorithms may enhance performance.

References

1. Prince, M., et al.: Recent global trends in the prevalence and incidence of dementia, and survival with dementia. Alzheimer's Res. Ther. **8**, 1 (2016)
2. Akter, L., Ferdib-Al-Islam: Dementia identification for diagnosing Alzheimer's disease using XGBoost algorithm. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), pp. 205–209 (2021)
3. Sharma, J., Kaur, S.: Gerontechnology—the study of Alzheimer disease using cloud computing. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 3726–3733 (2017)
4. Symptoms of dementia. https://www.nhs.uk/conditions/dementia/symptoms/
5. Hane, C., et al.: Predicting onset of dementia using clinical notes and machine learning: case-control study. JMIR Med. Inform. **8**(6), e17819 (2020)
6. Mar, J., et al.: Validation of random forest machine learning models to predict dementia-related neuropsychiatric symptoms in real-world data. J. Alzheimers Dis. **77**(2), 855–864 (2020)
7. Zhu, F., et al.: Machine learning for the preliminary diagnosis of dementia. Sci. Program. **2020**, 1–10 (2020)
8. So, A., et al.: Early diagnosis of dementia from clinical data by machine learning techniques. Appl. Sci. **7**(7), 651 (2017)
9. Bennasar, M., et al.: Cascade classification for diagnosing dementia. In: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2535–2540 (2014)
10. Mathkunti, N.M., Rangaswamy, S.: Machine learning techniques to identify dementia. SN Comput. Sci. **1**(3), 1–6 (2020). https://doi.org/10.1007/s42979-020-0099-4
11. Ferdib-Al-Islam, et al.: Hepatocellular carcinoma patient's survival prediction using oversampling and machine learning techniques. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 445–450 (2021)
12. Ferdib-Al-Islam, Ghosh, M.: An enhanced stroke prediction scheme using SMOTE and machine learning techniques. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6 (2021)

Sentiment Analysis of Real-Time Health Care Twitter Data Using Hadoop Ecosystem

Shaik Asif Hussain[1(✉)] and Sana Al Ghawi[2]

[1] Centre for Research and Consultancy, Middle East College, Muscat, Sultanate of Oman
shussain@mec.edu.om

[2] Department of Electronics and Communication Engineering, Middle East College, Muscat, Sultanate of Oman
sanaghawi@mec.edu.om

Abstract. The term "sentimental analysis" refers to classifying and categorizing user opinions depending on how they express their feelings about specific pieces of data. Trending concepts are gaining popularity, and Twitter is the most utilized to discover people's thoughts. This approach uses Hive ware, Sqoop, Hadoop, and Flume to capture real-time health information from Twitter via the system setup. Using the Flume agent, the keyword file on the Hadoop cluster obtains similar data, causing the resulting data to synch with HDFS (Hadoop Distributed File Systems). As one of the most well-known social media platforms, Twitter receives an enormous volume of tweets each day. There are various ways that this data can be analyzed in multiple ways, including for business or government purposes. Twitter's massive volume of data makes it difficult to store and process this information. Hadoop is a framework for dealing with big data, and it features a family of tools that may be used to process various types of data. The health care real-time tweets are used in this research. This work has been using Apache Flume to collect real-time tweets. To address further the proposed system, it is designed to perform sentimental analysis from the tweets, conversations, feelings, and activities on social media and determine the cognition and behavioral state of everyone. In the proposed work for the real-valued and current patterns of health tweets, sentimental analysis is used in real-time. Hadoop ecosystem tools like Hive and Pig are used for execution time and real-time tweet analysis. In real-time work, it is calculated that Pig is more efficient than Hive based on the trial results, as Pig takes less time to execute than Hive.

Keywords: Twitter · Hadoop · Health care · Hive · Tweets

1 Introduction

More than a billion individuals use Twitter every day, sending out hundreds of millions of tweets every minute at a rate of more than 100 million every hour. A relational SQL database is insufficient for analyzing and comprehending vast activities. A massively parallel and distributed system like Hadoop is ideal for handling this data type. The work focus on how Twitter data can be mined and exploited to make targeted, real-time,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 453–463, 2023.
https://doi.org/10.1007/978-3-031-27409-1_41

and informed decisions regarding or to find out what people think about a particular issue of interest. The proposed work concentrates on the mining and utilization of Twitter-generated data. Using sentiment analysis, businesses can see how effective and pervasive their marketing campaigns are. In addition, companies can examine the most popular hashtags currently trending. The potential uses for Twitter data are virtually limitless.

Using text analytics, opinion mining (also known as sentiment analysis) is used to glean information about people's feelings and thoughts from various data sources. Most of the time, the information needed to conduct this Sentimental analysis is gleaned from the internet and various social media platforms. This method uncovers the text's hidden emotions (sentiments) and then goes in-depth to examine them. The goal of dynamic analysis is to identify the thoughts represented and to find the expressed sentiments. It is possible to get real-time information about the most popular social topics via social media, and this data changes dynamically with time. Sentimental analysis of Twitter data can reveal how well people understand specific political and corporate issues. Sentimental analysis can also examine the user's perspective on a wide range of unstructured tweets (Fig. 1).

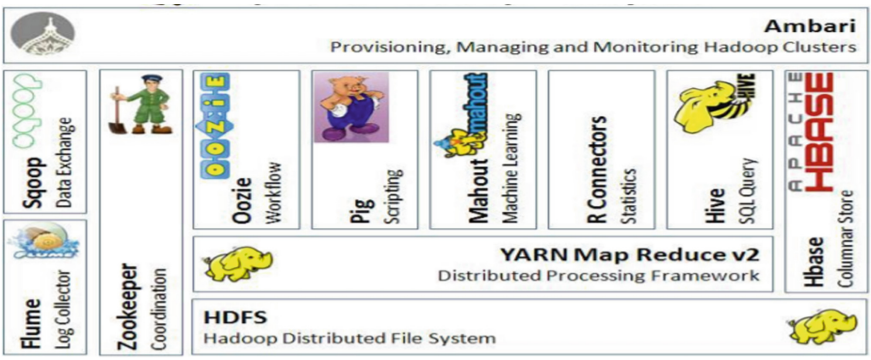


Fig. 1. Clearly describes Apache Hadoop Ecosystem (intellipaat.com)

Preprocessing, indexing, word occurrence monitoring, counting, and word clustering, are some of the methods now in use for finding the events [3, 4]. The developing approaches to concept (topic) detection concentrate only on global-scale issues, whereas fascinating existing ideas of lesser magnitude will receive minimal attention [5]. New shifts will be detected quickly in the evolving concepts, and users will be notified so that they may respond quickly. It's becoming increasingly difficult to prioritize and customize data sources because of the rapid influx of information.

It has become increasingly difficult for standard data analysis methods to do adequate analysis on big data sets. When it comes to processing big amounts of data, Hadoop has emerged as a robust architecture that can handle both distributed processing and distributed storage. In the Hadoop framework, MapReduce and Hadoop distributed file system (HDFS) are two of the most essential parts of it. To clearly understand the file system used by Hadoop can store and distribute data into memory blocks and easily move them across clusters using the MapReduce computing methodology. Additionally,

Apache provides various tools and components to meet the demands of developers. All of these is together referred to as the Hadoop Ecosystem. Hadoop's file system has been used to hold real-time streaming data on Indian political concerns.

1.1 Social Networking and Smart Solutions

The use social networking platforms to share their posts, feelings, and activities. It is not evident to find the influence of these social networks because it has attracted much attention. Relentlessly it is known that social networking has a direct impact on health care and destabilizes the situation as well. At an early age, students don't like to express their expressions or feelings directly as the result of hormones is too much. Hence the parents or elders should take responsibility for monitoring the children in their teenage as this age is prominent where children need suitable guidance and support. After reaching maturity level, students blame the parents that proper permission was not given, and the elders must monitor their children accurately so that they will not deviate from the wrong side.

The advent of technologies and fast-moving timelines has mined much information to understand the influence of the diffusion process. The sentiment analysis in big data processing can easily predict the behavioral situation and health care process. The big data framework stores the tweets, blogs, and conversations in the Application program interface (API), where the maximum entropy classifier of the supervised machine learning algorithm can classify and predict the situation in positive, negative, and neutral results.

The big data sets confirm that the health state scenarios as precision and recall. Data analytics plays a prominent role in analyzing and decision-making. The sentiment analysis is used to collect the feedback and focus on the tweets of certain words interpreted to determine the health and behavioral state. The study is carried out in two phases testing and training and testing step. The big data analytics of healthcare management performs social network data sharing using python and Hadoop map reduce.

2 Background and State of the Art

Literature analysis is performed to identify the need for and importance of the related work from other researchers and highlight the significance of sentimental analysis. The proposed work gives the feedback [1] from the user is quite essential in media networks. The social media network data is analyzed based on gender, location, and features. This data is performed for sentiment analysis and determine the product based on the score of everyone.

Sentimental analysis [2] used to analyze the data collected from twitter all these tweets are stored in an excel file and preprocessed and filtered using the Naïve Bayes classifier. This classifier calculates the sentimental score as positive, negative, and neutral emotions. The data analytics by collecting data from various sources such as Twitter, Facebook, and Instagram. The data collected from social media networks are fetched an open API. This data comprises the secret key, access token to analyze through Rest API. The data file is stored in.csv format and trained with sequence sets of health data.

The hive warehousing is used to store and process the data inbuilt with the MapReduce classifier. A software called flume is used to import and export processed data files. This paradigm is used to convert unstructured data into HDFS to solve the problems of big data analytics. In Hadoop ecosystems, sentiment analysis of tweets has become increasingly popular in the recent several years. Using item-based Collaborative Filtering methods and Apache Mahout, [3] suggested a recommender system. Streaming messages are broken down into smaller bits by the system, which then constructs a recommendation model for recommending news. A new recommender model is being built while the old one is being phased out. Proposed systems have been tested and found reliable and quick to respond. Using the Naive Bayes method and the MapReduce paradigm, [8] developed a system for classifying many tweets. Multi-node Hadoop cluster ranks tweets according to their subject matter. The naive Bayes method and MapReduce algorithm demonstrate how they can be combined.

Adverse medication reactions can be identified automatically using a dimension reduction technique known as Latent Dirichlet Allocation (LDA) [11]. The datasets used from any source of social media would be unstructured to analyze the sentiment data whether by supervised or unsupervised learning approaches, according to the authors of [12]. It is seen in the survey carried that unsupervised method achieves greater accuracy [25] of 82.26% by using multinomial Bayes Method when compared to supervised methods for 67% accuracy (Table 1).

Table 1. The below table gives the literature survey of the approaches used in big data.

Reference	Analysis approach	Data extraction	Preprocessing	Tweets/emoticons/posts /opinions/reviews	Imbalance of data
[13]	Mapreduce	Flume	Data cleaning	Tweets	Yes
[14]	Hive	Twitter4j	Data transformation	Tweets	No
[15]	Pig	Flume	Data reduction	Posts	Yes
[16]	Hive	Flume	Aggregation	Opinions	No
[17]	Pig	Flume	Normalization	Posts	Yes
[18]	Hive	Topsy	Data cleansing	Reviews	Yes
[19]	Hive	RestaPI	Missing data	Posts	No
[22]	Hive	Flume	Data transformation	Tweets	No
[23]	MapReduce	RestaPI	Numerosity reduction	Reviews	Yes

3 Design Methodology

Sentimental analysis can be applied to any knowledge area to describe how individuals feel about current concepts and technologies. Tweets include dynamic data that is updated

in real-time. Sentimental analysis is used to extract people's opinions from tweets on Twitter in this suggested work. To configure the system to obtain real-time data from Twitter, this procedure makes use of numerous big data tools, such as Hive, Flume, HDFS (Hadoop), and the analysis is carried out on the process that is most in need of it. Run Hive Meta store server and flume agent after Hadoop cluster is started first. A database table is established to show and perform sentimental and hashtag analysis on the yield tweets. Finally, the Hadoop environment's sentimental analysis reports are shown.

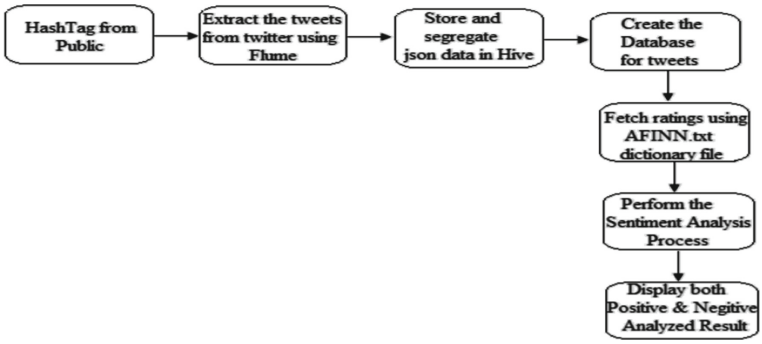


Fig. 2. Various frameworks Hive, Flume, HDFS used in processing sentimental analysis

Figure 2 depicts a conceptual diagram of the sentiment analysis method as proposed in the paper. In Flume, conf, specify the hashtag, and then utilise the flume environment to pull the live twitter data. Each of these tweets has a corresponding hashtag, and all the extracted tweets are saved to an HDFS directory for later retrieval. An hive-serdes-1.0-SNAPSHOT.jar file converts highly unstructured Twitter data to structured data using Hive. The tweets are stored in a database table created by the Hive. There is an enormous quantity of data in the Hive from this converted structured data. To execute sentiment analysis, the hive warehouse is mined for its sparse data. It's crucial to keep in mind the following: tweet id, tweets text, and followers. Splitting the text data into words based on id and storing it in a separate table called split words is the next step after gaining access to the necessary parameter information. Each word in the id changes as the procedure proceeds.

After then, consult your dictionary, which has an enormous number of positive and negative terms. Dictionary terms. Every word has a star next to it to indicate how important it is. Some words have a rating of $+5$ to $+1$, whereas many others have a rating of -5 to -1. Finally, you'll need id, split words table, and rating information from a dictionary. This is done by comparing the terms in both files that have the same number of letters. Calculate the average rating for each word after the process is complete. It is important to know which words have ratings more than zero and which words have ratings less than zero when calculating average ratings. Negative and positive ratings data were analyzed to identify the public's favorable and negative perceptions of the product. Finally, in the Hadoop environment, show the result. Methodology for applying sentiment analysis and subsequently displaying the analyzed results. There are good and negative values assigned to the tweets based on their overall attitudes. As part

of a sentiment analysis, these hashtags are used to find out what people feel about new technology or anything else. The following process was used to carry out the intended job. It is necessary to establish a Twitter application and gain access to the keys to obtain the tweets directly from the Twitter source. The tweets from Twitter can be accessed with the use of these keys. Bin/start-all.sh can be used to kick off the HDFS cluster before moving on to sentiment analysis.

3.1 Flume Collection of Tweets from Twitter

Large volumes of log data can be processed and transferred quickly and easily using Flume's distributed and accessible architecture. Its architecture is simple and adaptable, depending on the volume of data flowing through it. A simple expandable data model that can be used for online analytic purposes is employed. The various data sources that offer most of the information that needs to be evaluated include cloud servers, enterprise servers, application servers, and social networking sites. The log files include this information, which can be accessed. Flume is a strong and fault-tolerant system with a wide range of recovery and reliability options. Apache Flume is utilised to gather tweets from Twitter as a part of the planned study, which uses streaming tweets from the service. Buffering the tweets to an HDFS sink allows them to be pushed to a specific HDFS location. In the root directory of a flume, a flume.conf file is created for obtaining tweets from twitter. There is a setup of the sink, source, and channel. Our data comes from Twitter, and it goes to HDFS, which acts as both a drain and a source. The.com. Cloudera, flume, source, Twitter Source is supplied as part of the source setup. Next, all four Twitter tokens are exchanged. Finally, in the source settings, the keywords are passed, which results in the tweets being retrieved. In the sink configuration, the HDFS properties are set up. There are no more issues with the HDFS path or write batch size, format, or type. In the end, the memory channel configuration has been completed. Using the command below, four tokens, a Twitter source type, and keywords are all included. Now is the time to begin the actual execution process itself. Using the following queries, we can retrieve tweets from Twitter.

Step 1: In the first install the flume and send all the installation by changing the directory to flume home.
Step 2: In the following step capture and obtain the tweets from streaming data of twitter and interface it to flume as an agent.

Figure 3 views the fetching of tweets. Using SQL, Apache Hive's data warehouse software allows users to write, read, and maintain massive datasets in distributed storage. All the data in the storage structure has been pre-specified and is known in advance. The JDBC driver and command-line tool are used by the users to connect to the database. The Hive CLI can be used to access HDFS files directly. The hive terminal must be opened to do this, and the Hive's meta store must be started to store the metadata from hive tables.

3.2 Feature Extraction and Extraction of Hashtags

In this stage, known as preprocessing, many fields in Twitter tweets are examined. These include ids, text, entities, languages, time zones, and many more. We used tweet ids and

entities fields where the entity field has a member hashtag to identify popular hashtags. These two members are combined to perform further analysis on the tweet id.

After this phase, a sample of the outcome is shown with symbols, app indices and URL with different user mentions. In each hashtag object, there are two fields: a text field and an indices field where the hashtag appears.

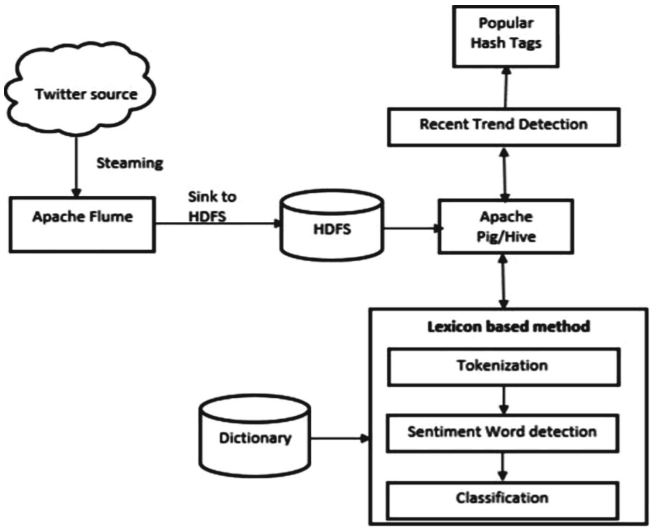


Fig. 3. System model for capturing and analyzing the tweets.

3.3 Sentiment Analysis

It is defined as an author's expression of opinion about any object or aspect of the subject matter. This technique's primary goal is the identification of opinion words in a body of text. After identifying opinion words, the sentiment values of these words are assigned. The text's polarity must also be determined, and that's the last step. You can have positive, negative, or no polarity at all. Sentiment classification has been accomplished using a Lexicon approach. During this step, the sentence is broken down into individual words. Tokenization is a term for this process. As a starting point, these words are tokenized and used to identify opinion words. Pig and Hive have been used to perform sentiment analysis on real-time tweets. The sentiment analysis is carried out using Algorithm 2. Hadoop's file system stores the trending topic tweets fetched by Apache Flume. These data must be loaded into Apache Pig to perform sentiment analysis. Pig can be used to identify the sentiment of tweets by following these steps.

3.4 Extracting the Tweets

As a result, Apache Pig's JSON-formatted tweets include not only the tweets themselves but also additional information such as the tweet id, the location from which the tweets

were posted and the tweet posting time. Only tweets are used for the sentiment analysis. We used the JSON Twitter data to extract Twitter id and tweet as a preprocessing step for sentiment analysis.

3.5 Tokenizing the Tweets

Using this technique, we can determine the positive and negative connotations of different words. Splitting a sentence into individual words is the only way to find sentimental words. Tokenization refers to the process of separating a continuous stream of text into individual words. The tweets that were collected in the previous step are tokenized and broken down into individual words. Sentiment analysis uses this tokenized list of words as a starting point for further processing.

3.6 Sentiment Word Detection and Classification of Tweets

There is a dictionary created of sentiment words to find sentiment words in tokenized tweets. Sentiment words are rated from 5 to −5 in this dictionary, with the higher the number, the better the word's meaning. Words rated 1 to 5 are considered positive, while words rated 5 to 1 are considered negative. To make it easier for you to find the tweets you're looking for, we've sorted them by tweet id. As a result, we must now compute the average of the tweets' ratings. Tweets' average rating (AR) is calculated by multiplying the number of words in a tweet by formula where TWP stands for the total number of words in a tweet. This is divided as the tweets into positive and negative ones based on the calculated average rating (Fig. 4).

```
Data: dataset: = Corpus of tweets, list of dictionary words with rating
Result: classification: = positive, negative or neutral
Notations: d_r: dictionary_word_rating, T:Tweets, C:Corpus
While T in C do
    while words in tweet do
        if word = = any phrase in dictionary then
            word_rating = d_r;
            continue;
        end
    end
    avg_rating = avg(word_rating)
    if avg_rating ≥ = 0.0 then
        Given tweet is positive
    end
    else if avg_rating < 0.0 then
        Given tweet is negative
    end
    else
        Given tweet is neutral
    end
end
```

Fig. 4. Pseudo algorithm sentiment classification

There are positive and negative tweets based on the number of stars they receive. Sentimental words are removed from tweets, and those that remain are categorized as neutral. Detection of emotional words Dictionary-based sentiment analysis is used for this purpose.

4 Simulated Results

Sentiment analysis and HiveQL processing are performed on tweets associated with trending topics stored in HDFS. Step-by-step figures are provided below.

4.1 Loading the Tweets and Feature Extraction

The below figure shows the Hadoop distributed file systems directory with category, Execution, hashtags, command, and a local host to represent.csv files (Fig. 5).

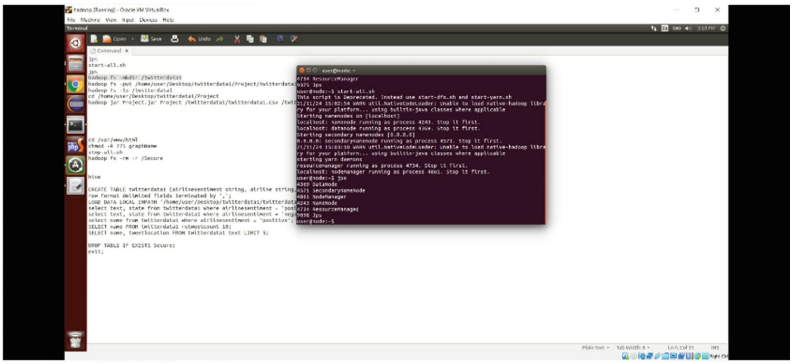


Fig. 5. Hadoop directory file system

A Hive UDF function is used to split the tweet into words to identify sentiment words. An array of words and a tweet id are both stored in a Hive table. We used a built-in UDTF function to extract each word from an array and create a new row for each word because an array contains multiple words. The id and word are stored in a separate table.

The loaded dictionary must be mapped to the tokenized words to rate them. Tables with ids, words, and dictionaries were joined in a left outer join. Words that match sentiment words in the dictionary are given ratings or NULL values if they match the sentiment words in the dictionary. The id, word, and rating are all stored in a hive table. We now have an id, a word, and a rating after completing the above steps. Then a "id" operation is performed to group all words in a tweet, after which an average operation is performed on the ratings given to each word in a tweet (Fig. 6).

Despite this, human assessors and computer systems will make very different errors, and so the figures are not completely comparable.

5 Conclusion

Twitter is the most widely used online platform for data mining. Twitter and Hadoop's ecosystems are used to perform sentimental analysis on the information that it is streamed online. As a result, in this research, a framework was incorporated into the live broadcast

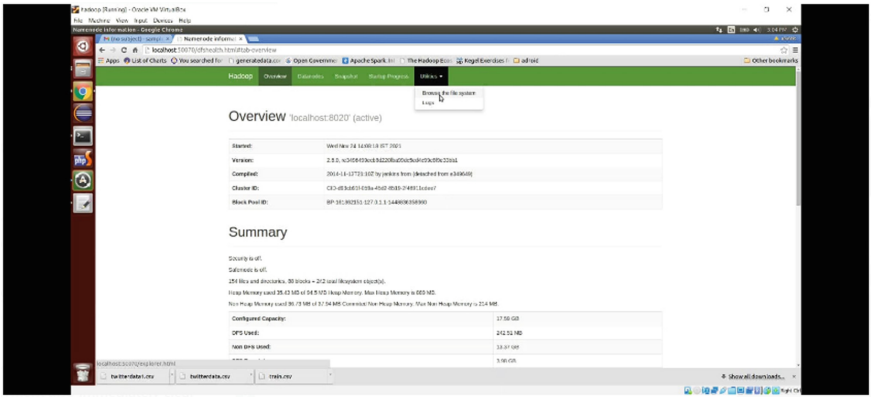


Fig. 6. Shows the local host and summary of system calculations

of Twitter information to discover the public's perceptions of each concept under investigation. Unstructured Twitter data was used for the sentimental analysis, and tweets were retrieved at the time they were posted. Live broadcasting information is retrieved from Twitter as part of the proposed methodology.

Sentiment analysis systems are judged on their correctness by how closely their results match human perceptions. Precision and recall over the two categories of negative and positive texts are commonly used to measure this. Research shows that only around 80% of the time do human raters agree with each other (see Inter-rater reliability). Although 70% accuracy in sentiment classification may not sound like much, a programme that achieves this level of accuracy is performing almost as well as people.

References

1. Danthala, M.K.: Tweet analysis: Twitter data processing using Apache Hadoop. Int. J. Core Eng. Manag. (IJCEM) **1**(11), 94–102 (2015)
2. Mahalakshmi, R., Suseela, S.: Big-SoSA: social sentiment analysis and data visualization on big data. Int. J. Adv. Res. Comput. Commun. Eng. **4**(4), 304–306 (2015)
3. Judith Sherin Tilsha, S., Shobha, M.S.: A survey on Twitter data analysis techniques to extract public opinion. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(11), 536–540 (2015)
4. Ramesh, R., Divya, G., Divya, D., Kurian, M.K.: Big data sentiment analysis using Hadoop. Int. J. Innov. Res. Sci. Technol. **1**(11), 26–35 (2015)
5. Kumar, P., Rathore, V.S.: Efficient capabilities of processing of big data using Hadoop map reduce. Int. J. Adv. Res. Comput. Commun. Eng. **3**(6), 7123–7126 (2014)
6. Furini, M., Montangero, M.: TSentiment: on gamifying Twitter sentiment analysis. In: IEEE ISCC 2016 Workshop: DENVECT, Messina (2016)
7. Olson, S., Downey, A.: Sharing clinical research data: workshop summary. In: National Academies Press Book Sharing Clinical Research Data: Workshop Summary (2013)
8. Barskar, A., Phulre, A.: Opinion mining of social data using Hadoop. Int. J. Eng. Sci. Comput. **6**(1), 3849–3851 (2016)
9. Kaushal, Koundal, D.: Recent trends in big data using Hadoop. Int. J. Inform. Commun. Technol. **8**(2), 39–49 (2019)

10. Wankhede, M.: Analysis of social data using Hadoop ecosystem. Int. J. Comput. Sci. Inf. Technol. **7**(1), 2402–2404 (2016)
11. Ganesh: Performance evaluation of cloud service with Hadoop for Twitter data. Indones. J. Electr. Eng. Comput. Sci. **13**(2), 392–404 (2019)
12. Singh, J.: Big data: tools and technologies in big data. Int. J. Comput. Appl. **112**(1), 6–10 (2015)
13. Vinutha, Raju, T.: An accurate and efficient scheduler for Hadoop MapReduce framework. Indones. J. Electr. Eng. Comput. Sci. **12**(2), 1132–1142 (2018)
14. Barskar, P.A.: Opinion mining of Twitter data using Hadoop and Apache Pig. Int. J. Comput. Appl. **158**(1), 1–6 (2017)
15. Rodrigues, A.P.: Sentiment analysis of social media data using Hadoop framework: a survey. Int. J. Comput. Appl. **151**(1), 119–123 (2016)
16. Patil, N.: Twitter sentiment analysis using Hadoop. Int. J. Innov. Res. Comput. Commun. Eng. **4**(4), 8230–8236 (2016)
17. Yan, P.: MapReduce and semantics enabled event detection using social media. J. Artif. Intell. Soft Comput. **7**(3), 201–213 (2017)
18. Ed-Daoud, Maalmi, A.K.: Real-time machine learning for early detection of heart disease using big data approach. In: 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), pp. 1–5. IEEE, Morocco (2019)
19. El Abdouli, A., Hassouni, L.: A distributed approach for mining Moroccan hashtags using Twitter platform. In: 2nd International Conference on Networking, Information Systems and Security 2019 Proceedings, pp. 1–10. IEEE, Morocco (2019)
20. Kumar, A., Singh, M.: Fuzzy string-matching algorithm for spam detection in Twitter. In: International Conference on Security and Privacy 2019. LNCS, vol. 26, pp. 289–301. Springer, Singapore (2019)
21. Alotaibi, S., Mehmood, R.: Sehaa: a big data analytics tool for healthcare symptoms and diseases detection using Twitter, apache spark, and machine learning. Appl. Sci. **10**(4), 1398–1406 (2019)
22. Kafeza, E., Kanavos, A.: T-PCCE: Twitter personality based communicative communities' extraction system for big data. IEEE Trans. Knowl. Data Eng. **32**(1), 1625–1638 (2019)
23. Tripathi, A.K., Bashir, A.: A parallel military dog-based algorithm for clustering big data in cognitive industrial internet of things. IEEE Trans. Ind. Inform. **17**(2), 2134–2142 (2020)
24. Wang, G., Liu, M.: Dynamic trust model based on service recommendation in big data. Comput. Mater. Contin. **58**(2), 845–857 (2019)
25. Ramaraju, R., Ravi, G.: Sentimental analysis on Twitter data using Hadoop with spring web MVC. In: Intelligent System Design, pp. 265–273. Springer, Singapore (2019)
26. Hussain, S.A.: Prediction and evaluation of healthy and unhealthy status of COVID-19 patients using wearable device prototype data. MethodsX **9**(27), 226–238 (2022)



A Review on Applications of Computer Vision

Gaurav Singh[1(✉)], Parth Pidadi[2], and Dnyaneshwar S. Malwad[1]

[1] AISSMS College of Engineering, Pune, India
gauravsingh13020@gmail.com
[2] Modern Education Society's College of Engineering, Pune, India

Abstract. Humans and animals acquire most of their information through visual systems. The sense of sight in the machine is provided by computer vision technology. Nowadays, vision-based technologies are playing a vital role in automating various processes. With the improvement in computation, powers, and algorithms computer vision performs different tasks such as object detection, human and animal action recognition, and image classification. The application of computer vision is growing continuously in various industries. In this paper, we summarize computer vision implementation in popular fields such as astronomy, medical science, the food industry, the manufacturing industry, and autonomous vehicles. We provide the different computer vision techniques and their performance in various areas. Finally, a brief overview is presented on future development.

Keywords: Computer vision · Object detection · Astronomy · Medical science · Manufacturing industry · Autonomous vehicle

1 Introduction

Artificial Intelligence's major objective is to make computers intelligent so that they can act intelligently. Artificial intelligence (AI) systems are more generic, have the ability to reason, and are more adaptable. Reasoning, learning, problem solving, linguistic intelligence, and perception are basic components of AI. Artificial intelligence is a field of interest for many researchers and is a technology which has applications in various fields such as language processing, robotics, speech recognition, computer vision as shown in Fig. 1 [1–4].

Natural language processing makes use of computational approaches to study, comprehend, and create human language content. NLP is now being utilized to develop speech-to-speech translation engines and spoken dialogue systems, as well as to mine social media for health and financial information and to discern sentiment and emotion toward products and services [1, 5]. Robotics is a field that deals with the planning and operation of robots, as well as the use of computers to control and prepare them. In the manufacturing industry, robots are used to speed up the assembly process [2]. Speech recognition can be described as a process of making a computer capable of understanding and respond to human speech [6]. Speaker verification and recognition are two components of speech recognition. The former needs to identify whether a sound is a database sample, whereas the latter requires determining which sound sample is in the database

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 464–479, 2023.
https://doi.org/10.1007/978-3-031-27409-1_42

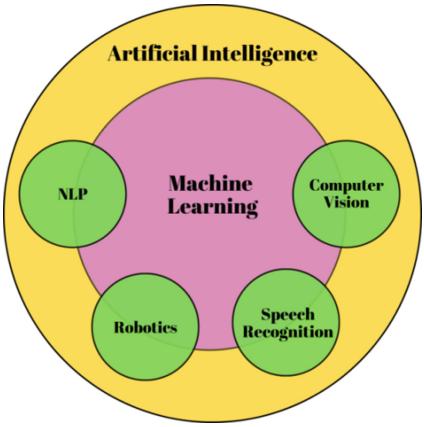


Fig. 1. Fields of artificial intelligence.

[3]. Acquisition of speech feature signals, preprocessing, feature extraction, biometric template matching, and recognition outcomes are all part of the speech recognition process [7]. In computer vision, a machine generates an understanding of image information by applying algorithms to visual information. The understanding can be converted into pattern observation, classification, etc. [4].

Ever since the introduction of the first digital computers in the 1960s, many people have been trying to perform image analysis using early age computers. A few early successes were in the form of character recognition, fingerprint and number plate recognition soon followed the list [8]. In the course of years image recognition using computers has come a long way and become more and more sophisticated. Computer Vision (CV) has applications in a wide variety of domains from the detection of stars and galaxies to self-driving cars [9–11]. Since the advent of deep learning, traditional computer vision can only perform only a handful tasks. With all this manual work, still the error margins were high. Machine learning introduced a solution to this problem of manual work [12, 13]. With ML algorithms, developer no longer needed to code every parameter in their CV application. Machine learning helped solve many problems which were historically impossible, but required many developers working at the same time on the same project and implementing ML algorithms for CV applications [14]. Deep learning (DL) provided with a different approach to ML algorithms, DL relies on neural networks to solve vision problems [14, 15]. Neural networks are general purpose functions that can solve problems that can be represented through examples. Neural network is able to extract common patterns from labeled data and transform them into mathematical equations that can be used to classify future data [16].

Over the last few years, AI has been widely used in service industries such as retail, e-commerce, entertainment, logistics, banking and finance services. There are various review papers on artificial intelligence which focuses on the application of various CV fields in direct consumer related services. Nowadays many researchers are focusing on implementation of CV in non-service industries. Computer vision is one of the emphatic fields in AI for manufacturing, medical science and astronomy. This paper presents a

comprehensive review on application of CV in emerging domains like astronomy, medical science, manufacturing, construction, food industry and review how CV techniques are applied in various domains to optimize, simplify and automate various manual cumbersome tasks. The paper discusses the development in computer vision technologies, their role in the growth of various sectors and addresses recommendations for further applications (Fig. 2).

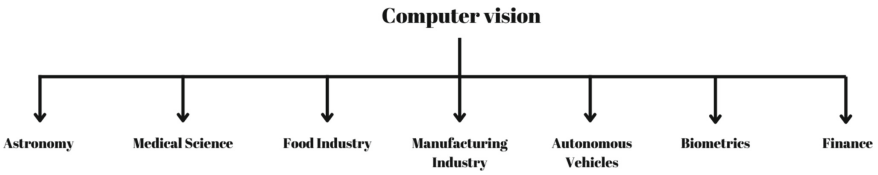


Fig. 2. Applications of computer vision in various fields

2 Computer Vision Applications

In this section we present a comprehensive review of application of CV in various domains. The reviewed work is then grouped based on the domain in which computer vision is applied. Table 1 gives a detail list of reviewed approaches and their main features.

2.1 Astronomy

The below section discusses the implementation of computer vision algorithms in astronomical applications and how CV is used in tackling problems from astronomy point of view. Most of the applications have used datasets generated from Sky Image Cataloging and Analysis System (SKICAT) [17], and the Palomar Observatory Sky Survey [18] and many other sky surveys. Classification of astronomical applications is shown in Fig. 3.

Object Classification

In scientific process, object classification is one of the essential steps, as it provides us with key insights to our datasets and helps make optimal decisions and minimize errors. Completeness and efficiency are two important quantities for astrophysical object classification. Classification of stars, galaxies and other astrophysical objects such as supernovae, quasars, stars and galaxies from photometric datasets is an important problem because it is an incommodious task to manually classify the objects into different categories. This is a challenging task as, stars are unresolved in the photometric datasets because of their large distance from earth, however, and galaxies being further away appear as extended sources. Moreover, other astrophysical objects also appear as point sources such as supernovae and quasars; hence, it becomes difficult to classify them. Classification based on machine learning and computer vision can accelerate workflow

Table 1. Summary of literature survey done in the field of computer vision.

Author	Methodology	Remark
Astronomy		
D. J. Mortlock et al.	Bayesian analysis	Study of quasars beyond redshift of z = 7.085
A. A. Collister and O. Lahav, S. C. Odewahn et al., David Bazell and Yuan Peng, S. C. Odewahn et al., David Bazell and Yuan Peng, Moonzarin Reza, R. Carballo et al., Ofer Lahav et al.	ANN	Classification of galaxy based on morphological data and photometric parameters, stellar/non-stellar objects
S. Dieleman et al.	CNN	Automating classification of galaxy based on morphological data
N. S. Philip et al., N. Mukund et al.	DBNN	Classifying galaxies using Difference Boosting Neural Network (DBNN)
N. Weir, U. M. Fayyad and S. Djorgovski, N. M. Ball et al., Joseph W. Richards et al.	Decision tree	Automated approach to galaxy classification for identification of galaxies in photometric datasets
Peng et al.	SVM	Quasar classification using several SVM with 93.21% efficiency
Medical science		
Pawan Kumar Upadhyay et al.	Coherent CNN	Obtained 97% accuracy for retinal disease detection
Nawel Zemmal et al.	Transductive SVM	Detection of glaucoma and feature extraction using grey level co-occurrence matrix
Shouvik Chakraborty et al.	SUFACSO	Detection of covid-19 using CT scan images
Maitreya Maity et al.	C4.5 decision tree	Detection of anemia due to reduced level of hemoglobin
Wilson F. Cueva	ANN	Using ANN to identify "Melanoma" with 97.51% accuracy
Food processing		

(*continued*)

Table 1. (*continued*)

Author	Methodology	Remark
Liu et al., 2016	PLSDA, PCA-BPNN, LS-SVM	Comparison of various methods for rice seed classification based on variety
Kaur and Singh, 2013	SVM	Comparison of various methods for rice seed classification based on quality
Olgun et al., 2016	SIFT + SVM	Classification of wheat grains
Kadir Sabanci, n.d.	ANN, SVM, decision tree, kNN	Classification of bread and durum wheat
Xia et al., 2019	MLDA, LS-SVM	Maize seed classification into 17 varieties
Huang and Lee, 2019	CNN	Classification of coffee by using CNN
Manufacturing		
Christos Manettas et al., Imoto et al.	CNN	Object orientation classification wafer surface defect detection for manufacturing processes
Scime and Beuth	Bag of key points using SVM	Detection of faults during additive manufacturing processes
Self-driving cars		
Gupta et al.	Deep learning	A survey of deep learning techniques for autonomous vehicles
Novickis et al., Chen et al.	CNN	Proposed architecture for pedestrian detection using multiple cameras
Muthalagu et al.	Linear regression	Improvements in lane detection systems

and help astrophysicists to focus on other important problems [19]. Odewahn et al. have discussed methods for the application of star/galaxy discrimination. In their work they have implemented artificial neural network and successfully classified stellar and non-stellar categories based on 14 element image parameters set. Simple numerical experiments were conducted to identify significant image parameters for separation of galaxies and stars and to illustrate the robustness of the model [18]. Classification of galaxies based on morphology using automated machine learning can be done using various ML algorithms like ANN, ET, DT, RF, kNN [20].

Support vector machine (SVM) a machine learning algorithm can be used to demonstrate identification of quasars in sky survey datasets like SDSS, UKIDSS, and GALEX.

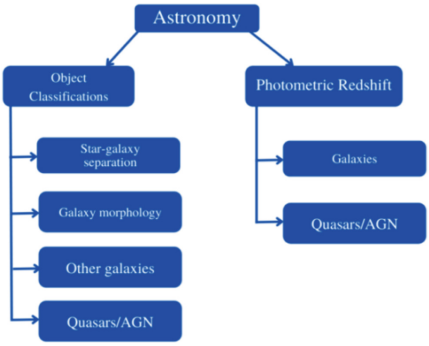


Fig. 3. Applications of computer vision in astronomy

Employment of a hierarchy of SVM classifier is suggested in this approach. According to the study, results obtained through experiments show using multiple SVM classifiers is more useful than using single SVM classifier for distinguishing astronomical objects. Cross validation for increasing confidence can be done by selecting candidates by using a previously mentioned approach [21].

Photometric Redshift for Various Astrophysical Objects

A photometric redshift is an estimate of an astronomical object's recession velocity made without measuring its spectrum, such as a galaxy or quasar [22]. Photometry is used to estimate the observed object's redshift, and hence its distance, thanks to Hubble's law [23] investigates a new method for predicting photometric redshifts based on artificial neural networks (ANNs). ANNs require a large spectroscopically identifiable training set, unlike the traditional template-fitting photometric red shift methodology.

Other Applications in Astronomy

Classification of variable stars has been the center of attraction of many astrophysicists. It is important to classify variable stars to reveal the underlying properties like mass, luminosity, temperature, internal and external structure, composition, evolution and other stellar properties. Richards et al. [24] presented tree-based classification approaches for variable stars. Random forests, classification and, regression trees and boosted trees algorithms are compared with previously used SVMs, Gaussian mixture models, Bayesian averaging of artificial neural networks, and Bayesian networks (Fig. 4).

The best classifier in terms of total misclassification rate is an RF with B = 1000 trees, which achieves a 22.8% average misclassification rate. The HSC–RF classifier with B = 1000 trees has the lowest catastrophic error rate of 7.8%.

2.2 Medical Science

The area of medical research has been transformed and has experienced huge evolution at an accelerated rate in numerous sectors of medical science such as neurological illness detection, facial recognition, retinal problems, and much more since the introduction of

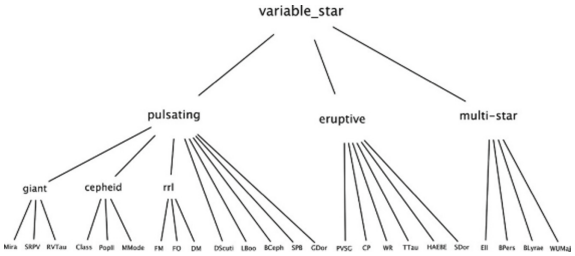


Fig. 4. Hierarchy from data set used for variable star classification [24].

machine learning and computer vision. Recent advancements in picture categorization and object identification can considerably assist medical imaging.

Detection of Retinal Disorders

Optical coherence tomography (OCT) pictures used to diagnose retinal disorders using machine learning methods [25]. Devastating diseases like cataract, glaucoma has become one of the diseases to cause eye blindness. Machine learning based models are used to reduce cumbersome task of eye diseases detection by automating the process. The genetic algorithm and a transductive SVM wrapper technique are utilized. The RIM-ONE data set, which acts as a benchmark, was utilized to validate the author's suggested approach. As a result, the RIM-ONE is utilized to improve algorithms. For this task, RIM-ONE R3 database is used. Seeing the feasibility of direct images classification, grey-level co-occurrence matrix (GLCM) and a descriptor vector is opted for feature extraction. The descriptor vector consists of thirteen suitable features extracted from the matrix [26]. Chakraborty and Mali [27], discussed methods for detection of COVID-19 using radiological image analysis of CT-scan images. The author has proposed a superpixel-based new technique to segmenting CT scan images is provided in this study to deal with this circumstance well and for speeding up the resting process of the novel coronavirus infection.

Analyzing pathophysiological changes in erythrocytes is crucial for detecting anemia early. Anemia is the most prevalent blood condition in which the red blood cell (RBC) or blood hemoglobin is deficient. Image processing tools have been used by the author. This study uses a thin blood smear to describe infected erythrocytes. 100 patients between the ages of 25 and 50 are chosen at random and blood samples are taken from each of them. Each blood sample is then processed into thin smear blood slides [28] (Fig. 5).

Skin cancer can be detected using computer aided technologies as demonstrated by various studies. As the scientific literature suggests skin cancer if not diagnosed at an early stage can be life-threatening, and early detection of skin cancer such as melanoma is an indication of high chances of survival. The present work was based only on the Asymmetry, Border, Color, and Diameter [29].

2.3 Food Grains

With continuous population expansion, the food business must continue to increase output while also enhancing product quality. To boost productivity, improvements in

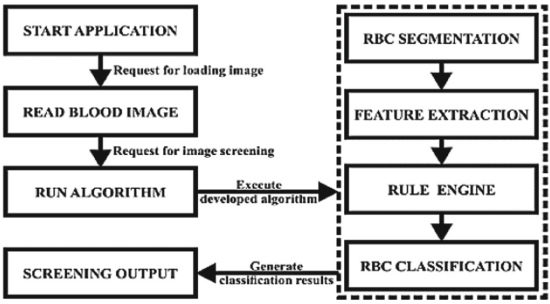


Fig. 5. Workflow diagram of the proposed screening system [28].

the manufacturing chain are essential. One of these advancements is the automation of food grain categorization, which has received a lot of attention in recent years as novel techniques to doing automatic classification have been proposed.

Rice

Rice seeds come in a variety of sizes, colors, shapes, textures, and constitutions, which can often be difficult to distinguish with the naked eye. Traditional rice variety discrimination methods rely mostly on chemical and field approaches, both of which are damaging, time-consuming, and complicated, and are not suited for sorting and online measurements. As a result, finding a nondestructive, easy, and quick approach for categorizing rice types would be extremely beneficial. The study suggests use of PLSDA, PCA-BPNN, and LS-SVM. Finally, using multispectral imaging in conjunction with chemometric techniques to detect rice seed types is a particularly appealing technique since it is nondestructive, simple, and rapid, and it does not require any preparation [30] (Fig. 6).

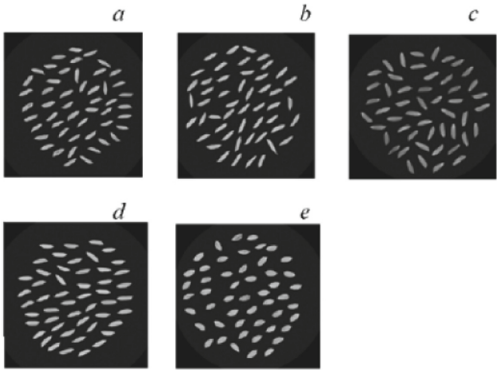


Fig. 6. Images of rice varieties.

Determination of quality of rice can depend on various factors, such as, color, density, shape, size, number of broken kernels and chalkiness. Human inspection of rice quality is

neither objective nor efficient. Many researches have used image processing to examine grain quality. Computer vision (CV) is a technology for inspection and assessment that is quick, inexpensive, consistent, and objective. Using Multi-Class SVM, the author offers a machine technique to grade rice kernels. The Support Vector Machine assisted in accurately grading and classifying rice kernels (better than 86%) at a low cost. Based on the findings, it can be determined that the method was adequate for categorizing and grading various rice types based on their internal and external qualities [31] (Fig. 7).

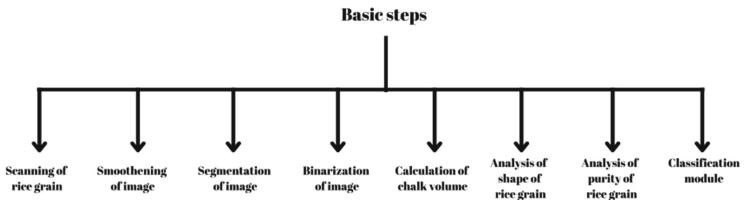


Fig. 7. Basic steps for grading of rice and classification [31].

Wheat

Wheat is a key food source worldwide, and it is widely farmed in most nations. It can adapt to a variety of habitats, including both irrigated and dry soil. Wheat manufacturing requires certified pure grain, and grains should not be combined with various genotypes throughout the production process. Commercially two groups are made for wheat classification: grain hardness and appearance. The study suggests an automated method that can accurately categorize wheat grains. The accuracy of DSIFT is examined for this purpose by focusing on the SVM classifier. Initially, k-means is applied to DSIFT features for clustering, then, by generating the Bag of Words of visual words; photos are represented using histograms of characteristics. The proposed technique achieves 88.33% rate by conducting an experimental research on a specific data set [32].

To extract the visual features of grains or things, computer vision systems employ image processing technologies. Computer vision and artificial intelligence (AI) can be used to give autonomous quality evaluation. As a result, a fast, unmanned system with excellent accuracy for grain classification may be constructed. A basic computer vision-based application is given that uses a multilayer perceptron (MLP)-based artificial neural network (ANN) to properly categorize wheat grains into bread or durum [33] (Fig. 8).

Corn

For determining quality of seeds and classifying them, seed purity can be used as an essential criterion. For the classification of seed types, for 1632 maize seeds (17 varieties), hyperspectral images between 400 and 1000 nm were obtained. The classification accuracy improved with use method of combining features based on MLDA wavelength selection method. Meanwhile, the classification model based on the MLDA feature transformation/reduction approach outperformed successive projections algorithm (SPA) with linear discriminant analysis (LDA) (90.31%) and uninformative variable

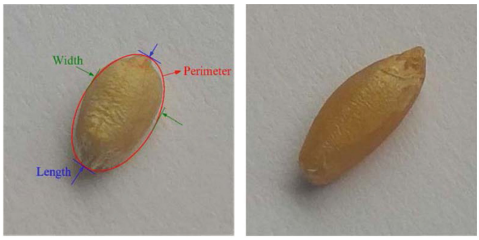


Fig. 8. Bread wheat versus durum wheat [33]

eliminations with LDA (94.17%) in terms of classification accuracy, and increased by 2.74% when compared to the mean spectrum [34].

Coffee

Coffee is one of the important commercial crops and a highly consumed drink in human culture, due to its high caffeine content. Huang and Leeused the Convolutional Neural Network (CNN), a prominent deep learning method, to preprocess photos of raw beans of coffee collected by image processing technology. CNN excels at extracting color and structure from photos. As a result, we can quickly distinguish between excellent and poor bean pictures, with incomplete blackness, brokenness, etc. The author used their own technology to swiftly determine which green beans were excellent and which were poor. Using this strategy, the time spent manually selecting coffee beans may be cut in half, and the creation of specialty coffee beans can be accelerated. Model based on CNN to distinguish between excellent coffee beans and poor coffee beans, a total accuracy of 93.343% was obtained with a false positive rate of 0.1007 [35, 36] (Fig. 9).

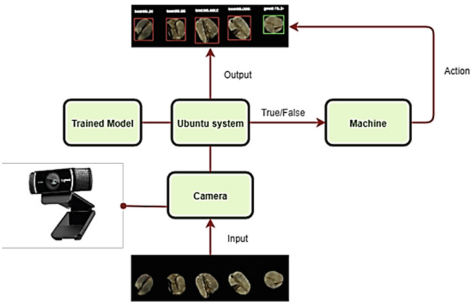


Fig. 9. Architecture of identification [35]

2.4 Manufacturing

Machine Learning-based Artificial Intelligence applications are commonly regarded as promising industrial technology. Convolutional Neural Networks (CNN) and other Deep

Learning (DL) techniques are effectively applied in various computer vision applications in manufacturing. Deep learning technology has recently improved to the point that it can conduct categorization at the level of a human, as well as give powerful analytical tools for processing massive data from manufacturing [37–39].

Study suggests using convolutional neural network (CNN) for classification of object orientation using synthetic data. A full, synthetic data-based positioning estimation of manufacturing components may be justified as a viable idea with potential applications in production. Several images of resolution 3000 × 4000 pixel are taken using camera placed on top of workbench [40].

Automatic defect categorization feature sorts defective photos into pre-defined defect classifications morphologically. To compare the obtained accuracy of automated defect categorization approach with the suggested approach, wafer-surface-defect SEM pictures from a real manufacturing plant were used. All anomalies were transformed to a consistent image of 128 × 128 pixels for the experiment. Four sets of defect image data were created [38]. The classification accuracy of the author's suggested approach and the commercially available conventional ADC system were 77.23% and 87.26% respectively (Fig. 10).

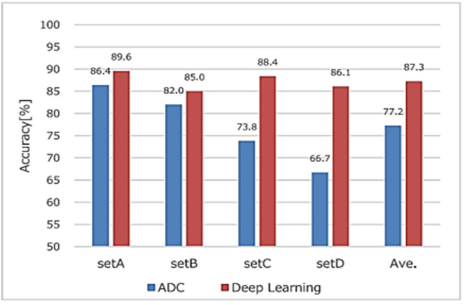


Fig. 10. Comparison of automated defect categorization and proposed method [38]

Additive Manufacturing, sometimes known as 3D printing, has seen tremendous growth in recent years, especially for equipment and techniques that produce various metal objects. Scime and Beuth [41] has implemented computer vision techniques for manufacturing carried out using Laser Powder Bed Fusion (LPBF) machines. The authors chose a machine learning strategy over human design of anomaly detectors because of its inherent flexibility. In 100% of circumstances; the algorithm is able to determine existence of zero anomalies with high accuracy. Finally, in 89% of situations, the algorithm is able to accurately detect the presence of an anomaly. This method is a unique Additive Manufacturing application of modern computer vision techniques.

2.5 Autonomous Vehicles

Various AI, ML and DL methods have acquired popularity and step forward as a result of recent advancements in these approaches. Self-driving vehicles are one such application, which is expected to have a significant and revolutionary influence on society and the

way people commute. However, in order for these automobiles to become a reality, they must be endowed with the perception and cognition necessary to deal with high-pressure real-life events, make proper judgments, and take appropriate and safe action at all times.

Autonomous cars are based on a shift from human-centered autonomy to entirely computer-centered autonomy, in which the vehicle's AI system regulates and controls all driving responsibilities, with human involvement required only when absolutely essential [42, 43] (Fig. 11).



Fig. 11. Automation levels by SAE.

Object Detection

Camera data is received by the camera object detection module. If the embedded computing capacity is adequate, each camera's data is analyzed by its own object detector; however, if resources are limited, a deep neural network model is applied to images obtained from merging multiple camera frames, which is currently YOLOv3 (real time object detection model) architecture. Object identification using 'RADAR' might potentially be done using a deep neural network model, which can recognize moving points and clusters. In SS-DNN the obtained images are processed by object detector as well as camera, however, it assigns a label to each pixel in the frame. The above steps are necessary to establish the car's mobility and also the parts of road markings and road signs that need to be taken out for CNN-based classifiers. Perception ANNs are used to construct three separate vehicle surrounding maps based on the data they collect [44] (Fig. 12).

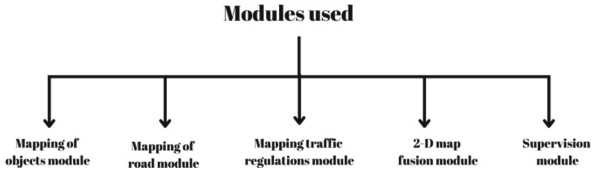


Fig. 12. Modules

Pedestrian Detection

A self-driving car's ability to identify pedestrians automatically and reliably is critical. The information is gathered while driving on city streets. In total 58 data sequences were retrieved from almost 3 h of driving on city streets throughout many days and illumination environments. There are a total of 4330 frames. The author designed and manufactured a one-of-a-kind test equipment rig to collect data for pedestrian detection

on the road. The data gathering system aboard the test vehicle may now be mobile thanks to this design.

Only the HOG and CCF algorithms for pedestrian detection are compared in this research. For detection on multiple scale levels, HOG features are integrated with SVM and the sliding window approach. CCF uses low-level information from a pre-trained CNN model cascaded with a boosting forest model such as Real AdaBoost as a classifier. The dataset contains 58 video sequences that have been labeled. The author has utilized 39 for training and the other 19 for testing. The experimental findings reveal that CCF outperforms HOG features substantially. In CCF approach when thermal and color images were used in combination resulted in peak performance leading to 9% log-average miss rate [45].

Lane Detection

Perception, planning, and control are the three basic building elements of self-driving automobile technology. The goal of this research is to create and test a perception algorithm that leverages camera data and computer vision to aid autonomous automobiles in perceiving their surroundings. Cameras are the most closely related technology to how people see the environment, and computer vision is at the foundation of perception algorithms. Though Lidar and radar systems are being employed in the development of perception technologies, cameras present us with a strong and less expensive means of obtaining information about our surroundings.

A strong lane recognition method is addressed in this work, which can estimate the safe drivable zone in front of an automobile [46]. Then, using perspective transformations and histogram analysis, the author presents an improved lane detecting approach which overcomes the limitations of minimalistic approach for lane detection [47]. Curved and straight can also be spotted using the above approach.

3 Limitations and Challenges

Choosing the right algorithm and finding the right dataset can be difficult in computer vision. Underfitting and overfitting can occur as a result of a small or large number of datasets. The amount of data required to improve accuracy by even a small margin is enormous. The majority of real-world data is unlabeled, and a great deal of effort is expended in labeling the data. To process photometric datasets, computer vision requires more computational power. A variety of limitations may arise as a result of poor camera quality [48].

4 Conclusion

Computer vision technology can be effectively implemented in industries which are depending on image and video information. Many industries are adopting AI to transform their business to next level for them computer vision is driving force. This review presents capability of computer vision in astronomy, medical science, food industry, manufacturing industry and autonomous vehicles. The algorithms and methods suitable

for each industry will be helpful guideline for the researcher working in that area. Computer vision is not only used for classification of objects on earth but also in the universe beyond Earth's atmosphere. This study aims to an inspiring map to implement computer vision in wide range of industries.

Acknowledgements. Not applicable.

Credit. Gaurav Singh: Conceptualization, Writing-Original Draft.
Parth Pidadi: Writing-Review and Editing, Resources.
Dnyaneshwar S. Malwad: Project administration.

Data Availability Statement. My manuscript has no associated data.

Compliance with Ethical Standards.

Conflict of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Hirschberg, J., Manning, C.D.: Advances in natural language processing. In: A Companion to Cognitive Science, pp. 226–234 (2008). https://doi.org/10.1002/9781405164535.ch14
2. Balakrishnan, S., Janet, J.: Artificial intelligence and robotics: a research overview (2020)
3. Zhang, X., Peng, Y., Xu, X.: An overview of speech recognition technology. In: Proceedings of the 2019 4th International Conference on Control, Robotics and Cybernetics (CRC), pp. 81–85 (2019). https://doi.org/10.1109/CRC.2019.00025
4. Feng, X., Jiang, Y., Yang, X., et al.: Computer vision algorithms and hardware implementations: a survey. Integration **69**, 309–320 (2019). https://doi.org/10.1016/j.vlsi.2019.07.005
5. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. J. Am. Med. Inform. Assoc. **18**, 544–551 (2011). https://doi.org/10.1136/amiajnl-2011-000464
6. Niemueller, T., Widyadharma, S.: Artificial intelligence—an introduction to robotics. Artif. Intell. 1–14 (2003)
7. Gaikwad, S.K., Gawali, B.W., Yannawar, P.: A review on speech recognition technique. Int. J. Comput. Appl. **10**, 16–24 (2010). https://doi.org/10.5120/1462-1976
8. Khan, A.A., Laghari, A.A., Awan, S.A.: EAI endorsed transactions machine learning in computer vision: a review. 1–11 (2021)
9. Badue, C., Guidolini, R., Carneiro, R.V., et al.: Self-driving cars: a survey. Expert Syst. Appl. **165**, 113816 (2021). https://doi.org/10.1016/j.eswa.2020.113816
10. Ball, N.M., Brunner, R.J., Myers, A.D., Tcheng, D.: Robust machine learning applied to astronomical data sets. I. Star-galaxy classification of the Sloan Digital Sky Survey DR3 using decision trees. 497–509
11. Ball, N.M., Loveday, J., Fukugita, M., et al.: Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. **1046**, 1038–1046 (2004). https://doi.org/10.1111/j.1365-2966.2004.07429.x

12. Kardovskyi, Y., Moon, S.: Automation in construction artificial intelligence quality inspection of steel bars installation by integrating mask R-CNN and stereo vision. Autom. Constr. **130**, 103850 (2021). https://doi.org/10.1016/j.autcon.2021.103850
13. Odewahn, S.C., Nielsen, M.L.: Star-galaxy separation using neural networks. **38**, 281–286 (1995)
14. Hanocka, R., Liu, H.T.D.: An introduction to deep learning. In: ACM SIGGRAPH 2021 Courses, SIGGRAPH 2021, pp. 1438–1439 (2021). https://doi.org/10.1145/3450508.3464569
15. Chai, J., Zeng, H., Li, A., Ngai, E.W.T.: Deep learning in computer vision: a critical review of emerging techniques and application scenarios. Mach. Learn. Appl. **6**, 100134 (2021). https://doi.org/10.1016/j.mlwa.2021.100134
16. Abiodun, O.I., Jantan, A., Omolara, A.E., et al.: State-of-the-art in artificial neural network applications: a survey. Heliyon **4**, e00938 (2018). https://doi.org/10.1016/j.heliyon.2018.e00938
17. Weir, N.: Automated star/galaxy classification for digitized POSS-II. **109**, 2401–2414 (1995)
18. Odewahn, S.C., Stockwell, E.B., Pennington, R.L., et al.: Automated star/galaxy discrimination with neural networks **103**, 318–331 (1992)
19. Ball, N.M., Brunner, R.J.: Data mining and machine learning in astronomy (2010)
20. Reza, M.: Galaxy morphology classification using automated machine learning. Astron. Comput. **37**,(2021). https://doi.org/10.1016/j.ascom.2021.100492
21. Peng, N., Zhang, Y., Zhao, Y., Wu, X.: Selecting quasar candidates using a support vector machine classification system 1 introduction. **2609**, 2599–2609 (2012). https://doi.org/10.1111/j.1365-2966.2012.21191.x
22. Zheng, H., Zhang, Y.: Review of techniques for photometric redshift estimation. Softw. Cyberinfrastruct. Astron. II **8451**, 845134 (2012). https://doi.org/10.1117/12.925314
23. Firth, A.E., Lahav, O., Somerville, R.S.: Estimating photometric redshifts with artificial neural networks 2 artificial neural networks. **1202**, 1195–1202 (2003)
24. Richards, J.W., Starr, D.L., Butler, N.R., et al.: On machine-learned classification of variable stars with sparse and noisy time-series data. Astrophys. J. **733**,(2011). https://doi.org/10.1088/0004-637X/733/1/10
25. Upadhyay, P.K., Rastogi, S., Kumar, K.V.: Coherent convolution neural network based retinal disease detection using optical coherence tomographic images. J. King Saud. Univ. – Comput. Inf. Sci. (2022). https://doi.org/10.1016/j.jksuci.2021.12.002
26. Zemmal, N., Azizi, N., Sellami, M., et al.: Robust feature selection algorithm based on transductive SVM wrapper and genetic algorithm: application on computer-aided glaucoma classification. Int. J. Intell. Syst. Technol. Appl. **17**, 310–346 (2018). https://doi.org/10.1504/IJISTA.2018.094018
27. Chakraborty, S., Mali, K.: A radiological image analysis framework for early screening of the COVID-19 infection: a computer vision-based approach. Appl. Soft Comput. **119**, 108528 (2022). https://doi.org/10.1016/j.asoc.2022.108528
28. Maity, M., Mungle, T., Dhane, D., Maiti, A.K., Chakraborty, C.: An ensemble rule learning approach for automated morphological classification of erythrocytes. J. Med. Syst. **41**(4), 1–14 (2017). https://doi.org/10.1007/s10916-017-0691-x
29. Cueva, W.F., Muñoz, F., Vásquez, G., et al.: Detection of skin cancer "Melanoma" through computer vision. pp. 1–4 (2017)
30. Liu, W., Liu, C., Ma, F., Lu, X., Yang, J., Zheng, L.: Online variety discrimination of rice seeds using multispectral imaging and chemometric methods. J. Appl. Spectrosc. **82**(6), 993–999 (2016). https://doi.org/10.1007/s10812-016-0217-1
31. Kaur, H., Singh, B.: Classification and grading rice using multi-class SVM. **3**, 1–5 (2013)
32. Olgun, M., Okan, A., Özkan, K., et al.: Wheat grain classification by using dense SIFT features with SVM classifier. **122**, 185–190 (2016). https://doi.org/10.1016/j.compag.2016.01.033

33. Sabanci, K., Kayabasi, A., Toktas, A.: Computer vision-based method for classification of the wheat grains using artificial neural network (2017)
34. Xia, C., Yang, S., Huang, M., et al.: Maize seed classification using hyperspectral image coupled with multi-linear discriminant analysis. Infrared Phys. Technol. **103077** (2019). https://doi.org/10.1016/j.infrared.2019.103077
35. Huang, N., Chou, D.-L., Lee, C.: Real-time classification of green coffee beans by using a convolutional neural network. In: 2019 3rd International Conference on Imaging, Signal Processing and Communication, pp. 107–111
36. Huang, N., Chou, D.-L., Wu, F.-P., et al.: Smart agriculture real-time classification of green coffee beans by using a convolutional neural network (2020)
37. Krizhevsky, A., Sutskever, I.: ImageNet classification with deep convolutional neural networks. In: Handbook of Approximation Algorithms and Metaheuristics, pp. 1–1432 (2007). https://doi.org/10.1201/9781420010749
38. Imoto, K., Nakai, T., Ike, T., et al.: A CNN-based transfer learning method for defect classification in semiconductor manufacturing. IEEE Trans. Semicond. Manuf. **32**, 455–459 (2019). https://doi.org/10.1109/TSM.2019.2941752
39. Wang, J., Ma, Y., Zhang, L., et al.: Deep learning for smart manufacturing: methods and applications. J. Manuf. Syst. **48**, 144–156 (2018). https://doi.org/10.1016/j.jmsy.2018.01.003
40. Manettas, C., Nikolaos, K.A.: Synthetic datasets for deep learning in computer-vision assisted tasks in manufacturing: a new methodology to analyze the functional and physical architecture of manufacturing existing pro. Procedia CIRP **103**, 237–242 (2021). https://doi.org/10.1016/j.procir.2021.10.038
41. Scime, L., Beuth, J.: Anomaly detection and classification in a laser powder bed additive manufacturing process using a trained computer vision algorithm. Addit. Manuf. **19**, 114–126 (2018). https://doi.org/10.1016/j.addma.2017.11.009
42. Inagaki, T., Sheridan, T.B.: A critique of the SAE conditional driving automation definition, and analyses of options for improvement. Cogn. Technol. Work **21**(4), 569–578 (2018). https://doi.org/10.1007/s10111-018-0471-5
43. Gupta, A., Anpalagan, A., Guan, L., Khwaja, A.S.: Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. Array **10**, 100057 (2021). https://doi.org/10.1016/j.array.2021.100057
44. Novickis, R., Levinskis, A., Science, C., et al.: Functional architecture for autonomous driving and its implementation (2020)
45. Chen, Z., Huang, X.: Pedestrian detection for autonomous vehicle using multi-spectral cameras. IEEE Trans. Intell. Veh. 1 (2019). https://doi.org/10.1109/TIV.2019.2904389
46. Muthalagu, R., Bolimera, A., Kalaichelvi, V.: Lane detection technique based on perspective transformation and histogram analysis for self-driving cars. Comput. Electr. Eng. **85**, 106653 (2020). https://doi.org/10.1016/j.compeleceng.2020.106653
47. Assidiq, A.A.M., Khalifa, O.O., Islam, R., et al.: Real time lane detection for autonomous vehicles. 82–88 (2008)
48. Khan, A.A., Laghari, A.A., Awan, S.A.: Machine learning in computer vision: a review. EAI Endorsed Trans. Scalable Inf. Syst. **8**, 1–11 (2021). https://doi.org/10.4108/eai.21-4-2021.169418



Analyzing and Augmenting the Linear Classification Models

Pooja Manghirmalani Mishra[1(✉)] and Sushil Kulkarni[2]

[1] Machine Intelligence Research Labs, Mumbai, India
pmanghirmalani@ieee.org
[2] School of Mathematics, Applied Statistics and Analytics, NMIMS, Mumbai, India

Abstract. Statistical learning theory offers an architecture needed for analysing the problem of inference, which includes, gaining knowledge, predictions, decisions or constructing models from a set of data. It is studied in a statistical architecture that is there are assumptions of statistical nature of the underlying phenomena. For predictive analysis, Linear Models are considered. These models tell about the relation between the target and the predictors using a straight line. Each linear model algorithm encodes specific knowledge, and works best when this assumption is satisfied by the problem to which it is applied. To generalize logistic regression to several classes, one possibility is to proceed in the way described previously for multi-response linear regression by performing logistic regression independently for each class. Unfortunately, the resulting probability estimates will not sum to one. In order to obtain proper probabilities, it is essential to combine the individual models for each class. This produces a joint optimization problem. A simple way is address multiclass problems also known as pair-wise classification. In this study, a classifier is derived for every pair of classes using only the instances from these two classes. The output on an unknown test example which is based on the class which receives maximum votes. This method has produced accurate results in terms of classification error. It is further used to produce probability estimates by applying a method called pair-wise coupling, which calibrates the individual probability estimates from the different classifiers.

Keywords: Linear Models · Learning Disability · Statistical Learning · Metric Structure · Linear Classification · Regression

1 Introduction

Statistical learning theory offers an architecture needed for analysing the problem of inference, which includes, gaining knowledge, predictions, decisions or constructing models from a set of data. It is studied in a statistical architecture that is there are assumptions of statistical nature of the underlying phenomena. For predictive analysis, Linear Models are considered. These models tell about the relation between the target and the predictors using a straight line. Each linear model algorithm encodes specific knowledge, and works best when this assumption is satisfied by the problem to which it is applied.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 480–492, 2023.
https://doi.org/10.1007/978-3-031-27409-1_43

To test the functionality of the linear models, a case study of Learning Disability is considered. Learning disability (LD) refers to a neurobiological disorder which affects a person's brain and interferes with a person's ability to think and remember [6]. The learning disabled frequently have high IQs. It is also not a single disorder, but includes disabilities in any of areas related to reading, language and mathematics [9]. LD can be broadly classified into three types. They are difficulties in learning with respect to read (Dyslexia), to write (Dysgraphia) or to do simple mathematical calculations (Dyscalculia) which are often termed as special learning disabilities [4]. LD cannot be cured completely by medication. Children suffering from LD are made to go through a remedial study in order to make them cope up with non-LD children of their age. For detecting LD, there does not exist a global method [7]. While considering the case study, we decided to consider the problem of LD which is a lifelong neuro-developmental disorder that manifests during childhood as persistent difficulties in learning to efficiently read, write or do simple mathematical calculations despite normal intelligence, conventional schooling, intact hearing and vision, adequate motivation and socio-cultural opportunity.

The present method available to determine LD in children is based on check lists containing the symptoms and signs of LD. This traditional method is time consuming, not accurate and obsolete also. Such LD identification facilities are much less at schools or even in cities. If the LD determination facility is attached with schools and the check-ups are arranged as a routine process, LD can be identified at an early stage. Under these circumstances, it is decided to carry out a research work in the topic in a view to increase the diagnostic accuracy of LD prediction. Based on the statistical machine learning tool developed, the presence and degree of learning disability in any child can be determined accurately at an early stage.

2 The Concept of Distance

This concept is basic in calculus [2]. The best way to define closeness is in terms of distance by taking points to be close if the distance between them is small. So, it can be used this to define closeness. Distance is originally a concept in geometry, but this can be made more general if we concentrate on the following three essential properties of distance.

- $(M\,1)$ the distance between two points x and y is a positive real number unless x and y are identical in which case the distance is zero.
- $(M\,2)$ the distance from x to y is the same as the distance from y to x.
- $(M\,3)$ if x, y, z are three points, then the distance between the two points x and z cannot exceed the sum of the remaining two distances.

One can use the properties to define distance function in a non-empty set x as a function $X \times X\ to\ \mathbb{R}$, the set of all real numbers. In mathematics, we have seen that the idea of limits involves the idea of closeness. One can say that the limits of a function l when the function takes the valued close to l in an appropriate set. Similarly, the ideas of continuity, differentiation and integration all require the notation of closeness. This concept is basic in calculus. The best way to define closeness is in terms of distance by

taking points to be close if the distance between them is small. Distance is originally a concept in geometry, but this can be made more general if we concentrate on the following three essential properties of distance.

- $(M\,1)$ the distance between two points x and y is a positive real number unless x and y are identical in which case the distance is zero.
- $(M\,2)$ the distance from x to y is the same as the distance from y to x.
- $(M\,3)$ if x, y, z are three points, then the distance between the two points x and z cannot exceed the sum of the remaining two distances.

One can use the properties to define distance function in a non-empty set x as a function $X \times X\ to\ \mathbb{R}$, the set of all real numbers.

2.1 Definition of Metric

Let X be a non-empty set. A function $d : X \times X \to \mathbb{R}$ is called a metric distance in X if for all $x, y, z \epsilon X$,

- $(M\,1)$ Non-Negative Condition $d(x, y) \geq 0$; $d(x, y) = 0\ iff\ x = y$
- $(M\,2)$ Symmetry $d(x, y) = d(y, x)$
- $(M\,3)$ Triangle Inequality $d(x, z) \leq d(x, y) + d(y, z)$

The pair (X, d) is called a metric space.

3 Metric on Linear Space

In the preceding section, the considered examples in which the metric was defined by making use of the properties of the range of the function. In each of the problems, the range was $\mathbb{R}$, and properties of $\mathbb{R}$ were used to define the metrics. When the sets under consideration are vector space, it is natural to metrize the sets to define metrics on the sets by using the structure of the vector space [11]. In particular, if in the space, each vector has a magnitude satisfying some 'nice' properties the space can be given a metric structure by employing this norm, such a space is then called a normal linear space.

3.1 Normal Linear Space

Definition: Let V be a vector space $\mathbb{R}$. A norm on V is a function $\|\| : V \to \mathbb{R}$ which satisfies the properties:

- $(N\,1)\|x\| \geq \forall x \epsilon V\ and\ \|x\| = 0\ iff\ x = 0$
- $(N\,2)\|x + y\| \leq \|x\| + \|y\|$
- $(N\,3)\|c.x\| = \|c\|.\|x\|\ where\ c\ is\ constant$

If the linear space v has a norm defined on it, we can use this function to define a metric d *and* v as follows:

$$d(x, y) = \|x - y\| \textit{ for all } x, y \epsilon V$$

To show that this is a matric on v, we proceed as follows:

1. $d(x, y) \geq 0 \textit{ for } \|x - y\| \geq 0 \textit{ by } (N1)$
   $d(x, y) = 0 \textit{ iff } \|x - y\| = 0 \textit{ iff } x = y \textit{ by } (N1)$
2. $d(y, x) = \|y - x\| = \| - (x - y)\| = |-1|\|x - y\| = \|x - y\| = d(x, y)$
3. $d(x, y) = \|x - z\| = \|x - y + y - z\| \leq \|x - y\| + \|y - z\| \textit{ by } (N2)$
   $\textit{therefore } d(x, z) \leq d(x, y) + d(y, z)$

This establishment that d is a metric. A vector space with this metric is called a normal linear space. Norm function $\|x\|$ on $\mathbb{R}^2$ can be defined as:

$$\|x\| = \sqrt{x_1^2 + x_2^2} \forall x \epsilon \mathbb{R}^2$$

This norm is called Euclidean norm or usual norm on $\mathbb{R}^2$ and the set $\mathbb{R}^2$ with the norm $\| \|$ is a normal vector space. Next defining the function:

$$d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R} \textit{ by } d(x, y) = \|x - y\|,$$

one can easily show that d is a metric on $\mathbb{R}^2$.

3.2 Metric Structure Concept

In this section, certain tools are presented that will help to construct SVM.

Open Ball and Open Set in Metric Space

Let (x, d) be a metric space and $p \in x$. We define an open ball at p with radius r as the set of all points of x whose distance from p is less than r, is denoted by $B(p, r)$. Thus:
$B(p, r) = \{x \epsilon X : d(x, p) < r\}$ *where* p is called the center and $r > 0$.

Open Ball in $\mathbb{R}^2$

Let $p \epsilon \mathbb{R}^2$ and r, be any positive real number. The set:

$$B(p, r) = \left\{x \epsilon \mathbb{R}^2 : d(x, p) < r\right\} \text{ where } d(x, p) = \|x - p\|,$$

is an open ball with center at p and radius at r.

(i) Here, we can take

$$\|x - p\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

In $\mathbb{R}$, the open ball

$$B(p, r) = \{x \epsilon \mathbb{R} : |x - p| < r\}$$
$$= (p - r, p + r)$$

This shows that every open ball in $\mathbb{R}$ is bounded open interval.

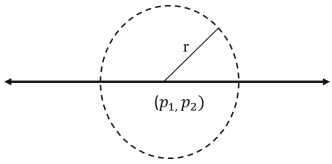


Fig. 1. Every open ball in $\mathbb{R}$ is bounded open interval

In $\mathbb{R}^2$, open ball

$$B(p, r) = \left\{x \epsilon \mathbb{R}^2 : \|x - p\| < r\right\}$$

$$= \left\{x \epsilon \mathbb{R}^2 : \sqrt{(x_1 - p_1)^2 + (x_2 - p_2)^2} < r\right\}$$

where $x = (x_1, x_2)$ *and* $p = (p_1, p_2) \epsilon \mathbb{R}^2$. This solves that an open ball is $\mathbb{R}^2$ is a set of all points inside the circle having centre (p_1, p_2) and radius r as shown in the Fig. 1.

(ii) we can also take: $\|x - p\| = |x_1 - p_1| + |x_2 - p_2|$

Let us take $p = (0, 0) \epsilon \mathbb{R}^2 : d((x_1, x_2), (0, 0)) < 1\}$

$$= \left\{(x_1, x_2) \epsilon \mathbb{R}^2 : |x - 0| + |x_2 - 0| < 1\right\}$$

$$= \left\{(x_1, x_2) \epsilon \mathbb{R}^2 : |x_1| + |x_2| < 1\right\}$$

$$= \{(x_1, x_2) : x_1 + x_2 < 1\} n \{(x_1, x_2) : -x_1 + x_2 < 1\} n \{(x_1, x_2) : x_1 - x_2 < 1\}$$
$$n\{(x_1, x_2) : -x_1 - x_2 < 1\}$$

Which is shown in the following Fig. 2:

Thus, open ball in $\mathbb{R}^2$, with given metric is 0 set of all points inside the oblique square having a given point on the plane as a centre and given positive real number as radius.

(iii) If $d(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$ then an open ball in $\mathbb{R}^2$ is the set of all points inside the square having a given point on the plane as a centre and given positive real number as radius.

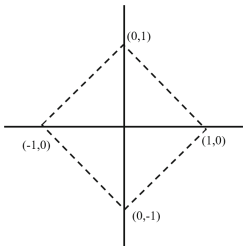


Fig. 2. Open ball in $\mathbb{R}^2$

Open Set in a Metric Space

Definition Let (X, d) be a metric space and $G \subset X$ each $\in G$, there is an open ball $B(p, r)$ such that $B(p, r) \subset G$. Note that every open ball is an open set in (x, d). Also, whole space $\mathbb{R}^n$ is an open set because every open ball of each point in $\mathbb{R}^n$ is a sub set of $\mathbb{R}^n$ i.e. for each $p \in \mathbb{R}^n$ for each $r > 0$. $B(p, r) \in \mathbb{R}^n$, according to the Haurdoff's property, there exists open balls $B(p, r_1)$ *and* $B(q, r_2)$ such that $B(p, r_1)\, n\, B(q, r_2) \neq \varnothing$ provided $p, q \in \mathbb{R}^n$ *and* $p \neq q$. It can be easily proved that union may infinitely open sets: $\{G_\alpha \subset \mathbb{R}^n, \alpha \in I\}$ is an open set in $\mathbb{R}^n$. A set $F \subset X$ is said to be closed in (x, d) iff its complement $F' = X - F$ is an open set.

4 Different Types of Points in Metric Structure

Definition (i) Let (x, d) be a metric space and $A \subset X$. A point $p \in A$ is an interior point of A if there exists $r > 0$ such that $B(p, r) \subset A$. The set of interior points of A is denoted by A^0 (Fig. 3).

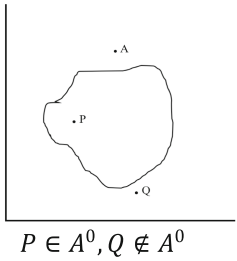


Fig. 3. Types of points in metric structure

Note that if $(\mathbb{R}^n, d)$ is a metric space, and $A \subset \mathbb{R}^n$, a point $p \in A$ is said to be an interior point of A if there exists an open set containing p in the definition. In $\mathbb{R}^2$, all points inside a circle are interior points. A^0 is infact an open set in X (Fig. 4).

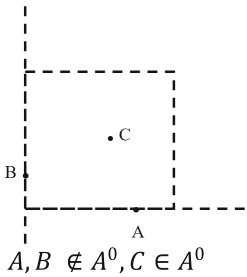


Fig. 4. Types of points in metric structure

Definition (ii) A point $p \in X$ is called exterior point of A of p is an interior point of $X - A$ i.e. A^C. In other words $p \in X$ is an exterior point if there exists $r > 0$ *such that* $B(p, r) \subset A^C$. A set of all exterior points of A is called exterior of A and is denoted by *ext* A. For example, Q is an exterior of A.

Definition (iii) A point $p \in X$ is called boundary point of A if any open ball centered at p has non-empty intersection with both A and its complement.

In other words, p is a boundary point of A if it is neither an interior point of A nor exterior point of A (i.e. A^C) i.e.:

$$\textit{for every } r > 0, B(p, r) \cap A \neq \varnothing, B(p, r) \cap A^C \neq \varnothing.$$

Set of all boundary points of A is denoted as $\delta(A)$.
For example, in $(\mathbb{R}^2, d)$:

(i) Every point $x \in B(p, r)$ is an interior point of $B(p, r)$.
Hence $[B(p, r)]^0 = B(p, r)$.
(ii) Every point $x \in \mathbb{R}^2$ satisfying $d(p, x) > r$, *where* $p \in \mathbb{R}^2$ is an exterior point of A.
(iii) Every point$x \in \mathbb{R}^2$ satisfying $d(p, r) = r$ *where* $p \in \mathbb{R}^2$ is boundary point of $A \subset (\mathbb{R}^2)$.

4.1 Connectedness

The line $\mathbb{R}$ is connected, but if we remove the point O from this set, it falls apart into two disconnected sets $(-\infty, 0)$ and $(0, -\infty)$. Observe that both the sets $(-\infty, 0)$ and $(0, -\infty)$ are open sets.

Definition A metric space (X, d) is said to be connected iff it cannot be expressed as a disjoint union of two sets which are both open. If $X = A \cup B, A \cap B = \varnothing$ and both A and B are open then X is not connected i.e. X has a separation. Let (X, d) be a metric space. A and B are two separated sets of X i.e.

$$X = A \cup B, A \cap B = \varnothing$$

where A, B are open. Constructing a hyperplane H that will separate A and B as shown in the figure. Let $P = A$ or B and L be the sets of all points on H. The perpendicular distance from a boundary point $b_p \in p$ *and* L os denoted by $d(b_p, L)$ and is defined by:

$$d(b_p, L) = \inf\{d(b_p, x) : x \in L\}$$

A and B has maximum width of separation if b_p satisfies the above condition.

For our case study, we consider the open sets LD1 and LD2 as shown in the Fig. 5. The points which are inside the LD1 are interior points where as the points which are not inside LD1 are exterior points. Points which are on boundary of LD1 are boundary points.

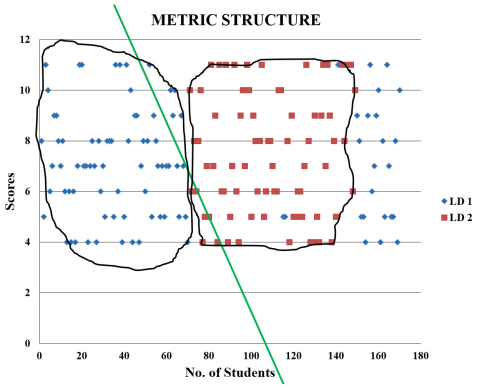


Fig. 5. Metric structure for the data set of LD1 versus LD2

5 The Two-Variable Linear Model

Simple Linear model can be expressed in terms of two-variables. Continuous variables x and y are taken on X and Y-axis. Each dot on the plot represents the x and y scores for an individual. The pattern clearly shows a positive correlation.

Straight line can be drawn through the data points so that complete dataset can be divided approximately into two classes. When we fit a line to data, we are using what we call a **linear model**. A straight line is often referred to as a **regression line** and the analysis that produces it is often called **regression analysis**. With y on as y-axis and x as x-axis; $y = a_0 + a_1 x$ is the regression line in where a_0 is the slope of the line and a_1 is the y-intercept.

6 Case Study: Learning Disability

6.1 Data Collection

With the help of LD centres and their doctors, a checklist containing the 79 most frequent signs and symptoms of LD is created for LD assessment. This checklist is then used for further studies there and on subsequent evaluation with the help of these professionals and from the experience gained, another checklist, reducing to 15 prominent attributes were evolved which is used in the present research work. This has led to the collection of a data set of 841.

6.2 Implementation and Results

Implementation

The system is implemented using Java. The experiments were conducted on a workstation with an Intel Core i3 CPU, 4 GHz, 2 GB of RAM, running on Microsoft Windows 10 Home Edition. A detailed study on the uses of different classification algorithms, viz. single layer perceptron, winnow algorithm, back propagation algorithm, LVQ, Naïve Bayes' classifier and J-48 classifier are used for the prediction of LD in children. The main drawback found in all these classification algorithms is that, there is no proper solution for handling the inconsistent or unwanted data in the data base and also the classifier accuracy is low. Hence, the classification accuracy has to be increased by adopting new methods of implementation by proper data pre-processing. Studies, as part of this research work, are conducted to achieve these goals. Linear regression being a simple method for numeric prediction, and it has been widely used in statistical applications for since several years. Generally, the low performance in linear models is due to their rigid structure which implies to linearity [1]. If the data shows a nonlinear dependency, the best-fitting straight line will be found, where "best" is interpreted as the least mean-squared difference. This line may not fit very well. However, linear models serve well as building blocks for more complex learning methods. The summary of outcomes of all the linear classifiers applied for classification of the data as LD or NLD is given below.

Table 1. Summary of output of all linear classifiers applied on the LDvsNLD database.

Method	Accuracy (%)	Correctness (%)	Coverage (%)
Single layer perceptron algorithm	93	92	92
Winnow algorithm	97	96	96
Learning vector quantization algorithm	96.1	95	95
Back propagation algorithm	86.54	86	86
J-48 classifier	88.8	87	87
Naïve Baye's classifier	94.23	95	95

The further work of this study will be to focus on the identification of the sub types of LDs and their overlaps which a general linear model fails to achieve.

7 Conclusion

In this research work, the prediction of LD in school age children is implemented through various algorithms. The main problem considered, in the work for analysis and solving, is the design of an LD prediction tool based on machine learning techniques. A detailed study on the uses of different classification algorithms show in Table 1 are used for the prediction of LD in children. The main drawback found in all these classification algorithms is that, there is no proper solution for handling the inconsistent or unwanted data in the data base and also the classifier accuracy is low. Hence, the classification accuracy has to be increased by adopting new methods of implementation by proper data preprocessing. Following is the derivation of models which will be implemented in future work to identify the type of LD and hence creating non-binary output.

7.1 The Linear Regression Model

The above equation may not divide the dataset into two classes exactly as well as there may be some points which may be closed to the line so we require one more component say ε. Thus, the equation can be written as:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

where y is the dependent or response variable and x is the independent or predictor variable. The random variable ε is the error term in the model [13]. Error is not a mistake but is a random fluctuation and describes the vertical distance from the straight line to each point [12]. Constants β_0 and β_1 are determined using observed values of x and y and make inferences such as confidence intervals and tests of hypotheses for β_0 and β_1. We may also use the estimated model to forecast or predict the value of y for a particular value of x, in which case a measure of predictive accuracy may also be of interest.

Simple linear regression model for n observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i . i = 1, 2, \ldots, n \tag{2}$$

In this case, there is only one x to predict the response y, and is linear in β_0 and β_1. Following assumptions are made.

(a) $E(\varepsilon_i) = 0$ gives $E(y_i) = \beta_0 + \beta_1 x_i$, for i = 1,2,…n
(b) $var(\varepsilon_i) = \sigma^2$ gives $var(y_i) = \sigma^2$, for i = 1,2,…n
(c) $cov(\varepsilon_i, \varepsilon_j) = 0$, gives $cov(y_i, y_j) = 0$, for i = 1,2,…n and $i \neq j$

The response y is often influenced by more than one predictor variable. A linear model relating the response y to several predictors has the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k + \varepsilon \tag{3}$$

The parameters $\beta_0, \beta_1 \ldots, \beta_k$ are called regression coefficients. ε provides random variation in y. With the help of above discussion, we are in position to discuss various methods of regression analysis in classifying data.

7.2 Numeric Prediction: Linear Regression

When all the attributes and the given class contain numeric model, linear regression is applied in order to express the class as a linear combination of the attributes with predetermined weights: x = w0 + w1 a1 + w1 a1 + … + wk ak, where x is the class; a_1, a_2,…, a_k are the attribute values; and w_0, w_1,…, w_k are weights and are calculated from the training data.

For a given database, first instance will have a class, say $x^{(1)}$, and attribute values $a_1{}^{(1)}$, $a_2{}^{(1)}$,…, $a_k{}^{(1)}$ where the superscript denotes that it is the first example. It is also convenient to assume an attribute, a_0 whose value is always 1. The predicted value for the first instance's class can be written as: $w_0a_0{}^{(1)} + w_1a_1{}^{(1)} + w_2a_2{}^{(1)} + \ldots + w_ka_k{}^{(1)} = \sum_{j=0}^{k} w_ja_j^{(1)}$.

This value is the predicted value for the first instance's class. The value to be considered for this study is difference between the predicted and the actual values. Through the method of linear regression, coefficients w_j could be chosen where there are $k + 1$ of them. Aim is to minimize the sum of the squares of these differences over all the training instances. Suppose there are n training instances; denote the ith one with a superscript (i). Then the sum of the squares of the differences is:

$$\sum_{i=1}^{n}(x^{(i)} - \sum_{j=0}^{k} w_ja_j^{(1)})^2$$

where the expression inside the parentheses is the difference between the ith instance's actual class and its predicted class. This sum of squares is what has to be minimized by choosing the coefficients appropriately.

7.3 Linear Classification: Logistic Regression

Logistic regression builds a linear model based on a transformed target variable. Suppose first that there are only two classes. Logistic regression replaces the original target variable:

Pr[1] $a_1, a_2,\ldots,a_k$], which cannot be approximated accurately using a linear function, with:

$$\log (\Pr[1]\, a_1,\ a_2, \ldots, a_k])\,(1 -\ \Pr[1]\, a_1,\ a_2, \ldots, a_k]).$$

The resulting values are no longer constrained to the interval from 0 to 1 but can lie anywhere between negative infinity and positive infinity. The transformed variable is approximated using a linear function just like the ones generated by linear regression. The resulting model is:

$$\Pr[1]\, a_1,\ a_2, \ldots, a_k] \ =\ 1/1 \ +\ \exp(-w_0 - w_1a_1 - \ldots - w_ka_k),$$

Just as in linear regression, weights must be found that fit the training data well. Linear regression measures the goodness of fit using the squared error. In logistic regression the *log-likelihood* of the model is used instead. This is given by:

$$\sum_{j=0}^{k}\left(1 - x^{(i)}\right)(1 - \Pr a_1^{(i)}, a_2^{(i)}, \ldots, a_k^{(i)}]) + x^{(i)}\log(pr[1]a_1^{(i)}, a_2^{(i)}, \ldots, a_k^{(i)})$$

where the $x^{(i)}$ are either zero or one. The weights wi need to be chosen to maximize the log-likelihood. There are several methods for solving this maximization problem. A simple one is to iteratively solve a sequence of weighted least-squares regression problems until the log-likelihood converges to a maximum, which usually happens in a few iterations.

To generalize logistic regression to several classes, one possibility is to proceed in the way described previously for multi-response linear regression by performing logistic regression independently for each class. Unfortunately, the resulting probability estimates will not sum to one. In order to obtain proper probabilities, it is essential to combine (couple) the individual models for each class. This produces a joint optimization problem. A simple way is address multiclass problems also known as pair-wise classification. Here, a classifier is built for every pair of classes using only the instances from these two classes. The output on an unknown test example which is based on the class which receives maximum votes. This method generally produces accurate results in terms of classification error. It can also be used to produce probability estimates by applying a method called pair-wise coupling, which calibrates the individual probability estimates from the different classifiers. The use of linear functions for classification can easily be visualized in instance space. The decision boundary for two-class logistic regression lies where the prediction probability is 0.5, that is:

$$\Pr[1][a_1,\ a_2, \ldots, a_k] = 1/1 + \exp(-w_0 - w_1 a_1 - \ldots - w_k a_k) = 0.5$$

this occurs when $-w_0 - w_1 a_1 - \ldots - w_k a_k = 0$. Because this is a linear equality in the attribute values, the boundary is a linear plane, or *hyperplane,* in instance space. It is easy to visualize sets of points that cannot be separated by a single hyperplane, and these cannot be discriminated correctly by logistic regression. Multi-response linear regression suffers from the same problem [1]. Each class receives a weight vector calculated from the training data.

Suppose the weight vector for class 1 is: $w_0^{(1)} + w_1^{(1)} a_1 + w_2^{(1)} a_2 + \ldots + w_k^{(1)} a_k$ and the same for class 2 with appropriate superscripts. Then, an instance will be assigned to class 1 rather than class 2 if:

$$w_0^{(1)} + w_1^{(1)} a_1 + w_2^{(1)} a_2 + \ldots + w_k^{(1)} a_k > w_0^{(2)} + w_1^{(2)} a_1 + w_2^{(2)} a_2 + \ldots + w_k^{(2)} a_k;$$

which is, it will be assigned to class 1 if:

$$\left(w_0^{(1)} - w_0^{(2)}\right) + \left(w_1^{(1)} - w_1^{(2)}\right) a_1 + \ldots + \left(w_k^{(1)} - w_k^{(2)}\right) a_k > 0$$

This is a linear inequality in the attribute values, so the boundary between each pair of classes is a hyperplane. The same holds true when performing pair-wise classification. The only difference is that the boundary between two classes is governed by the training instances in those classes and is not influenced by the other classes.

References

Frank, E., Hall, M.A., Witten, I.H.: The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", 4th edn. Morgan Kaufmann (2016)

Edwards, J.: Differential Calculus for Beginners (2016). ISBN: 9789350942468, 9350942461

Jain, K., Manghirmalani, P., Dongardive, J., Abraham, S.: Computational diagnosis of learning disability. Int. J. Recent Trends Eng. **2**(3), 64 (2009)

Jain, K., Mishra, P.M., Kulkarni, S.: A neuro-fuzzy approach to diagnose and classify learning disability. In: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), 28–30 Dec 2012. Advances in Intelligent Systems and Computing, vol. 236. Springer (2014)

Manghirmalani, P., Panthaky, Z., Jain, K.: Learning disability diagnosis and classification—a soft computing approach. In: World Congress on Information and Communication Technologies, pp. 479–484 (2011). https://doi.org/10.1109/WICT.2011.6141292

Manghirmalani, P., More, D., Jain, K.: A fuzzy approach to classify learning disability. Int. J. Adv. Res. Artif. Intell. **1**(2), 1–7 (2012)

Mishra, P.M., Kulkarni, S.: Classification of data using semi-supervised learning (a learning disability case study). Int. J. Comput. Eng. Technol. (IJCET) **4**(4), 432–440 (2013)

Manghirmalani Mishra, P., Kulkarni, S.: Developing prognosis tools to identify LD in children using machine learning techniques. In: National Conference on Spectrum of Research Perspectives (2014). ISBN: 978-93-83292-69-1

Manghirmalani Mishra, P., Kulkarni, S., Magre, S.: A computational based study for diagnosing LD amongst primary students. In: National Conference on Revisiting Teacher Education (2015). ISBN: 97-81-922534

Mishra, P.M., Kulkarni, S.: Attribute reduction to enhance classifier's performance-a LD case study. J. Appl. Res. 767–770 (2017)

Rolewicz, S.: Metric Linear Spaces. Monografie Mat. 56. PWN–Polish Sci. Publ., Warszawa (1972)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B **58**(1), 267–288 (1996)

Yan, X., Su, X.G.: Linear Regression Analysis: Theory and Computing (2009). ISBN: 13:978-981-283-410-2



Literature Review on Recommender Systems: Techniques, Trends and Challenges

Fethi Fkih[1,2(✉)] and Delel Rhouma[1,2]

[1] Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
f.fki@qu.edu.sa

[2] MARS Research Laboratory LR17ES05, University of Sousse, Sousse, Tunisia

Abstract. Nowadays, Recommender Systems (RSs) have become a necessity especially with the rapid increasing of the numerical data volume. In fact, internet's users need an automatic system that help them to filter the huge flow of information provided bay websites or even by research engines. Therefore, a Recommender System can be seen as an Information Retrieval system that can respond to an implicit user's query. The RS draw this implicit user's query based on a user's profile that can be created using some semantic or statistic knowledge. In this paper, we provide an in-depth literature review on main RS approaches. Basically, RS's techniques can be divide into three classes: collaborative Filtering-based, Content-based and hybrid approaches. Also, we show the challenges and the potential trends in this domain.

Keywords: Recommender System · Collaborative Filtering · Content-based Filtering · Hybrid Filtering · Sparsity

1 Introduction

Due to the massive expansion of the internet, the market size of e-commerce also expanded with hundreds of millions of items that need to be handled [1, 2]. This huge amount of items caused difficulty for the users to find the items suited to their preferences. In fact, an items' huge number will consume resources (time and materials). Thus, there is an urgent demand to help users to save their time in the search process and to find the items they are interested in. To reach this purpose, Recommender Systems (RS) have been emerged in the recent years especially in the e-commerce field.

Recommender Systems (RSs) are filtering techniques used to automatically provide suggestions for interesting or useful items to the users according to their personal preferences [3]. The recommendation system usually dealing with a large amount of data to give individualized recommendations useful to the users. Most of e-commerce sites such as Amazon.com and Netflix now use the recommendation systems as an integral part of their sites to effectively provide suggestions that directed toward the items that best meet the user's needs and preferences. In order to implement its core function, recommendation systems try to predict the most pertinent items for customers by accumulating

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 493–500, 2023.
https://doi.org/10.1007/978-3-031-27409-1_44

user preferences, which are either directly given as product ratings or are inferred by analyzing user behavior. Customers receive rated product lists as recommendations.

Although the significant improvement in the current Recommender Systems performances, they still suffer from many problems that limit their effectiveness, these problems are mainly related to the data proprieties used for building such systems. In fact, low-quality data will necessarily lead to a low-performance system.

In this paper, we provide an in-depth literature review on Recommender system main approaches. The paper is organized as follows: in Sect. 2 we introduce the popular approaches used in the recommendation field. In Sect. 3, we supply a discussion that shows the advantages and disadvantages of each approach. In Sect. 4, we provide some prospective topics for future study.

2 Recommender System Main Approaches

In order to implement its core function, recommendation systems intend to forecast the most appropriate items for users by identifying their tastes which can be expressed explicitly (as an item ratings) or implicitly (as an interpretation of the user behaviors). RS generally uses a range of filtering techniques to generate this list of recommendations, RSs developed based on a simple idea: to make decisions, people habitually trust recommendations of their friends or those sharing a similar taste. For example, consumers look to movie or book reviews when deciding what to watch or read [4]. In order to imitate this behavior, Collaborative Filtering-based RSs used algorithms to benefit from the recommendations produced by a community of users to provide some relevant recommendations to an active user. In fact, RS generally uses a range of filtering techniques to generate a list of recommendations, collaborative filtering (CF) and content-based filtering are widely implemented and used techniques adopted in the domain of RSs.

2.1 Collaborative-Filtering Based Approaches

Collaborative Filtering (CF) is a simple and widely used technique in the domain of RS, the rationale of this approach is to recommend to the user the items that other users with similar tastes liked. There is no need for any previous knowledge about users or items, the recommendations make based on interactions between them instead [5]. The idea of CF-based CF is finding users in a community sharing the same taste. If two users share similar rated items, then they can also share the same tastes [6]. Ratings scales can take different ranges depending on the dataset. Commonly, the rating scale can be ranged from 1 to 5: 1 means good and 5 means bad, or binary rating choices between agree/disagree or good/bad. To demonstrate this idea, let's use the example of a basic book recommendation tool that aids users in choosing a book to read. Using a scale of 1 to 5, where 1 is "bad" and 5 is "good," a user will rate books. The system then suggests further books that the user might like based on community ratings [7].

Normally CF-based approaches utilize the ratings of users for items. Therefore, the rating R consists of the association of two things: user u and item i. Consequently, the core task of a CF-based RS is to predict the value of the function $R(u, i)$ [1]. One way to visualize and handle ratings is as a matrix [8]. This matrix consists of rows represent

users, columns represent items (books, movies, etc.) and the intersection of a row and a column represents the user's rating. This matrix is processed in order to generate the recommendations. In fact, CF approaches can be classified into two main categories: memory-based and model-based techniques. This classification is according to how the rating matrix is processed.

2.1.1 Memory-Based Algorithms

Memory-based algorithms are the most well-known collaborative filtering algorithms. The items that were rated by a user are used to search for a neighbor that shares the same appreciation with him. When a neighbor of a user is found the preferences of neighbors are compared to generate recommendations [9].

User-Based The purpose of this technique is to predict the rating R(u,i) of a user u on an item i using the ratings given to item i by other users most similar to u, named nearest-neighbors that have similar rating patterns. The users perform the main role in the user-based method. In this context, GroupLens [10] is the earliest used user-based collaborative method.

Item-Based This technique tries to predict the rating of a user u on an item i using the ratings of u for items similar to item i. Using the same idea, the Item-based approach uses the similarity between items by looking into the set of items rated by the user and calculates the similarity degree on the item i. By taking a weighted average of the target user's ratings on these similar items [11].

2.1.2 Model-Based Algorithms

In contrast to Memory-based algorithms, which use the stored ratings directly in the prediction process, Model-based algorithms utilize prior ratings for the model training in order to enhance the performance of the CF-base Recommender System [9]. The general idea of the model-based is to build a model with the help of dataset ratings. After being trained with the existing data, the model is then used to forecast user ratings for new items. Examples of Model-based approaches include Matrix Completion Technique, Latent Semantic methods [9], and data mining techniques such as clustering and association rules. Algorithms for data mining and machine learning can be used to forecast user ratings for items or to figure out how to rank items for a user. The mostly used machine learning algorithms in model-based recommender systems.

Clustering Clustering algorithm tries to assign items to groups in which the items in the same groups are similar in order to discover meaningful groups that exist in the data. K-means clustering algorithm is the simplest and most commonly used algorithm. K-means partitions set of N items into k disjoint clusters that contain Nj of similar items. Zhu [12] presented a book recommendation system uses the K-means clustering method that classifies the users into groups then recommends books according to the user's group. Clustering is sure to improving efficiency by using a dimensionality reduction technique.

Association Rule Association rules focus on finding rules that will predict relationships between items based on the patterns of occurrences of other items in a transaction [13]. This technique helps to understand customers' buying habits and discover groups of items that are usually purchased together. The authors in [14] used an association rule mining approach for building an efficient recommendation algorithm. Their strategy adapted association rules to reduce the impact of attacks, where fake data is often purposefully inserted to affect the recommendations.

2.2 Content-Based Filtering

Content-based Recommender System recommends items based on a similarity measure that quantify the taste of other similar users with no previous knowledge about users or items. Content-based filtering, however, recommends Items using information about the item itself rather than on the preferences of other users. The RS attempts to recommend to the user items that are similar to previous liked items. The RS uses items' features to compute similarity values between items [1].

Mooney et al. [18] LIBRA is a content-based book recommendation system that utilizes information extraction and Naïve Bayes classifier. The system constructed through three steps: First, extracting relevant information from the unstructured content of a document. Therefore, learning user profile by using an inductive naive Bayesian text classifier to produce a ranked list of preferences. Next, producing recommendations by predicting the appropriate ranking of the remaining items based on the posterior probability of a positive classification.

2.3 Hybrid Filtering

Hybrid recommender systems combine many recommendation strategies together to enhance the performance of the RS or to deal with the cold-start problem. For example, collaborative filtering approaches suffer from new-item problems that they can't recommend new items that not rated yet. As the prediction of new items is based on semantic features that can be automatically extracted from the corresponding item, this problem can be overcome by content-based approaches [19].

The authors in [20] introduced a hybrid approach for designing a book recommendation system through combining Collaborative Filtering and Content-based techniques. The techniques have been combined using the mixed method where several recommendations provided by different techniques are merged. Content-based technique uses demographic features (age, gender) of the user profile as input to filter similar users in order to solve problems related to low-quality data, such that, cold start problem.

3 Discussion

Although Collaborative filtering approaches proved good efficiency in the Recommender Systems field, they suffer from many issues such as cold start, sparsity and scalability. Many of researchers proposed approaches to overcome these problems.

3.1 Cold-Start Problem

This problem occurs when a recommender is unable to make meaningful recommendations cause of the lack of information about a user or an item. For instance, if a new user doesn't rate any item yet then the recommender system is incapable to know his interests. Therefore, this problem can reduce the performance of the collaborative filtering [8]. Some researchers tried to overcome this issue by either getting users to rate items and choose their favorite categories at the start [16] or making recommendations using the user's demographic information, such as, his age, gender, etc. In this context, Kanetkar et al. [17] adopted a Demographic-based approach that makes demographic recommendations to give personalized recommendations by cluster users based on their demographic aspects. The authors in [15] introduce a technique that used semantic resources that can be integrated into the recommendation process which deal with the cold-start problem.

3.2 Sparsity Problem

CF fails to provide a relevant recommendation when data is very sparse. Generally, the items' number is very large than the users' number. In this case the majority of the user-item matrix elements take the value 0 (without rating). Many solutions have been proposed for overcoming data sparsity problem, such as, association rules. Burke [19] presented an improved Collaborative filtering technique for personalized recommendation. They proposed a framework of association rules mining and clustering technique that incorporates different types of correlations between items' or user's profiles. The fuzzy system is commonly used for fixing problems with collaborative filtering. In the same context, the authors in [20] proposed a hybrid fuzzy-based personalized recommender system, uses fuzzy techniques to deal with the sparsity problem and improve the prediction accuracy and to handle customer data uncertainty using linguistic variables which are used to describe customer preferences. The authors in [7] used the Fuzzy system to make up the sparsity problem with CF when the recommendation is unobtainable if a new item is added in user-item matrix. Also, the recommendation will be inaccessible if the related user community's information is insufficient.

3.3 Scalability

Scalability can be defined as the inability of the RS to provide recommendations in real-world datasets. Given the huge data flow on the internet, the number of users and items in the RS datasets is growing rapidly. In fact, large datasets enclose sparse data which avoids scalability of the Recommender System. In this context, the authors in [15] proposed a model that provides a scalable and efficient recommendation system by combined CF with association rule mining. The proposed model aims mainly to supply relevant recommendations to the user.

3.4 Approaches Evaluation

The main challenge in any application of Recommender System in the real world is to select the suitable approach that provide the best performance. However, the system's

performance is not the only criterion for choosing the appropriate RS. Table 1 show the advantages and the disadvantages of each approach that can help to select the best technique for a given application. As mentioned previously in this work, the data quality can visibly influence the RS performance. From our perspective, low quality data (as the majority of available dataset) which characterized by a high sparsity and low density should be handled before any further processing. The state of art proves that the RS domain needs new techniques for improving the data quality and solving issues related to cold start problem and data sparsity. Besides, content-based approach can provide solutions for the mentioned problems but it suffers from many challenges due to the big complexity of text mining and Natural Language Processing tools that can be involved to extract the missing information such as, gender, age [21–23], sentiments etc.

Table 1. Comparison between different RS approaches

Approach	Advantages	Disadvantages
Collaborative filtering (CF) techniques		
User-based	– Independent on the domain – Performance is improved over time – Serendipity – Absence of content analysis	– Data sparsity – Popular taste – Scalability – New-item problem – New-user problem – Cold-start problem
Item-based	– No content analysis – Domain independent – Performance is improved over time – Serendipity	– Data sparsity – Cold-start problem – New-item problem – Popular taste
Content-based (CB) techniques		
Attribute-based techniques	– No cold-start problem – No new item problem – User independence – Sensitive to changes of preferences – Provide transparency – Can explicitly listing content features – Can map from user needs to items	– Can't learn, no feedback – Only works with categories – Ontology modeling and maintenance is required – Over-specialization – Serendipity problem

4 Conclusion

In this paper, we provided an in-depth literature on the main approaches currently used in the recommendation field. In fact, we supplied an overview on the theoretical foundations

of the Recommender system. Also, we carried out a comparison between the different techniques and we highlighted for each one its advantages and disadvantages.

As a future work, we intend to integrate semantic resources such as, ontologies and thesaurus, into Recommender Systems. These resources can provide semantic knowledge can be extracted from textual content that can reduce the negative influence of the low-quality data on the Recommender System influence.

References

1. Fkih, F., Omri, M.N.: Hybridization of an index based on concept lattice with a terminology extraction model for semantic information retrieval guided by WordNet. In: Abraham, A., Haqiq, A., Alimi, A., Mezzour, G., Rokbani, N., Muda, A. (eds.) Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016). HIS 2016. Advances in Intelligent Systems and Computing, vol. 552. Springer, Cham (2017)
2. Fkih, F., Omri, M.N.: Information retrieval from unstructured web text document based on automatic learning of the threshold. Int. J. Inf. Retr. Res. (IJIRR) **2**(4) (2012)
3. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Recommender Systems Handbook, pp. 1–35. Springer, Boston, MA (2011)
4. Gandhi, S., Gandhi, M.: Hybrid recommendation system with collaborative filtering and association rule mining using big data. In: 2018 3rd International Conference for Convergence in Technology (I2CT). IEEE (2018)
5. Lee, S.-J., et al.: A movie rating prediction system of user propensity analysis based on collaborative filtering and fuzzy system. J. Korean Inst. Intell. Syst. **19**(2), 242–247 (2009)
6. Tian, Y., et al.: College library personalized recommendation system based on hybrid recommendation algorithm. Procedia CIRP **83**, 490–494 (2019)
7. Schafer, J.B., et al.: Collaborative filtering recommender systems. In: The Adaptive Web. Springer, Berlin, Heidelberg (2007)
8. Cacheda, F., et al.: Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender systems. ACM Trans. Web (TWEB) **5**(1), 2 (2011)
9. Fkih, F.: Similarity measures for collaborative filtering-based recommender systems: review and experimental comparison. J. King Saud Univ. - Comput. Inf. Sci. (2021)
10. Resnick, P., et al.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. ACM (1994)
11. Sarwar, B.M., et al.: Item-based collaborative filtering recommendation algorithms. Www **1**, 285–295 (2001)
12. Zhu, Y.: A book recommendation algorithm based on collaborative filtering. In: 2016 5th International Conference on Computer Science and Network Technology (ICCSNT). IEEE (2016)
13. Lin, W., Alvarez, S.A., Ruiz, C.: Collaborative recommendation via adaptive association rule mining. Data Min. Knowl. Disc. **6**, 83–105 (2000)
14. Sandvig, J.J., Mobasher, B., Burke, R.: Robustness of collaborative recommendation based on association rule mining. In: Proceedings of the 2007 ACM Conference on Recommender systems. ACM (2007)
15. Sieg, A., Mobasher, B., Burke, R.: Improving the effectiveness of collaborative recommendation with ontology-based user profiles. In: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems. ACM (2010)

16. Kurmashov, N., Latuta, K., Nussipbekov, A.: Online book recommendation system. In: 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO). IEEE (2015)
17. Kanetkar, S., et al.: Web-based personalized hybrid book recommendation system. In: 2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014). IEEE (2014)
18. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the Fifth ACM Conference on Digital Libraries. ACM (2000)
19. Burke, R.: Hybrid web recommender systems. In: The Adaptive Web, pp. 377–408. Springer, Berlin, Heidelberg (2007)
20. Chandak, M., Girase, S., Mukhopadhyay, D.: Introducing hybrid technique for optimization of book recommender system. Procedia Comput. Sci. **45**, 23–31 (2015)
21. Ouni, S., Fkih, F., Omri, M.N.: BERT- and CNN-based TOBEAT approach for unwelcome tweets detection. Soc. Netw. Anal. Min. **12**, 144 (2022)
22. Ouni, S., Fkih, F., Omri, M.N.: Novel semantic and statistic features-based author profiling approach. J. Ambient Intell. Hum. Comput. (2022)
23. Ouni, S., Fkih, F., Omri, M.N.: Bots and gender detection on Twitter using stylistic features. In: Bădică, C., Treur, J., Benslimane, D., Hnatkowska, B., Krótkiewicz, M. (eds.) Advances in Computational Collective Intelligence. ICCCI 2022. Communications in Computer and Information Science, vol. 1653. Springer, Cham (2022)



Detection of Heart Diseases Using CNN-LSTM

Hend Karoui(✉), Sihem Hamza, and Yassine Ben Ayed

Multimedia Information Systems and Advanced Computing Laboratory, MIRACL, University of Sfax, Sfax, Tunisia
Karouihend2020@gmail.com, sihemhz401@gmail.com, yassine.benayed@gmail.com

Abstract. The ElectroCardioGram (ECG) is one of the most used signals for the prediction of heart disease. Lots of research are based on the ECG signal for the detection of cardiac diseases. In this research, we propose a four phases method to the recognition of cardiac disease. The first phase is to remove noise and detect QRS complex using a pass-band filter. In the second phase, we segment the filtered signal. For third phase is to make a fusion of three types of characteristics such as Zero Crossing Rate (ZCR), entropy and cepstral coefficients. Those features extracted considered as input in the next step which is the combination between Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) proposed for multi-class classification of the ECG signal into five classes: Normal beat (N), Left bundle-branch block beat (L), Right bundle-branch block beat (R), premature Ventricular contraction (V), and Paced beat (P). The proposed model was evaluated using the MITBIH arrhythmia database and achieve an accuracy equal to 95.80%.

Keywords: Heart diseases · ECG signals · Features extraction · CNN · LSTM

1 Introduction

Cardiovascular diseases are now the leading cause of death in the world. According to World Health Organization (WHO), 17.7 million deaths are attributable to cardiovascular disease, which accounts for 31% of all deaths worldwide[1]. Lots of methods are used for the Recognition of cardiac diseases. Among these methods electrocardiography which allows the analysis of the heart's conduction system, in order to obtain information on the cardiac electrical activity. The recording of the conduction system is physically represented by an ElectroCardioGram (ECG) [1]. The importance of the ECG signal is due to the effect that the P, QRS, T waves constituting this signal [2] (shown in Fig. 1).

In recent years, the ECG signal have been used in several academic research for the detection of heart diseases using the deep learning or the machine learning techniques [3–6].

Hela and Raouf [3] proposed a clinical decision support system based on Artificial Neural Network (ANN) as a machine learning classifier and used time scale input

[1] https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 501–509, 2023.
https://doi.org/10.1007/978-3-031-27409-1_45

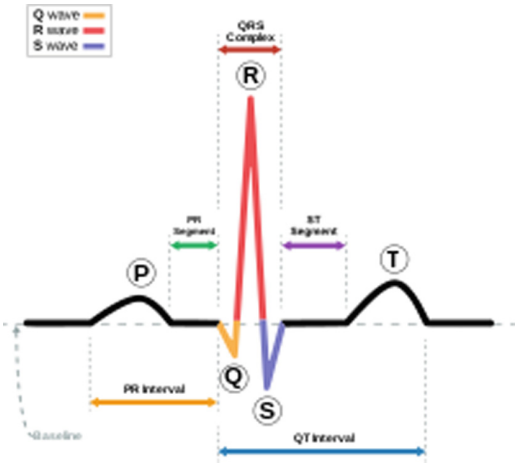


Fig. 1. QRS waves of ECG signal

features. The two types of extracted features are morphological and coefficients of Daubechies Wavelet Transform (DWT). For the proposed system is more accurate by using trainbr training algorithm which achieve an accuracy rate equal to 93.80%.

Savalia and Emamian [4] proposed a framework for the classification of 7 types of arrhythmias using CNN with 4 convolution layers and the Multilayer Perceptron (MLP) with four hidden layers. The high accuracy was obtained with the MLP (accuracy equal to 88.7%).

Rajkumar et al. [5] proposed a model CNN for the classification of ECG signal. The proposed model was tested with the MIT-BIH arrhythmia and achieve an accuracy equal to 93.6%.

Acharya et al. [6] presented a deep learning approach to automatically identify and classify the different types of ECG heartbeats. The model CNN was developed to automatically identify 5 different categories of heartbeats in ECG signals. This model was tested by using the MIT-BIH arrhythmia and achieve an accuracy of 94.03% and 93.47% with and without noise removal respectively.

This study proposes a method for detection heart diseases using the ECG signals. Our method presents four principal steps: preprocessing of ECG signal, segmentation of filtered signal. For step three we proposed to use three types of characteristics: ZCR, entropy and cepstral coefficients which gives a better result in [7]. Then we developed the CNN-LSTM model for the detection of heart diseases using MIT-BIH arrhythmia database.

The rest of this article is organized as follows. We introduce in the Sect. 2 the proposed approach. In Sect. 3, presents the experimental results and the comparative works. Finally, Sect. 4 summarizes this paper.

2 Proposed Work

In this paper, the proposed approach for the detection of heart diseases using ECG signal is presented in Fig. 2. This method consists of four steps: preprocessing of the ECG signal using band-pass filter to eliminate the noise and the application of pan-tompkins algorithm for better detection of QRS complex, the next step was the segmentation of the filtered signal, for the step of features extraction we proposed three types of characteristics ZCR, entropy and cepstral coefficients. The extracted features considered as input to Deep Neuronal Network (DNN) step.

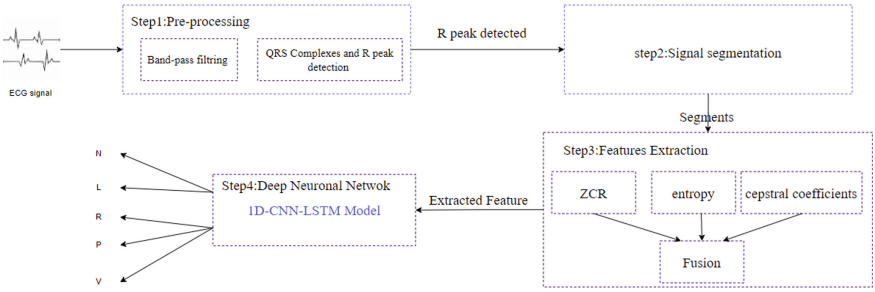


Fig. 2. Architecture of the proposed system

2.1 Pre-processing

The pre-processing of ECG signal is very important step to obtain a good classification. Many techniques have been developed for ECG signal pre-processing. In our study we used two pre-processing steps. First, we applied a filter pass-band to the ECG signal to eliminate the noise. The pan-tompkins was applied to detect the QRS complex and the R-peak. This algorithm was proposed by Jiapu Pan and Willis J. Tompkins in 1985 [8]. Figure 3 shows the R-peak detected.

2.2 Signal Segmentation

The mean idea of this technique is that each pre-processed ECG signal has several segments and for each segment there is an R-peak [9]. The used method in this study is R-centered it mean for each R peak we takes 0.5 s on the right of peak and 0.5 s on the left of peak to obtain a frame contain one peak. Figure 4 represent the result of segmentation obtained from filtered signal.

2.3 Features Extraction

Feature extraction is a fundamental step in the recognition process prior to classification.

In this article we proposed a combination of three types of characteristics ZCR, entropy, cepstral coefficients as are represented in Fig. 5.

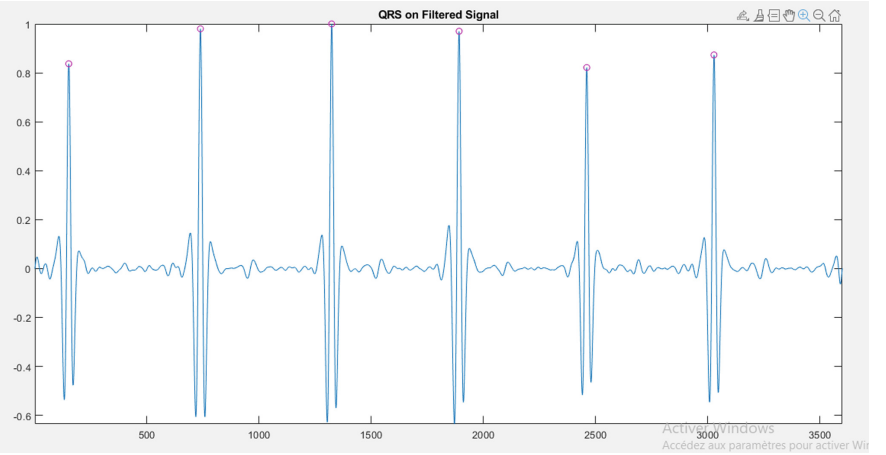


Fig. 3. R-peak detected

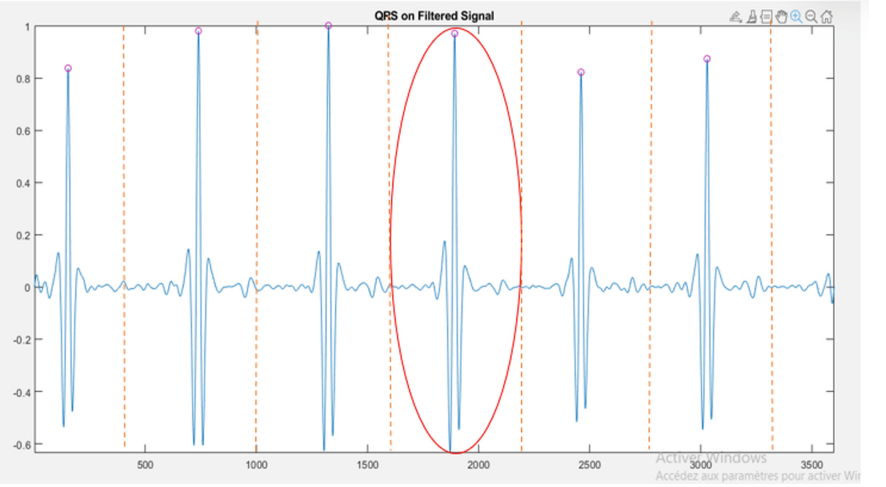


Fig. 4. Result of segmentation steps

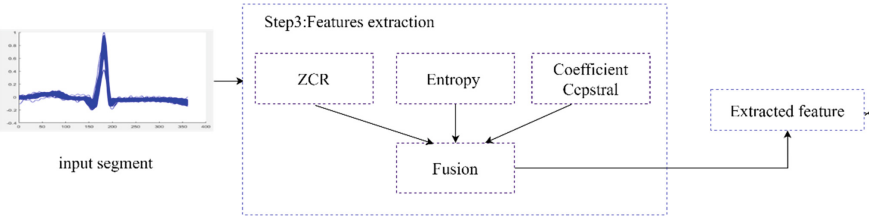


Fig. 5. Features extraction phases

- Zero Crossing Rate (ZCR)

  ZCR[2] is the measures how many times the waveform crosses the zero axis. It has been used in several fields noting the speech recognition. The ZCR defined according to the following equation:

$$ZCR = \frac{1}{N-1}\sum_{n=1}^{N-1} sign(s(n)s(n-1))$$

  With N represent the length of the signal s.

- Entropy

  Entropy is a measure of uncertainty and is used as a basis for techniques including feature selection and classification model fitting [9]. It defined according to the following equation:

$$H(x) = -\sum_{k} P(X_k) log_2[P(X_k)]$$

  With $x = \{X_k\}\ 0 \leq k \leq N-1$ and $P(X_k)$ is the probability of the X_k.

- Cepstral Coefficients

  Cepstral coefficients [9] are the most frequently used in the speech domain. The Fig. 6 shows the steps to follow to calculate cepstral coefficients.

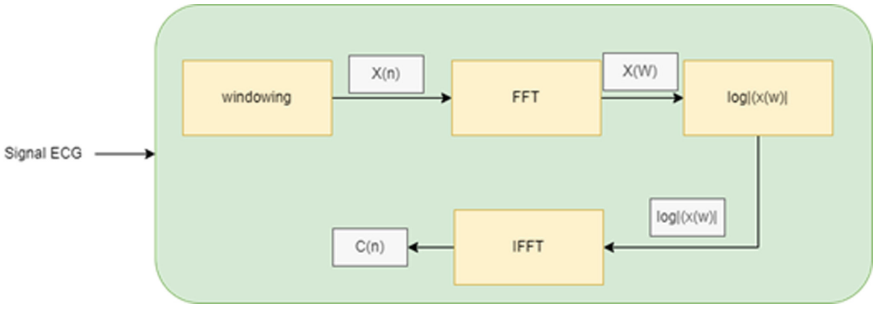


Fig. 6. The steps to follow to calculate cepstral coefficients

After the segmentation of filtered signal, we have calculated the following characteristics ZCR, entropy and cepstral coefficients based on the segmented signal, then we have combined the calculated values as represented in Fig. 5. The extracted features are taken as input to DNN in the next step.

2.4 Deep Neuronal Network (DNN)

In this study, we proposed 1D-CNN-LSTM model for the detection of heart disease based on the extracted features from the proposed ECG signal.

[2] https://www.sciencedirect.com/topics/engineering/zero-crossing-rate.

- Convolutional Neural Network (CNN)

  The convolutional neural network is a particular type of deep neural network developed for image classification. The CNN composed of convolutional, pooling and fully connected layers [10, 11].
- Long Short-Term Memory (LSTM)

  The model Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) used in deep learning. In addition, it is used for the classification of signals. His architecture consists of three gates: input gates, forget gates and output gates which includes the blocks of memory cells through which the signals flow [12, 13].
- CNN-LSTM

  In this study, we have proposed to use the combination of two models which are the CNN and the LSTM to automatically detect the heart disease from ECG signal with the combination of three characteristics (ZCR, entropy, cepstral coefficient).

  Figure 7 represent the first model CNN-LSTM.

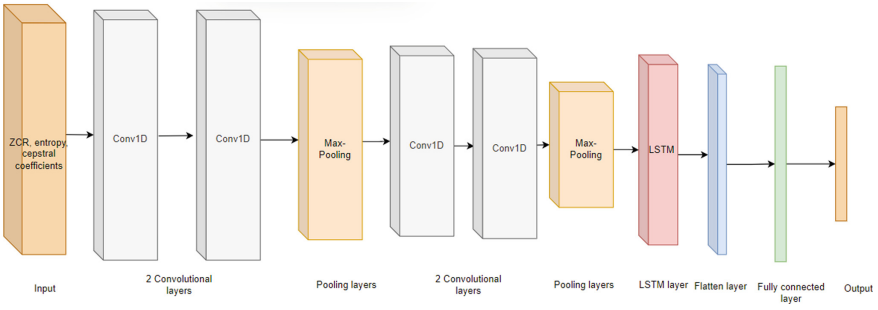


Fig. 7. Architecture of the first proposed model

Figure 8 represent the second model CNN-LSTM.

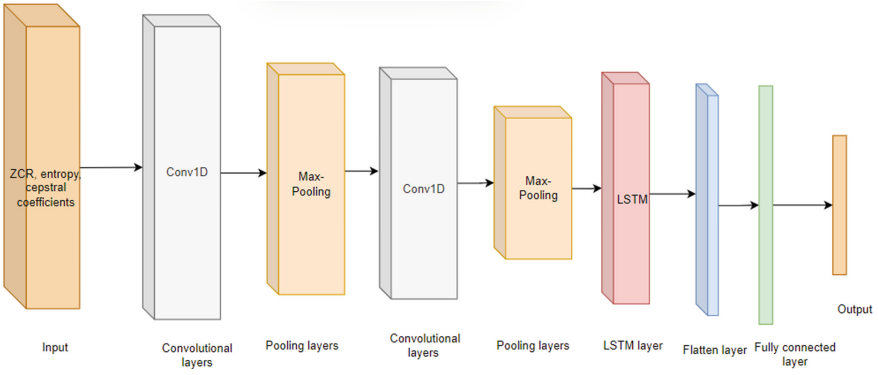


Fig. 8. Architecture of the second proposed model

The proposed model was evaluated in different experiments, they have the same input data which correspond to the vectors obtained from feature extraction, and they have a size equal to (1×1058). The first model consists of four convolution layers, two max-pooling layers, and one LSTM cell followed by flatten layer finally we found two fully connected layers. The two convolution layers are composed of (1×32) filters of size (1×5) each with padding same and we have used the Relu activation function. These two layers are followed by a max-pooling layer with size (1×5) and stride equal to 2. In the second convolution layers, we only changed the number of filters to 64 instead of 32 also we kept the same parameters and propriety (padding same, Relu activation function, max-pooling). The LSTM cell with 64 units is placed between the last max-pooling layer and the flatten layer. Concerning the classification layer, the first layer is fully connected and is composed of 64 neurons with a Relu activation function. The output layer has a vector of size 5 which is the number of classes considered in our system with an activation function softmax is used mainly for multi-class classification.

For the second model, we have reduced the number of layers while keeping the rest of the layers (max-pooling, LSTM, flatten) also we are keeping the same parameters such as the number of filters (1×32 for the first convolution layer, 64 for the second convolution layer, padding same, and the Relu activation function). This model consists of 2 layers of convolution each layer is followed by a max-pooling layer, a cell of LSTM, a layer of flatten, and two fully connected layers.

3 Experimental Results

3.1 Database

In this study, we have tested the suggested model on the MIT-BIH arrhythmia database available on the Physionet[3] website. It contains 48 half-hour excerpts from 47 subjects including 25 men aged 32–89 years and 22 women aged 23–89 years. For the 48 recordings, recordings 201 and 202 correspond to the same subject.

We use only 15 records from the database with a duration of 15 min for each of them. The MIT-BIH arrhythmia contains five classes (N, L, R, V and P).

In this research, we try to compare our result with the results achieved by the author Hela and Raouf [3]. Hence, we tested our model on 15 records from the MIT-BIH arrhythmia database.

In our paper, we proposed to develop two models (as explained in Sect. 2.4) for the detection of heart diseases using MIT-BIH arrhythmia database. The database was divided into training set and test set in the ratio of 80%, 20% respectively.

3.2 Results and Discussion

After pre-processing of ECG signal to detect complex QRS and R-peaks, the filtered signal was segmented by taking a duration of 1 min for each detected R-peak and then

[3] https://physionet.org/content/mitdb/1.0.0/.

for each segment we extracted the following characteristics: ZCR, entropy, cepstral coefficients.

The extracted features are considered as input for DNN phase which is the most interesting part of our work. We have tested the MIT-BIH arrhythmia with the first model which gives an accuracy equal to 93.67% then with the second model which gives an accuracy equal to 95.80% which is the best result obtained.

Indeed, we have achieved a 95.80% recognition rate using the second model (consists of 2 convolutional layers) on the MIT-BIH arrhythmia database (15 individuals). This recognition rate is higher than that obtained by Hela and Raouf [3]. Which use the same database MIT-BIH arrhythmia and that obtained a rate equal to 93.80%.

As mentioned in Sect. 1, the author Hela and Raouf [3] proposed a clinical decision support system based on Artificial Neural Network (ANN) as a machine learning classifier and uses time scale input features.

Based on the result obtained by the author Hela and Raouf [3], we have been used the fusion of three features (ZCR, entropy, cepstral coefficients) and we have been applied 1D-CNN-LSTM model (the second model) on the MIT-BIH arrhythmia and we obtained achieve subject accuracy of 95.80%.

The Table 1 represents some previous works which used the same database.

Table 1. Comparative with some previous studies used the MIT-BIH arrhythmia

Author	Classifier	Accuracy (%)
Hela and Raouf [3]	ANN	93.80
Savalia and Emamian [4]	MLP	88.7
Rajkumar et al. [5]	CNN	93.6
Acharya et al. [6]	CNN	94.03
Proposed approach	CNN-LSTM (first model)	93.67
	CNN-LSTM (second model)	95.80

In our research, we have proposed to use the ECG signal for the detection of heart diseases. Our proposed approach consists of four steps: pre-processing of the ECG signal using band-pass filter to eliminate the noise and the application of pan-tompkins algorithm for better detection of QRS complex, then we realized the segmentation which depends on the R peaks detection, for the step of features extraction we proposed three types of characteristics ZCR, entropy and cepstral coefficients. To evaluate our work, we have proposed to apply two models (CNN-LSTM) on the MIT-BIH arrhythmia database.

The best accuracy is achieved by the second model (consists of 2 convolutional layers) with an accuracy equal to 95.80%. This work was compared with another work used the same public database [3] which is obtained with the classifier an accuracy equal to 93.80%.

4 Conclusion

In this paper, we proposed a method to classify the ECG signal based on the use of a fusion of tree types of characteristics (ZCR, entropy, cepstral coefficients). The proposed model CNN-LSTM was evaluated in different experiments and tested with the MIT-BIH arrhythmia database which contains 5 classes. In the first experiment we achieve accuracy equal to 93.67%. For the second experiment we achieve a high accuracy 95.80% which is better than these obtained in the first experiment.

In the future, we will improve the results obtained by another model. Also, we can test another database such as Physikalisch-Technische Bundesanstalt (PTB) diagnostic database.

References

1. Celin, S., Vasanth, K.: ECG signal classification using various machine learning techniques. J. Med. Syst. **42**(12), 1–11 (2018)
2. Abrishami, H., et al.: P-QRS-T localization in ECG using deep learning. In: IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 210–213. Las Vegas, NV, USA (2018)
3. Hela, L., Raouf, K.: ECG multi-class classification using neural network as machine learning model. In: International Conference on Advanced Systems and Electric Technologies, pp. 473–478. Hammamet, Tunisia (2018)
4. Savalia, S., Emamian, V.: Cardiac arrhythmia classification by multi-layer perceptron and convolution neural networks. Bioengineering **5**(2), 35 (2018)
5. Rajkumar, A., et al.: Arrhythmia classification on ECG using deep learning. In: 5th International Conference on Advanced Computing and Communication Systems, pp. 365–369. India (2019)
6. Acharya, U., et al.: A deep convolutional neural network model to classify heartbeats. Comput. Biol. Med. **89**, 389–396 (2017)
7. Sihem, H., Yassine, B.A.: Toward improving person identification using the ElectroCardioGram (ECG) signal based on non-fiducial features. Multimed. Tools Appl. 18543–18561 (2020)
8. Fariha, M., et al.: Analysis of pan-tompkins algorithm performance with noisy ECG signals. J. Phys. **1532** (2020)
9. Sihem, H., Yassine, B.A.: An integration of features for person identification based on the PQRST fragments of ECG signals. Signal, Image Video Process. **16**, 2037–2043 (2022)
10. Oh, S.L., et al.: Comprehensive electrocardiographic diagnosis based on deep learning. Artif. Intell. Med. **103** (2020)
11. Swapna, G., et al.: Automated detection of diabetes using CNN and CNNLSTM network and heart rate signals. Procedia Comput. Sci. **132**, 1253–1262 (2018)
12. Islam, et al.: A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Inform. Med. Unlocked **20** (2020)
13. Verma, D., Agarwal, S.: Cardiac arrhythmia detection from single-lead ECG using CNN and LSTM assisted by oversampling. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 14–17 (2018)



Incremental Cluster Interpretation with Fuzzy ART in Web Analytics

Wui-Lee Chang(✉), Sing-Ling Ong, and Jill Ling

Drone Research and Application Center, University of Technology Sarawak, Sibu, Malaysia
wuileechang@gmail.com

Abstract. Clustering in web analytics extracts information from data based on similarity measurement on the data patterns, where similar data patterns are grouped as a cluster. However, the typical clustering methods used in web analytics suffer from three major shortcomings, viz., (1) the predefined number of clusters is hard to determined when new data are generated over time; (2) new data might not be adopted into the existing clusters; and (3) the information given by a cluster (centroid) is vague. In this study, an incremental learning method using the Fuzzy Adaptive Resonance Theory (Fuzzy ART) algorithm is adopted (1) to analyze the underlying structure (hidden message) of the data, and (2) to interpret cluster into an understandable and useful knowledge about user activity on a webpage. An experimental case study was conducted by capturing the integrated data from Google Analytics on the University of Technology Sarawak (UTS), Malaysia, website to analyze user activity on the webpage. The results were analyzed and discussed, and it shown that the information obtained at each cluster can be interpreted in term of cluster boundary at each feature space (dimension), whereas the user activity are explained from the cluster boundary without revisiting the trained data.

Keywords: Incremental Learning · Fuzzy ART · Clustering · Web Analytics

1 Introduction

Web analytics is a popular tool used in modern business models [1, 2] to automatically discover user activity by measuring and analyzing web data for improvements in digital marketing performance. Web data contain information that can be described with data patterns, which are attributed with features of web server logs [3, 4] that records user activity such as traffic sources or user visited sites, page viewed, and resources that link to the webpage. Each data pattern can be interpreted as a vector of multidimensional features that describe each user behavior or web activity [5]. The number of web data is increased upon each user visits and user activities are varied over time [5, 6]. Thus, a completely known or labelled data collection with labels is difficult to obtain [7].

Clustering (unsupervised classification) method [8], that groups unlabeled data set with similar data patterns into a number of cluster, is often used to extract hidden message from unlabeled data structure at each cluster. For instance, clustering in web analytics

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 510–520, 2023.
https://doi.org/10.1007/978-3-031-27409-1_46

application is often group similar data patterns from (server) log file and obtain the cluster labeling by analyzing the associated data at each cluster to better understanding user activity on a webpage [9]. K-means clustering [10, 11] is used to cluster text data on a webpage and Latent Dirichlet Allocation (LDA) [10] or normalized term frequency and inverse document frequency (NTF-IDF) [11] is used to extract the text labels at the clusters to understand the data. The data structure, either interpreted as the cluster labels or knowledge/information, are commonly extracted in post-processing, i.e., trained data that is associated at each cluster are further analyzed to extract the useful information (e.g., cluster size, labels and intervals) [8–17].

Despite the effectiveness of clustering is demonstrated in web analytics applications [9–11], it suffers from three common limitations of clustering, i.e., (1) the predefined number of clusters is hard to determined [18] when a complete (labelled) data set collection not available while new web data are generated over time [6]; (2) new (or unknown) data that are generated everyday might lead to performance degradation over time [19]; and (3) catastrophic forgetting [20] of the previously acquire data structure occurred when a cluster is no able to recognize it previously associated data, which might lead to imprecise knowledge/information interpretation after each learning.

In this study, an incremental learning approach, where trained data are discarded after each learning, is proposed to tackle the above mentioned problems using the Fuzzy Adaptive Resonance Theory (Fuzzy ART) [21] (1) to analyze the underlying structure (hidden message) of the data, and (2) to interpret data structure into an understandable description for user activity on a webpage. Motivated from previous works on incremental learning with clustering [14–17], clusters simplify the problems into groups (clusters) that can be visualize to further analyze the data structure over time. From the findings, each cluster that is attributed with a weight vector to represent the group (set) of similar data is insufficient to retain the previously acquired data structure, i.e., the previously associated data at a cluster is changed the weight vector is updated. Moreover, it is hard to obtain a good quality cluster with uniform data distribution (errors) over the clusters [8].

Fuzzy ART is a neural network-based model that learn new data one after another and discard them after each learning while tackle the stability-plasticity dilemma [21], where the clustering model is "stable" to recognize all trained data at each cluster and "plastic" to increase in the number of clusters to adopt new data. It is worth mentioning that the Fuzzy ART cluster is interpreted without the post-processing. Thus, it is crucial in extracting data structure from clusters from time-to-time without referring to the voluminous (trained) web data frequently. A case study is conducted using the integrated data from the Google Analytics tool of the website of University of Technology Sarawak (UTS), Malaysia, to obtain past (history) user activity to understand the type of users visiting the webpage.

The organization of the paper is as follows. Section 2 describes the background of the Fuzzy ART learning algorithm and structure. Section 3 proposes an incremental learning methodology-based Fuzzy ART clustering to analyze web analytics. Section 4 discusses the experimental results and findings, and Sect. 5 is the conclusion.

2 Fuzzy ART

The Fuzzy ART learning structure [21] is depicted in Fig. 1, which describes the input layer, layer 0, layer 1, and layer 2, respectively. On the input layer, an input pattern that is attributed with p features is denoted as $\boldsymbol{x} = \{x_1, \ldots, x_p\} \in \mathbb{R}^p$. At layer 0, each x_i, $i = 1, \ldots, p$, is normalized to a range of [0, 1] (i.e., denoted as a_i), and a new training data $\boldsymbol{A} = \left\{a_1, \ldots, a_p, a_1^{'}, \ldots, a_p^{'}\right\}$ is generated when p complement attributes (i.e., denoted as $a_i^{'} = 1 - a_i$) are added.

Layer 1 is called the short-term (temporary) memory, where $\boldsymbol{A}$ is fed for the learning process that includes category match, vigilance test, and growing. In this layer, $\boldsymbol{W}^1 = \left\{\boldsymbol{w}_1^1, \ldots, \boldsymbol{w}_c^1\right\}$ is duplicated from $\boldsymbol{W}^2 = \left\{\boldsymbol{w}_1^2, \ldots, \boldsymbol{w}_c^2\right\}$, where $\boldsymbol{W}^1 \equiv \boldsymbol{W}^2$. Each cluster is denoted as $\boldsymbol{w}_j^1 = \left[w_{j,1}, \ldots, w_{j,2p}\right]$ that describes the j-th cluster's prototype weight vector. The category match is conducted using Eq. 1,

$$T_j = \frac{\left|\boldsymbol{A} \wedge \boldsymbol{w}_j^1\right|}{\beta + \left|\boldsymbol{w}_j^1\right|} \in (0, 1), \tag{1}$$

where, $\boldsymbol{A} \wedge \boldsymbol{w}_j^1 = \min\left(\boldsymbol{A}, \boldsymbol{w}_j^1\right)$, $\left|\boldsymbol{w}_j^1\right| = \sum_{i=1}^{2p}\left(w_{j,i}^1\right)$, and $\beta \approx 0$ is a constant choice parameter. $T_j \approx 1$ when $\boldsymbol{A}$ is a member of $\boldsymbol{w}_j^1$, otherwise $T \approx 0$. A winner $J = \underset{j}{\text{argmax}}\left(T_j\right), j = 1, \ldots, c$, is determined among all clusters. If more than one T_j is maximal (i.e., J indexes are determined), T_j with the smallest j index is chosen as the final J index [21]. The vigilance test is a hypothesis test with a vigilance value $\rho \in [0, 1]$ to set (learn) or reset (not learn) $\boldsymbol{w}_J^1$ (as described using Eq. 2),

$$\frac{\left|\boldsymbol{A} \wedge \boldsymbol{w}_j^1\right|}{|\boldsymbol{A}|} \geq \rho, \tag{2}$$

where it is set if Eq. 2 is satisfied and otherwise reset. If it is set, update $\boldsymbol{w}_J^1$ with Eq. 3, where $\alpha \in [0, 1]$ is a constant learning rate parameter.

$$\boldsymbol{w}_J^1 = \alpha\left(\boldsymbol{A} \wedge \boldsymbol{w}_J^1\right) + (1 - \alpha)\left(\boldsymbol{w}_J^1\right) \tag{3}$$

If it is reset, repeat category match by omitting the previous J prototype weight vector and identify a new J.

The growing happens when the hypothesis test is not satisfied at all existing clusters, and a new cluster $\boldsymbol{w}_{c+1}^1$ is created, where $\boldsymbol{w}_{c+1}^1 = \boldsymbol{A}$.

Layer 2 is called the long-term memory, where $\boldsymbol{W}^2$ holds the previous cluster structure, and it is set as $\boldsymbol{W}^2 = \boldsymbol{W}^1$ after each learning process and for the next learning of new $\boldsymbol{A}$.

A Fuzzy ART cluster $\boldsymbol{w}_i^2$ structure can be represented with hyperbox (depicted in Fig. 2) that is interpreted with cluster intervals on each feature space, where lower bounds and upper bounds are described with $\boldsymbol{u}_j = \left[w_{i,1}^2, \ldots, w_{i,p}^2\right]$ and $\boldsymbol{v}_j =$

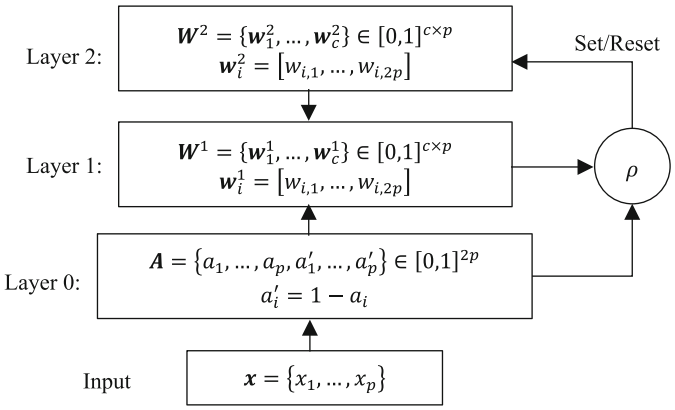


Fig. 1. Fuzzy ART learning structure

$\left[1 - w_{i,p+1}^2, \ldots, 1 - w_{i,2p}^2\right]$, respectively, from the prototype weight vector. Any data point that is bounded within the cluster intervals is recognized and associated to that cluster.

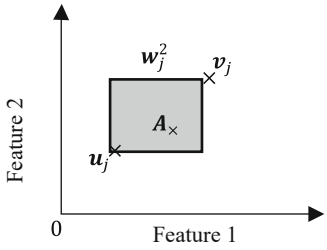


Fig. 2. A two-dimensional cluster (hyperbox) is labelled in grey box. A data point position is labelled with "×" symbol.

3 Proposed Methodology

The proposed web analytics methodology-based Fuzzy ART to understand web data from cluster is explained with the following steps.

Step 1: A web data pattern $\boldsymbol{x}_i = \{x_{i,1}, \ldots, x_{i,p}\}$ is fetched for learning, where p vector elements are determined that can reflect the user activity on the webpage. All feature elements are described in numerical data.

Step 2: $\boldsymbol{x}_i$ is normalized to a range of [0,1] using Eq. 4, where each feature element is divided by its respective maximum value of the feature space that is obtained from a set of data collection.

$$\boldsymbol{x}_i = \left\{ \frac{x_{i,1}}{\max(x_{*,1})}, \frac{\ldots}{\max(x_{*,\ldots})}, \frac{x_{i,p}}{\max(x_{*,p})} \right\} \tag{4}$$

Step 3: Determine a new training data vector $\boldsymbol{A}$.

$$\boldsymbol{A} = \left\{x_{i,1}, \ldots, x_{i,p}, x^{'}_{i,1}, \ldots, x^{'}_{i,p} | x^{'}_{i,*} = 1 - x_{i,*}\right\} \tag{5}$$

Step 4: Initialize a clustering model with either from an empty cluster or load previous model. An empty cluster is initiated with a cluster prototype weight vector that is normalized from $\boldsymbol{A}$, i.e., $W = \{\boldsymbol{w}_1 | \boldsymbol{w}_1 = \boldsymbol{A}\}$, and parameters of α, ρ, and β are determined in priori. The previous model consists of c prototype weight vectors, i.e., $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c\}$, and previously defined parameters of α, ρ, and β.

Step 5: Determine a winner among the clusters based on the matching function T_j in Eq. 6 and the winner J is determined with $\underset{j}{\text{argmax}} T_j$. $\boldsymbol{w}_J$ denotes the winner cluster for $\boldsymbol{A}$.

$$T_j = \frac{|\boldsymbol{A} \wedge \boldsymbol{w}_j|}{\beta + |\boldsymbol{w}_j|} \tag{6}$$

Step 6: A hypothesis test function H is evaluated at $\boldsymbol{w}_J$ to adapt (set) $\boldsymbol{A}$ or reject (reset) $\boldsymbol{A}$ using Eq. (7).

$$H_J = \frac{|\boldsymbol{A} \wedge \boldsymbol{w}_j|}{|\boldsymbol{A}|} \tag{7}$$

Step 7: Update $\boldsymbol{w}_J$ to adapt $\boldsymbol{A}$ if $H_J \geq \rho$ using Eq. (8). Otherwise, repeat Step 5 to determine another winner.

$$\boldsymbol{w}_J = \alpha(\boldsymbol{A} \wedge \boldsymbol{w}_J) + (1 - \alpha)(\boldsymbol{w}_J) \tag{8}$$

Step 8: Extract data structure from each cluster (that is denoted as I_j in Eq. (9)) at each feature space.

$$I_j = \left\{\left[w_{j,1}, 1 - w_{j,p+1}\right], \left[w_{j,2}, 1 - w_{j,p+2}\right], \ldots, \left[w_{j,p}, 1 - w_{j,p+p}\right]\right\}, \tag{9}$$

where cluster intervals are obtained with $\left[w_{j,*}, 1 - w_{j,p+*}\right]$.

Repeat Steps 1–8 for the next data patterns.

4 Experimental Case Study

A case study is conducted to evaluate the effectiveness of the proposed methodology by analyzing user activity on the website of University of Technology Sarawak (UTS), Malaysia. Web data are extracted from the Google Analytics tool that involve 14 selected features to describe the users. The web data taken from December 2019 to March 2022 are considered in the following analyses. The 14 selected features are the (1) visiting date, (2) daily total page viewed, (3) number of active users, (4) number of affinity users, (5) number of in-market segment users, (6) number of potential customer, (7) number of Malaysian users, (8) number of Non-Malaysian users, (9) number of English language

users, (10) number of Bahasa Malaysia language users, (11) number of Mandarin language users, (12) number of other languages users, (13) number of new users, and (14) number of returning users.

The distribution of the collected web data (852 data patterns) at each feature space, that are described in numerical representation and normalized to the range of [0,1], are first analyzed with box-and-whisker plot (depicted in Fig. 3) to understand the distribution. From Fig. 3, it is noted that outliers are commonly detected at each feature space. Due to the unlabeled web data are being used, all outliers are assumed the abnormal activity that worth for further investigation. For example, an outlier of very high (or very low) number of new users (at feature-13) indicates the effectiveness to attract new users. The same terminology is interpreted at features-2 to 14, while feature-1 indicates the daily activity.

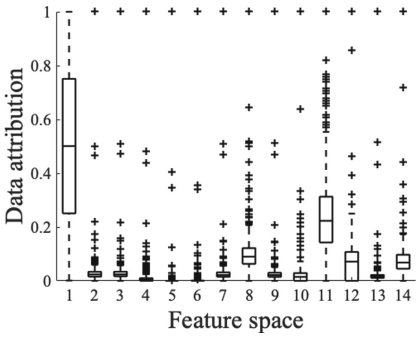


Fig. 3. A box-and-whisker plot of the collected web data on each of the 14 features. Features (1) to (14) are plotted from left to right sequence on the x-axis, while their normalized feature distributions are plotted on the y-axis. Symbol of "+" is used to indicate outlier or special case.

4.1 Data Structure

In this section, the intervals of a hyperbox are visualized using a rectangular box on the 14 features, as shown in Fig. 4, where $\rho = 0.8$, $\alpha = 1$ and $\beta = 0.002$ are set for the following discussion.

From Fig. 4, a notable character at each cluster can be observed and their interval values are depicted in Table 1. For example, cluster 1 (Fig. 4(a)), the cluster contains the information of features (1) visiting date contains of days 1 to 301, (2) daily total page viewed of 11 to 1700 (1.7k) views, (3) number of active users of 9 to 531 users, (4) number of affinity users of 0 to 3564 (3.6k) users, (5) number of in-market segment users of 0 to 120 users, (6) number of potential customer of 0 to 1066 (1.1k) users, (7) number of Malaysian users of 9 to 491 users, (8) number of Non-Malaysian users of 0 to 111 users, (9) number of English language users 8 to 444 users, (10) number of Bahasa Malaysia language users of 0 to 11 users, (11) number of Mandarin language users of 0 to 84 users, (12) number of other languages users of 0 to 9 users, (13) number of new users of 0 to 386 users, and (14) number of returning users of 9 to 226 users.

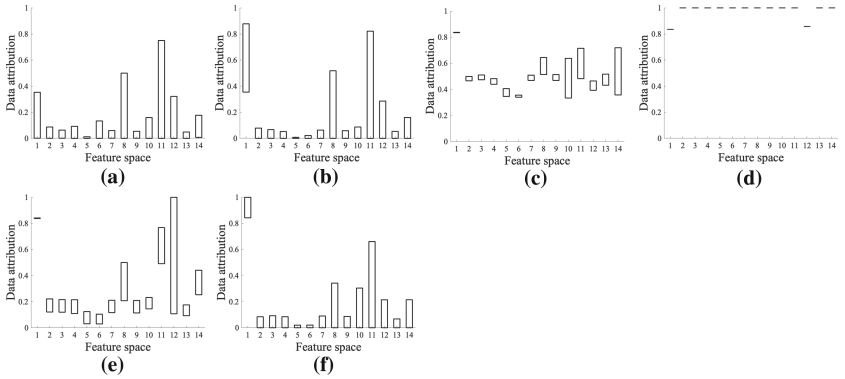


Fig. 4. The interpreted intervals of hyperbox of (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, and (f) cluster 6 at $\rho = 0.8$.

Table 1. Cluster intervals values interpreted from prototype weight vectors.

i	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	$\boldsymbol{u}_{1,i}$	$\boldsymbol{v}_{1,i}$	$\boldsymbol{u}_{2,i}$	$\boldsymbol{v}_{2,i}$	$\boldsymbol{u}_{3,i}$	$\boldsymbol{v}_{3,i}$	$\boldsymbol{u}_{4,i}$	$\boldsymbol{v}_{4,i}$	$\boldsymbol{u}_{5,i}$	$\boldsymbol{v}_{5,i}$	$\boldsymbol{u}_{6,i}$	$\boldsymbol{v}_{6,i}$
1	1	301	302	748	711	713	712	712	714	717	718	852
2	11	1.7k	0	1.5k	9k	9k	19k	19k	2k	4k	0	1.7k
3	9	531	0	562	4k	4k	8k	8k	1k	1.8k	0	777
4	0	3.6k	0	2k	17k	19k	39k	3.9k	4.2k	8.3k	0	3.3k
5	0	120	0	77	3.5k	4k	10k	10k	305	1.2k	0	201
6	0	1.1k	0	176	2.7k	2.8k	8k	8k	236	834	0	164
7	9	491	0	520	3.8k	4.1k	8.2k	8.2k	957	1.7k	0	741
8	0	111	0	115	114	143	222	222	46	111	0	76
9	8	444	0	480	3.8k	4.2k	8.2k	8.2k	933	1.7k	0	718
10	0	11	0	6	23	44	69	69	10	16	0	21
11	0	84	0	92	54	80	112	112	55	86	0	74
12	0	9	0	8	11	13	24	24	3	28	0	6
13	0	386	0	424	3.4k	4.1k	8k	8k	736	1.4k	0	539
14	9	226	0	204	456	918	1.3k	1.3k	323	563	0	274

The data associated within each cluster are shown in Fig. 5, to justify the data distribution within each cluster are bounded within the cluster intervals and the box-and-whisker plot are majority in normal skewed at all feature space, indicating a good representation of the clusters. Noted that Fig. 5 (d) contains only a single data patterns in cluster 4, that highlighted the outliers (refer to Fig. 3), and Fig. 5 (e) exhibited non-normal skewed to indicate the outliers determined in Fig. 3.

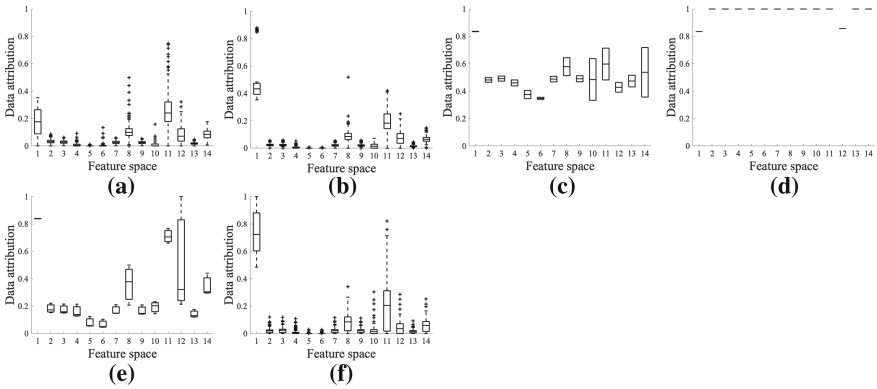


Fig. 5. The box-and-whisker plot on each 14 features with the associated data of (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, and (f) cluster 6 at $\rho = 0.8$.

4.2 Comparison of Data Structures

We further analyze the data structure obtained using K-Means algorithm and Evolving Vector Quantization (EVQ) [22] basic algorithm to compare with the previous results obtained using the proposed methodology.

The predefined number of clusters of K-Means is set to six clusters, as shown in Fig. 6, in the analysis. In the figure, trained data are revisited and clustered into their associative clusters, respectively, and the data distribution within each cluster is plotted with box-and-whisker plot. It is noted that the median of the box-and-whisker plot in cluster 4 of Fig. 6 (c) associating the outliers of the trained data (refer to Fig. 3). Even though the intervals are interpreted, but the real information contained within the cluster is limited. This can be observed through the illustration of the skewness of the box-and-whisker plots in Fig. 6 (c) that are majority skewed to the higher density of the data distribution.

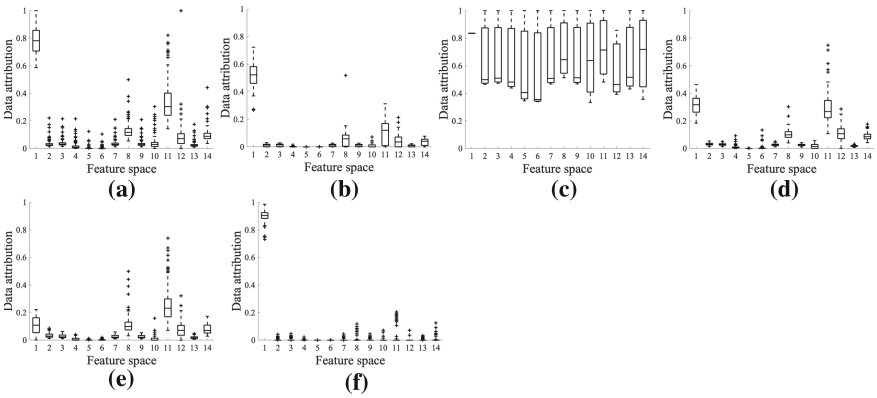


Fig. 6. The box-and-whisker plot on each 14 features with the associated data of (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, and (f) cluster 6 using K-Means of six clusters.

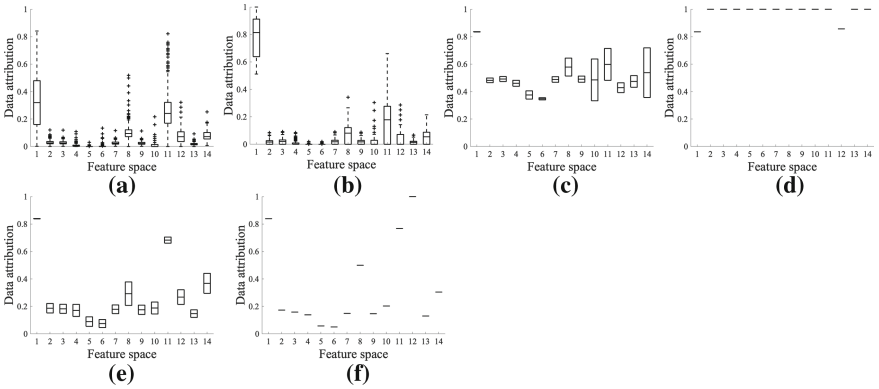


Fig. 7. The box-and-whisker plot on each 14 features with the associated data of (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, and (f) cluster 6 using Evolving Vector Quantization algorithm of $\alpha = 0.02$ and $h = 0.7$.

Two parameters of EVQ, i.e., the learning rate $\alpha = 0.02$ and the maximum cluster width $h = 0.7$, are set to obtain the six clusters, as shown in Fig. 7. Although EVQ keep track on its cluster width through data distribution within a cluster that is represented with a centroid weight vector, it suffers from catastrophic forgetting, where the previously recognized data are forgotten over time. Thus, the trained data are revisited and re-clustered into their associative clusters for the data distribution at each cluster to be plotted. The results have shown that clusters intervals of EVQ can be used to describe the data distribution (refer to Fig. 3), where the box-and-whiskers are mainly normal distributed.

4.3 Findings

From the analyses, the proposed methodology-based Fuzzy ART clustering with hyperbox prototype weight vectors are more informative with regards to cluster interval interpretation without the need to re-evaluate the trained data. The trained data that are generated over time, eventually obtained in high volume, and re-evaluation at the cluster post-processing for interpretation will become impractical over times.

5 Conclusion

This study proposed an incremental learning methodology-based Fuzzy ART clustering to discover the data structure on-the-fly, in web analytics application, through cluster (hyperbox) interval interpretation after each learning. The hyperbox intervals recognized trained data through their subset feature attributions, and the number of hyperboxes is increased to adapt new data. Thus, it is more practical to be applied for web analytics, where web data, which are generated over times, are learnt and discarded from after each learning and the trained information are retained within the hyperbox constantly.

While most of the web analytics applications aimed to predict user activity or behavior while browsing a webpage [9–11], the proposed methodology is feasible to be practically implemented as the knowledge discovery tool in the prediction function/model incrementally, without the need of re-training or re-learning the model.

References

1. Król, K.: The application of web analytics by owners of rural tourism facilities in Poland–diagnosis and an attempt at a measurement. J. Agribus. Rural Dev. **54**(4), 319–326 (2019)
2. Kö, A., Kovacs, T.: Business analytics in production management–challenges and opportunities using real-world case experience. In: Working Conference on Virtual Enterprises, pp. 558–566 (2021)
3. Nazar, N., Shukla, V.K., Kaur, G., Pandey, N.: Integrating web server log forensics through deep learning. In: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–6 (2021)
4. Terragni, A., Hassani, M.: Analyzing customer journey with process mining: from discovery to recommendations. In: 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 224–229 (2018)
5. Tamilselvi, T., Tholkappia Arasu, G.: Handling high web access utility mining using intelligent hybrid hill climbing algorithm based tree construction. Clust. Comput. **22**(1), 145–155 (2018). https://doi.org/10.1007/s10586-018-1959-8
6. Nasraoui, O., Soliman, M., Saka, E., Badia, A., Germain, R.: A web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE Trans. Knowl. Data Eng. **20**(2), 202–215 (2008)
7. Li, N., Shepperd, M., Guo, Y.: A systematic review of unsupervised learning techniques for software defect prediction. Inf. Softw. Technol. **122**(February 2019), 106287 (2020)
8. Sinaga, K.P., Yang, M.: Unsupervised K-means clustering algorithm. IEEE Access **8**, 80716–80727 (2020)
9. Fabra, J., Álvarez, P., Ezpeleta, J.: Log-based session profiling and online behavioral prediction in e-commerce websites. IEEE Access **8**, 171834–171850 (2020)
10. Janmaijaya, M., Shukla, A.K., Muhuri, P.K., Abraham, A.: Industry 4.0: Latent Dirichlet Allocation and clustering based theme identification of bibliography. Eng. Appl. Artif. Intell. **103**, 104280 (2021)
11. Chang, A.C., Trappey, C.V., Trappey, A.J., Chen, L.W.: Web mining customer perceptions to define product positions and design preferences. Int. J. Semant. Web Inf. Syst. **16**(2), 42–58 (2020)
12. Pehlivan, N.Y., Turksen, I.B.: A novel multiplicative fuzzy regression function with a multiplicative fuzzy clustering algorithm. Rom. J. Inf. Sci. Technol. **24**(1), 79–98 (2021)
13. Borlea, I.D., Precup, R.E., Borlea, A.B.: Improvement of K-means cluster quality by post processing resulted clusters. Procedia Comput. Sci. **199**, 63–70 (2022)
14. Chang, W.L., Tay, K.M., Lim, C.P.: Clustering and visualization of failure modes using an evolving tree. Expert Syst. Appl. **42**(20), 7235–7244 (2015)
15. Chang, W.L., Pang, L.M., Tay, K.M.: Application of self-organizing map to failure modes and effects analysis methodology. Neurocomputing **249**, 314–320 (2017)
16. Chang, W.L., Tay, K.M.: A new evolving tree for text document clustering and visualization. In: Soft Computing in Industrial Applications, vol. 223. Springer (2014)
17. Chang, W.L., Tay, K.M., Lim, C.P.: A new evolving tree-based model with local re-learning for document clustering and visualization. Neural Process. Lett. **46**(2), 379–409 (2017). https://doi.org/10.1007/s11063-017-9597-3

18. Khan, I., Luo, Z., Huang, J.Z., Shahzad, W.: Variable weighting in fuzzy k-means clustering to determine the number of clusters. IEEE Trans. Knowl. Data Eng. **32**(9), 1838–1853 (2019)
19. Su, H., Qi, W., Hu, Y., Karimi, H.R., Ferrigno, G., De Momi, E.: An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators. IEEE Trans. Ind. Inform. **18**(3), 1864–1872 (2020)
20. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: a continual structure learning framework for overcoming catastrophic forgetting. In: International Conference on Machine Learning, pp. 3925–3934 (2019)
21. Carpenter, G., Grossberg, S., Markuzon, N., Reynolds, J.H.: Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog. IEEE Trans. Neural Netw. **3**(5), 220–226 (1992)
22. Lughofer, E.: Evolving Fuzzy Systems Methodologies, Advanced Concepts and Applications, vol. 266 (2011)

TURBaN: A Theory-Guided Model for Unemployment Rate Prediction Using Bayesian Network in Pandemic Scenario

Monidipa Das[1(✉)], Aysha Basheer[2], and Sanghamitra Bandyopadhyay[2]

[1] Indian Institute of Technology (Indian School of Mines), Dhanbad 826004, India
monidipadas@hotmail.com
[2] Indian Statistical Institute, Kolkata 700108, India

Abstract. Unemployment rate is one of the key contributors that reflect the economic condition of a country. Accurate prediction of unemployment rate is a critically significant as well as demanding task which helps the government and the policymakers to make vital decisions. Though the recent research thrust is primarily towards hybridization of various linear and non-linear models, these may not perform satisfactorily well under the circumstances of unexpected events, e.g., during sudden outbreak of any infectious disease. In this paper, we explore this fact with respect to the current scenario of coronavirus disease (COVID) pandemic. Further, we show that explicit Bayesian modeling of pandemic impact on unemployment rate, together with theoretical insights from epidemiological models, can address this issue to some extent. Our developed theory-guided model for unemployment rate prediction using Bayesian network (TURBaN) is evaluated in terms of predicting unemployment rate in various states of India under COVID-19 pandemic scenario. The experimental result demonstrates the efficacy of TURBaN, which outperforms the state-of-the-art hybrid techniques in majority of the cases.

Keywords: Unemployment rate · Time series prediction · Bayesian network · Epidemiology · Theory-guided modeling

1 Introduction

Unemployment rate can be simply described as the percentage of individuals in the labour force who are currently unemployed in spite of having capability to work. This is one of the major social problems, which also works as a key driving force behind the slow-down of financial/economical growth of a country. Slowing down of the economy, in turn, reduces the demand of the enterprises for work, and thus leads to the consequence of increasing unemployment rate [6]. An accurate prediction of unemployment rate, therefore, is of paramount importance that helps in making appropriate decision and in designing effective plans

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 521–531, 2023.
https://doi.org/10.1007/978-3-031-27409-1_47

by the government and the various policy-makers. However, the prediction of any macroeconomic variable, like unemployment rate, is not a trivial task, since these are mostly non-stationary and non-linear in nature [3]. Though the combination of linear and non-linear prediction models in recent years have shown promising performance in this context, these are mostly univariate models, and therefore, fail to capture the influence from external factors [11]. Further, in the scenario of COVID-19 pandemic, the external influencing factors, especially, the disease spread pattern (in terms of daily increment/decrement of infected, recovered, and deceased case counts) itself need proper modeling so as to get the future status of the same. However, the modeling of the pandemic pattern is also not straightforward task and this requires adequate theoretical knowledge on epidemiology. Hence, there still remains huge scope of developing improved techniques of predicting unemployment rate while tackling these crucial issues.

Our Contributions: In this paper, we attempt to address the aforementioned challenges by developing a theory-guided unemployment rate prediction model based on Bayesian network, hereafter termed as TURBaN. The TURBaN is built on the base technology of theory-guided Bayesian network (TGBN) model, as introduced in our previous work [2]. The generative nonlinear modeling using Bayesian network helps TURBaN to handle the nonlinear nature of unemployment rate time series, and at the same time, takes care of the influence from multiple external factors. On the other side, the theoretical guidance makes TURBaN capable of better modeling the COVID spread pattern that may have direct or indirect influence on unemployment rate time series. Our major contributions in this regard can be summarized as follows:

- Exploring theory-guided Bayesian network (TGBN) for multivariate prediction of unemployment rate;
- Developing TURBaN as TGBN-based prediction model, capable of efficiently modeling the influence of pandemic on unemployment rate time series;
- Validating our model with respect to predicting monthly unemployment rate time series in nine states from various parts of India;

Our empirical study of prediction across different forecast horizons demonstrates superiority of TURBaN over several benchmarks and state-of-the-arts.

The rest of the paper is outlined as follows. Section 2 discusses on the various recent and related research works. Section 3 presents the methodological aspects of our proposed TURBaN. Section 4 describes the experimental details including dataset, baselines, set-up, and major findings. Finally, we conclude in Sect. 5.

2 Related Works

Prediction of unemployment rate is a widely explored area both from traditional statistical and modern machine learning perspectives. Among the various traditional models, the autoregressive integrated moving average (ARIMA) model [6,9], the Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH) model [8], and their variants have been most widely used for the unemployment rate forecasting purpose. However, majority of these models are linear,

and hence, not very suitable for longer term prediction of non-linear and non-symmetric time series, like the unemployment rate. On the other side, the modern machine learning approaches based on variants of artificial neural network (ANN) can inherently deal with the non-linearity in the unemployment rate time series, and therefore, have become popular in recent years [7]. However, the unemployment rate datasets can contain both linear and nonlinear components, and so, the decision cannot be made based on either of these models, separately. The recent research thrust is therefore found in developing hybrid models that combine both the linear and the nonlinear approaches to forecast the unemployment rate [1,3]. For example, in [3], the authors have proposed a 'hybrid ARIMA-ARNN' model, where the ARIMA model is applied to catch the linear patterns of the data set. Subsequently, the residual error values of the ARIMA model are fed to an auto-regressive neural network (ARNN) model to capture the nonlinear trends in the data. The model assumes that the linear and the nonlinear patterns in the unemployment rate time series data can be captured separately. Similar kinds of hybrid models have been explored in the work of Ahmed et al. [1] and Lai et al. [10] as well. Apart from the ARIMA-ARNN, these two works have also studied ARIMA-ANN (combination of ARIMA and ANN) and ARIMA-SVM (combination of ARIMA and support vector machine) models. All these hybrid models have been found to show far more promising forecast performance compared to the stand-alone statistical and machine learning approaches. However, these are primarily univariate models, and hence, are inherently limited to predict based on only the past values of unemployment rate and ignore the other external factors, like the disease infected case counts in the situation of a pandemic, which can directly or indirectly influence the unemployment rate. Though the statistical vector auto-regression (VAR) model and its variants can overcome this issue, these require extensive domain knowledge for manipulation of the input data so as to deal with the non-linearity in the dataset [11]. Contrarily, the machine learning models are more potent regarding the multivariate prediction of the unemployment rate time series.

Nevertheless, in the present context of COVID-19 pandemic, modeling the influence from daily infected, recovered, and deceased case counts is not simple, since these external variables themselves need appropriate modeling, which require adequate knowledge in epidemiology. It may be noted here that the recently introduced theory-guided data-driven approaches [2,4,5] have huge potentiality to tackle this intelligently. In this paper, we explore the same on prediction of unemployment rate time series with consideration to the effect of the spread of COVID-19. To the best of our knowledge, ours is the first to investigate the effectiveness of theory-guided model for unemployment rate prediction.

3 Proposed Model: TURBaN

An overview of our developed theory-guided model for unemployment rate prediction using Bayesian network (TURBaN) is shown in Fig. 1. As shown in the

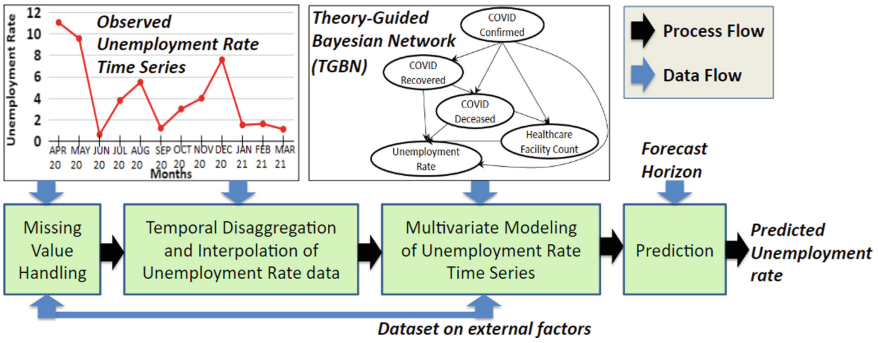


Fig. 1. An overview of process and data flow within TURBaN

figure, the model is comprised of four major steps, namely *missing value handling*, *temporal disaggregation followed by interpolation*, *multivariate modeling based on theory-guided Bayesian network*, and *prediction*. Each of these steps is further discussed below.

3.1 Missing Value Handling

TURBaN is developed in the recent background of COVID-19 pandemic. Therefore, the multivariate prediction of unemployment rate in TURBaN considers the disease development statistics, including the COVID confirmed case count ($\mathcal{CC}$), recovered case count ($\mathcal{RC}$), and deceased case count ($\mathcal{DC}$) on daily basis. The issue of missing value in these disease datasets and also in the unemployment rate dataset is handled by employing mean value substitution technique. We also consider the healthcare infrastructure data, in terms of the number of COVID-dedicated healthcare facility count ($\mathcal{HF}$) as another possible external influencing factor, which is assumed to remain the same in present days.

3.2 Temporal Dis-Aggregation and Interpolation

The prime objective of this step is to eliminate the data frequency issue. Note that the disease count datasets are available on daily basis, whereas the unemployment rate dataset is available on monthly basis. Temporal disaggregation followed by interpolation [12] converts the monthly unemployment rate data to daily scale, such that the mean of the interpolated data over each month remains the same as the original monthly unemployment rate value.

3.3 Multivariate Modeling of Unemployment Rate Tme Series

The multivariate modeling of unemployment rate in TURBaN is achieved by using the theory-guided Bayesian network (TGBN). In TGBN, the Bayesian network (BN) helps in generative modeling of the influence of external factors

on unemployment rate. The theoretical guidance regarding COVID development pattern is obtained from the epidemiological SIRD (Susceptible-Infected-Recovered-Dead) model [2], which is subsequently exploited by the TGBN to predict the unemployment rate in the context of pandemic (refer to Sect. 3.4).

The directed acyclic graph (DAG) of TGBN is obtained by employing score-based technique such that the network captures the causal relationships between the disease variables ($\mathcal{CC}$, $\mathcal{RC}$, $\mathcal{DC}$), unemployment rate ($\mathcal{UR}$), and healthcare facility count ($\mathcal{HF}$) at best. The DAG obtained by TURBaN, when applied on the Indian dataset, is shown in Fig. 1. As per this structure, the network parameters are computed in terms of conditional probability distributions:

$$\mathcal{P}(\mathcal{CC}) \sim \mathcal{N}(\theta_{0(\mathcal{CC})}, \sigma_{\mathcal{CC}}) \tag{1}$$

$$\mathcal{P}(\mathcal{RC}|\mathcal{CC}) \sim \mathcal{N}(\theta_{0(\mathcal{RC})} + \theta_{1(\mathcal{RC})}.\mathcal{CC}, \sigma_{\mathcal{RC}}) \tag{2}$$

$$\mathcal{P}(\mathcal{DC}|\mathcal{CC}, \mathcal{RC}) \sim \mathcal{N}(\theta_{0(\mathcal{DC})} + \theta_{1(\mathcal{DC})}.\mathcal{CC} + \theta_{2(\mathcal{DC})}.\mathcal{RC}, \sigma_{\mathcal{DC}}) \tag{3}$$

$$\mathcal{P}(\mathcal{HF}|\mathcal{CC}, \mathcal{DC}) \sim \mathcal{N}(\theta_{0(\mathcal{HF})} + \theta_{1(\mathcal{HF})}.CC + \theta_{2(\mathcal{HF})}.\mathcal{DC}, \sigma_{\mathcal{HF}}) \tag{4}$$

$$\begin{aligned}\mathcal{P}(\mathcal{UR}|\mathcal{HF}, \mathcal{CC}, \mathcal{RC}, \mathcal{DC}) \sim \mathcal{N}(\theta_{0(\mathcal{UR})} + \theta_{1(\mathcal{UR})}.\mathcal{HF} \\ + \theta_{2(\mathcal{UR})}.\mathcal{CC} + \theta_{3(\mathcal{UR})}.\mathcal{RC} + \theta_{4(\mathcal{UR})}.\mathcal{DC}, \sigma_{\mathcal{UR}})\end{aligned} \tag{5}$$

$\mathcal{N}$ indicates the Gaussian distribution, σ is the standard deviation associated with the nodes (subscript) and θs denote the parameters regulating the mean.

3.4 Prediction

In TGBN, the theoretical guidance from SIRD is utilized during the prediction step of TURBaN. As per the SIRD model, the total population ($\mathbb{N}$) at any time t is the sum of sub-population of Susceptible ($\mathcal{S}_t$), Infected ($\mathcal{I}_t$), Recovered ($\mathcal{R}_t$), and Dead ($\mathcal{D}_t$), and is governed by the following set of differential equations,

$$\mathcal{S}'_t = -\beta \mathcal{S}_t \mathcal{I}_t \tag{6}$$

$$\mathcal{I}'_t = \beta \mathcal{S}_t \mathcal{I}_t - \gamma \mathcal{I}_t - \mu \mathcal{I}_t \tag{7}$$

$$\mathcal{R}'_t = \gamma \mathcal{I}_t \tag{8}$$

$$\mathcal{D}'_t = \mu \mathcal{I}_t \tag{9}$$

Here, γ, β, and μ are the parameters indicating recovery rate per unit of time, infected rate per unit of time, and death rate per unit of time, respectively. $\mathcal{I}_t$ can be computed based on daily COVID case counts ($\mathcal{CC}_t$,$\mathcal{RC}_t$,$\mathcal{DC}_t$), as follows:

$$\mathcal{I}_t = \sum_{i=1}^{t} \mathcal{CC}_t - \left(\sum_{i=1}^{t} \mathcal{RC}_t + \sum_{i=1}^{t} \mathcal{DC}_t \right) \tag{10}$$

This disease development pattern, as presented through the SIRD model in eqs. 6-9, provides TURBaN a view of the pandemic situation in future and thereafter helps TGBN to predict the unemployment rate ($\mathcal{UR}$) accordingly.

Given the forecast horizon for the unemployment rate time series and the present healthcare infrastructure ($\mathcal{HF}$) of the study-region, TURBaN first employs the trained TGBN to separately infer the COVID case counts for each day t at the forecast horizon, and then the value that matches the best with the SIRD-predicted pandemic pattern is treated as the predicted COVID case count. This can be expressed as follows:

$$\mathcal{CC}_t = \underset{v_i \in \{v | \mathcal{P}(\mathcal{CC}_t = v | \mathcal{HF}) \geq th\}}{argmin} |v_i - \mathcal{CC}_t^{SIRD}| \tag{11}$$

$$\mathcal{RC}_t = \underset{v_i \in \{v | \mathcal{P}(\mathcal{RC}_t = v | \mathcal{HF}) \geq th\}}{argmin} |v_i - \mathcal{RC}_t^{SIRD}| \tag{12}$$

$$\mathcal{DC}_t = \underset{v_i \in \{v | \mathcal{P}(\mathcal{DC}_t = v | \mathcal{HF}) \geq th\}}{argmin} |v_i - \mathcal{DC}_t^{SIRD}| \tag{13}$$

where th indicates threshold that helps in sampling the most probable case counts from a set of values. These TGBN predicted $\mathcal{CC}_t$, $\mathcal{RC}_t$, $\mathcal{DC}_t$, together with the healthcare infrastructure condition ($\mathcal{HF}$) is considered to be the evidence ($\mathcal{E}$) for predicting the unemployment rate at t as follows:

$$\mathcal{UR}_t = \{v,\ \mathcal{P}(\mathcal{UR}_t = v | \mathcal{E}) = 1\} \tag{14}$$

All the $\mathcal{UR}_t$ values, averaged over each month, is treated as the TURBaN-predicted unemployment rate value for that month in the given forecast horizon.

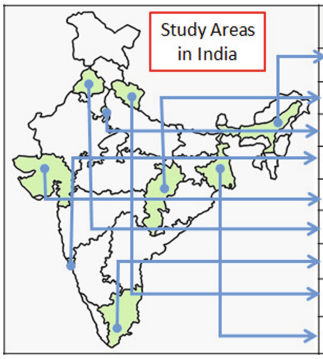
4 Experimentation

This section empirically evaluates TURBaN with consideration to the recent background of COVID-19 pandemic.

4.1 Study Area and Datasets

The empirical study is conducted to predict the unemployment rate in 9 selected states from various parts of India. A summary of the same is given in Fig. 2. The historical data of monthly unemployment rate in these states (refer Fig. 3) are collected from the Reserve Bank of India,[1] whereas the daily disease data and the healthcare infrastructure data are obtained from public sources.[2,3]

[1] HandBook of Statistics on Indian Economy (2020): https://www.rbi.org.in/.
[2] COVID case data: https://data.covid19india.org/.
[3] Healthcare facility data: https://www.indiastat.com/table/health/state-wise-number-type-health-facility-coronavirus/1411054.



States	COVID-19 dedicated Healthcare Facility Count	Population
Assam (AS)	378	34586234
Chhattisgarh (CT)	146	28566990
Delhi (DL)	155	18345784
Goa (GA)	46	1542750
Gujarat(GJ)	707	63907200
Punjab (PJ)	281	30141373
Tamil Nadu (TN)	1275	77841267
Uttarakhand (UK)	436	11250858
West Bengal (WB)	1208	99609303

Fig. 2. The states of India, studied for our experimentation purpose

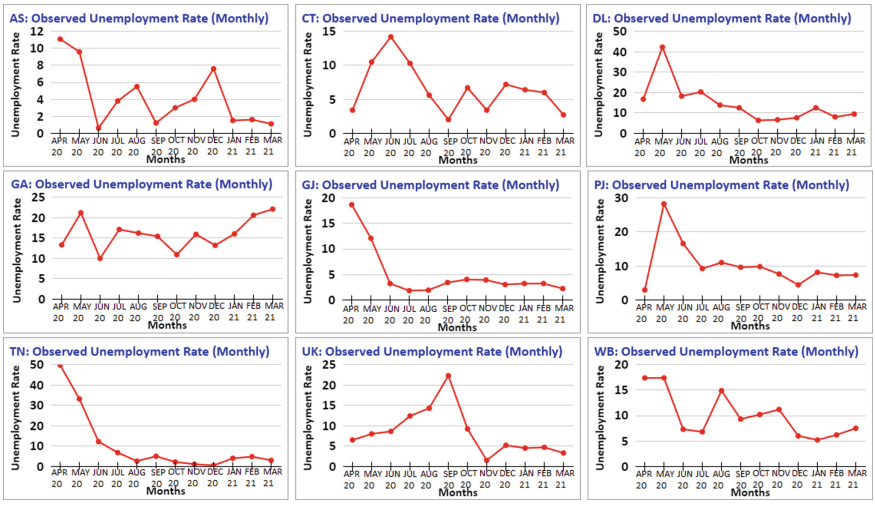


Fig. 3. Observed monthly unemployment rate data for the various states

4.2 Baselines and Experimental Set-up

Our devised TURBaN is evaluated in comparison with two traditional statistical benchmarks (ARIMA and GRACH), two machine learning benchmarks (ANN and Linear Regression), and two state-of-the-art hybrid models (ARIMA+ANN [1] and ARIMA+ARNN [3]). Further, to examine the effectiveness of hybridization with theoretical model, we also perform ablation study by eliminating the SIRD provided disease dynamics, and instead, using linear regression (LR) to get the future status of the pandemic. The model thus obtained is named as LR+BN. The training dataset is considered to have the data till March 2021, based on which we perform ***one month*** ahead and ***four month*** ahead prediction of the unemployment rate, respectively. All the models are executed in R-tool environment in Windows 64-bit OS (2.5 GHz CPU; 16 GB RAM).

4.3 Performance Metrics

The model performance has been measured using two popular evaluation metrics, namely the Normalized Root Mean Squared Error (NRMSE) and the Mean Absolute Percentage Error (MAPE). Mathematically, these can be expressed as:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(o_i - p_i\right)^2}}{max(o) - min(o)} \tag{15}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|o_i - p_i|}{o_i} \times 100\% \tag{16}$$

where, n is total no. of prediction days, o_i is the i-th observed unemployment rate; p_i is the respective predicted value; $max(o)$ and $min(o)$ denote the maximum and the minimum values of unemployment rate, found in observed data. In an ideal case of prediction, both NRMSE and MAPE values become 0.

Table 1. Comparative study of ***one month ahead*** prediction (i.e. prediction for APR-2021) of Unemployment Rate [boldface indicates the best performance per state]

Evaluation	Prediction	States								
Metrics	Model	AS	CT	DL	GA	GJ	PJ	TN	UK	WB
NRMSE	ARIMA	0.383	0.043	0.528	0.503	0.036	0.192	0.022	0.007	0.192
	GARCH	0.327	0.256	0.482	0.503	0.070	0.099	0.022	0.115	0.099
	LR	0.440	0.462	**0.244**	0.460	0.531	0.246	0.296	0.035	0.246
	ANN	0.266	0.207	0.526	0.537	0.068	0.190	**0.016**	**0.001**	0.190
	ARIMA+ANN	0.523	0.271	0.377	0.349	0.146	0.301	0.069	0.108	0.301
	ARIMA+ARNN	0.283	0.148	0.284	0.306	0.126	0.214	0.045	0.084	0.317
	LR+BN	0.376	0.284	0.360	0.507	0.188	0.187	0.160	0.110	0.187
	TURBaN	**0.161**	**0.014**	0.483	**0.244**	**0.025**	**0.066**	0.044	0.076	**0.066**
MAPE	ARIMA	20.083	0.177	0.696	0.378	0.333	0.918	0.478	**0.025**	0.013
	GARCH	17.168	1.040	0.635	0.378	0.661	0.475	0.471	0.400	0.159
	LR	23.086	1.878	**0.322**	0.345	4.986	1.180	6.342	0.123	0.308
	ANN	13.949	0.843	0.693	0.403	0.641	0.912	**0.353**	0.003	0.118
	ARIMA+ANN	27.449	1.103	0.497	0.262	1.370	1.445	1.475	0.373	0.582
	ARIMA+ARNN	14.864	0.604	0.374	0.230	1.179	1.027	0.974	0.289	0.521
	LR+BN	19.759	1.157	0.474	0.380	1.765	0.895	3.425	0.380	0.298
	TURBaN	**8.458**	**0.058**	0.637	**0.183**	**0.233**	**0.315**	0.933	0.262	**0.010**

4.4 Results and Discussions

The results of experimentation are summarized in Tables 1 and 2 and also depicted in Fig. 4. Following are the major findings that we obtain by analyzing the results.

– In is evident from Tables 1 and 2 that TURBaN outperforms all the other considered models in majority of the cases and with a large margin. Though ARIMA, ANN, and LR show promising performance in some of the instances,

our designed TURBaN, which is built on TGBN, offers a more consistent prediction producing average NRMSE of 0.07 and average MAPE of 0.67% only. This shows the benefit of considering external influence as well as theoretical guidance for unemployment rate prediction in a pandemic scenario.

– The superiority of TURBaN over the others is also clearly visible from Fig. 4. As per the figure, the TURBaN predicted unemployment rates of all the states

Table 2. Comparative study of ***four month ahead*** prediction (i.e. prediction for JUL-2021) of Unemployment Rate [boldface indicates the best performance per state]

Evaluation	Prediction	States								
Metrics	Model	AS	CT	DL	GA	GJ	PJ	TN	UK	WB
NRMSE	ARIMA	0.221	0.073	0.047	0.296	0.213	0.129	0.183	0.235	**0.008**
	GARCH	0.165	0.239	**0.021**	0.296	0.070	0.036	0.029	0.250	0.113
	LR	0.214	0.389	**0.020**	0.243	0.214	0.160	0.033	0.250	0.301
	ANN	0.104	0.191	0.065	0.298	0.068	0.127	0.034	0.136	0.088
	ARIMA+ANN	0.277	0.443	0.265	1.113	0.142	0.105	**0.0002**	0.363	0.497
	ARIMA+ARNN	0.097	0.209	0.146	0.183	0.107	0.118	0.013	0.192	0.259
	LR+BN	0.214	0.268	0.101	0.299	0.188	0.124	0.109	0.244	0.197
	TURBaN	**0.003**	**0.031**	**0.021**	**0.039**	**0.025**	**0.006**	0.007	**0.058**	0.020
MAPE	ARIMA	1.219	0.278	0.157	0.263	2.000	0.473	1.875	1.524	**0.014**
	GARCH	0.912	0.913	0.069	0.263	0.661	0.133	0.295	1.625	0.190
	LR	1.183	1.485	**0.066**	0.216	2.007	0.590	0.337	1.627	0.508
	ANN	0.574	0.728	0.217	0.265	0.641	0.469	0.350	0.881	0.149
	ARIMA+ANN	1.529	1.688	0.892	0.990	1.335	0.386	**0.002**	2.359	0.840
	ARIMA+ARNN	0.536	0.796	0.492	0.163	1.004	0.434	0.131	1.249	0.437
	LR+BN	1.185	1.022	0.341	0.266	1.765	0.456	1.120	1.588	0.333
	TURBaN	**0.018**	**0.119**	0.071	**0.035**	**0.233**	**0.021**	0.077	**0.374**	0.034

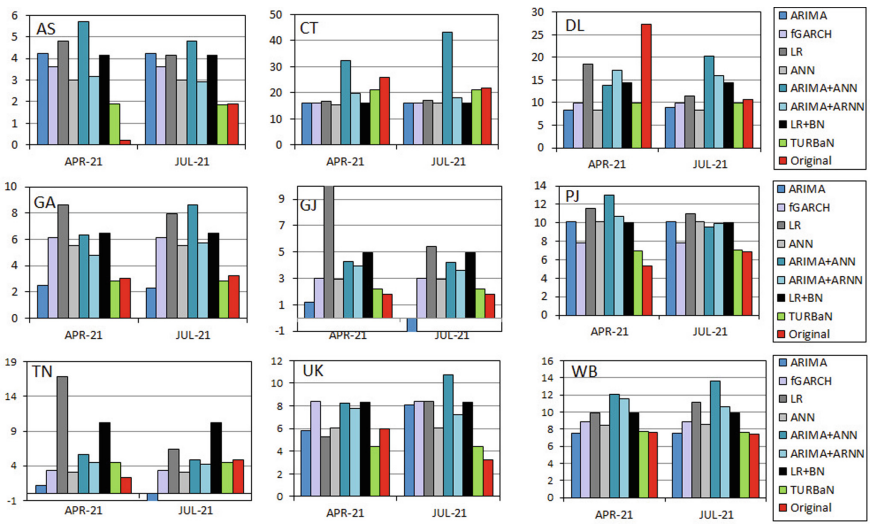


Fig. 4. Predicted versus Actual unemployment rate for the various states

for both April-2021 and July-2021 have the best match with the respective actual values. Moreover, the trend of change in the unemployment rate from April-2021 to July-2021 is also better captured by TURBaN, whereas that captured by the state-of-the-art hybrid models are surprisingly opposite. This again demonstrates the efficacy of TURBaN.
- Comparative study of TURBaN and LR+BN also shows that the theoretical guidance in TGBN helps TURBaN to substantially reduce prediction error (by 73%) than the case when the theoretical guidance is not used.

Overall, our study further establishes the effectiveness of theory-guided Bayesian analysis in the context of predicting unemployment rate under pandemic scenario. Note that, though we have evaluated TURBaN with respect to unemployment rate prediction in India under COVID-19 pandemic scenario, the model is applicable in the background of pandemic in other countries as well.

5 Conclusions

This paper has presented TURBaN as a hybridization of theoretical and machine learning models for unemployment rate prediction. The main contribution of this work remains in exploring theory-guided Bayesian network for multivariate prediction of unemployment rate across multiple forecast horizons. Rigorous experiment with Indian datasets reveals the superiority of TURBaN over several benchmarks and state-of-the-arts. Future scopes remain in further enhancing the model with added domain semantics to better tackle the underlying uncertainty.

Acknowledgment. We acknowledge the Research Grant from the National Geospatial Programme division of the Department of Science and Technology, Government of India.

References

1. Ahmad, M., Khan, Y.A., Jiang, C., Kazmi, S.J.H., Abbas, S.Z.: The impact of covid-19 on unemployment rate: an intelligent based unemployment rate prediction in selected countries of europe. Int. J. Finance Econ (2021)
2. Basheer, A., Das, M., Bandyopadhyay, S.: Theory-guided Bayesian analysis for modeling impact of covid-19 on gross domestic product. In: TENCON 2022–2022 IEEE Region 10 Conference, pp. 1–6 (2022)
3. Chakraborty, T., Chakraborty, A.K., Biswas, M., Banerjee, S., Bhattacharya, S.: Unemployment rate forecasting: a hybrid approach. Comput. Econ. **57**(1), 183–201 (2021)
4. Das, M., Ghosh, A., Ghosh, S.K.: Does climate variability impact COVID-19 outbreak? an enhanced semantics-driven theory-guided model. SN Comput. Sci. **2**(6), 1–18 (2021)
5. Das, M., Ghosh, S.K.: Analyzing impact of climate variability on COVID-19 outbreak: a semantically-enhanced theory-guided data-driven approach. In: 8th ACM IKDD CODS and 26th COMAD, pp. 1–9 (2021)

6. Gostkowski, M., Rokicki, T.: Forecasting the unemployment rate: application of selected prediction methods. Eur. Res. Stud. **24**(3), 985–1000 (2021)
7. Katris, C.: Forecasting the unemployment of med counties using time series and neural network models. J. Stat. Econ. Methods **8**(2), 37–49 (2019)
8. Katris, C.: Prediction of unemployment rates with time series and machine learning techniques. Comput. Econ. **55**(2), 673–706 (2020)
9. Khan Jaffur, Z.R., Sookia, N.U.H., Nunkoo Gonpot, P., Seetanah, B.: Out-of-sample forecasting of the canadian unemployment rates using univariate models. Appl. Econ. Lett. **24**(15), 1097–1101 (2017)
10. Lai, H., Khan, Y.A., Thaljaoui, A., Chammam, W., Abbas, S.Z.: Covid-19 pandemic and unemployment rate: a hybrid unemployment rate prediction approach for developed and developing countries of Asia. Soft Comput. 1–16 (2021)
11. Mulaudzi, R., Ajoodha, R.: Application of deep learning to forecast the South African unemployment rate: a multivariate approach. In: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, pp. 1–6. IEEE (2020)
12. Sax, C., Steiner, P.: Temporal Disaggregation of Time Series (2013). https://journal.r-project.org/archive/2013-2/sax-steiner.pdf



Pre-training Meets Clustering: A Hybrid Extractive Multi-document Summarization Model

Akanksha Karotia(✉) and Seba Susan

Department of Information Technology, Delhi Technological University, Delhi, India
akankshakarotia@gmail.com

Abstract. In this era where a large amount of information has flooded the Internet, manual extraction and consumption of relevant information is very difficult and time-consuming**.** Therefore, an automated document summarization tool is necessary to excerpt important information from a set of documents that have similar or related subjects. Multi-document summarization allows retrieval of important and relevant content from multiple documents while minimizing redundancy. A multi-document text summarization system is developed in this study using an unsupervised extractive-based approach. The proposed model is a fusion of two learning paradigms: the T5 pre-trained transformer model and the K-Means clustering algorithm. We perform the experiments on the benchmark news article corpus Document Understanding Conference (DUC2004). The ROUGE evaluation metrics were used to estimate the performance of the proposed approach on the DUC2004. Outcomes validate that our proposed model shows greatly enhanced performance as compared to the existent unsupervised state-of-the-art approaches.

Keywords: Multi-document · Extractive Summarization · Clustering Algorithm · Unsupervised Technique · Text Summarization

1 Introduction

In light of the exponential growth in information resources, the readers are overburdened with tons of data. Getting a piece of relevant information and summarizing it manually in a short period is a very challenging and time-consuming task for humans [1]. By eliminating information redundancy, we can save time and save resources, so it is imperative to remove information redundancy. To address this problem, text summarization has become an increasingly important tool. Text summarization is a task that is considered a sequence-to-sequence operation. Machine translation is one other application of sequence-to-sequence learning that has gained success and popularity over time [2]. The automatic text summarizer selects the most relevant and meaningful information and compresses it into a shorter version while preserving the original meaning [3]. The aim is to create short, to-the-point summaries that deliver the most important information and keep readers engaged, without taking more time to get the knowledge they require. Current researches on automatic text summarization focus on summarizing multiple

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 532–542, 2023.
https://doi.org/10.1007/978-3-031-27409-1_48

documents rather than single documents [4]. When a summary is derived from one document, it is referred to as a single-document summary, while when it is retrieved from a set of documents on a specific topic, it is referred to as a multi-document summary [5]. As of yet, almost all current extractive text summarization models lack the ability to efficiently summarize multi-document data. Consequently, this paper aims to manage this gap.

This work makes the subsequent main additions:

1. We propose an extractive-based unsupervised algorithm for extracting summaries from multiple documents based on the integration of the T5 pre-trained transformer model and the K-Means clustering algorithm.
2. We extracted the salient texts in two-stages: generation of abstractive summaries using the T5 pre-trained model followed by clustering of the abstractive summaries using the K-means clustering algorithm; the extracted salient text's benefits to generate the final meaningful summary.
3. We use the ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-S, and ROUGE-SU metrics to evaluate the performance of our proposed work with respect to nine existing unsupervised extractive-based summarization methods for multiple documents on the standard DUC2004 corpus.

The remaining sections of the paper are arranged as follows. Related work is described in Sect. 2. Detailed information about the proposed method is given in Sect. 3. Our experiments and results are presented in Sect. 4. Section 5 of this paper summarizes the work and outlines future directions.

2 Background

Natural language processing (NLP) chores have been refined by advancements in transformer-based frameworks [18, 19]. Encoder-decoder models have been deployed for generative language tasks such as abstractive-based summarization and question-answering systems [14, 15]. The competence of transformers to achieve semantic learning has also been enhanced significantly.

With the ability to automatically group similar items, one can discover hidden similarities and key concepts and organize immense quantities of information into smaller sets. Data can be condensed to a considerable extent using clustering techniques [16]. Since the documents in multi-document summarization are taken from different origins, they are verbose and redundant in articulating beliefs, and the summary contains only a few key points. Based on the distance between sentences, we can form semantic clusters that can be compressed into a particular sentence representing the important content of each cluster [17]. Applications of clustering are finding documents in similar categories, organizing enormous documents, redundant content, and recommendation models.

Using statistics derived from word frequencies and distributions, Luhn [6] used machine learning techniques to calculate the importance of sentences and created summaries based on the best sentence scores obtained. Edmundson and Wyllys proposed a method [7] that is an improvement on the Luhn method described above. What sets

this summarizer apart from other summarizers is that it requires three additional heuristics to measure sentence importance which includes bonus words, stigma words, and stopwords. Mihalcea and Tarau proposed a TextRank algorithm in [8]. Graph-based representations are utilized for summarizing content by calculating intersections between texts. LexRank uses eigenvector centrality, another node centrality method that was introduced in [9]. Based on common content between words and sentences, PageRank is a system for summarizing documents and identifying the central sentence in a document. Latent Semantic Analysis (LSA) [10] is a powerful summarization tool that identifies the patterns of relationships between terms and concepts using the singular value decomposition technique. Kullback-Leibler Divergence (KLD) [11] calculates the difference between two probability distributions. This approach takes a matrix of KLD values of sentences from an input document and then selects sentences with a lower KLD value to form a summary. SumBasic [12] uses the average sentence score to break links and iteratively selects the sentence containing the highest content word score. GenCompareSum [13] is a recently introduced method for single document summarization that first divides the document into sentences, and then combines these sentences into segments. Each segment is fed to a transformer model to generate multiple text fragments per segment. A similarity matrix is then computed between the text fragments and the original document using the BERT score. As a result, a score is generated per sentence by adding the values of the text fragments. A summary is formed by compiling the N sentences with the highest scores.

3 Methodology

This work is motivated in parts by the research work of Bishop et al. in [13], who introduced the model named GenCompareSum, described briefly in Sect. 2. The proposed model is an unsupervised hybrid extractive-based strategy for a multi-document news article. It is a fusion of the T5 pre-trained transformer model and the K-Means clustering technique. Figure 1 shows the proposed model architecture.

3.1 Extracting the Key Texts from Each Document Using the T5 Pre-trained Transformer Model

The dataset contains a total of T documents. It is organized into C folders, with an average of n documents per folder, i.e. $D = \{D_1, D_2, D_3, \ldots, D_n\}$. A T5 pre-trained transformer model was used to generate summaries for all documents separately $Sum = \{Sum_1, Sum_2, Sum_3, \ldots, Sum_n\}$. Each document has k sentences on average $D_\mathrm{i} = \{q_1, q_2, q_3, \ldots, q_k\}$. The output produced for each document from the T5 pre-trained transformer model is $D_i^{'} = \{q_1^{'}, q_2^{'}, q_3^{'}, \ldots, q_l^{'}\}$, where l refers to how many sentences there are. Combine the summaries obtained from n documents into one. This consolidated summary document shows key information from each document that guides subsequent modules in creating the final summary.

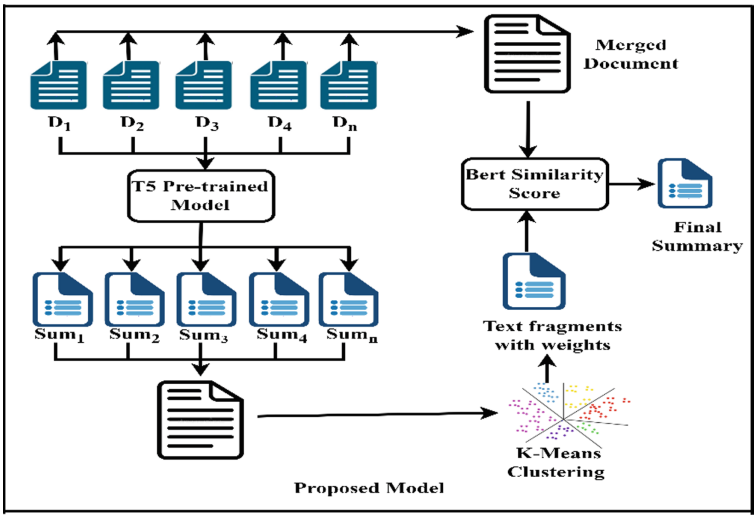


Fig. 1. Proposed model architecture

3.2 Extracting the Salient Texts Using the K-Means Clustering Strategy with the Help of Key Texts Extracted in the Above Section

The summary document generated in the above section is first tokenized into sentences and then pre-processed, such as removing stop words and converting uppercase letters to lowercase letters. Furthermore, to create the sentence vector, we used word embedding Word2Vec, a vector representation of all the words that make up the sentence, and used their average to arrive at a composite vector. After creating the sentence vector, we employed the K-Means clustering technique to group the sentence embeddings into a pre-defined number of clusters. In this case, we chose the number of clusters to be 7. Any cluster of sentence embeddings can be depicted as a group of sentences with the same meaning. These sentences have more or less the same information and meaning that can be expressed with only one sentence from each cluster. The vector representation of sentences with the lowest Euclidean distance from the cluster center represents the whole cluster. More the number of additional sentences in a group, the more vital that sentence is. Therefore, the text fragments obtained from i clusters $F = \{f_1, f_2, f_3, \ldots, f_i\}$ are associated with the count of sentences in that group, which is the weight w_t of the particular fragment f_t.

3.3 Final Summary Generation

Here, based on the BERT score we extract the sentences for the final summary. First, we create a merged document of the initial input documents. Next, we calculate the similarity between the text fragments obtained using the K-Means clustering technique and the merged input document using the BERT score. This similarity matrix is multiplied by the paired text fragment weights and summed to get the sentence scores. The top-scored sentences are fetched for the final summary and these extracted sentences are organized

in the ordered fashion as they were in the original document to get the final meaningful summary.

Equation (1) shows the formula used to calculate the similarity scores between the original document and text fragments to get the final sentence scores for each sentence r, where i is the number of clusters of text fragments that we get from the K-Means clustering algorithm and s is a sentence from the original merged document.

$$Sentence\ Score_r = \sum_{t=1}^{t=i} w_t * BERTSCORE(s_r, f_t) \tag{1}$$

4 Results and Evaluation

Google Colab online platform with 12 GB of RAM is used to perform the experiments. Our code is made available online[1] for facilitating future research. Results are evaluated on the DUC2004 dataset containing four human-generated summaries. We calculate the ROUGE scores between the four gold summaries and the system-generated summaries and average those scores to obtain the final ROUGE score as shown in Table 1 and Figs. 2, 3, 4, 5, 6, and 7. The results are presented in detail below.

4.1 Dataset Used

We have evaluated the performance of all the models on the dataset DUC2004 [20] for multi-document summarization. It contains a total of 500 news articles (documents) that are segregated into 50 folders, and each folder has ten documents on average. Each folder is associated with four different human-written summaries.

4.2 Performance Evaluation Metrics

ROUGE [21] is a performance metric used to evaluate the accuracy of generated summaries by comparing them to reference/gold summaries. It is a universally recognized benchmark used to evaluate machine-generated summaries and has also become the standard evaluation criterion for DUC collaborative work, making it popular for summary evaluation. Simply put, it works by comparing word occurrences in generated and referenced summaries.

4.3 Discussion on Results

Table 1 and Figs. 2, 3, 4, 5, 6, and 7 show the comparative analysis of various unsupervised state-of-the-art techniques, including the proposed system. Our model performs best with the highest ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-S, and ROUGE-SU F1 average scores of 34.013, 8.266, 2.951, 1.253, 10.366, 10.71. The least performing

[1] https://github.com/Akankshakarotia/Pre-training-meets-Clustering-A-Hybrid-Extractive-Multi-Document-Summarization-Model.

Table 1. Performance evaluation of various multi-document summarization models on the DUC2004 dataset using the ROUGE metric

Methods	Rouge-1	Rouge-2	Rouge-3	Rouge-4	Rouge-S	Rouge-SU
LSA	28.756 ± 0.311	3.837 ± 0.1857	0.804 ± 0.1047	0.241 ± 0.0469	6.244 ± 1.3214	6.758 ± 0.9288
Edmundson	30.272 ± 0.3581	5.587 ± 0.5128	1.506 ± 0.3232	0.567 ± 0.2134	7.961 ± 0.2085	8.24 ± 0.2176
Luhn	26.132 ± 0.4364	5.535 ± 0.1789	1.518 ± 0.0293	0.595 ± 0.0498	5.319 ± 0.2311	5.458 ± 0.2543
SumBasic	31.095 ± 0.7055	4.613 ± 0.1362	1.085 ± 0.1564	0.295 ± 0.1187	7.877 ± 0.4189	8.34 ± 0.4148
TextRank	26.146 ± 0.5377	5.71 ± 0.2947	1.481 ± 0.184	0.544 ± 0.11	5.394 ± 0.2701	5.563 ± 0.2827
Lead	32.634 ± 0.4501	6.704 ± 0.4952	1.991 ± 0.1885	0.783 ± 0.1283	9.248 ± 0.2846	9.571 ± 0.2819
Random	30.681 ± 0.3746	5.007 ± 0.2447	1.189 ± 0.1584	0.442 ± 0.084	8.088 ± 0.1344	8.426 ± 0.1316
GencopareSum	31.272 ± 1.5372	6.965 ± 0.6824	2.288 ± 0.3661	0.977 ± 0.2448	9.042 ± 0.9351	9.481 ± 0.9372
KLDivergence	30.079 ± 0.6514	6.441 ± 0.4692	1.584 ± 0.3825	0.557 ± 0.2391	8.167 ± 0.3169	8.49 ± 0.3251
Our model	**34.013 ± 0.8079**	**8.266 ± 0.7509**	**2.951 ± 0.4178**	**1.253 ± 0.2422**	**10.366 ± 0.4009**	**10.713 ± 0.3988**

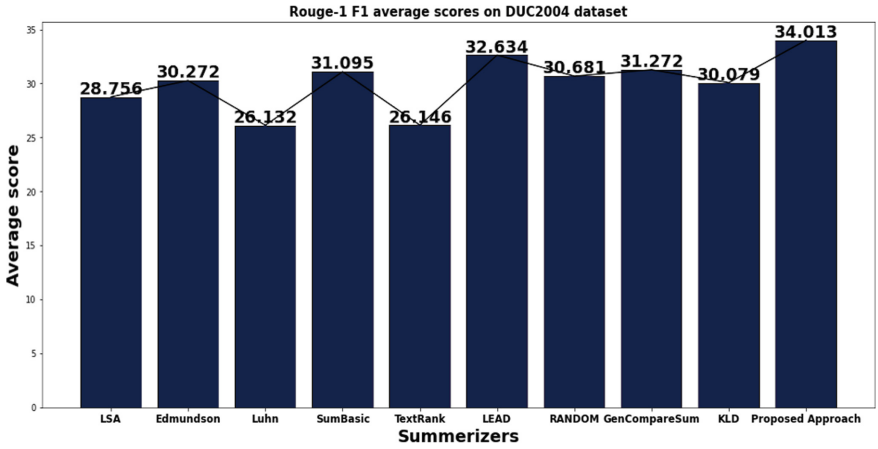


Fig. 2. Comparison between summarization models with ROUGE-1 F1 average score on the DUC2004 dataset

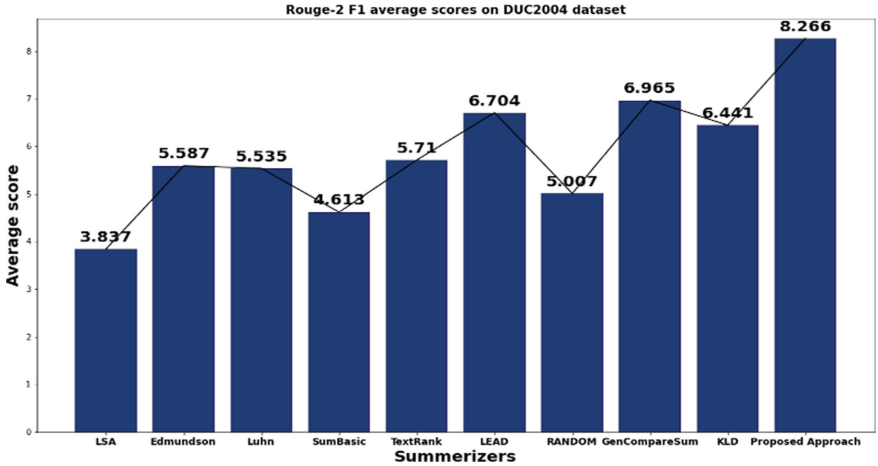


Fig. 3. Comparison between summarization models with ROUGE-2 F1 average score on the DUC2004 dataset

technique is Luhn method in ROUGE-1 F1 average score with 26.132, Latent Semantic Analysis (LSA) method in ROUGE-2 F1 average score with 3.837, LSA method in ROUGE-3 F1 average score with 0.804, LSA method in ROUGE-4 F1 average score with 0.241, Luhn method in ROUGE-S F1 average score with 5.319, Luhn method in ROUGE-SU F1 average score with 5.458. The deviation of ROUGE scores strongly depends on parameters such as the number of clusters selected to obtain salient texts. As shown in Table 2, we note that our model performed best when the number of clusters is 7. Figure 8 is the sample of the output summary from our model and the gold summary from the DUC2004 dataset.

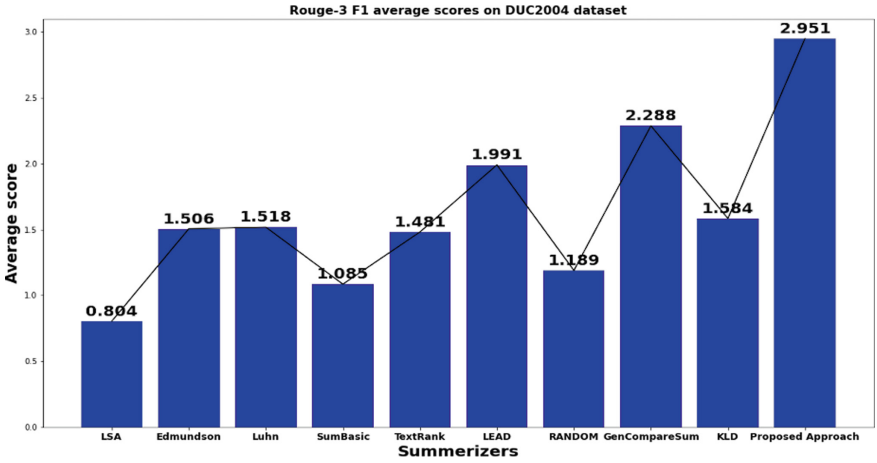


Fig. 4. Comparison between summarization models with ROUGE-3 F1 average score on the DUC2004 dataset

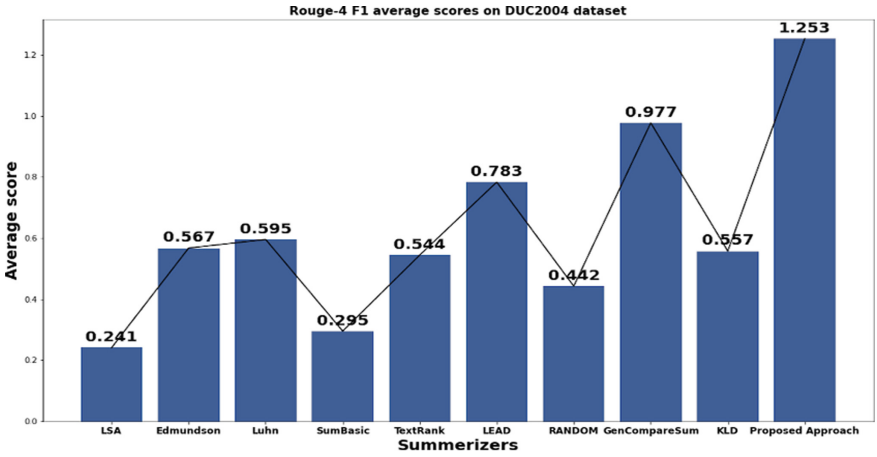


Fig. 5. Comparison between summarization models with ROUGE-4 F1 average score on the DUC2004 dataset

5 Conclusion

This work aims to develop an unsupervised extractive-based text summarization system for multi-documents. The developed system is two-stage, it uses both the T5 pre-trained transformer model and the K-Means clustering technique to get a salient text that helps to retrieve the most relevant sentences that can be represented as summaries for the

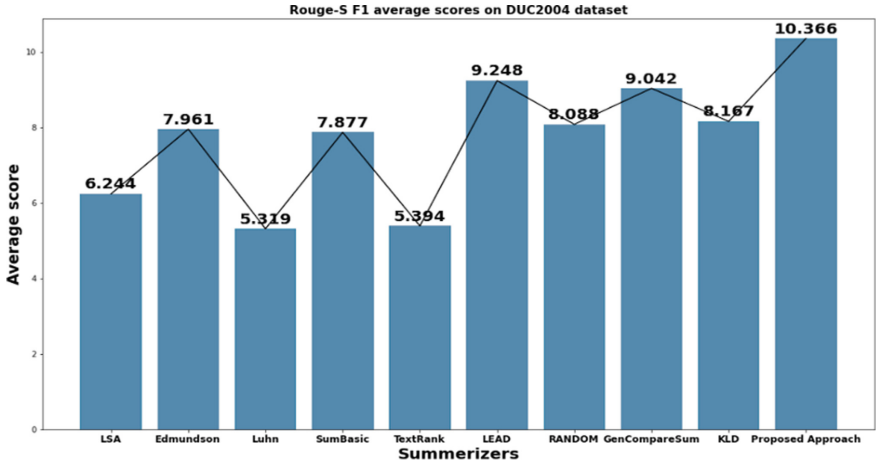


Fig. 6. Comparison between summarization models with ROUGE-S F1 average score on the DUC2004 dataset

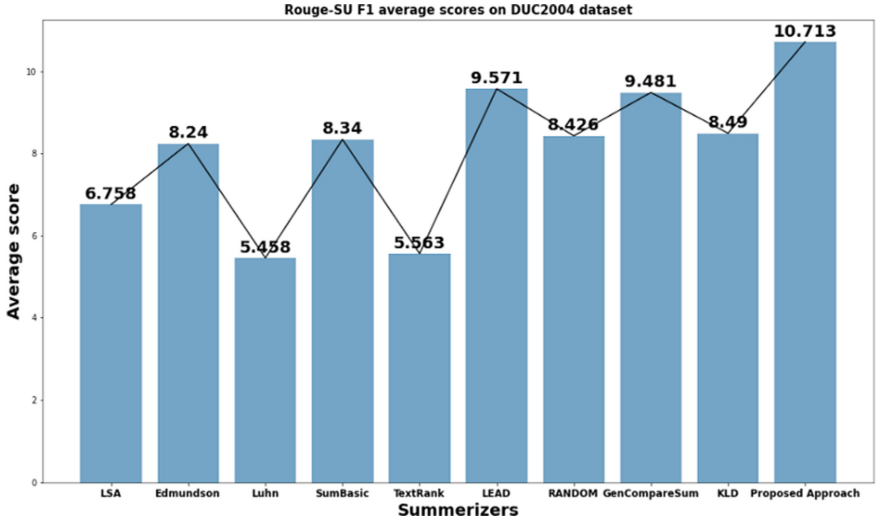
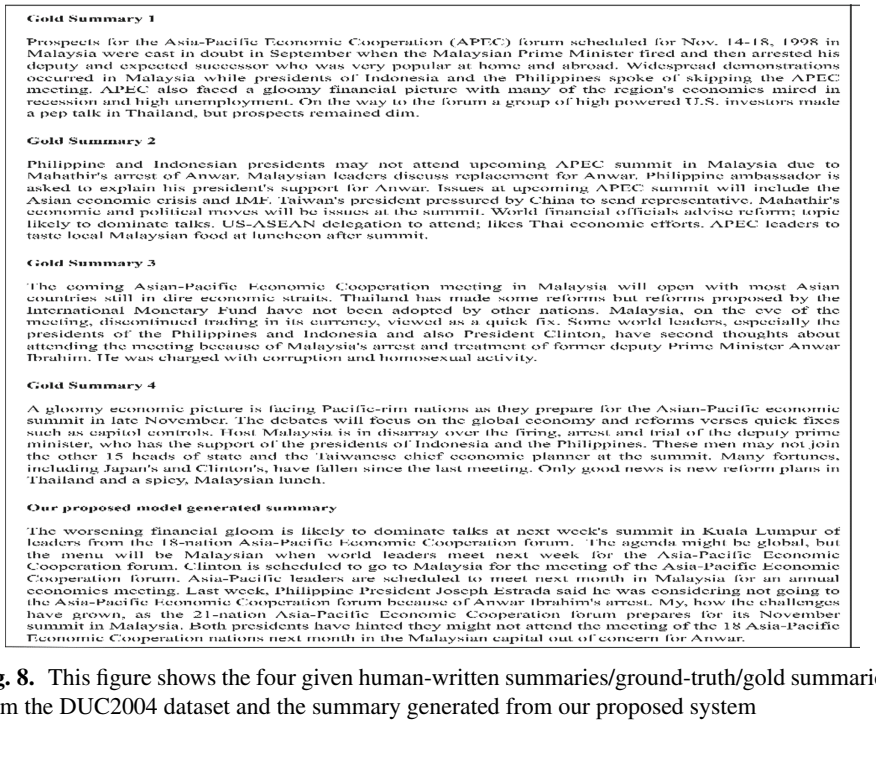


Fig. 7. Comparison between summarization models with ROUGE-SU F1 average score on the DUC2004 dataset

collection of documents. System performance was evaluated using the ROUGE metric on the DUC2004 benchmark dataset, which indicates that our proposed model outperformed all other unsupervised state-of-the-art summarization models. For further improvement, we will be exploring more combinations of pre-trained models and clustering algorithms from the vast literature available.

Table 2. ROUGE F1 average scores with different numbers of clusters on the DUC2004

Number of clusters	Rouge-1	Rouge-2	Rouge-3	Rouge-4	Rouge-S	Rouge-SU
3	33.185 ± 0.9684	7.553 ± 0.9602	2.585 ± 0.5183	1.1 ± 0.3483	10.003 ± 0.4343	10.349 ± 0.4357
5	33.387 ± 0.5468	7.776 ± 0.3528	2.651 ± 0.1398	1.167 ± 0.132	10.066 ± 0.2851	10.49 ± 0.2276
7	**34.013 ± 0.8079**	**8.266 ± 0.7509**	**2.951 ± 0.4178**	**1.253 ± 0.2422**	**10.366 ± 0.4009**	**10.713 ± 0.3988**
9	33.944 ± 0.6066	8.225 ± 0.4657	2.935 ± 0.2674	1.22 ± 0.161	10.284 ± 0.3443	10.625 ± 0.3433
11	33.878 ± 0.5115	8.071 ± 0.2487	2.787 ± 0.3742	1.216 ± 0.1949	10.288 ± 0.3194	10.63 ± 0.3164



Gold Summary 1

Prospects for the Asia-Pacific Economic Cooperation (APEC) forum scheduled for Nov. 14-18, 1998 in Malaysia were cast in doubt in September when the Malaysian Prime Minister fired and then arrested his deputy and expected successor who was very popular at home and abroad. Widespread demonstrations occurred in Malaysia while presidents of Indonesia and the Philippines spoke of skipping the APEC meeting. APEC also faced a gloomy financial picture with many of the region's economies mired in recession and high unemployment. On the way to the forum a group of high powered U.S. investors made a pep talk in Thailand, but prospects remained dim.

Gold Summary 2

Philippine and Indonesian presidents may not attend upcoming APEC summit in Malaysia due to Mahathir's arrest of Anwar. Malaysian leaders discuss replacement for Anwar. Philippine ambassador is asked to explain his president's support for Anwar. Issues at upcoming APEC summit will include the Asian economic crisis and IMF. Taiwan's president pressured by China to send representative. Mahathir's economic and political moves will be issues at the summit. World financial officials advise reform; topic likely to dominate talks. US-ASEAN delegation to attend; likes Thai economic efforts. APEC leaders to taste local Malaysian food at luncheon after summit.

Gold Summary 3

The coming Asian-Pacific Economic Cooperation meeting in Malaysia will open with most Asian countries still in dire economic straits. Thailand has made some reforms but reforms proposed by the International Monetary Fund have not been adopted by other nations. Malaysia, on the eve of the meeting, discontinued trading in its currency, viewed as a quick fix. Some world leaders, especially the presidents of the Philippines and Indonesia and also President Clinton, have second thoughts about attending the meeting because of Malaysia's arrest and treatment of former deputy Prime Minister Anwar Ibrahim. He was charged with corruption and homosexual activity.

Gold Summary 4

A gloomy economic picture is facing Pacific-rim nations as they prepare for the Asian-Pacific economic summit in late November. The debates will focus on the global economy and reforms verses quick fixes such as capitol controls. Host Malaysia is in disarray over the firing, arrest and trial of the deputy prime minister, who has the support of the presidents of Indonesia and the Philippines. These men may not join the other 15 heads of state and the Taiwanese chief economic planner at the summit. Many fortunes, including Japan's and Clinton's, have fallen since the last meeting. Only good news is new reform plans in Thailand and a spicy, Malaysian lunch.

Our proposed model generated summary

The worsening financial gloom is likely to dominate talks at next week's summit in Kuala Lumpur of leaders from the 18-nation Asia-Pacific Economic Cooperation forum. The agenda might be global, but the menu will be Malaysian when world leaders meet next week for the Asia-Pacific Economic Cooperation forum. Clinton is scheduled to go to Malaysia for the meeting of the Asia-Pacific Economic Cooperation forum. Asia-Pacific leaders are scheduled to meet next month in Malaysia for an annual economics meeting. Last week, Philippine President Joseph Estrada said he was considering not going to the Asia-Pacific Economic Cooperation forum because of Anwar Ibrahim's arrest. My, how the challenges have grown, as the 21-nation Asia-Pacific Economic Cooperation forum prepares for its November summit in Malaysia. Both presidents have hinted they might not attend the meeting of the 18 Asia-Pacific Economic Cooperation nations next month in the Malaysian capital out of concern for Anwar.

Fig. 8. This figure shows the four given human-written summaries/ground-truth/gold summaries from the DUC2004 dataset and the summary generated from our proposed system

References

1. Rezaei, A., Dami, S., Daneshjoo, P.: Multi-document extractive text summarization via deep learning approach. In: 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), pp. 680–685. IEEE (2019)
2. Mallick, R., Susan, S., Agrawal, V., Garg, R., Rawal, P.: Context-and sequence-aware convolutional recurrent encoder for neural machine translation. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp. 853–856 (2021)
3. Tsoumou, E.S.L., Lai, L., Yang, S., Varus, M.L.: An extractive multi-document summarization technique based on fuzzy logic approach. In: 2016 International Conference on Network and Information Systems for Computers (ICNISC), pp. 346–351. IEEE (2016)

4. Yapinus, G., Erwin, A., Galinium, M., Muliady, W.: Automatic multi-document summarization for Indonesian documents using hybrid abstractive-extractive summarization technique. In: 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1–5. IEEE (2014)
5. Hirao, T., Fukusima, T., Okumura, M., Nobata, C., Nanba, H.: Corpus and evaluation measures for multiple document summarization with multiple sources. In: Proceedings of the Twentieth International Conference on Computational Linguistics (COLING), pp. 535–541 (2004)
6. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
7. Edmundson, H.P., Wyllys, R.E.: Automatic abstracting and indexing—survey and recommendations. Commun. ACM **4**(5), 226–234 (1961)
8. Mihalcea, R., Tarau, P.: A language independent algorithm for single and multiple document summarization. In: Companion Volume to the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts (2005)
9. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)
10. Ozsoy, M.G., Alpaslan, F.N., Cicekli, I.: Text summarization using latent semantic analysis. J. Inf. Sci. **37**(4), 405–417 (2011)
11. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268 (2017)
12. Nenkova, A., Vanderwende, L.: The impact of frequency on summarization. MSRTR-2005-101 (2005)
13. Bishop, J., Xie, Q., Ananiadou, S.: GenCompareSum: a hybrid unsupervised summarization method using salience. In: Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 220–240 (2022)
14. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
15. Cachola, I., Lo, K., Cohan, A., Weld, D.S.: TLDR: extreme summarization of scientific documents. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4766–4777 (2020)
16. Yu, H.: Summarization for internet news based on clustering algorithm. In: 2009 International Conference on Computational Intelligence and Natural Computing, vol. 1, pp. 34–37. IEEE (2009)
17. Zhao, J., Liu, M., Gao, L., Jin, Y., Du, L., Zhao, H., Zhang, H., Haffari, G.: Summpip: unsupervised multi-document summarization with sentence graph compression. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1949–1952 (2020)
18. Goel, R., Vashisht, S., Dhanda, A., Susan, S.: An empathetic conversational agent with attentional mechanism. In: 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–4. IEEE (2021)
19. Goel, R., Susan, S., Vashisht, S., Dhanda, A.: Emotion-aware transformer encoder for empathetic dialogue generation. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 1–6. IEEE (2021)
20. https://www.kaggle.com/datasets/usmanniazi/duc-2004-dataset
21. Lin, C., Rey, M.: ROUGE: A Package for Automatic Evaluation of Summaries (2001)



GAN Based Restyling of Arabic Handwritten Historical Documents

Mohamed Ali Erromh[1,3], Haïfa Nakouri[2,3(✉)], and Imen Boukhris[1,3]

[1] University of Manouba, National School of Computer Science (ENSI), Manouba, Tunisia
{mohamedali.romh,imen.boukhris}@ensi-uma.tn
[2] University of Manouba, Ecole Supérieure de l'Economie Numérique (ESEN), Manouba, Tunisia
hayfa.nakouri@esen.tn,nakouri.hayfa@gmail.com
[3] Université de Tunis, LARODEC, ISG Tunis, Tunis, Tunisia

Abstract. Arabic handwritten documents consist of unstructured heterogeneous content. The information these documents can provide is very valuable both historically and educationally. However, content extraction from historical documents by Optical Character Recognition remains an open problem given the poor quality in writing. Furthermore, these documents most often show various forms of deterioration (e.g., watermarks). In this paper, we propose a Cycle GAN-based approach to generate a document with a readable font style from a historical Arabic handwritten document using a collection of unlabeled images. We used Arabic OCR for content extraction.

Keywords: Arabic Historical Text · Arabic Optical Character Recognition · Deep Learning · Generative Adversarial Network

1 Introduction

Documentation of knowledge using handwriting is one of the biggest achievements of mankind. Indeed, in the past, handwriting was the unique way of documenting important events and saving data. Accordingly, these historical documents are handwritten texts consisting of unstructured data with heterogeneous content. Indeed, a document can include different font sizes and types, and overlapping text with lines, images, stamps and sketches. Most often, the information that these documents provide is important both historically and educationally. For instance, they could help paleographers in manuscripts dating, classification and authentication; neurologists to detect neurological disorders, graphologists to analyse personality, etc.

Ancient handwritten documents have plenty of valuable information for historians and researchers hence the need of them to be digitized and of their content to be extracted [6]. This is not an easy task given that the quality of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 543–555, 2023.
https://doi.org/10.1007/978-3-031-27409-1_49

historical manuscripts is generally quite poor as the documents degrade over time, they also contain different ancient writing styles in different languages.

Despite machine printed documents are easy to be extracted by optical character recognition (OCR), the recognition of handwritten documents, especially historical ones, is still a scientific challenge due to the poor quality and low resolution of these documents. Besides, obtaining large sets of labeled handwritten documents is often the limiting factor to effectively use supervised Deep Learning methods for the analysis of these image-type documents.

In this context, the Generative Adversarial Network (GAN) can represent a solution. Indeed, the GAN has made a breakthrough and great success in many areas of computer vision research. It uses efficiently large unlabeled datasets for learning. It is based on a generator and a discriminator models. GANs can be used to generate synthesized images as well as to translate images from one domain to another, generate high definition images from low-definition ones, etc.

Most of the works dealing with handwritten documents consider Latin or Germanic languages. Arabic language, though, is slightly different since it exposes some challenges in writing such as the writing direction.

In this work, we propose a GAN-based approach to generate a document with a readable font style from a historical Arabic handwritten document using a collection of unlabeled images. We used Arabic OCR for content extraction.

This paper is organized as follows: Sect. 2 recalls the basic concepts of GANs and introduces their different types. Section 3 is dedicated to related works on handling handwritten documents namely those written in Arabic. Section 4 is devoted to our proposed GAN-based approach to restyle synthetic historical documents in Arabic. Section 5 presents the experiments results on original historical Arabic documents. Section 6 concludes the paper and exposes some future directions.

2 Generative Adversarial Networks

GANs are a class of deep generative models introduced by Goodfellow et al. [7] and have gained wide popularity and interest in many different application areas. A GAN model consists of two networks, namely a generator and a discriminator. The architecture of a GAN model in its original form is illustrated in Fig. 1. The generator typically generates data from initially random patterns. These generated fake observations are fed into the discriminator along with real observations. The discriminator acts as a classifier. It is trained to validate the authenticity of the input data, i.e., to distinguish real data from generated data. The crucial point is that the generator only interacts with the discriminator and has no direct access to real data. The generator thus learns from its failure based on the feedback of the discriminator and improves its performance in generating realistic data through the training process (backpropagation). The two networks contest with each other in a zero-sum game; hence, their goals are adversarial.

To assert that a GAN is successfully trained, the generated data has to fool the discriminator and the generated data samples should be as various as in a real-world data distribution.

There are several types of GANs. Each one has its specificity and its domain applications. In what follows, we cite some of them.

Vanilla GAN is the simplest type. The generator and the discriminator are simple multi-layer perceptrons (MLP). The algorithm tries to optimize a mathematical equation using stochastic gradient descent.

Deep convolutional GAN (DCGAN) [5] takes benefits from convolutional networks and GANs where the MLP is replaced with deep convolutional networks.

Conditional GAN (cGAN) [17] has shown its efficiency for more precise generation and differentiation of images by adding conditional parameters. Indeed, generators and discriminators are conditioned by some auxiliary information from other modalities (class labels, data, etc).

CycleGAN [21] is used for image-to-image translation. It allows transforming an image from one domain to another. It is based on cycle consistency loss to enable training without the need for paired data, meaning that no one-to-one mapping between the source and the target is needed.

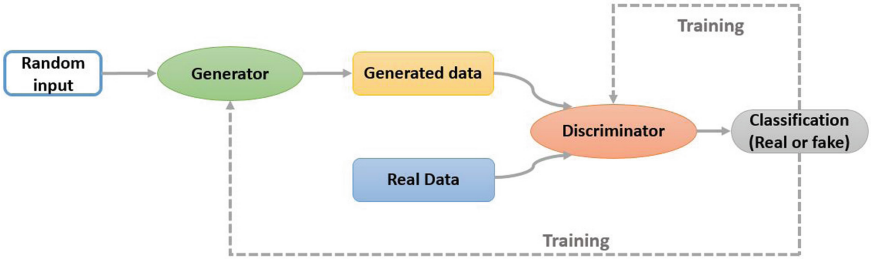


Fig. 1. Functioning of GAN

3 Related Works

Handwriting is an important way of communication through civilizations that has developed and evolved over time. Studying these documents is very important in many fields. For instance, they could help paleographers in manuscripts dating, classification and authentication; neurologists to detect neurological disorders, graphologists to analyse personality, etc.

However, while machine printed documents are easy to be extracted by optical character recognition (OCR) [15], handwritten documents recognition is still a scientific challenge especially for historical ones. Indeed, the access to these documents is reserved to only some experts. Moreover, their writing style is mostly characterized by distortion and pattern variation. Furthermore, these documents most often suffer from various forms of deterioration over time due to their aging and to the lack of preservation (e.g., watermarks, blurry fonts, overlays).

In literature, several methods were proposed to handle handwritten documents such as automatic text segmentation namely the conditional random fields (CAC) approach [16], data extraction namely the 2-phase incremental learning (AI2P) [2], curvelet image reconstruction (RIC) [10] and other works based on the generative adversarial network (GAN) on many languages such as Arabic [4], French [14] and Chinese [22].

For historical handwritten documents, GANs plays an important role in several tasks such as style transfer [21] or document enhancement [20] . Indeed, the Cycle GAN has shown effective results in generating Latin historical documents by providing a general framework that produces realistic historical documents with specific style and textual content/structure [21]. Besides, conditional GAN successfully restores images of severely degraded historical documents written in Croatian language [20]. It ensures significant document quality enhancement such as after watermark removal.

On the other hand, few works related to Arabic historical documents have been proposed. Alghamdi et al. [1] proposed a method for text segmentation of historical Arabic manuscripts using a projection profile. It is based on line and character segmentation based on the projection profile methods. Hassen et al. [8] investigate the recognition of sub-words in historical Arabic Documents using C-GRU which is an end-to-end system for recognizing Arabic handwritten sub-words. Hassen et al. [12] proposed a method for automatic processing of historical Arabic documents by identifying the authors and recognizing some words or parts of the documents from a set reference data.

To the best of our knowledge, no method for historical Arabic documents font restyling was proposed.

4 Proposed Framework

The idea of this work is to propose a method to automatically transcript the content of an ancient Arabic handwritten document using GANs by restyling the original image documents, that are challenging to read, to a more readable font style. Further, Arabic OCR are used for content extraction from these generated documents to evaluate to what extent the integrity of the original content was preserved. As depicted in Fig. 2, our method is based on four steps namely, data collection, data pre-processing, document restyling and content extraction.

4.1 Data Collection

The truth is, access to historical Arabic documents and their collection represents a heavy challenge in this work given their unavailability. Nevthertheless, our work uses basically two datasets: First the RASM 2018/2019 data set [11] which contains a selection of historical Arabic scientific manuscripts (10–19th century) digitized through the British Library Qatar Foundation Partnership. This first data set represents the historical Arabic images documents we aim

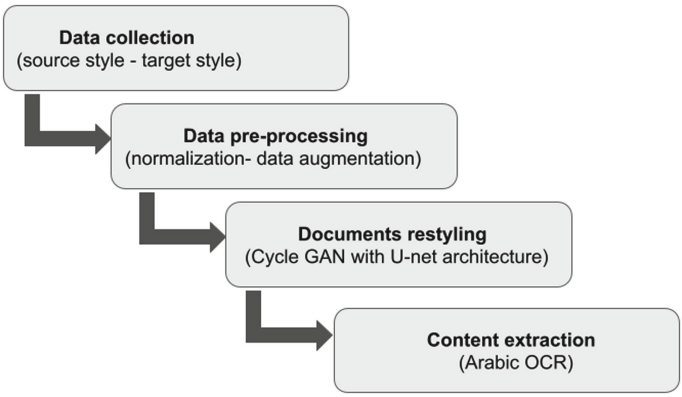


Fig. 2. Steps of the proposed method

to restyle. Second, the Nithar data set[1] which is a manually edited Egyptian dataset containing a diverse collection of cultural, historical and political intellectual essays. This second data set is solely used to spot the target font style to which the historical handwritten source style will be translated. In a nutshell, the exact content (text) of the first data set should be transcripted to the second data set's font style.

Based on these datasets, we split the data into four parts:

- trainA: it contains 70 historical handwritten images in .TIFF extension from RASM 2018/2019. It consists of the source domain dataset. It is a real unlabeled historical handwritten document presenting typical challenges of layout analysis and text recognition encountered in Arabic language as shown in Fig. 3. It contains a lot of margins and diagrams.
- trainB: it contains 70 images from Nithar. It consists of the target domain dataset.
- testA: 20 handwritten historical images from RASM 2018/2019 for the test phase.
- testB: 20 images from Nithar dataset for the test phase.

4.2 Data Pre-processing

In this phase, normalization and data augmentation will be used as data preprocessing methods.

- Normalization: the pixels of an image have intensity values in the range [0, 255] for each channel (red, green, blue) [18]. In order to eliminate this kind of skew on our data, we normalise images to have intensity values in the range [–1,1]. This is done by dividing by the mean of the maximum range (127.5)

[1] https://rashf.com/book/111111344.

and subtracting 1 (image/127.5–1). This function makes the features more consistent with each other and helps to improve prediction's performance. It is used to get the result faster as the machine has to process a smaller range of data.

- Data augmentation: having new training examples from existing data helps our learning models to generalize better. Since the access to historical handwritten documents is difficult, we increase the number of images through a random crop function. Data augmentation [19] is obtained by creating a random subset of the original image.

4.3 Cycle GAN-Based Restyling

Cycle GAN requires weak supervision and does not need paired images to perform style transfer from a source domain to a target domain. Thus, as depicted in Fig. 4, once we have pre-processed historical Arabic handwritten documents, we propose to use Cycle GAN to restyle them. Each neural network is a CNN



Fig. 3. Source domain image

and more precisely a U-net [24]. It consists of the encoder-decoder model with a skip connection between encoder and decoder.

Our proposed Cycle GAN will use a generator network that translates a historical document to a target domain (document written with the new font style) ($A2B$). The generator will take an image as input and outputs a generated more readable image.

The used activation function is ReLU. It produces a larger value if its inputs exceed a threshold. It is a non-linear activation function that is used in multi-layer neural networks.

After that, the discriminator A will distinguish between real or fake images produced by generator $A2B$.

Loss objective We train with a loss objective that consists of four different loss terms. In what follows, we will denote by: $\boldsymbol{A2B}$ to the generator; $\boldsymbol{A}$ to the discriminator; $\boldsymbol{x}$ to the source; $\boldsymbol{y}$ to the target; $\boldsymbol{z}$ to a real image; $\boldsymbol{m}$ to the number of images; $\boldsymbol{E_x}$ $\boldsymbol{p(x)}$ to an example of dataset source and $\boldsymbol{\alpha}(\boldsymbol{x})$ to the optimiser.

- **Identity loss** ($\boldsymbol{L_i}$): This loss term is used to regularize the generator [9]. It works like an identity classification function, given the actual model for each parent domain. The generator $A2B$ becomes free to change the hue between the source and target documents. $L_i(A2B)$is defined in Eq. 1.

$$L_i\left(A2B\right) = E_x \sim p(x)[|A2B_x - x|] \tag{1}$$

- **Cycle loss** (L_c): The cycle loss [13] limits the freedom of the GAN. Without it, there is no guarantee that a learned classification function correctly maps

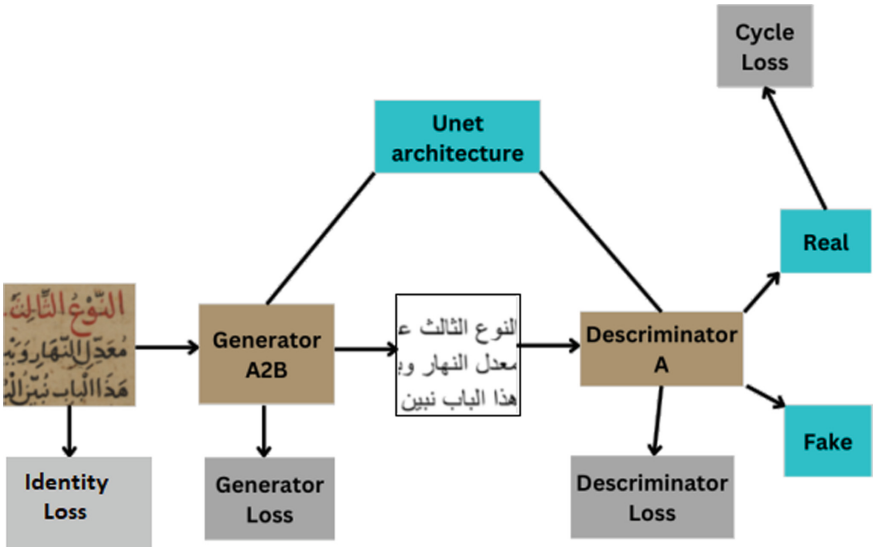


Fig. 4. Simplified view of the proposed Cycle GAN architecture

an individual x to the desired y. In addition, for each pair $x_i y_i \in X$ the Cycle GAN should be able to bring the image x_i back into the original domain X, i.e. $x_i \rightarrow A2B(x_i) \approx x_i$. As the nature of the cycleGAN is bidirectional the reverse mapping must also be fulfilled, i.e. $y \rightarrow A2B(y) \approx y_i$. The Cycle loss is defined in Eq. 2.

$$L_c\,(A2B) = E(a \sim p(a)[|\alpha(x) - A_y| + A2B_y - y] \tag{2}$$

- **Generator loss**: During the generator training, a random noise is sampled. The produced output is handled by the discriminator for classification as real or fake. Generator loss is calculated from the mapping of the discriminator which is rewarded if it succeeds in fooling the discriminator. The generator loss is calculated as shown in Eq. 3.

$$\bigtriangledown \Theta d \frac{1}{m}[log_A(x^i) + log(1 - A(A2B(z^i)))] \tag{3}$$

- **Discriminator loss**: The discriminator classifies both the real data and the fake data from the generator. It penalizes itself for misclassifying a real instance as fake, or a fake instance (created by the generator) as real. The discriminator loss is calculated as shown in Eq. 4.

$$\bigtriangledown \Theta A2B \frac{1}{m} \sum_{i=1}^{m} log(1 - A(A2B(z^i))) \tag{4}$$

Combined Network At this stage, we create a combined network to train the generator model. Here, the discriminator will be non-trainable. To train the generator network, we use cycle consistency loss and identity loss. Cycle consistency suggests that if we restyle a historical image to a new font style image, we should arrive at the original image. To calculate the cycle consistency loss, we first pass the input image x to generator $A2B$. Then, we calculate the loss between the image generated by generator $A2B$ and the input image x. Our goal is the same while taking image y as output to the generator $A2B$.

Results We augment the size of the dataset to 140 images for trainA (source) and for trainB (target) using a *random-crop* function. The augmented data is then used to train the model with a number of epoch=2000 using Adam optimizer [23]. 20 images from the RASM 2018/2019 dataset were used for the test (20 images from testA). We show in Fig. 7, the result of a synthetic image with a new style after 2000 epochs. Accordingly, we notice that the generated image result depends substantially on the number of epochs. Indeed, the bigger the number of epochs, the better the results. Generated images for 10, 500 and 2000 are presented respectively in Figs. 5, 6 and 7.





Fig. 5. Epochs = 10





Fig. 6. Epochs = 500





Fig. 7. Epochs = 2000

4.4 Arabic Data Extraction

To make sure that the original content is maintained intact after the restyling process, we have to extract the content of the generated images.

While many OCR methods have been proposed in literature and have been applied on Latin handwritten documents, Arabic language is still challenging because of its style and writing direction.

In our study, we compared three of the most used OCR methods for Arabic language namely PyMuPDF, Easy OCR, Arabic OCR (AOCR). As shown in Fig. 8, AOCR [3] gives better results. It is indeed a fast method able to better identify words with connected letters. Accordingly, in this work, we consider using AOCR for the extraction step.

5 Experiments

To show the usefulness of GANs in our approach, we applied AOCR directly on historical handwritten documents from the RASM 2018/2019 dataset (see Fig. 9). We compared the results with those found on the same documents generated with our GAN (see Fig. 10). We notice that without the use of GANs, there is no compatibility with the content of the source documents. After using Cycle GAN, results improved noticeably as most of the extracted words are effectively compatible to the original content. We notice that the words that were not preserved after changing the style are mostly crossed out words, overlapped, corrupted, etc.

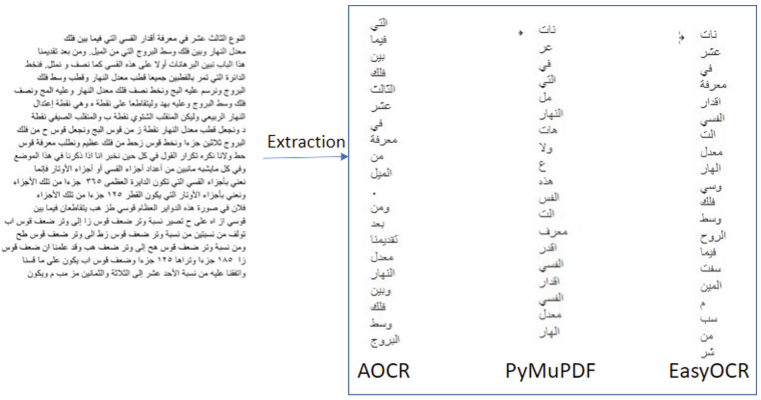


Fig. 8. Comparison between AOCR, PyMuPDF and EasyOCR

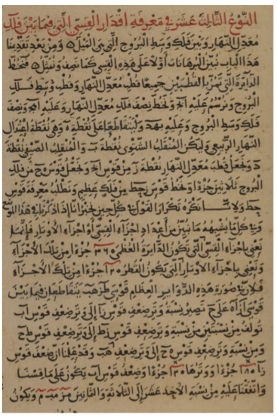


Fig. 9. Arabic OCR result on a historical image



Fig. 10. Arabic OCR result on a generated image

To assess the quality of the generated restyled documents and to ensure that the content integrity is preserved, we use the accuracy performance measure shown in Eq. 5. Actually, extracted words will be compared to those found with AOCR on manually Word-edited documents (Fig. 11) having the same content but with different layout and font style.

Accuracy is based on the distance between words extracted from the manually edited templates M_i and words extracted from the generated restyled documents G_i.

$$Accuracy = [\sum_{i=1}^{n} M_i - \sum_{i=1}^{n} G_i]/\text{Total Number of Words} \tag{5}$$

Table 1 shows the accuracy results of 20 images. As noticed, the accuracy demonstrate promising results. However, while some images gives excellent results (e.g. image 12, an accuracy of 94.03% is achieved), others have less better results (e.g. image 8, an accuracy of only 69.40% is found). We conclude that the quality of data extraction depends on the quality of the original document. In fact, when an image contains overlaps, colored words, margins, circles, etc., as in image 8, this negatively alters and compromises the extraction performance. Even though the extracted words of the generated documents are readable, AOCR still struggles with the extraction of some letters.

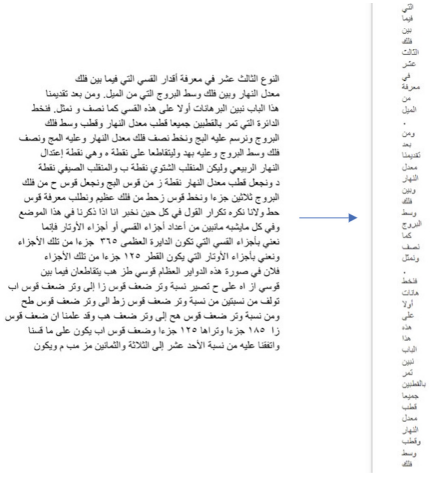


Fig. 11. Arabic OCR results on a word-edited image

With an average accuracy of 82.50%, we may conclude that, overall, the use of the Cycle GAN allows to produce a substantial and faithful style transformation of the historical source document to the target style while preserving the content.

Table 1. Accuracy results

Image 1	Image 2	Image 3	Image 4	Image 5	Image 6	Image 7	Image 8	Image 9	Image 10	Image 11
86.40%	70.53%	79.40%	88.20%	77.23%	74.30%	86.10%	69.40%	86.93%	83.89%	74.90%
Image 12	Image 13	Image 14	Image 15	Image 16	Image 17	Image 18	Image 19	Image 20	Average	
94.03%	89.50%	86.90%	80.30%	84.2%	84.60%	79.10%	85.60%	88.48%	82.50%	

6 Conclusion

Even though the content of Arabic historical handwritten documents is important, its extraction remains challenging. Applying directly OCR methods on

these documents is not interesting given their poor quality: different font sizes and types, text overlapping with lines, containing images, stamps and sketches. The proposed GAN based approach allows to restyle the historical handwritten source domain image to a more readable font style target image. To this end, four steps are proposed namely, data collection, data pre-processing, restyling using Cycle GAN and extraction using Arabic-OCR. The resulting images have a satisfactory quality able to be used for data extraction.

As future works, we plan to extend the framework to perform backwards by transforming any text document to a historical handwritten one. In addition, it will be interesting to consider other sources of Arabic historical data.

References

1. Alghamdi, A., Alluhaybi, D., Almehmadi, D., Alameer, K., Siddeq, S.B., Alsubait, T.: Text segmentation of historical arabic handwritten manuscripts using projection profile. In: 2021 National Computing Colleges Conference (NCCC), pp. 1–6. IEEE (2021)
2. Almaksour, A., Mouchère, H., Anquetil, E.: Apprentissage incrémental et synthèse de données pour la reconnaissance de caractères manuscrits en-ligne. In: Colloque International Francophone sur l'Ecrit et le Document, pp. 55–60. Groupe de Recherche en Communication Ecrite (2008)
3. Doush, I.A., AIKhateeb, F., Gharibeh, A.H.: Yarmouk arabic ocr dataset. In: 2018 8th International Conference on Computer Science and Information Technology (CSIT), pp. 150–154. IEEE (2018)
4. Eltay, M., Zidouri, A., Ahmad, I., Elarian, Y.: Generative adversarial network based adaptive data augmentation for handwritten arabic text recognition. Peer J. Comput. Sci. **8**, e861 (2022)
5. Fang, W., Zhang, F., Sheng, V.S., Ding, Y.: A method for improving cnn-based image recognition using dcgan. Comput., Mater. Contin. **57**(1), 167–178 (2018)
6. Fernández Mota, D., Fornés Bisquerra, A.: Contextual word spotting in historical handwritten documents. Universitat Autònoma de Barcelona (2015)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
8. Hassen, H., Al-Madeed, S., Bouridane, A.: Subword recognition in historical arabic documents using c-grus. TEM J. **10**(4), 1630–1637 (2021)
9. Hsu, C.C., Lin, C.W., Su, W.T., Cheung, G.: Sigan: Siamese generative adversarial network for identity-preserving face hallucination. IEEE Trans. Image Process. **28**(12), 6225–6236 (2019)
10. Joutel, G., Eglin, V., Emptoz, H.: Une nouvelle approche pour indexer les documents manuscrits anciens. In: Colloque International Francophone sur l'Ecrit et le Document. pp. 85–90. Groupe de Recherche en Communication Ecrite (2008)
11. Keinan-Schoonbaert, A., et al.: Ground truth transcriptions for training ocr of historical arabic handwritten texts. [“”] (2019)
12. Khedher, M.I., Jmila, H., El-Yacoubi, M.A.: Automatic processing of historical arabic documents: a comprehensive survey. Pattern Recognit. **100**, 107144 (2020)
13. Lei, Y., Harms, J., Wang, T., Liu, Y., Shu, H.K., Jani, A.B., Curran, W.J., Mao, H., Liu, T., Yang, X.: Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks. Med. Phys. **46**(8), 3565–3581 (2019)

14. Liu, X., Meng, G., Xiang, S., Pan, C.: Handwritten text generation via disentangled representations. IEEE Signal Process Lett. **28**, 1838–1842 (2021)
15. Memon, J., Sami, M., Khan, R.A., Uddin, M.: Handwritten optical character recognition (ocr): a comprehensive systematic literature review (slr). IEEE Access **8**, 142642–142668 (2020)
16. Montreuil, F., Nicolas, S., Heutte, L., Grosicki, E.: Intégration d'informations textuelles de haut niveau en analyse de structures de documents manuscrits non contraints. Document Numerique **14**(2), 77–101 (2011)
17. Pang, Y., Liu, Y.: Conditional generative adversarial networks (cgan) for aircraft trajectory prediction considering weather effects. In: AIAA Scitech 2020 Forum, p. 1853 (2020)
18. Perée, T., et al.: Implémentation d'un système d'imagerie multispectrale adapté au phénotypage de cultures en conditions extérieures et comparaison de deux méthodes de normalisation d'images (2019)
19. Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. **208**, 106236 (2021)
20. Souibgui, M.A., Kessentini, Y.: De-gan: a conditional generative adversarial network for document enhancement. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
21. Vögtlin, L., Drazyk, M., Pondenkandath, V., Alberti, M., Ingold, R.: Generating synthetic handwritten historical documents with ocr constrained gans. In: International Conference on Document Analysis and Recognition, pp. 610–625. Springer (2021)
22. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Adv. Neural Inf. Process. Syst. **32** (2019)
23. Zhang, Z.: Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), pp. 1–2. IEEE (2018)
24. Zhao, X., Yuan, Y., Song, M., Ding, Y., Lin, F., Liang, D., Zhang, D.: Use of unmanned aerial vehicle imagery and deep learning unet to extract rice lodging. Sensors **19**(18), 3859 (2019)



A New Filter Feature Selection Method Based on a Game Theoretic Decision Tree

Mihai Suciu[1(✉)] and Rodica Ioana Lung[2]

[1] Centre for the Study of Complexity and Faculty of Mathematics and Computer Science, Babes-Bolyai University, Cluj Napoca, Romania
mihai.suciu@ubbcluj.ro

[2] Centre for the Study of Complexity, Babes-Bolyai University, Cluj Napoca, Romania
rodica.lung@ubbcluj.ro

Abstract. A game theoretic decision tree is used for feature selection. During the tree induction phase the splitting attribute is chosen based on a game between instances with the same class. The assumption of the approach is that the game theoretic component will indicate the most important features. A measure for the feature importance is computed based on the number and depth of occurrences in the tree. Results are comparable and better in some cases than those reported by a standard random forest approach based also on trees.

1 Introduction

One of the key steps in data analysis is represented by feature selection. Any decision made based on the results of an analysis has to take into account the limitations naturally emerging from the data as well as the methods used to decide which are the features that are actually analysed. While feature selection is compulsory in the context of big data, its benefits can be envisaged also on smaller data-sets for which it represents a first step in the intuitive explanation of the underlying model.

Feature selection methods [5] are generally classified in three groups: filter methods, in which features are selected based on some metric indicating their importance [11], wrapper methods that consider subsets of the set of features evaluated by fitting a classification model [10], and embedded methods that intrinsically perform feature selection during the fitting stage, e.g. decision trees and random forests [17].

Decision trees are also widely used to validate feature selection methods [7]. Various real-world applications use them for testing and validation of filter selection methods. For example in network intrusion detection [15], stock prediction [16], nitrogen prediction in wastewater plant [1], code smell detection [9], are some of the approaches in which decision trees showed significant improvement in performance after feature selection.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 556–565, 2023.
https://doi.org/10.1007/978-3-031-27409-1_50

However, feature selection itself based on decision tree induction is one of the most intuitive approaches to assess the feature importance of a data set. As decision trees are build recursively and, at each node level some attribute(s) that best split node data have to be chosen, it is natural to assess that attributes involved in the splitting process are also important in explaining the data. Nevertheless, most feature selection methods that are based on decision trees ultimately use a form of random forest, i.e. multiple trees inducted on sampled data and attributes, in various forms and for different applications [8,14,17].

In this paper we assume that there is still room to explore in the use of a single decision tree for feature selection, as the performance of any approach naturally depends on the tree induction method. We propose the use of a decision tree that splits data based on a game theoretic approach to compute a feature importance and use it for selection. We compare our approach with a random forest filter selection method on a set of synthetic and real-world data.

2 A Game Based Decision Tree for Feature Selection (G-DTfs)

Consider a data set $(\mathcal{X}, \mathcal{Y})$, with $\mathcal{X} \subset \mathbb{R}^{n \times d}$ and $\mathcal{Y} \subset \{0,1\}^n$, such that each instance $x_i \in \mathcal{X}$ has label $y_i \in \mathcal{Y}$. If $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_d)$, with $\mathbf{X}_j \in \mathbf{R}^n$, we want to find a subset of $\{\mathbf{X}_1, \ldots, \mathbf{X}_d\}$ that best explains labels $\mathcal{Y}$.

In this paper we propose the use of the following game theoretic based decision tree in order to identify features/attributes of X that are most influential in separating the data into the two classes. At each node level, the attribute used to split the data is chosen by simulating the a game between the two classes. The tree is build recursively, top-down, starting with the entire (training) data-set at the root node. The following steps are used to split current node data (X, Y).
Check data First, check if the data in the node has to be split or not. The condition used is: if all instances have the same label, or if X contains only one instance, the node becomes a leaf.

2.1 Game Based Data Split at Node Level

If data (X, Y) in a node needs to be split, an axis parallel hyperplane is computed for each attribute $j = \overline{1,d}$ in the data in the following manner.
The node game Consider the following game $\Gamma(X, Y|j)$ composed of:

- the game has two players, L and R, corresponding to the two sub-nodes, and the two classes, respectively;
- the strategy of each player is to choose a β hyperplane parameter: β_L and β_R, respectively;
- the payoff of each player is computed in the following manner:

$$u_L(\beta_L, \beta_R|j) = -n_0 \sum_{i=1}^{n} (\beta_{1|j} x_{ij} + \beta_{0|j})(1 - y_i),$$

and

$$u_R(\beta_L, \beta_R|j) = n_1 \sum_{i=1}^{n} (\beta_{1|j} x_{ij} + \beta_{0|j}) y_i,$$

where

$$\beta = \frac{1}{2}(\beta_L + \beta_R)$$

and n_0 and n_1 represent number of instances having labels 0 and 1, respectively.

The payoff of the left player sums the coefficients that will be used in the construction of the hyper-plane for all instances having label 0 and minimizes this sum multiplied by their number in order to shift the products to the left of the axis. In a similar manner, the corresponding sum for instances having label 1 is maximized in order to shift them as far as possible from the instances with the other label. Coefficient β used to compute the payoffs is actually a linear combination of the strategies of the two players.

The Nash equilibrium of this game is represented by a β value that combines β_L and β_R in such a manner that none of the players can further shift their sums of products to the left or to the right, respectively, while the other maintains its choice unchanged. The equilibrium of the game can be approximated by using an imitation of the fictitious play [4] procedure.

Approximating the Nash equilibrium
The simplified fictitious play version used here to find a suitable value for β is implemented as follows: for a number of η iterations the best response of each player against the strategy of the other player is computed using some optimization algorithm. As we only aim to approximate β values that split the data in a reasonable manner the search stops after the number of iterations has elapsed. Each iteration the best response to the average of the strategies of the other player in the previous iterations is considered as the fixed one. The procedure is outlined in Algorithm 1.

Algorithm 1 Approximation of Nash equilibrium

Input: X, Y - data to be split by the node; j - attribute evaluated
Output: $X_{L|j}, y_{L|j}, X_{R|j}, y_{R|j}$, and β_j to define the split rule for the node based on attribute j;
Initialize β_L, β_R at random (standard normal distribution)
for η iterations: **do**
  Find $\beta_L = \underset{b}{\operatorname{argmin}}\, u_L(b, \beta_R)$;
  Find $\beta_R = \underset{b}{\operatorname{argmin}}\, u_R(\beta_L, b)$;
end for
$\beta_j = \frac{1}{2}(\beta_L + \beta_R)$
$X_{L|j} = \{x \in X | x_j^T \beta \leq 0\}$, $y_{L|j} = \{y_i \in y | x_i \in X_{L|j}\}$
$X_{R|j} = \{x \in X | x_j^T \beta > 0\}$, $y_R = \{y_i \in y | x_i \in X_{R|j}\}$

Selecting an attribute based on the Nash equilibrium

In order to select the attribute used to split node data the NE for each attribute $j = \overline{1,d}$ are approximated and corresponding sub-node data separated and further evaluated based on entropy gain. Whichever attribute returns the greatest entropy gain is selected for splitting data and the corresponding β_j is used to define the separating hyperplane.

2.2 Assigning Attribute's Importance for Feature Selection

Once the tree has been inducted, the importance of each feature in splitting the data can be considered based on weather the feature is used for splitting and the depth of the node that uses it. For each feature $j \in \{1 \dots d\}$ we denote by

$$\nu_j = \{\nu_{jl}\}_{l \in I_j},$$

the set containing the nodes that split data based on attribute j, with I_j the set of corresponding indexes in the tree, and let $\delta(\nu_{jl})$ be the depth of the node ν_{jl} in the decision tree, with values starting at 1 at the root node. Then the importance $\phi(j)$ of attribute j can be computed as:

$$\phi(j) = \begin{cases} \sum\limits_{l \in I_j} \frac{1}{\delta(\nu_{jl})}, & I_j \neq \emptyset \\ 0, & I_j = \emptyset \end{cases}. \tag{1}$$

Thus, the importance of an attribute depends on the depth of the node that uses it to split data. We assume that attributes that are used early in the induction may be more influential. Also, an attribute that may appear on multiple nodes with a higher depth may be influential and indicator $\phi()$ encompasses also this situation.

3 Numerical Experiments

Numerical experiments are performed on synthetic and real world data sets with various degree of difficulty in order to illustrate the stability of the proposed approach.

3.1 Experimental Set-Up

Data
We generate and use synthetic data sets with various degree of difficulty to test the stability of *G-DTfs*. For reproducibility and control on the generated synthetic data sets we use the `make_classification` function from the

`scikit-learn`[1] Python library [13]. To vary the difficulty of the generated data sets we use different values for the number of instances and number of attributes.

For real world data sets we use the Connectionist Bench (Sonar, Mines vs. Rocks) data set (R1) which has 208 instances and 60 attributes, the Parkinson's Disease Classification data set (R2) which has 756 instances and 754 attributes, and the Musk data set (version 1) (R3) which has 476 instances and 168 attributes. The data sets are taken from the UCI Machine Learning Repository [6].

All data sets used require the binary classification of the data instances and present different degrees of difficulty.

Parameter settings

For the synthetic data sets we use the parameters for `make_classification`: number of instances (250, 500, 1000), number of attributes (50, 100, 150), seed (500), the weight of each label (0.5—the data sets are balanced), and class separator (0.5—there is overlap between the instances of different classes). We create data sets with all combinations of the above parameters.

For G-DTfs we test different parameters: maximum depth of a tree (5, 10, 15), number of iterations for fictitious play (5).

We split each data set, synthetic or real, into $M = 10$ subsets. We report the results of G-DTfs and the compared approach on 10 independent runs on each data set used.

We compare the results of G-DTfs to the features selected by a Random Forest (RF) classifier [3]. For the RF classifier we set the parameters: number of estimators (100), split criterion (*gini* index), maximum depth of each estimator (this parameter takes the same value as G-DTfs maximum depth).

Performance evaluation

In order to evaluate the performance of G-DTfs, the stability indicator, SC, [2,12] is used. The stability indicator is based on the Pearson correlation between results reported on sampled data and indicates if the feature selection method is stable, i.e. how different/similar are the features selected based on different samples from the same data. As this is a desired characteristic of a feature selection method, we use it here to compare results reported by G-DTfs with a standard Random Forest (RF) approach for feature selection [13].

In order to compute the stability measure the data-set is split into M subsets by using resampling, and the feature selection method is applied on each subset, resulting M sets of features, which are represented as vectors Z_i, $i = \overline{1, M}$, with z_{ij} having the value 1 if feature j has been selected on the i^{th} sample, and 0 otherwise. The stability measure averages the correlations between all pairs of feature vectors, i.e.:

$$SC = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} Cor(Z_i, Z_j), \tag{2}$$

[1] Version `1.1.1`.

where $Cor(Z_i, Z_j)$ denotes the linear correlation between Z_i and Z_j. A high correlation indicates that the same features are identified as influential for all samples, while a correlation value close to 0 would indicate randomness in the selection of the features. If the score is used to evaluate feature selection methods, it indicates which one is more stable, with the higher the score, the better.

3.2 Numerical Results

Results are presented as mean and standard deviation of the SC score reported by the two methods for the various parameter settings tested (Tables 1 and 2). Results of a t-test comparing stability scores reported by the two methods accompany the data. For the synthetic datasets we find that in 14 settings G-DTfs are significantly better. Also, in most instances, the t-test is superfluous, as differences indicated by the mean and standard deviation values are obviously significant. This is however true in both ways: whenever RF results are better, the difference is also obviously significant.

The same situation appears in the case of real-world data (Table 2), with the additional notice that increasing the number of considered features appears to decrease the performance of RF and increases that of G-DTfs, in terms of stability. While it is true that a minimum number of features is desired, the behavior of a method when faced with larger numbers should be considered also.

The effect of the size of the feature set k on G-DTfs results is illustrated in Fig. 1 for two synthetic datasets, with 50 and 100 attributes, compared to that of RF. We find higher stability measures for G-DTfs with small tree depth (maximum depth of 3) and also the decreasing trend of RF stability measure. The influence of the tree depth on the same data-sets is illustrated in Fig. 2 for various k values. Results presented on these instances confirm that the stability score does not depend on the size of the tree after a certain threshold, which, for these data-sets, is around 5.

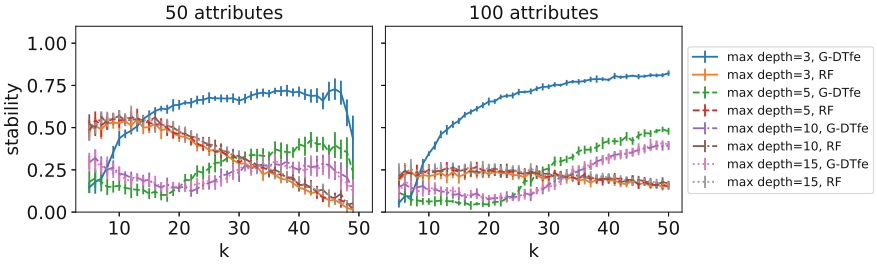


Fig. 1. Effect of parameter k on the stability of feature selection for G-DTfs and RF models with different values for the maximum depth parameter $(3, 5, 10, 15)$ on synthetic data sets with 50 attributes (left) and 150 attributes (right)

Table 1. Results for synthetic data sets, mean ± standard deviation over ten independent runs for the stability indicator for G-DTfs and RF. Data sets with 250 data instances and different number of attributes ($p_1 : 50, 100, 150$), with different maximum depth values ($p_2 : 5, 10, 15, 20$) and different values for the k parameter used in the feature selection procedure ($k : 30, 40$). A ($-$) indicates no significant difference between results, a ($\triangle$) symbol indicates that G-DTfe provides statistically better results and a ($\times$) symbol indicates that RF results are significantly better

p_1	p_2	k	G-DTfs	RF	Significance
50	5	30	0.33(±0.03)	0.33(±0.03)	–
	5	40	0.39(±0.05)	0.18(±0.04)	$\triangle$
	10	30	0.26(±0.04)	0.32(±0.03)	$\times$
	10	40	0.24(±0.04)	0.18(±0.03)	$\triangle$
	15	30	0.26(±0.05)	0.32(±0.03)	$\times$
	15	40	0.26(±0.04)	0.17(±0.03)	$\triangle$
	20	30	0.24(±0.05)	0.33(±0.03)	$\times$
	20	40	0.26(±0.05)	0.17(±0.03)	$\triangle$
100	5	30	0.26(±0.03)	0.22(±0.03)	$\triangle$
	5	40	0.41(±0.02)	0.20(±0.03)	$\triangle$
	10	30	0.16(±0.03)	0.23(±0.03)	$\times$
	10	40	0.32(±0.03)	0.20(±0.03)	$\triangle$
	15	30	0.16(±0.03)	0.23(±0.02)	$\times$
	15	40	0.33(±0.03)	0.19(±0.02)	$\triangle$
	20	30	0.16(±0.03)	0.24(±0.02)	$\times$
	20	40	0.32(±0.03)	0.20(±0.02)	$\triangle$
150	5	30	0.31(±0.03)	0.28(±0.02)	$\triangle$
	5	40	0.45(±0.02)	0.25(±0.02)	$\triangle$
	10	30	0.18(±0.02)	0.27(±0.03)	$\times$
	10	40	0.35(±0.03)	0.25(±0.02)	$\triangle$
	15	30	0.20(±0.03)	0.27(±0.02)	$\times$
	15	40	0.37(±0.02)	0.25(±0.03)	$\triangle$
	20	30	0.18(±0.02)	0.27(±0.03)	$\times$
	20	40	0.36(±0.03)	0.25(±0.02)	$\triangle$

4 Conclusions

The problem of identifying key features that can be used to explain a data characteristic is a central one in machine learning. Similar to other machine learning tasks, efficiency and simplicity are desired from practical approaches. In this paper a decision tree is used to assign an importance measure to features that can be used for their filtering. The novelty of the approach consists in using a game theoretic splitting mechanism for node data during the tree induction.

Table 2. Results for real world data sets, mean and standard deviation over ten independent runs for the stability indicator for G-DTfe and RF. Different real world data sets (R1-R3) for the G-DTfs and RF feature selection models with different maximum depth values (p_2 - $5, 10$) and different values for the k parameter used in the feature selection procedure. A ($-$) shows there is no statistical difference between the tested models, a ($\triangle$) symbol shows that G-DTfs provides statistically better results and a ($\times$) symbol indicates that RF results are significantly better

Data-set	p_2	k	G-DTfs	RF	Significance
R1	5	30	0.41($\pm$0.03)	0.39($\pm$0.04)	$\triangle$
	5	40	0.48($\pm$0.02)	0.31($\pm$0.02)	$\triangle$
	10	30	0.35($\pm$0.04)	0.40($\pm$0.03)	$\times$
	10	40	0.44($\pm$0.03)	0.31($\pm$0.03)	$\triangle$
R2	5	30	0.30($\pm$0.03)	0.42($\pm$0.03)	$\times$
	5	40	0.48($\pm$0.02)	0.42($\pm$0.02)	$\triangle$
	5	100	0.80($\pm$0.01)	0.35($\pm$0.01)	$\triangle$
	5	150	0.86($\pm$0.01)	0.31($\pm$0.02)	$\triangle$
	5	200	0.89($\pm$0.01)	0.29($\pm$0.01)	$\triangle$
	10	30	0.08($\pm$0.01)	0.43($\pm$0.03)	$\times$
	10	40	0.10($\pm$0.01)	0.43($\pm$0.03)	$\times$
	10	100	0.61($\pm$0.01)	0.35($\pm$0.01)	$\triangle$
	10	150	0.73($\pm$0.01)	0.31($\pm$0.01)	$\triangle$
	10	200	0.79($\pm$0.01)	0.28($\pm$0.01)	$\triangle$
R3	5	30	0.34($\pm$0.02)	0.45($\pm$0.02)	$\times$
	5	40	0.52($\pm$0.02)	0.45($\pm$0.02)	$\triangle$
	5	100	0.77($\pm$0.02)	0.30($\pm$0.02)	$\triangle$
	5	150	0.78($\pm$0.02)	0.14($\pm$0.02)	$\triangle$
	10	30	0.12($\pm$0.02)	0.46($\pm$0.01)	$\times$
	10	40	0.17($\pm$0.02)	0.45($\pm$0.02)	$\times$
	10	100	0.57($\pm$0.03)	0.33($\pm$0.02)	$\triangle$
	10	150	0.51($\pm$0.02)	0.17($\pm$0.03)	$\triangle$

The importance of a feature is assigned based on the position of the node(s) that is used for splitting data. While using a single decision tree yielded results comparable and even better than a standard random forest approach, an open research direction consists in exploring a forest of game theoretic based decision trees for feature selection.

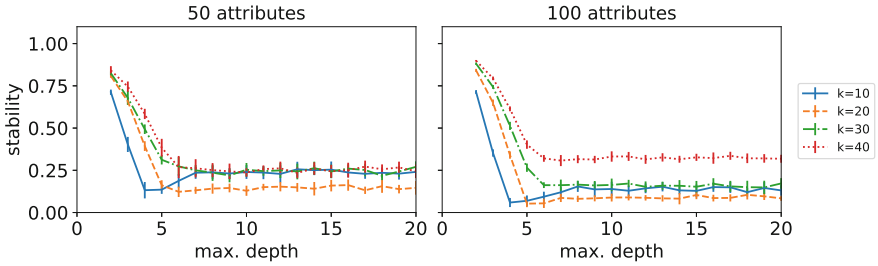


Fig. 2. Effect of parameter maximum depth on the stability of feature selection for G-DTfs on synthetic data sets with 50 attributes (left) and 150 attributes (right) and different values for parameter k $(10, 20, 30, 40)$

Acknowledgments. This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project number PN-III-P1-1.1-TE-2021-1374, within PNCDI III

References

1. Bagherzadeh, F., Mehrani, M.J., Basirifard, M., Roostaei, J.: Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. J. Water Process. Eng. **41**, 102,033 (2021)
2. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M.: Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. Data Anal. **143**, 106,839 (2020)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Brown, G.W.: Iterative solution of games by fictitious play. Act. Anal. Prod. Alloc. **13**(1), 374–376 (1951)
5. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
6. Dua, D., Graff, C.: UCI Machine Learning Repository (2017)
7. Hoque, N., Singh, M., Bhattacharyya, D.K.: EFS-MI: an ensemble feature selection method for classification. Complex Intell. Syst. **4**(2), 105–118 (2018)
8. Huljanah, M., Rustam, Z., Utama, S., Siswantining, T.: Feature selection using random forest classifier for predicting prostate cancer. IOP Conf. Ser.: Mater. Sci. Eng. **546**(5), 052,031 (2019). IOP Publishing
9. Jain, S., Saha, A.: Rank-based univariate feature selection methods on machine learning classifiers for code smell detection. Evol. Intell. **15**(1), 609–638 (2022)
10. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1), 273–324 (1997)
11. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A.: A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(4), 1106–1119 (2012)
12. Nogueira, S., Brown, G.: Measuring the stability of feature selection. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) Machine Learning and Knowledge

Discovery in Databases, pp. 442–457. Springer International Publishing, Cham (2016)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
14. Saraswat, M., Arya, K.V.: Feature selection and classification of leukocytes using random forest. Med. Biol. Eng. Comput. **52**(12), 1041–1052 (2014). https://doi.org/10.1007/s11517-014-1200-8
15. Sheen, S., Rajesh, R.: Network intrusion detection using feature selection and decision tree classifier. In: TENCON 2008–2008 IEEE Region 10 Conference, pp. 1–4 (2008)
16. Tsai, C.F., Hsiao, Y.C.: Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis. Support Syst. **50**(1), 258–269 (2010)
17. Wang, S., Tang, J., Liu, H.: Embedded unsupervised feature selection. Proc. AAAI Conf. Artif. Intell. **29**(1) (2015)



Erasable-Itemset Mining for Sequential Product Databases

Tzung-Pei Hong[1,2(✉)], Yi-Li Chen[2], Wei-Ming Huang[3], and Yu-Chuan Tsai[4]

[1] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan
tphong@nuk.edu.tw
[2] Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan
[3] Department of Electrical and Control, China Steel Inc., Kaohsiung, Taiwan
[4] Library and Information Center, National University of Kaohsiung, Kaohsiung, Taiwan
yjtasi@nuk.edu.tw

Abstract. Erasable-itemset mining has become a popular research topic and is usually used for product production planning in the industry. If some products in a factory may be removed without critically affecting production profits, the set composed of them is called an erasable itemset. Erasable-itemset mining is to find all the removable material sets for saving funds. This paper extends the concept of erasable itemsets to consider customer behavior with a sequence of orders. We consider the scenario that when an item (material) is not purchased, a product using that material cannot be manufactured, and clients will cancel all their orders if at least one such order exists. We propose a modified erasable-itemset mining algorithm for solving the above problem. Finally, experiments with varying thresholds are conducted to evaluate the execution time and mining results of the proposed algorithm considering customer behavior.

Keywords: Customer Behavior · Data Mining · Downward Closure · Erasable Itemset Mining · Sequential Product Database

1 Introduction

Data mining techniques are used in various databases, such as transactional, time-series [14], relational, and multimedia [15]. The techniques include association-rule mining [1, 2, 13], sequential-pattern mining [6, 22], utility mining [18, 19], classification [4], clustering, and so on. Association-rule mining is the most well-known concept for finding interesting knowledge patterns. Several approaches were designed for it. Among them, the Apriori algorithm [1, 2] and the FP-tree [13] are two commonly used to extract hidden patterns from a transactional database. Their purpose is to find frequent-item combinations and derive association rules using the frequent ones. The mining process uses two user-defined thresholds, minimum support and minimum confidence.

Compared with frequent-itemset mining in transaction databases, erasable-itemset mining is usually used in factory production planning. Along with economic issues,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 566–574, 2023.
https://doi.org/10.1007/978-3-031-27409-1_51

managers in a factory may need to consider the trade-off of sale profits, material costs, and fund flow. That is, they focus on the maximum utility of each material when products are manufactured in schedule planning. Finding the material combinations with low profits forms an important issue, and we call these combinations erasable itemsets. In 2009, Deng et al. [7] proposed the erasable mining problem used to analyze the production plan. It used a user-defined threshold, called gain ratio, to decide which material combinations are regarded as erasable itemsets.

In some applications, the order of events is important, such as the order of treatments in medical sequences in hospitals and sequences of items purchased by customers in retail stores. The above cases may have different meanings for distinct orders. Neither the Apriori nor the FP-tree method mentioned above considers the sequential relationship between events or elements. To solve this problem, Agrawal and Srikant proposed the task of sequential-pattern mining [3] to manage this issue. It is an eminent solution for analyzing sequential data.

Similarly, the sequence in which a customer gives invoices affects the sale benefits if some materials are erased. This paper thus defines a new erasable-itemset mining problem, which considers customer order sequences in product databases. For example, computer components such as hard drives, graphics cards, memory, etc., are used to compose a whole personal computer in a 3C market. When one of the orders for a customer is canceled because this order contains one material in an erased itemset, it will cause the customer to cancel the rest of the orders. It is because when a component is in shortage, the personal PC cannot be finished. Therefore, we propose an approach to find erasable itemsets from sequential product databases considering this scenario. The property of downward closure is adopted in the proposed method to increase the mining performance.

2 Related Works

Many researchers have been devoted to discovering hidden patterns from a transactional database. Agrawal and Srikant [2] firstly proposed the concept of frequent-pattern mining. They also designed a level-by-level approach, called the Apriori algorithm, to find frequent patterns in transactional databases. It processed a database multiple times to perform many checks during mining. After that, the FP-Growth method was introduced to overcome the disadvantage of mining efficiency [13]. It used a tree structure, called FP-tree, to store frequent items and their frequencies in the database to reduce the number of database scans. By traversing the FP tree, unnecessary candidate itemsets could not be generated.

The concept of erasable itemset mining was then introduced by Deng et al. to find the less profitable materials from production planning in a factory production plan [7]. They also designed a method called *META* to solve this problem. It first defined the gain ratio as the evaluation threshold and then checked derived candidate itemsets to determine whether their gain ratios were less than the user-defined threshold. If a candidate itemset satisfies the condition, it is regarded as an erasable itemset. In recent years, many algorithms for solving the erasable-itemset mining problem have been developed, such as the *MERIT* [8], the *MERIT+* [17], the *dMERIT+* [17], the *MEI* [16], the *MEIC* [21]

and the *BREM* [12] algorithms. They were introduced to improve the performance efficiency for this problem. Besides, with inserting new data along with time, the knowledge obtained from the old database is no longer applicable. Hong et al. thus proposed an incremental mining algorithm for erasable itemsets [9], which was similar to the *FUP* algorithm [5] for association rules. Then, the ε-quasi-erasable-itemset mining algorithm [10] was introduced to utilize the concept of the pre-large itemset [11] to enhance the performance efficiency in the incremental erasable-itemset mining process.

In the past, Agrawal and Srikant proposed sequential pattern mining [3] to find the frequent subsequences from sequence databases applied in telecommunication, customer shopping sequences, DNA or gene structures, etc. In this paper, we define a new erasable-itemset mining problem for product-sequence databases and design an algorithm based on the *META* algorithm to solve it.

3 Problem Definition

Table 1 is an example of a product-order database from a manufacturer with a client identifier and order time. Here, assume that each order contains only one product, and a customer can order more than once. The items represent the materials used to produce a product, and the profit is the earning from producing a product.

Table 1. An example of a product-order database

OID	Order time	CID	PID	Items	Profit
o_1	2022/07/17 09:02	c_1	p_1	{*A, B*}	20
o_2	2022/07/19 08:11	c_2	p_2	{*A, C*}	30
o_3	2022/08/11 12:13	c_3	p_3	{*B, C*}	20
o_4	2022/08/20 13:15	c_1	p_4	{*A, C*}	30
o_5	2022/08/21 08:07	c_3	p_5	{*A, C, F*}	80
o_6	2022/12/11 19:11	c_2	p_6	{*A, E*}	50
o_7	2022/12/13 07:00	c_2	p_7	{*A, D*}	70

The product-order database can be converted into a product-sequence database *S* according to clients and order time. The orders with the same client are sequentially listed and shown in Table 2.

Each sequence denotes one or more material sets from the orders of the same client. For example, $<\{A, B\}, \{A, C\}>$ represents the material items used to produce two products. Now, the *Profit* field is the sum of the profits of the products in a sequence, different from the original product database.

Some related definitions are described below for the presented erasable-itemset mining for a sequential product database.

Table 2. An example of a product-sequence database

SID	Item-set sequence	Profit
s_1	$<\{A, B\}, \{A, C\}>$	50
s_2	$<\{A, C\}, \{A, E\}, \{A, D\}>$	150
s_3	$<\{B, C\}, \{A, C, F\}>$	100

Definition 1 Let I_j be the union of the itemsets appearing in a sequence s_j. The gain of an itemset X, denoted $gain(X)$, is defined as follows:

$$gain(X) = \sum_{\{s_j|(X \cap I_j) \neq \varnothing\}} s_j.profit.$$

Take the 1-itemset $\{B\}$ in Table 2 as an example. The item $\{B\}$ appears in s_1 and s_3, and its gain is $50 + 100$, which is 150. Take the 2-itemset $\{AB\}$ as another example. Its contained item $\{A\}$ or $\{B\}$ exists in the three sequences, s_1, s_2, and s_3, respectively. Thus, its gain is $50 + 150 + 100$, which is 300.

Definition 2 The gain ratio of an itemset X, denoted $gain_ratio(X)$, is defined as follows:

$$gain_ratio(X) = \frac{gain(X)}{total_gain(S)},$$

where *total_gain(S)* is the sum of the profits of all the sequences in the given product sequence database S.

Take the itemset $\{B\}$ in Table 2 as an example. The total gain in Table 2 is calculated as $50 + 150 + 100$, which is 300. From the above derivation, $gain(B) = 150$. Thus, the gain ratio of $\{B\}$ is 150/300, which is 0.5.

Definition 3 An erasable itemset X for a product sequence database S is an itemset with its gain ratio less than or equal to a given maximum gain-ratio threshold.

Take the above $\{B\}$, $\{D\}$ and $\{BD\}$ as examples. Their gain ratios are 0.5, 0.5, and 1, respectively. Assume the user-specified maximum gain-ratio threshold λ is set at 0.6. Then $\{B\}$ and $\{D\}$ are erasable itemsets, but $\{BD\}$ is not.

We may use the downward-closure property to solve the problem efficiently. Below, we formally derive some theorems about the property for our proposed problem.

Theorem 1 *Let X and Y be two itemsets. If Y is a superset of X ($X \subseteq Y$), then $gain(X) \leq gain(Y)$.*

Proof Since $X \subseteq Y$, $X \cap I_j \subseteq Y \cap I_j$ for each j. Thus, $\{s_j|(X \cap I_j) \neq \emptyset\} \subseteq \{s_j|(Y \cap I_j) \neq \emptyset\}$. According to Definition 2, we can derive the following:

$$\sum_{\{s_j|(X \cap I_j) \neq \varnothing\}} s_j.profit \leq \sum_{\{s_j|(Y \cap I_j) \neq \varnothing\}} s_j.profit.$$

This means $gain(X) \leq gain(Y)$.

Theorem 2 *Let X and Y be two itemsets. If Y is a superset of X ($X \subseteq Y$) and X is not erasable in this problem, then Y is not erasable.*

Proof If X is not erasable, then $gain(X) > total_gain(D) * \lambda$. According to Theorem 1, when Y is a superset of X, we have $gain\ (Y) \geq gain(X)$. From the two inequalities above, we know $gain(Y) \geq gain(X) > total_gain(D) * \lambda$. Thus, Y is not erasable.

Theorem 3 *Let X and Y be two itemsets. If X is a subset of Y ($X \subseteq Y$) and the itemset Y is erasable in this problem, X must also be erasable.*

Proof If Y is erasable, then $gain(Y) \leq total_gain(D)*\lambda$. According to Theorem 1, when X is a subset of Y, we have $gain(X) \leq gain(Y)$. From the two inequalities above, we know $gain(X) \leq gain(Y) \leq total_gain(D)*\lambda$. Thus, X is erasable.

4 The Proposed Algorithm

We propose an algorithm to solve the above mining problem. It is described as follows.

The erasable-itemset mining algorithm for a product database with customers' orders

Input: A product-order database D and a maximum gain-ratio threshold λ.

Output: A set of all erasable itemsets E.

Step 1: Convert the product order database D to the corresponding sequence database S, with the profit of each sequence in S being the sum of the profits of the product orders in that sequence.

Step 2: Initially, set the variable j to 1, which records the number of items in the currently processed itemset.

Step 3: Set the candidate 1-itemsets as the items appearing in the product order database.

Step 4: Let C_j denote all the candidate j-itemsets.

Step 5: Calculate the gain of each j-itemset in C_j, which contains j-items according to the following formula:

$$gain(X) = \sum_{\{s_j | (X \cap I_j) \neq \varnothing\}} s_j.profit.$$

Step 6: For each j-itemset in C_j, if its gain ratio is less than or equal to λ, place the j-itemset in E_j, which contains all the erasable j-itemsets.

Step 7: Use E_j to generate all candidate $(j + 1)$-itemsets through the join operator, where all the j-subsets of any $(j + 1)$-itemset must exist in E_j.

Step 8: If C_{j+1} is empty, do the next step; Otherwise, set $j = j + 1$ and go to Step 5.

Step 9: Output the union of E_1 to E_j as the final mining result.

Table 3. The parameters of the datasets

Parameter	Description
C	The average number of orders for each customer
T	The number of distinct materials in the dataset
D	The total number of customers in the dataset
r	The maximum gain-ratio threshold

5 Experiments

To evaluate the performance of the proposed method, we used the IBM generator [23] to generate sequential test datasets with designated parameters. The parameter depiction of the datasets is shown in Table 3. Each itemset generated is regarded as a product with its profit randomly generated from 50 to 500.

Varying r thresholds used in a fixed dataset with $C(10)$, $T(25)$, and $D(50K)$ were applied to evaluate the execution time and mining results of the proposed algorithm. The test datasets are listed in Table 4.

Table 4. The datasets used to analyze the influence of r on the algorithm for this problem

Dataset	$\|C\|$	$\|T\|$	$\|D\|$	r (%)
C10T25D50K	10	25	50,000	4
C10T25D50K	10	25	50,000	8
C10T25D50K	10	25	50,000	12
C10T25D50K	10	25	50,000	16
C10T25D50K	10	25	50,000	20
C10T25D50K	10	25	50,000	24
C10T25D50K	10	25	50,000	28

The program is written in Java 12.0.2 and executed on an Intel Core i5-7400M machine with a 3.00 GHz CPU and 16 GB RAM. The running times for the proposed algorithm on distinct thresholds are shown in Fig. 1. Besides, Fig. 2 reveals the numbers of derived candidates and mined erasable ones on different thresholds. From the results shown in Figs. 1 and 2, with the threshold value increased, the proposed method could derive more candidates and erasable itemsets. Thus, the execution time increases along with raising the threshold.

We also compared our results with those from the original definition of erasable-itemset mining, where each product instead of each customer is considered in calculating gain values. The mining results obtained by the META algorithm are shown in Fig. 3. Comparing the results in Figs. 2 and 3, the proposed mining problem is more strict

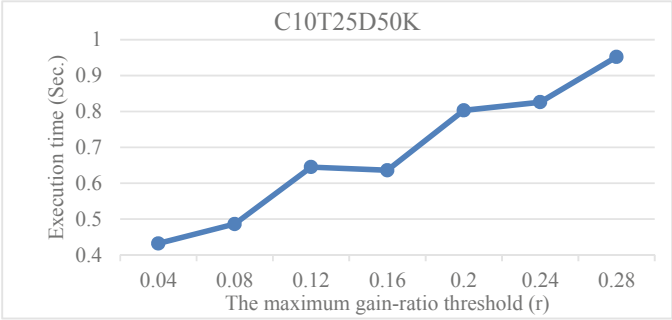


Fig. 1. Runtime for datasets in Table 4 by the proposed algorithm

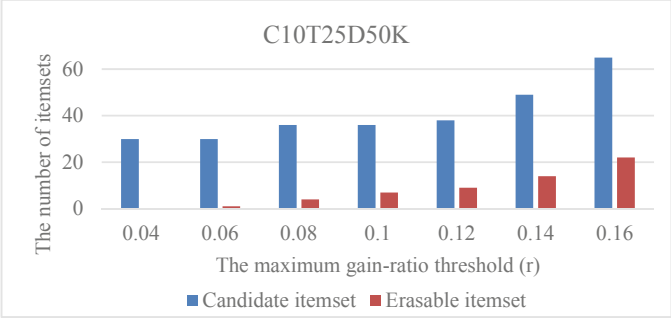


Fig. 2. Mining results for datasets in Table 4 by the proposed algorithm

than the original one and can get fewer but more relevant erasable itemsets for product sequence databases.

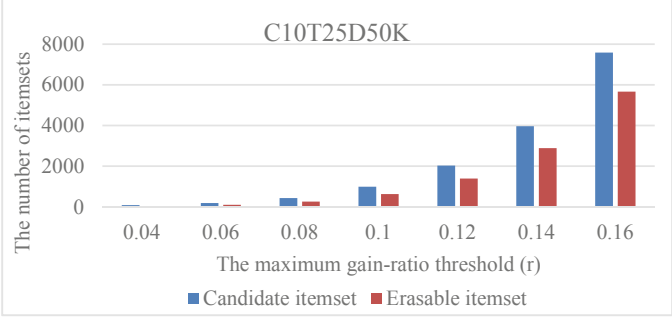


Fig. 3. Mining results for datasets in Table 4 by the META algorithm

6 Conclusions and Future Work

This paper defines the erasable-itemset mining problem for product sequence databases. We propose a method with erasable mining to consider customer behavior with a sequence of orders. When one of the orders from a customer is canceled due to the erased set, all the orders of the customer will be withdrawn. The downward-closure property for this new mining problem is also derived and used in the proposed algorithm to save execution time. The experimental results reveal that the synthetic databases with different parameters affect the execution time significantly. In the future, we will use the tree structure to improve the performance further. Besides, we will run more experiments for datasets with different parameters.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: The 27th ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: The 20th Very Large Data Bases Conference, pp. 487–499 (1994)
3. Agrawal, R., Srikant, R.: Mining sequential patterns. In: The 11th International Conference on Data Engineering, pp. 3–14 (1995)
4. Athira, S., Poojitha, K., Prathibhamol, C.: An efficient solution for multi-label classification problem using apriori algorithm (MLC-A). In: The 6th International Conference on Advances in Computing, Communications and Informatics, pp. 14–18 (2017)
5. Cheung, D.W., Han, J., Ng, V.T., Wong, C.Y.: Maintenance of discovered association rules in large databases: an incremental updating technique. In: The 12th International Conference on Data Engineering, pp. 106–114 (1996)
6. D'andreagiovanni, M., Baiardi, F., Lipilini, J., Ruggieri, S., Tonelli, F.: Sequential pattern mining for ICT risk assessment and management. J. Log. Algebr. Methods Program. **102**, 1–16 (2019)
7. Deng, Z.H., Fang, G.D., Wang, Z.H., Xu, X.R.: Mining erasable itemsets. In: The 8th International Conference on Machine Learning and Cybernetics, pp. 67–73 (2009)
8. Deng, Z.H., Xu, X.R.: Fast mining erasable itemsets using NC_sets. Expert Syst. Appl. **39**(4), 4453–4463 (2012)
9. Hong, T.P., Lin, K.Y., Lin, C.W., Vo, B.: An incremental mining algorithm for erasable itemsets. In: The 15th IEEE International Conference on Innovations in Intelligent Systems and Applications (2017)
10. Hong, T.P., Chen, L.H., Wang, S.L., Lin, C.W., Vo, B.: Quasi-erasable itemset mining. In: The 5th IEEE International Conference on Big Data, pp. 1816–1820 (2017)
11. Hong, T.P., Wang, C.Y., Tao, Y.H.: A new incremental data mining algorithm using pre-large itemsets. Intell. Data Anal. **5**(2), 111–129 (2001)
12. Hong, T.P., Huang, W.M., Lan, G.C., Chiang, M.C., Lin, C.W.: A bitmap approach for mining erasable itemsets. IEEE Access **9**, 106029–106038 (2021)
13. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. ACM SIGMOD Rec. **29**(2), 1–12 (2000)
14. Huang, C.F., Chen, Y.C., Chen, A.P.: An association mining method for time series and its application in the stock prices of TFT-LCD industry. In: The 4th Industrial Conference on Data Mining, pp. 117–126 (2004)

15. Kundu, S., Bhar, A., Chatterjee, S., Bhattacharyya, S.: Multimedia data mining and its relevance today—an overview. Int. J. Res. Eng., Sci. Manag. **2**(5), 994–998 (2019)
16. Le, T., Vo, B.: MEI: an efficient algorithm for mining erasable itemsets. Eng. Appl. Artif. Intell. **27**, 155–166 (2014)
17. Le, T., Vo, B., Coenen, F.: An efficient algorithm for mining erasable itemsets using the difference of NC-Sets. In: The 43th IEEE International Conference on Systems, Man, and Cybernetics, pp. 2270–2274 (2013)
18. Nawaz, M.S., Fournier-Viger, P., Song, W., Lin, J.C.W., Noack, B.: Investigating crossover operators in genetic algorithms for high-utility itemset mining. In: The 13th Asian Conference on Intelligent Information and Database Systems, pp. 16–28 (2021)
19. Singh, K., Singh, S.S., Kumar, A., Biswas, B.: TKEH: an efficient algorithm for mining top-k high utility itemsets. Appl. Intell. **49**(3), 1078–1097 (2018). https://doi.org/10.1007/s10489-018-1316-x
20. Srikant, R., Agrawal, R.: Mining sequential patterns: generalizations and performance improvements. In: The 5th International Conference on Extending Database Technology, pp. 1–17 (1996)
21. Vo, B., Le, T., Pedrycz, W., Nguyen, G., Baik, S.W.: Mining erasable itemsets with subset and superset itemset constraints. Expert Syst. Appl. **69**, 50–61 (2017)
22. Wang, X., Wang, F., Yan, S., Liu, Z.: Application of sequential pattern mining algorithm in commodity management. J. Electron. Commer. Organ. **16**(3), 94–106 (2018)
23. IBM Quest Data Mining Projection: Quest synthetic data generation code. http://www.almaden.ibm.com/cs/quest/syndata.htm (1996)



A Model for Making Dynamic Collective Decisions in Emergency Evacuation Tasks in Fuzzy Conditions

Vladislav I. Danilchenko(✉) and Viktor M. Kureychik

South Federal University, Taganrog, Russia
{vdanilchenko,vmkureychik}@sfedu.ru

Abstract. Quantitative assessment in collective behavior and decision-making in fuzzy conditions is crucial for ensuring the health and safety of the population, ensuring effective response to various emergencies. The task of modeling and predicting behavior in fuzzy conditions, as is known, has increased complexity due to a large number of factors from which an NP-complete multi-criteria task is formed. There is a difficulty in determining the quantitative assessment of the influence of fuzzy factors using a mathematical model. The paper proposes a stochastic model of human decision-making to describe the empirical behavior of subjects in an experiment simulating an emergency scenario. The developed fuzzy model combines fuzzy logic into a conventional model of social behavior. Unlike existing models and applications, this approach uses fuzzy sets and membership functions to describe the evacuation process in an emergency situation. To implement the proposed model, the process of social behavior during evacuation, independent variables are determined. These variables include measurements related to social factors, in other words, the behavior of individual subjects and individual small groups, which are of fundamental importance at an early stage of evacuation. The results of modeling the proposed model of decision-making in odd conditions are carried out, quantifying the degree of optimality of human decisions and determining the conditions under which optimal or quasi-optimal decisions are made. Modeling has shown acceptable results of the proposed approach in solving the problem of evacuation in emergency situations in fuzzy conditions.

Keywords: Evacuation · Human factor · Risk management · Decision-making · Fuzzy conditions

1 Introduction

Currently, special attention is paid to a number of issues in the field of evacuation. The task under consideration includes understanding how the population reacts to evacuation signals, how individual groups of people react to an obvious risk and how such groups of people make decisions about protective actions as a result of various emergency situations (emergencies). The available literature is quite informative in this area [1–7]. In this study, the task of forming a model for making dynamic collective decisions in evacuation tasks

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 575–584, 2023.
https://doi.org/10.1007/978-3-031-27409-1_52

in emergency situations in fuzzy conditions is considered, highlighting important aspects of decision-making about evacuation, discussing research on prevention, risk perception and research specifically devoted to evacuation [3–5].

2 Evacuation Planning

This study examines two main aspects: predicting the behavior of a group in an emergency situation, making decisions more effectively than using simple random decisions; factors influencing the choice of a chain of decisions.

The article is aimed at solving the problem of evacuation in emergency situations in fuzzy conditions by using machine learning interpretation tools. This approach will improve the efficiency of forecasting the evacuation options of the group and will reveal the factors affecting the effectiveness of forecasting.

In this paper, to simplify the description of the algorithm and the behavior model of the group, the members of the group under consideration will be considered as agents with individual characteristics.

Within the framework of the considered decision-making modeling model, aspects have been adopted, which will be disclosed in more detail later.

Agents have two strategies of behavior: normal and reaction stage. Agents are in the normal stage when they perform their pre-emergency actions. Agents in the reaction stage are those who reacted to an emergency situation either by investigation or evacuation. This assumption is based on the model proposed in the paper [4], which showed that evacuation behavior can be classified into various behavioral states.

The normal stage is characterized by certain actions, such as:

- Proactive evacuation, agents move from an unprotected area to a safe place outside of that area before a disaster occurs.
- Shelter: Agents move to shelters inside a potentially unprotected area.
- Local shelter: agents move to higher levels (for example, upper floors) of multi-storey buildings, for example in case of flooding.

In the case of the reaction stage, the following actions occur:

- Rescue: moving the injured with the help of rescue services to get out of the danger zone.
- Escape: salvation by the escape of the victim himself, in order to escape from danger after its onset.

Pre-evacuation planning and preparation are necessary to ensure effective and successful mass evacuation of the endangered population. With the approach of a natural disaster, an expert or a group of experts (depending on the complexity of the task) needs to make a decision on evacuation. After the decision to evacuate is made, evacuation plans should be drawn up.

The agents involved in the evacuation behave rationally, and their transitions from the normal to the reaction stage are controlled by a binary decision-making process,

such behavior can be described using mathematical models based on graph theory. Agents make decisions based on available information and signals during an emergency, following a series of steps: perception, interpretation and decision-making [5, 6]. Thus, on the basis of interpreted information and prompts, passengers can decide whether to switch from normal to the reaction stage.

The decision-making process is influenced by both environmental factors (external) and individual characteristics of agents (internal). Decision-making by agents depends on perceived information, such an influence is called external factors. However, the characteristics of agents (for example, previous experience, physical and mental state, and vigilance) can play a key role, since these internal factors can influence how an individual agent perceives, interprets, and makes decisions [6].

This study uses models based on the binary structure of the decision-making process, which is an approach to modeling that allows us to investigate how several internal and external factors influence the decisions of both individual agents and groups.

3 Definition of Fuzzy Conditions and Analysis of Existing Solutions

Fuzzy logic is a logical–mathematical approach that allows you to present approximate, rather than accurate, reasoning of people. It provides a simple way of reasoning with vague, ambiguous and inaccurate input data or knowledge that fits the context of risk and crisis management.

Fuzzy logic is expressed in linguistic rules that are used as "IF the input variable is a fuzzy set, THEN the output variable is a fuzzy set". Fuzzy inference systems handle it as follows:

- Fuzzification: at this stage, clear input data is transformed into fuzzy data, the degree of belonging of clear input data to pre-defined fuzzy sets.
- Conclusion: you can combine input data using logical fuzzy rules, which allows you to determine the degree of reliability of the data.
- Defuzzification: Defuzzification is required when it is necessary to obtain a clear number as an output from a fuzzy system.

To describe the degrees of truth, a fuzzy variable must contain several fuzzy sets. One set has one membership function, the arguments of the function must correspond to certain values, and the resulting solution must be within the specified range [0;1], this parameter reflects the degree of truth of the solution. The fuzzy inference system uses fuzzy theory as the main computational tool for implementing complex nonlinear mapping. Based on the reviewed works [3–7], it is possible to identify common parameters for describing the membership function: similarity, preference, and uncertainty. The similarity is reflected in the fuzzy analysis of cluster groups and their systems. Preference characterizes one of the tools in the decision-making process. The uncertainty parameter shows the degree of reliability of decisions obtained at the desired stage by expert systems or machine learning methods. The parameters similarity, preference and uncertainty do not exclude each other, and can be combined into a multi-criteria fuzzy decision-making system. In the works [8–12] the main properties and uncertainty of the

behavior of individual agents of the group are described. The rules for the formation of a fuzzy decision-making system are discussed in detail in the works [10–12], the parameters of the developed fuzzy rules are formulated using real data.

The analysis of fuzzy rules shows that this topic is relevant and is described in a limited list of sources of modern literature. The main sources describing fuzzy rules in the field of evacuation behavior are considered. The formulated fuzzy rules are used to obtain linguistic fuzzy rules that can fully describe the uncertainty of agents' behavior using machine learning methods.

4 Dynamic Decision Making Model

The proposed dynamic decision-making model uses fuzzy logic to control the evacuation of agents. Fuzzy sets and rules are defined for the behavior of each agent, which is influenced by the external environment and individual characteristics. Environmental factors and individual characteristics of agents are analyzed in the framework of determining the main aspects affecting the decision-making process by agents, as shown in Fig. 1. The decision-making process has a multi-level hierarchical structure. For example, decisions can be made based on the influence of the environment, psychological foundations and physiological parameters.

As shown in Fig. 2, this article uses fuzzy logic and machine learning methods to model the process of cognition. The factors influencing the behavior of individual agents are modeled as fuzzy input data. For example, the agent's current speed, the agent's position, the relative route of the main group. All these factors can influence the formation of the individual status of each agent in the next iteration.

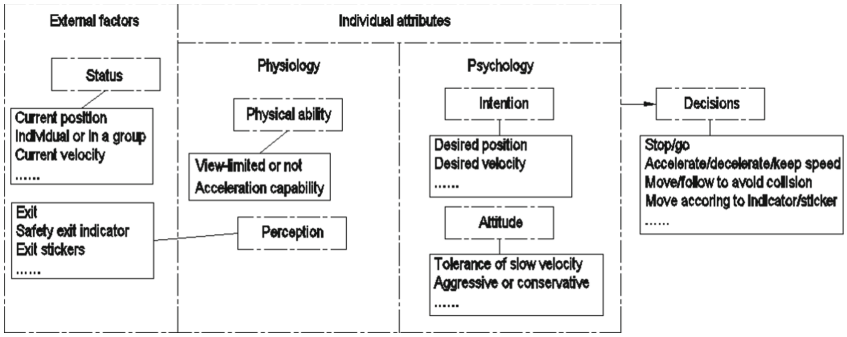


Fig. 1. Decision-making process

The "perception" factor includes the exit location, visibility of the safe exit sign/exit sticker, nearby agents and various obstacles. The "intent" factor contains the value of the movement speed and the coordinates of the agent's position. The "attitude" factor contains individual qualities of character and stress resistance of each agent. Different combinations allow agents to make different decisions, for example, whether he should walk or stop, to which position he should move and whether he should move according to the safe exit sign/exit stickers.

Machine learning algorithms try to "classify" or identify agent selection models based on observed data. An integral part of machine learning is an objective function that displays input output data and criteria for evaluating the efficiency of the algorithm.

$$y = f(x|\varphi) \tag{1}$$

where φ it is a vector of agent parameters for a machine learning model.

Machine learning classifiers can be divided into two main categories, i.e. hard classification and soft classification. The rigid classification seeks to sort through all possible solutions, while the soft classification predicts conditional probabilities for different classes and outputs the resulting solution with a probability fraction. With the help of a soft classification, it is possible to estimate the probability of choosing each option at an individual level, which gives much more information than the methods of a complete search. In other words, it is necessary to evaluate:

$$f(x|\varphi) = P(\arg maxg(x|\varphi)), \tag{2}$$

if $g_k(x|\varphi) = P(k)$, where $k \in \{0, 1\}$.

Interpreted or explicable machine learning is becoming increasingly important in the broad field of machine learning [5, 7, 11]. Machine learning methods can be roughly divided into two main categories, including model-dependent and model-independent. Model-independent methods are usually more flexible, which makes it possible to use a wide range of performance evaluation criteria for various machine learning models.

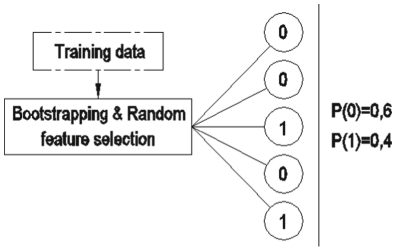


Fig. 2. Dynamic decision-making model with soft prediction mechanism

By means of the considered objective function with a partial dependence of the criteria for evaluating the effectiveness of the solution, it is possible to graphically display the dependence between the input data and the predicted probabilities [7–10].

To properly initialize partial dependency graphs, let's assume that we need to define a dependency x_s, $S \subseteq \{1, \ldots, p\}$, on the results of soft forecasting (probability of choice). It is worth noting that it is necessary to take into account the probability of choice g_k, where $k \in \{0, 1\}$. Partial dependence between x_s and g_k defined by the formula (3)

$$g_{ks}(x_s) = \frac{1}{N} \sum_{i=1}^{n} g_k\left(x_s, x_{C(i)}\right), \tag{3}$$

where $x_{C(i)}$, $(i = 1, \ldots, N)$, this variable determines the average marginal effect on the predicted probability of choosing each agent.

In many previous studies, this approach was used to quickly identify nonlinear relationships between output data and response data for machine learning models, for example, in the framework of solving a problem (black box) [7–10].

5 Algorithm for Making Dynamic Decisions

The selection and calibration of the objective function is based on the simulation results. In this article, three criteria are used from which the target fiction is formed: triangular, standard deviation and Gaussian function [8, 9]:

$$triangmf\,(x) = \begin{cases} 0,1 & (x < a)||(x > c),(x == b) \\ (x-a)/(b-a) & (x < b) \\ (c-x)/(c-b) & \infty \end{cases} \tag{4}$$

$$sigmf\,(x) = \frac{1}{1+e^{x}}, \tag{5}$$

$$gausmf\,(x) = e^{-\left((x-c)^2/2\right)}, \tag{6}$$

where a, b, c parameters that reflect the angle of increase of the graph of the objective function.

The process of preliminary formation of the objective function improves the quality of the solutions obtained, it is necessary to form a vector of criteria for the objective function taking into account each fuzzy factor, as shown in Fig. 3.

Step 1. The fuzzy component is divided into several linguistic groups. The time allocated for rest can also be divided into several linguistic groups.

Step 2. The time allocated for rest can be determined by modeling [7–12].

Step 3. In accordance with the formed pre-selection function and a set of fuzzy rules, a rest mechanism with a periodicity system is initialized. After each stage of rest, the target function is calibrated in accordance with the current indicators obtained.

An example of the formation of an objective function based on a fuzzy rest criterion is considered, for the remaining criteria, target functions are also formed and a certain algorithm for data processing and calibration of the main objective function is performed.

6 Experimental Part of the Study

As an example, the work models a cinema room. The simulation room is shown in Fig. 4. The room contains two exits, one in front, the other in the back. The personal characteristics of each agent were taken into account individually for each decision made. A relevant event of each agent is his decision to respond to an emergency situation and a change in the state of at least one of the other visible participants or belonging to his/her personal group (i.e. the group with which the participant attends the film).

The shaded squares represent agents who are in contact with the agent in question (making the decision). The remaining squares represent agents belonging to the personal group of the decision-making agent.

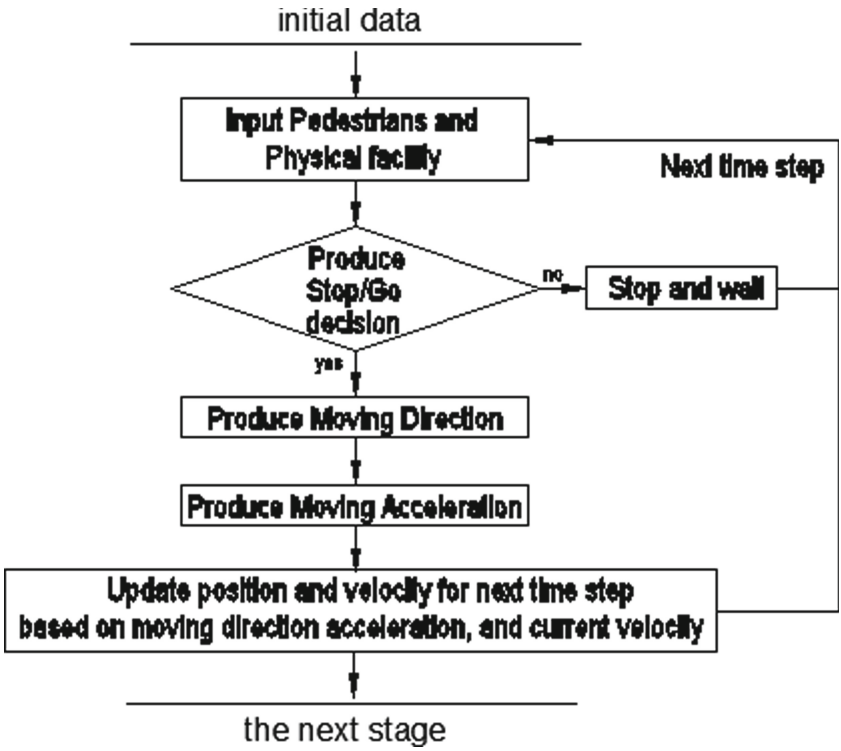


Fig. 3. Block diagram of the decision-making algorithm

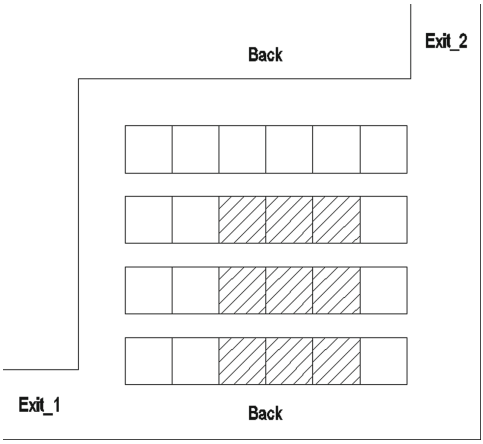


Fig. 4. Simulated room

In this paper, the objective function F_1 is used as an indicator of the performance of the dynamic decision-making model in fuzzy conditions

$$F_1 = 2\frac{P \cdot R}{P + R}, \tag{7}$$

where P - this is the ratio of solutions with an indicator of the objective function satisfying the efficiency criterion to the total number of positive solutions, but R - solutions with an indicator of the objective function satisfying the efficiency criterion for all solutions obtained, including sampling errors. F_1 - this is a weighted average of the ratio of variables P and R.

According to the simulation results, the most effective model under agents = 400. The graphs show different personal parameters of the agents when modeling the model. It is worth noting that the parameter P (accuracy) is a more important indicator than R in the case of evacuation, since there is a large number of false positives or erroneous calls, this causes low accuracy, which can increase the level of false positive decisions.

A model with a high accuracy parameter seems to be more optimal, while the value of the objective function seems to be the best metric, taking into account the specified vector of criteria.

In the Figs. 5 and 6 the results of modeling the objective function and criteria are shown P, R.

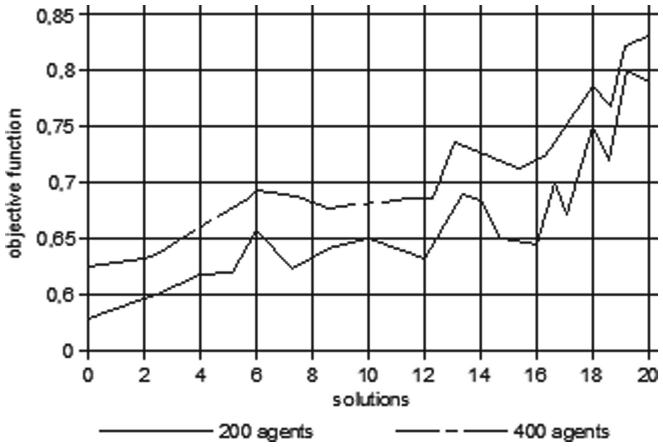


Fig. 5. Modeling the objective function

In this work, an optimal model is obtained, mainly based on an estimate of the value of the objective function. This model has one of the best solutions within the given criteria.

7 Conclusion

As part of this work, we modeled and interpreted decision-making before evacuation using machine learning interpretation tools in fuzzy conditions. The conducted tests have

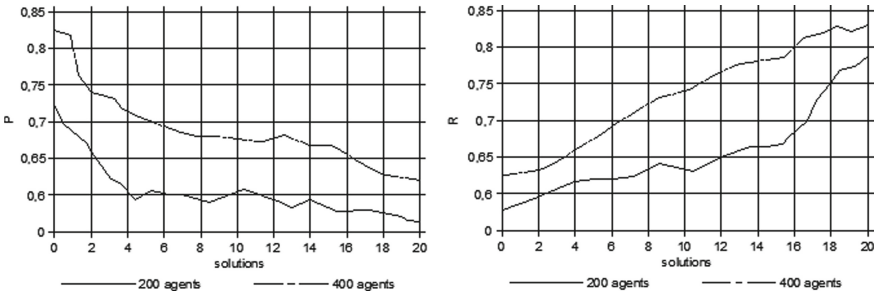


Fig. 6. Modeling criteria P, R

shown that the proposed algorithm for making dynamic decisions in fuzzy conditions can improve the result by using fuzzy rules for modeling the movements and behavior of the team when making dynamic decisions.

Acknowledgment. The research was funded by the Russian Science Foundation project No. 22-71-10121, https://rscf.ru/en/project/22-71-10121/ implemented by the Southern Federal University.

References

1. Gerasimenko, E., Rozenberg, I.: Earliest arrival dynamic flow model for emergency evacuation in fuzzy conditions. IOP Conf. Ser.: Mater. Sci. Eng. **734**, 1–6 (2020)
2. Reneke, A.: Evacuation decision model. US Department of Commerce, National Institute of Standards and Technology. https://nv-pubs.nist.gov/nistpubs/ir/2013/NIST.IR.7914.pdf (2013)
3. Kuligowski, E.D.: Human behavior in fire. In: The Handbook of Fire Protection Engineering, pp. 2070–2114. Springer (2016). https://doi.org/10.1007/978-1-4939-2565-058
4. Kuligowski, E.D.: Predicting human behavior during fires. Fire Technol. **49**(1), 101–120 (2013). https://doi.org/10.1007/s10694-011-0245-6
5. Akter, T., Simonovic, S.P.: Aggregation of fuzzy views of a large number of stakeholders for multi-objective flood management decision-making. J. Environ. Manag. **77**, 133–143 (2005)
6. Greco, S., Kadzinski, M.V., Mousseau, V., Slowinski, L.: ELECTREGKMS: robust ordinal regression for outranking methods. Eur. J. Oper. Res. **214**(1), 118–135 (2011)
7. Gerasimenko, E., Kureichik, V.V.: Minimum cost lexicographic evacuation flow finding in intuitionistic fuzzy networks. J. Intell. Fuzzy Syst. **42**(1), 251–263 (2022)
8. Sheu, J.B.: An emergency logistics distribution approach for quick response to urgent relief demand in disasters. Transp. Res. Part E-Logist. Transp. Rev. **43**, 687–709 (2007)
9. Zhao, Yan, X., Van Hentenryck, P.: Modeling heterogeneity in mode-switching behavior under a mobility-on-demand transit system: an interpretable machine learning approach. arXiv preprint arXiv:1902.02904 (2019)
10. McRoberts, B., Quiring, S.M., Guikema, S.D.: Improving hurricane power outage prediction models through the inclusion of local environmental factors. Risk Anal. **38**(12), 2722–2737 (2018). 10. 1111/risa.12728
11. Chai, C., Wong, Y.D., Er, M.J., Gwee, E.T.M.: Fuzzy cellular automata models for crowd movement dynamics at signalized pedestrian crossings. Transp. Res. Rec.: J. Transp. Res. Board **2490**(1), 21–31 (2015)

12. Zhao, Hastie, T.: Causal interpretations of black-box models. J. Bus. Econ. Stat. 1–19 (2019) (just-accepted). https://doi.org/10.1080/07350015.2019.1624293

Conversion Operation: From Semi-structured Collection of Documents to Column-Oriented Structure

Hana Mallek[1(✉)], Faiza Ghozzi[2], and Faiez Gargouri[2]

[1] Miracl Laboratory, University of Sfax, Sfax, Tunisia
mallekhana@gmail.com
[2] Miracl Laboratory, University of Sfax, ISIMS, Sfax, Tunisia
faiza.ghozzi@isims.usf.tn, faiez.gargouri@usf.tn

Abstract. Over the last few years, NoSQL databases have been a key for several problems for storing Big Data sources as well as implementing data warehouses (DW). With decisional systems, NoSQL column-oriented structure can provide relevant results for storing a multidimensional structure, where the relational databases cannot be able to handle the semi-structured data. In this research paper, we attempt to model a conversion operation in the ETL process, which is responsible for Extracting, Transforming and Loading data into the DW. Our proposed operation is handled in the ETL extraction phase, which is responsible for converting a series of semi-structured data to column-oriented structure. In the implementation phase, we propose a new component using Talend open studio for Big Data (TBD), which helps ETL designer to convert semi-structured data into column-oriented structure.

Keywords: Conversion operation · column-oriented · ETL process

1 Introduction

Over the years, the amount of information seems to increase exponentially, especially with the blooming grouth of new technologies such as smart devices and social networks like; Twitter, Facebook, Instagram, etc. Thereby, the term "Big Data" arose. Since the amount of information exceeds the management and storage capacity of conventional data management systems, several areas, namely the decision-making area, has to take into account this data growth. Nevertheless, it is obvious that various issues and challenges arise for the decision-making information system, mostly at the level of the integration system ETL (Extract-Transform-Load). It is noteworthy that massive data storage is a problem tackled by several researchers in order to find a good alternative for the classical models (relational databases, flat files, etc.) which are rigid and support only structured data. Indeed, NoSQL models are considered as a solution for the limitations of these typical models which are known as shameless databases. In the decision-making context, researchers faced several challenges when analyzing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 585–594, 2023.
https://doi.org/10.1007/978-3-031-27409-1_53

massive data, such as the heterogeneity of data sources, filtering uncorrelated data, processing unstructured data, etc. Furthermore, researchers Sharma et al. [12] handled the use of different NoSQL models and demonstrated the importance of these models for enterprises in order to improve scalability and high availability. From this perspective, several research works attempted to elaborate solutions to convert typical models (relational, flat files, etc.) to one or more NoSQL models. The main objective of this paper is to model the conversion operation of ETL processes which aims to convert a semi-structured data to a column-oriented structure. In this regard, we introduce the formal structure, algorithm and the implementation of this operation in the context of Big Data as a solution.

The remainder of this paper is organized as follows: Sect. 1 exhibits related works. Section 2 displays a formal model of the conversion operation and identifies the proposed algorithm. Section 3 foregrounds the experimental results illustrated through Talend for Big Data (TBD) invested to test our new component. Section 4 wraps up the closing part and displays some concluding remarks.

1.1 Related Works

The majority of works in the literature provide a solution with column-oriented, document-oriented or other NoSQL structures. Hence, many works took advantage of the column-oriented structure and considered it as a good alternative for classical structures (relational, CSV, XML, etc.), such as Chung et al. [7], who invested the column-oriented structure to implement the *JackHare* Framework that provides relational data migration to the column-oriented structure (HBase). Instead of replacing relational databases with NoSQL databases, Liao et al. [11] developed a data adaptation system that integrates these two databases. In addition, this system provides a mechanism for transforming a relational database to a column-oriented database (HBase). In the decision-making context, we find several works which take advantage of the NoSQL structure in order to accomplish different objectives such as Boussahoua et al. [5,8] who emphasized in their works that column-oriented NoSQL model is suitable for storing and managing massive data, especially for BI queries. The works presented above asserted that the relational structure should not be lost through the column-oriented structure since it is simple and easy to understand. Indeed, the document-oriented structure can handle more complex forms of data. This structure does not follow a strict structure, where the key-value pairs of a JSON, XML, etc. document can always be stored. Many researchers choose the document-oriented structure, in order to preserve the structure of a large semi-structured data collection (JSON, XML, etc.) such as geographic data Bensalloua et al. [3], or a large heterogeneous data collection (data lake) with [1]. Moreover, authors Yangui et al. [13] reported a conceptual modeling of ETL processes to transform a multidimensional conceptual model into a document-oriented model (MongoDB) through transformation rules. Both column-oriented and document-oriented structures offer several merits. In the BI context, authors Chevalier et al. [6] highlighted the reliability of the column-oriented NoSQL database

(HBase) over the document-oriented database MongoDB in terms of time load to implement OLAP systems. Several other works opted to provide developers the choice of using the NoSQL structure (column-oriented, document-oriented, etc.) depending on their needs in order to maintain the diversity of the data structure. Several researchers developed a Framework that supports more than one NoSQL model such as Banerjee et al. [2], Kuszera et al. [10] and Bimonte et al. [4]. These research works are quite relevent but the conversion operation of semi-structured data to NoSQL structure is not handled especially in decisional systems with the exception of the work of Yangui et al. [13].

2 Formal Model of the Conversion Operation

The main objective of our conversion operation $Conv$ is to apply a conversion rules on a collection Col of semi-structured documents with JSON type to have as output a column-oriented table Tab.

A column-oriented table Tab is defined by $(N_{Tab}, CF_{Tab}, L_{Tab}, RK_{Tab})$ where:

- N_{Tab} is the name of column-oriented table.
- $CF_{Tab} = \{CF_1, .., CF_j, ..CF_m\}$ is a set of column families.
- $L_{Tab} = \{L_1, .., L_k, .., L_n\}$ is a set of lines.
- $RK_{Tab} = \{RK_1, .., RC_c, .., RK_n\}$ is a set of identifiers where RK_c is the identifier of the line L_k.

A column family FC_j is defined by $(N_{FC}, \mathscr{C}_{FC}, \mathscr{L}_{FC})$ where:

- N_{FC} is the name of the column family.
- $\mathscr{C}_{FC} = \{C_1, .., C_k, .., C_p\}$ is a set of columns that represents a FC_j where, the data access of a FC_j is done through it.
- $\mathscr{L}_{FC} = \{L_1\}, .., \{L_h\}, .., \{L_n\}$ is a set of lines that represents the set of values of the columns C_k belonging to a family of columns FC_j.

The formalization of a collection of documents Col is defined as a set of documents $d_i(\ Col = d_1, .., d_i, .., d_{nc})$. Where $nc = \mid Col \mid$ is the size of the collection. In this paper, we adopt the definition of Ben Hamadou et al. [9] of the semi-structured document. Each document is described as a key-value pair, where the key identifies the JSON object (document) and the value refers to the document which can either be atomic or complex.

A document $d_i \in Col$, $\forall i \in [1, nc]$, is defined as a pair (key, value): $d_i = (k_{di}, v_{di})$. Where, k_{d_i} is a key that identifies the document d_i in the collection Col, v_{d_i} is the value of the document. The value v can be atomic or complex form (This definition is detailed in the research paper of Ben Hamadou et al. [9]).

The conversion operation is modeled as follows:

$$Conv(Col) = Tab$$

2.1 Conversion Rules

In order to ensure the conversion operation, a set of rules needs to be respected. These rules of transformation are summarized as follows:

- **Rule 1**: Each document d_i is transformed to a column family CF_j; where the atomic values v are transformed to columns C_k and the content of its values v are the values of the lines.
- **Rule 2**: Each key of a document K_{di} is transformed to a Row Key RK.
- **Rule 3**: Each atomic value v is transformed to a column C_k and the content of its value v is the value of the row.
- **Rule 4**: Each complex value v of object type and containing only atomic values is transformed to a column family CF_j; the attributes a_l are transformed to columns C_k and the values v_l correspond to the values of the lines.
- **Rule 5**: Each complex value v of object type and has non-empty objects; v_l is transformed to a column family CF_j; the composed objects v_l are transformed to column families CF_j in a recursive way through applying the previous rules (3 and 4).
- **Rule 6**: Each complex value v of array type and not empty, is transformed into a new column through a recursive call in the case where the values v_l are atomic; if these values are of object type and not empty, a new table is created where each value is transformed into a column family CF_j in a recursive way through applying the preceding rules (3, 4 and 5) (Fig. 1).

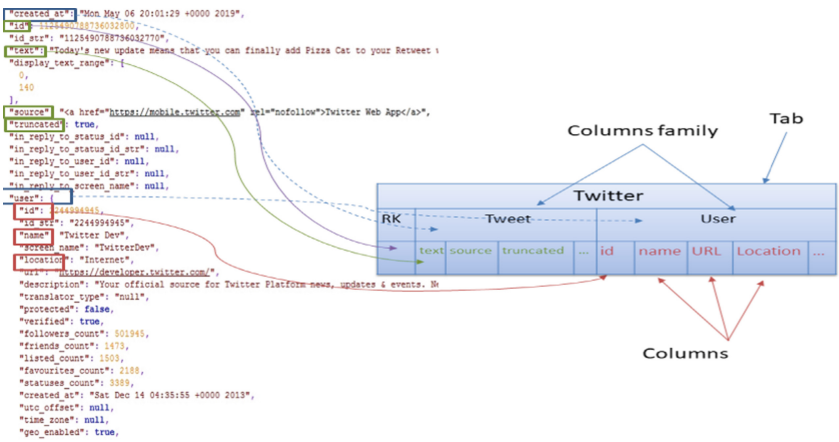


Fig. 1. Conversion operation from JSON structure to column-oriented structure

2.2 Conversion Operation Algorithm

The conversion procedure Algorithm 1 is launched by the procedure *ConversionCollection (collection)*, as illustrated in the algorithm below.

The procedure *ConversionCollection (collection)* rests upon the following steps:

- First, it allows to create a column-oriented table $TableName$ where the table name is the name of the collection $Collection.Name$ and the column family CF is the name of a document (lines 1 and 2).
- secondly, a loop allows to browse the documents d_i of a collection (line 3). For each document d_i, the Row key RK takes as value the key of a document Kd_i (line 4). We call the recursive procedure *Conversion* ($di.v$, *TableName*), in which $di.v$ is the value of a document (line 5).

Algorithm 1: Conversion of a collection of semi-structured documents

Input: *collection*
Output: converted collection
1 $TableName \leftarrow Collection.Name$;
2 CreateTable($TableName, CF$);
3 **foreach** *document* d_i **do**
4 $RK \leftarrow Kd_i$;
5 Conversion ($d_i.v$,$TableName$) ;

The Algorithm 2 of the procedure *Conversion(di.v, TableName)* is defined in the algorithm below which presents a series of steps ensuring the fullfillment of the objective of the conversion operation:

- This algorithm requires as input a value *val* (of type value) which can be either object or array and the name of the target table ($TableName$) oriented columns.
- An initialization is performed for the variable CF which represents the name of the family of columns. This variable takes the content of the variable *val*, and the variable $Table$ takes the name of the table from the variable $TableName$ (line 1).
- Afterwards, the existence of the table ($Table$) and columns family CF is tested. If the table does not exist, we create a table named "Table", using the procedure ($CreateTable(Table, CF)$). If this table exists and the column family does not exist, we call the update procedure ($MAJTable(Table, CF)$) to add the column family CF (line 7).
- For each value v in a complex value *val*, we test if the value v is an atomic value, then v is added in the list *LVA*. If the value v is a complex value of an object type, then v is added to the list *LObj*. If the value v is a complex value with type Array, then v is added to the list *LArray* (lines 10–18).
- We test if the list of atomic values *LVA* is not null then, each atomic value is transformed into a column through the procedure AddColumns(v,Table) and its content is filled through the function AddValue(v_l,Table) (lines 19–22).

Algorithm 2: The Conversion operation algorithm

```
   Input: val,TableName
   Output: converted documents
 1 CF ← val, Table ← TableName;
 2 LVA, LObj, LArray = null;
 3 if Table exists then
 4     if CF exists then
 5         Continue
 6     else
 7         MAJTable(CF,Table)
 8     else
 9         CreateTable (Table,CF)
10 foreach value v in val do
11     if v is a VA then
12         LVA ← LVA + v ;                    /* List of atomic values */
13     else
14         if v is an object then
15             LObj + v ;                         /* List of objects  */
16         else
17             if v is an array then
18                 LArray ← LArray + v ;          /* List of Array */
19 if LVA is not null then
20     foreach attribute in LAtt do
21         AddColumn(v,Table);
22         AddValue(Val,Table)
23 if LObj is non-null then
24     foreach object in LObj do
25         Conversion(object,Table);              /* Recursive call */
26 if LArray is non null then
27     foreach value v_l in LArray do
28         if the value v_l is of type object non null then
29             Conversion(object,NewTable);       /* Recursive call  */
30             AddColumn(KeyRef,Table)
31         else
32             Conversion(array,Table);           /* Recursive call  */
```

– We test if the list of complex values is an object type *LObj*. Then, we make a recursive call of the procedure $Conversion(object, Table)$ for each object in the list (lines 23–25).
– We test if the list of the complex values of type array *LArray* is not null. We test if there are complex values of object type. Then, we create a new

table and for each object we make a recursive call of the procedure *Conversion(object,NewTable)* in order to add a column of referencing $CleRef$ by the identifier of the object (lines 26 until 32).
– If there are no complex values, we make a recursive call with the following structure $Conversion(array, Table)$ sa as to add a new column.

3 Experiments

We are mainly concerned with the first phase of the ETL process, which is responsible for ensuring the extraction of semi-structured data and guarantees the execution of the conversion operation. In this phase, we shall ensure the execution of both algorithms of our two operations through a new component with Talend for Big Data. This component allows us to read a list of JSON files from the social network Twitter. Afterwards, we shall apply conversion rules in order to get a column-oriented structure. Subsequently, the designer has the possibility to select how to store the converted tables either:

– **On a single table**: in this case, the conversion operation is completed without performing a physical partitioning, in order to obtain a single table with several column families.
– **On several tables**: in this case the conversion operation fragments (physically) the tables and each table will be processed in a separate way.

3.1 Experimental Protocol

In this section, we illustrate the different configurations we used to evaluate the performance of our new component. In this case, we performed all our experiments on an Intel I7 i7-7500U processor with a clock speed by 2.90 GHz. The machine had 12 GB RAM and a 1 TB SSD.

3.2 A New Component with Talend for Big Data

The talend for Big Data tool offers a workspace for the developer to create a new component according to their needs. In our case, we create a new component called "tJSONToHBase" which grants the designer the freedom to model the conversion operation from semi-structure data (JSON) to column-oriented structure through HBase NoSQL database. The creation of the component tJSONToHBase starts with a description of the XML file: "tJSONToHBase_java.xml". We summarize the various characteristics of this component in Table 1. This component belongs to the "BigDimETL/Extraction" family in the talend palette. It is considered as an input component.

Table 1. The descriptive characteristics of the component "tJSONToHBase"

XML tags	Description
Family	BigDimETL/Extraction
Connectors	MAX_INPUT = "0", MIN_OUTPUT = "1"
Parameters	Input file or folder Requested input scheme Conversion mode The required column families and columns
Advanced parameters	Column family with corresponding table

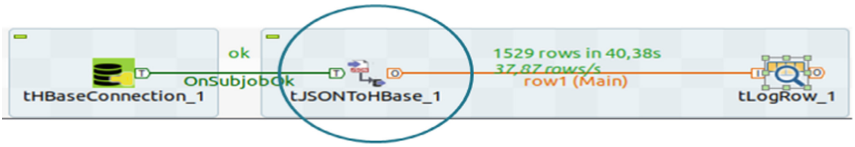


Fig. 2. Conversion from JSON structure to column-oriented structure

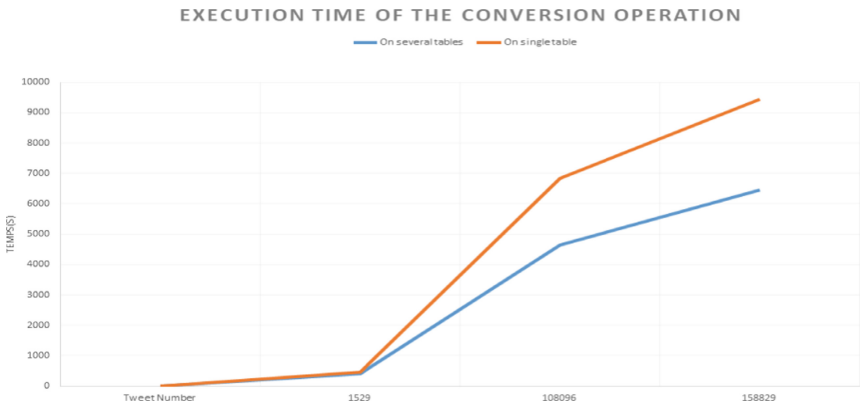


Fig. 3. Variation of the execution time of the conversion process

3.3 Evaluation of the Conversion Process

The "tJSONToHBase" component execution is illustrated in the Fig. 2.

Our component offers the possibility to choose the storage mode either on a single table or on several tables.

Table 2 portrays respectively the variation of the execution time compared to the number of tweets and compared to the number of tweets processed per second with the conversion mode on several tables and on a single table. We report in the Fig. 3 the measurements of the execution time compared to the number of tweets of the Table 2. It is to be noted that the conversion mode on a single table is more efficient than on several tables. The importance of this conversion is inferred through the number of Tweet/s as well as on the execution time which is less important than as that on several tables.

In this respect, it is worth noting that the processing speed of the tweets (3.73 and 27.7) for the small collection is very low. This speed stabilizes for the large collections, which is applicable for both modes.

Table 2. Variation of the execution time for the two conversion modes

Tweet number	Conversion mode			
	On a single table		On several tables	
	Tweet/s	Execution time	Tweet/s	Execution time
1529	27.27	59.08	3.73	409.82
108096	49.32	2191.52	19.34	4650.6
158829	53.06	2993.62	18.73	6451.65

4 Conclusion

At this stage, we would assert that in this research paper, we have elaborated the formalisation and the algorithm of conversion operation in the ETL processes. The developed solution yields the migration from a semi-structured documents to column-oriented structure for implementing multidimensional DW. Through experimentation, our solution is developed through identifying a new component using Talend for Big Data for ETL designer in order to convert semi-structured collection of JSON type into HBase database. As future work, we intend to model all operations in ETL processes for developing DW for Big Data.

References

1. Abdelhedi, F., Jemmali, R., Zurfluh, G.: Ingestion of a data lake into a NOSQL data warehouse: the case of relational databases. In: Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 3, pp. 25–27 (2021)
2. Banerjee, S., Bhaskar, S., Sarkar, A., Debnath, N.C.: A unified conceptual model for data warehouses. Ann. Emerg. Technol. Comput. (AETiC) **5**(5) (2021)
3. Bensalloua, C.A., Benameur, A.: Towards NOSQL-based data warehouse solution integrating ECDIS for maritime navigation decision support system. Informatica **45**(3) (2021)
4. Bimonte, S., Gallinucci, E., Marcel, P., Rizzi, S.: Data variety, come as you are in multi-model data warehouses. Inf. Syst. **104**, 101734 (2022)
5. Boussahoua, M., Boussaid, O., Bentayeb, F.: Logical schema for data warehouse on column-oriented NoSQL databases. In: Benslimane, D., Damiani, E., Grosky, W.I., Hameurlain, A., Sheth, A., Wagner, R.R. (eds.) DEXA 2017. LNCS, vol. 10439, pp. 247–256. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64471-4_20
6. Chevalier, M., El Malki, M., Kopliku, A., Teste, O., Tournier, R.: Implementing multidimensional data warehouses into NOSQL. In: Advances in Databases and Information Systems—19th East European Conference, ADBIS 2015, Poitiers, France (2015)
7. Chung, W., Lin, H., Chen, S., Jiang, M., Chung, Y.: Jackhare: a framework for SQL to NOSQL translation using MapReduce. Autom. Softw. Eng. **21**(4), 489–508 (2014)

8. Dehdouh, K., Boussaid, O., Bentayeb, F.: Big data warehouse: building columnar NOSQL OLAP cubes. Int. J. Decis. Support Syst. Technol. (IJDSST) **12**(1), 1–24 (2020)
9. Hamadou, H.B., Ghozzi, F., Péninou, A., Teste, O.: Querying heterogeneous document stores. In: 20th International Conference on Enterprise Information Systems (ICEIS 2018), vol. 1, pp. 58–68 (2018)
10. Kuszera, E.M., Peres, L.M., Fabro, M.D.D.: Toward RDB to NOSQL: transforming data with metamorfose framework. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 456–463 (2019)
11. Liao, Y.T., Zhou, J., Lu, C.H., Chen, S.C., Hsu, C.H., Chen, W., Jiang, M.F., Chung, Y.C.: Data adapter for querying and transformation between SQL and NOSQL database. Fut. Gener. Comput. Syst. **65**(C), 111–121 (2016)
12. Sharma, S., Shandilya, R., Patnaik, S., Mahapatra, A.: Leading NOSQL models for handling big data: a brief review. IJBIS **22**(1), 1–25 (2016)
13. Yangui, R., Nabli, A., Gargouri, F.: ETL based framework for NoSQL warehousing. In: European, Mediterranean, and Middle Eastern Conference on Information Systems, pp. 40–53. Springer (2017)



Mobile Image Compression Using Singular Value Decomposition and Deep Learning

Madhav Avasthi(✉), Gayatri Venugopal, and Sachin Naik

Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India
avasthi.madhav97@gmail.com

Abstract. Mobile images generate a high amount of data therefore, efficient image compression techniques are required, which can compress the image while maintaining image quality. This paper proposes a new lossy image compression technique that maintains the psychovisual redundancy using Singular Value Decomposition (SVD) and Residual Neural Networks (ResNet50). Images are compressed by SVD using the rank K of the image. However, it is difficult to predict the correct value of K to compress the image as much as possible while maintaining the quality of the image. First, a relation between the energy of the image and the number of focal points in an image is derived, using which 1500 images are compressed using SVD while maintaining psychovisual redundancy. This data is then used to train ResNet-50 to predict the rank values. The proposed method obtained a compression ratio of 41.9%, with 86.35% accuracy of rank prediction for the entire dataset. Therefore, the proposed method can predict the correct value of rank K, and hence automate the compression process while maintaining the psychovisual redundancy of the image.

Keywords: Deep Learning · Image Compression · Psychovisual Redundancy · Residual Neural Network (ResNet-50) · Singular Value Decomposition (SVD)

1 Introduction

Image data generated by mobile phones has increased exponentially in the past decade with the emergence of high-resolution cameras in the industry. An increase in the number of multimedia files can be observed according to Cisco's report on internet traffic, which forecasts that global traffic will grow 2.75 times between 2016 and 2021 [6]. However, in the past few decades, researchers have realized that the storage capacities have reached their upper limit due to the limitations of the laws of Physics [1]. Furthermore, the data transmission rate was not at par with the available storage capacity [28].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 595–606, 2023.
https://doi.org/10.1007/978-3-031-27409-1_54

Therefore, with 2.5e+9 GBs of data being produced daily, storing and transmitting data has become a significant challenge [2]. To solve this problem, data compression algorithms can be applied. It is a process of representing data using fewer numbers of bits than the original representation [11]. Over the years, various image compression studies have been suggested to compress data for further processing to reduce the storage cost or transmission time [5,7,8]. The objective of image compression is to reduce the number of bits required to represent an image either for transmission purposes or storage purposes [6].

Image compression can be categorized into lossy compression and lossless compression [7]. During Lossy compression, some part of the original data is lost, but it still possesses good fidelity. In Lossless compression, the original image is restored, and there is no distortion observed at the decoding stage, although the compression rate is often low [33]. The original image can be retrieved from the compressed image in lossless compression methods, whereas in lossy methods, some of the data is permanently lost, and hence the original image cannot be recovered after compression [8].

In this paper, we discuss the use of singular value decomposition (SVD) [8] for image compression, a lossy image compression technique that allows the image to be broken into multiple matrices and hold on to the singular value of the image. It is necessary while releasing the values which are not in order to retain the image quality and derive a relation between the number of focal points, brightness of the image, and rank versus energy graph of the image. This relation helps to attain higher compression ratios for mobile images while maintaining psychovisual redundancy, i.e., the ability to remove the unwanted information from the image that is not recognizable to the human eye, hence no change can be determined by the naked eye. Further, the compression process is automated by predicting the rank of the image by using ResNet-50, a convolution neural network used to produce efficient compression, keeping other performance parameters discussed further constant [29].

2 Performance Parameters

2.1 Compression Ratio (CR)

The bits required to represent the original size of the image with respect to the number of bits required to represent the compressed size of the image is called the compression ratio [4]. The compression ratio gives the details about the number of times the image has been compressed. R = n1/n2, where n1 represent the number of bits required for the original image and n2 represent the number of bits required for the compressed image.

2.2 Mean Square Error (MSE)

The MSE is the cumulative squared error between the compressed and the original image [4].

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{1}$$

2.3 Peak Signal to Noise Ratio (PSNR)

The PSNR is used to quantify the quality of reconstruction of an image [4]. The original data acts as a signal, and the error incurred is the noise produced due to compression. In order to compare the compression, it is used to approximate the human perception of reconstruction quality. Therefore, in some cases,

$$PSNR = (10) \log_{10} \frac{(255)^2}{MSE} \tag{2}$$

A higher PSNR value is considered good as it indicates that the ratio of Signal to Noise is high [4].

As is visible from the two formulas presented above, PSNR and MSE are inversely proportional to each other. This means a high PSNR value (in dB) indicates a low error rate, and the reconstruction may look closer to the original picture [4].

3 Theory

3.1 Singular Value Decomposition

Colored images are a combination of three matrices, namely red, green, and blue, which contain numbers that signify the intensity values of the pixels in an image. Singular Value Decomposition (SVD) decomposes a given matrix into three matrices called the U, I, and V. U and V are orthogonal matrices, whereas I is a diagonal matrix that contains the singular values of the input matrix in descending order. The rank of the matrix is determined by the non-zero elements in the diagonal matrix, I [3]. Compression is performed using a minor rank that is obtained by eliminating small singular values to approximate the original matrix.

3.2 ResNet-50

It is a convolutional neural network that is a variant of the ResNet model [29]. ResNet stands for Residual networks. It contains 48 convolutional layers connected with an average pooling and a max pooling layer. To achieve higher accuracy in deep learning, a higher number of layers are required in the neural network. However, increasing the number of layers is not an easy task, as when the layers are increased, there exists a problem of vanishing gradient. However, ResNet-50 solves this problem using "skip connections." As the name suggests, skip connections skip the convolution layer, and add the input of one layer to the output of another. This allows the next layer to perform at least as good as the previous layer and solves the notorious problem of vanishing gradient [29].

4 Literature Survey

Dasgupta and Rehna [8] points out that even though the processing power has increased with time, the need for effective compression methods still exists. For this, they propose compression using SVD (Singular Value Decomposition) which is used to reveal the fundamental structure of the matrix. SVD is used to remove redundant data. The algorithm proved successful with a compression to a great extent with only a limited decrease in the image quality. However, the authors did not give due diligence to the decompression of the image and extraction of the original image via decompression.

Babu and Kumar [9] discusses the disadvantage of fixed-length code as it can only be compressed to a specific limit. The author proposed Huffman coding as an efficient alternative to the existing system [30]. To restore the original image, the decoder makes use of a lookup table. In order to decode the compressed image, the algorithm compares each bit with the information available in the lookup table. When the metadata of the image matches with the information in the lookup table, the transmitted metadata is recognized as unique. The results clearly show that the proposed algorithm takes 43.75 microseconds (μs) to perform decoding, which is 39.60 less than the tree-based decoder for a dataset of 12×32.

Vaish and Kumar [10] also finds the requirement for more efficient ways to compress images. To do so, they propose a new method that uses Huffman code and Principal component analysis (PCA) for the compression of greyscale images. First, the compression is performed using PCA, where the image is reconstructed using a few numbers of principal components (PCs), removing the insignificant PCs. Further, quantization with dithering is performed which helps reduce contouring. The proposed method results are compared with JPEG2000. The results prove the successful as it provides better compression compared to JPEG2000, also generating a higher PSNR value from the same.

Rufai et al. [11] discusses lossless image compression techniques in the medical imaging field. They realized that JPEG2000 was available for the same, but was difficult to apply, as it involved complex procedures. The author proposed using Huffman code and SVD to regenerate the original image precisely. In the proposed method, SVD is used to compress the image to remove the low, singular values. Then, the generated image is further compressed using the Huffman code, and the final compression ratio was produced by multiplying the ratio from both algorithms. The results proved proposed algorithm to have low MSE and high PSNR compared to JPEG2000. However, this method cannot be used on colored images, which tends to be a significant drawback.

Bano and Singh [12] discussed the importance of security required for the storage and transmission of digital images. They studied various data hiding techniques and executed an algorithm based on the block-based cipher key image encryption algorithm. The authors tested the algorithm in order to obtain results that offered higher complexity and higher PSNR value. They concluded that there was a feeble loss in the quality of the image. Furthermore, the authors approached the concept of encryption of the image. They concluded that encryp-

tion and steganography together, improved the security of data and the proposed algorithm is viable for hardware testing in order to test its speed.

Erickson et al. [13] explains the importance and use of machine learning in medical imaging. The new algorithms could work with changes made in data. Therefore, the algorithm is selected based on the viability of the given dataset and time and space efficiency. The authors emphasized the importance of deep neural networks and CNN for image classification. The recent developments in deep neural networks and CNN have revolutionized the field of visual recognition tasks. However, the authors could not explain the application of CNN in visual recognition tasks and medical imaging and which techniques are used to perform it.

Narayan et al. [14] demonstrated compression of radiographic images using feedforward neural networks. They took the existing techniques into consideration and devised a training set by sampling an image. The experiment showed that compression ratios of up to 32 with SNR of 15-19db could be achieved. The authors realized that this approach had a few weaknesses. They devised an alternative training strategy that used a training set that resembled data from the actual image. The result was equally good as the image dependent and may also have better generalization characteristics.

The authors realized that they got very little insight into the internal operation of alexnet [15]. Hence, another study [16] proposed a visualization technique to show which inputs in each layer excite an individual feature map. This visualization technique helped them understand how the features are developing at each layer. The technique beat the single model result of alexnet by 1.7%, which was the highest in the model. However, the test on other datasets proved that the method is not universal.

Krizhevsky et al. [17] realized that even with one of the largest available datasets - ImageNet, the problem of object detection could not be solved due to the extreme complexity of the task. Therefore, a model with prior knowledge to compensate for the unavailable data was required. This problem was solved by using the Convolutional Neural Network (CNN). They used ReLU (Rectified Linear Units) nonlinearity function [31] which solved the problem of saturation and increased the training speed by several times. The architecture proposed had top-1, and top-5 error rate of 16.4%, nearly 9% less than the previously available technology. However, the limited GPU memory and small datasets available proved to be a hindrance in the research.

In [18], the authors observed that object detection performance that was measured on the PASCAL VOC had stagnation in the results. In the proposed method, they used a selective search algorithm, as opposed to a CNN algorithm, where it first breaks down the image into pieces and then groups it into bottom ups manner so as to group pixels of similar features. The extracted features are used to classify the warp image patch using support vector machines. The results showed a 10% increase over secDPM results. On the ILSVRC 2013 dataset, the proposed model gave 31.4% mean average precision. However, the authors did not consider the time and space efficiency.

The author in [19] built upon the previous work and stated several problems in the previously available RCNN model. The training stage of RCNN was multistage, and CNN was applied to each object region proposal. The proposed fast RCNN model applied convolutions once to an image to extract image features. Further, the extracted image features were added to the convolution feature map with the use of an object region proposal. The results showed that the mean precision score increased from 4% to 6% in different datasets. However, the author did not consider the region proposal state, which could hamper the speed of the model.

In [20], the authors observed that the introduction of CNN and RCNN (regions with Convolutional Neural Networks) in visual recognition tasks was slow and had a lot of redundancy. To solve this problem, the authors came up with a new U-shaped architecture where they first used downsampling to decrease the dimensions of the image and obtain a well-defined context. They then upsampled it, which led to a large number of feature channels. The new architecture proved to be efficient as the error rate was least for the three types of data sets. However, the bottleneck structure, had slow learning process.

In [21], the authors realized that the region proposal state became a bottleneck region in the model. To solve this problem, instead of selective search to make proposals, the authors proposed to use RPN (region proposal network). RPN used an anchor box using predetermined anchor boxes of several distinct sizes. The image was fed to a convolutional layer in this model, and a feature map was extracted from it. The results on PASCAL VOC 2007 and 2012 datasets show that the proposed method decreased the run time by 1,630 milliseconds and increased the efficiency. However, the author did not consider the time complexity in this case.

In [22], the authors proposed an algorithm that made use of a CNN trained with the backpropagation algorithm and lifted wavelet transformation. This was a comparative analysis where the algorithm suggested by the authors was weighed against the feed-forward neural network. The study was carved into three parts where the first part applied a compression engine similar to the Karhunen-Loeve transformation [32], which worked as a multilayer ANN. The second method used a single neural network for the compression algorithm. The results showed an inverse relation between PSNR and sub-image block size and compression ratio but a direct relation with the neurons considered.

Liu and Zhu [23] trained a binary neural network with a high compression rate and high accuracy. To begin with, the authors considered adopting a gradient approximation in order to reduce the gradient mismatches that occurred during the forward and backward propagation. Later, multiple binarization was applied to the activation value, and the last layer was binarised. This subsequently improved the overall compression rate. The results showed that the compression rate improved from 13.5 to 31.6, and the accuracy increased by 6.2%,15.2%, and 1% when weighed against XNOR, BNN, and Bi-Real, respectively.

Nian et al. [24] proposed a method to reduce the compression artifacts by pyramid residual convolutional network (PRCNN). Three modules, such as the pyramid, RB, and reconstruction, were used to achieve the goal. The process started with the first convolution layer and gives an output of 64 feature maps which were sent to pyramid modules that involve down-sampling, RB (Residual Block), and upsampling. Later, the branches of R1(High-level), R2(Middle-level), and R3(Low-level) were generated to learn different level features. Further, RB was used to preserve the low-level features, and the results from the R1 branch were used for reconstruction. Although the results provided by the authors indicated that the method improved PSNR and SSIM, it reduced the visual quality.

Liu et al. [25] proposed an image compression algorithm based on the DWT-BP neural network. The authors took an original image, performed first-order waviest transform, and used the decomposed wavelet coefficient as the training set of the BP neural network that also served as the sample set of the output layer. Additionally, the output of the compressed data by the BP network was further quantized. The paper follows an experimental comparative analysis which showed that the algorithm suggested by the authors was 10.04% more effective in terms of compression than the traditional BP (Back Propagation) neural network.

In [26], the author realized that although image compression using SVD gives promising results, it is difficult to predict the correct value of rank K using which the image is compressed to maintain its quality. They proposed a study on three images containing two faces and one monument where they calculated the error value at different levels of K and compared the results. The results showed that rank K should be at least 30% of the size of the image, and an image could be compressed to 77.7% of its original size. However, the authors should have given due diligence to the number of focal points in an image which can play a huge role in predicting the compression ratio.

In [27], authors noticed that the requirement for image compression has increased with time, where the quality of the image is also maintained. In order to solve this problem, they used SVD, where they extracted the RGB matrix from the image and removed the redundant data. Finally, the three matrices were combined again to form a compressed image. This method was performed on two images in two different formats (jpg and png). The results showed that, on average, the images could be reduced by 28.5%. However, the experiment was performed only on two images of similar features. Hence, these results cannot be considered universal for all images.

5 The Dataset

The dataset used for this paper was created with over 1500 high-resolution mobile phone images collected from 18 different handsets from different kinds of users ranging in 10 different cities. To depict the real-life use case of the dataset, factors like randomness, image quality, the level of image enhancement, and image type were considered. The age group that was considered for this dataset was 16–24.

All the images collected were normalized, and the singular value graph and the cumulative singular value graph were studied for each image. The rank of each image is extracted from the graphs. Further, after performing compression, the PSNR values and compression ratios were extracted and stored.

6 Methodology

In the given system, we first created an efficient program that compressed the image in size while maintaining the psychovisual redundancy using singular value decomposition. Since the dataset created for this paper contained colored images, every image was separated into three matrices, namely red, green, and blue, and the mean of these matrices was calculated to perform SVD on the entire image instead of being calculated for each matrix separately. This was used to plot the singular value graph and cumulative singular value graph for each image using the diagonal matrix, which provided a relation between the energy of the image to the rank of the image. The relation helped us to determine the exact value of the rank at which the image could be compressed while maintaining psychovisual redundancy. Furthermore, the images were compressed using the determined rank value by again performing SVD on it. This time each matrix, i.e., red, green, and blue, was compressed separately and finally combined to generate a sharp colored image as output.[1] Performance parameters like PSNR, MSE, and Compression Ratio were calculated over the original and compressed image, and the data is stored for the dataset creation.

6.1 Relation between Energy Versus Rank Graph and The Number of Focal Points

The shape of the cumulative singular value graph or the energy to the rank of the image graph is similar to the logarithmic curve in the first quadrant. The energy of the image tends to increase rapidly for the first few rank values, and then this line becomes parallel to the x-axis after reaching the maximum of the energy value. Hence, as shown in Fig. 1, after a certain point, there is no change in the energy of the image with a change in the rank of the image. The point beyond which there is no change in the energy with the increase in the rank of the image is considered the rank value. This gives the highest compression ratio while maintaining psychovisual redundancy as the energy has already reached its maximum value for the image.

In order to achieve a maximum compression ratio, we derived a relation between the number of focal points in an image and the energy versus rank graph of the image. The relation states that the graph achieves the peak instantaneously, signifying a massive increase in the energy of the image for a small change in the rank of the image when the images have less number of focal points.

[1] Link to the Image Dataset and Algorithm: https://github.com/MadhavAvasthi/Image-compression-using-SVD-and-Deep-learning.

However, as the number of focal points increases, there is a gradual increase in the energy of the image and an increase in the rank of the image until it reaches its peak, after which the characteristics remain unchanged. This relationship served as the fundamental for selecting mobile images dataset as the images clicked by mobile phones have less number of focal points.

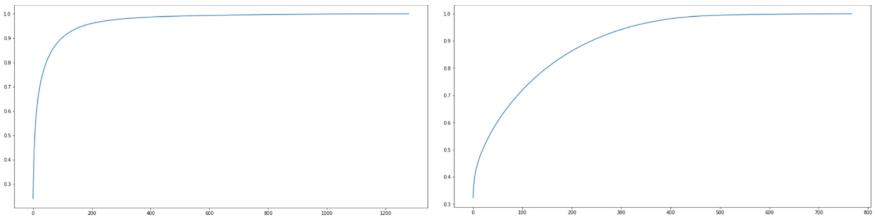


Fig. 1. (Left) Energy of the image versus Rank of the image graph for less number of focal points. (Right) Energy of the image versus Rank of the image graph for higher number of focal points.

6.2 Use of ResNet-50 to Predict the Rank

After developing an efficient algorithm for image compression using SVD and the derived relationship between the number of focal points and the energy versus rank graph, predicting the rank value still acted as a bottleneck to the algorithm as it would always require human intervention to evaluate the rank of the image using the graph. In order to automate the process, we used ResNet-50 to predict the rank of the image.

The input to our model is the actual rank values and the images from the dataset created. The images are provided as input to the pre-trained ResNet-50 with the pre-trained weights, as for this project, we have used transfer learning and then again given as input to ResNet-50 however, in order to keep the originality of the neural network, the first 15 layers are frozen before the input. The first few layers in neural network architecture are used to find the edges and shapes of the images in the given input dataset. The algorithm was tested with a batch size of 16, 32,64, and 100, and according to the results, the batch size of the images is kept to be 64.

7 Results and Discussions

We trained our model on the dataset created by us for the experiment. We believe that this dataset is better compared to any of the available online datasets as the images are not preprocessed, and the dataset provides a real-world replica of different types of mobile images. In order to maintain the psychovisual redundancy, we kept the PSNR values of the images higher than 80 in order to keep MSE

values minimum as PSNR and MSE values are inversely related. The images in the dataset used were of different kinds like landscape photography, portrait photography, selfies, fashion photography, travel photography, architecture photography, street photography, etc. The average PSNR value for the dataset is 93.74, stating that the MSE is negligible between the original and the compressed images. This helps in establishing pyschovisual redundancy between the two types of images. However, the average compression ratio for the entire dataset is 58.1% which means that the entire dataset has been reduced in size by 41.9%.

Further, in order to predict the rank of the image, ResNet-50 was used. We studied the effect of change of rank values on the image dataset and realized that a difference of 10% of rank would lead to a change of 4-5 % on compression ratio with a change of 1-2 PSNR values and a difference of 15% of rank value would lead to a change of 6-8% of compression ratio with a change of 2-5 PSNR values. Hence, for this paper, we realized that the accuracy of the regression problem cannot be measured as per the usual norms and therefore proposed our own method to do so.

According to the method, any image for which the difference between the actual rank and predicated rank value is less than 15% can be considered an accurate prediction as the change in the compression ratio and PSNR value is within a limit. Therefore, the accuracy of the overall algorithm is 86.35%. Hence, as shown in Fig. 2, the psychovisual redundancy is maintained for the image.



Fig. 2. (Left) Original Image. (Center) Compressed image with the actual rank of 320, PSNR=93.1 and CR=48.48. (Right) Compressed image with the predicted rank of 326, PSNR=93.37, and CR=47.56.

8 Conclusion and Future work

In this paper, we have presented an image compression algorithm using SVD and ResNet-50. The proposed algorithm used SVD to remove the unwanted singular values and re-construct the image while maintaining the quality of the image. Further, the rank for each image was collected, which was used to train the

ResNet-50 model that can predict the rank of the image, thereby automating the entire procedure while maintaining pyschovisual redundancy. This enables us to compress any image clicked by a mobile phone and compress it to nearly 42% without losing the quality of the image visible to the naked human eye. This can enable the user to enhance the memory efficiency of their phones by performing the compression while maintaining the quality of the image. This can help the user to decrease their dependency on cloud storage like google drive, Microsoft one drive, etc., thereby increasing the data privacy for the user. Further, the accuracy of the regression algorithm can be improved in the future, while the relation between energy versus rank graph and the number of focal points can also be used and studied with various kinds of datasets.

References

1. Pereira, G.: In: Schweiger, G. (ed.) Poverty, Inequality and the Critical Theory of Recognition, vol. 3, pp. 83–106. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45795-2_4
2. Rahman, M.A., Hamada, M., Shin, J.: The impact of state-of-the-art techniques for lossless still image compression. Electronics **10**(3), 360 (2021)
3. Bovik, A.C.: Handbook of Image and Video Processing. Elsevier Academic Press (2005)
4. Li, C., Bovik, A.C.: Content-partitioned structural similarity index for image quality assessment. Signal Process.: Image Commun. **25**(7), 517–526 (2010)
5. Patel, M.I., Suthar, S., Thakar, J.: Survey on image compression using machine learning and deep learning. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (2019)
6. Vaish, A., Kumar, M.: A new image compression technique using principal component analysis and Huffman coding. In: 2014 International Conference on Parallel, Distributed and Grid Computing (2014)
7. Sandeep, G.S., Sunil Kumar, B.S., Deepak, D.J.: An efficient lossless compression using double Huffman minimum variance encoding technique. In: 2015 International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT) (2015)
8. Dasgupta, A., Rehna, V.J.: JPEG image compression using singular value decomposition. In: International Conference on Advanced Computing, Communication and Networks, vol. 11 (2011)
9. Babu, K.A., Kumar, V.S.: Implementation of data compression using Huffman coding. In: 2010 International Conference on Methods and Models in Computer Science (ICM2CS-2010) (2010)
10. Vaish, A., Kumar, M.: A new image compression technique using principal component analysis and Huffman coding. In: 2014 International Conference on Parallel, Distributed and Grid Computing (2014)
11. Rufai, A.M., Anbarjafari, G., Demirel, H.: Lossy medical image compression using Huffman coding and singular value decomposition. In: 2013 21st Signal Processing and Communications Applications Conference (SIU) (2013)
12. Bano, A., Singh, P.: Image encryption using block based transformation algorithm. Pharma Innov. J. (2019)
13. Erickson, B.J., Korfiatis, P., Akkus, Z., Kline, T.L.: Machine learning for medical imaging. RadioGraphics **37**(2), 505–515 (2017)

14. Narayan, S., Page, E., Tagliarini, G.: Radiographic image compression: a neural approach. Assoc. Comput. Mach. 116–122 (1991)
15. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. Proc. Eur. Conf. Comput. Vision, Sep. **2014**, 818–833 (2014)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009)
17. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. (2012)
18. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014)
19. Girshick, R.: Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (2015)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
22. Shukla, S., Srivastava, A.: Medical images Compression using convolutional neural network with LWT. Int. J. Mod. Commun. Technol. Res. **6**(6) (2018)
23. Liu, S., Zhu, H.: Binary convolutional neural network with high accuracy and compression rate. In: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (2019)
24. Nian, C., Fang, R., Lin, J., Zhang, Z.: Artifacts reduction for compression image with pyramid residual convolutional neural network. In: 3rd International Conference on Video and Image Processing (ICVIP 2019). Association for Computing Machinery, pp. 245–250 (2019)
25. Liu, S., Yang, H., Pan, J., Liu, T.: An image compression algorithm based on quantization and DWT-BP neural network. In: 2021 5th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2021). Association for Computing Machinery, pp. 579–585 (2021)
26. Halim, S.A., Hadi, N.A.: Analysis Of Image Compression Using Singular Value Decomposition (2022)
27. Abd Gani, S.F., Hamzah, R.A., Latip, R., Salam, S., Noraqillah, F., Herman, A.I.: Image compression using singular value decomposition by extracting red, green, and blue channel colors. Bull. Electr. Eng. Inform. **11**(1), 168–175 (2022)
28. Campardo, G., Tiziani, F., Iaculo, M.: Memory Mass Storage, 1st edn. Springer, Heidelberg (2011)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
30. Rudberg, M.K., Wanhammar, L.: High speed pipelined multi level Huffman Decoding. In: IEEE International Symposium on Circuits and Systems, ISCA' 7 (1997)
31. Nair, V., Hinton, G.: Rectified Linear Units Improve Restricted Boltzmann Machines. ICML (2010)
32. Pratt, W.K.: Karhunen-Loeve transform coding of images. In: Proceedings of 1970 IEEE International Symposium on Information Theory (1970)
33. Li, C., Li, G., Sun, Y., Jiang, G.: Research on image compression technology based on Bp neural network. In: 2018 International Conference on Machine Learning and Cybernetics (ICMLC) (2018)



Optimization of Traffic Light Cycles Using Genetic Algorithms and Surrogate Models

Andrés Leandro(✉) and Gabriel Luque

ITIS Software, University of Malaga, Malaga, Spain
{arleandroc,gluque}@uma.es

Abstract. One of the main ways to solve the traffic problem in urban centers is the optimization of traffic lights, which has no trivial solution. A promising approach includes using metaheuristics to obtain accurate traffic light schedules, but calculating the quality (i.e., fitness) of a traffic network configuration has a high computational cost when high precision is required. This study proposes using surrogate models as an efficient alternative to approximate the fitness value without significantly reducing the overall quality of results. We have implemented a multi-step process for evaluating candidate surrogates, which includes validating results in a traffic network instance. As a result, we have identified several configurations of surrogate models that considerably improve the time to calculate the fitness with competitive estimated values.

Keywords: Genetic algorithms · Surrogate Models · Traffic Light Signals

1 Introduction

One of the main problems in city centres is vehicular traffic, which causes a general loss in quality of life and increases pollution [11]. The complexity of large-scale urban planning makes solutions to this problem non-trivial and challenging. A particularly successful approach to solve this issue is synchronizing traffic lights so that vehicle traffic flow is maximized and the time they're still is minimized.

Traffic design engineers frequently use simulators and optimization techniques to improve traffic systems. This paper uses metaheuristics to generate traffic light network configurations in combination with a microscopic simulator (specifically, one called "Simulation of Urban Mobility", SUMO [6]) which estimates the quality of those configurations. However, the use of this tool carries a high computational cost (requiring, in realistic scenarios, hundreds of hours [8]).

This research is partially funded by the Universidad de Málaga; under grant PID 2020-116727RB-I00 (HUmove) funded by MCIN/AEI/10.13039/501100011033; and TAILOR ICT-48 Network (No 952215) funded by EU Horizon 2020 research and innovation programme.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 607–617, 2023.
https://doi.org/10.1007/978-3-031-27409-1_55

Although there are several approaches to mitigate this situation, one way to reduce resource consumption is the implementation of surrogate models, or metamodels, which reduce the time it takes to evaluate a candidate configuration [7]. These models closely approximate the objective function but are less resource-intensive. While there has been some research on the use of surrogate models with metaheuristic algorithms applied to traffic analysis [5], many of these works don't directly compare the performance of the surrogate with the model when estimating the fitness (i.e. quality) of a configuration. The main contribution of this work is the appraisal of several possible surrogate models. It includes a statistical comparison of their estimation errors along with an empirical validation of the final selection, followed by a performance evaluation against the simulator in the same context; that is, calculating the fitness of candidate solutions in a Genetic Algorithm (GA). We are using this technique since it has obtained promising results in the past [4].

The rest of this paper is structured as follows: Sect. 2 describes the approach to using metaheuristics for traffic analysis, emphasizing the use of SUMO and its possible drawbacks. Section 3 establishes the methodology we followed to incorporate surrogate models in this approach. Section 4 presents and discusses the results of applying said methodology. Section 5 ends with a series of conclusions and gives some options for future research.

2 The Scheduling Traffic Lights Problem

The flow of vehicles in a specific city is a complex system, mainly coordinated by traffic light cycles. This paper tackles the optimal scheduling of all traffic lights in a specific urban area. The formulation is based on the proposals of García-Nieto et al. in [4]. The mathematical model is pretty straightforward, codifying each individual as a vector of integers, where each integer represents the phase duration of a particular state of the traffic lights in a given intersection.

Along with the duration of each phase, this model also considers the time offset for each intersection. Traffic managers use this value to allow synchronization between nearby junctions, a key factor in avoiding constant traffic flow interruptions on central routes. This change allows the modelling of more realistic scenarios. Still, it increases the problem's complexity since the number of decision variables grows in proportion to the number of intersections.

Once a solution is generated with a particular approach, evaluating its quality is necessary. For this, we have selected the software SUMO [6] to obtain the base data that can be used to calculate each solution's fitness. After the simulation, some of the statistics output by SUMO are combined on the objective function presented in Eq. 1. This function was proposed and has been used in other works researching the problem of traffic light scheduling [4].

$$f_{\text{obj}} = \frac{T_{\text{trip}} + T_{\text{sw}} + V_{\text{NR}}\, T_{\text{sim}}}{V_{\text{R}}^2 + P} \tag{1}$$

where T_{trip} is the total travel time for all vehicles. T_{sw} represents the total time vehicles are still. V_{R} is the number of vehicles which arrived at their destination,

and V_{NR} are the vehicles that didn't arrive at their destination by the given max time, T_{sim}, of a simulation. Finally, P is the ratio for the duration of green traffic lights versus red ones. We should note that values to minimize are in the numerator, while those to maximize are on the denominator. With this, the problem becomes a minimization task.

Since Eq. 1 must be computed for each candidate solution, it's necessary to run a full SUMO simulation for each solution generated by the genetic algorithm. This makes the consumption of computational resources spike massively and it becomes essential to propose ways that reduce the time required for fitness evaluation without significantly decreasing the quality of the calculations, such as the one presented in this paper.

3 Experimental Methodology

Next, we describe the following aspects of the experimental design: the dataset, the models used, the model selection process and how it was incorporated into the algorithm flow.

3.1 Dataset

As part of the experimental process, we worked with an instance of SUMO that had two intersections and 16 phases in total (both values set the size of the search space). The reduced size of this instance allows us to make a more detailed analysis. Although the scenario is small, the search space is multimodal and has approximately 52^{16} potential solutions, which can cause difficulties for many optimization techniques. As discussed, the solutions are integer vectors, where each value indicates the duration of one phase of the traffic lights at an intersection. The value of the elements also varies according to the configuration of the phase: they are generally between 8 and 60 seconds. However, some particular phases are fixed values (4, or multiples of it).

A *Latin Hypercube Sampling* (LHS) mechanism [9] was used to obtain a dataset from this instance. In total, N samples were generated. That value arises from $N = P + I$, where P is the population maintained by the genetic algorithm and I is the maximum number of iterations to run. Then, this value is equivalent to the number of solutions that would be generated during one run of the GA.

We also tested the hypothesis presented by Bhosekar [1] that increasing the number of samples used to train the surrogated model improves the results obtained. Testing that hypothesis in this scenario was considered relevant due to the high time required to run SUMO. Using N as the optimal value, additional datasets were generated for $\frac{N}{2}$, $\frac{N}{10}$ and $\frac{N}{20}$, to verify whether the quality of the models degraded with a lower number of samples.

3.2 Surrogate Models

We selected three different candidate models: an RBF model [2], a kriging (KRG) one [2] and one based on least-squares approximation (LS) [10]. The first two

models were chosen for their successful use in previous research on related problems, while the last one was mainly selected for its simplicity and training speed.

Since the parameter values for each model can affect its performance, we followed Forrester's suggestions [3] and tried several variants of each model, changing their parameterization and comparing the quality of their prediction. These variants were also tested with the datasets of different sizes.

The RBF model approximates a multivariable objective function with a linear combination of radial-basis functions, which are singular and univariable functions. The main benefit of this model type is its quick training and prediction of new values. For this model, the following hyperparameters variants were tried: `d0`, the scaling coefficient (the values tried were: 1.0, 2.0 and 4.0); `poly_degree`, which indicates if a global polynomial is added to the RBF, and its degree (the variants were not adding one, adding a constant and adding a linear polynomial), and `reg`, the regularization coefficient for the values of the radial-basis function (tested values were: 1e-10, 1e-11 and 1e-09).

The kriging model owes its name to Danie G. Krige, and is also based on a Gaussian function but, in this case, it's used to calculate the correlation for a stochastic process. For this model, the parameters considered to be tested were: `corr`, the correlation function (tested values were: `squar_exp`, `abs_exp` and `matern52`); `poly`, a deterministic function added to the process (tested values: `constant`, `linear` and `quadratic`), and `theta0`, the starting value of θ used by the correlation functions (tested values: 1e-1, 1e-2 and 1e-3).

Finally, the LS model adjusts the coefficients of a simple linear function. Although this model is less accurate [2], we considered it for this paper in order to evaluate its performance against more elaborate models like RBF and KRG, since the model is so simple that its execution is extremely fast. The LS model has no parameters, so there were no variants.

3.3 Model Selection

Before the genetic algorithm validation, we followed a process to select the metamodels according to their prediction accuracy. For this, 220 variants were tested considering the parameterization of the models and the dataset size. In concrete, we used 108 variants for the kriging model, 108 for the RBF and 4 for the LS.

To evaluate the accuracy of predictions made by a model, we followed the suggestions made by Forrester [3], using a k-fold Cross Validation process with $k = 10$, and mean squared error (MSE) [9] as a measure of the prediction errors. Afterwards, we used a non-parametric statistical test named Kruskal-Wallis (KW) to evaluate the differences between the variants and if they were statistically significant (with p-value <0.05). To complete the selection process, we followed the next steps:

1. For the MSE sets which reject the null hypothesis, we discard the variants with the higher average MSE.
2. For each remainder variant, we selected one for each type of surrogate with the lowest average MSE.

3. We applied a graphical analysis to these models (using a newly generated dataset). The models that were found invalid were discarded.

The final models were used for experimental validation using the genetic algorithm, a process described in the next section.

3.4 Integration in the Genetic Algorithm

We defined four different ways of using the simulator or the metamodels:

1. Only using SUMO. This configuration presents a baseline of efficiency to surpass and a goal line of quality to achieve.
2. Only using the surrogate. The final solution was recalculated with SUMO to obtain the real fitness.
3. Starting the search with the surrogate and changing the fitness calculation method to SUMO in the iteration $i_s = \frac{I}{2}$. The objective was to test if it was possible to determine a middle ground in terms of efficiency and quality between only using SUMO and only using the surrogate for fitness evaluation.
4. Similar to the previous one, but this one starts the search using only SUMO and then switching to the metamodel.

Since executing a GA is a stochastic process, we used ten seeds for the random number generator to establish a more robust statistical basis for our analysis.

4 Experimental Results

This section presents the results obtained. The code was developed in Python using the Surrogate Modeling Toolbox library [2]. Tests were done on a Windows 10 computer with a Core i5-6600K CPU, 16 GB RAM and 240 GB SSD.

4.1 Dataset

We started by generating the datasets used to train the models. Based on the methodology, the baseline number of samples to generate is $N = 10,050$, and three variants of 500, 1000 and 5000 samples were also created in order to validate the quality of models trained with them. Table 1 shows a summary of the generated datasets. We can see that reducing the number of samples improves the median's quality but also increases the deviation (however, the changes are not significant).

4.2 Model Selection

This section details the process followed to evaluate the candidate metamodels and their variants to select those to be assessed empirically with the GA.

Table 1. Statistics for the fitness values of the sampled solutions.

N	Median	Standard Deviation
500	1.26459	1.03156
1000	1.26250	1.02129
5000	1.27057	1.00747
10050	1.28241	1.00243

Preliminary Analysis To select the most suitable surrogate, we had to evaluate all variants (220 combinations). However, since the validation of each combination is very time-consuming, we ran preliminary executions to identify invalid variants. This allowed a reduction of the total combinations to 112.
Numerical Analysis. The main factor in selecting the model is its accuracy (measured using the MSE). Besides quality (see Table 2), we also registered the time it took to train each model. This was done to evaluate the impact of parameter choice and dataset size in the time it takes to prepare a surrogate (see Table 3).

Table 2. Identified variant clusters in CV tests.

Surrogate	N	Average MSE	Average Std. Dev.
KRG	500	2.14477	0.03737
KRG	1000	1.80102	0.00668
KRG	5000	1.93886	0.03679
LS	500	1.72522	N/A
LS	1000	1.47682	N/A
LS	5000	1.53926	N/A
LS	10050	1.54074	N/A
RBF(1)	500	1.19029	0
RBF(1)	1000	0.93834	0
RBF(1)	5000	1.00288	0
RBF(1)	10050	0.98769	2E-5
RBF(2)	500	3.83887	0
RBF(2)	1000	3.41965	1.49998E-05
RBF(2)	5000	3.63170	0.00043
RBF(2)	10050	3.56462	0.00188

Before discussing these results, we must notice that the most relevant result in our statistical analysis was that the only sets which reject the null hypothesis with p-value < 0.05 were the RBF models when their parameters are changed. This means at least one parameter makes significant changes in the MSE, which

Table 3. Training Time Stats by Model and Dataset Size.

Surrogate	N	Median (s)	Std. Deviation (s)
KRG	500	169.739	41.211
KRG	1000	688.054	122.924
KRG	5000	26829.037	4654.559
LS	500	0.023	0.0
LS	1000	0.058	0.0
LS	5000	0.179	0.0
LS	10050	0.511	0.0
RBF	500	0.512	0.283
RBF	1000	0.906	0.900
RBF	5000	21.564	135.069
RBF	10050	87.261	698.185

will be determined by the rest of the analysis. The remainder combinations cannot reject the null hypothesis with enough certainty, so we can conclude that a given choice of parameters or dataset size will not drastically affect the quality of the predictions. These results support the decision to discard the configurations with the highest training times to focus on others with a slightly higher MSE but that are much quicker to train.

As we said, Table 2 shows a summary of the prediction quality for each model. The rows for the LS model do not have standard deviation because there's only one element in each group. As we can see, most models are grouped by the number of samples used to train them, with the variants in parameter values being of little influence in the final results (the statistical tests also support this). A notable exception is the RBF model, where the `poly_degree` parameter causes a massive difference in the results. The "RBF(1)" cluster represents all the variants with `poly_degree = 0` and the "RBF(2)" those with `poly_degree = -1`. The RBF(1) cluster shows the lowest MSE values for all evaluated metamodels. Verifying this result was important since this surrogate has the second fastest training times. This verification is presented later in this section.

Another interesting result is the trend where all models trained with $N = 1000$ samples produced the lowest MSE for that surrogate. This dataset size seems to be a tradeoff point for the surrogates, giving enough values to closely approximate the objective function but not too many to over-fit the metamodel and reduce its generalisation ability.

Visual Analysis. Next, we validated the behaviour of the models using a graph. Figure 1 has predictions by the lowest MSE variants of each surrogate type, trained with $N = 1000$, for a new dataset generated with LHS. The solutions have been sorted by the fitness calculated by SUMO (the blue line) to easily visualise differences in the calculated values.

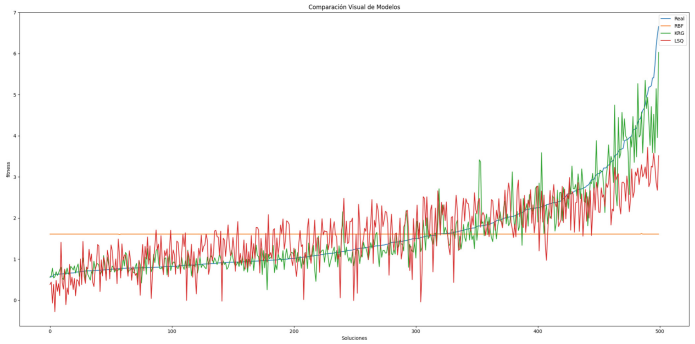


Fig. 1. Visual comparison of predictions ($N = 1000$).

The main result is that the most promising model, the RBF (orange line), is incapable of predicting new fitness values, returning a constant value. We hypothesise that it is caused by over-fitting during the training, making it incapable of generalisation. So, this model was discarded.

The LS model, the one with the second lowest MSE, seems to follow the trend of actual fitness more closely, but there is a significant deviation at many points of the graph, especially for the most extreme values.

Finally, despite having the highest MSE, the kriging model approximates the fitness calculated by SUMO very closely, especially in extreme fitness values. This behaviour can be a positive feature since better approximations for high fitness values help the model escape non-promising regions of the solution space much faster. Also, a better approximation of low fitness values allows the model to exploit the promising ones.

We verified these results with other graphs using different surrogate variants (not shown here because of space constraints). The general behaviour of the metamodels is the same as that shown in Fig. 1.

Final Selection. Considering the behavior observed in Fig. 1, all variants of the RBF model are discarded from being used with the genetic algorithm. We selected the following representatives of the LS and KRG models for experimental validation with the GA (the statistical tests show that none of the variants for these surrogates is significantly better than the rest):

- LS: $N = 1000$. No parameters.
- KRG: $N = 1000$, `corr = 'matern52', poly = 'constant', theta0 = 0.1`

4.3 Models Integration into Genetic Algorithm

This next section details the use of the candidate metamodels as part of the genetic algorithm as presented in Sect. 3.4. For each surrogate, we established the following three configurations for their use in the GA:

1. Estimating the fitness using only the surrogate.
2. Starting with SUMO and changing to the surrogate during the execution.
3. Starting with the surrogate and changing to SUMO during the execution.

As a control value for the comparison, we included a configuration using only SUMO to calculate the fitness. Therefore, we have seven configurations to test and 70 individual executions of the GA (because of the ten seeds used). For each execution, we registered the fitness of the best solution and the time it took to find it. Table 4 shows the median and standard deviation for each configuration.

Table 4. Final Fitness found and time to find it for every configuration of the GA.

Configuration	Final Fitness	$\sigma_{fitness}$	Time (s)	σ_{time} (s)
Only SUMO	0.41856	0.00477	749.483	482.423
Only LS	4.78811	0.48140	0.105	0.022
Only KRG	0.58588	0.10626	13.199	3.132
SUMO → LS	2.89284	0.06902	1418.557	65.470
SUMO → KRG	0.59083	0.06817	1427.703	269.100
LS → SUMO	0.41939	0.00606	652.356	268.599
KRG → SUMO	0.41928	0.00505	504.671	321.125

The LS-based surrogate predictions have the lowest quality by far, approximately 11 times larger than the SUMO proposals and 1.5 times larger than the second-worst configuration. Despite being the most efficient (almost 7,000 times faster to find the best solution than using SUMO alone), the massive disparity in results makes it unsuitable as a surrogate for the simulator. As expected, the model's simplicity prevents it from being able to accurately predict new values in distributions that do not follow a linear model.

The KRG surrogate also substantially improves execution time, being almost 57 times faster than SUMO. However, it presents a reduction in efficiency of approximately 40% with respect to the calculation with SUMO alone. Although this loss is significant, the solution has a fitness in the same order of magnitude as that offered by the algorithm only using SUMO and reaching it with a much reduced time. Therefore, this model could be a good candidate if a fast system is needed with a moderate reduction in quality.

Hybrid configurations starting with SUMO and then switching to the metamodel have predicted values which are still significantly worse in both cases than the calculated ones by SUMO. Although this does not make them attractive as potential substitutes, their efficiency rules them out: it takes almost twice as long to find the best solution. To explain this behaviour, we recorded the i_{best} iteration in which the final solution was obtained (see Table 5).

We can see that SUMO found the best solution much earlier than the hybrid configurations. A possible explanation is that the surrogates' approximation error misleads the search. It not only reduces the quality of the final solution

Table 5. Average i_{best} for the configurations.

Configuration	$\overline{i_{best}}$
Only SUMO	2751.7
SUMO → LS	5575.1
SUMO → KRG	9751.5

(which has lower real fitness than the found by the Only SUMO configuration) but also increases the total runtime since the GA keeps identifying "better" solutions (even though the actual best solution was found previously and discarded by the surrogate). Because of this, we can conclude that hybrid configurations where the GA execution begins by using SUMO and changing to the surrogate are inadequate for traffic network instances where SUMO can find the best solution before using the surrogate, so the latter can't provide much help.

Conversely, the configurations starting with the surrogate and switching to SUMO have a similar quality to the Only SUMO configuration (0.01% deviation) and produce substantial efficiency improvements (13% for LS → SUMO and 33% for KRG → SUMO). This behaviour is because models can approximate the fitness well enough to reduce the search space to a smaller region, which SUMO can easily exploit. Since the surrogate models execute much faster, this search space reduction also happens in less time. Taking this into account, these hybrid approaches can provide optimizations that might be interesting to explore further to reduce the required time for the traffic simulator.

In our previous analyses, we do not consider the effect of each model's training time, which is significant for some of them. However, it's important to note that the improvement in time is reported for every execution of the GA. A GA requires many executions for tuning the parameters or gathering statistical data, since it's a non-deterministic process and several runs are mandatory. Since the training process is only executed once, the larger the number of runs of GA, the smaller the effect of training time.

5 Conclusions and Future Works

Using metaheuristics with traffic simulators has provided handy tools to approach the traffic light optimization problem. However, it has introduced other issues due to the high time required by the simulator. To address this problem, this paper proposes using surrogate models to approximate the value of the simulator and describes the approach to train, select and evaluate such models in comparison with the simulator in a genetic algorithm.

A statistical evaluation was performed in conjunction with a graphical analysis for the model selection process. This analysis identified that some surrogates which could be promising due to their low error, such as RBF, were unsuitable for use because the predicted values deviated too much from the actual values. This type of analysis can be valuable to complement other error measures.

Experimental validation with the genetic algorithm showed that a metamodel could be much faster than a simulator in estimating fitness values. Still, this speed represents little if the proposed value is not close enough to the simulator's ones. With this in mind, one approach that may offer a good middle ground is a configuration where the surrogate and the simulator alternate, with one narrowing the search space and the other refining the search. These approaches proved to be able to approximate the actual value relatively well and reduce the computation time very significantly.

This work opens multiple research lines. First, the results should be validated in other larger instances. Then there are many possibilities, such as using other techniques, other surrogate models or modifying the interactions between them (e.g. retraining the model with some of the new solutions generated during the search).

References

1. Bhosekar, A., Ierapetritou, M.: Advances in surrogate based modeling, feasibility analysis, and optimization: a review. Comput. Chem. Eng. **108**, 250–267 (2018)
2. Bouhlel, M.A., Hwang, J.T., Bartoli, N., Lafage, R., Morlier, J., Martins, J.R.R.A.: A Python surrogate modeling framework with derivatives. In: Advances in Engineering Software p. 102662 (2019)
3. Forrester, A., Sobester, A., Keane, A.: Engineering Design Via Surrogate Modelling: A Practical Guide. Wiley, New York (2008)
4. García -Nieto, J., Ferrer, J., Alba, E.: Optimising traffic lights with metaheuristics: reduction of car emissions and consumption. In: IJCNN, pp. 48–54 (2014)
5. Liang, Y., Ren, Z., Wang, L., Liu, H., Du, W.: Surrogate-assisted cooperative signal optimization for large-scale traffic networks. Knowl.-Based Syst. **234**, 107542 (2021)
6. Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E.: Microscopic traffic simulation using SUMO. In: The 21st IEEE International Conference on Intelligent Transportation Systems. IEEE (2018)
7. Osorio, C., Bierlaire, M.: A simulation-based optimization framework for urban transportation problems. Oper. Res. **61**(6), 1333–1345 (2013)
8. Segredo, E., Luque, G., Segura, C., Alba, E.: Optimising real-world traffic cycle programs by using EC. IEEE Access **7**, 43915–43932 (2019)
9. Tong, H., Huang, C., Minku, L.L., Yao, X.: Surrogate models in evolutionary single-objective optimization: a new taxonomy and experimental study. Inf. Sci. **562**, 414–437 (2021)
10. Wang, S., Liu, Y., Zhou, Q., Yuan, Y., Lv, L., Song, X.: A multi-fidelity surrogate model based on moving least squares: fusing different fidelity data for engineering design. Struct. Multidiscip. Optim. **64**(6), 3637–3652 (2021). https://doi.org/10.1007/s00158-021-03044-5
11. Zhang, K., Batterman, S.: Air pollution and health risks due to vehicle traffic. Sci. Total Environ. **450–451**, 307–316 (2013)



The Algorithm of the Unified Mechanism for Encoding and Decoding Solutions When Placing VLSI Components in Conditions of Different Orientation of Different-Sized Components

Vladislav I. Danilchenko(✉), Eugenia V. Danilchenko, and Viktor M. Kureychik

South Federal University, Taganrog, Russia
{vdanilchenko,lipkina,vmkureychik}@sfedu.ru

Abstract. An approach to solving the problem of placing components on a crystal based on a Genetic Algorithm (GA) with a mechanism for encoding and decoding chromosomes is described, which allows creating an algorithmic environment in the field of genetic search for solving the problem of placing VLSI components with a mechanism for encoding and decoding chromosomes. The purpose of this work is to find ways to place components on a crystal based on the bioinspired search theory and the mechanism of encoding and decoding of solutions. The scientific novelty lies in the development of a modified genetic algorithm for bioinspired automated design, where the modification consists in the application of a new encoding and decoding mechanism in the conditions of different orientation of different-sized components. The statement of the problem in this paper is as follows: optimize the bioinspired search for the placement of VLSI components. To do this, a new encoding and decoding algorithm is proposed for components with different sizes. The fundamental difference from the known approaches is in the application of the modified mechanism of encoding and decoding of solutions in bioinspired automated design. It is worth noting that the work uses a mechanism for the separation of components on a crystal, based on a modified GA, taking into account a combination of criteria and the total length of the interconnects and the area of the crystal. Thus, the problem of creating modified algorithms and software for automated placement of VLSI components is currently of particular relevance.

Keywords: Genetic algorithms · Evolutionary calculations · Automated design systems · Schematic diagram · Topology · Coding solutions · Decoding solutions · Optimization algorithms

1 Introduction

The rapid development of technologies increases the importance of the modeling process in computer-aided design systems. The peculiarity of CAD matrix VLSI is the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 618–624, 2023.
https://doi.org/10.1007/978-3-031-27409-1_56

widespread use of macro modeling and decomposition methods, since the hierarchical structure of the design object contains a large number of such components. Moreover, even at the lower level there are logical components that represent quite large groups of connected devices. It is important that the electrical parameters and characteristics of logic components are determined only during their development and are repeatedly used during the modeling of a wide range of VLSI [1].

Component is understood as an element or subsystem of the designed circuit. Successful VLSI design, CAD programs should cover the modeling of devices, circuits and logic, placement and tracing, verification and test generation.

Modern CAD systems are complex and provide design up to the system level [2–4]. The advantage of designing matrix circuits is manifested in the fact that device modeling is used only at the stages of developing the crystal design and logic element circuits. To solve these problems, it is necessary, as noted, to switch to hierarchical design.

When the VLSI system design is completed, its structure and functional specifications for each structural unit are established. Then the logical design of each block is performed and the results are entered into the VLSI design database. At the same time, standard circuit components can be developed using circuit modeling, which is entered into the component library in the database.

When the logic circuit is detailed, the logic is verified (verified) by logic and circuit modeling programs. Then the procedures of automated topology synthesis are performed. The variants of the scheme obtained during topological design, after clarifying the parasitic parameters, are evaluated using circuit and logic modeling programs.

The calculated values of the criterion exceed the specified value, the topological parameters are changed and the procedures of the logical design are repeated. The cycle repeats until the specified requirements are met and the specified criteria vector is taken into account.

The main requirement when placing components on a crystal is to create conditions that take into account the criteria for the total length of the joints and the area of the crystal. To fulfill this requirement, it is necessary to take into account the characteristic properties of the crystal structure, which have a significant impact on the results of the formation of the VLSI topology.

Modern systems on a chip contain a large set of complex multi-dimensional components with different types of orientations on the crystal. Such components with non-standard shapes cannot be assembled into regular rows or matrices, they require different-sized placement, which creates difficulties in using the VLSI matrix type [1–6].

2 Optimal Search Criteria

The process of solving the problem of placing VLSI components is the transformation of an electrical circuit consisting of components with predefined input parameters and interconnections into a topology with specified geometric positions in the matrix structure of the block. In this case, the criteria for the effectiveness of solutions are: the number of the total length of the interconnects and the area of the crystal. These criteria are indirectly related to each other. As part of the task of placing VLSI components, a

vector of criteria is considered, where the criterion of the total length of the interconnects and the crystal area tends to a minimum, which is confirmed by the formulas [2–4].

$$\lim_{K_l \to 0} f(K_l) = P_l, \tag{1}$$

$$\lim_{K_p \to 0} f\left(K_p\right) = P_p, \tag{2}$$

where K_l—criterion for the total length of interconnects, K_p—crystal area, P_l and P_p—limits of functions $f(K_l)$ and $f\left(K_p\right)$ or the criterion for the completion of the algorithm.

These criteria are equivalent and they can be combined into one minimization function [5].

$$f = f(K_l) + f\left(K_p\right), \tag{3}$$

where f—objective function.

3 Encoding and Decoding Mechanism

Mismatched components with irregular placement do not always fit in size to each other, while part of the crystal area is not used. Within the framework of such a combination of criteria, it is necessary to comply with the rules: modules should not intersect, they should be located within the crystal field.

To find the optimum among alternative solutions in polynomial time, various iterative strategies are used, which make it possible to obtain an optimal or quasi-optimal solution. It is worth noting that this is often quite enough. Such heuristic algorithms produce iterative improvement of the objective function with various initial constraints and water data. Which leads to the formation of the necessary solution in one of the alternative placement sets.

Algorithms built on the basis of natural systems (genetic, bioinspired, etc.) require intensive calculations. However, the genetic algorithm has the potential to reduce the required amount of computational costs. A genetic algorithm capable of processing large sets of solutions (populations). This process is well parallelized to speed up the calculation process. This leads to the fact that iterative methods work in a narrow area, and the genetic algorithm allows you to expand the corresponding search space and place the necessary emphasis on priority search areas. The effectiveness of a genetic algorithm depends on the correct choice of genetic operators, the encoding and decoding mechanism, and a set of input parameters. In this regard, the paper proposes to apply a new mechanism for encoding and decoding solutions to the structure of the genetic algorithm.

Consider the method to encode data within the task of placing components on a chip. Currently, various methods are used to encode the location of components on the work field. For example, binary trees are used to describe the position of blocks on regular grids, in which the leaves are blocks and each internal node defines a vertical or horizontal connection with its two descendants [6–8].

The results of modern research show that the description of placement in matrix structures is relatively studied, but the description of the placement of different-sized

components is a rather complex process, which leads to the need to develop new algorithms and modify existing ones. The paper uses a directed diagonal tree of representation of schemes in the framework of the description of the placement of different-sized VLSI components [7–9].

A tree is a graph G = (K, E), which has vertices K—components, edges E—interconnections between two components. Directed edges E form pairs—parent and descendant. Each pair of parents has the possibility to have no more than two descendants: the left edge P_l and the right P_r. The coding chromosome is an object description of the hierarchy of the tree based on links, which is shown in the Fig. 1. Each vertex of the tree under consideration has the possibility to have no more than two descendants, which is due to the type of binary tree. This is due to the peculiarities of the topology construction procedure. In this tree it is necessary to explicitly encode both descendants: P_l and P_r. It is worth noting that in the absence of one of the descendants, it is necessary to determine which descendant is implemented, since they may not be equivalent. The classical encoding and decoding mechanism do not allow taking into account such a descendant characteristic.

The components placed on the working field of the crystal have rectangular shapes with different sides. In this regard, it is necessary to determine how the component should be positioned—horizontally or vertically. To describe this position of the component, the coding bit is used in the work: 1—vertically, 0—horizontally. It is worth noting that for the criterion of the total length of the interconnects, both the connections of the central contacts of the components and the contacts at the borders or inside each component are taken into account. In this case, it is necessary to encode not two possible positions of the component, but four, in other words, it is necessary to use two bits, as shown in the Fig. 2.

The coding chromosome contains a list of components, each of which has a unique name and two address links to its descendants. It is worth noting that also, each component must store a location bit. The information accompanying each component, such as overall dimensions, is not included in the coding chromosome, because it does not affect the configuration of the main tree. Therefore, this information is stored in a single copy.

Thus, to encode a directed tree with n components, the minimum required number of Q bits, the calculation of which can be described by the formula [9–12].

$$Q = 3n\left\lceil \log_2 n \right\rceil + 2n = n(3\left\lceil \log_2 n \right\rceil + 2). \tag{4}$$

4 Example of the Encoding and Decoding Mechanism

An example of the proposed encoding and decoding mechanism and based on it, the algorithm for building a tree is shown in the Fig. 3.

It is worth noting that the coordinates of the free space are determined not only by the contour, but also by the dimensions of the component in question [17–20]. At the same time, the time complexity of the described algorithm for converting a tree into a topology has a linear order of $O(n^2)$, where n is the number of components.

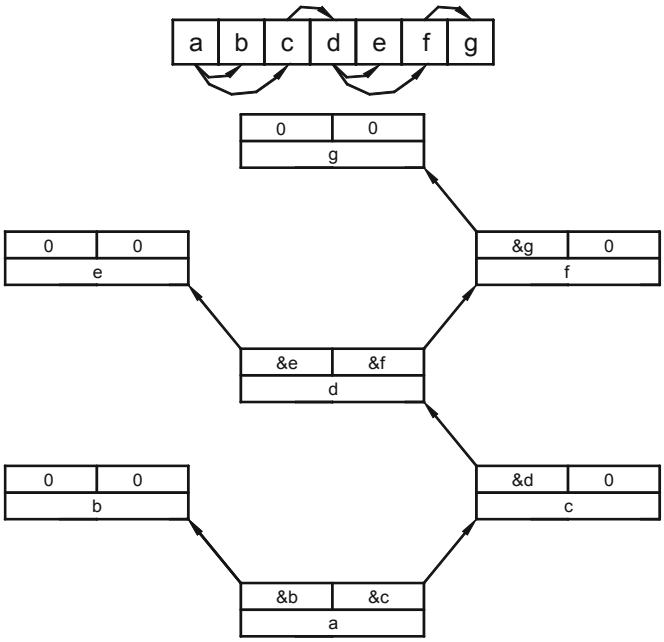


Fig. 1. An example of a description of the encoding and decoding mechanism in the problem of VLSI topology formation in conditions of different orientation of different-sized components

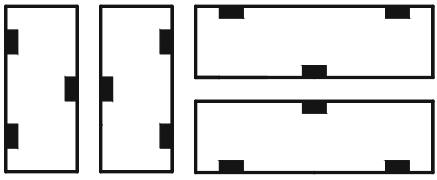


Fig. 2. Location of components

In the classical version, the tree is decoded in several stages. First, the tree is transformed into a graph, from which a vertical tree is obtained and then a horizontal tree again. For the developed diagonal tree, such transformations are not required, because the resulting topology is obtained in one stage of topology construction. The proposed encoding and decoding mechanism allow to significantly reduce the time spent on designing.

Use of the proposed that the use of the proposed mechanism for encoding and decoding alternative solutions will reduce the calculation time due to the unification of the transmitted data through the combined architecture.

5 Conclusion

The proposed modification of GA differs from the existing methods of encoding and decoding chromosomes. The proposed coding method differs from the existing ones

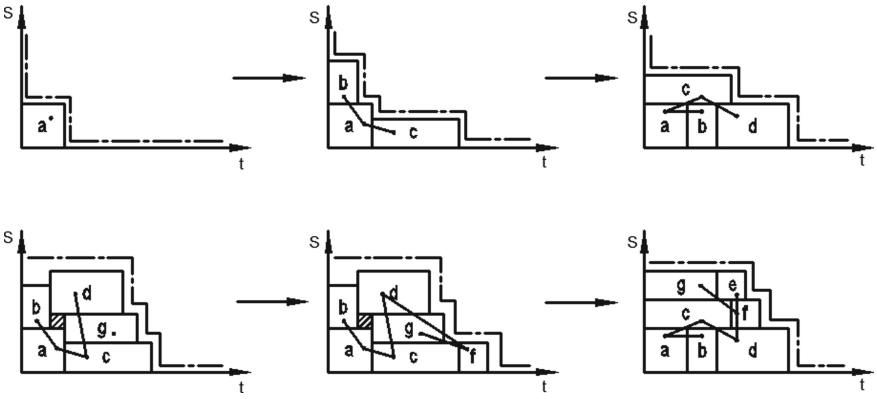


Fig.3. An example of the proposed encoding and decoding mechanism

in that it takes into account the size and location of the components. The decoding method allows you to reduce the time of topology construction by reducing the running time of the algorithm. It is worth noting that the use of the proposed mechanism for encoding and decoding alternative solutions will reduce the calculation time by unifying the transmitted data through various levels of design procedures in the framework of the task of placing VLSI components.

Acknowledgment. The reported study was funded by RFBR, project number № 20–37–90151.

References

1. Danilchenko, V.I., Kureychik, V.M.: Genetic algorithm of VLSI placement planning//Izvestiya SFU. Technical sciences **2**, 75–79 (2019)
2. Lebedev, B.K., Lebedev, V.B.: Planning on the basis of swarm intelligence and genetic evolution//Izvestiya SFU. Tech. Sci. **4** (93), 25–33 (2009)
3. Kovalev, A.V.: Method of designing high-speed asynchronous digital devices with low power consumption//News of universities. Electron. (1), 48–53 (2009)
4. Bova, V.V., Kureychik, V.V.: Integrated subsystem of hybrid and combined search in design and control tasksd//Izvestiya SFU. Tech. Sci. 12 (113), 37–42 (2010)
5. Lebedev, O.B.: Hybrid partitioning algorithm based on the ant colony method and collective adaptation. //Integrated models and soft computing in Artificial Intelligence. Collection of scientific papers of the V-th International Scientific and Technical Conference, vol. 2, T.1, pp. 620–628. Physical education, M (2009)
6. Bushin, S.A., Kovalev, A.V.: Models of power consumption of asynchronous CMOS VLSI functional blocks. Izvestiya SFU. Tech. Sci. 12(83), 198–200 (2009)
7. Bushin, S.A.: Method of reducing energy consumption in asynchronous VLSI blocks. Materials of X VNTC students and postgraduates technical cybernetics, radio electronics and control. Vol. 2, pp. 37–38. Publishing House of TTI SFU, Taganrog (2010)
8. Danilchenko, Y.V., Kureichik, V.M.: Bio-inspired approach to microwave circuit design. In: IEEE East-West design & test symposium (EWDTS), pp. 362–366. (2020)

9. Sokolov, A.A., Dobush, I.M., Sheerman, F.I., Babak, L.I., etc.: Complex-functional blocks of broadband radio frequency amplifiers for single-chip L- and S-band receivers based on SiGe technology. In: 3rd International scientific conference "ECB and Electronic Modules" (International Forum "Microelectronics-2017"), pp. 395–401. Technosphere, Alushta—Moscow (2017)
10. Kureichik, V.M., Lebedev, B.K., Lebedev, O.B.: Hybrid evolutionary algorithm of planning VLSI. In: Proceedings of the 12th annual genetic and evolutionary computation conference, GECCO '10 12th Annual genetic and evolutionary computation conference, GECCO-2010. Sponsors: Assoc. Comput. Mach., Spec. Interest Group Genet., Evol. Comput. (ACM SIGEVO), pp. 821–822. Portland, OR (2010)
11. Bushin, S.A., Kovalev, A.V.: The evolutionary method of placing multi-dimensional VLSI blocks Izvestiya SFU. Tech. Sci. **17**(83), 45–53 (2010)
12. Zhabin, D.A., Garays, D.V., Kalentyev, A.A., Dobush, I.M., Babak L.I.: Automated synthesis of low noise amplifiers using s-parameter sets of passive elements, Asia-Pacific Microwave Conference (APMC 2017), Kuala Lumpur, Malaysia (2017) (accepted for publication)
13. Kovalev A.V.: The evolutionary method of task distribution in systems-on-a-chip to reduce energy consumption Proceedings of the International scientific and technical conferences "Intelligent systems" and "Intelligent CAD". -M.: Physical education, T.1, pp. 102–103 (2010)
14. Babak, L.I., Kokolov, A.A., Kalentyev, A.A.: A new genetic-algorithm-based technique for low noise amplifier synthesis, european microwave week 2012, pp. 520–523. Amsterdam, The Netherlands (2012)
15. Mann, G.K.I., Gosine, R.G.: Three-dimensional min–max-gravity based fuzzy PID inference analysis and tuning. Fuzzy Sets Syst.**156**, 300–323 (2005)
16. Bova, V.V., Kuliev, E.V., Shcheglov, S.N.: Evaluation of the effectiveness of the method of searching for associative rules for big data processing tasks. News of the SFU. Technical sciences. Thematic issue Intelligent CAD. (2020)
17. Furber, S.B.: and Day P. Four-phase micropipeline latch control circuits // IEEE Trans VLSI Syst. **4**, 247–253 (1996)
18. Furber, S.: Computing without clocks: Micropipelining the ARM processor//Asynchronous digital circuit design. In: Birtwistle, G., Davis, A. (eds.), pp. 211–262. Springer-Verlag, New York (1995)
19. Alpert C.J., Mehta D.P., Sapatnekar S.S.: Handbook of algorithms for physical design automation. CRC Press, New York (2009)
20. Neupokoeva N.V., Kureychik V.M.: Quantum and genetic algorithms for the placement of EVA components. Publishing house of TTI SFU, Monograph—Taganrog (2010)

Machine Learning-Based Social Media Text Analysis: Impact of the Rising Fuel Prices on Electric Vehicles

Kamal H. Jihad[1,2], Mohammed Rashad Baker[3(✉)], Mariem Farhat[4], and Mondher Frikha[1]

[1] ENETCOM, Universite de Sfax, Sfax, Tunisia
[2] Electronic Computer Center, University Presidency, University of Kirkuk, Kirkuk, Iraq
[3] Department of Software, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq
mohammed.rashad@uokirkuk.edu.iq
[4] The Higher Institute of Information and Communication Technologies, Carthage University, Tunis, Tunisia

Abstract. Recently, oil costs and environmental concerns have risen dramatically. Additionally, growing urbanization, urban mobility, and employment face several difficulties. Developing Electric Vehicles (EVs) is one of the crucial solutions. However, adopting EVs has remained difficult despite favorable consumer attitudes. This article analyzes public opinions and what they express on Twitter about EVs. We selected social media in this study because it is one of the most prevalent platforms for expressing opinions or viewpoints. Therefore, our goal is to calculate how frequently the occurrence of words is to understand the concerns and interests of the consumers about EVs. To achieve that, we utilized three Machine Learning (ML) models: Random Forest (RF), Decision Tree (DT), and Naïve Bais (NB). We evaluated each model in terms of accuracy and the Matthews Correlation Coefficient. The results demonstrate that all the models can achieve excellent accuracies, especially the DT model, which can give the best outcomes.

Keywords: Electric vehicles · Fuel prices · Machine learning · Twitter

1 Introduction

Recently, worldwide interest in Electronic Vehicles (EVs) has increased, particularly in light of the significant increase in gasoline costs and the global fuel crisis [1]. Furthermore, with Russia's invasion of Ukraine, global fossil fuel prices began to rise and surged in early 2022. Therefore, governments have taken various temporary steps to alleviate the impact of increasing energy bills on consumers and businesses. However, designing cost-effective support policies remains a significant challenge for policymakers [2].

Analyzing people's opinions and perspectives is essential to help the decision-making process. Accordingly, decision-makers rely on tools to help them to understand people's opinions and perspectives. Many users have embraced Twitter as a universal platform

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 625–635, 2023.
https://doi.org/10.1007/978-3-031-27409-1_57

for worldwide news dissemination, article sharing, and social interaction. Consequently, such a high-volume, massive data is created on Twitter every second. While the data can be employed for substantial analysis and interpretation, noisy text data is the greatest obstacle to data analysis. This informal word method is noisy data that cannot be processed by Natural Language Processing (NLP) techniques [3].

Machine learning (ML) is an efficient method for understanding human behavior [4]. It is a relatively recent and effective data-processing method for scientific projects [5]. Recently, ML models have gained widespread popularity in various fields of text classification. High dimensionality is typical in the case of classifying social media messages. This increases the time and processing power required to execute some fundamental ML models [6]. In the paper, we analyze people's opinions about Evs on Twitter. We utilize three ML classifiers: Naive Bayes (NB), Random Forest (RF), and Decision Tree (DT).

2 Related Work

2.1 EVs Related Researches

Paris Declaration on Electro-Mobility stated that, in 2030, the EVs deployment would reach more than 100 million [7]. This has heightened the necessity for market participants to efficiently evaluate the industry and get value from online data such as social media [8]. Suresha and Tawari used topic models and the Valence Aware Dictionary to evaluate 45,000 tweets. On Twitter, the authors determined that "Tesla" was one of the top EV hashtags [9]. Similarly, according to Bhatnagar and Choubey's Twitter-based sentiment analysis utilizing TF-IDF scores and an NB classifier, the hashtag "#Tesla" had a better positive sentiment than other manufacturers [10].

Cofman et al. [11] determined that despite considerable performance improvements, most governments' EV adoption targets could not be realized. The authors in [12], Christidis and Focas, found that wealth, educational achievement, and urbanization substantially affected EV adoption in the European Union. Additionally, they determined that regional variations and local conditions significantly impact EV purchases. Soltani-Sobh et al. [13] studied EV adoption in the USA. They concluded that government incentives, urban roads, and electricity prices significantly affect EV adoption.

The elements influencing the public acceptability of electric cars in Shanghai (China) were investigated by Wang et al. [14]. A study conducted in Thailand by Thananusak et al. [15] to examine the performance of some factors. The factors included driving safety, speed, and range. The authors demonstrated that these factors were more crucial than charging infrastructure availability and financial considerations. Tu and Yang [16] researched Taiwanese consumers. The authors discovered that the availability of resources and opinions from customers' surroundings and their environmental consciousness impact consumer EV purchasing intentions. To understand the factors affecting EV purchases, Li et al. [17] conducted a systematic study of 1846 papers.

Kim et al. [18] conducted a study in Korea evaluating consumer intentions for EVs purchase. They found that consumers' intentions to purchase EVs were significantly impacted by prior experience driving EVs, along with factors like perception of government incentives, parking availability, educational achievement, and EVs number in the household.

2.2 Machine Learning Methods

Numerous ML models have been used to make predictions based on a particular dataset. Typically, these models are utilized for error detection and diagnosis. Li and Oechtering presented a privacy-aware Bayesian Network (BN) to find future trustworthy IoT applications [19]. Using Bayesian approaches, Viegas et al. [20] presented a system that might forecast occurrences, like power outages, in the smart grid. In addition to a Bayesian strategy, the RF algorithm was used to anticipate occurrences of interest. Similarly, Xu and Xu [21] studied developing a Bayesian Network-based health management diagnostic system for space avionics. Ali et al. [22] developed a BN model to choose the most proper links for power quality meter placement. Additionally, Yu et al. [23] developed a method capable of monitoring the health of hybrid systems with several faults. Using BA, the model can anticipate if a fault has occurred and the potential remedies for that problem.

For multiclass feature selection, Ramona et al. [24] developed a model using a Support Vector Machine (SVM). The model aimed to optimize the complexity and cost. The author built the model using kernel class separability and kernel target alignment. The findings demonstrated kernel class separability is less optimized than kernel target alignment. Nevertheless, the models only perform well with linearly separable data and poorly with nonlinear data classification. Wang et al. [25] used SVM, NB, and RF algorithms to perform an emotional analysis method on COVID-19 data collected from Sina Weibo, China. Using a Deep Neural Network, Turkish Tweets have been classified by Mansur ALP et al. [26] into basic emotions.

In this article, a set of ML models is utilized to calculate how frequently the occurrence of words is to understand the concerns and interests of consumers about EVs. Cross-validation is then utilized to examine various models' performance.

3 Materials and Method

In this section, we discuss the main parts of the proposed structure. Figure 1 illustrates the structure of the proposed ML model.

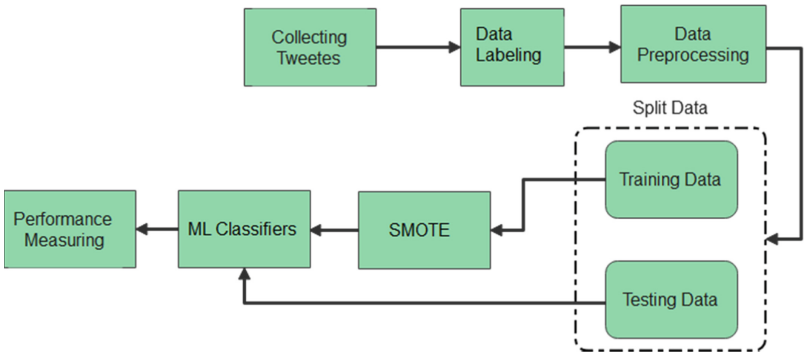


Fig. 1. The proposed ML structure.

3.1 Data Collection

In social media data analysis, several data collection methods were proposed. Social media microblogs, such as Twitter, make user tweets publicly. To investigate public perception and data classification for EVs, we developed a script to collect related data from Twitter. Instead of the official Twitter API, It has been designed and written in Python. In data collection, we used most related keywords used as hashtags on Twitter and keywords used by Google trends.

Table 1 shows the list of hashtags used to gather the dataset. Additional elements were retrieved from the collected tweets after data collection. For example, hashtags, language, tweet date, and tweet ID. These were stored as values separated by commas for analysis.

Table 1. The related used hashtags

#Electriccars	#Electricvehicles	#Hybridcars
#Electriccars	#evcharging	#HybridElectricCars
#NewElectricCarfriendly	#EVroadtrip	#getthepower
#EV	#EVgo	#Hybridelectricvehicle
#Builtforabetterdcodrivings	#DriveElectric	

In the experiment, we collected 292,666 tweets from 2022/03/01 to 2022/09/30 amid the global oil price hike after the Russia-Ukraine war to measure its impacts during the crisis. After filtration, 133,399 tweets were left as a final dataset.

3.2 Data Preprocessing

3.2.1 Data Cleaning

In the context of Twitter data for natural language processing, we consider the following as noise: Duplicate tweets, names of users, Hashtags, URLs & Whitespaces, and non-standard text and symbols. Following the elimination of noise, further preprocessing procedures were carried out:

1. **Lower casing**: To decrease dimensionality and enhance subject coherence, all words were switched to lowercase.
2. **Stop words removal**: The search engine is programmed to ignore some words such as “the”, “a”, and “an”. Since these terms are superfluous, they may be eliminated from the database, leaving users with essential information.
3. **Links Removal**: All links associated with a tweet were removed. For example, numbers, URLs, special characters, punctuations, and HTML tags.
4. **Lemmatization**: Words were reduced to their word roots by eliminating their suffixes and affixes. This procedure seeks to construct lemmas or meaningful versions of these words. For example, ‘going’ was lemmatized as ‘go’. This process was implemented on the whole dataset.

3.2.2 Data Labeling

Data labeling (or data annotation) is adding metadata, or tags, to raw data to show a machine learning model the target attributes—answers—it is expected to predict. There are different ways to perform data annotation. The style choice depends on the problem statement's complexity, the amount of data to be tagged, the size of a data science team, financial resources and available time.

Here we used Vader sentiment analysis to label the final dataset. If a single tweet were calculated as more significant than 0, it would be marked as positive. On the other hand, if the tweet were less than 0, it would be marked as a negative tweet. In the end, we have two binary classes (0 and 1) 0s represent negative tweets, and 1s represent positive tweets. Table 2 shows the labeled dataset and class tweets counts.

Table 2. Details of the collected dataset

No. of positive tweets	105601
No. of negative tweets	27798
Total	133399

3.2.3 Smote

Table 2 shows the total number of positive tweets after labeling was 105601 tweets, and the total number of negative tweets after labeling was 27798. We used the 80:20 ratio to split our dataset into training and testing for our experiment. Table 2 shows that our dataset is not balanced. Therefore, if we apply the selected models, the predictive results will be biased toward one class instead of another. Accordingly, we need a technique to make balancing our dataset. We used Synthetic Minority Oversampling Technique (SMOTE). SMOTE, or Synthetic Minority Oversampling Technique, is an additional method for oversampling the minority class. Consequently, SMOTE examines minority class examples and uses k nearest neighbor to identify a random nearest neighbor; a synthetic instance is then generated randomly in feature space [27].

We applied SMOTE technique to our training dataset to avoid biases towards any class. Table 3 lists the final data after applying SMOTE upsampling technique.

Table 3. Before and after applying SMOTE on the training dataset

	Before	After
No. of positive tweets	84428	84428
No. of negative tweets	22291	84428

3.3 ML Classification Methods

1. **Naïve Bayes**: It is a supervised ML algorithm that can address classification issues based on the Bayes theorem. Including a high-dimensional training dataset, NB is widely utilized in text classification [E]. It is one of the simplest and most efficient classification methods, and it aids in developing ML models that can generate predictions rapidly. It is a probabilistic model that anticipates based on the probability of the object.
2. **Decision Tree**: It is a Supervised learning algorithm that can be implemented for regression and classification; widely used in classification problems. The structure of DT is a tree-structured classifier as follows. Internal nodes represent dataset features. Then, decision rules are the branches, where each leaf node refers to the outcome. Furthermore, there are two types of nodes: Decision Node and Leaf Node. While the Decision Node has multiple branches and is utilized for decision-making, the Leaf Node does not contain branches and is the output of those decisions
3. **Random Forest**: An ML algorithm develops a forest of decision trees [15]. It can be used for both regression and classification. To enhance the result, RF can be operated by combining learning models. Decision trees are utilized to improve classification accuracy.

4 Models Evaluation Metrics

Several criteria are utilized to evaluate the performance of the different classifiers. The correct predictions are represented by the word "True" as True Positive (TP) and True Negative (TN). The incorrect predictions are referred to by the word "False" as False Positive (FP) and False Negative (FP) [28].

1. **Accuracy (CA)**: This parameter computes the percentage of accurate predictions compared to the total number of samples and can be written as "$AC = (TP + TN)/(TP + TN + FP + FN)$.
2. **Precision (PR)**: It enables assessing how much repeatability a model contains. This might inform the user that the same result would be obtained if the measurements were performed under the same circumstances. It is formulated as "$PR = TP/(TP + FP)$".
3. **Recall**: Also called sensitivity, it displays the ratio of the number of related examples retrieved to the number of relevant instances that might have been picked. The recall is written as "$Recall = TP/(TP + FN)$".
4. **F1-Score**: It essentially brings in both the *FPs* and *FNs* to weigh errors in decision-making. It is referred to as the Harmonic mean of recall and precision. It is worth mentioning that even if our precision is remarkably high, having a low recall will always dominate and bring down the F1-Score necessarily and vice-versa. It is formulated as "$F1 - score = 2[(Precision * Recall)/(Precision + Recall)$".

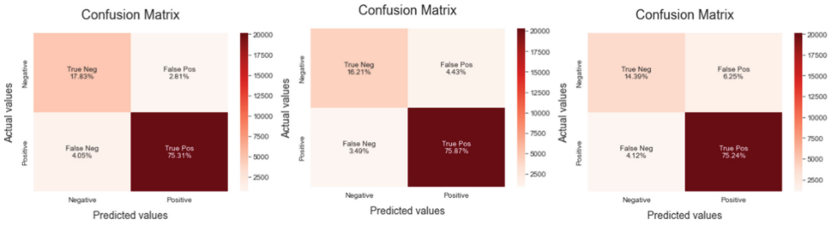
5 Performance Evaluation

Accuracy and F1-score derived on confusion matrices are two of the most often used metrics in binary classification problems. Nevertheless, These statistical metrics may provide dangerously optimistic and exaggerated conclusions, mainly when applied to unbalanced data sets. Instead, the MCC is a more reliable statistical rate. MCC would get a high score only if the prediction performed well in each of the confusion matrix's four areas ($TP, TN, FP, and\ FN$) proportionally to the size of positive and negative elements in the dataset. In our experimental results, we used MCC to overcome this issue after applying the ML models to our dataset to predict the performance of the proposed models.

Table 4 illustrate the achieved accuracy of the models. DT model is the most significant predictive value with 0.93. Additionally, the model achieved the highest MCC with 0.796 scores, whereas RF and NB reached 0.754 and 0.672, respectively. An in-depth evaluation of the optimal experiment was done using the confusion matrix. It aids in deriving explanations for how the model performed on a test. Confusion matrices are used to compare predicted and exact classes to visualizing the suggested classifier's accuracy. Figure 2 depicts the confusion matrix of the models.

Table 4. Results of applied ML models

	Class	Precision	Recall	F1-score	Accuracy	MCC
DT	0	0.81	0.86	0.84	0.93	0.796
	1	0.96	0.95	0.96		
RF	0	0.82	0.79	0.80	0.92	0.754
	1	0.94	0.96	0.95		
NB	0	0.87	0.70	0.74	0.90	0.672
	1	0.92	0.95	0.94		



DT RF NB

Fig. 2. The confusion matrix of the proposed ML models

Lastly, we draw the ROC curve, which depicts the trade-off between specificity and sensitivity. Classifiers perform better when their curves are further to the top-left corner.

In other words, it is the likelihood that a randomly selected positive instance will be covered more than a randomly selected negative instance. As shown in Fig. 3, the RF model achieved a better AUC value with a 0.958 score. In contrast, the NB and DT models achieved 0.935 and 0.911, respectively.

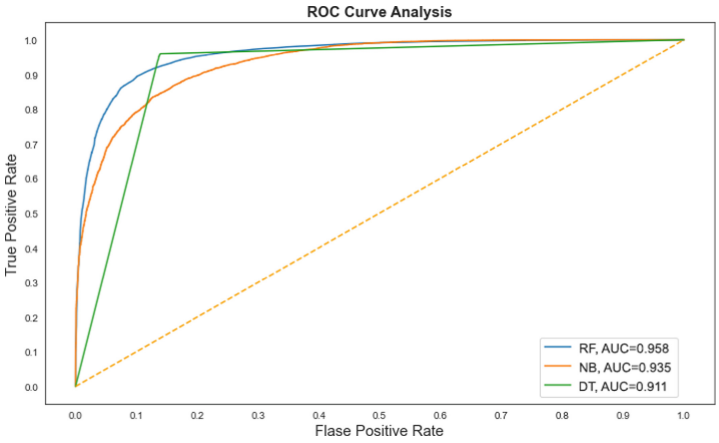


Fig. 3. The ROC curve analysis of proposed ML models

To demonstrate the superiority of the current study against the state-of-the-art counterparts, we compared our models with the efforts [29–31] and [32]. The comparison shows that our models outperformed them in the limitations. Furthermore, they achieved better evaluation of all metrics (i.e., precision, recall, f1-score and accuracy). Specifically, in the DT model, the findings reached 96, 95, 96 and 93%. Table 5 shows the details of the comparison between the ML models and ours.

Table 5. Comparison of our study with counterparts [29–31] and [32].

Study	Model	Case study	Limitation	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
[29]	DT	The battery in the EVs	The accuracy of the algorithms needs to be improved	84	80	82	72
	NB			95	91	93	88
	KNN			90	86	88	80
	SVC			79	95	86	76

(*continued*)

Table 5. (*continued*)

Study	Model	Case study	Limitation	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
[30]	DT	Charging EVs	1. ML algorithms run randomly 2. The dataset is not precise and large 3 The model needs to be practical	95	94	---	90.1
	RF			94	95	---	90.3
	NB			75	79	---	63.5
	KNN			88	92	---	83.3
	SVM			89	92	---	83.7
	DNN			91	87	---	88.2
	LSTM			96	97	---	95.3
[31]	Machine learning model	Predict the availability of an EV to perform V2H Services	There is a limited understanding of what impacts EVs	–	–	–	85.00
[32]	LR	Forecast the energy consumption of EVs	There are some limitations to conquer, including maximum substantially the charging time and the necessity for public charging	–	–	–	17.86
	DT			–	–	–	99.91
	NB			–	–	–	–
Our Study	DT	People's opinions about EVS in light of the rise in fuel	-----	96	95	96	93
	RF			94	96	95	92
	NB			92	95	94	90

6 Conclusion

In this article, we analyzed public sentiments and how they are expressed on Twitter in this work. The goal was to identify commonly appearing words better to understand consumers' interests and concerns about electric vehicles. Three ML models: RF, DT, and NB, were used. The accuracy and MCC of the models were evaluated. The models produced accurate results across the board, and where RF model produced better findings. The study demonstrated the importance of applying ML and NLP to the economy in times of crisis. The ambiguity from an economic perspective and content posted by another tweeter on social media are two typical problems. As a result, we concluded that

using ML models for classification problems can get significant accuracy. In addition, using oversampling techniques like SMOTE can give more reliable results without bias to one class against another.

Nevertheless, one limitation of this study is that its findings are based only on Twitter. Different social media platforms must be considered for better understanding and categorizing thoughts.

References

1. Tian, X., et al.: A bibliometric analysis on trends and characters of carbon emissions from transport sector. Transp. Res. Part D Transp. Environ. **59**, 1 (2018). https://doi.org/10.1016/j.trd.2017.12.009
2. Celasun, O., et al.: Surging energy prices in europe in the aftermath of the war: How to support the vulnerable and speed up the transition away from fossil fuels. IMF Work. Pap. **2022**, 1 (2022). https://doi.org/10.5089/9798400214592.001
3. Li, W., Xu, H.: Text-based emotion classification using emotion cause extraction. Expert Syst. Appl. **41**, 1742–1749 (2014). https://doi.org/10.1016/j.eswa.2013.08.073
4. Sang, Y.N., Bekhet, H.A.: Exploring factors influencing electric vehicle usage intention: An empirical study in malaysia. Int. J. Bus. Soc. **16**, 57–74 (2015). https://doi.org/10.33736/ijbs.554.2015
5. Yuvalı, M., Yaman, B., Tosun, Ö.: Classification comparison of machine learning algorithms using two independent CAD datasets. Mathematics. **10**, 311 (2022). https://doi.org/10.3390/math10030311
6. Hassan, S.U., Ahamed, J., Ahmad, K.: Analytics of machine learning-based algorithms for text classification. Sustain. Oper. Comput. **3**, 238–248 (2022). https://doi.org/10.1016/j.susoc.2022.03.001
7. Seddig, K., Jochem, P., Fichtner, W.: Integrating renewable energy sources by electric vehicle fleets under uncertainty. Energy **141**, 2145–2153 (2017). https://doi.org/10.1016/j.energy.2017.11.140
8. He, W., Tian, X., Tao, R., Zhang, W., Yan, G., Akula, V.: Application of social media analytics: A case of analyzing online hotel reviews. Online Inf. Rev. **41**, 921–935 (2017). https://doi.org/10.1108/OIR-07-2016-0201
9. Suresha, H.P., Kumar Tiwari, K.: Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data. Asian J. Res. Comput. Sci, 13–29 (2021). https://doi.org/10.9734/ajrcos/2021/v12i230278
10. Bhatnagar, S., Choubey, N.: Making sense of tweets using sentiment analysis on closely related topics. Soc. Netw. Anal. Min. **11**(1), 1–11 (2021). https://doi.org/10.1007/s13278-021-00752-0
11. Coffman, M., Bernstein, P., Wee, S.: Electric vehicles revisited: a review of factors that affect adoption. Transp. Rev. **37**, 79–93 (2017). https://doi.org/10.1080/01441647.2016.1217282
12. Christidis, P., Focas, C.: Factors affecting the uptake of hybrid and electric vehicles in the European union. Energies **12**, 3414 (2019). https://doi.org/10.3390/en12183414
13. Soltani-Sobh, A., Heaslip, K., Stevanovic, A., Bosworth, R., Radivojevic, D.: Analysis of the Electric vehicles adoption over the United States. In: Transportation Research Procedia, pp. 203–212. Elsevier (2017). https://doi.org/10.1016/j.trpro.2017.03.027
14. Wang, N., Tang, L., Pan, H.: Analysis of public acceptance of electric vehicles: An empirical study in Shanghai. Technol. Forecast. Soc. Change. **126**, 284–291 (2018)
15. Thananusak, T., Rakthin, S., Tavewatanaphan, T., Punnakitikashem, P.: Factors affecting the intention to buy electric vehicles: Empirical evidence from Thailand. Int. J. Electr. Hybrid Veh. **9**, 361–381 (2017). https://doi.org/10.1504/IJEHV.2017.089875

16. Tu, J.C., Yang, C.: Key factors influencing consumers' purchase of electric vehicles. Sustain. **11**, 3863 (2019). https://doi.org/10.3390/su11143863
17. Li, W., Long, R., Chen, H., Geng, J.: A review of factors influencing consumer intentions to adopt battery electric vehicles (2017). https://doi.org/10.1016/j.rser.2017.04.076
18. Kim, J.H., Lee, G., Park, J.Y., Hong, J., Park, J.: Consumer intentions to purchase battery electric vehicles in Korea. Energy Policy **132**, 736–743 (2019). https://doi.org/10.1016/j.enpol.2019.06.028
19. Li, Z., Oechtering, T.J.: Privacy-aware distributed Bayesian detection. IEEE J. Sel. Top. Signal Process. **9**, 1345–1357 (2015)
20. Viegas, J.L., Vieira, S.M., Melicio, R., Matos, H.A., Sousa, J.M.C.: Prediction of events in the smart grid: Interruptions in distribution transformers. In: Proceedings—2016 IEEE International power electronics and motion control conference, PEMC 2016, pp. 436–441. IEEE (2016). https://doi.org/10.1109/EPEPEMC.2016.7752037
21. Xu, L., Xu, J.: Integrated system health management-based progressive diagnosis for space avionics. IEEE Trans. Aerosp. Electron. Syst. **50**, 1390–1402 (2014)
22. Ali, S., Wu, K., Weston, K., Marinakis, D.: A machine learning approach to meter placement for power quality estimation in smart grid. IEEE Trans. Smart Grid. **7**, 1552–1561 (2016). https://doi.org/10.1109/TSG.2015.2442837
23. Yu, M., et al.: Scheduled health monitoring of hybrid systems with multiple distinct faults. IEEE Trans. Ind. Electron. **64**, 1517–1528 (2017). https://doi.org/10.1109/TIE.2016.2619322
24. Ramona, M., Richard, G., David, B.: Multiclass feature selection with kernel gram-matrix-based criteria. IEEE Trans. Neural Networks Learn. Syst. **23**, 1611–1623 (2012). https://doi.org/10.1109/TNNLS.2012.2201748
25. Li, L., et al.: Characterizing the propagation of situational information in social media during covid-19 epidemic: a case study on weibo. IEEE Trans. Comput. Soc. Syst. **7**, 556–562 (2020). https://doi.org/10.1109/TCSS.2020.2980007
26. Tocoglu, M.A., Ozturkmenoglu, O., Alpkocak, A.: Emotion analysis from turkish tweets using deep neural networks. IEEE Access. **7**, 183061–183069 (2019). https://doi.org/10.1109/ACCESS.2019.2960113
27. Baker, M.R., Mahmood, Z.N., Shaker, E.H.: Ensemble learning with supervised machine learning models to predict credit card fraud transactions. Rev. d'Intelligence Artif. **36**, 509–518 (2022)
28. Gozudeli, Y., Karacan, H., Yildiz, O., Baker, M., Minnet, A., Kalender, M., Akcayol, M.: A new method based on Tree simplification and schema matching for automatic web result extraction and matching. In: Proceedings of the international multi conference of engineers and computer scientists, (2015)
29. Haripriya, S., Esakki Vigneswaran, E., Jayanthy, S.: Battery management system to estimate battery aging using deep learning and machine learning algorithms, (2022). https://doi.org/10.1088/1742-6596/2325/1/012004
30. Shibl, M., Ismail, L., Massoud, A.: Machine learning-based management of electric vehicles charging: Towards highly-dispersed fast chargers. Energies. **13**, (2020). https://doi.org/10.3390/en13205429
31. Aguilar-Dominguez, D., Ejeh, J., Dunbar, A.D.F., Brown, S.F.: Machine learning approach for electric vehicle availability forecast to provide vehicle-to-home services. Energy Rep. **7**, 71–80 (2021). https://doi.org/10.1016/j.egyr.2021.02.053
32. Balaiah, G., Dhanasree, V. P., Jyothi, M., Varun, K., Chowhan, D.U.: Predicting Charge Consumption of Electric Vehicles Using Machine Learning. J. Algebr. **13**, 2087–2095 (2022)

MobileNet-Based Model for Histopathological Breast Cancer Image Classification

Imen Mohamed ben ahmed[1], Rania Maalej[2(✉)], and Monji Kherallah[3]

[1] Faculty of Sciences, University of Gafsa, Gafsa, Tunisia
[2] Medical School of Sfax, University of Sfax, Sfax, Tunisia
rania.maalej@medecinesfax.org
[3] Faculty of Sciences, University of Sfax, Sfax, Tunisia
monji.kherallah@fss.usf.tn

Abstract. Nowadays, Breast cancer is a massive health problem worldwide. To fight against this disease, we propose a high-performance Computer-Aided Diagnosis system using deep learning. Specifically, we focus on the classification of histopathological images of breast cancer into two classes (benign and malignant). For that, we present a Mobilenet-based breast cancer classification model. This model is trained with a new extended Breakhis dataset, which is created by applying some data augmentation techniques. According to the experiments, our proposed model gives a very competitive result and the accuracy reaches 0.9.

Keywords: Breast cancer classification · MobileNet · Deep learning · Data augmentation · BreakHis dataset

1 Introduction

Breast cancer is a disease due to the uncontrolled growth of certain cells in the breast. It is a major health issue and the leading cause of female cancer deaths in the world.

To fight against this disease and to improve the survival rate, the development of automatic medical imaging processing is becoming a necessity. Indeed, this field is a rapidly expanding area where the problem of automatic interpretation medical images is a pressing need. And given the large number of medical imaging devices, it becomes tedious to process this huge amount of information, hence the need for artificial intelligence, especially the deep learning field. In this context, a multitude of datasets are collected and a variety of deep neural networks have been proposed.

In this work, we are interested in the classification of histopathological images of breast cancer into two binary classes (benign and malignant). To achieve this goal, and to obtain a robust model, we choose Breakhis [1] as the dataset and the pre-trained network Mobilenet [2] as the features extractor.

This paper is organized as follows: Sect. 2 describes the relevant previous works. Section 3 presents BreakHis dataset, the used data augmentation techniques, and the proposed classification model. In Sect. 4 we give and analyze experiment results, and at the end, conclusions and future works are given in Sect. 5.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 636–643, 2023.
https://doi.org/10.1007/978-3-031-27409-1_58

2 Related Works

With the importance of breast cancer classification in histopathological imaging, there are several studies in the existing literature, and the most recent ones are based on the deep learning field. The most common works used the Convolutional Neural Network [3] or the pre-trained models such as AlexNet [4], GoogleNet [5], ResNet [6], and VGG16 [7]. Indeed, Spanhol et al. [8] presented a novel strategy for training the Alexnet-CNN architecture, based on the extraction of patches obtained randomly or by a sliding window mechanism, to process high-resolution textured images. Experimental results obtained on the BreaKHis dataset showed high performance and the accuracy reached 85.6%.

In [9], Seo, et al. proposed a novel Primal-Dual Multi-Instance SVM classification method, which allows scaling to a wide range of features. Histopathological images are segmented into patches. The feature vector is extracted through the Parameter Free Threshold Statistics (PFTAS) method for each patch. The PFTAS method extracts texture features by counting the number of black pixels in the neighborhood of a pixel. Experiment results on the BreaKHis dataset showed an improved accuracy of 89.8%.

In [10] a Deep Convolution Generative Adversial Network (DCGAN) is applied to give the number of consistent images in the minority class (benign) with that in the majority class (malignant). In addition, the pre-trained DenseNet201 model is used, and features are extracted from the lower layers of DensNet201, via a global average pooling (GAP). These features are passed throw the SoftMax layer to classify breast cancer. The proposed architecture was evaluated using histopathological images from the BreakHis database and showed promising results with 96% with the 40 $\times$ magnification.

Saini and Susan [11] used the Deep Convolutional Generative Adversarial Network (DCGAN) for minority data augmentation in the initial phase of their experiments. DCGAN is used to generate high-quality synthetic fake images from the available distribution of minority data. Then, the new dataset, with a balanced class distribution, is got across the deep transfer network. To enhance the network performance, the proposed VGG16 deep transfer architecture is followed by the batch normalization, 2D convolutional (CONV2D) layer, 2D Global Average Pooling, dropout, and dense layers. The model is evaluated using a two-class BreaKhis provided at four magnification levels and the best accuracy was 96.5%.

Ibraheem et al. [12] proposed three parallel CNN branches (3PCNNBNet) for breast cancer classification through histopathological images. This network offered several advantages, such as learning the high-level and low-level features by considering local and global features simultaneously. They also deployed deep residual blocks using skip connections to help the proposed model overcome the vanishing gradient problem and to improve training and testing. The proposed 3PCNNB-Net architecture was evaluated using histopathological images from the BreakHis database. The 3PCNNB-Net architecture achieved promising results, including a maximum accuracy of 97.04% with a 200 $\times$ magnification.

Zou et al. [13], introduced a novel attention high-order deep network (AHoNet) by simultaneously embedding attention mechanism and high-order statistical representation into a residual convolutional network, this technique allows the network to capture more discriminant deep features for breast cancer pathological images. Experiments on the benchmark BreakHis dataset for different magnification factors: 40X, 100X, 200X, and

400X validate the effectiveness of the proposed deep network in terms of the high scores obtained compared to the state-of-the-art deep networks, in fact, AHoNet gets the optimal image-level classification accuracies of 99.09% (Table 1).

Table 1. Compared results with CNN-based methods on the BreakHis dataset at the image level

System	Data augmentation	Features extraction	Classification	Magnification	Accuracy (%)
Spanhol et al. [8]	–	AlexNet CNN	Softmax layer	all	85.6
Seo et al. [9]	–	PFTAS method	Primal-Dual Multi-Instance SVM	200X	89.8
Djouima et al. [10]	DCGAN + rotation; shear; zoom; horizontal flip; fill mode; width shift; height shift	DenseNet201	Softmax layer	40X	96
Saini and Susan [11]	DCGAN	VGG16 + CNN	Softmax layer	40X	96.5
Ibraheem et al. [12]	−Three zoom ranges −Rotation at 90° −Horizontal and vertical flipping	Three parallel CNN branches + residual blocks	Softmax layer	200X	97.04
Zou et al. [13]	−Simple cropping −Horizontal and vertical flipping −Rotation (90, 120, and 180) −Cutmix method for data amplification	ResNet1825 + attention mechanism + high-order statistical representation	Softmax layer	200X	99.09

Table 2. The Breakhis image distribution by the magnification factor

Class		Benign	Malignant	Total
Magnification factors	40×	625	1370	1995
	100×	644	1437	2081
	200×	623	1390	2013
	400 ×	588	1232	1820

3 System Overview

3.1 BreakHis Histopathological Breast Cancer Dataset

The BreakHis [1] is an open-source dataset available. It consists of 7909 clinical breast tumor histopathological images of 700 × 460-pixel size, respectively, including 2480 benign tumors (adenoma, fibroadenoma, trichome tumors, and tubular adenoma) images and 5429 malignant tumors (ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma) images at four magnifications of 40, 100, 200, and 400 as shown in Table 2.

Figure 1 illustrates four different magnification images of breast tissue sections containing malignant tumors from the BreakHis dataset.

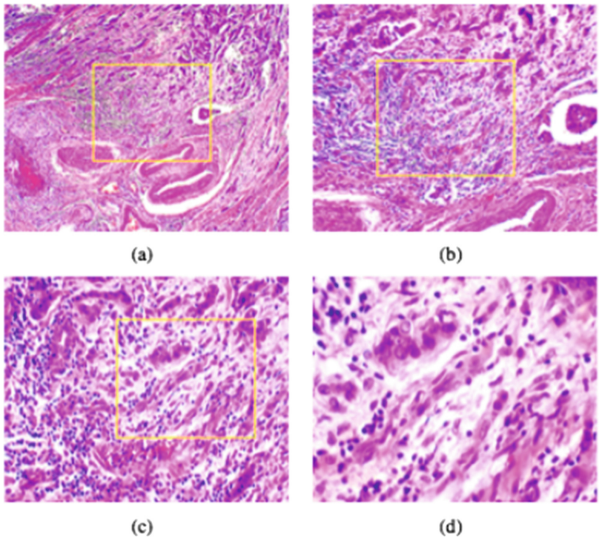


Fig. 1. A slide of a breast malignant tumor seen in different magnification factors: **a** 40×, **b** 100× **c** 200×, and **d** 400× [1].

3.2 Data Augmentation

The size of the dataset plays a very important role in achieving excellent performance in a deep-learning model. Therefore, the augmentation of data enhances network performance and overcomes the overfitting problem.

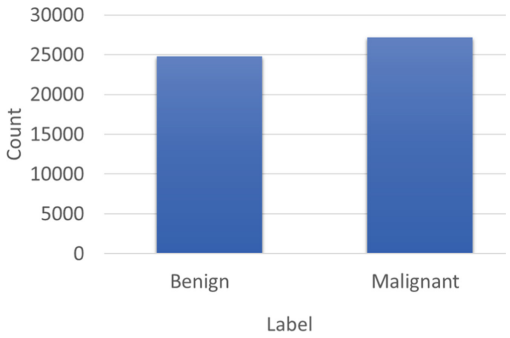


Fig. 2. The balanced class distribution in the new extended Breakhis dataset

In this study, multiple data augmentation techniques are applied, creating multiple versions of each image. Moreover, the horizontal flip transformation is used to generate synthetic high-quality fake images from the available distribution of minority data (benign). So, the class distribution is balanced. Then some other techniques are used to increase the total number of images in this new extended dataset, in fact, we use the vertical flip, rotation, cropping, and sharpening techniques so, as shown in Fig. 2, the total number of images becomes 51945 (24800 Benign and 27145 Malignant) (Table 3).

3.3 System Architecture

First, The Breakhis dataset is randomly shuffled to be divided into training, testing, and validation sets. Then, the images are resized to 224×224 pixels to comply with the input size requirements of MobineNet V3 [2] used for feature extraction. The MobileNet aims to use depthwise separable convolutions to make lighter deep neural networks, so the computational cost is less than the regular convolutional networks. However, MobileNet V3 uses a depth (number of features) multiplier in the convolutional layers to adjust the tradeoff between accuracy and latency.

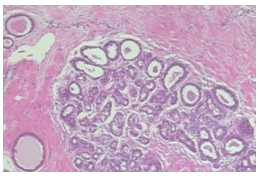
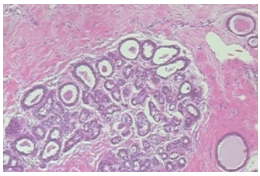
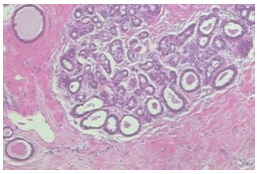
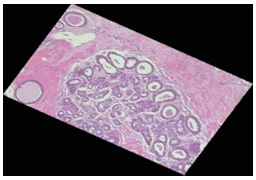
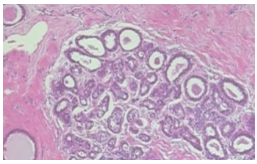
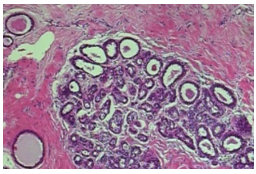
For classification, we propose a fully connected network with four dense layers, three Batch normalization layers, and three dropout layers with a rate of 0.5.

Our proposed model to classify the histopathological images of breast cancer into benign and malignant classes is presented in Fig. 3.

4 Experiment Results

To train our model, we choose the Adam optimizer [14] and use the 5-Fold Cross Validation Technique, we perform categorical cross entropy [15] as a cost function and we set the batch size parameter as 32.

Table 3. Data augmentation techniques applied on image "SOB_B_A-14-22549AB-40–017" from the Breakhis dataset

Image	Transformation	Image	Transformation
	Raw image		Rotation
	Horizontal flip		Crop
	Vertical flip		Sharpen

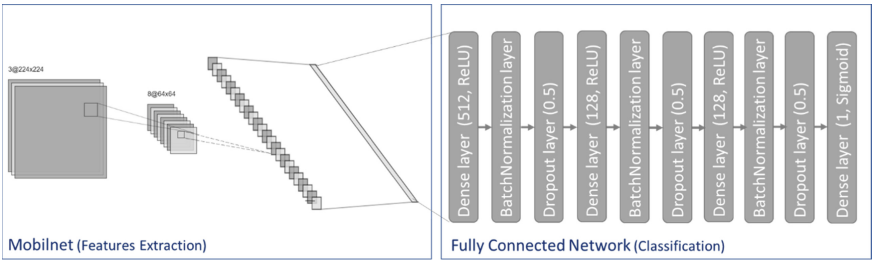


Fig. 3. Proposed system architecture

The best accuracy during the training is 0.99. Figure 4 visualizes the model performance history during the training and validation steps.

In addition, to analyze the recognition results on the testing set, we make the confusion matrix for our proposed model, as shown in Fig. 5. It's pretty obvious that our method has a very good discriminative effect for the histopathological images of breast cancer and the accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{4594 + 4785}{4594 + 4785 + 486 + 524} = 0.9$$

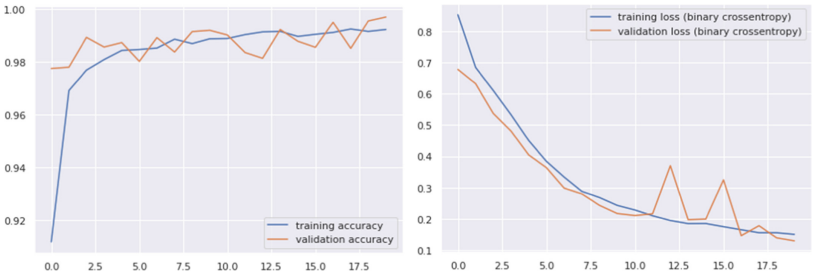


Fig. 4. Accuracy and Loss during training and validation.

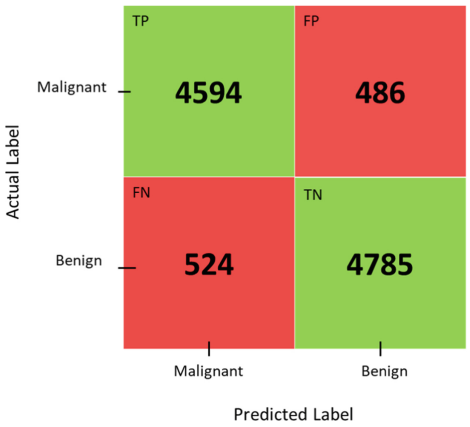


Fig. 5. Confusion matrix

5 Conclusion

This paper presents a Mobilenet-based model for breast cancer classification from histopathological images. This model is trained with a new extended Breakhis dataset, which is created by applying some data augmentation techniques. Based on the experiments, our proposed model gives a very competitive result. Indeed, the accuracy reaches 0.9. In future work, we propose to improve the network by increasing its depth in order to classify the subclasses of the BreastHis dataset.

References

1. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng., **63** (7), 1455-1462 (2016)
2. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Adam, H.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314–1324. (2019)
3. Ting, F.F., Tan, Y. J., Sim, K.S.: Convolutional neural network improvement for breast cancer classification. Expert. Syst. Appl. **120**, 103-115 (2019)

4. Hassan, S.A., Sayed, M.S., Abdalla, M.I., Rashwan, M.A.: Breast cancer masses classification using deep convolutional neural networks and transfer learning. Multimedia Tools and Applications **79**(41–42), 30735–30768 (2020). https://doi.org/10.1007/s11042-020-09518-w
5. Yao, X., Wang, X., Karaca, Y., Xie, J., Wang, S.: Glomerulus classification via an improved googlenet. IEEE Access. **8**, 176916-176923 (2020)
6. Li, J., Zhang, J., Sun, Q., Zhang, H., Dong, J., Che, C., Zhang, Q.: Breast cancer histopathological image classification based on deep second-order pooling network. In: 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–7). IEEE (2020)
7. Albashish, D., Al-Sayyed, R., Abdullah, A., Ryalat, M.H., Ahmad Almansour, N.: Deep CNN Model based on VGG16 for Breast cancer classification. In: 2021 International Conference on Information Technology (ICIT), pp. 805-810. (2021)
8. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, et L.: Breast cancer histopathological image classification using Convolutional Neural Networks. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2560-2567. Vancouver, BC, Canada, (2016)
9. Seo, H., Brand, L., Barco, L.S., Wang, et H.: Scaling multi-instance support vector machine to breast cancer detection on the BreaKHis dataset. Bioinform. **38** (Supplement_1), i92-i100 (2022)
10. Djouima, H., Zitouni, A., Megherbi, A.C., Sbaa, et S.: classification of breast cancer histopathological images using DensNet201. In: 2022 7th International Conference on Image and Signal Processing and their Applications (ISPA), pp. 1-6. Mostaganem, Algeria (2022)
11. Saini, M., Susan, S.: Deep transfer with minority data augmentation for imbalanced breast cancer dataset. Appl. Soft Comput. **97**, 106759 (2020)
12. Ibraheem, A.M., Rahouma, K.H., Hamed, H.F.A.: 3PCNNB-Net: Three parallel cnn branches for breast cancer classification through histopathological images. Journal of Medical and Biological Engineering **41**(4), 494–503 (2021). https://doi.org/10.1007/s40846-021-00620-4
13. Zou, Y., Zhang, J., Huang, S., Liu, B.: Breast cancer histopathological image classification using attention high-order deep network. Int. J. Imaging Syst. Technol., **32**(1), 266-279 (2022)
14. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. (2014). arXiv preprint arXiv:1412.6980
15. Hoskisson, R.E., Hitt, M.A., Johnson, R.A., Moesel, et D. D.: Construct validity of an objective (entropy) categorical measure of diversification strategy. Strat. Manag. J. **14**(3), 215-235(1993)



Investigating the Use of a Distance-Weighted Criterion in Wrapper-Based Semi-supervised Methods

João C. Xavier Júnior, Cephas A. da S. Barreto(✉), Arthur C. Gorgônio, Anne Magály de P. Canuto, Mateus F. Barros, and Victor V. Targino

Federal University of Rio Grande do Norte, Lagoa Nova Campus, Natal 59078900, Brazil

cephasax@gmail.com

Abstract. This paper proposes the use of a more elaborated criterion for both selecting and/or labelling unlabelled instances of a wrapping-based SSL method. In order to assess the feasibility of the proposed method, an empirical analysis will be conducted, in which the proposed Self-training versions are compared to other Self-training versions, some existing SSL methods (i.e. LLGC and GRF), and also a supervised classifier using different proportions of the training set (10 and 90%). The experimental results have shown that the Distance-weighted Criterion has improved the performance of the Self-training method, especially when this criterion is used on the labelling process.

Keywords: Semi-supervised learning · Self-training · selection · labelling

1 Introduction

Besides the continuous growth of computing power, several problems do not have sufficient labelled data to build a strong hypothesis, which leads to poor performance of computational methods and, in turn, these problems have become a challenge for new researches. Semi-supervised Learning (SSL) methods are an interesting alternative to search for efficient solutions to problems with few labelled data. Among the existing SSL methods, the Self-training method [9] has attracted attention in different application domains. It is a wrapping-based method that builds its hypothesis in a wrapping phase that iteratively selects and labels unlabelled instances. Despite its efficiency, the wrapper-based methods can wrongly select and/or label several instances. Since it is an iterative process, these errors are propagated to the next iterations, leading to a snowball effect and, as a consequence, impairing its performance. As a result, some efforts have been made to improve the selection and labelling process of the wrapper-based methods.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 644–654, 2023.
https://doi.org/10.1007/978-3-031-27409-1_59

Recent studies on this subject focus mainly on using an additional apparatus to reduce selection and labelling errors [13] or creating new approaches to select and label instances more accurately [1,2]. In [2], for instance, the authors used a different way to compute the reliability of the unlabelled instances, using a distance-weighted criterion as measure for selecting unlabelled instances. This new measure was applied to select instances in Self-training and Co-training methods, and the obtained results surpassed the standard versions of both methods. In the same context, in [1], an agreement-based measure was presented. The main idea is to propose an efficient way to define the reliance of unlabelled instances based on the performance of a classifier ensemble. According to the authors, the Self-training methods with the proposed measure obtained exceptional results compared to existing SSL methods.

It is important to emphasise that the majority of the wrapper-based SSL studies proposed extensions for selecting unlabelled instances. Nevertheless, an efficient labelling process is also very important to the performance of a wrapper-based SSL method since a wrong label can strongly deteriorate the performance of this method. Therefore, this paper proposes the use of a more elaborated criterion in both selecting and labelling instances in a wrapper-based SSL method. The main aim is to further improve the performance of the wrapper-based SSL methods.

In order to do that, the idea of a distance-weighted criterion (DwC) is used for selecting and/or labelling instances. Therefore, the DwC is adapted for its use in the labelling process. In addition, as the DwC criterion proposes a combination of two measures (confidence and distance), it is also possible to combine any criterion with distance measure to label instances in wrapper-based methods. Then, this work also proposes the use of the agreement measure, as originally proposed in [1], in the DwC criterion. These two proposed approaches can be applied to any wrapper-based SSL method. However, in this paper, they will be applied to Self-training.

Finally, an empirical analysis will compare these new Self-training versions to some Self-training versions (standard and with random selection), some existing SSL methods (i.e., LLGC and GRF), and a supervised classifier using different proportions of training set (i.e., 10 and 90%). This analysis will be done using 35 well-known classification datasets.

2 Related Work

As previously mentioned, the majority of Self-training extensions use selection and labelling criteria based on confidence prediction or distance measure. In addition, recent studies have proposed different strategies to improve the performance of Self-training, with special attention to the selection process. Thus, this section will present some studies related to the mentioned criteria and studies designed to improve the performance of the Self-training method.

Confidence prediction and distance metrics are the most common approaches for the selection of unlabelled instances in wrapper-based SSL methods. In [5], for

instance, the authors used a selection threshold combined with the capabilities of a decision tree as a basis for their Self-training version. In [8], the authors used the confidence prediction along with data density peaks to build a model. Additionally, differential evolution was also used to discover data structure, aiming to help the training, selection and labelling processes.

Distance metrics have also been used to select unlabelled instances in Self-training. In [4], for example, a method to solve video classification tasks was proposed and it uses distance metrics to select the most similar instances at each iteration. In another study [6], the authors also used distance metrics to build the confidence metric of the selection process. This research uses a K-NN algorithm as a noise filter to select only the nearest instances.

Several studies have implemented adaptations in the selection and labelling criteria, aiming to improve the performance of the Self-training method. In [7], for instance, the authors proposed a dynamic confidence threshold to avoid the selection of unreliable instances. This approach, named FlexCon-CS, defines the confidence threshold using the main classifier performance at the previous iteration. The results have shown that this process improved the effectiveness of the Self-training method, mainly when few instances are labelled.

Two recent studies proposed new criteria for selecting unlabelled instances in wrapper-based SSL methods. In [2], for instance, the authors proposed a combined selection criterion, aiming to achieve a better selection performance. This criterion, named Distance-weighted Criterion (DwC), is defined by weighting the confidence value with a distance metric. Thus, this work takes advantage of the two most traditional selection criteria being used together. The proposed approach was applied to Self-training and Co-training and the obtained results were promising, specially for the Self-training method. In [1], a novel criterion was presented. This work used the agreement reached by a classifier ensemble as a selection criterion. The authors applied their proposal to Self-training and Co-training methods. The results showed that the use of this novel criterion improves the performance of the Self-training and Co-training methods.

Overall, it is possible to note that the use of confidence prediction or distance metrics is the most common approach in wrapper-based SSL methods. In addition, these studies proposed extensions with focus on the selection process. Nevertheless, an efficient labelling process also plays a fundamental role in the performance of a SSL method. Unlike the majority of SSL methods, this paper proposes a more elaborated criterion to be used in both selection and labelling instances in the wrapping process.

3 The Proposed Approach

As already mentioned, this paper proposes the use of a more elaborated criterion in the selection and/or labelling of instances in wrapper-based SSL methods. With a more accurate selection or labelling criterion, an improvement in the performance of the SSL method is expected. In order to do this, three Self-training versions are proposed.

- Selection (S): In this proposed method, the proposed criterion is applied to measure all unlabelled instances. After that, the instances with the best values are selected to be labelled. This proposed version is similar to the majority of all extended versions of Self-training.
- Labelling (L): In this version, the proposed criterion is used to define the class that will be assigned to the newly selected instance. In order to do this, this criterion is calculated to each class of a problem. Then, the class with the highest value is assigned to that instance. In this version, as this criterion is used only in the labelling process, the selection process is performed as in the original Self-training method.
- Selection and Labelling (SL): In this version, the proposed criterion is used in both selection and labelling processes.

As it can be observed, the proposed criterion is used in the selection and/or labelling processes. In order to do this, this criterion has to be selected. In this paper, we selected the DwC (Distance-weighted Criterion) metric. In fact, we selected two DwC-based criteria. The first one, named Distance-weighted Criterion-Confidence (DwC-C), uses the confidence and distance to compose the DwC metric. The second one, named Distance-weighted Criterion-Agreement (DwC-A), uses agreement and distance to compose the DwC metric.

3.1 Distance-Weighted Criterion (DwC)

The use of DwC in the selection process was originally defined in [2] and it is presented in Eqs. 1 and 2.

$$\mathrm{DwC}_i = \max(\forall_{j \in C}\ \mathrm{DwC}_{ij}) \tag{1}$$

where:

$$\mathrm{DwC}_{ij} = \ \mathrm{W}_{ij} \times \frac{1}{d_j} \tag{2}$$

where DwC_{ij} is the distance-weighted criterion of instance i to class j; W_{ij} defines the W criterion value for an instance i to the j-th class; C is the set of classes for that problem; and d_j is the distance between instance i and the centroid of class j in the labelled set.

In Eqs. 1 and 2, the W value for an instance i to a particular class j is weighted by its distance to the centroid of the $j-th$ class in the labelled set, providing the DwC_{ij} value. The DwC_{ij} value is computed to all classes, and the highest DwC_{ij} value is used as the DwC_i.

Herewith, the DwC criterion considers both W and distance measure to define the value assigned to a unlabelled instance. It is important to emphasise that the acronym W represents an abstract criterion. Thus, the W criterion can be any criterion that can be used in a SSL method. In this paper, we will investigate two criteria, confidence prediction and agreement.

3.2 Distance-weighted Criterion-Confidence (DwC-C)

The confidence prediction is usually delivered by a classifier and it is provided to all classes of a problem. This measure is used in the DwC criterion as the W parameter in Eq. 2.

The use of confidence prediction in the selection process has already been presented in [2]. In this paper, we adapt this criterion to be also used as a labelling criterion. In order to label instances using DwC-C, based on Eqs. 1 and 2, the following steps are adopted. Let a classification problem have two possible classes (e.g. "a" and "b"). Now, suppose that the confidence prediction for one instance is 0.345 to class "a" and 0.655 to class"b". In addition, suppose that the distances between this instance and the centroids of classes "a" and "b" are, respectively, 0.25 and 4.84. Therefore, the DwC-C criterion for that instance will be 1.38 (class "a") and 0.135 (class "b"). In this sense, as 1.38 is higher than 0.135, that instance will be labelled with the class label "a".

There are three possible Self-training versions with DwC-C. The first one, St-dwc-C(S), uses DwC-C only for selecting unlabelled instances. The second one, St-dwc-C(L), uses DwC-C only for labelling, whereas the third and last version, St-dwc-C(SL), uses DwC-C for selecting and labelling.

3.3 Distance-Weighted Criterion-Agreement (DwC-A)

Classifier ensemble is a classification structure that consists of a set of classifiers (base classifiers), organised in a parallel way, which provides their outputs to a combination method, which gathers all predictions and provides the system output. This system has been widely used in the literature to build a model with better predictive accuracy than a single base classifier [11]. One important measure that can be extracted from a classifier ensemble is agreement. It reflects the level of similarity between the classification predictions obtained by the base classifiers. In order to do this, it counts the number of similar votes among all base classifiers. The main aim of using agreement is to reflect the output consensus of a group, which tends to achieve a more assertive classification.

In [1], the agreement criterion has been used as a selection criterion. On the other hand, this paper presents DwC-A, which is a criterion that combines agreement and a distance measure to build a more robust criterion. The DwC-A value is calculated using Eqs. 1 and 2. Agreement is used in the DwC criterion as the W parameter in Eq. 2. By doing this, the agreement reached by a classifier ensemble is weighted by a distance measure to build the criterion value used for selection and/or labelling. Finally, there are also three DwC-A versions for Self-training, St-dwc-A(S), St-dwc-A(L) and St-dwc-A(SL).

4 Experimental Methodology

This empirical analysis is based on an 10×10-fold cross validation method, and the main steps of the Experimental Framework is described as follows.

1. split the dataset into *10* stratified folds;
2. separate fold 1 for testing (Test set—T);
3. use the remaining folds (2–10) to training, dividing it into 10% of the labelled set (L) and 90% of the unlabelled set (U);
4. build a learning model by performing the SSL implementation using U and L;
5. validate the learning model using T and save the obtained results;
6. repeat steps 2–6 (changing the fold used for testing) until all folds have been used as test set.

This process is repeated 10 times with different data distribution in the folds, generating 100 values. The final result is the average of all 100 values.

In addition, it is applied to 35 classification datasets to evaluate the feasibility of the proposed approach. These datasets are well known and available to download from the UCI Repository.[1] Table 1 presents the description of all datasets.

Table 1. Description of the datasets

No	Dataset	*Inst*	*Att*	*Class*	Type	No	Dataset	*Inst*	*Att*	*Class*	Type
d1	Abalone	4177	8	28	C,N	d19	Musk	6598	168	2	N
d2	Adult	32561	14	2	C,N	d20	Nursery	12960	8	5	C
d3	Arrhythmia	452	260	13	N	d21	Ozone	2536	73	2	N
d4	Automobile	205	26	7	C,N	d22	Pen-digits	10992	16	10	N
d5	Car	1728	6	4	N	d23	Pima	768	9	2	N
d6	Cnae	1080	857	9	N	d24	Planning	182	13	2	N
d7	Dermatology	366	34	6	N	d25	Seeds	210	7	3	N
d8	Ecoli	336	7	8	C,N	d26	Semeion	1593	256	10	N
d9	German-credit	1000	20	2	C,N	d27	Solar-flare	1389	10	6	C,N
d10	Glass	214	10	7	N	d28	Spectf-heart	267	44	2	N
d11	Haberman	306	4	2	N	d29	Tic-tac-toe	958	9	2	C
d12	Hill-valley	606	101	2	N	d30	Twonorm	7400	21	2	N
d13	Ilpd	583	10	6	N	d31	Vehicle	946	18	4	N
d14	Image-seg	2310	19	7	N	d32	Waveform	5000	40	3	N
d15	Kr-vs-kp	3196	36	2	C	d33	Wilt	4839	6	2	N
d16	Mammography	961	6	2	N	d34	Wine	4898	11	11	N
d17	M-features	2000	64	10	N	d35	Yeast	1484	8	10	N
d18	Mushroom	8124	22	2	C						

The obtained results will be presented and discussed in two parts. The first analysis compares eight Self-training versions, six proposed versions (e.g. St-dwc-C (S), St-dwc-C (L), St-dwc-C (SL), St-dwc-A (S), St-dwc-A (L) and St-dwc-A (SL)), a standard version (St-std), and a version with random selection (St-rand). In the second analysis, the best proposed versions (one based on dws-C and one

[1] https://archive.ics.uci.edu/ml/datasets.php.

based on dws-A) will be compared to two existing SSL methods (i.e. LLGC and GRF), and to a supervised classifier using two training set proportions (i.e. 10 and 90%).

The accuracy rate will be used as predictive measure. Additionally, a statistical analysis is performed, using the Friedman and the Nemenyi post-hoc tests [3]. The results of these tests are presented graphically through the critical difference (CD) diagram. This diagram shows the statistical difference between results and places the best results on the left. All methods were developed using the Weka API[2]; a Decision Tree (with confidence factor = 0.05) as base classifier; and 10% was the proportion of unlabelled instances to be selected at each iteration. The second part of the analysis uses four additional methods: two graph-based methods, Learning With Local and Global Consistency (LLGC) [10] and Gaussian Random Fields (GRF) [12], and a Decision Tree with two training proportions (i.e. 10 and 90%). These last two methods also used a confidence factor = 0.05.

5 Experimental Results

This section details the experimental results of the proposed Self-training versions. This analysis is divided in two parts, in which the first one presents the results of all eight Self-training versions. The second part performs a comparative analysis with some existing SSL methods.

5.1 Experimental Results—First Analysis

This subsection presents the results of all eight Self-training versions, standard (St-std), random selection (St-rand), St-dwc-C(S), St-dwc-C(L), St-dwc-C(SL), St-dwc-A (S), St-dwc-A(L) and St-dwc-A(SL). In these acronyms, S stands for the selection process while L stands for the labelling process. Finally, the performance metric used to assess the analysed methods is accuracy. The results will be presented in Table 2. It is important to emphasize that the set of experiments were performed over all 35 sdatasets described in Table 1. Nevertheless, for simplicity reasons, Table 2 presents the average accuracy over all 35 datasets; the average ranking over all datasets (Rank); and the overall number of wins (Wins) for each version.

From Table 2, in relation to the DwC-C versions (columns 4–6), the St-dwc-C(L) version delivered the best result in 3 out of 3 criteria with 71.12% (Avg), 4.49 (Rank) and 3 (wins). In terms of the DwC-A versions (columns 7–9), the St-dwc-A(L) version, that uses the DwC-A approach in the labelling process, obtained the best results in 3 out of 3 criteria, delivering 73.40% (Avg), 2.89 (Rank) and 12 (wins). In fact, St-dwc-A(L) achieved the best performance among all versions, in terms of accuracy.

The statistical results of the first analysis, presented in Fig. 1, show that there is a significant difference between the best proposed version (St-dws-A(L)) and the last three Self-training versions (St-dws-C(SL), St-rand and St-std).

[2] https://www.cs.waikato.ac.nz/ml/weka/.

Table 2. Results—*proposed self-training and baselines*—accuracy

	St-std	St-rand	St-dwc C(S)	St-dwc C(SL)	St-dwc C(L)	St-dwc A(S)	St-dwc A(SL)	St-dwc A(L)
Avg (%)	70.49	70.25	70.97	70.64	71.12	71.84	70.89	**73.40**
Rank	5.17	5.03	4.57	4.91	4.49	4.14	4.57	**2.89**
Wins	4	3	0	3	3	6	6	**12**

In summary, in this first analysis, we can state that St-dwc-A(L) obtained the best overall results. This version uses the DwC-A criterion in the labelling process, indicating that the labelling process with more elaborated criteria can improve the results of a Self-training method even more. It is important to emphasise that all proposed versions surpassed the standard (std) and random (rand) versions in almost all three analysed criteria (avg, rank and wins). The only exception is the DwC-C versions, which St-std surpassed in one criterion (wins).

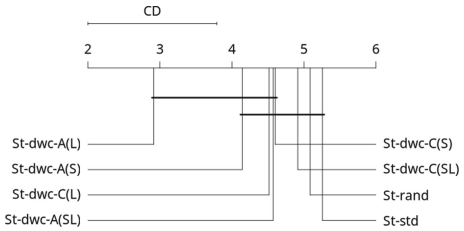


Fig. 1. CD diagram—first part of the experiments

5.2 Experimental Results—Second Analysis

The second analysis compares the two best proposed versions (St-dwc-C(L) and St-dwc-A(L)) with two existing SSL methods and with a supervised method using two training proportions. Table 3 presents the accuracy results of the second analysis. According to this table, it is possible to state that J48 trained with 90% of the original training set obtained the best results in 3 out of 3 criteria, with 79.71% (Avg), 1.57 (Rank) and 27 wins. In addition, the version St-dwc-A(L) obtained the second best result in 2 out of 3 criteria, with 73.40% (Avg) and 2.71 (Rank). Finally, in the third place, the version St-dwc-C(L), with 71.12% (Avg) and 3.6 (Rank).

The statistical results of the second analysis, presented in Fig. 2, reinforces the expected superiority of J48-90%. This supervised method was statistically superior to almost all competing methods. The only exception was the St-dwc-A(L) version that obtained a result statistically similar to J48-90%.

Table 3. Results—second part of the experiments—accuracy

	GRF	LLGC	J48-10%	J48-90%	St-dwc-C(L)	St-dwc-A(L)
Avg (%)	53.85	47.32	70.55	**79.71**	71.12	73.40
Rank	4.26	4.97	3.69	**1.57**	3.60	2.71
Wins	6	2	0	**27**	0	2

In summary, this second analysis presents an expected scenario with the superiority of J48-90%, including the statistical difference highlighted by the CD diagram. This comparison works as a benchmark. The results of the best proposed versions can indicate if the performance improvement promoted by the DwC approaches can take the wrapper-based SSL to the performance level of a supervised method. The statistical similarity pointed out by 2 between J48-90% and St-dwc-A(L) is an example of a result that shows that using a more elaborate labelling criterion (e.g. DwC-A) can take the Self-training performance to the level of a supervised method.

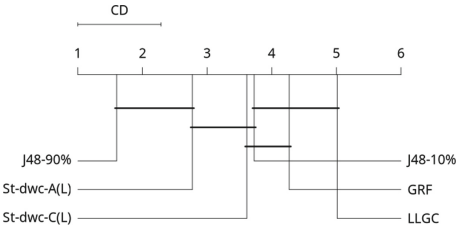


Fig. 2. CD diagram—second part of the experiments

Finally, all proposed approaches increased the Self-training performance, specially when the combined criterion (DwC-C or DwC-A) was used in the labelling process. This is an indication of the importance of the labelling process in a wrapper-based SSL method, since the use of a more elaborated criterion leads to an improvement in the performance of almost all proposed versions.

6 Final Remarks

This paper proposed the use of a more elaborated criterion in the wrapping phase of the Self-training method. In order to do this, we defined the use of a more elaborated criterion for the selection and/or labelling of unlabelled instances in this method. In order to assess the feasibility of the proposed approach, an empirical analysis was conducted, using the Self-training method, 35 classification datasets and being evaluated by accuracy.

Through the empirical analysis, we could observe that the proposed approaches positively affected Self-training's performance, especially when the

distance weighted criterion (DwC-C or DwC-A) was used in the labelling process. The St-std-A(L) version statistically outperformed the St-std method, in terms of accuracy. Additionally, all the proposed versions obtained better results than St-std and St-rand. Then, it is possible to conclude that the proposed approaches positively affected the performance of the Self-training method.

The proposed versions surpassed the performance of J48-10% and graph-based SSL methods. Despite the superiority of J48-90%, the statistical analysis pointed out that the proposed version St-dwc-A(L) obtained a similar performance. This result is promising because it indicates that a more elaborated criterion used in the labelling process can make the Self-training performance similar to a supervised method.

References

1. Barreto, C.A.d.S., Canuto, A.M.d.P., Xavier-Júnior, J.C., Gorgônio, A.C., Lima, D.F., da Costa, R.R.: Two novel approaches for automatic labelling in semi-supervised methods. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
2. Barreto, C.A., Gorgônio, A.C., Canuto, A.M., Xavier-Júnior, J.C.: A distance-weighted selection of unlabelled instances for self-training and co-training semi-supervised methods. In: Brazilian Conference on Intelligent Systems, pp. 352–366. Springer (2020)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**(Jan), 1–30 (2006)
4. Suzuki, T., Kato, J., Wang, Y., Mase, K.: Domain adaptive action recognition with integrated self-training and feature selection. In: 2013 2nd IAPR Asian Conference on Pattern Recognition, pp. 105–109. IEEE, Naha, Japan (2013)
5. Tanha, J., van Someren, M., Afsarmanesh, H.: Semi-supervised self-training for decision tree classifiers. Int. J. Mach. Learn. Cybern. **8**(1), 355–370 (2015). https://doi.org/10.1007/s13042-015-0328-7
6. Triguero, I., SÃez, J.A., Luengo, J., García, S., Herrera, F.: On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. Neurocomputing **132**, 30–41 (2014). https://doi.org/10.1016/j.neucom.2013.05.055
7. Vale, K.M., Canuto, A.M.d.P., Gorgônio, F.L., Lucena, A.J., Alves, C.T., Gorgônio, A.C., Santos, A.M.: A data stratification process for instances selection in semi-supervised learning. In: International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
8. Wu, D., Shang, M., Luo, X., Xu, J., Yan, H., Deng, W., Wang, G.: Self-training semi-supervised classification based on density peaks of data. Neurocomputing **275**, 180–191 (2018). https://doi.org/10.1016/j.neucom.2017.05.072
9. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd Annual Meeting on Association for Computational Linguistics, pp. 189–196 (1995). https://doi.org/10.3115/981658.981684
10. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, pp. 321–328 (2004)
11. Zhou, Z.H.: Ensemble methods: foundations and algorithms. Chapman and Hall/CRC (2012)

12. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 912–919 (2003)
13. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5982–5991 (2019)



Elections in Twitter Era: Predicting Winning Party in US Elections 2020 Using Deep Learning

Soham Chari(✉), Rashmi T, Hitesh Mohan Kumain, and Hemant Rathore

Department of CS and IS, K K Birla Goa Campus, BITS Pilani, India
{h20210029,h20210020,h20210068,hemantr}@goa.bits-pilani.ac.in

Abstract. Twitter has become one of the leading platforms for people to express their political opinions during elections. Understanding public sentiment by on-ground election polling is expensive and time-consuming. Recently, tweets have gained popularity for analyzing public sentiment toward political parties. However, a large number of tweets have to be analyzed to simulate real-world elections effectively. Most previous works on Twitter sentiment analysis have trained on small tweet datasets, and most have failed to consider negative sentiment tweets. Considering only the positive tweets does not produce the actual on-ground sentiment. In this study, we have analyzed a large corpus of positive, negative, and neutral sentiment tweets about political parties in the US elections 2020 and predicted the election outcome. We constructed three distinct classification models using *Bidirectional LSTM, GRU, and Hybrid CNN-LSTM* to classify tweets into positive, negative, or neutral sentiments. Our results show that Bidirectional LSTM achieved an accuracy of 95.99%, outperforming other deep learning approaches and related works. A custom net-score metric is proposed that considers both positive and negative sentiments. The Democratic Party outperformed the Republican Party with a higher net score, indicating that the *Democratic Party* is predicted to win. VADER algorithm was used to find the winning margin, which considered polarity, compound sentiment score, and retweet count. We found that the *Democratic Party* would lead by a marginal gap of +6.250%, which is close to the actual election results.

1 Introduction

Views, opinions, and feelings of a person play a vital role in their critical decision-making. Analyzing the sentiments of a large group of people can predict the overall sentiment of an entire village, town, or even nation. Thus, sentiment analysis of social media such as Twitter can provide crucial insights into electoral preferences, governance, consumer markets, and stock markets. Many researchers like Tumasjan et al., and Boutet et al. compared the volume of tweets that mention different election candidates/parties [5,22]. They obtained voting results close to on-ground electoral polls. Tumasjan et al. did a word count-based analysis [22].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 655–667, 2023.
https://doi.org/10.1007/978-3-031-27409-1_60

They found that tweets' overall sentiment comes close to that displayed during ground-level electoral programs and media coverage. However, word count-based text analysis (e.g., LIWC) has some limitations. Since tweets are essentially phrases or sentences, considering only the word count will miss vital contextual and semantic information.

The Valence Aware Dictionary for sentiment Reasoning (VADER) was developed to counter the above problem [11]. VADER is a dictionary-based approach that uses grammatical and syntactic rules applied by humans while displaying intense sentiment. The rules are generalizable for different text types, like tweets, movie reviews, and news editorials. Several studies have used the VADER algorithm or similar techniques for sentiment analysis of election-related tweets. However, none of the previous works (to the best of our knowledge) have performed descriptive analysis to predict the winning margin using the VADER algorithm.

Sentiment analysis has been used to predict election outcomes by considering only positive sentiment tweets towards a political party or both positive and negative tweets. Most of the authors have considered only positive tweets for predicting the likely winner of elections [8,17,19]. Using only positive sentiment tweets has the same disadvantage as using tweet volume alone, as shown in [19]. If only tweet volume were considered to predict the winning party, negative sentiment tweets against a party would be erroneously added to the tweet count. Similarly, if only positive sentiment tweets are considered to predict the winning party, we would miss vital negative tweets that could clearly show preference against a party. Even negative sentiments play an essential role in political events. Considering both positive and negative sentiments in a large tweet dataset can help us understand the elections better and bring us closer to on-ground election sentiment. Previous studies like Ramteke et al. have used only 56,037 tweets for training and testing their model for US Elections 2020 [17]. Nugroho et al. have used around 11K tweets for their US elections tweet dataset [14]. On the other hand, we have analyzed 1,506,097 tweets for this study. We have considered both positive and negative tweets in the net score to predict the winning party. All these factors ensured that our study was comparable with actual election results.

In this work, we have used different deep learning approaches, namely, Bidirectional Long Short-Term Memory (Bidirectional LSTM), Gated Recurrent Unit Networks (GRU), and Hybrid Convolutional Neural Networks-Long Short Term Memory (Hybrid CNN-LSTM) for sentiment analysis. We train the different classification models on tweets related to the US Elections 2020 to classify each tweet into one of the three sentiment classes (*Positive Sentiment*, *Negative Sentiment*, *Neutral Sentiment*) towards each political party. In addition, sentiment scores obtained using the VADER algorithm are used to perform a comparative analysis of the political parties to predict the winning margin. We propose the net-score metric that predicts the likely winner of the elections considering both positive and negative sentiments. We observed that the *Democratic Party* would lead by a large margin in sentiment score when retweets were considered.

Our key contributions are summarized as follows:

- We propose a framework to predict the winning party of US Elections 2020 and the corresponding winning margin. We used tweets to analyze the political sentiments of people and gauge the favorability of one party against others.
- We designed a dictionary-based approach using the VADER algorithm and developed three deep learning models (Bidirectional LSTM, CNN-LSTM, and GRU) to classify US Elections 2020 tweets into either positive, negative, or neutral sentiment.
- We evaluate and compare the performance of each deep learning-based classification model. Bidirectional LSTM achieves the highest accuracy of 95.99% and an AUC of 0.96.
- We also develop a custom net-score metric by considering both positive and negative tweets towards each party to find the winning party. Our descriptive analysis achieved a marginal gap close to the actual election results.

2 Related Work

Sentiment analysis has become one of the core techniques for analyzing public opinion, even in political events such as elections. Many researchers have used Twitter sentiment analysis to evaluate the general sentiment towards political parties. Subramanian et al. suggested that dictionary-based and deep learning algorithms are efficient methods for sentiment analysis [21]. Dictionary-based approaches such as VADER rely on a dictionary that maps lexical features to emotion intensities known as *sentiment scores* [11]. It is sensitive to polarity (positive/negative) and emotional intensity. Chaudhry et al. used VADER to perform sentiment analysis of tweets before and after the US Elections 2020 [7].

Deep learning algorithms (e.g., RNNs, CNNs) started gaining popularity due to their ability to build more context-aware and domain-specific models for sentiment analysis. Hidayatullah et al. performed Twitter sentiment analysis for the 2019 Indonesian Presidential Election by utilizing three machine learning (SVM, Logistic Regression, Multinomial Naive Bayes) and five deep learning algorithms (LSTM, CNN, CNN-LSTM, GRU-LSTM, and BiLSTM) [10]. The Bidirectional LSTM model outperformed both the machine learning and deep learning algorithms. Gaikar et al. proposed an LSTM-based sentiment analysis for the Indian Lok Sabha Elections 2019 [8]. Researchers like Pedipina et al. proposed GRU-based frameworks to understand the winning chances of any political party and the political response of people to the Delhi Elections [15]. Convolutional Neural Networks (CNNs) have been gaining popularity in sentence-based applications due to their ability to extract local features in the text. Liao et al. proposed the idea of a shallow CNN for semantic analysis of 10,662 Twitter data and achieved an accuracy of 74.5% [13]. A larger tweet dataset could help achieve better classification performance, as discussed in [13]. In tweets (similar to sentences), looking at the nearby words would only convey some meaning. We also need to consider and analyze long-term patterns in the text. Literature suggests

that hybrid CNN-LSTM models achieve better results in such situations. Jain et al. proposed a hybrid CNN-LSTM model for Twitter airline sentiment analysis that outperformed other methods like CNN, LSTM, Naive-Bayes classifier, SVM, Logistic Regression, and Decision Trees [12].

Similar studies have been performed for the US Elections 2020. Xia et al. trained a Multi-Layer Perceptron (MLP) on the Sanders Benchmark Twitter dataset and tested it on 260,498 US Elections 2020 tweets [23]. They predicted a less than 1% difference in ratings between the two parties. Similarly, Chandra et al. trained BERT and LSTM models on the IMDB dataset and tested them on the US Elections 2020 dataset [6,18]. They separated the tweets based on geolocation and predicted the state-wise election outcome. The LSTM model performed better while training using the IMDB dataset. However, the BERT model better predicted the Trump, Biden, and Contentious States for US Elections 2020. Singh et al. extracted tweets related to US Elections 2020 and performed sentiment analysis using BERT, achieving an accuracy of 94% [20]. Chaudhry et al. used the Naive Bayes technique to understand the shift in sentiment before and after the US Elections 2020 [7]. They performed sentiment classification and predicted the margin of victory in 50 different states.

Our study performs VADER-based descriptive analysis by considering factors like compound sentiment score, polarity, and retweet count together. These factors collectively provide greater insight into understanding the overall sentiment in the Twitter community. However, most of the previous works have analyzed these factors independently. We also apply distinct deep learning techniques (Bi-LSTM, CNN-LSTM, GRU) for tweet-wise sentiment classification of US Elections 2020 Twitter data.

3 Proposed Framework

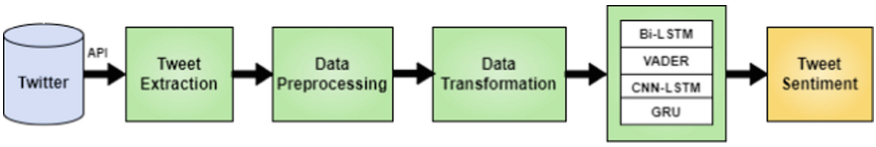


Fig. 1. Proposed Framework for Sentiment Analysis of US Elections 2020 Tweets

Figure 1 explains the overall proposed approach for sentiment analysis of US Elections 2020 tweets to predict the winning party. We extracted 6,000,000 tweets using Twitter API that correspond to Tweet IDs in the US Elections 2020 dataset [3,18]. Then the extracted tweets and metadata were integrated with the US elections dataset. The unimportant columns, duplicate, missing, and non-English tweets were removed from the collected data, resulting in 1,506,097 tweets. The tweets were further preprocessed by reducing them to lowercase,

removing noisy data such as URLs, HTML elements, and hashtags. Since the deep learning models cannot process raw text data, the tweets were transformed into numerical word embeddings using the GloVe pre-trained word embeddings model (discussed in Sect. 4.1). Different deep-learning algorithms (BiLSTM, GRU, and CNN-LSTM) were used to classify the tweets into different sentiment classes. The VADER algorithm was used to generate sentiment scores for each tweet. The scores were then used for descriptive analysis of sentiment in the tweet data. Section 4.2 describes the Bidirectional LSTM-based approach to classify the transformed tweets into different sentiment classes (Positive, Negative, and Neutral). Section 4.3 explains the VADER algorithm for sentiment scores. Section 4.4 describes the hybrid CNN-LSTM classifier, and Sect. 4.5 discusses the GRU-based classifier for sentiment classification. Tweets related to each political party are then classified into their corresponding sentiment. Aggregated sentiment towards a political party is calculated using a net-score metric based on the number of tweets in each class. The net score indicates which party is most likely to win the US Elections 2020. The experimental results and comparison of each method are discussed in Sect. 5.

4 Experimental Setup

This section will discuss the data preprocessing steps, classification models & their architecture, and the performance metrics used to evaluate the models.

4.1 Data Preprocessing

US Elections 2020 dataset is a collection of tweet metadata compiled by Sabuncu et al. dated from July 1, 2020, to November 11, 2020 [18]. It consists of tweet metadata such as tweet ID, party name, positivity/negativity score for each tweet, and an overall sentiment score. The Tweet ID is used to identify each tweet uniquely. We used Tweet IDs to extract the actual tweets using the Twitter API. The extracted data was integrated with the US elections dataset. The tweets from deleted/banned user accounts were unavailable through the Twitter API. Hence, only the tweets from legitimate user accounts as per the Twitter guidelines were considered for the dataset integration [4]. The integrated dataset was then reduced by removing unimportant fields such as user IDs, negativity and positivity scores, entities, number of tokens, and tweet IDs. Non-English, duplicate tweets, and missing values were also removed from the dataset. The final dataset consisted of $1,506,097$ tweets. The tweets were further preprocessed by converting the text to lowercase and removing URL links, HTML tags, hashtags, user mentions, and special characters. This helps in reducing noise in the dataset. Since models cannot handle direct text data, the tweets must be initially transformed into numerical data. Each tweet is tokenized into a set of individual words. Once the dataset is split into train, test, and validation sets, the tokenized training data is used to create a vocabulary of words. The vocabulary is used to number each unique word, resulting in a number sequence representation for

each tweet. The number sequences are padded with zeros to bring all sequences to the same length.

Direct usage of sequence representation will result in many dimensions (up to maximum tweet length). Plus, classifier models must learn complex relationships between words before classifying tweets into sentiment classes. This will result in a massive number of learnable parameters and the curse of dimensionality. Transfer learning can reduce this by using GloVe pre-trained word embeddings model [16]. GloVe is trained with massive text datasets to derive co-occurrence relationships between words, and thus it can reduce this burden from the classifier models. Sentiment scores are used to label the tweets as positive, negative, or neutral. After sufficient experimentation, the threshold for tweet classification was set to 0.5. Tweets with scores below -0.5 were labelled as negative, above 0.5 as positive, and tweets between -0.5 and 0.5 (inclusive) as neutral.

4.2 Bidirectional LSTM Classifier

LSTM is an extension of recurrent neural networks that can understand the context within long sequences and learn long-term dependencies. They maintain this information using a cell state in addition to the existing hidden state of RNNs. This cell state helps in long-term memory by storing and loading previous events. Bidirectional LSTMs contain two LSTM units, one which processes input in the forward direction and the other processes input in the backward direction. This helps the output layer to get information from past and future states simultaneously.

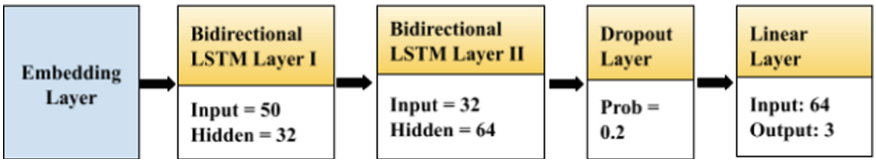


Fig. 2. Bidirectional LSTM Model Architecture

Figure 2 shows the architecture of the Bidirectional LSTM model followed in the paper. Firstly, a word embedding layer is the input layer that takes two arguments: input size and embedding size. The input size is the size of the vocabulary of the dataset. This size was restricted to 250,000 tokens to follow computational constraints. Furthermore, while building the vocabulary, GloVe pre-trained word vectors were used [16]. After the embedding layer, a stack of two Bidirectional LSTM layers with 32 hidden units was added. The number of hidden units was fixed to 32 after performing sufficient experimentation. A dropout layer was added after these Bidirectional LSTM layers. The dropout layer helps prevent overfitting by randomly setting a set of internal units to 0. Finally, a dense, fully connected layer was the output layer with an input size

of 64 and an output size of 3 (for Positive, Negative, and Neutral sentiment classes).

4.3 VADER Algorithm for Sentiment Analysis

The VADER algorithm works as a dictionary-based approach [11]. The tweets are preprocessed and tokenized, and each word/token is mapped with a sentiment score.

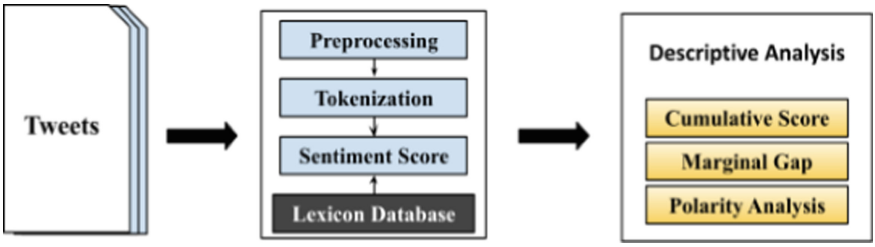


Fig. 3. Framework for VADER-based Sentiment Analysis

Figure 3 shows the overall framework followed for the VADER algorithm-based analysis. After data preprocessing, tweets were separated based on their corresponding party preference into four sets, namely, *Republicans*, *Democrats*, *Both Party*, and *Neither Party*. The VADER algorithm calculates positive, negative, and neutral sentiment scores and a compound sentiment score for each tweet. The compound score is the normalized sum of positive, negative, and neutral sentiment scores. The compound scores are used to calculate the mean compound score for each set, the marginal gap (which is used to find the winning margin), and the sentiment polarity after considering the retweet counts (i.e., the number of times a particular tweet was retweeted). The number of retweets indicates the popularity of the tweet's sentiment and how far it has traveled through Twitter. Analysis of sentiment score and retweet count reveals the overall sentiment of the Twitter community towards each political party.

4.4 Hybrid CNN-LSTM Classifier

Hybrid CNN-LSTM combines two crucial properties of CNNs and LSTMs. CNNs help find local context within data, while LSTMs find long-term sentence relations. This is very useful for classifying sentence-based tweet data. The preprocessed tweet data is converted to 50-dimension word embeddings using the GloVe pre-trained model. These non-trainable weights form the embedding layer in our model. The model consists of two convolution blocks, a bidirectional LSTM block, and dense layers with softmax for classification. In total, the model has 65,155 trainable parameters. The first convolution block has a convolution layer

with 128 filters and kernel size 3, followed by a max-pooling layer with kernel size 3. The second convolution block has a convolution layer with 64 filters and kernel size 4 followed by a max pool layer with kernel size 2. LSTM block consists of a single bidirectional LSTM layer of 32 units to learn long-term dependencies in the tweet. It has a dropout of 0.3, so the model tries different neural paths to reduce error and perform better generalization. Output from the LSTM layer is fed to a dense layer with 16 neurons with a dropout of 0.3. A final dense layer of three neurons uses softmax to generate the probabilities of whether the tweet exhibits positive, negative, or neutral sentiment.

4.5 GRU Classifier

Gated Recurrent Unit is similar to an LSTM but has fewer parameters as it lacks the output gate, unlike an LSTM. GRUs are also very capable of understanding long-term dependencies. They perform comparatively well in terms of time consumption and memory utilization. We followed a model architecture similar to the Bidirectional LSTM model. Again, while building the vocabulary, GloVe pre-trained word vectors were used [16]. The model architecture consisted of an embedding layer with the same configuration as LSTM. It is followed by a stack of two GRU layers with 64 hidden units and a dropout of 0.2. Finally, it has a dense, fully connected layer as the output layer.

4.6 Performance Measures

We have used accuracy, precision, recall, F1 score, and ROC-AUC score to evaluate the performance of our proposed classification models. *Accuracy* is calculated as the ratio of the number of correct predictions to the total count. The *precision* score is the number of correct positive predictions to the total number of positive predictions. The *recall* score is the number of correct positive predictions to the total number of actual positives, and it denotes the sensitivity of the predictions to the actual labels. The *F1-score* conveys the balance between precision and recall. The *ROC-AUC* (Area Under the Receiver Operating Characteristic Curve) score measures the ability of the model to distinguish between different classes.

The *Compound Score* (CS) is the sum of positive, negative & neutral sentiment scores, normalized between −1 (extremely negative) and +1 (extremely positive). We have used *Marginal Gap* (MG) to understand the relative favorability for a political party and to calculate the winning margin [2]. It is the mean compound score difference between significant parties in their respective range. Mathematically, the marginal gap is obtained as follows:

$$MG = \frac{Avg.\ CS\ of\ A - Avg.\ CS\ of\ B}{Range} \times 100 \tag{1}$$

where A and B are political parties under consideration.

The net sentiment score is used to predict the likely outcome of the election. The net sentiment score can be defined as the difference between the positive

to the total ratio (PvT ratio) and the negative to total ratio (NvT ratio) [17]. Mathematically:

$$Net\ Score_i = \frac{P_i\ -\ N_i}{TotalCount_i} \tag{2}$$

where, i is the political party, P is Number of tweets predicted as Positive Sentiment for party i, N is Number of tweets predicted as Negative Sentiment for party i, TotalCount is the total number of tweets about party i.

5 Experimental Results

This section will discuss results of different sentiment classification models and VADER algorithm-based analysis.

Table 1. Performance Comparison of the Deep Learning Models for Sentiment Classification

Approach	Accuracy	Precision	Recall	F1 Score	ROC AUC
Bidirectional LSTM	**95.99**%	**96.00**%	**96.11**%	**96.06**%	**96.99**%
CNN-LSTM	91.16%	91.51%	91.21%	91.33%	93.20%
GRU	95.68%	95.74%	95.72%	95.73%	96.71%

We have performed various experiments using different classification algorithms for sentiment analysis of a large number of tweets about political parties in the US Elections 2020. Table 1 explains the performance of each of our classification models. Testing was done with a separate set of tweets that were not used during training and evaluation. The Bidirectional LSTM model achieved the highest accuracy of 95.99%, followed by the GRU model (95.68%) and the CNN-LSTM model (91.16%). Bidirectional LSTM also achieved the highest precision of 96.00%, the highest recall of 96.11%, the highest F1 score of 96.06%, and the highest ROC AUC score of 96.99%. It was followed closely by the GRU model in terms of precision (95.74%), recall (95.72%), F1 score (95.73%), and ROC AUC score (96.71%). The CNN-LSTM model trailed with a precision of 91.51%, recall of 91.21%, F1 score of 91.33%, and ROC AUC score of 93.20%. In conclusion, Bidirectional LSTM outperformed CNN-LSTM and GRU models for sentiment analysis of US election 2020 tweets out of the three state-of-the-art deep learning classifiers. It performed better concerning all the performance metrics.

Since the BiLSTM model performed the best, it was used to predict the winning party in the US Elections 2020 using the net score metric. A net score of -0.1912 was obtained for the Republican Party and -0.0767 for the Democratic Party, calculated using equation 2. Based on these results, net positive sentiment towards the Democratic Party is higher than that of the Republican

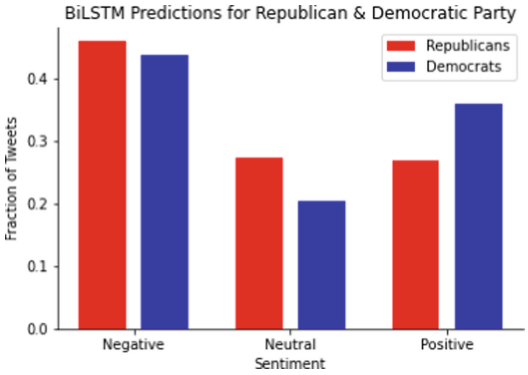


Fig. 4. BiLSTM Predictions for Democratic and Republican Tweets

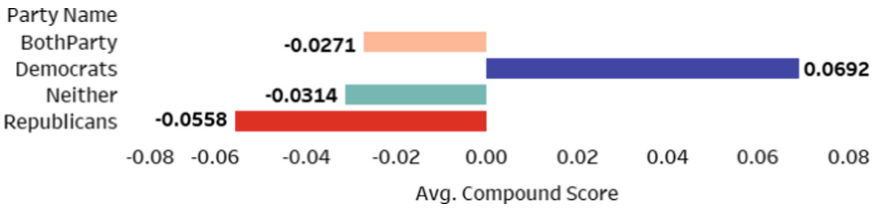


Fig. 5. Average Compound Sentiment Score for the Democratic and Republican Parties

Party. Figure 4 shows the tweet classification results for the Republican and Democratic parties.

Figure 5 shows the mean compound score obtained using the VADER algorithm. The distance between two points defines the gap between them, considering it under its range determines the margin. The marginal gap calculated using equation-1 between Democrats and Republicans is 6.25%. This indicates that the Democratic Party is leading with a winning margin of 6.25% compared to the Republicans, which is very close to the actual US Elections 2020 results [1]. As per the vote count, Biden won 81,284,666 votes, whereas Trump won 74,224,319. Therefore the margin of vote count observed is 7,060,347, which is about 4.54%. Secondly, the sentiment polarity of the Democratic Party is positive, while that of the Republicans is negative.

Figure 6 shows each party's average compound score after considering the retweet counts. It is obtained by averaging the (compound score * retweet count) for each tweet. Democrats lead by a large margin when retweets are considered. Figure 6 indicates that people primarily retweet negative sentiments towards political parties. Secondly, the Republican Party faced high negative sentiment in the Twitter community compared to the Democratic Party.

Table 2 compares our proposed work with existing literature that predicts election results. Most of the studies have considered a limited number of tweets

Table 2. Comparison with Existing Literature

Author(s)	Tweet Volume	Approach	Accuracy	Predicts Win Margin	Tweets for Winning Party Prediction
Hasan [9]	6250	NB WSD	79%	No	No
Hidayatullah [10]	115931	Bi-LSTM	84.60%	No	No
Pedipina [15]	5452	GRU	85%	No	No
Gaikar [8]	15000	LSTM	-	No	P
Singh [20]	-	BERT	94%	No	Both P & N
Xia [23]	260498	MLP	81.53%	Yes, < 1%	Both P & N
Chandra [6]	1170000	BERT,LSTM	88.1%	Yes, State-Wise	P, Ne & N
Chaudhry [7]	18432811	Naive Bayes	94.58%	Yes, State-wise	Both P & N
Proposed	1506097	Bi-LSTM	95.99%	Yes, 6.25%	Both P & N

* P - Positive Tweets. N - Negative Tweets. Ne - Neutral Tweets

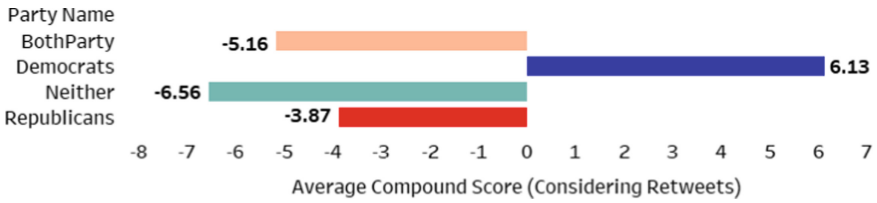


Fig. 6. Average Compound Sentiment Score considering Retweets

in predicting the winning party, which might not lead to an accurate prediction of the on-ground election results [8–10,15,23]. On the other hand, they have not made any predictions for the winning margin or used only positive sentiment tweets. Some studies have predicted the winning margin in each state separately and did not provide such insights into the overall election results [6,7]. Comparatively, our proposed work considers a large volume of tweets, achieves better performance predicting the tweet sentiment, and considers both positive and negative sentiments in predicting the winning party using the net score metric. Also, our descriptive analysis yields a margin of win close to the actual election results.

6 Conclusion

Sentiment analysis of social media platforms (like Twitter) in the political domain has become essential in understanding public opinion for political events (like elections). This study performed sentiment analysis on US Election 2020 Twitter data using Bidirectional LSTM, VADER algorithm, CNN-LSTM, and GRU and classified tweets into either positive, negative, or neutral sentiments. The Bidirectional LSTM model outperformed other deep learning approaches with an accuracy of 95.99%. This model's predictions gave a net score of −0.1912 for the Republican Party and −0.0767 for the Democratic Party. Since the net score for the Democratic is greater than the Republican Party, the Democratic Party is predicted to be the likely winner of the US Elections 2020. The marginal gap observed was +6.25%, and a large margin was observed while considering

the retweet counts in VADER. This shows that the Democratic Party is most likely to win the US Elections 2020 with a winning margin of +6.25%.

As an extension of this study, election data from other social media platforms could also be considered. Secondly, it could be helpful to include emojis and emoticons in tweets as they hold a weighted sentiment value. Since many tweets are sarcastic, their sentiment might get misinterpreted by the classification models. Adding sarcasm detection algorithms to the framework could help improve the sentiment classification models.

References

1. Presidential Election Results: Biden Wins. www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html (2020). Accessed Jan 2022
2. Marginal Gap using Relative Difference. www.en.wikipedia.org/wiki/Relative_change_and_difference/ (2021). Accessed: Jan 2022
3. Twitter Developer API. https://developer.twitter.com/en/products/twitter-api/ (2021). Accessed Oct 2021
4. Twitter's Platform Manipulation and Spam Policy. www.help.twitter.com/en/rules-and-policies/platform-manipulation (2021). Accessed Oct 2021
5. Boutet, A., Kim, H., Yoneki, E.: What's in your tweets? I know who you supported in the UK 2010 general election. In: International AAAI Conference on Web and Social Media, vol. 6, pp. 411–414 (2012)
6. Chandra, R., Saini, R.: Biden vs Trump: modeling US general elections using BERT language model. IEEE Access **9**, 128494–128505 (2021)
7. Chaudhry, H.N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z.I., Shoaib, U., Janjua, S.H.: Sentiment analysis of before and after elections: twitter data of US election 2020. Electronics **10**(17), 2082 (2021)
8. Gaikar, D., Sapare, G., Vishwakarma, A., Parkar, A.: Twitter sentimental analysis for predicting election result using LSTM neural network. Int. Res. J. Eng. Technol. (IRJET) **06**, 3665–3670 (2019)
9. Hasan, A., Moin, S., Karim, A., Shamshirband, S.: Machine learning-based sentiment analysis for twitter accounts. Math. Comput. Appl. **23**(1), 11 (2018)
10. Hidayatullah, A.F., Cahyaningtyas, S., Hakim, A.M.: Sentiment analysis on twitter using neural network: Indonesian presidential election 2019 Dataset. In: IOP Conference Series: Materials Science and Engineering, vol. 1077. IOP (2021)
11. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
12. Jain, P.K., Saravanan, V., Pamula, R.: A hybrid CNN-LSTM: a deep learning approach for consumer sentiment analysis using qualitative user-generated contents. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **20**(5), 1–15 (2021)
13. Liao, S., Wang, J., Yu, R., Sato, K., Cheng, Z.: CNN for situations understanding based on sentiment analysis of twitter data. Procedia Comput. Sci. **111**, 376–381 (2017)
14. Nugroho, D.K.: US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis. In: 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 136–141. IEEE (2021)

15. Pedipina, S., Sankar, S., Dhanalakshmi, R.: Sentimental analysis on twitter data of political domain. In: Computer Networks, Big Data and IoT, pp. 205–216. Springer (2021)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: The Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
17. Ramteke, J., Shah, S., Godhia, D., Shaikh, A.: Election result prediction using Twitter sentiment analysis. In: International Conference on Inventive Computation Technologies (ICICT). vol. 1, pp. 1–5. IEEE (2016)
18. Sabuncu, I.: USA Nov. 2020 Election 20 Mil. Tweets (with Sentiment and Party Name Labels) Dataset (2020). www.dx.doi.org/10.21227/25te-j338
19. Shi, L., Agarwal, N., Agrawal, A., Garg, R., Spoelstra, J.: Predicting us primary elections with twitter (2012). In: Workshop Social Network and Social Media Analysis: Methods, Models and Applications (NIPS) (2012)
20. Singh, A., Dua, N., Mishra, V.K., Singh, D., Agrawal, A., et al.: Predicting elections results using social media activity a case study: USA presidential election 2020. In: 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 314–319. IEEE (2021)
21. Subramanian, R.R., Akshith, N., Murthy, G.N., Vikas, M., Amara, S., Balaji, K.: A survey on sentiment analysis. In: 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 70–75. IEEE (2021)
22. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: International AAAI Conference on Web and Social Media, vol. 4, pp. 178–185 (2010)
23. Xia, E., Yue, H., Liu, H.: Tweet sentiment analysis of the 2020 US presidential election. In: Companion Proceedings of the Web Conference 2021, pp. 367–371 (2021)



Intuitionistic Multi-criteria Group Decision-Making for Evacuation Modelling with Storage at Nodes

Evgeniya Gerasimenko(✉) and Alexander Bozhenyuk

Southern Federal University, Taganrog, Russia
{egerasimenko,avbozhenyuk}@sfedu.ru

Abstract. In this paper, we consider an algorithm for emergency decision-making in fuzzy intuitionistic environment. To transport the maximum number of aggrieved from the dangerous area to the safe destination, a dynamic flow model with transit arc capacities is constructed. The intermediate nodes of the network can store the flow in order for the flow to be maximized. Uncertain experts' evaluations and high level of hesitance are incorporated into the decision-making process as fuzzy intuitionistic numbers. Multi-attribute group decision-making is used to rank the intermediate shelters to evacuate the maximum possible number of aggrieved. In the method, experts have different weights for different attributes, which allows considering the degree of experts' competence for different attributes. The attribute weights are not known beforehand and are defined during the algorithm. A case study is conducted to illustrate evacuation of the maximum number of aggrieved with intermediate location at nodes with limited capacities in order to transport evacuees to the safe destination based on modified fuzzy intuitionistic TOPSIS.

Keywords: Intermediate storage · Intuitionistic fuzzy sets · TOPSIS

1 Introduction

Throughout the world history, disasters and hazard events have spontaneously occurred and caused severe damage to life, property and society. Therefore, countries all over the world pay great attention to emergency. Hazard events are divided into natural, man-made and technological [1].

Emergency decision-making is one of the most important parts of decision theory. Owing to complex environment, lack of information about alternatives, it is difficult to give the exact evaluations of attributes. Moreover, experts often express hesitation and uncertainty while making decisions. In this regard, many valuable tools have been developed to simulate uncertainty while decision-making. Fuzzy sets proposed by Zade [2] indicated an expert's uncertainty in the form of a membership function, which shows the degree of belonginess of the element to the set. Later, various extensions of fuzzy sets were proposed which represent various degrees of experts' doubts about the specific value of membership degree. The following are representatives: type-2 fuzzy sets, fuzzy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 668–679, 2023.
https://doi.org/10.1007/978-3-031-27409-1_61

multisets, intuitionistic fuzzy sets, intuitionistic soft fuzzy sets, linguistic arguments, hesitant fuzzy sets. Intuitionistic fuzzy set consists of membership degree of an element to the set, non-membership degree and degree of hesitation [3]. Type-2 fuzzy set presents the membership of a given element as a fuzzy set. Type-*n* fuzzy set generalizes type-2 fuzzy set allowing the membership to be type- n-1 fuzzy set. In fuzzy multiset, the elements can be repeated more than once. Hesitant fuzzy set appears when a decision-maker has some possible values of attributes and is not sure what to choose so that using a set of possible membership degrees to assess the attribute [4]. In this paper, experts' evaluation will be presented as fuzzy intuitionistic numbers in order to rank the shelters for evacuation.

Multiple and conflicting objectives inherent in decision-making along with ambiguity and uncertainty make decision-making problems complex and difficult [5]. Multi-attribute group decision making is widely used in decision theory since the single experts cannot provide the true evaluations of each attribute. The tasks of real-life emergency decision-making are becoming complex and require much specific knowledge. Therefore, the experience of multiple experts is needed to make reasonable decisions. Experts' weights are often considered to be equal [6, 7] or given beforehand [8, 9], which can lead to the incorrect results. Due to various parameters incorporated into the decision-making process, experts should evaluate various attributes using different weights [10].

In TOPSIS, experts evaluate the alternatives based on the values of closeness coefficients. These values are defined based on positive and negative ideal solutions. The best alternative is considered to be the nearest to the positive ideal alternative and the farthest from the negative ideal alternative. Authors [11] applied fuzzy sets and their extensions and its extensions to handle uncertainty while making decisions based on TOPSIS.

The main contribution of this study is a fuzzy maximum lexicographic dynamic flow algorithm based on the multiple attribute group decision making method. The difference of the method from existing that it allows us to rank the terminals during evacuation based on TOPSIS in intuitionistic fuzzy setting.

2 Basic Concepts and Definitions of Intuitionistic Fuzzy Sets

Fuzzy sets were introduced by L. Zade in order to describe inherent in reasoning and evaluations uncertainty. Fuzzy sets handle membership of an element to a set to indicate the grades of uncertainty. Intuitionistic fuzzy sets as a generalization of fuzzy sets were proposed by Atanassov in 1986 [3]. In intuitionistic fuzzy set, there are membership function and non-membership function which show the hesitance of a decision-maker. In addition, there is an intuitionistic index that indicates the level of expert's uncertainty.

Definition 1 . *Let* $X \neq 0$ *be a reference set. An intuitionistic fuzzy set* $\tilde{A}$ *is defined as.*

$$\tilde{A} = \{\big(x, \mu_{\tilde{A}}(x), \nu_{\tilde{A}}(x)\big); x \in X\}, \tag{1}$$

where $\mu_{\tilde{A}}(x) : X \to [0, 1]$ and $\nu_{\tilde{A}}(x) : X \to [0, 1]$ satisfy the condition $0 \leq \mu_{\tilde{A}}(x) + \nu_{\tilde{A}}(x) \leq 1$, for each $x \in X$.

An intuitionistic index $\pi_{\tilde{A}}(x), 0 \leq \pi_{\tilde{A}}(x) \leq 1$, that indicates the degree of uncertainty is:

$$\pi_{\tilde{A}}(x) = 1 - \mu_{\tilde{A}}(x) - \nu_{\tilde{A}}(x). \tag{2}$$

Let $\tilde{A} = (\mu_{\tilde{A}}, \nu_{\tilde{A}})$ and $\tilde{B} = (\mu_{\tilde{B}}, \nu_{\tilde{B}})$ be IFS of the set X; then the basic operations with IFS are defined as follows:

$$\tilde{A} \oplus \tilde{B} = (\mu_{\tilde{A}}(x) + \mu_{\tilde{B}}(x) - \mu_{\tilde{A}}(x)\mu_{\tilde{B}}(x), \nu_{\tilde{A}}(x) \times \nu_{\tilde{B}}(x), x \in X) \tag{3}$$

$$k\tilde{A} = (1 - (1 - \mu_{\tilde{A}}(x)^k, \nu_{\tilde{A}}(x)^k, x \in X, k > 0. \tag{4}$$

To compare IFS the score function is used [10]. Let $s(\tilde{A}) = \mu_{\tilde{A}} - \nu_{\tilde{A}}, s(\tilde{A}) \in [-1, 1]$ be the score if the IFS $s(\tilde{A})$. If the scores are equal, the accuracy functions are implemented, where $f\left(\tilde{A}\right) = \mu_{\tilde{A}} + \nu_{\tilde{A}}, f\left(\tilde{A}\right) \in [0, 1]$.

The distance [10] between two IFS $\tilde{A} = (\mu_{\tilde{A}}, \nu_{\tilde{A}})$ and $\tilde{B} = (\mu_{\tilde{B}}, \nu_{\tilde{B}})$ is defined as follows:

$$d\left(\tilde{A}, \tilde{B}\right) = \frac{1}{2}\left(\left|\mu_{\tilde{A}} - \mu_{\tilde{B}}\right| + \left|\nu_{\tilde{A}} - \nu_{\tilde{B}}\right|\right) \tag{5}$$

The task of determining the maximum number of evacuees flow the dangerous area to the shelter with people' s storage at intermediate destinations is given as a model (6)-(8). Equation (8) gives the upper bounds of flow for each node at each time period. The model given by Eqs. (6)-(8) gives the ranked set of intermediate nodes with storage to transfer the aggrieved to the safe destination $x_1 \subseteq x_2 \subseteq \ldots x_m$, where x_1 has the highest priority and x_m—the lowest one. This ranked set will be found by multiple attribute intuitionistic fuzzy group decision making algorithm based on TOPSIS. Each node has node capacity $\tilde{x}(\theta)$. Each arc has a time-depended assigned fuzzy arc capacity $\tilde{u}_{ij}(\theta)$ and traversal time $\tau_{ji}(\theta)$.

$$val(\tilde{\gamma}, T) \to \max = \sum_{\theta=0}^{T} \sum_{x_j \in \Gamma(s)} \tilde{\xi}_{sj}(\theta) \geq \sum_{\theta=\tau_{st}}^{T} \sum_{x_k \in \Gamma^{-1}(t)} \tilde{\xi}_{kt}(\theta + \tau_{st}) \tag{6}$$

$$Subjectto: \sum_{\theta=\tau_{ij}}^{\vartheta} \sum_{x_k \in \Gamma^{-1}(i)} \tilde{\xi}_{ki}\left(\theta - \tau_{ij}\right) - \sum_{\theta=0}^{\vartheta} \sum_{x_j \in \Gamma^{1}(i)} \tilde{\xi}_{ij}(\theta) \geq \tilde{0},$$

$$\forall\left(x_i, x_j\right) \neq \{s, t\}, \vartheta \in T, \tag{7}$$

$$0 \leq \widetilde{\xi}_{ij}(\vartheta) \leq \tilde{u}_{ij}(\vartheta), \forall\left(x_i, x_j\right) \in \tilde{A}, \in T. \tag{8}$$

3 Emergency Evacuation in Fuzzy Environment

3.1 MAGDM Algorithm in Intuitionistic Environment for Ranking the Shelters for Evacuation

Let us consider a multi-attribute group decision-making problem in intuitionistic environment for ranking the shelters for evacuation. In group decision-making, several experts are needed to evaluate the alternatives in order to get reasonable decisions. Let $\{C_1, C_2, \ldots C_t\}$ be the set of experts, $\{A_1, A_2, \ldots A_m\}$—the set of alternatives, $\{B_1, B, \ldots B_n\}$ be the set of attributes.

Present the Algorithm for finding the relative order of alternatives in intuitionistic fuzzy conditions as a MAGDM problem [10].

Step 1. Present experts' evaluation in the form of decision matrices $D^k = (\alpha_{ij}^k)_{m\times n}$, where $(\alpha_{ij}^k) = (\mu_{ij}^k, \nu_{ij}^k)$.

Step 2. Compose the positive ideal decision matrix $D^+ = (\alpha_{ij}^+)_{m\times n}$ and the negative ideal decision matrices $D^d = (\alpha_{ij}^d)_{m\times n}$ and $D^u = (\alpha_{ij}^u)_{m\times n}$, where $\alpha_{ij}^+ = (\sum_{k=1}^t \alpha_{ij}^k)/t$, $i = 1, 2, .., n$, $k = 1, 2, ..t$, $\alpha_{ij}^d = \min\limits_{1\le k\le t}\{\alpha_{ij}^{(k)} | \alpha_{ij}^{(k)} \le \alpha_{ij}^+\}$, $\alpha_{ij}^u = \max\limits_{1\le k\le t}\{\alpha_{ij}^{(k)} | \alpha_{ij}^{(k)} \ge \alpha_{ij}^+\}$.

Step 3. Compose the collective decision matrix $D = (\alpha_{ij})_{m\times n}$ according to the values of closeness coefficients applying intuitionistic fuzzy weighted averaging operator. To do it, firstly, find the distances between the expert's evaluation α_{ij}^k and positive ideal α_{ij}^+ along with the negative ideal matrices α_{ij}^d and α_{ij}^u by Eq. (5).

$$d_{ij}^+ = \frac{1}{2}\left(\left|\mu_{ij}^k - \mu_{ij}^+\right| + \left|\nu_{ij}^k - \nu_{ij}^+\right|\right) d_{ij}^d = \frac{1}{2}\left(\left|\mu_{ij}^k - \mu_{ij}^d\right| + \left|\nu_{ij}^k - \nu_{ij}^d\right|\right)$$

$$d_{ij}^u = \frac{1}{2}\left(\left|\mu_{ij}^k - \mu_{ij}^u\right| + \left|\nu_{ij}^k - \nu_{ij}^u\right|\right)$$

Define the closeness coefficients of α_{ij}^k: $c_{ij}^k = \frac{d_{ij}^u + d_{ij}^d}{d_{ij}^u + d_{ij}^d + d_{ij}^+}$.

The collective decision matrix $D = (\alpha_{ij})_{m\times n}$ consists of elements $\alpha_{ij} = w_{ij}^{(1)}\alpha_{ij}^{(1)} + \cdots + w_{ij}^{(t)}\alpha_{ij}^{(t)}$, where an expert's weight C_k regarding the attribute B_j for the alternative A_i: $w_{ij}^{(k)} = \frac{c_{ij}^{(k)}}{\sum_{k=1}^t c_{ij}^{(k)}}$, $w_{ij}^{(k)} \ge 0$, $\sum_{k=1}^t \sum_{k=1}^t c_{ij}^{(k)}{}_{ij}^{(k)} = 1$

Step 4. Find the attribute weight vector w_j based on the principle: the closer to fuzzy positive ideal value and farther from the intuitionistic fuzzy negative ideal, the large the weight is $w_j = \frac{c_j}{\sum_{j=1}^n c_j} = \frac{\sum_{i=1}^m c_{ij}}{\sum_{j=1}^n \sum_{i=1}^m c_{ij}}$

where c_{ij} defines the closeness coefficient of experts' collective assessment α_{ij} regarding its distances to the positive ideal value $\alpha_j^+ = (1, 0)$ and the negative ideal value $\alpha_j^- = (0, 1)$, $c_{ij} = \frac{d(\alpha_{ij}, \alpha_j^-)}{d\left(\alpha_{ij}, \alpha_j^+\right) + d\left(\alpha_{ij}, \alpha_j^-\right)}$.

Step 5. Determine the weighted decision matrix $D^{'} = (\alpha_{ij}^{'})_{m\times n}$, where $\alpha_{ij}^{'} = w_j\alpha_{ij}$, $W = (w_1, w_2, \ldots, w_n)$ be the weight vector.

Step 6. Calculate the distance d^+ and d^- of each alternative's collective evaluation value to intuitionistic fuzzy positive ideal evaluation $A^+ = (\alpha_1^+, \alpha_2^+, \ldots, \alpha_n^+)$ and intuitionistic fuzzy negative ideal evaluation value $A^- = (\alpha_1^-, \alpha_2^-, \ldots, \alpha_n^-)$.

$$d_i^+ = \sum_{j=1}^{n} d\left(\alpha_{ij}^{'}, \alpha_j^+\right), i = 1, \ldots, m, d_i^- = \sum_{j=1}^{n} d\left(\alpha_{ij}^{'}, \alpha_j^-\right), i = 1, \ldots, m$$

Step 7. Calculate each alternative's closeness coefficient $c_i = \frac{d_i^-}{d_i^- + d_i^+}$.

Step 8. Determine the rank of alternatives based on the alternatives' closeness coefficients [9].

3.2 Emergency Evacuation Based on the Maximum Dynamic Flow Finding

Present the Algorithm for emergency evacuation based on the maximum dynamic flow finding [12, 13].

Step 1. Transform the initial dynamic network $\tilde{G}$ into a time-spaced network $\tilde{G}^*$ by copying every node and arc at the specific time period along with converting the intermediate capacitated node x_i into the nodes x_i^+ and x_i^- with the arc capacity $\tilde{u}\left(x_i^+, x_j^-, \theta, \theta\right) = q(x_i)$.

Step 2. Pass the flow along the augmenting paths in the residual network $\tilde{G}^{*r}$.

Step 2.1. The If $\overset{\sim}{\xi}^{*r}\left(x_i^*, x_j^*, \theta, \vartheta\right) < \tilde{u}^*\left(x_i^*, x_j^*, \theta, \vartheta\right)$ in $\tilde{G}_e^{*r}$, then $\tilde{u}^{*r}\left(x_i^*, x_j^*, \theta, \vartheta\right) = \tilde{u}^*\left(x_i^*, x_j^*, \theta, \vartheta\right) - \overset{\sim}{\xi}^*\left(x_i^*, x_j^*, \theta, \vartheta\right)$. If $\overset{\sim}{\xi}^*\left(x_i^*, x_j^*, \theta, \vartheta\right) > \tilde{0}$, then $\tilde{u}^{*r}\left(x_j^*, x_i^*, \vartheta, \theta\right) = \overset{\sim}{\xi}^*\left(x_i^*, x_j^*, \theta, \vartheta\right)$.

2.2. If the path exists, move to the step 2.3

2.3 If there is no path to the sink, the maximum flow without intermediate storage to the destination

n t is found, turn to step 2.4.

Step 3. Pass the flow $\overset{\sim}{\sigma}^* = min[\tilde{u}^{*r}\left(x_i^*, x_j^*, \theta, \vartheta\right)]$, turn to the step 2.5.

Step 4. Find the augmenting paths from the intermediate nodes that allow storage to the sink T in priority order of nodes based on fuzzy intuitionistic TOPSIS method. The sink t has the highest priority; then there is the intermediate node x_i with the highest among others $q(x_i) > \tilde{0}$.

4.1 If a path exists, move back to the step 2.3

4.2 If there is no path, the maximum flow to the sink t is found, move to step 2.6

Step 5. Transform the evacuation flows: 1) for arcs joining $(x_j^{*\mu r}, \vartheta)$ and $(x_i^{*\mu r}, \theta)$, decrease the flow value $\overset{\sim}{\xi}^{\mu}\left(x_i^*, x_j^*, \theta, \vartheta\right)$ by the value $\overset{\sim}{\sigma}^{*\mu}$. The total flow is $\overset{\sim}{\xi}^{\mu}\left(x_i^*, x_j^*\theta, \vartheta\right) - \overset{\sim}{\sigma}^{*\mu}$. Move back to the step 2.2. 2) for arcs joining $(x_i^{*\mu r}, \theta)$ and $(x_j^{*\mu r}, \vartheta)$, increase the flow value $\overset{\sim}{\xi}^{\mu}\left(x_i^*, x_j^*, \theta, \vartheta\right)$ by the value $\overset{\sim}{\sigma}^{*\mu}$. Total flow value is $\overset{\sim}{\xi}^{\mu}\left(x_i^*, x_j^*, \theta, \vartheta\right) + \overset{\sim}{\sigma}^{*\mu}$ and turn to the step 2.2

Step 6. Remove dummy sinks and shelters. Turn to the original network.

4 Case Study

In this section, we provide a case-study to simulate the emergency decision-making [14] in order to evacuate the maximum number of aggrieved from the dangerous area s and transport them to the safe shelter t. The evacuation is performed from the stadium Zenit in Saint Petersburg, Russia to the safe area. The safe pattern of evacuation considers storage at nodes so that to transport the maximum possible number of evacuees. Figure 1 shows the initial emergency network with the dangerous area s and the shelter t. Figure 2 represents the real network in the form of a fuzzy graph within the time horizon $\mathrm{T} = 4$.



Fig. 1. Real evacuation network.

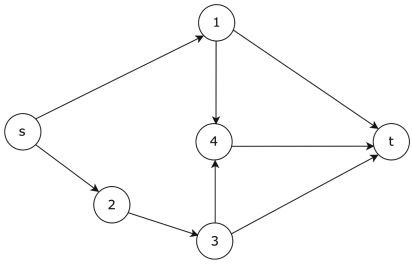


Fig. 2. Graph image of the real network.

Transit fuzzy arc capacities and traversal time parameters are given in Table 1.

Owing to the complexity of a decision-making task, incomplete information about the emergency, four decision makers C_i $(i = 1,...,4)$ are asked to assess the priority

order of intermediate nodes x_1, x_2, x_3, x_4 for pushing the flow to the sink. Inherent uncertainty of decision-making problems makes experts to hesitate and be irresolute about the choice of membership function. Therefore, intuitionistic fuzzy assessments towards four attributes: the level of reachability (B_1), capacity of destination nodes (B_2), reliability (security) (B_3), and total expenses (B_4), are used to rank intermediate nodes.

The attribute weight vector W is unknown and will be determined by the principle that the attribute whose evaluation value is close to the positive ideal evaluation and far from negative ideal evaluation values has a large weight.

To evacuate the maximum people from the dangerous area s to the safe destination t, we find the maximum s-t flow. Firstly, convert the dynamic network into the static (Fig. 3) by expanding the nodes and arcs of the network in time dimension.

Table 1. Transit fuzzy arc capacities and traversal time parameters.

T	Arc capacities, traversal times											
	(s, x_1^+)	(x_1^+, x_1^-)	(x_1^-, x_3^+)	(x_3^+, x_3^-)	(x_3^-, x_4^+)	(x_4^-, x_4^+)	(s, x_2^+)	(x_2^-, x_2^+)	(x_2^-, x_4^+)	(x_2^-, t)	(x_3^-, t)	(x_4^-, t)
0	$\tilde{90}, 1$	$1\tilde{00}, 0$	$\tilde{97}, 1$	$1\tilde{00}, 0$	$\tilde{72}, 1$	$\tilde{87}, 0$	$1\tilde{30}, 1$	$1\tilde{40}, 0$	$\tilde{72}, 1$	$\tilde{80}, 0$	$\tilde{70}, 1$	$\tilde{65}, 1$
1	$\tilde{95}, 1$	$1\tilde{00}, 0$	$\tilde{95}, 1$	$1\tilde{00}, 0$	$\tilde{70}, 1$	$\tilde{87}, 0$	$1\tilde{35}, 1$	$1\tilde{40}, 0$	$\tilde{70}, 1$	$\tilde{80}, 1$	$\tilde{70}, 1$	$\tilde{65}, 1$
2	$\tilde{80}, 1$	$1\tilde{00}, 0$	$1\tilde{02}, 1$	$1\tilde{00}, 0$	$\tilde{70}, 1$	$\tilde{87}, 0$	$1\tilde{00}, 1$	$1\tilde{40}, 0$	$\tilde{56}, 1$	$1\tilde{05}, 1$	$\tilde{70}, 1$	$\tilde{67}, 1$
3	$\tilde{80}, 2$	$1\tilde{00}, 0$	$\tilde{90}, 1$	$1\tilde{00}, 0$	$1\tilde{10}, 1$	$\tilde{87}, 0$	$1\tilde{00}, 1$	$1\tilde{40}, 0$	$\tilde{55}, 1$	$\tilde{55}, 1$	$1\tilde{10}, 1$	$\tilde{68}, 1$
4	$\tilde{75}, 1$	$1\tilde{00}, 0$	$\tilde{90}, 1$	$1\tilde{00}, 0$	$1\tilde{10}, 2$	$\tilde{87}, 0$	$\tilde{90}, 1$	$1\tilde{40}, 0$	$\tilde{60}, 2$	$\tilde{50}, 1$	$1\tilde{00}, 1$	$\tilde{68}, 1$

Secondly, find the augmenting paths to transport the flows in the time-expanded network. A series of paths with the corresponding flow distribution is found and the maximum s-t flow without intermediate storage is shown in Fig. 4 Therefore, the total maximum s-t flow in the network without intermediate storage is $5\tilde{05}$ flow units.

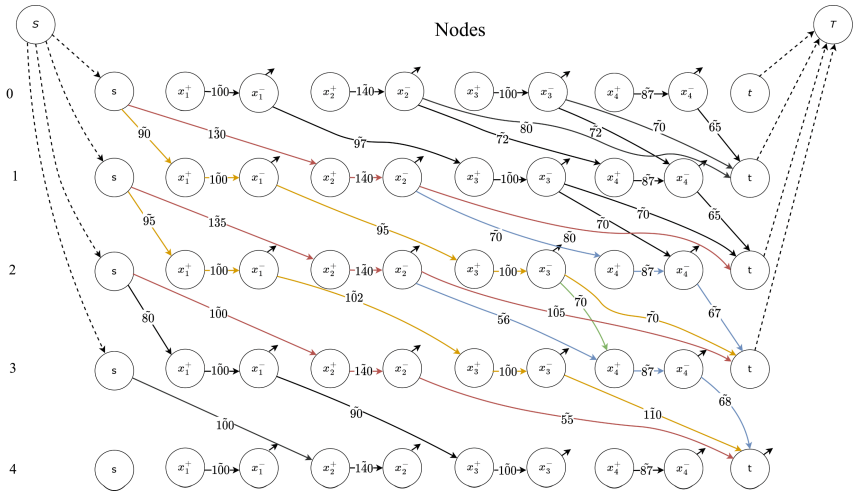


Fig. 3. The time-expanded network.

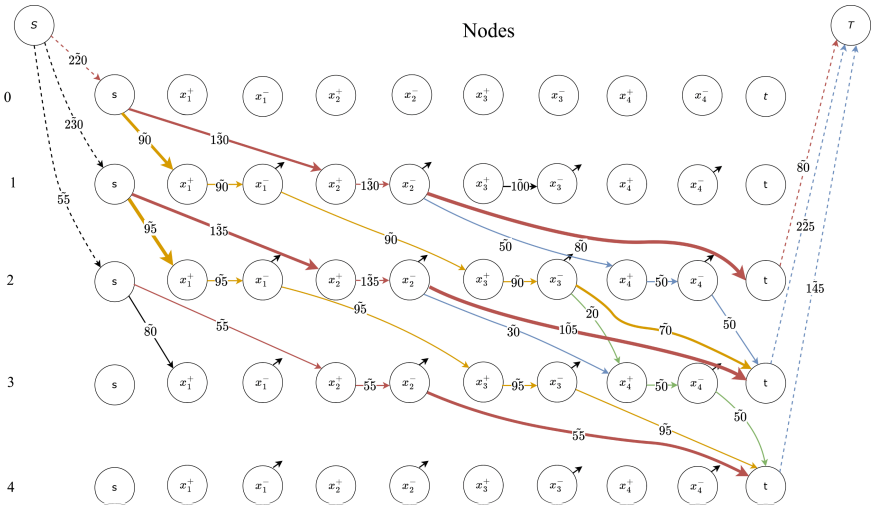


Fig. 4. Network with maximum flow without intermediate storage.

To find extra flows with intermediate storage we should define the order of intermediate nodes for evacuation the aggrieved in Fig. 4. Four experts provide the assessments of alternatives concerning attributes in Table 2.

Following the steps of the intuitionistic TOPSIS, calculate intuitionistic fuzzy negative ideal (Tables 3–4) and positive ideal (Table 5) decision matrices. Intuitionistic fuzzy collective and weighted decision matrices are performed in Tables 6–7.

According to the step 6 the distances of alternatives' evaluation values to the values A^+ and A^- are $d_1^+ = 2.975$, $d_2^+ = 2.993$, $d_3^+ = 3.057$, $d_4^+ = 3.263$, $d_1^- = 1.025$, $d_2^- = 1.007$, $d_3^- = 0.943$, $d_4^- = 0.737$. The relative closeness coefficients: $c_1 = 0.256$, $c_2 = 0.252$, $c_3 = 0.236$, $c_4 = 0.184$. The alternatives thus are ranked as: $x_1 \succ x_2 \succ x_3 \succ x_4$.

Then, push the additional flow values which are stored at nodes to evacuate the maximum number of aggrieved. Finally, we have the paths:1) $S \to s^2 \to x_1^{3+} \to x_1^{3-} \to T$ with $\tilde{80}$ units. 2) $S \to s^2 \to x_2^{3+} \to x_2^{3-} \to T$ with $\tilde{45}$ units. 3) $S \to s^3 \to x_2^{4+} \to x_2^{4-} \to T$ with $\tilde{100}$ units.

The maximum flow with intermediate storage is $\tilde{730}$ flow units, which is shown in Fig. 5.

5 Conclusion and Future Study

The paper illustrates the approach to evacuation of the maximum amount of aggrieved from the dangerous area to the safe destination so that the intermediate nodes can store the evacuees. This method enables maximizing the total amount of flow by pushing the maximum amount of flow from the source. The order of nodes for transporting aggrieved to the sink is found by MAGDM algorithm in intuitionistic environment based on TOPSIS. Group decision-making is required since one expert cannot have enough

Table 2. Intuitionistic fuzzy decision matrix of the DMs

	B_1	B_2	B_3	B_4
C_1				
x_1	(0.5, 0.4)	(0.7, 0.3)	(0.4, 0.4)	(0.8, 0.1)
x_2	(0.7, 0.2)	(0.3, 0.5)	(0.6, 0.3)	(0.7, 0.1)
x_3	(0.4, 0.3)	(0.6, 0.3)	(0.8, 0.1)	(0.5, 0.2)
x_4	(0.3, 0.6)	(0.2, 0.7)	(0.7, 0.1)	(0.4, 0.5)
C_2				
x_1	(0.6, 0.2)	(0.5, 0.4)	(0.5, 0.3)	(0.6, 0.3)
x_2	(0.5, 0.3)	(0.2, 0.6)	(0.4, 0.4)	(0.8, 0.1)
x_3	(0.5, 0.3)	(0.4, 0.3)	(0.6, 0.2)	(0.7, 0.1)
x_4	(0.2, 0.6)	(0.4, 0.5)	(0.5, 0.3)	(0.7, 0.2)
C_3				
x_1	(0.3, 0.5)	(0.5, 0.2)	(0.6, 0.3)	(0.9, 0.1)
x_2	(0.5, 0.3)	(0.6, 0.2)	(0.5, 0.3)	(0.8, 0.1)
x_3	(0.4, 0.5)	(0.7, 0.1)	(0.6, 0.3)	(0.4, 0.5)
x_4	(0.2, 0.6)	(0.3, 0.5)	(0.4, 0.2)	(0.5, 0.4)
C_4				
x_1	(0.2, 0.6)	(0.3, 0.6)	(0.7, 0.1)	(0.8, 0.1)
x_2	(0.5, 0.4)	(0.7, 0.2)	(0.4, 0.3)	(0.6, 0.1)
x_3	(0.3, 0.6)	(0.5, 0.3)	(0.3, 0.4)	(0.6, 0.2)
x_4	(0.4, 0.4)	(0.4, 0.5)	(0.2, 0.5)	(0.7, 0.1)

Table 3. Intuitionistic fuzzy negative ideal decision matrix D^u

	B_1	B_2	B_3	B_4
x_1	(0.6,0.2)	(0.7,0.3)	(0.7,0.1)	(0.9,0.1)
x_2	(0.7,0.2)	(0.7,0.2)	(0.6,0.3)	(0.8,0.1)
x_3	(0.5,0.3)	(0.7,0.1)	(0.8,0.1)	(0.7,0.1)
x_4	(0.4,0.4)	(0.4,0.5)	(0.7,0.1)	(0.7,0.1)

professional knowledge of each aspect of evacuation to make reasonable decisions. Experts' weights on various attributes in the method are unknown and determined by the principle that the attribute whose evaluation value is close to the positive ideal evaluation and far from negative ideal evaluation values has a large weight. The proposed method handles intuitionistic fuzzy values of experts' assessments because of inherent hesitation in exact membership degrees. This technique enables experts to consider the degree of

Table 4. Intuitionistic fuzzy negative ideal decision matrix D^d.

	B_1	B_2	B_3	B_4
x_1	(0.2,0.6)	(0.3,0.6)	(0.4,0.4)	(0.6,0.3)
x_2	(0.5,0.4)	(0.2,0.6)	(0.4,0.4)	(0.6,0.1)
x_3	(0.3,0.6)	(0.4,0.3)	(0.3,0.4)	(0.4,0.5)
x_4	(0.2,0.6)	(0.2,0.7)	(0.2,0.5)	(0.4,0.5)

Table 5. Intuitionistic fuzzy positive ideal decision matrix D^+.

	B_1	B_2	B_3	B_4
x_1	(0.421, 0.394)	(0.521, 0.322)	(0.564, 0.245)	(0.799, 0.131)
x_2	(0.560, 0.291)	(0.491, 0.331)	(0.482, 0.322)	(0.737, 0.100)
x_3	(0.404, 0.405)	(0.564, 0.228)	(0.613, 0.221)	(0.564, 0.211)
x_4	(0.280, 0.542)	(0.330, 0.544)	(0.482, 0.234)	(0.595, 0.251)

Table 6. Intuitionistic fuzzy collective decision matrix D

	B_1	B_2	B_3	B_4
x_1	(0.426, 0.395)	(0.531, 0.305)	(0.565, 0.249)	(0.809, 0.121)
x_2	(0.550, 0.295)	(0.503, 0.319)	(0.482, 0.320)	(0.742, 0.100)
x_3	(0.407, 0.400)	(0.563, 0.235)	(0.619, 0.219)	(0.569, 0.203)
x_4	(0.274, 0.552)	(0.337, 0.534)	(0.486, 0.230)	(0.604, 0.243)

Table 7. Intuitionistic fuzzy weighted decision matrix $D^{'}$

B_1	B_2	B_3	B_4
(0.108, 0.825)	(0.161, 0.759)	(0.187, 0.708)	(0.404, 0.518)
(0.152, 0.777)	(0.150, 0.767)	(0.151, 0.753)	(0.345, 0.487)
(0.102, 0.827)	(0.175, 0.714)	(0.213, 0.686)	(0.231, 0.608)
(0.064, 0.884)	(0.091, 0.864)	(0.152, 0.694)	(0.251, 0.643)

membership, non-membership and hesitation. A case study is conducted to simulate the evacuation of the maximum number of evacuees with storage at intermediate. MAGDM algorithm in intuitionistic environment based on TOPSIS is used to rank the shelters for

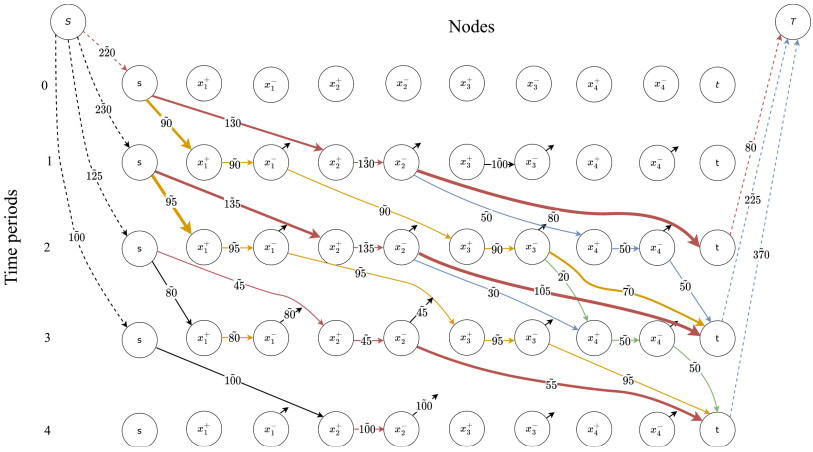


Fig. 5. Network with maximum flow with intermediate storage.

evacuation. Abstract flow models in fuzzy environment will be proposed to evacuate the maximum amount of people as a part of the future research.

Acknowledgments. The research was funded by the Russian Science Foundation project No. 22–71-10121, https://rscf.ru/en/project/22-71-10121/ implemented by the Southern Federal University.

References

1. Kittirattanapaiboon, S.: Emergency evacuation route planning considering human behavior during short—and no-notice emergency situations. Electron. Theses Diss., 3906 (2009)
2. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
3. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**(1), 87–96 (1986)
4. Xu, Z.: Hesitant fuzzy sets theory, studies in fuzziness and soft computing. 314 (2014)
5. Ren, F., Kong, M., Zheng, P.: A new hesitant fuzzy linguistic topsis method for group multi-criteria linguistic decision making. Symmetry **9**, 289 (2017)
6. Su, W.H., Zeng, S.Z., Ye, X.J.: Uncertain group decision-making with induced aggregation operators and Euclidean distance. Technol. Econ. Dev. Econ. **19**(3), 431–447 (2013)
7. Wang, W.Z., Liu, X.W., Qin, Y.: Multi-attribute group decision making models under interval type-2 fuzzy environment. Knowl.-Based Syst. **30**, 121–128 (2012)
8. Pang, J.F., Liang, J.Y.: Evaluation of the results of multi-attribute group decision-making with linguistic information. Omega **40**(3), 294–301 (2012)
9. Hajiagha, S.H.R., Hashemi, S.S., Zavadskas, E.K.: A complex proportional assessment method for group decision making in an interval-valued intuitionistic fuzzy environment. Technol. Econ. Dev. Econ. **19**(1), 22–37 (2013)
10. Yang, W., Chen, Z., Zhang, F.: New group decision making method in intuitionistic fuzzy setting based on TOPSIS. Technol. Econ. Dev. Econ. **23**(3), 441–461 (2017)
11. Park, J.H., Park, I.Y., Kwun, Y.C., Tan, X.G.: Extension of the TOPSIS method for decision making problems under interval-valued intuitionistic fuzzy environment. Appl. Math. Model. **35**(5), 2544–2556 (2011)

12. Gerasimenko, E., Kureichik, V.: Minimum cost lexicographic evacuation flow finding in intuitionistic fuzzy networks. J. Intell. Fuzzy Syst. **42**(1), 251–263 (2022)
13. Gerasimenko, E., Kureichik, V.: Hesitant fuzzy emergency decision-making for the maximum flow finding with intermediate storage at nodes lecture notes in networks and systems **307**, 705–712 (2022)
14. Tian, X., Ma, J., Li, L., Xu, Z., Tang, M.: Development of prospect theory in decision making with different types of fuzzy sets: A state-of-the-art literature review. Inf. Sci. **615**, 504–528 (2022)



Task-Cloud Resource Mapping Heuristic Based on EET Value for Scheduling Tasks in Cloud Environment

Pazhanisamy Vanitha(✉), Gobichettipalayam Krishnaswamy Kamalam, and V. P. Gayathri

Kongu Engineering College, Perundurai, Tamil Nadu, India
{vanitha.it,gayathri.it}@kongu.edu

Abstract. The most popular and highly scalable computing technology, cloud computing bases its fees on the amount of resources used. However, due to the increase in user request volume, task scheduling and resource sharing are fetching key needs for active load sharing of a capability between cloud resources. This will improve the overall performance of cloud systems. The above aspects contribute to the development of standard, heuristic, and meta-heuristic algorithms, as well as other task scheduling techniques. The job planning problem is typically solved using heuristic job planning algorithms like Min-Min, MET, Max-Min and MCT. This research proposes a novel hybrid cloud computing method based on Min-Min and Max-Min experimental procedures. When using the Cloudsim simulator, this algorithm has been evaluated with a number of optimization settings, including makespan, average source usage, load sharing, typical waiting time, and in-parallel execution of short-duration activities and long-duration tasks. The proposed algorithm TCRM_EET experimental results computed and performance analyzed are carried out based on the analytical benchmark.The findings demonstrate that the proposed method outperforms Min-Min and Max-Min for such values.

Keywords: Load sharing · Heuristic algorithms · Makespan · Resource sharing · Min-Min · Max-Min · Task scheduling

1 Introduction

Along with development and expansion demand for information technology, cloud computing is becoming a viable option for both personal and business needs. It provides customers with a vast array of virtualized cloud resources that are available on demand, via remote access, and for pay-per-use over the internet anywhere in the world [1].

Additionally, when compared to other computing technologies, cloud computing has a number of advantages and traits. It is a vast network access that is elastic, virtualized, affordable, resource-pooling, independent of device or location, always accessible from anywhere via the internet or private channels. It reduces expensive costs associated with data centre construction, upkeep, disaster recovery, energy use, and technical staff.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 680–689, 2023.
https://doi.org/10.1007/978-3-031-27409-1_62

Therefore, maximising their use is key to achieving higher throughput and making cloud computing viable for large-scale methods and groups [2–4].

Cloud computing adoption comes in a variety of formats. Public; in this kind, consumers can access cloud resources in a public way using web browsers and an Internet connection. Private clouds, like intranet access in a network, are created for a particular group or organisation and only allow that group's members to access them. In essence, a hybrid cloud mixes with the merged clouds being a mixture of public and private clouds. It is shared by two or more businesses with similar cloud computing necessities [2].

Cost, security, accessibility, user task completion times, accessibility, adaptability, and performance tracking, the need for a continual and quick regular access to the internet and reliable Task scheduling, scaling, interoperability, QoS management, service concepts, VM allocation and migration, and transportability and effective load sharing are a few of the problems and challenges associated with cloud computing. Job scheduling, resource sharing, and load sharing are generally regarded as the top issues in cloud computing and broadcast network since they significantly improve the performance of the system as a whole. [2, 3, 5–7].

A brand-new hybrid scheduling algorithm has been put forth in this research. Its name is Hybrid Max-Min and Min-Min Algorithm (HAMM). As implied by its name, it relies on Max-Min and Min-Min, two classic experiential procedures, to take use of their advantages and get around their drawbacks. When compared to Min-Min and Max-Min, it typically performs better in the following areas: makespan, average consumption, typical waiting time, effective parallel performance between short jobs and long activities, load balancing, and average execution time.

The essential task scheduling techniques are described in Sect. 2 and Sect. 3 of the remaining portion of this study. The proposed algorithm, together with a flowchart and pseudocode, are presented in Sect. 4. Section 5 defines simulation and analysis, including the Cloudsim simulator tool. Results and discussions are performed in Section 6. Section 7 concludes by describing the results and the next steps.

1.1 Scheduling of Tasks

Scheduling of tasks, is mainly focus on properties and nature of the algorithm to be formulated with mapping of resources like users' tasks are assigned to cloud resources which are currently available. This could be done with suitable time period and also utilizing the resource in a best way.

For the scheduler, overall performance can be measured with certain parameters which include makespan reduction, best resource usage, effective workload distribution across resources and other factors all affect the scheduler's overall performance [8]. The Meta-Task has a crucial function in task scheduling as well. It is a group of tasks that the system, which in this case is the cloud provider, got from various users. Meta tasks may have comparable properties or share some characteristics with one another [9].

Normally, this process could be divided as three categories namely resource discovery, selecting the resource and task submission. Broker communicates and provides meaningful data to the cloud resources. Submission of the task; during this stage, the assignment is given to the chosen resource to be carried out and scheduled [11].

Traditional, heuristic, meta-heuristic, and other classifications of work scheduling and load balancing algorithms exist [10]. Task scheduling is regarded as an NP-hard problem.—To identify the best answers to problems, complete and heuristic algorithms work best [6, 11, 12].

Every user wants their task to be done as quickly as possible, so an effective scheduler will allocate users evenly and concurrently among all tasks, avoiding starvation for any tasks or users [13]. The benefits and drawbacks of various heuristic algorithms are covered in the section that follows.

2 Heuristic Algorithms

Algorithms applied with Heuristics are subset of different mode that are ideal for cloud job allocation. Approach relies upon schedule's finish time. The following points address MCT, OLB, MET, Min-Min, and Max-Min task mapping heuristic as examples of heuristic algorithms.

A. Mode scheduling (Immediate):

It is sometimes dubbed internet mode. Jobs are executed straight from the front of the queue in this mode.

Opportunistic Load Balancing:

The task allocation and execution is done randomly and OLB will allocate the unexecuted task to the resources currently available. The task completion or execution time are not taken care by OLB [14, 15].

B. Mode scheduling (Batch):

Jobs allocated as batches within the specified time.

Max-min and Min-Min Heuristic Algorithm:

In Max-Min algorithm, a task is selected with higher completion time rather than less completion time preferred in the min-min algorithm. For the unexecuted tasks, remaining time to be calculated and completion time gets updated accordingly. The procedure continues until the task gets completed. For the concurrent task execution and increase makespan, max-min is better and precedes with all the heuristic algorithms. The issue faced and need to be addressed in max-min algorithm is starvation [16–18].

In min-min algorithm, small tasks with good execution time given preference to avoid starvation problem in max-min. There are many improved algorithms came into serve the purpose like to calculate the average task execution time.

3 TCRM_EET Algorithm

When there are numerous more short tasks than large ones, the Min-Min algorithm proves to be worse. In vice-versa, the Max-Min approach proves to be worse. For instance, undertakes numerous small tasks concurrently with few long task. In this instance, the long task's execution time is probably what determines the system's makespan.

For instance, undertakes numerous long tasks concurrently with few small tasks. In this instance, the long task's execution time is probably what determines the system's makespan.

In Fig. 1, our proposed task scheduling TCRM_EET is presented and mapped using $AverageEET_i$. Tasks count whose $AverageEET_i$. Larger than makespan value is $\geq (\boldsymbol{\mu}/2)$, then tasks are listed in descending order, otherwise tasks listed in ascending order.

Based upon count value, tasks were grouped in tasks set TS either in ascending order or descending order. To select a CRj for scheduling Ti, figure out the minimum completion time of each tasks Ti on all CRj as follows.

$$CT_i = \min\left(EET_{ij} + RT_j\right), 0 \leq j \geq \mu(1)$$

where, EET_{ij} *represents the excepted execution time of tasks* T_i *on* CR_j

RT_j *represents the ready time of* CR_j *after completing execution of Previously assigned tasks*

The Pseudocode of proposed TCRM_EET algorithm:

```
// Stage 1: Grouping the tasks in MT in either ascending or descending order in
Tasks Set TS.
    For all tasks (Ti) in the Meta Task Set (MT)
        Compute makespan using Min-Min algorithm
    For all Tasks (Ti) 0 ≤ i ≥ λ do
        Compute TETi = EETij, (for each task in all cloud resources).
        Compute Average EETi = TETi / μ
    End For
    For all Tasks (Ti) 0 ≤ i ≥ λ do
        If (makespan < Average EETi )
            count=count+1
    End For
    If (count>=(λ/2))
        Group all the tasks Ti in descending order of Average EETi in Tasks Set TS
    Else
        Group all the tasks Ti in ascending order of Average EETi in Tasks Set TS
// Stage 2: Assigning/Scheduling Task (Ti) with Minimum Completion Time (MCTi) to
Cloud Resource (CRj) that gives minimum MCT.
    Loop until all tasks in (TS) have been mapped.
        For every task Ti in TS,
            Find the Cloud Resource CRj with Minimum Completion Time
    (MCTij)
            Assign Ti to the cloud resource CRj that gives the MCT.
            Delete the task Ti from the Tasks Set TS.
        End For
        Update the cloud resource Ready Time (RTj)
    End Loop
    compute makespan = max(MCTi) 0 ≤ i ≥ λ
```

Fig. 1. xxx

3.1 An Illustration

Consider a scenario with eight tasks and eight cloud resources as a basic outline. In Table 1, the tasks Ti, CRj, and EETij are shown.

Table 1. Consistent high task high machine heterogeneity

Tasks (Ti)/Cloud Resource (CRj)	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8
T1	25,137.5	52,468.0	150,206.8	289,992.5	392,348.2	399,562.1	441,485.5	518,283.1
T2	30,802.6	42,744.5	49,578.3	50,575.6	58,268.1	58,987.9	85,213.2	87,893.0
T3	242,727.1	661,498.5	796,048.1	817,745.8	915,235.9	925,875.6	978,057.6	1,017,448.1
T4	68,050.1	303,515.9	324,093.1	643,133.7	841,877.3	856,312.9	861,314.8	978,066.3
T5	6,480.2	42,396.7	98,105.4	166,346.8	240,319.5	782,658.5	871,532.6	1,203,339.8
T6	175,953.8	210,341.9	261,825.0	306,034.2	393,292.2	412,085.4	483,691.9	515,645.9
T7	116,821.4	240,577.6	241,127.9	406,791.4	1,108,758.0	1,246,430.8	1,393,067.0	1,587,743.1
T8	36,760.6	111,631.5	150,926.0	221,390.0	259,491.1	383,709.7	442,605.7	520,276.8

For the given scenario in Table 1, the makespan value obtained using Min-min algorithm is 379155.5

Step 1: Average execution time of each task Ti on all cloud resource CRj computed as shown in Table 2.

Table 2. Ti and average EETi

Tasks (Ti)	Average EETi
T1	283,685.5
T2	58,007.9
T3	794,329.6
T4	609,545.5
T5	426,397.4
T6	344,858.8
T7	792,664.7
T8	265,848.9

Step 2: Tasks Ti are ordered based on the value of makespan found in Min-min algorithm 379155.5. The count of tasks whose average EETi is greater than makespan value is >= (number of tasks/2), then the tasks are arranged in descending order, otherwise the tasks are arranged in ascending order. For the scenario given, tasks listed for scheduling in Task Set TS and are shown in Table 3.

Table 3. Task Set TS—scheduling order

T3	T7	T4	T5	T6	T1	T8	T2

Step 3: Now the tasks Ti in the tasks set TS is taken one by one and allocated to the cloud resource CRj whose completion time is minimum and Table 4. Presents the tasks, cloud resources mapping for Min-Min algorithm and the proposed TCRM_EET algorithm.

Table 4. Ti and CRj mapping

Min-Min			TCRM_EET algorithm		
Tasks T_i	Cloud resource CR_j allocated	Expected completion time ECT_{ij}	Tasks T_i	Cloud resource CR_j allocated	Expected completion time ECT_{ij}
T5	R1	6480.2	T3	R1	242727.1
T1	R1	31,617.7	T7	R2	240577.6
T2	R2	42,744.5	T4	R1	310777.2
T8	R1	68378.3	T5	R3	98105.4
T4	R1	136,428.4	T6	R4	306034.2
T7	R3	241,127.9	T1	R3	248312.2
T6	R2	253086.4	T8	R5	259491.1
T3	R1	379,155.5	T2	R6	58987.9
Makespan−379,155.5			*Makespan−310777.2*		

As can be evident, the proposed algorithm TCRM_EET ought to decide the tasks to be scheduled in the descending order based on each task Ti average EETi and from Table 4, it clearly depicts proposed algorithm TCRM_EET achieves minimum makespan and better CR utilization than Min-min heuristic.

4 Results and Discussion

Simulation is carried out for 12 different possible characteristice of ETC matrix, task and resource heterogeneity, and consistency. ETC matrix value is generated as an avearge of 100 ETC matrices for all 12 possible characteristic combination. The size of the genertaed matix is $\tau^*\mu$, where $\tau = 512$ tasks and $\mu = 16$ cloud resources.

The proposed algorithm TCRM_EET experimental results computed and performance analyzed are carried out based on the analytical benchmark.

12 ETC matrix entities are fundamentally defined as u-x-yyzz.k, u—defines division allocation uniformly for creating ETC twelve matrix instances. x-specifies consistency, value x is c-consistent, i-inconsistent, pc-partially consistent, yy-represents task heterogeneity, zz-represents resource heterogeneity. Makespan of the heuristic (TCRM_EET)

compared with the existing heuristics for twelve ETC matrix instances is shown in Figs. 2, 3, 4, 5, and 6.

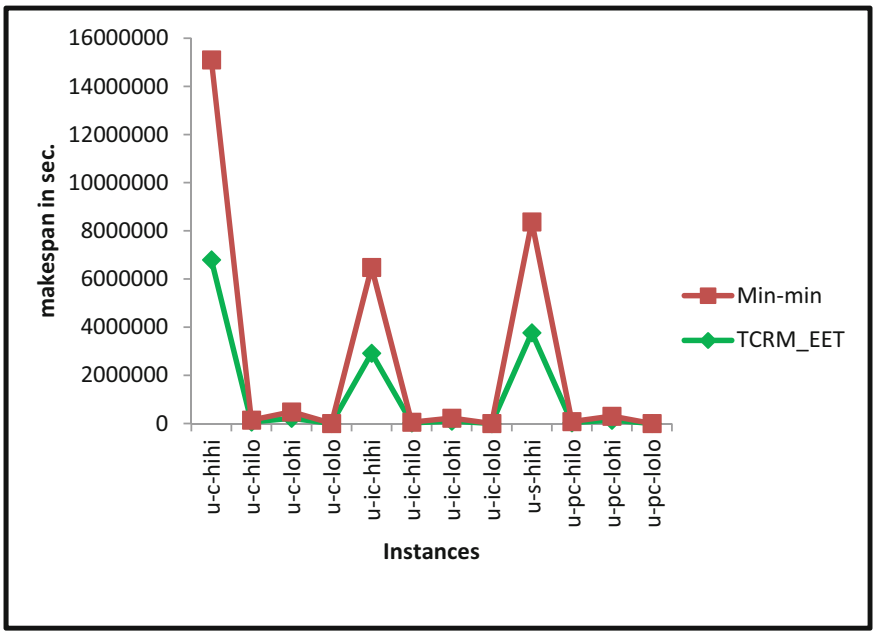


Fig. 2. Makespan values

As seen from the graphical representation, comparison results depict that proposed algorithm (TCRM_EET) performs better than Min-Min and it has a shorter makespan.

5 Conclusion and Future Work

Mapping of cloud users tasks to the available heterogeneous cloud resources is the primary concern in distributed cloud environment to bring out an efficient performance in cloud system. This paper delvers an efficient heuristic technique that combines the advantage of both Min-Min and Max-Min heuristic. Experimental evaluation of proposed heuristic TCRM_EET brings out efficient performance in mapping the tasks to the appropriate cloud resource. Heuristic TCRM_EET achieves better utilization rate of cloud resources, and least makespan. The proposed approach follow static scheduling of tasks in cloud environment. In cloud environment, service providers follow a pay per use go strategy, so in future an efficient scheduling strategy to be considered satisfying the cost efficiency in allocating the tasks to the cloud resources there by providing the customer a service with minimum makespan and reduced cost for servicing. From the service provider point of view, consideration is required in better utilization of resources to gain the cost for the cloud resources provided by them. Future consideration can deal with scheduling of tasks to be performed dynamically.

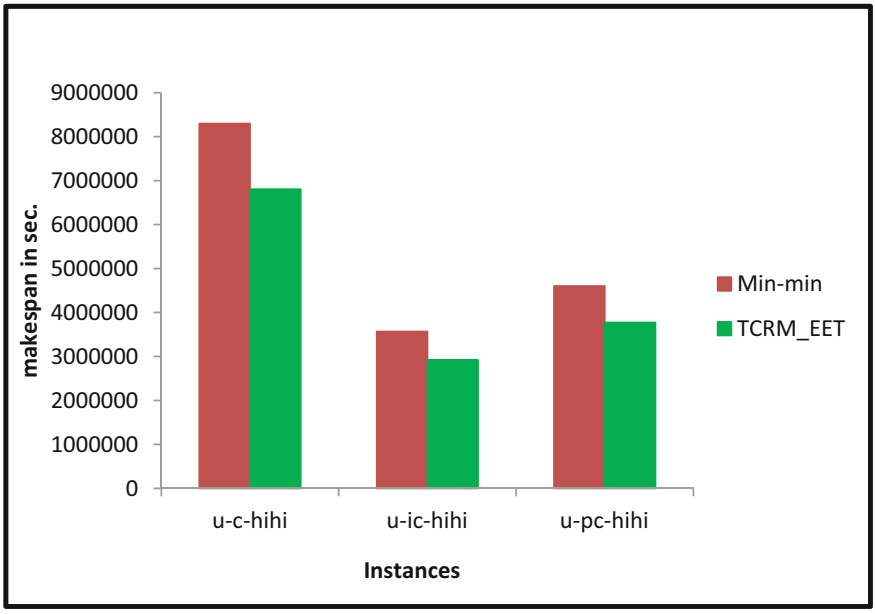


Fig. 3. Makespan—High Task/Cloud Heterogeneity

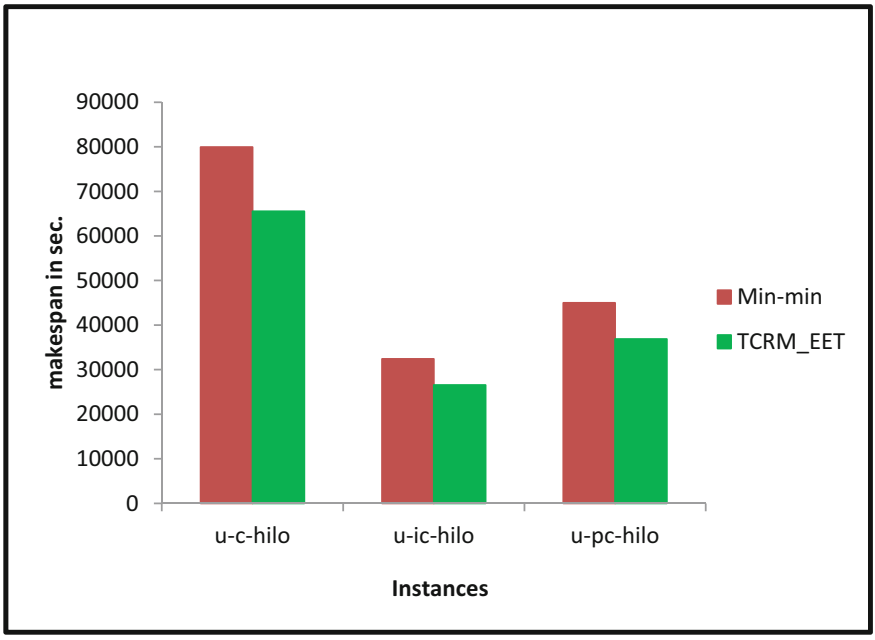


Fig. 4. Makespan: High Task and Low Cloud Heterogeneity

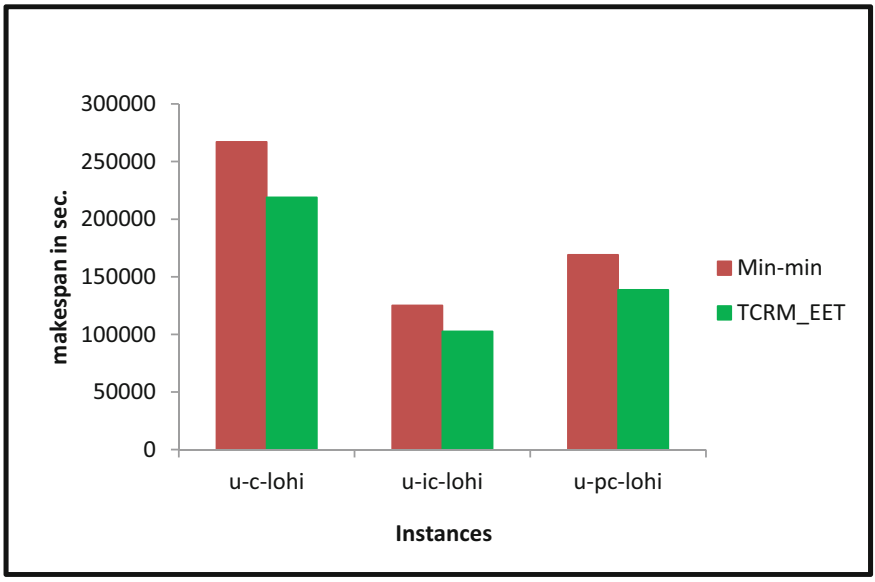


Fig. 5. Makespan—Low Task and High Cloud Heterogeneity

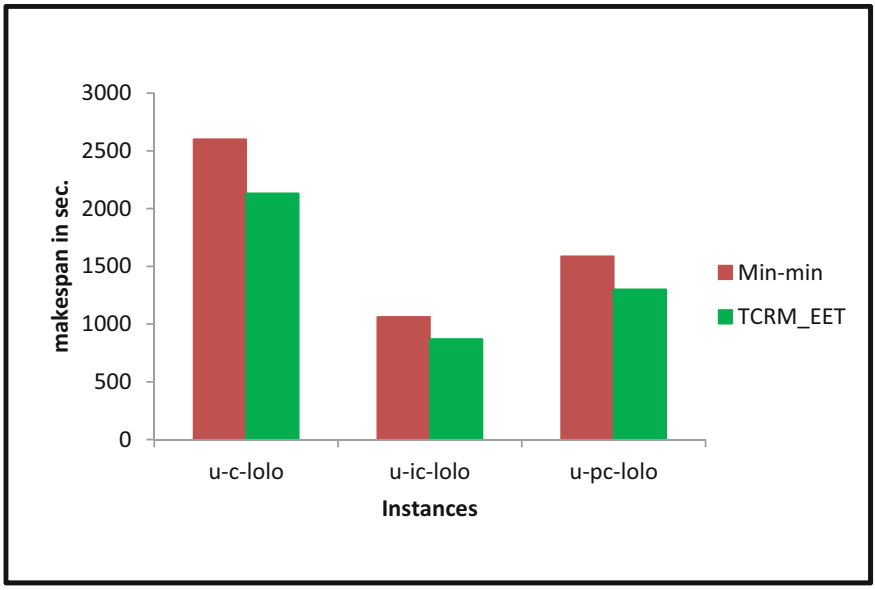


Fig. 6. Makespan—Low Task/Cloud Heterogeneity

References

1. Shah, M. N1., Patel, Y.: A survey of task scheduling algorithm in cloud computing. Int. J. Appl. Innov. Eng. & Manag (IJAIEM), **4**(1),(2015)
2. Ramana, S., Murthy, M.V.R., Bhaskar, N.: Ensuring data integrity in cloud storage using ECC technique. Int. J. Adv. Res. Sci. Eng., BVC NS CS 2017, **06**(01), 170–174 (2017)
3. Mathur, P., Nishchal, N.: Cloud Computing: New challengeto the entire computer industry. In: International conference on parallel, distributed and grid computing (PDGC–2010)
4. Alugubelli, R.: Data mining and analytics framework for healthcare. Int. J. Creat. Res. Thoughts (IJCRT). **6**(1), 534–546 (2018), ISSN:2320–2882
5. Srinivasa, R.S.K.: Classifications of wireless networking and radio. Wutan Huatan Jisuan Jishu, **14**(11), 29–32 (2018)
6. Ahmad, I., Pothuganti, K.: Smart field monitoring using toxtrac: a cyber-physicalsystem approach in agriculture. International conference on smart electronics and communication (ICOSEC) pp. 723–727, (2020)
7. Balne, S., Elumalai, A.: Machine learning and deep learning algorithms used to diagnosis of Alzheimer's: Review. Materials Today: Proceedings (2021). https://doi.org/10.1016/j.matpr.2021.05.499
8. Koripi. M.: 5G Vision and 5g standardization. Parishodh J. **10**(3), 62–66 (2021)
9. Koripi. M.: A Review on secure communications and wireless personal area networks (WPAN). Wutan Huatan Jisuan Jishu, 17(7), 168–174 (2021)
10. Srinivasa, R.S.K.: A Review on wide variety and heterogeneity of iot platforms. Int. J. Anal. Exp. Modal Anal., **12**(1), 3753–3760 (2020)
11. Bhaskar, N., Ramana, S., Murthy, M.V.R.: Security tool for mining sensor networks. Int. J. Adv. Res. Sci. Eng., BVC NS CS 2017, **06**(01), 16–19 (2017). ISSN No: 2319–8346
12. Koripi, M.: A review on architectures and needs in advanced wireless communication technologies. J. Compos. Theory, **13**(12), 208–214 (2020)
13. Srinivasa, R.S.K.: Infrastructural constraints of cloud computing. Int. J. Management. Technol. Eng. **10**(12), 255–260 (2020)
14. Kamalam, G.K., Sentamilselvan, K.: SLA-based group tasks max-min (gtmax-min) algorithm for task scheduling in multi-cloud environments. In: Nagarajan, R., Raj, P., Thirunavukarasu, R. (eds.) Operationalizing multi-cloud environments. EICC, pp. 105–127. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-74402-1_6
15. Kamalam, G.K., Sentamilselvan, K.: Limit value task scheduling (lvts): an efficient task scheduling algorithm for distributed computing environment. Int. J. Recent. Technol. Eng. (IJRTE). **8**(4), 10457−10462 (2019)
16. Kamalam, G.K., Anitha, B., Mohankumar, S.: Credit score tasks scheduling algorithm for mapping a set of independent tasks onto heterogeneous distributed computing. Int. J. Emerg. Technol. Comput. Sci. & Electron (IJETCSE). **20**(2), 182–186, (2016)
17. Kamalam, G.K., Murali Bhaskaran, V.: A new heuristic approach: min-mean algorithm for scheduling meta-tasks on heterogeneous computing systems. Int. J. Comput. Sci. Netw. Secur. **10** (1), 24–31 (2010)
18. Kamalam, G.K., Murali Bhaskaran, V.: An improved min-mean heuristic scheduling algorithm for mapping independent tasks on heterogeneous computing environment. Int. J. Comput. Cogn. **8**(4), 85–91 (2010)



BTSAH: Batch Task Scheduling Algorithm Based on Hungarian Algorithm in Cloud Computing Environment

Gobichettipalayam Krishnaswamy Kamalam(✉), Sandhiya Raja, and Sruthi Kanakachalam

Kongu Engineering College, Tamil Nadu, Perundurai, India
{sandhiya.it,sruthi.it}@kongu.edu

Abstract. Cloud computing is an on-demand computing service that enables the accessibility of information systems resources, notably data management and computational power, without the user being involved directly in direct active administration. Large clouds frequently contain services that are distributed across numerous locations, each of which is a data centre, also known as a cloud centre. The fundamental reason for cloud computing's appeal is the on-demand processing service, which allows users to pay only for what they use. Thus, cloud computing benefits customers in various ways through the internet. Cloud deployment models include SAAS, IAAS and PAAS. A lot of research is being done on IAAS because all consumers want a complete and appropriate allocation of requirements on the cloud. As a result, a major primary objectives of cloud technology is providing excellent remote access to resources so that advantage or profit may be maximised. The proposed methodology Batch Task Scheduling Algorithm based on Hungarian Algorithm (BTSAH) efficient locates a cloud resource that better suits to the constraint of tasks grouped in batches depending on the availability of cloud resource to achieve better utilization of resource, minimum completion time of tasks termed makespan. The proposed approach gains the advantage of Hungarian algorithm in achieving efficient scheduling technique. This paper compares and contrasts simulation analysis of the most popular and extensively used cloud computing scheduling algorithms Min-Min scheduling methodology.

Keywords: Task scheduling · Hungarian method · Min-Max scheduling

1 Introduction

Cloud computing has emerged as an important and popular technology across today's globe. Cloud customers have become more reliant on cloud services in recent years, necessitating the provision of high-quality, efficient, and dependable services. Implementing these services can be accomplished through a variety of techniques. Scheduling task is one of the most significant elements [1, 2].

The scheduling process entails allocating resources to certain tasks in order to complete them efficiently. The primary goals of scheduling are effective resource use; optimizing the server usage allocated to tasks; optimizing the resource utilization; load

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 690–702, 2023.
https://doi.org/10.1007/978-3-031-27409-1_63

balancing; and completing the activities with greater priority while minimizing both completion time and average waiting time. Some scheduling methods also consider QOS parameters. Furthermore, the primary benefits of scheduling are improved performance and increased system throughput. Making span, load balance, deadlines, processing time, and sustainability are all frequent parameters in scheduling algorithms. The experimental results reveal that some standard scheduling methods do not perform well in the cloud and that there are some challenges with implementing them in the cloud. Scheduling algorithms are classified into two modes: batch and online. The jobs in the first category are organized into a predetermined set based on their arrival in the cloud. Batch mode scheduling approaches include FCFS, SJF, RR, Min-Min, Max-Min, and RASA. Jobs are scheduled solely at their arrival time in the second mode, which is online. The example of scheduling tasks in online mode is most fit task [2–4].

The proposed work comprises completing a comparative analysis of research for the most prevalent task scheduling algorithms, namely FCFS, STF, LTF, and RR utilizing the CloudSim simulator toolkit and accounting for time and space shared scheduling allocation principles. The time duration for tasks in Vm is utilized to derive algorithm performance metrics [5–7].

The quality and speed of the schedule are the primary concerns of task scheduling algorithms. The Min-Min method essentially finishes the job first and has the relatively short overall finishing time, giving it the benefit of simplicity and relatively short completion time. The scheduling mechanism Min—Min is investigated in this study. The results show that the proposed approach works well in a cloud computing context [8].

The Min-Min (Minimum-Minimum completion time) method is a type of heuristic dynamic task scheduling system. The main goal of this approach is to generate a large number of jobs that may be assigned to run on the fastest resources. The min-min method is a fundamental job scheduling mechanism in the environment of cloud computing. This approach uses all currently available system resources to estimate the minimal time duration for each job. The work that takes the least amount of time to complete is chosen and assigned to the appropriate processor. After deleting the newly mapped job, the method is continued till all scheduled task sets are emptied [9, 10].

Scheduling jobs in order of priority is really a challenge since each task needs to be completed in a short amount of time. Priority should be taken into account by some work scheduling algorithms. There are several algorithms that consider job priority in order to handle this challenge. This problem may be solved using the combinatorial optimization technique. The Hungarian Method is one example of this algorithm. This approach solves the specified issue in polynomial time. Denes Konig and Jeno Egrary, two prominent Hungarian mathematicians, invented this approach in 1957. This strategy can be used in cloud technology to improve scheduling results [11, 12].

Furthermore, the study provides a comprehensive literature on various job scheduling methods in the cloud computing environment. The following is how this paper is structured: Section 2 conducts a literature study, and Sect. 3 explains the BTSAH algorithm. Section 4 outlines the findings and discussion; Section 6 highlights the conclusions and future work [8, 9].

1.1 Literature Review

Cloud computing is a modern technology that uses the internet to fulfil users in many ways. Cloud providers primarily offer three types of services: SAAS, PAAS, and IAAS. Numerous studies on infrastructure are conducted since all customers want adequate cloud resource allocation. A crucial issue to consider in the cloud is the scheduling of jobs according to requirements. For priority-based scheduling, there are several methods available. The allocation of jobs and resources is focused on the Hungarian model, which priorities resources and employment to satisfy the requirements. The complexity and time requirements of the Hungarian approach are different from those of the conventional methods [1, 2].

Scheduling algorithms demonstrate the critical function they play in the environment of cloud computing in determining a potential timetable for the work. Because the goal is to achieve the shortest total execution time, existing literature has demonstrated that the task scheduling issue is NP-Complete. The Hungarian method, a well-known optimization technique, is the foundation of the proposed pair-based job scheduling solution for cloud computing environments. By modelling the suggested approach and contrasting it with three already-in-use algorithms: first-come, first-served; the Hungarian method with lease period; and the Hungarian method with reversed lease period in 22 distinct datasets, the performance assessment demonstrates that, when compared to current methods, the suggested approach yields a superior layover time [2, 4].

Using the internet and a pay-per-use model, cloud computing distributes data and computational resources. Using this, software gets automatically updated. Scheduling in computing is a technique for allocating tasks to resources that can complete them after they have been specified through some mechanism. It could consist of virtual compute components like threads, processors, or flows of data planned on hardware resources such as CPUs. In cloud computing, the primary issue that lowers system performance is task scheduling. A task-scheduling method must be effective in order to boost system performance. Current task scheduling algorithms are focused with task available resources, CPU resources, processing time, and computational costs. An effective task-scheduling method helps to decrease wait time. In addition, this algorithm uses fewer resources and takes less time to execute. All jobs must be independent of one another according to the proposed algorithm, and when a task is scheduled for execution, it automatically completes it [3, 4, 20].

A cloud is made up of a number of virtual machines that can be used for both storage and computing. The efficient delivery of remote and geographically dispersed resources is the primary goal of cloud computing. Scheduling is one of the difficulties the cloud, which is always evolving, encounters. A computer system's ability to do work in a particular order is governed by a set of principles known as scheduling. When the circumstances and the nature of tasks change, a competent scheduler adjusts their scheduling approach. For task execution efficiency and comparability with FCFS and Round Robin Scheduling, a Generalized Priority method was introduced in the research work. Testing the technique in the cloud Sim toolkit reveals that it performs better than other conventional scheduling algorithms [4, 5].

Manufacturing scheduling is becoming increasingly important as production shifts from restricted variety to high volume to a large variety of low volume. Manufacturing

scheduling issues cannot be solved using the Hungarian algorithm for resource allocation because this algorithm's solutions may contradict the rules governing the priority of the processes that make up specific manufacturing tasks. Multiple approaches for assigning values to the periods of particular machines assigned to processes are presented in this research in order to employ the Hungarian approach to scheduling challenges. According to early assessments, it is anticipated that a scheduler based on a Hungarian algorithm can provide effective schedules when machine limitations are not difficult and scheduling horizons are sufficiently large in comparison to the durations of jobs [5, 6].

A new crucial point in the importance of network computing is cloud computing. It offers increased productivity, significant expandability, and quicker and easier programme development. The current programming approach, the upgraded IT architecture, and the execution of the new approach to business comprise the fundamental content. Task scheduling algorithms have a direct influence on quality and timeliness of the schedule. The Min-Min algorithm is simple and has least time duration, and initially it just performs the job with the least overall finishing time [6, 8].

Cloud computing is a popular computing paradigm that provides high dependability and on-demand resource availability. User's requirements are met by creating a virtual network with the necessary settings. The necessity for optimum resource use of the cloud resources, however, has grown urgent given the constantly growing pressure on these resources. The suggested work analyses the feasibility of the Hungarian algorithm for load transfer in cloud to FCFS. The computations, which were done in CloudSim, show a significant improvement in a variety of performance metrics. When the Hungarian method was compared to FCFS, the end time of a given work schedule was decreased by 41%, as well as the overall runtime decreased by 13% [7, 15].

Cloud-based computing resources that are accessible over the internet offer simple and on-demand network connectivity. With the use of cloud services, individuals and companies may effortlessly access hardware and software, including networks, storage, servers, and applications that are situated remotely. To ensure optimal resource consumption, efficiency, and shorter turnaround times, the jobs submitted to this environment must be completed on time, utilizing the resources available, which calls for an effective task scheduling algorithm for allocating the tasks properly. Small-scale networked systems can utilize the use of Max-min and Min-min. By scheduling large tasks ahead of smaller ones, Improved Max-min and Improved Max-min seek to accomplish resource load balancing [8, 16].

In the world of cloud computing, scheduling user activities is a highly difficult operation. The scheduling of lengthy tasks might not be possible with the min-min method. As a result, this work proposes an enhanced min-min method that is based on the three requirements as well as the min-min algorithm. The dynamic priority model, service cost, and service quality are the three restrictions imposed by the simulation experiment using the freeware CloudSim. The experimental findings demonstrate that it may boost resource utilization rate, enable extended tasks to run in a fair amount of time, and satisfy user needs when compared to the conventional min-min method [9, 17].

With a utility computing paradigm where customers pay according to utilization, cloud computing systems have seen a considerable increase in popularity in recent years.

The key objectives of cloud computing is to maximize profit while enabling effective remote access to resources. Thus, scheduling which focuses on allocating activities to the available resources at a specific time is the main challenge in developing cloud computing systems. Job scheduling is critical for improving cloud computing performance. The key issues in job scheduling is the allocation of workloads among systems in order to optimise QoS metrics. This study provides a simulated comparison of the most well-known and widely used task scheduling algorithms in cloud computing, notably the FCFS, STF and RR algorithms [10, 18].

2 BTSAH Algorithms

In Fig. 1, our proposed heuristic BTSAH is figured out. Cloud user's tasks are divided into batches based on the availability of cloud resources. If the number of cloud resources available for servicing is 'μ', then the number of batches for scheduling will be $\frac{\mu}{\lambda}$where, λrepresentsthenumberoftaskstobescheduled. Batch of tasks are scheduled one batch after the other. The ready time of the cloud resources are updated as soon as one batch of tasks is scheduled. An optimal assignment of tasks and cloud resource is performed using Hungarian algorithm [13, 14, 19].

BTSAH Algorithm

Step1: Form a batch of μ * μ matrix comprising the ETC matrix value of the tasks selected for scheduling
Step2: Identify the minimum value in each row and reduce the values in row by that value.
Step3: Identify the minimum value in each column and reduce the values in column by that value.
Step4: Find the smallest count of horizontal line and vertical line which is used for marking zeros in μ * μ matrix. If the count equals the size of the matrix μ, assignment that results is called optimal and exists among the zeros. Go to step6. If it is less than μ, go to Step 5.
Step5: to create more zeros; minimum value among the values that are uncovered by the lines in Step4 are identified as 'k'. Reduce uncovered values by value 'k', and increase values that are covered twice by value 'k'. Go to step 4.
Step6: For each row select the column with zero, and assign the corresponding task to the corresponding cloud resource, if the resulting assignment reduces makespan, then make the assignment, otherwise keep the previous assignment itself.
Step7: Now update the ready time of all cloud resource and take another batch of μ tasks and proceed to step1 until all the tasks are scheduled.
Step8: compute makespan

$$\text{makespan} = \max\left(\text{CT}\left(\text{TS}_i, \text{CL}_j\right)\right), 0 \le i < \tau \text{ and } 0 \le j < \mu$$

Fig. 1. The Pseudo-code of proposed BTSAH algorithm

A. Evaluation Parameters

Metrics to bring out the importance and significance compared with the existing bench mark algorithm is makespan. The makespan is to define the scheduling strategy of

the tasks by considering the time taken for the completion of tasks Ti grouped in batches in TSi submitted to the cloud resources CRk present in the distributed computing environment is calculated using the formula stated below

$$\text{makespan}[k] = \sum_{i=1}^{t} \text{ETC}_{ik}$$

The overall completion time of the entire batch of tasks, is computed as:

$$\text{makesapn} = \max(\text{makespan}[k], 1 \leq k \leq \mu)$$

Bench mark data set details for the simulation environment is presented below in Table 1.

Table 1. Simulation Environment

Benchmark Model	Descriptions
Size of matrix ETC	$\tau*\mu$
Unit of tasks	τ
Cloud resources	μ
Instance count	12
Matrix count in each instance	100
Number of batches	τ / μ

Analysing the efficiency in terms of time taken for execution depends on identifying the appropriate cloud resource for the corresponding tasks. Since the proposed approach follows a batch of tasks and scheduled using Hungarian technique leads to $O(\mu^3)$. It is time efficient obviously comparing a scheduling problem when the number of combinations of tasks and cloud resources to be considered for making a best choice seems to NP-complete problem.

B. An Illustration

Consider a scenario with six tasks and three cloud resources as a basic outline. In Table 1, the tasks Ti, CRj, and EETij are shown.

For the given scenario in Table 2, the six tasks are divided into two batches comprising three tasks each and are performed scheduling in two batches and are shown in Fig. 2 and 3.

Table 3. Presents the tasks, cloud resources mapping for Min-Min algorithm and the proposed BTSAH algorithm.

From Fig. 1 and 2, Table 2 it clearly delivers that BTSAH heuristic performs better mapping using Hungarian approach and brings out least makespan and better CR utilization than Min-min heuristic.

Table 2. Consistent low task low machine heterogeneity

Tasks/Cloud resource	CR1	CR2	CR3
T1	70.1	111.7	117.6
T2	55.4	70.6	72.5
T3	104.0	106.8	118.7
T4	113.6	161.2	186.4
T5	46.0	53.0	54.5
T6	29.5	33.2	80.5

3 Results and Discussion

Simulation work is carried out for 12 different possible characteristice of ETC matrix, task and resource heterogeneity, and consistency. ETC matrix value is generated as an avearge of 100 ETC matrices for all 12 possible characteristic combination. The size of the genertaed matix is τ*μ, where $\tau = 512$ tasks and $\mu = 16$ cloud resources.

Proposed heuristic BTSAH experimental results computed and performance analyzed are carried out based on the analytical benchmark.

12 ETC matrix entities are fundamentally defined as u-x-yyzz.k, u- defines division allocation uniformly for creating ETC twelve matrix instances. x-specifies consistency, value x is c-consistent, i-inconsistent, pc-partially consistent, yy-represents task heterogeneity, zz-represents resource heterogeneity. Makespan of heuristic (BTSAH) compared with the existing heuristic Min-Min for twelve ETC matrix instances is shown in Fig. 4, 5, 6, 7, and 8.

Graphical representation presents the comparison results of heuristic (BTSAH and Min-Min) and proposed heuristic BTSAH performs better than Min-Min and it has a least makespan.

4 Conclusion and Future Work

Common huddle in cloud region is scheduling resource and tasks efficiently. The proposed methodology Batch Task Scheduling Algorithm based on Hungarian Algorithm (BTSAH) efficient locates a cloud resource that better suits to the constraint of tasks grouped in batches depending on the availability of cloud resource to achieve better utilization of resource, minimum completion time of tasks termed makespan. The proposed approach gains the advantage of Hungarian algorithm in achieving efficient scheduling technique. Thus, the proposed heuristic technique BTSAH meet out the scheduling decisions efficiently considering a batches of tasks satisfying time efficiency and evaluation metric makespan resulting in minimum value. Better utilization of resources achieved through Hungarian approach in BTSAH algorithm instead of mapping a numerous tasks to same cloud resource. BTSAH meets out for static scheduling. Future consideration to be dealt with the dynamic environment considering arrival time of tasks to perform dynamic scheduling with QoS constraints to meet the pay per use policy to achieve cost

Batch-1 Tasks for Scheduling

Tasks/Cloud Resource	CR1	CR2	CR3
T1	70.1	111.7	117.6
T2	55.4	70.6	72.5
T3	104.0	106.8	118.7

Step 1: Subtract minimum of every row

Tasks/Cloud Resource	CR1	CR2	CR3
T1	0.0	41.6	47.5
T2	0.0	15.2	17.1
T3	0.0	2.8	14.7

Step 2: Subtract minimum of every column

Tasks/Cloud Resource	CR1	CR2	CR3
T1	0.0	38.8	32.8
T2	0.0	12.4	2.4
T3	0.0	0.0	0.0

Step 3: subtract the smallest uncovered entry from all uncovered rows. Smallest entry is 2.4.

Tasks/Cloud Resource	CR1	CR2	CR3
T1	-2.4	36.4	30.4
T2	-2.4	10.0	0.0
T3	0	0	0

Step 4: add the smallest entry to all covered columns. Smallest entry is 2.4.

Tasks/Cloud Resource	CR1	CR2	CR3
T1	0	0	36.4	30.4
T2	~~0~~	~~0~~	10	0
T3	~~2.4~~	0	0

Now, the zeros are covered with three lines which is equal to the size of the matrix, hence the optimal assignment after batch 1 – is

Tasks	Allocated Cloud Resource	ECT
T1	CR1	70.1
T2	CR3	72.5
T3	CR2	106.8

Fig. 2. Batch-1 tasks and cloud resource mapping

efficiency for cloud consumers and better cloud resource utilization rate for cloud service providers.

Batch-2 Tasks for Scheduling

Tasks/Cloud Resource	CR1	CR2	CR3
T4	113.6	161.2	186.4
T5	46.0	53.0	54.5
T6	29.5	33.2	80.5

Step 1: After updating Ready time of cloud resource

Tasks/Cloud Resource	CR1	CR2	CR3
T4	183.7	268.0	258.9
T5	116.1	159.8	127.0
T6	99.6	140.0	153.0

Step 2: Subtract minimum of every column

Tasks/Cloud Resource	CR1	CR2	CR3
T1	0.0	43.9	64.3
T2	0.0	3.3	0.0
T3	0.0	0.0	42.5

Step 2: Subtract minimum of every row

Tasks/Cloud Resource	CR1	CR2	CR3
T1	0.0	84.3	75.2
T2	0.0	43.7	10.9
T3	0.0	40.4	53.4

Now, the zeros are covered with three lines which is equal to the size of the matrix, hence the optimal assignment after batch 1 – is

Tasks	Allocated Cloud Resource	ECT
T4	R1	183.7
T5	R3	127
T6	R2	140

Fig. 3. Batch-2 tasks and cloud resource mapping

Table 3. Ti and CRj Mapping

Min-Min			BTSAH Algorithm		
Tasks Ti	Cloud Resource CRj allocated	Expected Completion Time ECTij	Tasks Ti	Cloud Resource CRj allocated	Expected Completion Time ECTij
T6	CR1	29.5	T1	CR1	70.1
T5	CR2	53	T2	CR3	72.5
T2	CR3	72.5	T3	CR2	106.8
T1	CR1	99.6	T4	CR1	183.7
T3	CR2	159.8	T5	CR3	127
T4	CR1	213.2	T6	CR2	140
Makespan – 213.2			Makespan – 183.7		

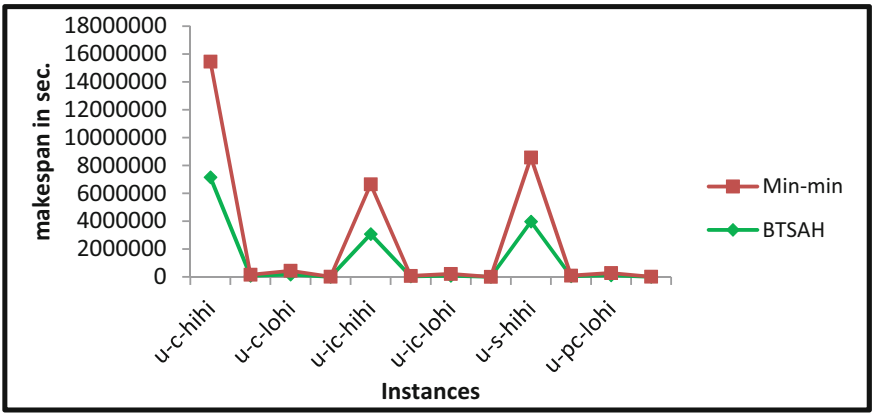


Fig. 4. Makespan values

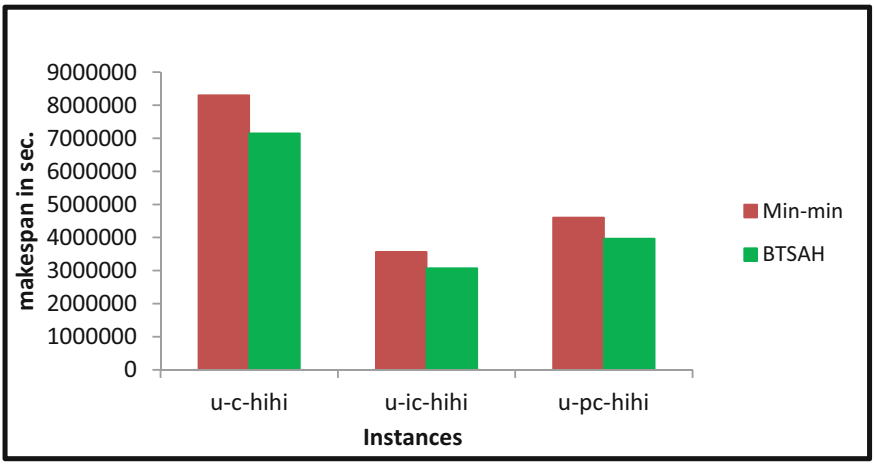


Fig. 5. Makespan - High task/cloud heterogeneity

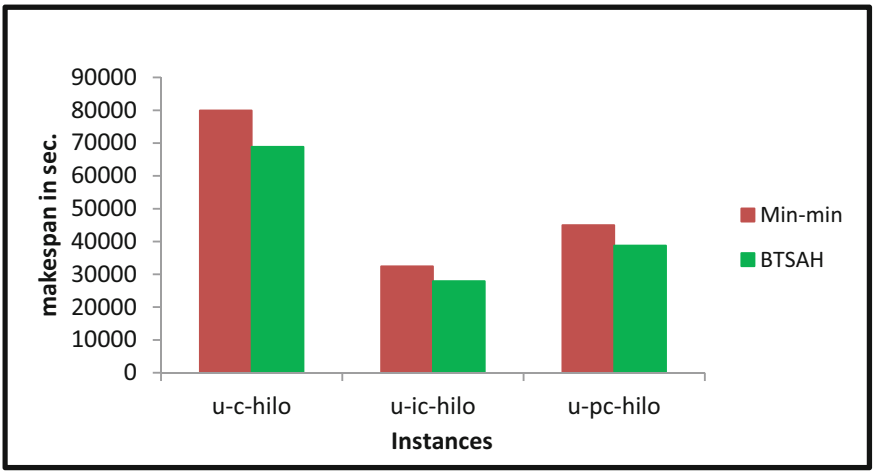


Fig. 6. Makespan: High task and low cloud heterogeneity

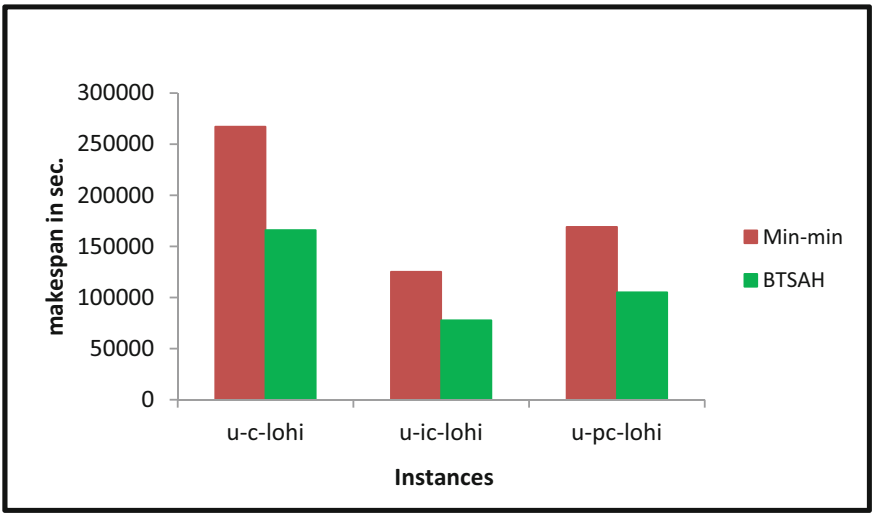


Fig. 7. Makespan - Low task and high cloud heterogeneity

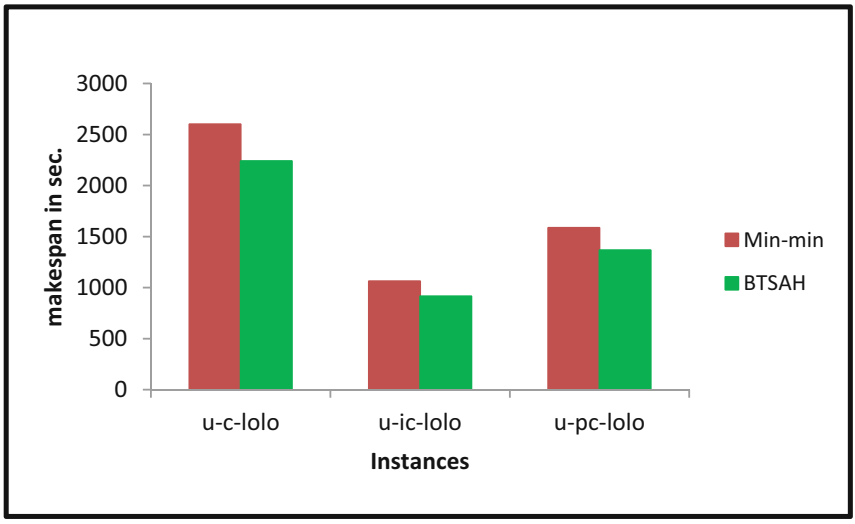


Fig. 8. Makespan—Low task/Cloud heterogeneity

References

1. Patel, R.R., Desai, T.T., Patel, S.J.: Scheduling of jobs based on Hungarian method in cloud computing. In: 2017 International conference on inventive communication and computational technologies (ICICCT). IEEE (2017)
2. Panda, S.K., Nanda, S.S., Bhoi, S.K.: A pair-based task scheduling algorithm for cloud computing environment. J. King Saud Univ.-Comput. Inf. Sci. **34**(1), 1434–1445 (2022)

3. Razaque, A., et al.: Task scheduling in cloud computing. in 2016 IEEE long island systems, applications and technology conference (LISAT). IEEE (2016)
4. Agarwal, D., Jain, S.: Efficient optimal algorithm of task scheduling in cloud computing environment. arXiv preprint arXiv:1404.2076, (2014)
5. Tamura, S., et al.: Feasiblity of hungarian algorithm based scheduling. In: 2010 IEEE international conference on systems, man and cybernetics. IEEE (2010)
6. Wang, G., Yu, H.C.: Task scheduling algorithm based on improved Min-Min algorithm in cloud computing environment. In: Applied mechanics and materials. Trans Tech Publ (2013)
7. Bala, M.I., Chishti, M.A.: Load balancing in cloud computing using Hungarian algorithm. Int. J. Wirel. Microw. Technol. **9**(6), 1–10 (2019)
8. Sindhu, S., Mukherjee, S.: Efficient task scheduling algorithms for cloud computing environment. In: International conference on high performance architecture and grid computing. Springer(2011)
9. Liu, G., Li, J., Xu, J.: An improved min-min algorithm in cloud computing. In: Proceedings of the 2012 International conference of modern computer science and applications. Springer (2013)
10. Alhaidari, F., Balharith, T., Eyman, A.-Y.: Comparative analysis for task scheduling algorithms on cloud computing. In: 2019 International conference on computer and information sciences (ICCIS). IEEE(2019)
11. Kamalam, G.K., Sentamilselvan, K.: SLA-based group tasks max-min (gtmax-min) algorithm for task scheduling in multi-cloud environments. In: Nagarajan, R., Raj, P., Thirunavukarasu, R. (eds.) Operationalizing Multi-Cloud Environments. EICC, pp. 105–127. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-74402-1_6
12. Kamalam, G.K., Sentamilselvan, K.: Limit value task scheduling (lvts): an efficient task scheduling algorithm for distributed computing environment. Int. J. Recent. Technol. Eng. (IJRTE), **8**(4), 10457−10462 (2019)
13. Kamalam, G.K., Anitha, B., Mohankumar, S.: Credit score tasks scheduling algorithm for mapping a set of independent tasks onto heterogeneous distributed computing. Int. J. Emerg. Technol. Comput. Sci. & Electron (IJETCSE), **20**(2), 182–186 (2016)
14. Kamalam, G.K., Murali Bhaskaran, V.: A new heuristic approach: min-mean algorithm for scheduling meta-tasks on heterogeneous computing systems. Int. J. Comput. Sci. Netw. Secur. **10**(1), 24–31 (2010)
15. Kamalam, G.K., Murali Bhaskaran, V.: An improved min-mean heuristic scheduling algorithm for mapping independent tasks on heterogeneous computing environment. Int. J. Comput. Cogn. **8**(4), 85–91, (2010)
16. Ahmad, I., Pothuganti, K.: Smart field monitoring using toxtrac: a cyber-physicalsystem approach in agriculture. In: International conference on smart electronics and communication (ICOSEC), pp. 723–727, (2020)
17. Balne, S., Elumalai, A.: Machine learning and deep learning algorithms used to diagnosis of Alzheimer's: Review. Materials Today: Proceedings (2021). https://doi.org/10.1016/j.matpr.2021.05.499
18. Koripi, M.: 5G Vision and 5g standardization. Parishodh J. **10**(3), 62–66 (2021)
19. Koripi, M.: A review on secure communications and wireless personal area networks (WPAN). Wutan Huatan Jisuan Jishu, **17** (VII), 168–174, (2021)
20. Srinivasa, R.S.K.: A Review on wide variety and heterogeneity of iot platforms. Int. J. Anal. Exp. Modal Anal., **12**(1), 3753–3760 (2020)



IoT Data Ness: From Streaming to Added Value

Ricardo Correia(✉), Cristovão Sousa, and Davide Carneiro

Escola Superior de Tecnologia e Gestão, Politécnico do Porto, Porto, Portugal
{8150214,cds,dcarneiro}@estg.ipp.pt,
ricardo.correia.a@icloud.com

Abstract. The industry 4.0 paradigm has been increasing in popularity since its conception, due to its potential to leverage productive flexibility. In spite of this, there are still significant challenges in industrial digital transformation at scale. Some of these challenges are related to Big Data characteristics, such as heterogeneity and volume of data. However, most of the issues come from the lack of context around data and its lifecycle. This paper presents a flexible, standardized, and decentralized architecture that focuses on maximizing data context through semantics to increase data quality. It contributes to closing the gap between data and extracted knowledge, tackling emerging data challenges, such as observability, accessibility, interoperability, and ownership.

1 Introduction

In the recent past, the Internet of Things (IoT) has emerged as a revolutionary paradigm for connecting devices and sensors. This allows visibility and automation of an environment, opening the path to industrial process optimization which might lead to improved efficiency and increase flexibility [28]. When that paradigm was applied to the industry world it became the fourth industrial revolution [33], seeking to improve efficiency and provide visibility over not only the machines and products but also the whole value chain. The benefits of this new age of industrialization, also known as Industry 4.0, has been enabling small, medium, and large companies to improve their ways of working, thereby increasing quality and quantity of the product and services while reducing costs [5].

The adoption of IoT in the industry has been steadily increasing not only vertically, but also horizontally. Vertical growth is driven by adding all kinds of sensors, wearables, and actors, estimating the market to grow to 102.460 million USD by the year 2028 [29]. This is because more clients and business departments are interested in the data available. In contrast, horizontal growth has been stimulated by the integration of multiple companies, producing information to the same data repository [22,26,28]. With machine learning, heavy computation processes, and powerful visualization tools, the data collected is empowered to enhance process efficiency and predictability across workstations, resulting in a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 703–713, 2023.
https://doi.org/10.1007/978-3-031-27409-1_64

massive increase in productivity and lower costs [1,3]. However, without a scalable architecture in place to extract and improve data quality, the data gathered within the environment becomes an asset difficult to convert into value. This leads to what data scientists describe as a Data Swamp [17]. The quality of data extracted is a crucial factor in the success of an IIoT environment since the data obtained will heavily contribute to key business decisions and even automated actions within the production floor [6,22]. It is possible that such scenarios could result in monetary losses or even security risks if not handled correctly. For this reason, one cannot rely solely on the sensors to produce quality data, since many of them are pruned to failure [20,21]. Instead, a resilient architecture capable of identifying faulty data, managing data quality metrics, and ensuring confidence in the englobing environment must be implemented. Adding quality restrictions to the gathered data allows users to promote a much more productive communication between machines, processes, people, and organisations.

One of the most significant aspects of data quality is the observability level that can be inferred from it [27,30]. This is especially relevant when the data is getting more and more complex due to transformations and relationships. For this reason, an architecture designed to cope with IIoT demands must include a feature to provide data observability at a large scale, thus providing much-needed insights into the data.

The main purpose of this work is to develop an architecture capable of facing today's data management challenges, with a focus on iteratively enriching metadata with context through semantics. To be successful, the architecture must meet additional requirements. These include decentralized components, centralized infrastructure, resilient, accessible, and observable data and metadata repository, with lineage capabilities, data ownership information, and scalable data transformation tools.

Besides these points, the architecture should also conform to the existing reference architecture principles [9,15], such as modular design, horizontal scalability, adaptable and flexible, and performance efficient.

2 Architectures for Data Management

The data collected from IIoT environments exhibits some Big Data characteristics that need significant effort to be effectively used for value creation [13]. For this reason, in order to be successful in designing an IIoT data management and governance solution it was necessary to analyse existing state-of-the-art architectures with focus on capabilities to close the gap between data and knowledge. Due to the significant volume and heterogeneity of data within these environments, the first strategy analyzed was the implementation of a Data Lake [12].

A Data Lake is a centralized and scalable repository containing considerable amounts of data, either in its original format or as the result of transformation, which needs to be analyzed by both business experts and data scientists [2]. Data Lakes have the capability of processing voluminous and quickly generated unstructured data [25]. The Data Lake should be architected to be divided into

sections, which Bill Inmon refers to as data ponds [17] and some other researchers refer to as zones [31]. As a result, data separations facilitate data lifecycle management and, therefore, data quality. Having the Data Lake lake separated into different sections allows for a more scalable solution since each section can grow separately. As an example, the raw data station and its process pipeline can be scaled to boost a fast data application, providing quick data insights and sacrificing the processed data level; this can be extremely useful depending on the context. Using this architecture, an archival data level can be deployed [17], so old generated data may be stored in a cheaper storage system, allowing the data to be kept as long as possible.

While this architecture was born as a result of the necessity from some data warehouses limitations [27], it has been at the center of the Data Engineering community. However, if on one hand it has been praised by many, on the other hand there are many reports in which this architecture has failed miserably, creating monumental *Data Swamps* [17]. That allowed some products to grow, such as Delta Lake, to respond to such necessities. This most recent iteration of the Data Lake architecture aims to provide more functionality to the Data Lake, contributing features such as stream and batch processing unification, cloud-based storage, and real-time data availability [4].

Although the Data Lake architecture does offer some interesting features that can be used to construct an efficient data management tool, there are some requirements for data quality that this centralized data storage couldn't easily fulfill, metadata management, interoperability and overall data quality [24].

To cope with this, a new data architecture design pattern emerged, The Data Fabric. This novel architecture addresses ingestion, governance, and analytics features within uncontrolled growth of contexts of data [34]. The solution proposed aims to assist in end-to-end data integrations and data discovery through data virtualization, with the assistance of API and data services data across the organization allowing for better discoverability and integration within the environments even if they are in old legacy systems [35]. Instead of proposing a completely revamped design for data management, the data fabric simply seeks to create a virtualization layer on top of each data interaction, such as transformation, storage, and serving, creating a global uniformization to facilitate data access [16]. The development of a Data Fabric inspired architecture allowed for the creation of multiple services and APIs that can interact with multiple data types across multiple tools. This proved efficient for context and semantic metadata management.

Another data architecture that shares a similar purpose to Data Fabric is the Data Mesh. This architecture follows a more decentralised approach and has a significant focus on handling data as a product.

Data Mesh aims to restructure the organizational structure around data utilization, following the concepts laid out by Domain Driven Design [14] to introduce addressing data ownership issues, creating less friction when trying to produce valued information mitigating the problems encountered in Big Data Lakes and warehouses [10].

In order to face the information extracting challenges and inter-divisional problems, the Data Mesh approach proposes a paradigm shift comparable to the microservices architecture, focusing on treating data as a product, and creating data teams to handle the whole subset of data belonging to the business domain, rather than dividing them into teams for the different data processes, such as collection, transformation, and provisioning. This leads to increased ownership of the data itself by the teams, thus leading to more agility when it comes to producing knowledge [11]. Additionally, the Data Mesh paradigm does not aim to replace any of the architectures for data management, it instead aims to restructure the organization around it. This will allow each data team to use the preferred data structure for the specific domain. With these changes, teams are also more incentivized to maintain data quality. This is because they are the owners of that domain of data that will be served to other data teams and customers [23]. These emergent data architecure design patterns for data governance are driven by quality issues of data. The first step was to consider a large centralized repository that handled all the data within the environment, allowing for data quality processes to be applied to the data within. Then a series of services were created to interact with the consumed data and context metadata. And finally, the last iteration allows for multiple features that improve data quality, interoperability, observability, reusability and visibility.

3 Data Observability Challenges within IIoT Data Management

There have been significant efforts to create and iterate data management architectures to close the gap in the level of knowledge that can be extracted from raw data. However, all of those architectures have faced similar problems when formulating their solution [6,37]. In the context of this work, we're interested in exploring the following aspects of data quality: traceability, trust, fit for use, context and semantic value, interoperability and reusability. Different strategies can be used to formulate solutions to these problems, and the prime strategy was the implementation of strong data observability practices.

Derived from the original concept of system observability, data observability has followed the same practices that made systems successful in their monitoring practices. Instead of tracking logs, traces, and metrics as is usual for system observability, data observability practices aim to monitor other concepts such as freshness, distribution, volume, schema, and lineage to prevent downtime and ensure data quality [27]. Each dimension of data observability aims to enrich data quality in different ways. **Freshness** seeks to understand how up-to-date the ingested data is, as well as the frequency at which it is updated. **Distribution** is a dimension that establishes the limits of data values, defining the accepted values that reading can have, and defining outliers and potentially flawed data and sources. **Volume** refers to the completeness of the data, identifying possible data shortages and sources that stopped sending data downstream. **Schema** monitoring keeps track of data structure definitions and changes. And lastly,

lineage is one of the most critical dimensions. This is because it allows us to have traceability of data transformations since its origin, allowing the user to identify possible breaking points and which systems might be impacted.

The dimensions of data observability allowed for some of the challenges to be tackled in an efficient way. Freshness promotes data quality in the fit for use and trust dimension. Distribution and volume improves context value and trust in data. In addition, schema monitoring and lineage allowed for better context value, traceability, and data interoperabilty.

In order to further enhance data quality, the FAIR data principles [32] were incorporated into the solution to address the identified challenges. The FAIR principles were designed to help design a data management platform. These principles emphasize the capacity of computation systems to find, access, interoperate, and reuse data with minimal human intervention. They enable scale with the increasing volume, complexity, and velocity of data [32]. A successful data management strategy is not a goal in itself, but rather a conduit for innovation and discovery within such structures. The authors present four fundamental principles that guide their approach. The dimensions encountered in the FAIR data movement are findability, accessibility, interoperability, and reusability. Each of these dimensions has the goal of improving and facilitating the usage of data. The authors propose a set of techniques to achieve these principles, such as having metadata indexed in a searchable resource, having it accessible via a standard communication protocol, and metadata representation using a standard language so it can be acted upon. These principles led to the development of a service to host context and semantic metadata. The use of such a service enabled continuous data quality improvements through streaming data processes and data observability practices. This newly introduced service also allowed for easier data interoperability and reusability since it hosted all the metadata needed for such use cases.

In order to understand which metadata was kept in the service, first the concept of definition of data must be introduced.The definition of data (DoD) is one of the most critical problems that arises when discussing IIoT data management platforms and one that was highly focused during development. Data and data quality definitions need to be established so the designed solution can fit the needs of the environment.

The definition of data in the IIoT environment englobes sensor readings and context metadata related to the sensor reading, being complemented with representations of other business aspects [19,36]. One key aspect of the definition is the context metadata that can be used to enhance data quality, and that will be maintained within the context service previously described. Data within these IIoT environments have been identified as particularly challenging to handle. Karkouch [18] has identified data quality characteristics and challenges such as uncertainty, erroneous, voluminous, continuous, correlated, and periodic, which contribute to this fact.

In order to better understand the context around data, and how it can help to increase data quality, the concept was approached in twofold, each focus-

ing on concepts inspired by separate dimensions of data observability. In the first place, statistical and computed metadata are presented, which are automatically generated through computing processes along with data processing. In this section, fields that relate to dimensions such as freshness, distribution, and volume can be found. These fields include metrics such as medians, minimums and maximums, outlier percentages, time evolution, missing data percentages, throughput, and much more. The second section centers on the semantic value of the data. The information in this section focuses not only on schema monitoring and lineage dimensions of data observability but also on interoperability capabilities by constructing a semantic net relating entities from the environment to each other. This perspective of data quality is one of the most influential and impactful categorizations within data quality. This is because it strongly contributes to the contextualization of the data within the whole environment of data, improving visibility, interoperability, and discoverability [7,8]. During the construction of this metadata, questions such as what, when, who, why, and how should be asked, and then, additional values should be continually added to enrich the data context that can be valuable for data utilization. Some examples of semantic metadata in IoT include sensor creation date, sensor brand, sensor expiration date, IP address, battery level, owner, location, sensors in the same room, and much more.

4 An Intelligible Data Mesh Based Architecture for IIoT Environments

Given the challenges that were identified, it was designed a FAIR data compliant architecture that focuses on metadata management and standardization to elevate the value of the data. This architecture can be visualized in Fig. 1.

The entry point of the architecture is the Context Broker. Besides being responsible for receiving all the raw data from the IIoT environment, this component is also responsible for the key aspect of metadata management. When raw data comes from the IIoT environment it is ingested by the context broker and is automatically enriched with the context metadata. The base semantics and context management should be managed beforehand, so the semantic graph can have a wider reach.

The context broker can also function as documentation for all the sensors and relations within the environment, and when a newly added component is integrated, the semantics graph should be updated with the revised values, keeping a realistic view of the monitored environment. This environment documentation is used with the aid of shared smart data models.

The component that is responsible for the data storage and delivery is the data gateway. This piece is designed according to the event sourcing pattern, to retain the full history of data, and maximize the interoperability and reusability aspects of the FAIR data principles. Ideally, the data gateway should be decentralized and meant to hold all the data across all the domains and stages of the data lifecycle, providing enough flexibility to satisfy specific business needs,

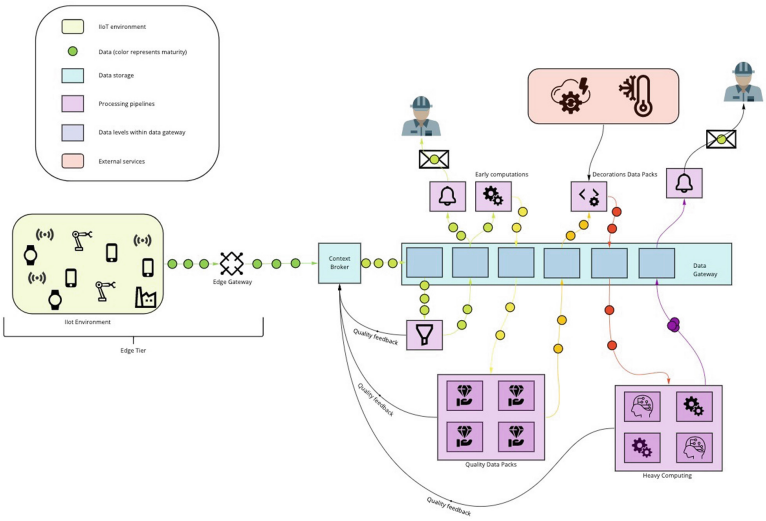


Fig. 1. Data ness architecture

and facilitating discoverability, access, and findability, thus enhancing the other two principles of FAIR data. This design enables the data gateway to be the central point of data access, allowing for processing pipelines to move the data around, third party projects to use the stored data, and data visualization tools to empower environment observability.

The final component implemented in the architecture is the plug-and-play pipeline design. These pipelines are meant to connect to the data gateway and move data around it, performing all necessary computations in between, so data can be iteratively converted into knowledge. The computations that can take place within a pipeline include, but are not limited to, filtration systems, machine learning model building, automated actions, ETL processes, alerting, and data quality metrics. Among the most significant pipeline types are the data context enrichment pipelines, which will take data from a data stage, add context information in the form of *data packs*,[1] and output the newly computed data back to the data gateway to be used in a more mature data stage.

To ensure data lineage capabilities, all pipelines should annotate data with metadata stating which computation has taken place. Such metadata should include information such as pipeline identifier, timestamp that data was consumed, timestamp that data was produced, initial value, and output value. Such data lineage metadata should belong to a shared and common pipeline model that needs to be maintained. This is so the pipelines can be more easily understood so they can be applied across multiple business divisions and data stages.

[1] Data packs represent the information that is added or modified by the pipelines to the data that is processed. The information added can be related to data quality metrics or context information.

Said model should include values such as pipeline name, description, data input requirements, data output format, input model, and ownership.

The pipelines may output results to a diverse range of destinations, such as:

- Data gateway, in the form of the next iteration of enhanced data, passes data to the next data stage.
- Context broker, with updated data context and freshly calculated data quality metrics. This path is specially significant because it allows for context iterations, allowing for continuous progress in data quality reflecting changes within the environment.
- External services, such as alerts, environment updates, or monitoring systems.

The flexibility of pipeline development allows for abstractions in the form of parametrizable variables, which empowers reusability. Also, pipeline development should aim to create simple and business-focused code, following the single responsibility principle which allows for a lower development cycle, high cohesion, and also increasing reusability.

The presented architecture assures the most significant components of the reference architectures of today [9,15], such as: context management with device management and defined ontology, data management with ingestion and provisioning capabilities, analytics processes, visualization support, and decentralization.

5 Conclusions and Future Work

IIoT data management environments enclose many different challenges today. New patterns and technologies emerge, bringing security concerns about the data held, rising the need to understand where the data came from and how it affects the business. All these problems can be boiled down to an understanding of data and, more specifically, its ever-evolving context.

We discussed an architecture that addresses these problems. Design of this system focuses on iteratively enhancing data quality with decentralized components and centralized infrastructure, providing a data management reference system to contribute to the reliability of data quality within the Industry 4.0 paradigm. The purposed architecture follows FAIR data design principles to cope with data observability challenges, towards value added data governance within IIoT realtime environemnts. The results of this research work are to be incorporated into a reference methodology for the development of data quality oriented big data architectures in industry.

Acknowledgments. This work has been supported by national funds through FCT—Fundação para a Ciência e Tecnologia through project EXPL/CCI-COM/0706/2021.

References

1. Adi, E., Anwar, A., Baig, Z., Zeadally, S., Adi, E., Anwar, A., Baig, Z., Zeadally, S.: Machine Learning and Data Analytics for the IOT (2020)
2. Alserafi, A., Abell, A.: Towards information profiling?: Data lake content metadata management (2016). https://doi.org/10.1109/icdmw.2016.0033
3. Ambika, P.: Machine learning and deep learning algorithms on the industrial internet of things (iiot). Adv. Comput. **117**, 321–338 (2020). https://doi.org/10.1016/BS.ADCOM.2019.10.007
4. Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., nski, M.S., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., Xin, R., Zaharia, M., Berkeley, U.: Delta lake: High-performance acid table storage over cloud object stores (2020). https://doi.org/10.14778/3415478.3415560, https://doi.org/10.14778/3415478.3415560
5. Boyes, H., Hallaq, B., Cunningham, J., Watson, T.: The industrial internet of things (iiot): An analysis framework. Comput. Ind. **101**, 1–12 (2018). https://doi.org/10.1016/J.COMPIND.2018.04.015
6. Byabazaire, J., O'hare, G., Delaney, D.: Data quality and trust: review of challenges and opportunities for data sharing in iot. Electronics (Switzerland) **9**, 1–22 (2020). https://doi.org/10.3390/electronics9122083
7. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Sci. J. **14** (2015). https://doi.org/10.5334/DSJ-2015-002/METRICS/, http://datascience.codata.org/articles/10.5334/dsj-2015-002/
8. Ceravolo, P., Azzini, A., Angelini, M., Catarci, T., Cudré-Mauroux, P., Damiani, E., Keulen, M.V., Mazak, A., Keulen, M., Mustafa, J., Santucci, G., Sattler, K.U., Scannapieco, M., Wimmer, M., Wrembel, R., Zaraket, F.: Big data semantics. J. Data Semant. (2018)
9. Cosner, M.: Azure iot reference architecture—azure reference architectures—microsoft docs (2022). https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/iot
10. Dehghani, Z.: How to move beyond a monolithic data lake to a distributed data mesh (2019). https://martinfowler.com/articles/data-monolith-to-mesh.html
11. Dehghani, Z.: Data mesh principles and logical architecture (2020). https://martinfowler.com/articles/data-mesh-principles.html
12. Dixon, J.: Pentaho, hadoop, and data lakes (2010). https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
13. Diène, B., Rodrigues, J.J.P.C., Diallo, O., Hadji, E.L., Ndoye, M., Korotaev, V.V.: Data management techniques for internet of things (2019)
14. Evans, E.: Domain-Driven Design: Tackling Complexity in the Heart of Software. Addison-Wesley (2004)
15. IBM: Internet of things architecture: Reference diagram—ibm cloud architecture center (2022). https://www.ibm.com/cloud/architecture/architectures/iotArchitecture/reference-architecture/
16. IBM: What is a data fabric?—ibm (2022). https://www.ibm.com/topics/data-fabric
17. Inmon, B.: Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump, 1st edn. Technics Publications, LLC, Denville, NJ, USA (2016)
18. Karkouch, A., Mousannif, H., Al, H., Noel, T.: Journal of network and computer applications data quality in internet of things: a state-of-the-art survey. J. Netw. Comput. Appl. **73**, 57–81 (2016)

19. Kim, S., Castillo, R.P.D., Caballero, I., Lee, J., Lee, C., Lee, D., Lee, S., Mate, A.: Extending data quality management for smart connected product operations. IEEE Access **7**, 144663–144678 (2019). https://doi.org/10.1109/ACCESS.2019.2945124
20. Kodeswaran, P., Kokku, R., Sen, S., Srivatsa, M.: Idea: a system for efficient failure management in smart iot environments* (2016). https://doi.org/10.1145/2906388.2906406, http://dx.doi.org/10.1145/2906388.2906406
21. Lin, Y.B., Lin, Y.W., Lin, J.Y., Hung, H.N.: Sensortalk: an iot device failure detection and calibration mechanism for smart farming. Sensors (Switzerland) **19** (2019). https://doi.org/10.3390/s19214788
22. Liu, C., Nitschke, P., Williams, S.P., Zowghi, D.: Data quality and the Internet of Things. Computing **102**(2), 573–599 (2019). https://doi.org/10.1007/s00607-019-00746-z
23. Machado, I.A., Costa, C., Santos, M.Y.: Data mesh: concepts and principles of a paradigm shift in data architectures. Procedia Comput. Sci. **196**, 263–271 (2021). https://doi.org/10.1016/j.procs.2021.12.013
24. Mehmood, H., Gilman, E., Cortes, M., Kostakos, P., Byrne, A., Valta, K., Tekes, S., Riekki, J.: Implementing big data lake for heterogeneous data sources, pp. 37–44. Institute of Electrical and Electronics Engineers Inc. (2019). https://doi.org/10.1109/icdew.2019.00-37
25. Miloslavskaya, N., Tolstoy, A.: Big data , fast data and data lake concepts 2 big data concept. **88**, 300–305 (2016). https://doi.org/10.1016/j.procs.2016.07.439
26. Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M.S., Upadhyay, R., Martynenko, A.: Iot, big data and artificial intelligence in agriculture and food industry. IEEE Internet of Things J. 1–1 (2020). https://doi.org/10.1109/jiot.2020.2998584
27. Moses, B.: The rise of data observability: architecting the future of data trust. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, p. 1657. WSDM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3488560.3510007, https://doi.org/10.1145/3488560.3510007
28. Oktian, Y.E., Witanto, E.N., Lee, S.G.: A conceptual architecture in decentralizing computing, storage, and networking aspect of iot infrastructure. IoT **2**, 205–221 (2021). https://doi.org/10.3390/iot2020011
29. Reports, V.: Industrial internet of things (iiot) market is projected to reach usd 102460 million by 2028 at a cagr of 5.3% - valuates reports (2022). https://www.prnewswire.com/in/news-releases/industrial-internet-of-things-iiot-market-is-projected-to-reach-usd-102460-million-by-2028-at-a-cagr-of-5-3-valuates-reports-840749744.html
30. Shankar, S., Parameswaran, A.G.: Towards Observability for Production Machine Learning Pipelines (2021)
31. Sharma, B.: Architecting Data Lakes: Data Management Architectures for Advanced Business Use Cases Ben (2018)
32. Wilkinson, M.D.: Comment: The fair guiding principles for scientific data management and stewardship (2016). https://doi.org/10.1038/sdata.2016.18, http://figshare.com
33. Xu, M., David, J.M., Kim, S.H.: The fourth industrial revolution: opportunities and challenges. Int. J. Financ. Res. **9** (2018). https://doi.org/10.5430/ijfr.v9n2p90, http://ijfr.sciedupress.com, https://doi.org/10.5430/ijfr.v9n2p90
34. Yuhanna, N.: Big data fabric drives innovation and growth—forrester (2016). https://www.forrester.com/report/Big-Data-Fabric-Drives-Innovation-And-Growth/RES129473

35. Yuhanna, N., Szekely, B.: Ty—forrester surfacing insights in a data fabric with knowledge graph (2021)
36. Zhang, L., Jeong, D., Lee, S., Al-Masri, E., Chen, C.H., Souri, A., Kotevska, O.: Data quality management in the internet of things. Sensors **21**, 5834 (2021). https://doi.org/10.3390/S21175834, https://mdpi.com/1424-8220/21/17/5834/htm
37. Zicari, R.V.: Big data: challenges and opportunities (2014). http://odbms.org/wp-content/uploads/2013/07/Big-Data.Zicari.pdf



Machine Learning-Based Social Media News Popularity Prediction

Rafsun Jani[1], Md. Shariful Islam Shanto[1], Badhan Chandra Das[2], and Khan Md. Hasib[1(✉)]

[1] Bangladesh University of Business and Technology, Dhaka, Bangladesh
khanmdhasib@bubt.edu.bd
[2] Florida International University, Miami, FL, USA

Abstract. The Internet has surpassed print media such as newspapers and magazines as the primary medium for disseminating public news because of its rapid transmission and widespread availability. This has made the study of how to gauge interest in online stories a pressing concern. Mashable News, one of the most popular blogs in the world, is the main source for the dataset used in this study, which was collected at the UCI data repository. Random forest, logistic regression, gaussian naive bayes, k-means and multinomial naive bayes are the four kinds of machine learning algorithms used to forecast news popularity based on the number of times an item has been shared. Gaussian naive bayes provides the most accurate predictions, at 92%. The findings suggest that Gaussian naive bayes method improves prediction and outlier detection in unbalanced data.

Keywords: News Popularity Prediction · Machine Learning Classifiers

1 Introduction

At present, many people depend on social media to get connected with their friends, news reading, entertainment, and other's activities. Social media is becoming more and more popular every day for the casting of news because this news first arrives from people from print media or TV channels. Even getting popular on social media news has many reasons but one of the most popular reasons is the news can be read easily from a cell phone or any hand-operated device in a short time by getting connected to the internet. We know every aspect of the internet has been largely influenced by social media. Moreover, people get useful resources and information from social media. When a person reads any article on social media, he may see the comments made by other users,

Supported by organization x.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 714–725, 2023.
https://doi.org/10.1007/978-3-031-27409-1_65

and since these comments are made by different individuals, no other organization or individual has any power over them. Therefore, users can make decisions about whether this news is fake or not which makes the biggest difference from other newscast mediums. Those news articles are been considered popular and propagated to many users. Earlier big agencies or large broadcasting houses had dominated though which is decreasing nowadays. people are not only dependent on particular sources of news e.g. TV channels or newspapers it's more open nowadays and a good headline or title connects more people (sometimes many sources have subscriptions). And so as people get many things in a single platform so they more rely on it. The news reaches the users and their participation by reading, commenting, and sharing had created value. Its direct feedback and readers' acceptance are always important. As it's been called the nerve of society. It's important for the news to reach the reader for proper acceptability and to make the news acceptable and important to the reader. Many times when important news reaches the reader it does not get much acceptance just because of the lack of proper title and headlines. Therefore, proper titles and headlines will play an effective role in the acceptability of this paper to the reader. Considering this fact, in this paper we studied how a good title can an important role to spread the reach of that particular news by employing machine learning algorithms. These are the main contributions made by this study.

- We focused on mainly titles and headlines for news reach on social media.
- A unique approach to predicting the outcome of a news reach on social media is proposed by introducing a well-known machine learning algorithm that gives good predicting accuracy.
- We perform an extensive experiment on a UCI machine learning repository data set containing 100000 news posts based on title and headlines. Then we apply our proposed framework.

The paper is organized as follows. The related works are discussed in Sect. 2. The architecture of the Proposed model is described in Sect. 3. Section 4 shows the experiment and the result analysis. Finally, we draw a conclusion with the discussion of the paper in Sect. 5.

2 Related Works

In recent years, the popularity of social media news has emerged as one of the most talked-about subjects among many eminent scholars worldwide. This is because news gets popular with its reader and the reader starts reading it by being attracted to the title or headline of the news. So, we worked out the purpose of popularizing news to a reader, based on the news title and headline. This study is one of those acknowledged works concerning the key forecast of the growing popularity of developed countries [1].

2.1 Predicting Popularity on Social Media News

Namous et al. [2] applied many machine learning algorithms to make popularity predictions. were used namely: RandomForest, Support vector machine, NB, and the most typical mining technique employed for classification. They used 39000 articles from the mashable website with a large and recently collected data set. they got the best model prediction from RandomForest and the neural network and achieved a 65% accuracy with optimized parameters.

Liu et al. [3] Analyzing internet news sparks widespread academic attention for predicting news popularity. They use a china website named free but for the data source during 2012-16 and 7850 news articles are eventually acquired. They suggest five characteristics that forecast popularity. and finally, predict news popularity in two aspects one is whether the news will be popular and another one is how many vicious the news ultimately attracts.

Deshpande et al. [4] is focused on news popularity prediction on performed improvement based. They are taking some criteria to judge the popularity of news such as the number of comments, number of shares, and the number of likes. This research is taken a data set of 39,797 news articles collected from the UCI repository. They thought like, comments, and sharing are more important for getting popularity and which data is a collection from an online news website. After all, they used three different learning algorithms, and It turns shown that adaptive boosting is the best model predictor for other algorithms, with 69% accuracy and 73% F-measure.

Hensinger et al. [5] focused on the textual information. Only terms that could be discovered in article titles and descriptions were employed in their experiments. Their suggested model conducts a pairwise comparison with a maximum of 85.74% for words paired with subject tags as features and 75.05% for words as a bag of words.

Wicaksono et al. [6] showing how to increase accuracy in measuring the popularity of online news. They used 61 attributes and 39,797 instances of online news data set downloaded from UCI machine learning site. This paper used to predict online news popularity few machine learning methods such as Random Forest, Support vector machine (SVM), and performance increase used by grid search and Genetic algorithm. After all, They got time measurement in seconds.

Fernandes et al. [7] is showing how to grow interested in predicting online news popularity. They proposed an intelligent decision support system. During 2 years period, they collected 39,000 articles from a widely used news source. This paper performed windows evaluation and user testing of five states of-the-art under a distinct matrix. Random Forest produced the best outcome overall. After all, One of the most crucial tasks is to assess the significance of the Random Forest Inputs and expose the keywords-based characteristics.

Rathord et al. [8] discusses various algorithms used in the process of popularity prediction of news articles. and the best result got from the Random Forest classification algorithm. They have predicted the popularity based on the number of shares, Likes, and used 39,644 articles in this paper with 59 attributes.

This paper used some popular algorithms which gave the best performance but from all over algorithms give accurate predictions by Random Forest.

2.2 Predicting Fake News on Social Media

Kesarwani et al. [9] They are presented with a specific frame to predict fake news on social media used by data mining algorithms (KNN). and Used a Total number of 2,282 posts where 1669 posts are "true" and 264 posts are "No factures" for the data set. and the dataset after pre-processes divided two-part and used only the k-NN algorithm. As a result, the algorithm achieved an approximate 79% classification accuracy on the test set.

All these works discussed the prediction of news popularity and used some well-known machine learning techniques, including NB, Random Forest, KNN, and SVM, with 65–73% accuracy. In several articles, neural networks and boosting methods were utilized, although the accuracy was similar. They have been concentrating on based on how popular news is on social media aspects like Likes, Comments, and Shares. However, our attention is on news reach on social media, which is based on news titles and headlines. Because, in our opinion, readers first select news titles or headlines before reading the news; otherwise, they skip it. And our paper is much more significant for the internet news portal organization because they can foresee the ideal title or headline for the news. As a result, readers of social media news and social media users alike become interested in the news. So, that is what distinguishes our work from that of others.

3 Proposed System

The proposed system gets started with the collection of datasets from the UCI machine learning repository. As shown in Fig. 1 some pre-processing tasks had been performed on the collected data to convert them into sequences. Our data labeling is completed as soon as we perform data pre-processing. Then we have included three new output columns, and each one shows a high, moderate, or low reach based on the data it contains. And for prediction, we later use the newly created output columns. Next, we select the appropriate features for our model through feature selection. Then, some state-of-the-art sequence-to-sequence models had been trained and tested on the collected data.

3.1 Dataset Collection and Preprocessing

The model was created using the 'Multi-Source Social Feedback of online News feeds' from UCI Machine Learning Repository. The data was gathered from two reputable news sources (Google News and Yahoo! News) as well as three social networking sites (Linkedin, Facebook, and Google+). There are four topics in it: Palestine, Economy, Microsoft, and Obama [10]. Following that, the data received reveals the number of shares related to each distinct URL, which is

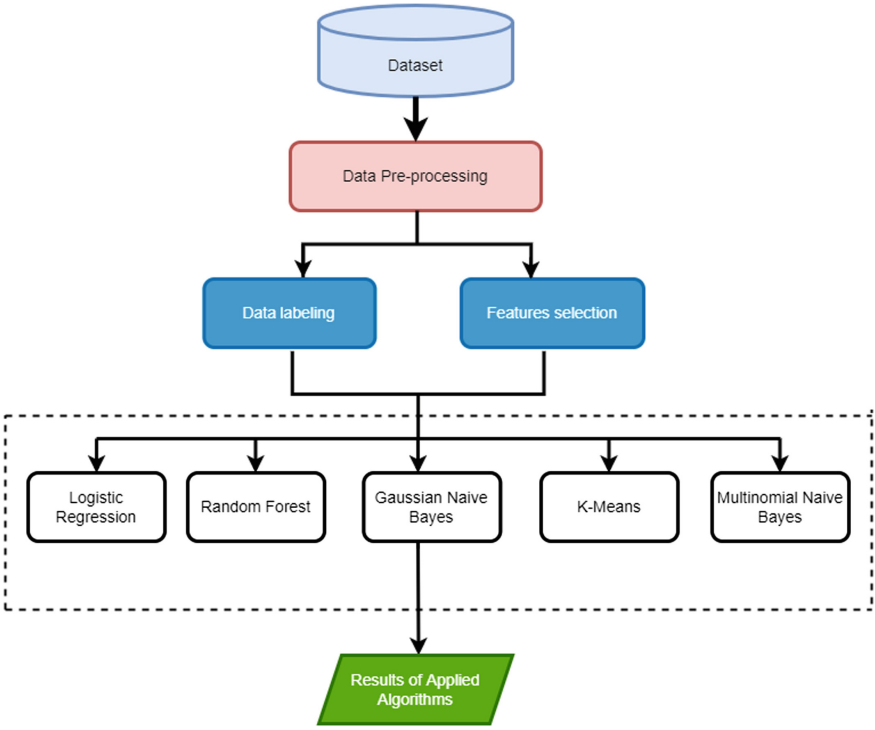


Fig. 1. Overview of proposed methodology for news reach prediction

employed as a popularity metric. The final value for Social Media's popularity contains the sharing value news item for 72 h from publishing time. 11 characteristics that are displayed explain each news article in Table 1.

1. After collecting the data, we perform some pre-processing tasks e.g. remove the invalid, duplicate, and null values.
2. Verifying and selecting data type for each attribute.
3. Compromised the attributes which do not make sense.
4. Based on popularity, we categorized the news items as high, moderate, and low for each social media's popularity feedback.

After pre-processing more than 80% of instances (around Eighty thousand) remain valid.

3.2 Applied Frameworks

In the second phase of our proposed system, To forecast, we use numerous machine learning methods to measure the performance of each social media. First, we describe the concepts of Random Forest (RF), Logistic Regression (LR),

Table 1. Name, type of data, and description of data variables in the news data file

Variable	Type	Description
IDLink	*Numeric*	The identifier for each news article specifically
Title	*String*	Title of the news story as stated by the authorized media source
Headline	*String*	According to the authorized media sources, the news item's headline
Source	*String*	Publishes the news item's original source
Topic	*String*	To find the content in the official media sources, use the following search term
Publish Date	*Timestampe*	The news items' publishing date and time
Sentiment Title	*Numeric*	The score of the news item's title's sentiment
Sentiment Headline	*Numeric*	The score for the article's headline's sentiment
Facebook	*Numeric*	Final score of the news item's popularity based on
GooglePlus	*Numeric*	Final score of the news item's popularity based on Google+
LinkedIn	*Numeric*	According to social media outlet LinkedIn, the news item's final popularity score

Gaussian Naive Bayes (GNB), K-means, and Multinomial Naive Bayes (MNB). Then we configure our models and apply those models to the pre-processed data.

Random Forest: The random forest algorithm has seen great success as a general-purpose regression and classification technique. The method has demonstrated good performance in situations when the number of variables is significantly more than the number of observations. It combines numerous randomized decision trees and averages out their predictions [11]. The algorithm is efficient in handling missing values however, it can be overfitted. In order to create the model more quickly or with greater predictive potential, RandomForest uses the hyper-parameter [12]. Large datasets may be processed fast using the computationally effective Random Forest approach [13].

Logistic Regression: For categorical outcomes, which are often binary, logistic regression models are used to analyze the impact of predictor variables[19]. A multiple or multivariable logistic regression model is used when there are several factors [14].

Binary logistic regression: In the case of categorizing an object into two potential outcomes, binary logistic regression was previously described. It is an either/or solution when this idea is normally expressed as a 0 or a 1.

Multinomial logistic regression: A model known as multinomial logistic regression allows for the classification of items into many classes. Before the model runs, a group of three or more preset classes is set up.

Ordinal logistic regression: It is necessary to rank the classes in the ordinal logistic regression model when there are several categories into which an object might be categorized. The ratio of classes is not required. The separation between each class might differ.

Gaussian Naive Bayes: Gaussian naive Bayes allows features with continuous values and models them all as following a Gaussian distribution. Thus, Gaussian naive Bayes has a slightly different approach and can be efficient. Its classifier works well and can be used for a variety of classification problems. Since we use classification data for news reach prediction on social media and it's given so much better results than others classifier algorithms. So, given a training dataset of N input variable X with corresponding target variable t (Gaussian) naive Bayes assume that the (class conditional dataset) is normally distributed.

$$P\left(X\|Z=C,\mu_c,\sum{}_c\right)=N\left(X\|\mu_c,\sum{}_c\right) \tag{1}$$

Where $\sum$ is the class-specific con- variance matrix and μ is the class-specific mean vector. This method is quite helpful for categorizing huge datasets. The method makes the assumption that each characteristic in the classification process operates independently of the others [15]. The algorithm's effectiveness in categorization is due to its computations having a low error rate.

K-means: Unsupervised machine learning methods that are widely used include K-means clustering. Unsupervised algorithms often construct an interface from the dataset using just input vectors and excluding references to previously known labeled results. The cluster sum of squares is reduced to assign each data point to its corresponding cluster. Every data point is then assigned to the closest cluster using the k-mean procedure, which first determines the K number of centroids. K-means clustering to extract and analyze the properties of news content [16].

Multinomial Naive Bayes: A common bayesian learning technique in natural language processing is the multinomial Naive Bayes algorithm [17]. The method, which predicts the tag of a text such as an email or newspaper article, is based on the Bayes theorem. For a given sample, it determines the probabilities of each item and then outputs the article with the highest probability. It was based on the subsequent formula

$$P\left(A\|B\right)=P\left(A\right)*P\left(B\|A\right)/P\left(B\right) \tag{2}$$

A news item in a newspaper may express a variety of emotions or have the predisposition to be positive or negative, therefore the article's content can be actively utilized to assess the reader's reaction [18].

3.3 Models' Configuration

In this paper, for all the models described above, we have applied five popular machine learning algorithms, such as Random Forest, Logistic Regression, Gaussian naive Bayes, Multinomial naive Bayes, and K-means. K-means is an unsupervised learning technique, as we are aware. We used four different cluster types in this case, with a maximum number of 10 and a minimum number of 2.

however, given that we got good performance when we used Cluster 10. These Algorithms perform well good predictions for any classification problem. But we got a good performance from Gaussian naive Bayes and Logistic Regression. In this prediction, we used to evaluate 80% data for training and 20% data for testing. Gaussian naive Bayes performs evaluation better than Logistic Regression. Gaussian naive Bayes accurately classify datasets and properly evaluate them.

4 Experiment and Result Analysis

On the UCI machine learning repository dataset [10], multiple categorization algorithms were compared in an experiment. We used five well good machine-learning algorithms in this experiment. But we got a good performance from Gaussian naive Bayes and Logistic Regression.

4.1 Experiment Setup

For data pre-processing Such as, removing *–1*, *null* value clear, remove *duplicate* value, the python programming language is being used, visualization of each comparable part of data, experiment and evaluations of the algorithms. Used UCI machine learning dataset. And five machine learning algorithms are implemented.

4.2 Features Selection

Features Selection is carrying importance for improving prediction results. Since it is a method of predicting Social media news reach, that's why effective feature selection is so much important. In this paper, we used two feature selection methods one is *selectKBest* and another one is *linear Regression*. We have used it for different types of output like *Facebook*, *Linkedin*,*Google plus*. For each type of output features selection selected some common features Fig. 2. which improved our model prediction rate.

4.3 Experiment Result

Several algorithms were used in our experiment, and the results are presented in Fig. 3. Precision, Recall, and F-measure is used as evaluation methods. These metrics were determined using the confusion matrix presented in Table 2. The evaluation measurements' formula is shown in Eqs. 3, 4, and 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

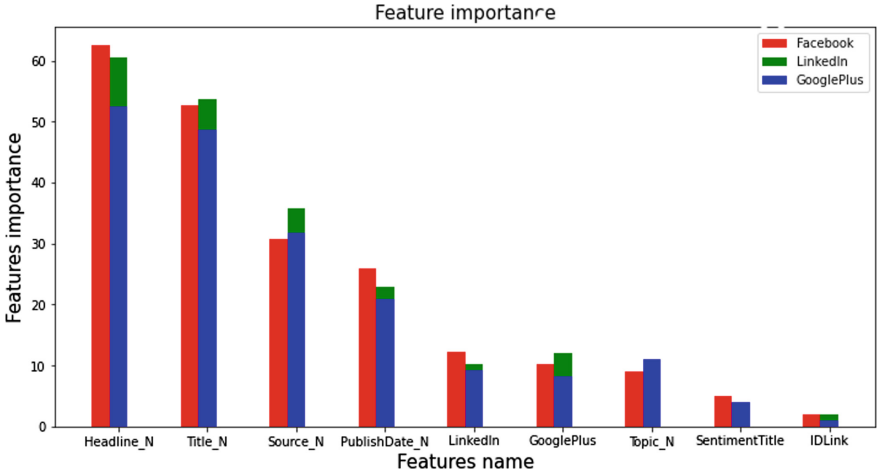


Fig. 2. Selection of important features

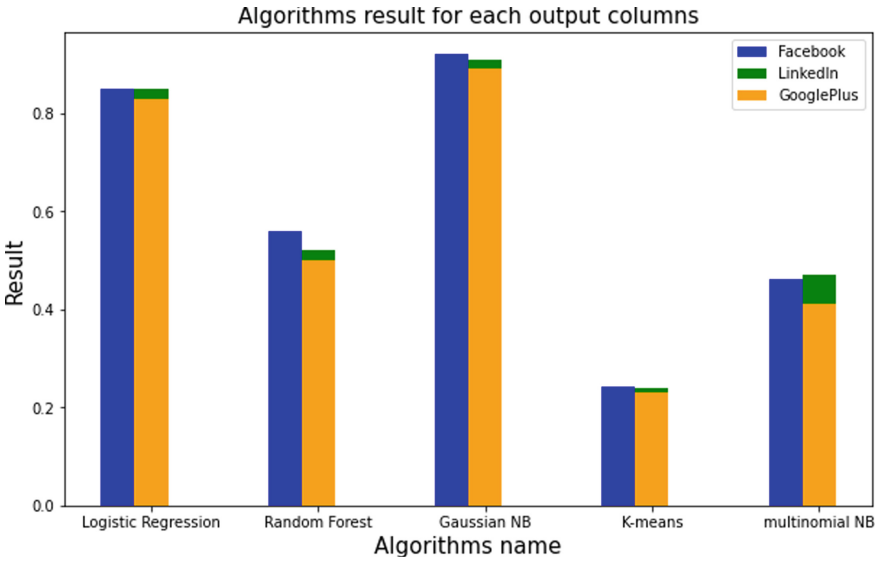


Fig. 3. Prediction results showing for different social media platforms

Table 2. Confusion matrix

2*		Predicted Class	
		P' (Positive)	N' (Negative)
2* Actual Class	P (Positive)	True Positive (TP)	False Negative (FN)
	N (Negative)	False Positive (FP)	True Negative (TN)

We represented our experiment results, precision, recall, and F-measures displayed in Table 3

Table 3. Experiment result

	Accuracy	Precision	Recall	F-measure	Output types
Random forest	0.56%	0.55	0.62	0.58	**Facebook**
Logistic regression	0.85%	0.91	0.96	0.94	
Gaussian Naive Bayes	**0.92%**	**0.99**	**0.88**	**0.93**	
Multinomial Naive Bayes	0.46%	0.51	0.65	0.57	
K-means	0.241	0.62	0.54	0.59	
Random forest	0.52%	0.52	0.59	0.56	**LinkedIn**
Logistic regression	0.85%	0.91	0.97	0.93	
Gaussian Naive Bayes	**0.91%**	**0.98**	**0.90**	**0.91**	
Multinomial Naive Bayes	0.47%	0.50	0.61	0.59	
K-means	0.240	0.61	0.55	0.67	
Random forest	0.50%	0.53	0.61	0.61	**GooglePlus**
Logistic regression	0.83%	0.89	0.95	0.91	
Gaussian Naive Bayes	**0.89%**	**0.97**	**0,98**	**0.92**	
Multinomial Naive Bayes	0.42%	0.55	0.59	0.61	
K-means	0.229	0.58	0.54	0.69	

5 Conclusion and Discussion

News has been counted as popular if it gets popular on Social Media. In this investigation, we used a UCI online news popularity dataset to do exploratory data analysis and machine learning prediction. The number of shares was changed into a popularity/unpopularity classification issue by data preprocessing techniques including normalization and principal component analysis, significantly bettering the quality of the Dataset. The headline and title are the key factors to a user being willing to read an article. To predict popularity we used Random Forest, Logistic Regression, Gaussian naive Bayes, and K-mean. Among all the algorithms, Gaussian naive Bayes reached the highest accuracy with 92%. Outcome yielded from our proposed method in Fig. 3 and Table 3. implies that we may better accurately forecast news reach by minimizing biases in social media data. We analyzed our outcomes and found that the output we got before data leveling gave much better output after data leveling. And data pre-processing played a vital role in the approached model. The dataset contains limited categories of news, so we hope to work on every possible category in the future by establishing a method, the forecast before publishing may be integrated with the user reactions (React, comment, share) after publication to predict more precisely.

References

1. Wu, B., Shen, H.: Analyzing and predicting news popularity on twitter. Int. J. Inf. Manag. **35**(6), 702–711 (2015). https://doi.org/10.1016/j.ijinfomgt.2015.07.003
2. Namous, F., Rodan, A., Javed, Y.: Online news popularity prediction. In: 2018 Fifth HCT Information Technology Trends (ITT), pp. 180–184 (2018). https://doi.org/10.1109/CTIT.2018.8649529
3. Liu, C., Wang, W., Zhang, Y., Dong, Y., He, F., Wu, C.: Predicting the popularity of online news based on multivariate analysis. In: 2017 IEEE International Conference on Computer and Information Technology (CIT), pp. 9–15 (2017). https://doi.org/10.1109/CIT.2017.36
4. Deshpande, D.: Prediction evaluation of online news popularity using machine intelligence. In: 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1–6 (2017)
5. Hensinger, E., Flaounas, I., Cristianini, N.: Modelling and predicting news popularity. Pattern Analysis and Applications **16**(4), 623–635 (2013)
6. Wicaksono, A.S., Supianto, A.A.: Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction. Int. J. Adv. Comput. Sci. Appl. **9**(12) (2018)
7. Fernandes, K., Vinagre, P., Cortez, P.: A proactive intelligent decision sup- port system for predicting the popularity of online news. In: Portuguese Conference on Artificial Intelligence, pp. 535–546. Springer, Berlin (2015)
8. Rathord, P., Jain, A., Agrawal, C.: A comprehensive review on online news popularity prediction using machine learning approach. Trees **10**(20), 50 (2019)
9. Kesarwani, A., Chauhan, S.S., Nair, A.R.: Fake news detection on social media using k-nearest neighbor classifier. In: 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 1–4 (2020). IEEE
10. Moniz, N., Torgo, L.: Multi-source social feedback of online news feeds (2018). arXiv:1801.07055
11. Biau, G., Scornet, E.: A random forest guided tour. Test **25**(2), 197–227 (2016). https://doi.org/10.1007/s11749-016-0481-7
12. Oshiro, T.M., Perez, P.S., Baranauskas, J.A.: How many trees in a random forest? In: International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 154–168. Springer, Berlin (2012)
13. Hasib, K.M., Towhid, N.A., Alam, M.G.R.: Online review based sentiment classification on bangladesh airline service using supervised learning. In: 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1–6 (2021)
14. Hasib, K.M., Rahman, F., Hasnat, R., Alam, M.G.R.: A machine learning and explainable ai approach for predicting secondary school student performance. In: 2022 IEEE 12th Annual Computing and Communica- tion Workshop and Conference (CCWC), pp. 0399–0405 (2022)
15. C. Olakoglu, N., Akkaya, B.: Comparison of multi-class classification algorithms on early diagnosis of heart diseases. In: y-BIS 2019 Conference Book: Recent Advances N Data Sc Ence and Bus Ness Analyt Cs, p. 162 (2019)
16. Liu, J., Song, J., Li, C., Zhu, X., Deng, R.: A hybrid news recommendation algorithm based on k-means clustering and collaborative filtering. J. Phys.: Conf. Ser. **1881**, 032050 (2021). IOP Publishing
17. Jahan, S., Islam, M.R., Hasib, K.M., Naseem, U., Islam, M.S.: Active learning with an adaptive classifier for inaccessible big data analysis. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2021)

18. Singh, G., Kumar, B., Gaur, L., Tyagi, A.: Comparison between multinomial and bernoulli naïve bayes for text classification. In: 2019 International Conference on Automation, Computational and Technology Management (ICACTM), pp. 593–596 (2019). https://doi.org/10.1109/ICACTM.2019.8776800
19. Hasib, K.M., Tanzim, A., Shin, J., Faruk, K.O., Mahmud, J.A., Mridha, M.F.: BMNet-5: a novel approach of neural network to classify the genre of bengali music based on audio features. IEEE Access **10**, 108545–108563 (2022). https://doi.org/10.1109/ACCESS.2022.3213818



Hand Gesture Control of Video Player

R. G. Sangeetha, C. Hemanth(✉), Karthika S. Nair, Akhil R. Nair, and K. Nithin Shine

School of Electronics Engineering, Vellore Institute of Technology, Chennai, India
{Sangeetha.rg,hemanth.c}@vit.ac.in, {karthikanair.s2020,
akhilnair.r2020,nithinshine.k2020}@vitstudent.ac.in

Abstract. The rise of ubiquitous computing has expanded the role of the computer in our daily lives. Though computers have been with us for several decades, still we follow the same, old, primitive methods such as a mouse, keyboard, etc. to interact with them. In addition, a variety of health issues are brought on by a person's continual computer use. In the study of language, hand gestures are a crucial part of body language. The usage of a hand-held device makes human-computer interaction simple. The proposed work aims to create a gesture-controlled media player wherein we can use our hands and control the video played on the computer.

Keywords: Gesture · Video

1 Introduction

Everyone relies on computers to complete the majority of their tasks. keyboard and mouse are the two main input methods. but the continual and continuous use of computers has led to a wide range of health issues that are affecting many people. a desirable way of user-computer interaction is the direct use of the hands as an input device [1]. since hand gestures are a fully natural way to communicate, they do not negatively impact the operator's health the way that excessive keyboard and mouse use does [2, 3].

This research implements a gesture-based recognition technique for handling multimedia applications. in this system, a gesture recognition scheme is been proposed as an interface between humans and machines. here, we make a simple arduino-based hand gesture control using ultrasonic sensors and photo-sensors which automatically increase or decrease the screen brightness (according to the room brightness), play/pause a video, increase or decrease the volume, go to the next video, etc. in a video player with the help of hand gestures.

Three Ultrasonic sensors and an ldr is used for this work. the sensors and arduino board can be fixed on the computer, and the movement of hands towards and away from the screen can be detected by the three ultrasonic sensors and hence used to control the video player. the program code is written in arduino programming language. python language is also used to interface between arduino and the video player.

The Ultrasonic sensors will be fixed on the top of the laptop or computer and the arduino board will be behind and connected to the laptop or computer with the help of a usb cable. the hand gestures are linked to the vlc media player using the short-cut keys

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 726–735, 2023.
https://doi.org/10.1007/978-3-031-27409-1_66

using in the keyboard. for example, we use a space bar to play/pause a video and an up and down arrow to increase or decrease volume. this is linked with the help of python language.

The Ultrasonic sensor will detect the gestures by calculating distance with the help of travel time and speed of sound and this is calculated in the arduino code [4]. the ldr detects the room brightness and produces adaptive brightness in the vlc player. the room brightness and the screen brightness of the laptop are directly proportional to each other. this hardware implementation can be used in any laptop or pc but this is restricted to vlc media players only.

1.1 Need for Hand Gesture Control

This work aims to control the key pressings of a keyboard by using hand gestures. This technique can be very helpful for physically challenged people because they can define the gesture according to their needs. Even if our keyboard is disabled or has any issues, this can be of great help. Interaction is simple, practical, and requires no additional equipment when gestures are used. It is possible to combine auditory and visual recognition [5, 6]. But in a noisy setting, audio commands might not function [8].

2 Design/Implementation

Gesture-controlled laptops are becoming increasingly well-known recently. By waving our hands in front of our computer or laptop, we can control several features using a method known as leap motion. In this work, we build a Gesture control VLC media player using Ultrasonic sensors by combining the Power of Arduino and Python. An adaptive brightness feature is also added.

2.1 Design Approach

VLC Media Player is a free and open-source, cross-platform multimedia player. VLC Media player shortcuts are great for saving time. Several common actions can be performed without even moving the mouse or clicking on the menu buttons. The hotkeys are great for quick video playback actions. In this work three Ultrasonic sensors are fixed above the laptop screen, one in the middle and the other two at each end.

2.2 Economic Feasibility

The total development cost for this implementation is less than 500 Rs., which is quite low when considering its advantages. It does not require any additional operation cost and can easily be fixed on a normal existing computer. The ultrasonic sensor is vulnerable to environmental conditions such as dust, moisture, and aging of the diaphragm and hence has a short lifetime.

2.3 Technical Feasibility

This implementation can work on a normal Windows computer. A minimum of 400MB of disk space is required to install Python IDLE and Arduino IDE along with the VLC media player. It requires very less processing power. The main disadvantage is that; the applications that are running in the background-Python and the Arduino along with the sensors will consume a lot of power.

2.4 Operational Feasibility

This setup is quite simple and can be easily fixed on the monitor. The ultrasonic sensors can detect hands at a distance of 5 to 40 cm away from them. The 16MHz speed of the microcontroller provides a very quick response time and the gestures are quite simple.

3 System Specifications

3.1 Hardware Specifications

The Arduino microcontroller board has sets of digital and analog input/output (i/o) pins that can connect to different expansion boards (called "shields"), breadboards (used for prototyping), and other circuits and sensors. this work makes use of an arduino uno R3.

An ultrasonic sensor called the HC-SR04 that is used in this research has a transmitter and a receiver. To determine the distance from an item, this sensor is employed. Here, the distance between the sensor and an item is determined by the time it takes for waves to transmit and receive. This sensor makes use of non-contact technologies and sound waves. This sensor allows the target's required distance to be determined accurately and without causing any damage.

A 5mm LDR is used to detect ambient brightness. A light-dependent resistor, commonly referred to as a photoresistor or LDR, is a component whose resistance depends on the electromagnetic radiation that strikes it. When light strikes them, their resistance is reduced, and in the dark, it is increased. When a constant voltage is provided and the light intensity is raised, the current begins to increase.

3.2 Software Specifications

1. Arduino IDE (1.8.13)–The Arduino Integrated Development Environment (IDE) is a cross-platform application (for Windows, macOS, Linux) that is written in functions from C and C + +. It is used to write and upload programs to Arduino-compatible boards.
2. Python DLE (3.9.5)–IDLE (Integrated Development and Learning Environment) is an integrated development environment (IDE) for Python.
3. PIP (21.1.1)–pip is a package-management system written in Python used to install and manage software packages. It connects to an online repository of public packages, called the Python Package Index.
4. PyAutoGui library–Used to programmatically control the mouse & keyboard. Installed using PIP.

5. Serial library–Used to interface Python with the Serial monitor of Arduino.
6. Screen_brightness_control–A Python tool for controlling the brightness of your monitor programmatically.
7. Time-IT's a Python library to uses delay functions. programmatically.

4 Results and Discussions

1. **Gesture to play or pause the video**

Use the left and right sensors to perform this gesture. First, the left and the right sensor detect an obstacle in front of it. When we keep both our hands in front of the left and right sensors, the video will be paused. Similarly, the video will be played if the same action is repeated as shown in Fig. 1.

When the distance between the hand and the left and right sensor is greater than 10 cm or lesser than 40 cm, it prints "Play/Pause" in the serial monitor, and the python code will receive this command and mimic the keyboard key pressing of "Space bar" and so the video will be paused. The same procedure will be repeated to play the video.

When we take the output for the left and right sensor, it prints "Play/Pause" when we keep both our hands in front of the left and right sensors thus indicating that both the sensors have been detected and the video has been either played or paused, provided the distance between the hand and both the sensors are greater than 10 cm and lesser than 40 cm.

2. **Gesture to take a snapshot of the video**

We use the left and center sensor to perform this gesture. First, the left and centre sensor detect an obstacle in front of it. When we keep both our hands in front of the left and center sensor, a snapshot of the video will be taken. When the distance between the hand and the left and center sensor is greater than 10 cm or lesser than 40 cm, it prints "Snap" in the serial monitor, and the python code will receive this command and mimic the keyboard key pressing of "Shift + S" and so the snapshot of the video will be taken as shown in Fig. 2.

When we take the output for the left and center sensor, it prints "Snap" when we keep both our hands in front of the left and center sensors thus indicating that both the sensors have been detected and the snapshot of the video has been taken, provided the distance between the hand and both the sensors are greater than 10 cm and lesser than 40 cm.

3. **Gesture to full screen the video**

We use the right and center sensors to perform this gesture. First, the right and center sensors detect an obstacle in front of it. When we keep both our hands in front of the left and center sensor, the video will change to full-screen mode. The same action should be repeated to exit the full-screen mode as shown in Fig. 3.



Fig. 1. Gesture for play and pause the video

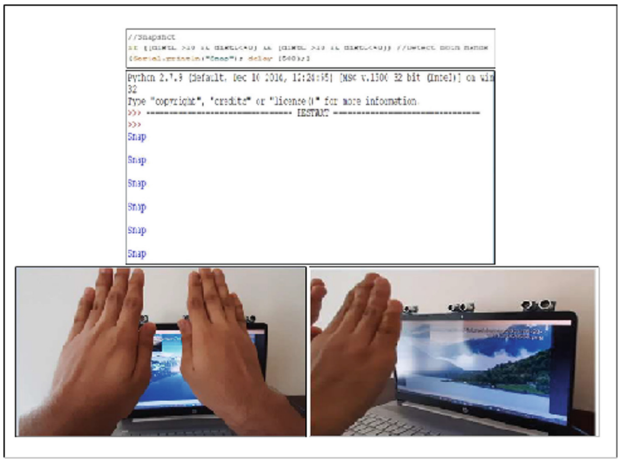


Fig. 2. Gesture to take snapshot

When the distance between the hand and the right and center sensor is greater than 10 cm or lesser than 40 cm, it prints “Fscreen” in the serial monitor, and the python code will receive this command and mimic the keyboard key pressing of “f” and so the video will play in the full-screen mode. The same procedure will be repeated to exit the full-screen mode.

When we take the output for the right and center sensor, it prints “Fscreen” when we keep both our hands in front of the right and center sensors thus indicating that both the sensors have been detected and the video is in full-screen mode, provided the distance between the hand and both the sensors are greater than 10 cm and lesser than 40 cm.

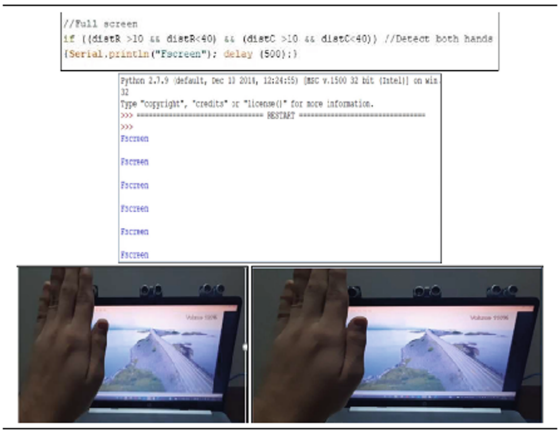


Fig. 3. Gesture to maximize the screen

4. **Gesture to increase and decrease the volume**

We used the left sensor to perform this gesture. First, the left sensor detects an obstacle in front of it. When we move our hand toward the left sensor, the volume of the video will increase. Likewise, the volume will get decreased when we slowly take our hand away from this sensor as shown in Fig. 4.

When the distance between the hand and the left sensor is greater than or equal to 5 cm and less than or equal to 40 cm, it first waits for 100 milli seconds for hand hold time. Using the calculate_distance() function, it first finds the distance between our hand and the left sensor. If it is greater than or equal to 5 cm and less than or equal to 40 cm, it prints "Left Locked" in the serial monitor. Then a loop runs as long as the distance is less than or equal to 40 cm. First, it calculates the distance between the left sensor and our hand. If the distance is less than 10 cm, it prints "Vup" in the serial monitor and the python code will receive this command and mimic the keyboard key pressing of "ctrl + up" and so the volume of the video is increased. Then it waits for 300 ms and the gesture will be performed again depending on our hand motion. Likewise, if the distance is more than 20 cm, it prints "Vdown" and the python code will receive this command and mimic the keyboard key pressing of "ctrl + down" and so the volume of the video decreases. Then again it waits for 300 ms.

5. **Gesture to change the aspect ratio of the display**

We used the center sensor to perform this gesture. First, the center sensor detects an obstacle in front of it. When we move our hand toward the center sensor, the aspect ratio of the display will change as shown in Fig. 5.

When the distance between the hand and the center sensor is greater than or equal to 5 cm and less than or equal to 40 cm, it first waits for 100 milli seconds for hand hold time. Using the calculate_distance() function, it first finds the distance between our

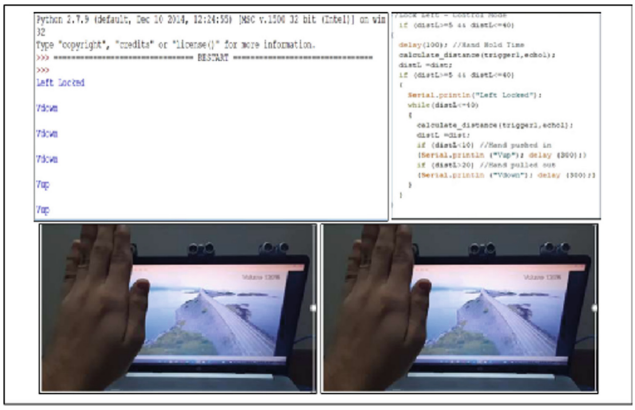


Fig. 4. Gestures for volume control

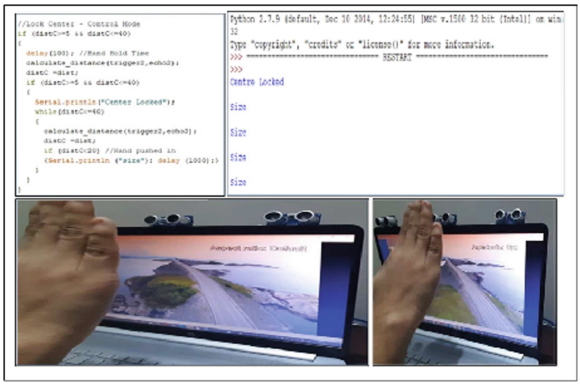


Fig. 5. Gesture to change the aspect ratio

hand and the center sensor. If it is greater than or equal to 5 cm and less than or equal to 40 cm, it prints “Center Locked” in the serial monitor. Then a loop runs as long as the distance between the hand and the center sensor is less than or equal to 40 cm. It.

calculates the distance between our hand and the sensor. If the distance is less than 20 cm, it prints “size” in the serial monitor, and the python code will receive this command and mimic the keyboard key pressing of “a” so the aspect ratio of the display changes each time. Then it waits for 1000 ms and will continue again.

6. **Gesture to rewind or forward the video**

We used the right sensor to perform this gesture. First, the right sensor detects an obstacle in front of it. When we move our hand toward the right sensor, the video will rewind. Likewise, when we gradually take our hands away from the right sensor, the video will be forwarded as shown in Fig. 6.

When the distance between the hand and the right sensor is greater than or equal to 5 cm and less than or equal to 40 cm, it first waits for 100 milli seconds for hand hold time. Using the calculate_distance() function, it first finds the distance between our hand and the right sensor. If it is greater than or equal to 5 cm and less than or equal to 40 cm, it prints "Right Locked" in the serial monitor. Then a loop runs as long as the distance is less than or equal to 40 cm. First, it calculates the distance between the right sensor and our hand. If the distance is less than 20 cm, it prints "Rewind" in the serial monitor, and the python code will receive this command and mimic the keyboard key pressing of "ctrl + left" and so the video rewinds. Then it waits for 300 ms and will continue again. Likewise, if the distance is more than 20 cm, it prints "Forward" and the python code will receive this command and mimic the keyboard key pressing of "ctrl + right" and so the video forwards. Then it waits for 300 ms.

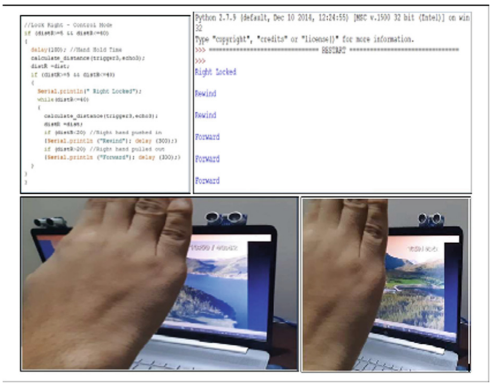


Fig. 6. Gesture to forward or rewind the video

7. **Adaptive Brightness feature**

An LDR is used to perform this feature. The Voltage drop across the LDR is inversely proportional to the ambient light intensity. The voltage drop at high light intensity is less than 0.3V which increases when the intensity decreases and reaches up to 4.9V. This change is converted into a percentage and sets the screen brightness of the monitor accordingly. A minimum threshold brightness of 20% is kept so that it does not turn completely dark even in extremely low ambient brightness. The screen brightness increases by 1% for a voltage drop of 0.05V. The corresponding output voltage and light intensity at normal room brightness and when a light source is brought near the LDR are shown in the figure below. The output voltage (received by the A0 pin) ranges from 0-5V and is displayed on a scale of 0–1024 in the output terminal of Python IDLE. This can be seen in Fig. 7.

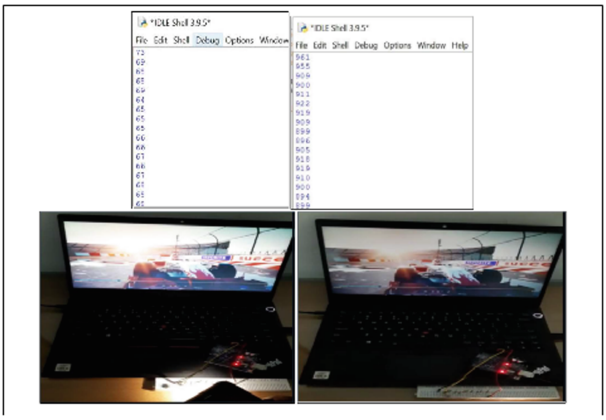


Fig. 7. Gesture to control the adaptive brightness

5 Conclusion

For controlling the VLC player's features, the program defines a few gestures. Depending on the desired function, the user will make a gesture as input. Since users can design the gesture for certain commands in accordance with their requirements, the program is more useful for people. The usage of hand gestures can be expanded to playing games and opening applications available on the device. This sort of interaction can make the stressful lives of people easier and more flexible. We also need to extend the system for some more types of gestures as we have implemented it for only 7 actions. However, we can use this system to control applications like PowerPoint presentations, games, media player, windows picture manager, etc.

References

1. Tateno, S., Zhu, Y., Meng, F.: Hand gesture recognition system for in-car device control based on infrared array sensor. In: 2019 58th Annual conference of the society of instrument and control engineers of japan (sice), pp. 701–706. (2019). https://doi.org/10.23919/SICE.2019.8859832
2. Jalab, H.A., Omer, H.K.: Human-computer interface using hand gesture recognition based on neural network. In: 2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW), pp. 1–6. (2015). https://doi.org/10.1109/NSITNSW.2015.7176405
3. Tsai, T.-H., Huang, C.-C., Zhang, K.-L.: Design of hand gesture recognition system for human-computer interaction. Multimedia Tools and Applications **79**(9–10), 5989–6007 (2019). https://doi.org/10.1007/s11042-019-08274-w
4. Haratiannejadi, K., Selmic, R.: Smart glove and hand gesture-based control interface for multi-rotor aerial vehicles in a multi-subject environment. IEEE Access, **8**, 227667–227677. https://doi.org/10.1109/ACCESS.2020.3045858
5. Parkale, Y.V.: Gesture-based operating system control. Second international conference on advanced computing & communication technologies **2012**, 318–323 (2012). https://doi.org/10.1109/ACCT.2012.58

6. Harshitaa, A., Hansini, P., Asha, P.: Gesture based Home appliance control system for Disabled People. In: 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1501–1505. (2021). https://doi.org/10.1109/ICESC51422.2021.9532973
7. Abdelnasser, H., Harras, K., M. Youssef, "A Ubiquitous WiFi-Based Fine-Grained Gesture Recognition System. In: IEEE Transactions on Mobile Computing, vol. 18, no. 11, pp. 2474–2487. (2019). https://doi.org/10.1109/TMC.2018.2879075



Comparative Analysis of Intrusion Detection System using ML and DL Techniques

C. K. Sunil(✉), Sujan Reddy, Shashikantha G. Kanber, V. R. Sandeep, and Nagamma Patil

Department of Information Technology, National Institute of Technology Karnataka, Surathkal 575025, India
sunilchinnahalli@gmail.com

Abstract. Intrusion detection system (IDS) protects the network from suspicious and harmful activities. It scans the network for harmful activity and any potential breaching. Even in the presence of the so many network intrusion APIs there are still problems in detecting the intrusion. These problems can be handled through the normalization of whole dataset, and ranking of feature on benchmark dataset before training the classification models. In this paper, used NSL-KDD dataset for the analysation of various features and test the efficiency of the various algorithms. For each value of k, then, trained each model separately and evaluated the feature selection approach with the algorithms. This work, make use of feature selection techniques like Information gain, SelectKBest, Pearson coefficient and Random forest. And also iterate over the number of features to pick the best values in order to train the dataset.The selected features then tested on different machine and deep learning approach. This work make use of stacked ensemble learning technique for classification. This stacked ensemble learner contains model which makes un-correlated error there by making the model more robust.

Keywords: Autoencoders · Feature Selection · Gradient Boosting · Information Gain · Machine learning · Pearson coefficient · SelectKBest

1 Introduction

In this modern age of technology, it is essential to protect networks from potential security threats since many people have high access to Internet systems. This has given rise to a lot of security concerns due to the high availability of the Internet. Systems can be attacked with malicious source code, which can be in various forms like viruses, worms, and Trojan horse; over time, it's becoming much harder to detect intrusion in systems using only techniques like firewalls and encryption.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 736–745, 2023.
https://doi.org/10.1007/978-3-031-27409-1_67

Intrusion detection system acts as network-level protection for computer networks. Intruders use weaknesses in networks, such as poor internet protocols, bugs in source code, or some network flaws, to breach security. Intruders may try to access more content than what is possible with their current rights, or hackers who try to steal sensitive and private data from the user's system. There are two types of intrusion detection systems: Signature-based and anomaly-based. Signature-based identification relies on examining network packet flows and compares them with configured signatures of previous attacks. The anomaly detection technique works by comparing given user parameters with behavior that deviates from a normal user. This paper has proposed many methods to improve the performance of Intrusion Detection Systems using Machine learning techniques. It makes use of Precision, Accuracy, Recall, and F1-Score to evaluate how a model performs.

This paper makes use of feature selection and extraction techniques like SelectKBest, Random Forest, Pearson Coefficient, and Information Gain. Once the best features are selected using above mentioned methodology, those features are tested on different machine learning classification algorithmic models.

The aim of this work is to use feature selection methods to remove insignificant features from data and then apply ML algorithms for intrusion detection.

1. Feature selection method was carried out using algorithms like selectKbest, Information gain, Pearson coefficient, and Random forest feature selection technique.
2. For classification, used different ML models like XGBoost classifier(XGB classifier), Random Forest classifier(RF classifier), Autoencoders
3. A comparison is performed using different ML model's results using Precision, Accuracy, Recall, and F1-Score with respect to each Machine learning model.
4. This work also compared the effect using k best feature with respect to accuracy for separately for specific feature selection techniques on the validation dataset.
5. In this work, design a novel ensemble model with widely varying base layer models to ensure that the models make uncorrelated errors and then compare the proposed approach with the state-of-the-art works.

2 Literature Survey

The authors [1] discuss about feature selection using various machine learning algorithms to perform comparative analysis. The authors used hybrid intrusion detection systems, which are created by stacking multiple classifiers together. The following algorithms to perform analysis k-NN, Naive Bayes, SVM, NN, DNN, and Auto-encoder were used to detect the best-suited algorithm for the prediction. The paper [2] discusses the ways of combining feature selection and machine learning techniques to perform comparative analysis most effectively.

Although the current IDS has more advantages in terms of network protection and attack prevention, with the ever-developing complex network architecture and updated attacks, most of the traditional IDS rely on rule-based pattern

matching and the classical Machine learning approach [3]. [4] considers real-time intrusion based detection system. They used a dynamic changing model, and when data is gathered, it uses the XGBoost technique to ensure maximum results are obtained. The authors [5] applied the machine learning model in real-life activity. The authors used genetic algorithms and decision trees, which are then used to automatically generate rules which are used to classify network connections.

Alzam et al. [6] use the pigeon hole-inspired optimizer for feature selection and decision tree for the classification. The drawback of this model is that they have not to bench marked their approach on other machine learning and deep learning model. Iearcatino et al. [7] make use of autoencoders for getting the compressed feature representation for the dataset later it is trained on the machine learning model for prediction. Feature selection techniques are not well utilized; a simple statistics-based approach has been used to select the feature, which is not a robust method for feature selection. Results are not compared with the ensemble technique.

In the proposed work addresses the limitations of all these papers by considering uncorrelated models in our ensemble model while also using a superior feature selection algorithm based on robust experimentation.

3 Methodology

The dataset NSL-KDD [8] consists of around 1,30,000 traffic records. These are divided into training and test datasets. The dataset had many classes, and we combined some of the classes into one single super class, as mentioned in Table 1. This is done to efficiently train the ML model since having a lot of classes can lead to poor results. We merge similar intrusion attacks into a single attack to reduce the number of classes. One more reason to merge classes is the high-class imbalance that will exist for the classes with fewer instances. This can lead to problems while training the dataset, so it is prevented by merging classes. The dataset consists of four attack-type classes and one normal-type class, which signify the type of the request.

3.1 Data Pre-processing

The given data is normalized before it is sent into the model for further training. We used a standard min-max scalar for this purpose, which will center the data around the median, as shown in Eq. 1.

Table 1. Details of Normal and Attack classes in NSL-KDD dataset

Class	Attack type	Train	Test
Normal	Normal	67343	9711
Denial-of-service	Teardrop, Back, Land, pod, Smurf NeptuneApache2, Worm, Udpstorm, processtable	45927	7458
Probe	Saint, Satan, Ipsweep, Portsweep, MSCAM, Nmap	11656	2421
Remote to user	named, Guess-passwd, Imap, phf, multihop, warexcelent, Ftp_write, spy, Snmpguess, Xlock, XSnoop, Httptunnel,Sendmail	995	2754
User to Root attack	Rootkit, Buffer-overflow, LoadModule, Sql-attack, perl, Xterm, Ps	52	200
Total		125973	22544

$$x - scaled = \frac{(x - min)}{max - min} \tag{1}$$

x-scaled represents the scaled value of x after applying the scaling methods, min represents the minimum among the column, the max value represents the maximum in the given column, and x is the value which you want to scale. Once the prepossessing is performed, we visualize the dataset to check the distribution around the mean. None of the datasets were found to have normal distribution around the mean. **Feature selection abbreviations**: PC—Pearson coefficient [9], IG—Information gain [10], RF—Random Forest feature selection [11], and SKB—Select K Best **Model abbreviations:** RF—Random Forest Classifier [12], XGB—Extreme Gradient Boosting [13], and DT—Decision Tree Classifier[14]

3.2 Best Feature Selection

The dataset dimension is reduced to a lesser dimension, selecting the best features among the existing data. Basically, we define a function for all these feature selection methods. The function takes k as input and returns the reduced dimension dataset features back to the calling function; while doing so, it selects the best k features among all available data features. We have used SelectKBest, Information gain, Random forest, feature selection, and Pearson coefficient for the best feature selection.

3.2.1 Select K Best The SelectKBest class scores the features of the dataset using a function and then retains only the k highest ranking features. In our case, we use the f_regression function given in sklearn library The SelectKBest classifier simply scores the options employing a function and then removes nearby "k" highest evaluation options. So, as an example, if the chi-square is passed as a score performed, SelectKBest can use the chi-square to compute the relation between every feature of "X" (Actual Data set without label) and "y" (assumed to be category labels). A small chi-square value means the feature is independent of y. An out-sized worth can mean the feature is non-randomly associated with y, so it probably supplies vital data. Solely k options are preserved; for the SelectKBest classifier, negative values can not be accepted, so this can not be used for the Z-score.

3.2.2 Information Gain The information gained will be used to find the best features from the dataset. Information gain technique is used to generate a preferred sequence of attributes which is used to narrow down the state of a random variable. An attribute with high mutual information is preferred over other attributes. As in Eq. 2.

$$IG(D, x) = \ddot{H}(D) - H(D, x) \tag{2}$$

where IG(D, x) is the information gained for the dataset D for the given variable x. H(D) is the entropy of the whole dataset before any partition. H(D,x) is condition randomness for given variable x.

3.2.3 Pearson Coefficient The Pearson coefficient has been used to find the best features in the dataset. It is used to calculate the dependence of two variables on each other. If two variables have a dependence very close to 1, either of them can be removed to reduce the number of features to be trained. The correlation coefficient is used to find the linear relationship between the two attributes in the given dataset. The final values are within *1* and *–1*, where *1* indicates a strong positive linear relationship between attributes. *–1* indicates a strong negative linear relationship between attributes. A result of zero indicates no relationship at all.

The equation for Pearson's coefficient as in

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

3.2.4 Random Forest Feature Selection Random forest classifiers can also be used for feature selection. It is a multi-tree-based approach where each tree is built on how well a split will increase the node purity, and it tries to reduce the impurity of all the built trees in the forest. The first and last nodes will have the highest and lowest increase in purity. Nodes with the highest increase in purity will be used for splitting the first time, while nodes with the lowest increase in purity will be used as a split at the end of the tree. We can compute the importance of each node by taking the average of the importance of nodes in each tree.

3.3 Machine Learning Models

For classification purposes, we have used different machine-learning classification techniques. We used a stacked ensemble model, random forest classifier, and autoencoder for training the model.

3.3.1 Stacked Ensemble Model We created the stacked ensemble model, which consists of four machine-learning classifiers. We train the autoencoder, neural network, and random forest parallelly. Later, it is passed into the Xtreme gradient boosting model for voting-based classification.

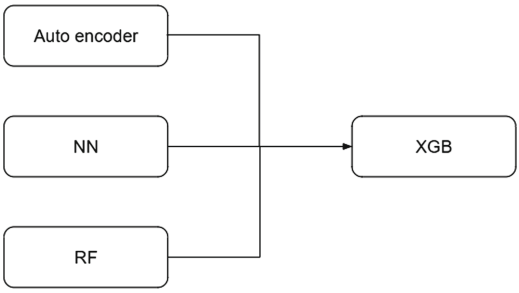


Fig. 1. Ensemble-model architecture

3.3.2 Random Forest Random forest use bootstrap sampling to obtain a subset of features to build a particular decision tree in an entire forest. This will ensure that no particular decision tree will be overfitting since no tree considers all the features. Random forests trade bias for variance. The final predictions are obtained by bagging by assigning equal importance to each decision tree.

3.3.3 Neural Networks This is a deep-learning model. We consider a 4-layer deep neural network with 16, 8, 8, and 1 node respectively. Rectified Linear Unit activation function is used at every hidden layer. The output layer contains the sigmoid activation function. The output is a probability between 0 and 1, indicating the probability that the input instance is an attack.

3.3.4 Auto Encoders Autoencoder consists of two components-encoders and decoders. This autoencoder is trained in an unsupervised fashion to learn a compressed version of the input. This compressed version of the input eliminates any noise that the input might have. For the purpose of classification, we discard the decoder portion. We take the compressed input from the encoder and feed it to a neural network that performs classification. This phase is supervised. So autoencoders for classification used in this work contain both supervised and unsupervised phases.

3.3.5 Extreme Gradient Boosting (XGB) In this work, XGB is used to combine the predictions of all base layer models. This is responsible for obtaining a non-linear combination of the individual base layer models. Soft-ensembling is used as the probabilities extracted from each model are combined. In XGB, multiple decision trees are built iteratively. Each tree is built by assigning a greater weightage to the misclassified instances from the previous decision tree. Finally, boosting is performed to combine the precision of each decision tree. The weight is decided by the weighted accuracy of the training instances. XGB has support for parallel processing, making it a very effective algorithm that can be sped up with GPU-based parallel processing techniques.

3.4 Why Ensemble Learning?

When multiple models with widely different training methodologies are used, we can expect them to make uncorrelated errors. This is the main motivation behind using the ensemble model; in this work, we used three different kinds of models in the base layer of our ensemble model. Neural networks are a deep learning model that is trained completely in a supervised manner. An Autoencoder is a deep learning model that has two components, one of which is trained in a supervised manner and the other in an unsupervised manner. Random forests is a machine learning model working with a completely different methodology. Hence, we can expect them to make uncorrelated errors. Figure 1 depicts the proposed ensemble model.

4 Experiments and Analysis

In this work, we combined the test and train dataset, which is given separately in the dataset, shuffled it before training, and then used the above-mentioned feature selection method to extract the k best feature. Further, run it for all possible values of k from 15 to 40. Later, computed the validation accuracy of the model with the validation dataset and then selected the one model which is performing best. In this work, the selected k-best method was performed best in comparison to other methods. In this work, selected the best features according to this model and formed the new training dataset. Later, we trained the proposed machine learning model on a newly formed dataset. This work has used different machine learning algorithms like XGBoost, Random forest and neural networks, autoencoders, and stacked attention networks. The evaluation metrics used in this work are Accuracy, Recall, Precision, and F1-Score.

Table 2. Accuracy of models for feature selectors on validation dataset

No. of feature	RF	PC	SKB	IG
15	99.10	99.53	**99.93**	99.80
20	99.58	99.85	99.93	99.92
25	99.87	99.92	99.93	99.93
30	99.87	99.93	99.93	99.93
35	99.93	99.93	99.93	99.93

Table 3. Evaluation metrics on trained models

Evaluation metric	RF	NN	AE	XGB
Accuracy	82.69	87.42	87.79	88.81
Recall	97.04	96.98	96.97	96.68
Precision	72.28	78.75	79.30	81.02
F-1 score	82.85	86.92	87.25	88.16

This work has used different machine learning algorithms like XGBoost, Random forest and neural networks, autoencoders, and stacked attention networks.

4.1 Training with Selected K Features

Once get the top k features using the above-mentioned feature selection algorithms, then, train these extracted features on four models (a) Random Forest Classifier (b) Neural Networks (c) Auto-encoder (d) Ensemble model with XGB classifier. From Table 2, It is observed that the highest accuracy obtained is 99.93% and there are many cells that contain the same value. If multiple models have the same accuracy, it would be beneficial to select the model which has this accuracy for the minimum number of features. This can help train the model faster while making sure accuracy doesn't get affected. This work selected the K value was 15 and the SelectKBest feature selection algorithm for training purposes.

4.2 Result and Analysis

Once the model is trained with a selected number of features on different models, we plot the following metrics. RF-Random Forest, NN-Neural Network, AE-Autoencoder, XGB-Ensemble model with XGB classifier

From the above results (Tables 3 and 4), It is noted that the Neural network, autoencoder and ensemble model have similar accuracies, but the random forest has an accuracy of 82.69, this can be attributed to the fact that while creating a random forest, some similar decision trees could be formed which makes it so that

Table 4. Comparison with SOTA and ensemble

Model	Accuracy
AE-supervised [7]	84.21
Random forests	82.69
AE	87.79
FFNN	87.42
Ensemble-with XGB	88.21

the model has duplicate results from similar models, thus reducing the overall information to train. It is observed that the ensemble model performs the best considering all the evaluation metrics. This can be attributed to the fact that the Ensemble model's results undergo a non-linear combination through XGB so that it can extract the best results from the models. The ensemble model has the highest F1 score, Precision, and accuracy when compared to other models. It also helps that all the models in the ensemble model are independent and make uncorrelated errors. In the context of this problem, since there exists a class imbalance, F1-score would be the best parameter to compare models, and we can see that the Ensemble model performs the best, followed by Autoencoder. Autoencoder has a similar performance as that ensemble model, which implies that maximum weightage has been given to the autoencoder when the non-linear combination takes place

5 Conclusion and Future Work

It is observed that feature selection algorithms have a significant effect on the result of the model. Even 15 features could be a good representation of the whole dataset thus reducing the training time of the model and also allowing us to host this with much less memory consumption. The model performance saturates after a certain K value of selected k features pointing to the fact that many features after this point are not relevant while training the model and don't contribute any important information. The model is trained with a Random forest classifier, Neural network, autoencoder, and ensemble model with an XGB classifier. The ensemble model performs a non-linear combination of these models to take the best information from each of these models. We see that this ensemble model outperforms all the component models.

References

1. Rashid, A., Siddique, M.J., Ahmed, S.M.: Machine and deep learning based comparative analysis using hybrid approaches for intrusion detection system. In: 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), pp. 1–9 (2020). IEEE
2. Ali, A., Shaukat, S., Tayyab, M., Khan, M.A., Khan, J.S., Ahmad, J., et al.: Network intrusion detection leveraging machine learning and feature selection. In: 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), pp. 49–53. IEEE (2020)
3. Gao, N., Gao, L., Gao, Q., Wang, H.: An intrusion detection model based on deep belief networks. In: 2014 Second International Conference on Advanced Cloud and Big Data, pp. 247–252. IEEE (2014)
4. Sangkatsanee, P., Wattanapongsakorn, N., Charnsripinyo, C.: Practical real-time intrusion detection using machine learning approaches. Comput. Commun. **34**(18), 2227–2235 (2011)
5. Sinclair, C., Pierce, L., Matzner, S.: An application of machine learning to network intrusion detection. In: Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99), pp. 371–377. IEEE (1999)

6. Alazzam, H., Sharieh, A., Sabri, K.E.: A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. Expert Syst. Appl. **148**, 113249 (2020)
7. Ieracitano, C., Adeel, A., Morabito, F.C., Hussain, A.: A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. Neurocomputing **387**, 51–62 (2020)
8. Aggarwal, P., Sharma, S.K.: Analysis of kdd dataset attributes-class wise for intrusion detection. Procedia Comput. Sci. **57**, 842–851 (2015)
9. Kirch, W. (ed.): Pearson's Correlation Coefficient, pp. 1090–1091. Springer, Berlin (2008)
10. Shaltout, N., Elhefnawi, M., Rafea, A., Moustafa, A.: Information gain as a feature selection method for the efficient classification of influenza based on viral hosts. Lect. Notes Eng. Comput. Sci. **1**, 625–631 (2014)
11. Kursa, M., Rudnicki, W.: The all relevant feature selection using random forest (2011)
12. Leo, B.: Random Forests, vol. 45. Springer, Berlin (2001)
13. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
14. Rokach, L., Maimon, O.: Decision Trees **6**, 165–192 (2005)



A Bee Colony Optimization Algorithm to Tuning Membership Functions in a Type-1 Fuzzy Logic System Applied in the Stabilization of a D.C. Motor Speed Controller

Leticia Amador-Angulo(✉) and Oscar Castillo

Tijuana Institute of Technology, Tijuana, Mexico
gloria.amador@tectijuana.edu.mx, ocastillo@tectijuana.mx

Abstract. In this research a Bee Colony Optimization algorithm (BCO) for stabilization of a D.C Motor Speed Controller is presented. The first idea of the BCO is to find of the optimal design of the Membership Functions (MFs) in the Type-1 Fuzzy Logic System (T1FLS). BCO algorithm shows excellent results when real problems are analyzed in the Fuzzy Logic Controller (FLC). Some types of indices performance implemented in the field of the control are used. With the goal of verifying the efficiency of the BCO a comparative with other bio-inspired algorithms for the stabilization of the case study.

Keywords: Fuzzy sets · Bee · Fuzzy logic controller · Speed · Uncertainty

1 Introduction

In the last years, the techniques in the implementation of the meta-heuristics algorithm have introduced stabilization and control in solving complex problems. some problems studied with the bco algorithm are; Arfiani et al. in [1] this algorithm is proposed with the hybridization in a k-means algorithm, Cai et al. in [2] study an improved bco by optimizing the clusters initial values, Chen et al. in [3] apply a bco based on quality-of-life health, Cubranić-dobrodolac et al. in [4] presents a bco to measuring speed control in vehicles, Jovanović et al. in [5] presents a bco to control the traffic, Selma et al. in [6] presents a hybridization of anfis controller and bco applied in control, and Wang et al. in [7] study an improved bco for airport freight station scheduling.

The main contribution is highlighting the good results in this algorithm to optimize the speed in the FLC problem, the real problem is simulated with a FLC to find the smallest error in the simulation.

The organization in each section is presented below. Section 2 shows some important Related Works. Section 3 outlines the bio-inspired algorithm. Section 4 shows the study case. Section 5 outlines the design presented of the T1FLS. Section 6 shows the results and a comparative analysis with others bio-inspired algorithms, and, Sect. 7 shows some important conclusions and some recommendations to improve this paper.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 746–755, 2023.
https://doi.org/10.1007/978-3-031-27409-1_68

2 Related Works

The real problem studied is called "dc speed motor controller", several authors is interesting in this problem, for example; in [8] this study case is analyzed with a raspberry pi 4 and python by Habil et al. in [9] a pid controller is tuning for this real problem by Idir et al., in [10] a real-time pid controller is designed with this problem by le Thai et al. in [11] some experimentally robustness is studied with this real problem by Prakosa et al., in [12] a particle swarm optimization (PSO) tuning is studied to stabilize this study case by Rahayu et al. and in [13] an interval linear quadratic regulator and its application is applied to this real problem by Zhi et al.

The BCO is an efficient technique used for several authors for some mention, in [14] a bco is used to find the values in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters with an it3fls, in [15] an effective bco for distributed flowshop, in [16] a bco model to construction site layout planning, in [17] a bco applied big data fuzzy c-means, and in [18] a bco and its applications.

3 Bee Colony Optimization Algorithm

The first idea in developing this bio-inspired algorithm was Teodorović Dušan. The function in the BCO algorithm consists in explores collective intelligence through honey bees in the recollection of the nectar [18]. This algorithm is characterized by having tow phases: ***backward pass*** and ***forward pass.*** Each bee can have one of three roles, such as; follower bee, scout bee and bee [19]. Equations 1–4 express the dynamics of BCO algorithm;

$$P_{ij,n} = \frac{\left[\rho_{ij,n}\right]^{\alpha} \cdot \left[\frac{1}{d_{ij}}\right]^{\beta}}{\sum_{j \in A_{i,n}} \left[\rho_{ij,n}\right]^{\alpha} \cdot \left[\frac{1}{d_{ij}}\right]^{\beta}} \tag{1}$$

$$\mathrm{D}_{\boldsymbol{i}} = \mathrm{K} \cdot \frac{Pf_{\,i}}{Pf_{\,colony}} \tag{2}$$

$$Pf_{\,i} = \frac{1}{L_I}, L_i = TourLength \tag{3}$$

$$\mathrm{Pf_{colony}} = \frac{1}{\mathrm{N_{Bee}}} \sum_{\mathrm{i=1}}^{\mathrm{N_{Bee}}} \mathrm{Pf_i} \tag{4}$$

A bee has a probability (k) in a node (i), Eq. 1 express this behavior, where (j) is the following node selected, all nodes in a neighborhood is expressed by $N^k{}_i$, $\rho_{\,\mathrm{ij}}$ indicates a rating value, β expresses the exploration in the algorithm (next node to visit), respect to, d_{ij} indicates the value for the heuristic distance and α represents in the actual iteration the best solution. Equation 2 represents the duration for a bee, where the waggle dances is expressed by K [20]. A bee (i) has in the execution a probability score expressed by $Pf_{\,i}$ in Eq. 3, and $Pf_{\,colony}$ indicates the average of the probability in all colony and is expressed by Eq. 4. The Fig. 1 illustrates the step to step of the BCO algorithm.

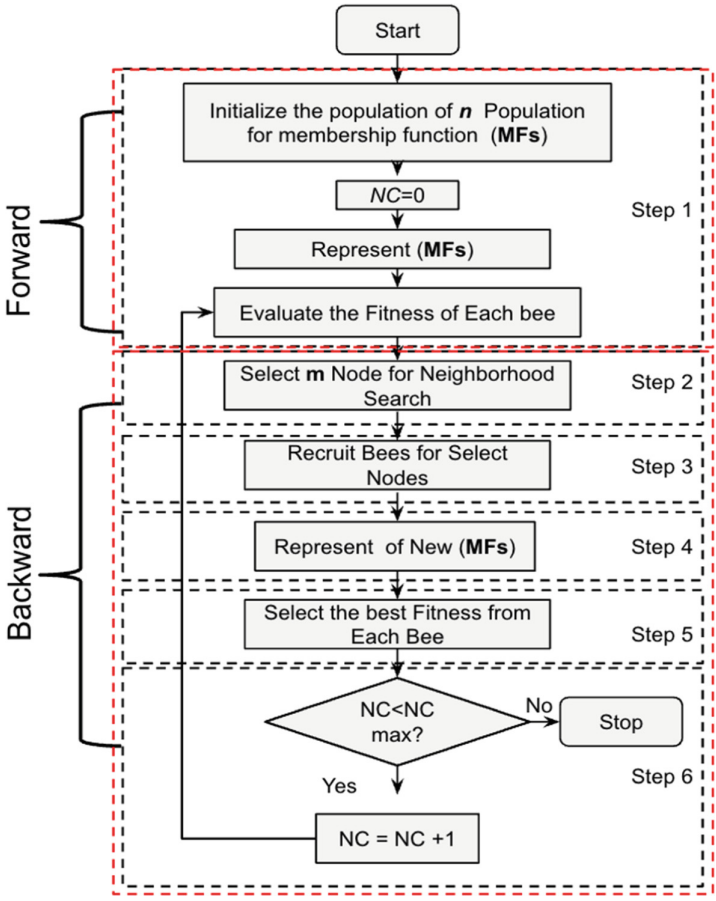


Fig. 1. Illustration step to step in the BCO algorithm.

4 Fuzzy Logic Controller Problem

The problem statement presents a real problem called DC speed motor is very popular in control controller. Figure 2 illustrates the initial state of the references, where the main objective consists in moving starting from an initial state at a speed based in 40 rad/s, and the model in the FLC is represented in Fig. 3.

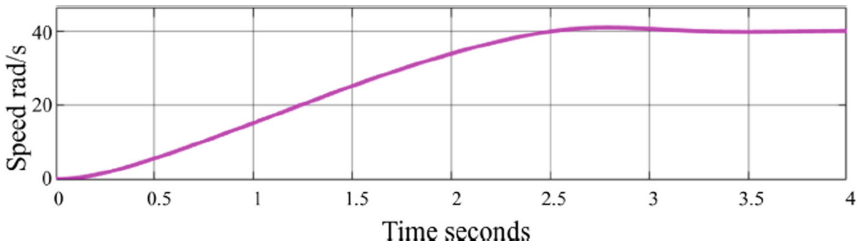


Fig. 2. Behavior in the speed response with the initial model.

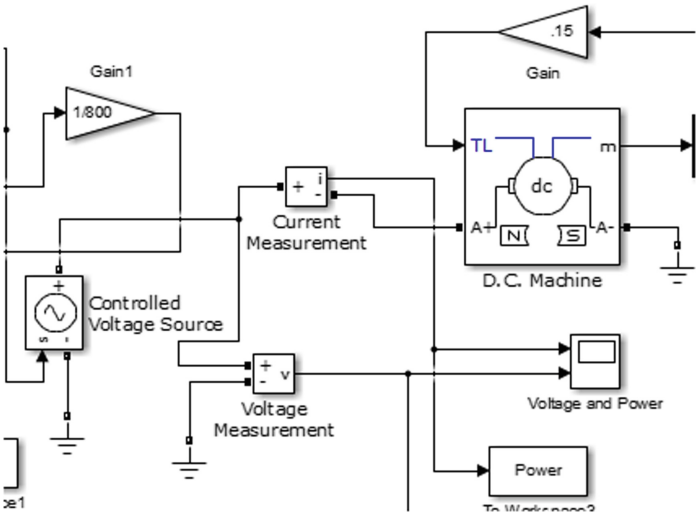


Fig. 3. Model of control for the studied problem.

5 Proposed Design of the T1FLS

5.1 State of the Art for the T1FLS

Zadeh proposed the main idea of the FLS in 1965 [21, 22]. Mamdani in 1974 propose a case of Fuzzy Controller to implementation of the FLS [23]. Figure 4 shows the graphic description of a T1FLS.

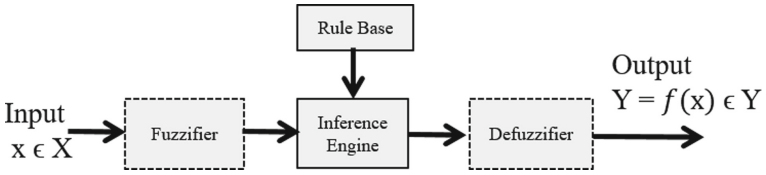


Fig. 4. Architecture of a T1FLS.

A T1FLS in the universe x is characterized by a MF $u_A(x)$ taking values on the interval [0, 1] and can be defined by Eq. (5) [21, 22].

$$A = \{(x, \mu_A(x)) | x \in X\} \tag{5}$$

5.2 Proposed Design of the T1FLS

The main architecture is designed with a Mamdani type of system, the Fig. 5 shows the visual representation (inputs and output) and the distribution of the each MFs (triangular and trapezoidal) with the names for each linguistic variable, and Fig. 6 represents the 15 rules that contain the T1FLS.

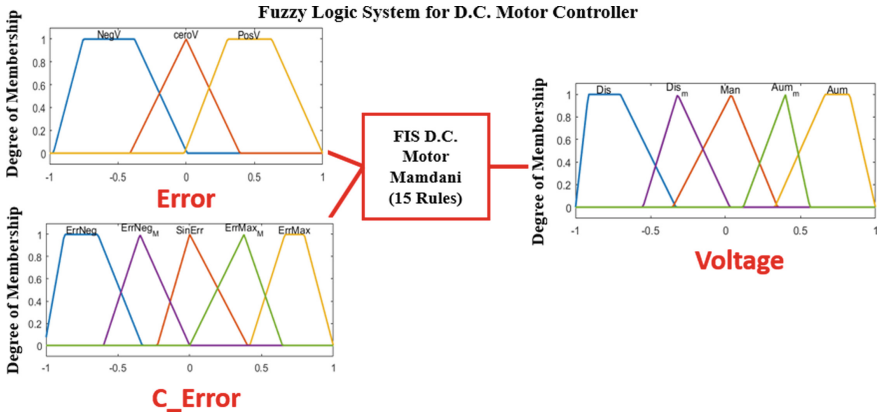


Fig. 5. Design of the proposed T1FLS.

1.- if (**Error** is NegV) and (**C_Error** is ErrNeg) then (**Voltage** is Dis)
2.- if (**Error** is NegV) and (**C_Error** is SinErr) then (**Voltage** is Dis)
3.- if (**Error** is NegV) and (**C_Error** is ErrMax) then (**Voltage** is Dis_m)
4.- if (**Error** is CeroV) and (**C_Error** is ErrNeg) then (**Voltage** is Aum_m)
5.- if (**Error** is CeroV) and (**C_Error** is ErrMax) then (**Voltage** is Dis_m)
6.- if (**Error** is PosV) and (**C_Error** is ErrNeg) then (**Voltage** is Aum_m)
7.- if (**Error** is PosV) and (**C_Error** is SinErr) then (**Voltage** is Aum)
8.- if (**Error** is PosV) and (**C_Error** is ErrMax) then (**Voltage** is Aum)
9.- if (**Error** is CeroV) and (**C_Error** is SinErr) then (**Voltage** is Man)
10.- if (**Error** is NegV) and (**C_Error** is ErrNeg_M) then (**Voltage** is Dis)
11.- if (**Error** is CeroV) and (**C_Error** is ErrNeg_M) then (**Voltage** is Aum_m)
12.- if (**Error** is PosV) and (**C_Error** is ErrNeg_M) then (**Voltage** is Aum)
13.- if (**Error** is PosV) and (**C_Error** is ErrMax_M) then (**Voltage** is Aum)
14.- if (**Error** is CeroV) and (**C_Error** is ErrMax_M) then (**Voltage** is Dis_m)
15.- if (**Error** is NegV) and (**C_Error** is ErrMax_M) then (**Voltage** is Dis)

Fig. 6. Proposed Fuzzy Rules for the T1FLS.

A bee indicates the possible solution, in this case, the number the parameters that represent each MFs in the T1FLS, for this real problem a total of 45 values and Fig. 7 represents of the vector solution.

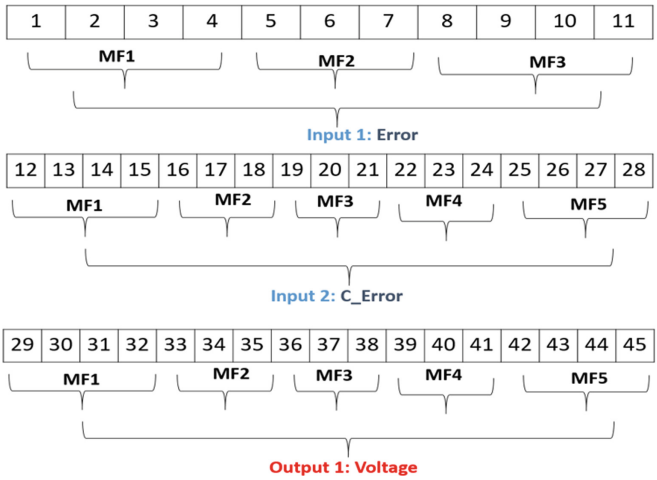


Fig. 7. Distribution of the values in the each MFs (Vector solution).

6 Results in the Experimentation

The main parameters for the BCO algorithm are represented in Table 1.

Table 1. Main values in the parameters for the BCO algorithm

Parameters	Values
Population (N)	50
Follower Bee	25
α	0.5
β	2.5
Iterations	30

The function fitness used in BCO algorithm is Root Mean Square Error (RMSE) and is expressed by Eq. (6).

$$\varepsilon = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(X_t - \widehat{X}_t\right)^2} \tag{6}$$

Others metrics to evaluate the efficiency in the results for the FLCs are presented by Eqs. (7–11).

$$ISE = \int_0^{\infty} e^2(t)dt \tag{7}$$

$$IAE = \int_0^{\infty} |e(t)|dt \tag{8}$$

$$ITSE = \int_0^{\infty} e^2(t)tdt \tag{9}$$

$$ITSE = \int_0^{\infty} |e(t)|tdt \tag{10}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{Y}_i - Y_i\right)^2 \tag{11}$$

A total of 30 experiments were executed. The average (AVG) is presented in Table 2 for minimum values to find by BCO algorithm.

Table 2. Final errors for the BCO algorithm

PerformanceIndexes	Best	Worst	AVG	σ
ITAE	4.19E-03	3.78E + 02	1.48E + 02	2.26E + 01
ITSE	2.52E-06	2.12E + 04	4.25E + 02	1.47E + 03
IAE	1.53E-03	3.83E + 03	1.72E + 01	3.49E + 02
ISE	8.23E-07	6.77E + 03	1.61E + 02	4.68E + 02
MSE	1.95E-07	2.01E + 03	9.32E + 01	1.82E + 02

Table 2 shows the best MSE that BCO algorithm to find is of **1.95E-07**, which represents an important stabilization in the speed of the FLC.

Table 3 shows a comparative with other algorithms to help in the demonstration of the good results to find in this paper, such as; Chicken Search Optimization (CSO), Fuzzy Harmony Search (FHS) and Fuzzy Differential Evolutional (FDE).

Table 3. Comparison between the BCO, FHS and FDE algorithms

Performance indexes	*BCO*	CSO [24]	FHS [25]	FDE [25]
Best	3.69E-02	1.38E-02	2.36E-01	2.73E-01
Worst	7.87E + 00	9.17E + 00	7.00E-01	6.06E-01
AVG	1.04E + 00	5.18E-01	4.52E-01	4.35E-01

Table 3 presents the cso with a rmse value of **1.38e-03**, fhs with a **2.36e-01** and fde with a **2.73e-01**, comparing results the rmse to bco is of **3.69e-02** value. The best results

is with the cso regarding bco, fhs and fde algorithm. The metric of the average for all simulations is better with fhs algorithm with a value of **4.52e-01** instead for the bco the result to find is the **1.04e + 00**. Figure 8 shows the convergence for this algorithm, and Fig. 9 shows the speed response in the real problem.

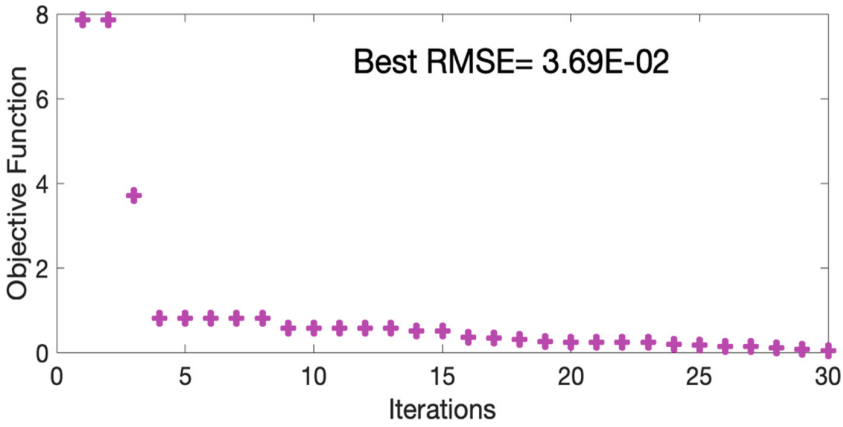


Fig. 8. Best Convergence on the results in the proposal.

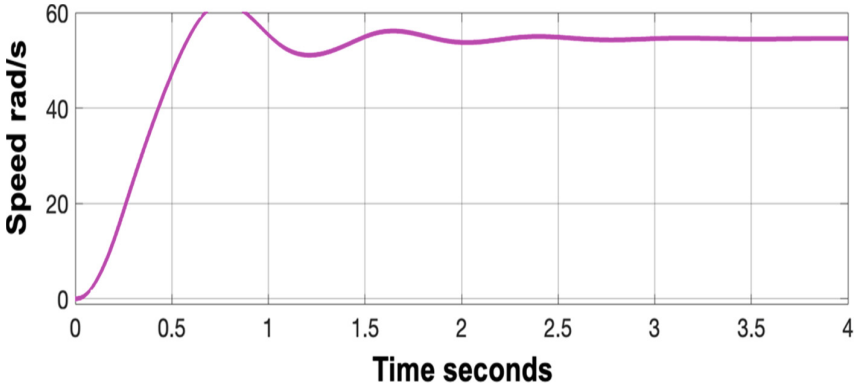


Fig. 9. Speed response in the real problem with the BCO algorithm.

7 Conclusions

The main conclusion is highlighting the efficiency of the proposed algorithm based on a real problem for the fuzzy controller. A stabilization on the speed is shown on the results (see Fig. 7). in this paper, an important analysis in the comparative with three metaheuristics algorithms was possible to realize with cso, fhs and fde (see Table 3), the main conclusion is to analyze that the bco algorithm obtain excellent results with the

fitness functions metric with a value of 3.69e-02 compared to cso with 1.38e-02 theses two algorithms present excellent results, regarding a fhs of 2.36e-01 and fde of 2.73e-01. A strategy to improve this research is to add perturbation or disturbance in the flc with the main objective of exploiting in greater depth the results of the bco algorithm. Other idea is to increase the extension of the fuzzy sets (fs) with the interval type—2 fls and to be able to analyze in more detail the levels of the uncertainty in the real problem.

References

1. I. Arfiani, H. Yuliansyah and M.D. Suratin, "Implementasi Bee Colony Optimization Pada Pemilihan Centroid (Klaster Pusat) Dalam Algoritma K-Means". *Building of Informatics, Technology and Science (BITS)*, vol.3, no.4, pp. 756–763, 2022
2. Cai, J., Zhang, H., Yu, X.: Importance of clustering improve of modified bee colony optimization (MBCO) algorithm by optimizing the clusters initial values. J. Intell. & Fuzzy Syst., (Preprint), 1–17
3. Chen, R.: Research on motion behavior and quality-of-life health promotion strategy based on bee colony optimization. J. Healthc. Eng., 2022
4. Čubranić-Dobrodolac, M., Švadlenka, L., Čičević, S., Trifunović, A., Dobrodolac, M.: A bee colony optimization (BCO) and type-2 fuzzy approach to measuring the impact of speed perception on motor vehicle crash involvement. Soft. Comput. **26**(9), 4463–4486 (2021). https://doi.org/10.1007/s00500-021-06516-4
5. Jovanović, A., Teodorović, D.: Fixed-time traffic control at superstreet intersections by bee colony optimization. Transp. Res. Rec. **2676**(4), 228–241 (2022)
6. Selma, B., Chouraqui, S., Selma, B., Abouaïssa, H.: Design an Optimal ANFIS controller using bee colony optimization for trajectory tracking of a quadrotor UAV. J. Inst. Eng. (India): Ser. B, 1–15 (2022)
7. Wang, H., Su, M., Zhao, R., Xu, X., Haasis, H.D., Wei, J., Li, H.: Improved multi-dimensional bee colony algorithm for airport freight station scheduling. arXiv preprint arXiv:2207.11651, (2022)
8. Habil, H.J., Al-Jarwany, Q. A., Hawas, M. N., Nati, M.J.: Raspberry Pi 4 and Python based on speed and direction of DC motor. In: 2022 4th Global Power, Energy and Communication Conference (GPECOM), pp. 541–545. IEEE (2022)
9. Idir, A., Khettab, K., Bensafia, Y.: Design of an optimally tuned fractionalized PID controller for dc motor speed control via a henry gas solubility optimization algorithm. Int. J. Intell. Eng. Syst. **15**, 59–70 (2022)
10. Le Thai, N., Kieu, N.T.: Real-Time PID controller for a DC motor using STM32F407. Saudi J Eng Technol **7**(8), 472–478 (2022)
11. Prakosa, J. A., Gusrialdi, A., Kurniawan, E., Stotckaia, A. D., Adinanta, H.: Experimentally robustness improvement of DC motor speed control optimization by H-infinity of mixed-sensitivity synthesis. Int. J. Dyn. Control., 1–13, 2022
12. Rahayu, E.S., Ma'arif, A., Çakan, A.: Particle Swarm Optimization (PSO) tuning of PID control on DC motor. Int. J. Robot. Control. Syst. **2**(2), 435–447 (2022)
13. Zhi, Y., Weiqing, W., Jing, C., Razmjooy, N.: Interval linear quadratic regulator and its application for speed control of DC motor in the presence of uncertainties. ISA Trans. **125**, 252–259 (2022)
14. Amador-Angulo, L., Castillo, P. Melin, P., Castro, J. R.: Interval Type-3 fuzzy adaptation of the bee colony optimization algorithm for optimal fuzzy control of an autonomous mobile robot. Micromachines, **13**(9), 1490 (2022)

15. Huang, J.P., Pan, Q.K., Miao, Z.H., Gao, L.: Effective constructive heuristics and discrete bee colony optimization for distributed flowshop with setup times. Eng. Appl. Artif. Intell. **97**, 104016 (2021)
16. Nguyen, P.T.: Construction site layout planning and safety management using fuzzy-based bee colony optimization model. Neural Comput. Appl. **33**(11), 5821–5842 (2020). https://doi.org/10.1007/s00521-020-05361-0
17. Razavi, S.M., Kahani, M., Paydar, S.: Big data fuzzy C-means algorithm based on bee colony optimization using an Apache Hbase. Journal of Big Data **8**(1), 1–22 (2021). https://doi.org/10.1186/s40537-021-00450-w
18. Teodorović, D., Davidović, T., Šelmić, M., Nikolić, M.: Bee colony optimization and its Applications. Handb. AI-Based Metaheuristics, 301–322 (2021)
19. Biesmeijer, J. C., Seeley, T. D.: The use of waggle dance information by honey bees throughout their foraging careers. Behav. Ecol. Sociobiol. **59**(1), 133–142 (2005)
20. Dyler, F.C.: The biology of the dance language. Annu. Rev. Entomol. **47**, 917–949 (2002)
21. Zadeh, L.A.: The concept of a Linguistic variable and its application to approximate reasoning. Part II, Information Sciences **8**, 301–357 (1975)
22. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. **1**(1), 3–28 (1978)
23. Mamdani, E.H.: Application of fuzzy algorithms for control of simple dynamic plant. In Proceedings of the Institution of Electrical Engineers **121**(12), 1585–1588 (1974)
24. Amador-Angulo, L., Castillo, O.: Stabilization of a DC motor speed controller using type-1 fuzzy logic systems designed with the chicken search optimization algorithm". In: International conference on intelligent and fuzzy systems, pp. 492–499. Springer, Cham (2021)
25. Castillo, O., et al.: A high-speed interval type 2 fuzzy system approach for dynamic parameter adaptation in metaheuristics. Eng. Appl. Artif. Intell. **85**, 666–680 (2019)



Binary Classification with Genetic Algorithms. A Study on Fitness Functions

Noémi Gaskó(✉)

Faculty of Mathematics and Computer Science, Centre for the Study of Complexity, Babeş-Bolyai University, Cluj-Napoca, Romania
noemi.gasko@ubbcluj.ro

Abstract. In this article, we propose a new fitness function that can be used in real-value binary classification problems. The fitness function takes into account the iteration step, controlling with it the importance of some elements of the function. The designed genetic algorithm is compared with two other variants of genetic algorithms, and with other state-of-the-art methods. Numerical experiments conducted both on synthetic and real-world problems show the effectiveness of the proposed method.

Keywords: Classification problem · Genetic algorithm · Fitness function · Synthetic dataset · Real-world dataset

1 Introduction and Problem Statement

Classification, an essential task in machine learning, aims to sort data into different classes. Examples of application possibilities include speech recognition [5], protein classification [6], handwriting recognition [2], face recognition [3], etc.

Supervised classification problems can be divided into several classes [4] proposes a taxonomy in which four properties are investigated: structure, cardinality, scale, and relation category features. In the next section, we will focus on binary classification problems, where the features take real values.

Formally, the binary classification problem can be described as follows: for a given set of input data $X = (x_1, x_2, \ldots, x_n)^T$, where $x_i \in \mathbf{R}^p, p > 1$ and for a given set of labels $Y = \{y_1, y_2, \ldots, y_n\}$, where $y_i \in \{0, 1\}$ (y_i corresponding to x_i) the problem consists in finding a model that makes a good prediction from the input data X to the Y.

Several algorithms have been used for classification problems, such as support vector machines, decision trees, and logistic regression. Surveys describing these methods include [8,9,15].

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, project number 194/2021 within PNCDI III.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 756–761, 2023.
https://doi.org/10.1007/978-3-031-27409-1_69

GA was successfully used for feature selection (for example, in [1,17]), and for binary classification as well. In [12] a genetic algorithm is combined with an Adaboost ensemble-based classification algorithm for binary classification problems; the new algorithm is called GA(M)-EQSAR [11] proposes a genetic algorithm based on an artificial neural network that is compared with log maximum likelihood gradient ascent and root-mean-square error to minimise gradient descent algorithms. In [14] a parallel genetic algorithm is presented to solve a nonlinear programming form of the binary classification problem [10] proposes a prediction model of bankruptcy prediction based on a binary classification model and a genetic algorithm.

The main goal of this article is to use a genetic algorithm to solve the binary classification problem. Genetic algorithms (GA) are a powerful optimisation tool in continuous and discrete problems. The essential tasks in designing a GA consist in finding a good representation (encoding) and in defining the fitness function, which are not trivial tasks in solving complex problems. In this article, we design a new fitness function and we compare the results obtained with the new function with the existing ones from the literature.

The rest of the paper is organised as follows: the next section describes the genetic algorithm and the proposed fitness function. Section three presents the obtained numerical results, and the article ends with a conclusion and recommendations for further work.

2 Proposed Method—Genetic Algorithm

In the next section, we present the essential parts of the genetic algorithm.

Encoding The chromosome represents a classification rule. For encoding, we use a vector representation with length $L = 2 \cdot NF + 1$, where NF is the number of features. When the features are real values, for each feature two genes are used in the chromosome. If in these pairs the first value is greater or equal to the second value, we do not take into account the classification rule. The last gene of the chromosome can only take the values of 0 and 1, which is the classification label.

Example 1. Let us consider a simple example with three real-value features (f_1, f_2,f_3) in the interval $[-1, 1]$, and the following chromosome:
0.21 0.11 −0.39 0.42 0.37 0.81 1. In the first pair, the first element is greater than the second one; therefore, we do not take into account in the classification rule, which will be as follows:

IF $f_2(i) > -0.39$ and $f_2(i) < 0.42$ and $f_3(i) > 0.37$ and $f_3(i) < 0.81$ THEN class = 1

ELSE class = 0,

where in $f_k(i), k \in \{1, 2, 3\}$, i represents an instance (a row) of the problem.

Fitness function Before the definition of the fitness function, we describe the basic classification classes:

- true positive (TP)—the actual class is Y, and the predicted class is also Y
- false positive (FP)—the actual class is Y, and the predicted class is not Y
- true negative (TN)—the actual class is not Y, and the predicted class is not Y
- false negative (FN)—the actual class is not Y, but the predicted class is Y

The proposed fitness function takes into account, among other factors, precision ($\frac{TP}{TP+FP}$) and sensitivity ($\frac{TP}{TP+FN}$). The fitness function takes into account the iteration number (denoted by gen_{nr}) as well, and the precision counts more in every step, while the sensitivity has less importance after some generations:

$$f_1 = gen_{nr} \cdot \frac{TP}{TP+FP} + \frac{1}{gen_{nr}} \frac{TP}{TP+FN} + TP + TN$$

In the following, we present two fitness functions designed for classification problems. These fitness functions will be used for comparisons. In [16], the following fitness function is proposed:

$$f_2 = \frac{TP}{TP + A \cdot FN} \cdot \frac{TN}{TN + B \cdot FP},$$

where $0.2 < A < 2$ and $1 < B < 20$ are two parameters.

In [13] the following fitness function is proposed:

$$f_3 = w_1 \cdot PA + w_2 \cdot CPH + w_3 \cdot S,$$

where w_1, w_2, w_3 are parameters, $PA = \frac{TP}{TP+FP}$ is predictive accuracy, CPH is the comprehensibility, the difference between maximum number of coalitions and the actual number of conditions, and $S = \frac{TP}{TP+FN}$ is the sensitivity.

2.1 Genetic Operators Standard operators are used; uniform mutation and crossover is used; and for selection, elitist selection is applied.

3 Numerical Experiments

Data sets For numerical experiments, synthetic and real-world data are used. For synthetic data the scikit-learn[1] Python library is used. Synthetic data was generated with different difficulty level (a smaller value of class separator indicates a harder classification problem). Two real world data sets where used: in the case of the data banknote authentication data were extracted from images from genuine and forged banknote-like specimens. The Haberman's survival data set contains cases on the survival of patients who had undergone surgery for breast cancer from a study conducted at the University of Chicago's Billings Hospital.

Table 1 presents the basic properties of the used data sets, the number of instances and the number of attributes.

[1] https://scikit-learn.org/stable/.

Table 1. Synthetic and real-world data sets used for the numerical experiments

Data set	No. instances	No. attributes
Synthetic 1	100	4 (seed=1967, class_separator=0.5)
Synthetic 2	100	4 (seed=1967, class_separator=0.1)
Synthetic 3	100	4 (seed=1967, class_separator=0.3)
Synthetic 4	100	4 (seed=1967, class_separator=0.7)
Synthetic 5	100	4 (seed=1967, class_separator=0.9)
Synthetic 6	100	3 (seed=1967, class_separator=0.5)
Synthetic 7	100	3 (seed=1967, class_separator=0.1)
Synthetic 8	100	3 (seed=1967, class_separator=0.3)
Synthetic 9	100	3 (seed=1967, class_separator=0.7)
Synthetic 10	100	3 (seed=1967, class_separator=0.9)
Banknote [7]	1372	4
Haberman's [7]	306	3

Parameter setting For implementation of the genetic algorithm, we use a public Python code.[2] The used parameters are the following: the population size is 40, the maximum number of generations is 500, the mutation probability is 0.1, and the crossover probability is 0.8. The rest of the parameters are the same as in the basic downloaded code.

Performance evaluation For the performance evaluation, we use normalised accuracy, the fraction of correctly detected classes over the total number of predictions. To obtain normalised accuracy, we use a ten-fold cross-validation, where 90% of the data is used to fit the data and 10% is used to test the performance of the algorithm.

Comparisons with other methods For comparison we use three variants of the GA algorithm: GA_1—the genetic algorithm with our proposed fitness function, GA_2 and GA_3 with the above described f_2 respectively f_3 fitness functions. We compare these variants of genetic algorithms with four well-known classifiers from the literature: Logistic Regression (LR), the k-nearest-neighbour classifier (kNN), the Decision Tree classifier (DT), and the Random Forest classifier (RF).

Results Table 2 presents the average and standard deviation of obtained accuracy values for ten independent runs. We used a Wilcoxon ranksum statistic test in order to decide if there exits a statistical difference between the compared methods.

In the case of real-world data sets the four state-of-the art methods (LR, kNN, DT, RF) outperformed the proposed genetic algorithm. Regarding the synthetic data sets, in harder classification problems (where the value of the

[2] downloaded from https://pypi.org//project/geneticalgorithm/, last accessed 1/09/2022.

class separator is smaller) the proposed GA performed as well as the classic state of the art algorithms. From the three variants of the genetic algorithms, GA_1 outperformed GA_2 in one case, and outperformed GA_3 in four cases.

Table 2. Average values and standard deviation of the normalized accuracy over 10 independent runs. A (*) indicates the best result based on the Wilcoxon ranksum statistical test (more stars in a line indicate no statistical difference)

Dataset	GA_1	GA_2	GA_3	LR	kNN	DT	RF
Synthetic1	$0.49 \pm 0.16^*$	$0.55 \pm 0.14^*$	$0.45 \pm 0.15^*$	$0.54 \pm 0.18^*$	$0.58 \pm 0.18^*$	$0.53 \pm 0.00^*$	$0.57 \pm 0.14^*$
Synthetic2	$0.53 \pm 0.16^*$	$0.55 \pm 0.10^*$	$0.46 \pm 0.15^*$	0.37 ± 0.21	$0.58 \pm 0.14^*$	$0.51 \pm 0.00^*$	$0.53 \pm 0.11^*$
Synthetic3	$0.53 \pm 0.16^*$	$0.52 \pm 0.21^*$	0.45 ± 0.15	$0.50 \pm 0.16^*$	$0.53 \pm 0.17^*$	$0.51 \pm 0.00^*$	$0.59 \pm 0.14^*$
Synthetic4	$0.55 \pm 0.15^*$	$0.55 \pm 0.14^*$	0.45 ± 0.15	0.58 ± 0.18	$0.64 \pm 0.20^*$	$0.67 \pm 0.00^*$	$0.66 \pm 0.16^*$
Synthetic5	0.55 ± 0.15	0.56 ± 0.16	0.47 ± 0.13	$0.72 \pm 0.12^*$	$0.73 \pm 0.14^*$	$0.65 \pm 0.00^*$	$0.71 \pm 0.14^*$
Synthetic6	$0.51 \pm 0.16^*$	$0.52 \pm 0.10^*$	0.44 ± 0.13	$0.54 \pm 0.18^*$	$0.58 \pm 0.18^*$	$0.53 \pm 0.00^*$	$0.58 \pm 0.18^*$
Synthetic7	$0.55 \pm 0.15^*$	$0.51 \pm 0.11^*$	$0.47 \pm 0.16^*$	0.37 ± 0.21	$0.58 \pm 0.14^*$	$0.49 \pm 0.00^*$	$0.49 \pm 0.14^*$
Synthetic8	$0.57 \pm 0.14^*$	0.46 ± 0.15	0.43 ± 0.16	0.50 ± 0.16	$0.53 \pm 0.17^*$	$0.56 \pm 0.00^*$	$0.59 \pm 0.15^*$
Synthetic9	0.45 ± 0.15	0.46 ± 0.15	0.45 ± 0.16	0.58 ± 0.18	0.64 ± 0.20	$0.67 \pm 0.00^*$	$0.68 \pm 0.12^*$
Syntetic10	0.49 ± 0.16	0.52 ± 0.12	0.50 ± 0.12	0.72 ± 0.12	$0.73 \pm 0.14^*$	0.66 ± 0.00	$0.76 \pm 0.13^*$
Banknote	0.47 ± 0.06	0.51 ± 0.10	0.56 ± 0.04	0.99 ± 0.01	$1.00 \pm 0.00^*$	$0.98 \pm 0.00^*$	$0.99 \pm 0.00^*$
Haberman's	0.37 ± 0.19	0.54 ± 0.22	0.71 ± 0.08	0.71 ± 0.08	$0.71 \pm 0.08^*$	$0.64 \pm 0.00^*$	$0.70 \pm 0.09^*$

4 Conclusions and Further Work

Genetic algorithms are useful optimisation methods in several challenging problems. The use of GAs in classification problems is a straightforward choice. In this article, we propose a new fitness function that can be used in binary real-value classification problems. The designed genetic algorithm is compared with other variants of GA, with different fitness functions, and with other state-of-the-art methods. Numerical experiments conducted on synthetic and real-world problems show the effectiveness of the proposed method.

Other fitness functions will be investigated in future studies, as well as an extension of the multiclass classification problem.

References

1. Babatunde, O.H., Armstrong, L., Leng, J., Diepeveen, D.: A genetic algorithm-based feature selection (2014)
2. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Convolutional neural network committees for handwritten character classification. In: 2011 International Conference on Document Analysis and Recognition, pp. 1135–1139. IEEE (2011)
3. Connolly, J.F., Granger, E., Sabourin, R.: An adaptive classification system for video-based face recognition. Inf. Sci. **192**, 50–70 (2012)

4. Czarnowski, I., Jędrzejowicz, P.: Supervised classification problems-taxonomy of dimensions and notation for problems identification. IEEE Access **9**, 151386–151400 (2021)
5. Desai, N., Dhameliya, K., Desai, V.: Feature extraction and classification techniques for speech recognition: a review. Int. J. Emerg. Technol. Adv. Eng. **3**(12), 367–371 (2013)
6. Diplaris, S., Tsoumakas, G., Mitkas, P.A., Vlahavas, I.: Protein classification with multiple algorithms. In: Panhellenic Conference on Informatics, pp. 448–456. Springer, Berlin (2005)
7. Dua, D., Graff, C.: UCI Machine Learning Repository (2017). https://archive.ics.uci.edu/ml
8. Kesavaraj, G., Sukumaran, S.: A study on classification techniques in data mining. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–7 (2013)
9. Kumar, R., Verma, R.: Classification algorithms for data mining: a survey. Int. J. Innov. Eng. Technol. (IJIET) **1**(2), 7–14 (2012)
10. Min, J.H., Jeong, C.: A binary classification method for bankruptcy prediction. Expert Syst. Appl. **36**(3), 5256–5263 (2009)
11. Pendharkar, P.C.: A comparison of gradient ascent, gradient descent and genetic-algorithm-based artificial neural networks for the binary classification problem. Expert Syst. **24**(2), 65–86 (2007)
12. Perez-Castillo, Y., Lazar, C., Taminau, J., Froeyen, M., Cabrera-Pérez, M.Á., Nowe, A.: Ga (m) e-qsar: a novel, fully automatic genetic-algorithm-(meta)-ensembles approach for binary classification in ligand-based drug design. J. Chem. Inf. Model. **52**(9), 2366–2386 (2012)
13. Robu, R., Holban, S.: A genetic algorithm for classification. In: Recent Researches in Computers and Computing-International Conference on Computers and Computing, ICCC. vol. 11 (2011)
14. To, C., Vohradsky, J.: Binary classification using parallel genetic algorithm. In: 2007 IEEE Congress on Evolutionary Computation, pp. 1281–1287. IEEE (2007)
15. Umadevi, S., Marseline, K.J.: A survey on data mining classification algorithms. In: 2017 International Conference on Signal Processing and Communication (ICSPC), pp. 264–268. IEEE (2017)
16. Vivekanandan, P., Nedunchezhian, R.: A new incremental genetic algorithm based classification model to mine data with concept drift. J. Theor. Appl. Inf. Technol. **21**(1) (2010)
17. Yang, J., Honavar, V.: Feature subset selection using a genetic algorithm. In: Feature Extraction, Construction and Selection, pp. 117–136. Springer, Berlin (1998)

SA-K2PC: Optimizing K2PC with Simulated Annealing for Bayesian Structure Learning

Samar Bouazizi[1,3(✉)], Emna Benmohamed[1,2], and Hela Ltifi[1,3]

[1] Research Groups in Intelligent Machines, National School of Engineers (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia
bouazizi.samar@gmail.com, emna.benmohamed@enis.tn, hela.ltifi@ieee.org

[2] Computer Department of Cyber Security, College of Engineering and Information Technology, Onaizah Colleges, P.O. Box 5371, Onaizah, Kingdom of Saudi Arabia

[3] Computer Science and Mathematics Department, Faculty of Sciences and Techniques of Sidi Bouzid, University of Kairouan, Kairouan, Tunisia

Abstract. Bayesian Network is an efficient theoretical model to deal with uncertainty and knowledge representation. Its development process is divided into two stages: (1) learning the structure and (2) learning the parameters. In fact, defining the optimal structure is a big difficulty that has been extensively investigated and still needs improvements. We present, in this paper, an extension of the existing K2PC algorithm with Simulated Annealing optimization for node ordering. Experimentations on well-known networks show that our proposal can extract the original's closest topology efficiently and reliably.

Keywords: Bayesian network · K2PC · Simulated annealing · Structure learning

1 Introduction

Bayesian networks (BNs) are frequently used in a variety of fields, such as risk analysis, medical diagnosis, agriculture, machine learning, etc. [10, 11], and this due to their ability to represent probabilistic knowledge over a set of variables considered as uncertain. A BN is a graphical model built over a set of random variables [9, 13]. It is denoted as $BN = (Gr, P)$, with P indicates the distributions' probability and Gr a directed acyclic graph. $Gr = (No, Ed)$, *where* $No = (No_1, No_2, \ldots, No_n)$ represents the nodes having discrete or continuous values. The dependence between the connected parent and child nodes is represented by a set of directed edges. Expert knowledge and reasoning modeling have made substantial use of causal probabilistic networks [19].

Finding the best structure for the dataset is an NP-hard task. The fundamental explanation is the rapidly increase in the number of structures that can exist. As a result, numerous BN structure learning methods, including [17] have been introduced. Three main approaches have been suggested: (1) constraint-based approaches, (2) score-based approaches, and (3) hybrid approaches. The second one includes the most often utilized

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 762–775, 2023.
https://doi.org/10.1007/978-3-031-27409-1_70

algorithms, like the K2 algorithm [9] and its improvement K2PC [6, 7]. These two algorithms employ a greedy heuristic search strategy for skeleton construction, and their effectiveness is primarily determined by the received variables' order.

Because of the importance of node ordering, numerous approaches have been suggested, that are classified as evolutionary and heuristic [15]. "*Who learns better BN structures*?" " [19]. To answer this challenge, we suggest an optimization of the K2PC (extension of K2 algorithm). Our proposal is to use the K2PC algorithm in conjunction with the SA algorithm to resolve the node ordering issue. To validate our proposal, we conduct experiment simulations on well-known networks.

The remainder of this paper begins with Sect. 2 to recall basic concepts. Section 3 describes a novel method for proper BN skeleton learning based on SA optimization. Section 4 describes the experimental results obtained utilizing well-known networks. Section 6 contains the conclusion and future works.

2 Theoretical Background

2.1 BN Structure Learning

Building BN can be done in two ways: (1) manually with expert assistance, or (2) automatically using learning algorithm. This later involved two stages: defining the structure and estimating the parameters. The qualitative knowledge representation is formed by structure learning, and the quantitative knowledge representation is formed by parameters learning [3–5]. It is the context of BN structure learning that interests us. It enables the explicit graphical representation [12] of the causal link between dataset variables [14]. Several BN structure-learning algorithms have been introduced in literature. These can be grouped into three broad categories: (1) *constraint-based approach*: its algorithms rely on conditional independence which involves conducting a qualitative investigation of the dependence and independence nature between variables and try to identify a linkage that reflects their relationships. In [21], the authors suggested new combination of PC algorithm and PSO, and then considered structure priors for ameliorating the PC-PSO performance. As illustrated in the experimentation the proposed approach has achieved superior results in terms of BIC scores compared to the other methods. (2) *Score-based approach*: its algorithms are used to generate a graph maximizes a given score. The score is frequently described as a metric of how well the data and graph fit together. Examples are the MWST (*Maximum Weight Spanning Tree*), GS (*Greedy Search*), and K2 algorithm. These algorithms use several scores such as BDe (Bayesian Dirichlet Equivalent) and BIC (*Bayesian Information Criterion*), and (3) *Hybrid approach*: it includes local search producing a neighborhood covering all interesting local dependencies using independence tests. Examples are MMMB (*Max Min Markov Blanket*) and MMPC (*Max Min Parents Children*). In [13], the researchers introduced new order method-based on BIC score that has been employed to generate the proper node order for K2 algorithm. This improvement allows to efficiently learn the BN topology, and the obtained results prove the performance of such combination.

Why K2 algorithms: several studies, including [1] and [22], stated that the score-based approach includes the most commonly utilized types of algorithms. K2 [8] being one of the more effective and frequently applied [16, 22]. It is a greedy search method that

is data-driven for structure learning. Determining the node ordering as input allows it to improve the learning effectiveness and to significantly reduce the computing complexity. Several works have been proposed to improve the k2 nodes ordering issue such as [13]. As described, the authors proposed an improvement of the structure learning approach by suggesting a novel method for node order learning that is based on the BIC score function. As provided in this work, the proposed method dramatically allows to minimize the node order space and to produce more effective and stable results. For this, we are interested in the K2PC algorithm.

2.2 K2PC Algorithm

The aim of the K2PC is to improve the node ordering of the K2 for parents and children search [6, 7] (cf. Fig. 1).



Fig. 1. Parents and children research spaces of K2PC algorithm.

In Fig. 1. We depict the recommended strategy for finding parents and children. Thus, the search space is separated into two sub-spaces: (1) the parents space of (Pred(Xi): Xi predecessors) and the (2) children space (Succ(Xi): Xi successors). As a result, the K2PC is divided into two search phases: those for parents (marked in blue) and those for children (marked in green). In fact, Fig. 2 illustrates the key directives of the K2PC.

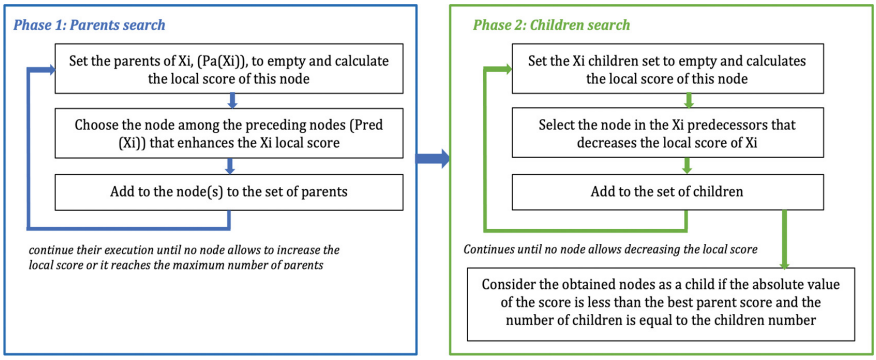


Fig. 2. The K2PC process [6].

The K2PC algorithm is presented by algorithm 1:

Algorithm 1 . K2PC algorithm.

Input: data, nodes order
Output: DAG
1.for i from 1 to N
2. for j from 1 to i−1 do
3. scorePa(i) ←calculateScore(i,j)
4. if scorePa(i) > scoreMaxPa(i) & long(Pa(i)) < nbPa) then
5. scoreMaxPa(i) ← scorePa(i)
6. Pa(i) ←j
7. end if
8. end for
9. addEdges(i, Pa(i))
10. end for
11. for i from 1 to N do
12. for j from i+1 to N do
13. scoreChild(i) ←calculateScore(i,j)
14. if scoreChild(i) < scoreMinChild(i) & long(Child(i)) < nbChild) then
15. scoreMinChild(i) ← scoreChild(i)
16.end if
17.end for
18. if |scoreMinChild(i)| < |scoreMaxPa(i)|then
19. Child(i) ←j
20. addEdges(i, Child(i))
21. end if
22. end for

Parents search

Children search

2.3 Simulated Annealing

One of the most used heuristic methods for treating optimization issues is the Simulated Annealing (SA) algorithm [17]. Based on a modeling of the metallurgical annealing process of a heated solid item, it is a random algorithm. A solid (metal) is heated to an extremely high temperature during the annealing process, allowing the atoms in the molten metal to move freely around one another. However, as the temperature decreases, the atoms' motions are restricted. It is a technique for handling significant combinatorial optimization issues. The Metropolis algorithm is applied iteratively by the SA to produce a series of configurations that tend toward thermodynamic equilibrium [20]. In the Metropolis algorithm, we start from a given configuration, and we make it undergo a random modification. If this modification reduces the objective function (or energy of the system), it is directly accepted; Otherwise, it is only accepted with a probability equal to $exp^{-\Delta E/T}$; where ΔE is the change in energy state and T is the temperature. This rule is called Metropolis criterion [18]. The flowchart presented in Fig. 3 present the SA procedure which begins with an initialization of temperature and initial solution which will be chosen either randomly or with a heuristic then, it generates a neighboring solution by testing with the temperature, if the current temperature is lower than the new temperature then the new solution is retained otherwise it will be tolerated with a probability and these steps are repeated until the maximum number of iterations is reached or the temperature becomes lower than zero.

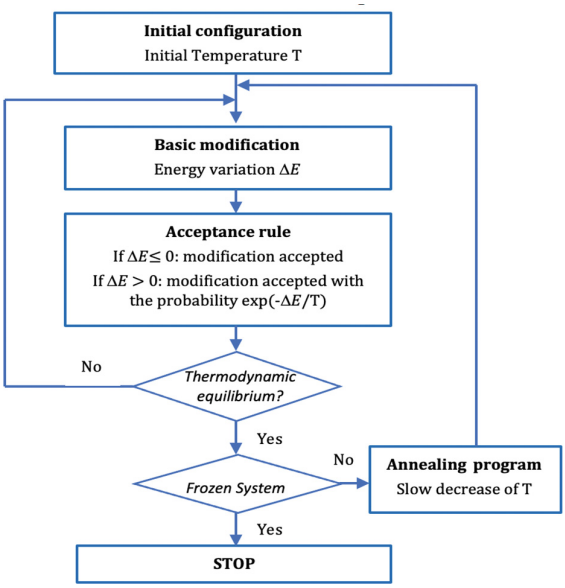


Fig. 3. SA functioning procedure [20]

SA is a popular tool for tackling numerous optimization issues. It has been introduced for providing effective optimization solution in engineering, scheduling, decision problems, etc. As a result of its versatility in modeling any type of decision variable and its high global optimization capabilities, SA is employed in this paper.

3 Proposed SA-K2PC

Our idea is to introduce an improved score-based method for building the optimal Bayesian network structure. Hence, we propose a SA optimization of the recently extended version of the widely used K2 algorithm, which is the K2PC.

As previously mentioned, K2PC algorithm have proven its efficiency compared to other existing K2 versions. However, it is highly sensitive to the order of the nodes initially entered [7]. For this reason, we think that the SA optimization can give a better specification of the K2PC nodes order to arrive at the most correct structure: we named this combination the SA-K2PC. Its steps are presented by Fig. 4.

As presented in Fig. 4, our algorithm begins with an initialization of a randomly chosen solution. Then, it uses the K2PC algorithm which searches for the structure by randomly using different input orders and calculates each time the corresponding BIC score and with the correct order found by the SA algorithm. Our new algorithm returns the correct structure and the BIC score of the original graph and the learned graph.

the simulated annealing therefore seeks the order which makes the structure learned by the K2PC algorithm, closer to the original which will be the result of our algorithm with the BIC scores, calculated using the Eq. (1):

$$Score\ BIC\ (B, D)\ =\ \log L(D|\theta^{MV}, B) - 1/2Dim(B)LogN \tag{1}$$

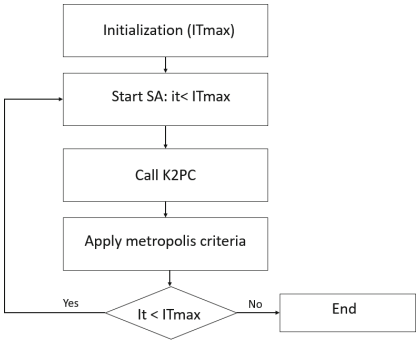


Fig. 4. SA-K2PC process.

Algorithm 3. SA-K2PC.

1. **Data:** Nit, M, NS, T, //N: maximum iteration number, NS: number of sub-iterations, T: temperature, M: model
2. **Initialization:** T=T0 initial temperature
3. S=S0 // Initial solution obtained by a random choice
4. For each iteration i from 1 to N iter Do // Find a neighboring solution S' of S
5. For it < NS Do
6. Call K2PC code “Computation of a neighboring solution; // Call K2PC code
7. If score(S') > score(S) Then
8. Accept S'; S=S';
9. Else
10. Calculate ΔTc=Tc'-Tc; //Apply Metropolis Rule
11. If P=e(-ΔT) > random0,1 then
12. Accept S'; S=S';
13. End if
14. End if
15. If score(S) > score(BestS) then
16. BestS = S;
17. End if
18. End for
19. Return BestS // return best order and best score
20. End for

Our algorithm returns as results the BIC score of the original graph, the BIC score of the new graph and the best choice of the order to generate the best learned structure as shown in the Fig. 5 for the ASIA 1000 base with number of iterations equal to 20 and number of sub-iterations equal to 10.

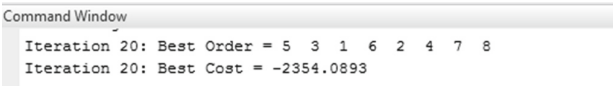


Fig. 5. Best order and score returned by proposed SA.

This algorithm also returns a representative graph of the scores compared to the different iterations executed for the same condition (cf. Fig. 6).

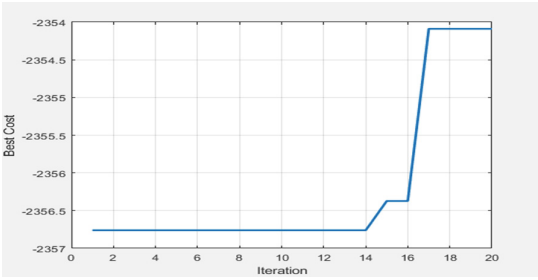


Fig. 6. Best order and score returned by the proposed SA-K2PC.

Figure 6 represents the variation of the score according to the iteration number. We notice from this figure that the score becomes maximum from iteration number 17.

4 Experimental Results and Evaluation

4.1 Used Reference Networks

For the SA-K2PC algorithm test, we will use three well-known databases (small, medium and large). Table 1 presents them.

Table 1. Used databases

Base	Number of cases	Number of nodes	Number of arcs
Asia	250/500/1000/2500/5000/100000	8	8
Alarm	250/500/1000/2500/5000/100000	37	46
Cancer	250/500/1000/2500/5000/100000	5	4

4.2 Structural Difference Based Evaluation

To evaluate the SA-K2PC performance, we test it basing on metrics visible in Tables 2 and 3.

Table 2. Used metrics for comparison

Edge	Description
RE (Reversed edges)	An edge that exists in both graphs (original and learned) but the arrow direction is reversed
CE (Correct edges)	An edge that appears in the original and learned graphs where the arrow direction is the same in the both graphs
AE (Added edges)	Is an edge that is not found in the original graph
DE (Deleted edges)	An edge that exists in the original graph but does not exist in the learned one
SD (Structural difference)	The addition of the arcs not correctly learned. It is the sum of the arcs added, reversed, and deleted

Table 3. Structural difference evaluation of SA-K2PC

Networks	Samples	CE	DE	RE	AE	SD
Cancer	*250*	2	0	2	0	2
	500	3	0	1	0	1
	1000	4	0	0	0	0
	2000	4	0	0	0	0
	3000	4	0	0	0	0
	5000	4	0	0	0	0
	10000	4	0	0	0	0
Asia	*250*	6	1	0	0	2
	500	7	1	0	0	1
	1000	4	0	4	0	4
	2000	5	0	3	0	3
	3000	6	1	1	0	2
	5000	5	0	3	0	3
	10000	6	0	3	1	4
Alarm	*250*	14	13	19	22	54
	500	17	12	17	24	53
	1000	25	6	15	21	42
	2000	14	9	23	26	58
	3000	16	9	21	23	53
	5000	17	23	23	21	52
	10000	18	9	19	24	52

For ASIA: SA-K2PC gives 4−7 CE and low SD (between 1 and 4), which can be considered as interesting evaluation results. For CANCER: for a maximum number of iterations equal to 20 and the number of sub-iterations equal to 10, there are no errors for the cases of 1000, 2000, 5000 and 10000. The structure is correctly learned and SA-K2PC is quite effective for this case. For ALARM: The results cannot be considered as the best ones since the SD is high and the number of CE can be considered as average.

Several existing research works deal with graph or structure learning for the BN [3, 14–16]. We will compare our results with these works. Table 6 presents the AE, DE, RE and CE generated by our optimized algorithm compared to those of [6, 17, 20, 22] for the ASIA and ALARM databases. Results marked in bold represent the best obtained results and those marked with a star (*) represent the second-best values (Table 4).

For ASIA, our proposal returns the better result for ASIA 5000, the same result as ITNO-K2PC for ASIA 10000 and for the other two cases our proposal returned the second-best results with low SD: thus, our proposal returned good results.

For ALARM, our proposal returns best results compared to [2]. The other results are average but are not the best ones.

4.3 Effectiveness Evaluation

A second method to test the effectiveness of the proposed SA-K2PC, is the accuracy metric, which can be calculated using the following equations:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{2}$$

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \tag{3}$$

and

$$\mathbf{F}1 = 2 * \frac{\boldsymbol{précision} * \boldsymbol{Recall}}{\boldsymbol{précision} + \boldsymbol{Recall}} \tag{4}$$

Table 5 presents the effectiveness results of the SA-K2PC compared to related works.

For ASIA, we can retain that SA-K2PC returns either the best value or the second-best value for precision, recall and F1. Therefore, we can conclude that our proposal is effective. For ALARM, we notice that SA-K2PC returns good results, but not the best in comparison with the other proposals, except [2] in some cases.

Table 4. ASIA and ALARM databases comparison for structural difference evaluation

		Asia				Alarm			
		1000	2000	5000	10000	1000	2000	5000	10000
Tabar et al. [22]	CE	4	5	5	6	38	39	41	41
	DE	0	0	0	0	2	1	1	1
	RE	4	3	3	3	8	8	8	8
	AE	0	0	1	1	4	4	7	7
	SD	4	3	4	4	14	13	16	16
Ko et al. [16]	CE	5	5	5	5	38	39	40	40
	DE	0	0	0	0	4	2	2	2
	RE	3	3	3	3	4	4	4	4
	AE	1	1	1	1	9	9	13	15
	SD	4	4	4	4	17	15	19	21
Ai [2]	CE	4	4	4	4	23	23	24	24
	DE	1	1	1	1	3	3	2	2
	RE	2	2	1	1	28	21	21	20
	AE	3	3	3	3	34	34	32	30
	SD	6	6	5	5	59	55	55	52
Benmohamed et al. [7]	CE	7	7	6	6	38	39	38	38
	DE	0	0	1	1	6	5	6	6
	RE	1	1	1	1	2	2	0	0
	AE	0	0	1	1	10	10	9	10
	SD	1	1	3	3	18	17	17	16
SA-K2PC	CE	6*	6*	**7**	**6**	25*	14	17	18
	DE	1*	19	**1**	**1**	6	9	8	9
	RE	1	1	0	1	15	23	23	19
	AE	0	0	0	1	21	26	21	24
	SD	2*	2*	1	3	42	58	52	52

Table 5. ASIA and ALARM databases comparison for effectiveness evaluation

		ASIA				ALARM			
		1000	2000	5000	10000	1000	2000	5000	10000
Tabar et al. [22]	TP	4	3	3	3	38	39	41	41
	FN	0	0	0	0	2	1	1	1
	FP	4	3	4	3	12	12	15	15
	SD	4	3	4	4	14	13	16	16
	Precision	0.5	0.623	0.555	0.555	0.75	0.765	0.732	0.732
	Recall	1	1	1	1	0.95	0.975	0.976	0.976
	F1	0.667	0.769	0.714	0.714	0.844	0.847	0.36	0.836
Ko et al. [16]	TP	4	3	3	3	38	39	40	40
	FN	0	0	0	0	4	2	2	2
	FP	5	4	4	4	13	13	17	19
	SD	5	4	4	4	17	15	19	21
	Precision	0.444	0.555	0.555	0.555	0.745	0.75	0.702	0.678
	Recall	1	1	1	1	0.9	0.951	0.952	0.952
	F1	0.615	0.714	0.714	0.714	0.815	0.839	0.808	0.792
Ai [2]	TP	4	4	4	4	23	23	24	24
	FN	1	1	1	1	3	3	2	2
	FP	5	5	4	4	56	55	52	50
	SD	6	6	5	5	59	58	55	52
	Precision	0.444	0.444	0.444	0.444	0.291	0.295	0.316	0.324
	Recall	0.8	0.8	0.8	0.8333	0.885	0.885	0.923	0.923
	F1	0.571	0.571	0.571	0.666	0.438	0.442	0.47	0.8
(Benmohamed et al. 2020) INTO-K2PC [5]	TP	7	7	6	6	38	39	38	38
	FN	0	0	1	1	6	5	6	6
	FP	1	1	2	2	12	12	11	10
	SD	1	1	3	3	18	1	17	16
	Precision	0.875	0.875	0.75	0.75	0.76	0.764	0.775	0.791
	Recall	1	1	0.857	0.857	0.864	0.886	0.864	0.864
	F1	0.933	0.933	0.8	0.8	0.806	0.82	0.817	0.826

(*continued*)

Table 5. (*continued*)

		ASIA				ALARM			
		1000	2000	5000	10000	1000	2000	5000	10000
SA-K2PC	TP	6	6	7	6	25*	14	17	18
	FN	1	1	1	1	6	9	8	9
	FP	1	1	0	2	26	49	44	43
	SD	2	2	1	3	42	58	52	52
	Precision	0.857*	0.857*	0.875	0.75	0.49	0.26	0.33	0.35
	Recall	0.857*	0.857*	0.875	0.857	0.80	0.60	0.68	0.66
	F1	0.910*	0.910*	0.875	0.799*	0.607	0.362	0.444	0.457

Our experiments showed that SA-K2PC gives very good results for small and medium databases (Cancer and Asia) and average results for the large database (Alarm)[1].

5 Conclusion

Our work essentially concerns the BN structure learning. We have chosen the K2PC algorithm considered effective in literature, but its weakness is in the sensitivity to the order given as input, hence we have chosen Simulated Annealing (SA) which is targeted to solving this problem. Our proposal consists of two phases: the first is to seek the best order at the input and the second is to learn the BN structure using the best order returned by the first phase. We tested our proposal using different evaluation methods and with three well-known networks. We concluded that the SA-K2PC combination can be considered effective, especially for small and medium databases.

In future works, we plan to improve our proposal by further optimizing our algorithm to generate better results, especially for large databases (such as HAILFINDER including 56 nodes and 66 arcs and DIABETES including 413 nodes and 602 arcs) and then applying it to real cases data.

[1] True positives (TP) indicate the number of correctly identified edges. False positives (FP) represent the number of incorrectly identified edges. False negatives (FN) refer to the number of incorrectly identified unlinked edges.

References

1. Amirkhani, H., Rahmati, M., Lucas, P.J., Hommersom, A.: Exploiting experts' knowledge for structure learning of Bayesian networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2154–2170 (2016)
2. Ai, X.: Node importance ranking of complex networks with entropy variation.". Entropy **19**(7), 303 (2017)
3. Bouazizi, S., Ltifi, H.: Improved visual analytic process under cognitive aspects. In: Barolli, L., Woungang, I., Enokido, T. (eds.) AINA 2021. LNNS, vol. 225, pp. 494–506. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-75100-5_43
4. Benjemmaa, A., Ltifi, H., Ben Ayed, M.: Multi-agent architecture for visual intelligent remote healthcare monitoring system. In: International conference on hybrid intelligent systems, pp. 211–221. Springer, Cham(2016)
5. Benjemmaa, A., Ltifi, H., Ayed, M.B.: Design of remote heart monitoring system for cardiac patients. In: Advanced information networking and applications, pp. 963–976. (2019)
6. Benmohamed, E., Ltifi, H., et Ben Ayed, M.: A novel bayesian network structure learning algorithm: best parents-children. In: 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 743–749. IEEE (2019)
7. Benmohamed, E., Ltifi, H., et Ben Ayed, M.: ITNO-K2PC: An improved K2 algorithm with information-theory-centered node ordering for structure learning. J. King Saud Univ.-Comput. Inf. Sci., (2020)
8. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks form data. Mach. Learn. **9**, 309–347 (1992)
9. Ellouzi, H., Ltifi, H., BenAyed, M.: 2015, New multi-agent architecture of visual intelligent decision support systems application in the medical field. In: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications, pp. 1–8. IEEE (2015)
10. Ltifi, H., Benmohamed, E., Kolski, C., Ben Ayed, M.: Adapted visual analytics process for intelligent decision-making: application in a medical context. Int. J. Inf. Technol. & Decis. Mak. **19**(01), 241–282 (2020)
11. Ltifi H., Ben Ayed M., Kolski, C., and Alimi, A. M.: HCI-enriched approach for DSS development: the UP/U approach. In: 2009 IEEE Symposium on Computers and Communications, pp. 895–900. IEEE (2009)
12. Ltifi, H., Ayed, M.B., Trabelsi, G., Alimi, A.M.: Using perspective wall to visualize medical data in the Intensive Care Unit. In: 2012 IEEE 12th international conference on data mining workshops, pp. 72–78. IEEE (2012)
13. Lv, Y., Miao, J., Liang, J., Chen, L., Qian, Y.: BIC-based node order learning for improving Bayesian network structure learning. Front. Comp. Sci. **15**(6), 1–14 (2021). https://doi.org/10.1007/s11704-020-0268-6
14. Huang, L., Cai, G., Yuan, H., Chen, J.: A hybrid approach for identifying the structure of a Bayesian network model. Expert Syst. Appl. **131**, 308–320 (2019)
15. Jiang, J., Wang, J., Yu, H., Xu, H.: a novel improvement on K2 algorithm via markov blanket. In: Poison identification based on Bayesian network, pp. 173–182. Springer (2013)
16. Ko, S., Kim, D.W.: An efficient node ordering method using the conditional frequency for the K2 algorithm. Pattern Recognition Lett. **40**, 80–87 (2014)
17. Kirkpatric, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science. **220** (67180), (1983)
18. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E.: Equation of state calculations by fast computing machines. Journal of Chemical Physics, 21, (1953)
19. Scutari, M., Graafland, C.E., Gutiérrez, J.M.: Who learns better Bayesian network structures: accuracy and speed of structure learning algorithms. Int. J. Approximate Reasoning **115**, 235–253 (2019)

20. Sun, Y., Wang, W., Xu, J.: An new clustering algorithm based on QPSO and simulated annealing, (2008)
21. Sun B., Zhou Y., Wang J., Zhang, W.: A new PC-PSO algorithm for Bayesian network structure learning with structure priors. Expert. Syst. Appl., **184**, 115237 (2021)
22. Tabar, V.R., Eskandari, F., Salimi, S., et al.: Finding a set of candidate parents using dependency criterion for the K2 algorithm.". Pattern Recogn. Lett. **111**, 23–29 (2018)



A Gaussian Mixture Clustering Approach Based on Extremal Optimization

Rodica Ioana Lung(✉)

Centre for the Study of Complexity, Babes-Bolyai University, Cluj Napoca, Romania
rodica.lung@ubbcluj.ro

Abstract. Many machine-learning approaches rely on maximizing the log-likelihood for parameter estimation. While for large sets of data this usually yields reasonable results, for smaller ones, this approach raises challenges related to the existence or number of optima, as well as to the appropriateness of the chosen model. In this paper, an Extremal optimization approach is proposed as an alternative to expectation maximization for the Gaussian Mixture Model, in an attempt to find parameters that better model the data than those provided by the direct maximization of the log-likelihood function. The behavior of the approach is illustrated by using numerical experiments on a set of synthetic and real-world data.

1 Introduction

Gaussian Mixture Model (GMM) is a clustering model that uses the multivariate normal distribution as representation for data clusters [19,24]. Parameters of the model are estimated by using expectation maximization, which maximizes the log-likelihood function. If there is enough data available and the normality assumptions are met, it is known that this approach yields optimal results. However, there are many situations in which the available data may not be suitable for this method, even though the gaussian mixture model may be useful in representing the clusters. For such situations deviations from the optimal value of the log-likelihood function may be beneficial, and this paper attempts to explore such possible situations.

There are many practical applications that use GMM to model data, because, if successful, it offers many theoretical advantages in further analyses. We can find examples in image analysis [11], sensor fault diagnosis [25], driving fatigue detection [2,26], environment [14,17], health [12], etc.

Solutions for the clustering problem can be evaluated by using internal quality measures for clusters [15]. An example of such an index that is often used to evaluate the performance of an algorithm is the Silhouette Score (SS) [20]. The SS combines the mean intra-cluster distance of an instance to the mean distance to the nearest cluster. Higher values indicate better cluster separation. In

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 776–785, 2023.
https://doi.org/10.1007/978-3-031-27409-1_71

many applications that use GMM, results reported indicate a higher SS value for various applications: detecting abnormal behavior in smart homes [3], aircraft trajectory recognition [13], HPC computing [4], customer churn [22], analysis of background noise in offices [7], and for a image recommender system for e-commerce [1], etc.

GMM has also been extensively used on medical applications such as to analyse COVID-19 data [10,23] with silhouette score as performance indicator; GMM models have reported best silhouette scores for medical document clustering [6] on processed data extracted from PubMed. Other applications in which GMM results are evaluated based on the SS include: manual muscle testing grades [21], insula functional parcellation [27], where it is used with an immune clonal selection algorithm, clustering of hand grasps in spinal cord injury [8], etc.

In this paper, an attempt to estimate parameters for the Gaussian mixture model by using the silhouette coefficient in the fitness evaluation process of an extremal optimization algorithm is proposed. The Gaussian mixture model assumes that clusters can be represented by using multivariate normal distribution, and its parameters consist on the mean and covariance matrices for these distributions. The standard approach to estimate parameters is to use expectation maximization (EM) by which the log-likelihood function is maximized. Instead of using EM, extremal optimization algorithm is used to evolve means and covariance matrices in order to improve the silhouette score of the clusters. To avoid local optimal solutions, a small perturbation of the data is added during search stagnation. Numerical experiments are used to illustrate the behavior of the approach.

2 Noisy Extremal Optimization—GM

The clustering problem can be expressed in the following manner: we are given a data set $\mathbb{D} \subset \mathbb{R}^{nxd}$. An element $x_i \in \mathbb{D}$ is called an instance of the data and it belongs to $\mathbb{R}^d$. Attributes, or features of the data are the column vectors X_j containing the j^{th} component of each instance, with $X_j \subset \mathbb{R}^n$. Intuitively, when clustering the data we try to find instances that are somehow grouped together, i.e. that are forming clusters of data. The criterion by which data is considered as grouped depends on the approach. Clusters are usually denoted by C_l, $l = \overline{1,k}$, where k denotes their number. k may be given a priori or it may be deduced during the search process.

2.1 Gaussian Mixture Model

The GMM model represents clusters by using the multivariate normal distribution [24]. Thus, each cluster C_i is represented by using the mean and covariance matrix and the likelihood function is used to determine the probability that an instance belongs to a cluster. The aim is to find for each cluster the mean $\mu_i \in \mathbb{R}^d$ and the covariance matrix $\Sigma_i \in \mathcal{M}_{d\times d}$ that best describes the data. The

corresponding probability density function for the entire data-set is

$$f(x) = \sum_{i=1}^{k} f(x|\mu_i, \Sigma_i) P(C_i) \tag{1}$$

where k is the number of clusters, and $P(C_i)$ are the prior probabilities or mixture parameters. The prior probabilities, as well as the mean and covariance matrices, are estimated by maximizing the log-likelihood function $P(\mathbb{D}|\theta)$ where

$$\theta = \{\mu_1, \Sigma_1, P(C_1), \ldots, \mu_k, \Sigma_k, P(C_k)\} \tag{2}$$

with $\sum_{i=1}^{k} P(C_i) = 1$ and

$$P(\mathbb{D}|\theta) = \prod_{j=1}^{n} f(x_j). \tag{3}$$

The log-likelihood function

$$\ln P(\mathbb{D}|\theta) = \sum_{j=1}^{n} \ln \left(\sum_{i=1}^{k} f(x_j|\mu_i, \Sigma_i) P(C_i) \right) \tag{4}$$

is maximized and

$$\theta^* = \arg\max_{\theta} \ln P(\mathbb{D}|\theta)$$

is used to describe data clusters. Finding θ^* is usually performed by using the *expectation maximization (EM)* approach. EM computes the posterior probabilities $P(C_i|x_j)$ of C_i given x_j as:

$$P(C_i|x_j) = \frac{f_i(x_j) P(C_i)}{\sum_{a=1}^{k} f_a(x_j) P(C_a)}. \tag{5}$$

$P(C_i|x_j)$ is denoted by w_{ij} and is considered the weight, or contribution of point x_j to cluster C_i. w_{ij} is the probability used to assign instance x_j to cluster C_i.

The EM algorithm consists of three steps: initialization, expectation, and maximization, which are succinctly described in what follows as the EO approach is based on them:

(*i*) the means μ_i for each cluster C_i are randomly initialized by using a uniform distribution over each dimension X_a. Covariance matrices are initialized with the identity matrix, and $P(C_i) = \frac{1}{k}$.

(*ii*) In the expectation step posterior probabilities/weights $w_{ij} = P(C_i|x_j)$ are computed using Eq. (5).

(*iii*) In the maximization step, model parameters μ_i, Σ_i, $P(C_i)$ are re-estimated by using posterior probabilities (w_{ij}) as weights. The mean μ_i for cluster C_i is estimated as:

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} w_{ij}}, \tag{6}$$

and the covariance matrix Σ_i of C_i is updated using:

$$\Sigma_i = \frac{\sum_{j=1}^{n} w_{ij} \overline{x}_{ji} \overline{x}_{ji}^T}{\sum_{j=1}^{n} w_{ij}}, \tag{7}$$

where $\overline{x_{ji}} = x_j - \mu_i$.

The prior probability for each cluster $P(C_i)$ is computed as:

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n}. \tag{8}$$

The expectation (ii) and maximization (iii) steps are repeated until there are no differences between means updated from one step to the other. Predictions are made based on posterior probabilities w_{ij}.

2.2 Noisy Extremal Optimization Gaussian Mixture

Extremal optimization—EO is a stochastic search method based on the Bak-Sneppen model of self-organized criticality [5,16]. EO is suitable for problems in which the solution can be represented by components with individual fitness values. Its goal is to find an optimal configuration by randomly updating the worst component of the current solution (Algorithm 1).

Thus, to use EO, we need to define the search domain, and subsequently the solution encoding, the objective function f, and fitness functions f_i to evaluate each component of the solution. Within nEO-GM the EO is used to search for the optimal positions of the clusters' means and covariance matrices. The posterior probabilities are computed in the same manner as EM. Thus, an individual s encodes:

$$s = \{\mu_1, \Sigma_1, \ldots, \mu_k, \Sigma_k\},$$

with $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d \times d}$. Matrices Σ_i have to be symmetric and positive semi-definite. Each mean and covariance matrix μ_i and Σ_i characterize a component C_i, $i = 1 \ldots, k$, so we can also write $s = (C_1, \ldots, C_k)$.

Algorithm 1 Outline of general EO

1: **Input:** Search domain D, objective function f, component fitness functions f_i, $i = 1, \ldots, k$;
2: Randomly generate potential solution s;
3: Set $s_{best} = s$;
4: **for** a number of iterations **do**
5: evaluate $f_i(s)$, $i = 1, \ldots, k$;
6: find component s_i with the worst fitness;
7: replace s_i with a random value;
8: **if** $f(s) > f(s_{best})$ **then**
9: Replace $s_{best} = s$;
10: **end if**
11: **end for**
12: **Output:** s_{best}.

The initialization (line 2, Algorithm 1) is performed with parameters estimated using EM, as there is no reason not to start the search with a good solution. The only drawback of this approach is that the EO may not be able to deviate from this solution.

The fitness $f_i(C_i)$ of each component is computed as using the average intracluster distance of C_i. Thus, in each EO iteration, the cluster having the highest intracluster distance is randomly modified by altering the mean and covariance matrix. The overall objective function $f(s)$ to be maximized is the silhouette coefficient SS computed as follows: for each point x_i the silhouette coefficient sc_i based on configuration $s = (C_1, \ldots, C_k)$ is:

$$sc_i(s) = \frac{\nu_{out}^{\min}(x_i) - \nu_{in}(x_i)}{\max\{\nu_{out}^{\min}(x_i), \nu_{in}(x_i)\}}, \tag{9}$$

where $\nu_{out}^{\min}$ is the mean distance from x_i to all points in the closest cluster:

$$\nu_{out}^{\min} = \min_{j \neq \hat{y}_i} \frac{\sum_{x \in C_j} ||x_i - x||}{n_j} \tag{10}$$

and n_j is the size of cluster C_j. $\nu_{in}(x_i)$ is the mean distance from x_i to points in its own cluster $\hat{y}_i$:

$$\nu_{in}(x_i) = \frac{\sum_{x \in C_{\hat{y}_i}, x \neq x_i} ||x_i - x||}{n_{\hat{y}_i} - 1}. \tag{11}$$

For an instance x_i, $sc_i \in [-1, 1]$; a value closer to 1 indicates that x_i is much closer to other instances within the same cluster than to those in closest one. A value close to 0 indicates that x_i may lay somewhere at the boundaries of two clusters. A value closer to -1 indicates that x_i is closer to another cluster, so it may be miss-clustered. The silhouette coefficient SS averages sc_i values across all instances:

$$SS(s) = \frac{1}{n} \sum_{i=1}^{n} sc_i(s). \tag{12}$$

2.3 Noise

In order to increase the diversity of the search, considering that there is only one configuration s, and to avoid premature convergence, whenever there are signs that the search stagnates a small perturbation is induced in the data by adding a small noise randomly generated following a normal distribution with mean zero and a small standard deviation σ. This noise mechanism is triggered with a probability equal to the number of iteration no change has taken place (line 8, Algorithm 1) divided by a number—parameter of the method. The search on the modified data set takes place for a small number of iteration, after which the data set is restored.

3 Numerical Experiments

Numerical experiments are performed on a set of synthetic and real-world data sets. The synthetic data sets are generated by using the `make_classification` function from the `sklearn` package in `Python` [18]. The real world data sets used are presented in Table 1. nEO-GM reports the SS score of the best solution and its value is compared with the corresponding score of the solution found by EM on the same data set. As external indicator, the NMI is used to compare the clusters reported by the algorithms with those that are considered as 'real' ones. For each data-set 10 independent runs of nEO-GM are performed. Statistical significance of differences in results for both SS and NMI scores is evaluated by using a t-test.

Table 1. Real world data-sets and their characteristics, all available on the UCI machine learning repository [9].

No.	Name	Instances	Attributes	Classes
1	Cryoptherapy	90	6	2
2	Cervical	72	19	2
3	Immunotherapy	90	7	2
4	Plrx	182	12	2
5	Transfusions	748	4	2
6	Forest	243	12	2

Table 2 presents the characteristics and results reported for the synthetic data-sets. The class separator parameter (on the columns) with values 1, 2, and 5, controls the overlapping of clusters in the data-set. Figure 1 illustrates the effect of this parameter on a data set with 500 instances and 2 attributes. We find nEO-GM to be more efficient for the more difficult data sets, with many identical results to EM for the well separated data.

Results reported on the real-world data sets are also compared with other two standard clustering methods: K-means and Birch [18,24]. Table 3 presents the result of the t-test comparing SS values; an * indicates that nEO-GM NMI value is significantly better. The table also presents the SS value for the 'real' clustering structure, and we find that in some situations this value is actually negative. In the same situations we find that while SS values of nEO-GM are significantly worse than those of other methods, the NMI values are significantly better, indicating the potential of using the intra-cluster density as the fitness of components during the search of identifying the underlying data structure.

Table 2. Numerical results reported on the synthetic data-sets. p-values of the t-test comparing SS values reported by nEO-GM compared with the baseline EM results. A line indicates no difference in the numerical results. An (*) indicates significant difference in NMI values.

Instances	Attributes	k	1	2	5
100	3	3	2.330475e-02	0.011942	–
	6	6	1.839282e-02*	0.026713	–
	9	9	1.302159e-01	0.071069	–
200	3	3	1.997630e-03	0.044480*	–
	6	6	7.605365e-03	0.018128	–
	9	9	6.253126e-02	0.130290*	–
500	3	3	5.950287e-04	0.013106	–
	6	6	3.840295e-04	0.001085	0.171718
	9	9	1.077090e-03*	0.006823	–
1000	3	3	5.854717e-09	0.026761*	0.101399
	6	6	1.557011e-04	0.000696	0.027319*
	9	9	1.424818e-04	0.000011	–

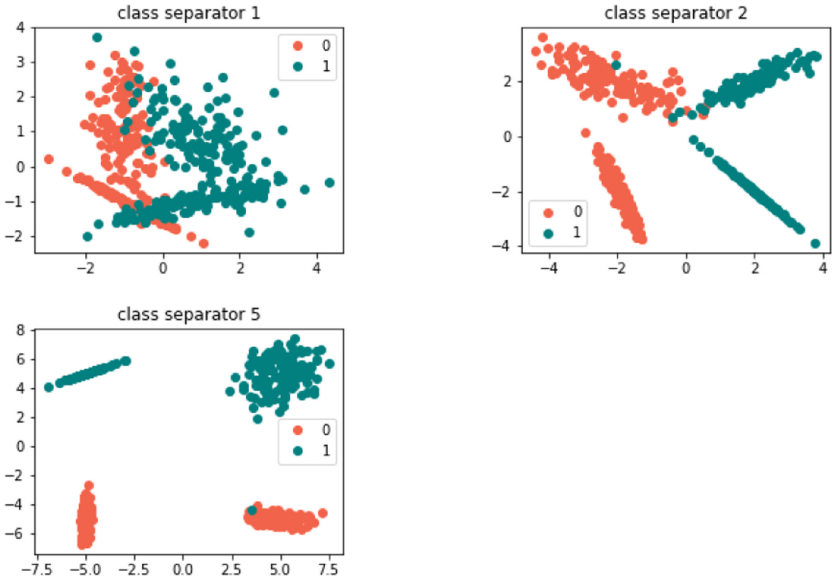


Fig. 1. Example of data generated with different class separator values, controlling the overlap of the clusters.

Table 3. Results reported for the real-world data sets. p values resulted from the t-test comparing SS values reported by nEO-GM with three other methods are presented. An * indicates significant differences in NMI values also. The column SS reports the value of the SS indicator computed on the 'real' cluster structure of the data.

	Data	SS	EM	K-means	Birch
1	Cryoptherapy	0.072783	0.782437	0.166044	0.166044
2	Cervical	0.160617	0.363157	0.750102*	0.249712
3	Immunotherapy	–0.121060	0.171718	0.999996*	0.999996*
4	Plrx	–0.013145	0.000593*	0.103138 *	0.008488 *
5	Transfusions	0.178546	0.171718	1.000000*	1.000000 *
6	Forest	0.237199	0.000571	1.000000*	1.000000*

4 Conclusions

An extremal optimization for estimating the parameters of a Gaussian mixture model is presented. The method evolves the means and covariance matrices of clusters by maximizing the silhouette coefficient, and minimizing the intra-cluster distance. A simple diversity preserving mechanism consisting of inducing a noise in the data for small periods of time is used to enhance the search. Results indicate that this approach may better identify overlapping clusters. Further work may include mechanisms for including the number of clusters into the search of the algorithm.

Acknowledgments. This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS—UEFISCDI, project number PN-III-P4-ID-PCE-2020-2360, within PNCDI III.

References

1. Addagarla, S., Amalanathan, A.: Probabilistic unsupervised machine learning approach for a similar image recommender system for E-commerce. Symmetry **12**(11), 1–17 (2020)
2. Ansari, S., Du, H., Naghdy, F., Stirling, D.: Automatic driver cognitive fatigue detection based on upper body posture variations. Expert Syst. Appl. 203 (2022). https://doi.org/10.1016/j.eswa.2022.117568
3. Bala Suresh, P., Nalinadevi, K.: Abnormal behaviour detection in smart home environments. In: Lecture Notes on Data Engineering and Communications Technologies, vol. 96, p. 300 (2022). https://doi.org/10.1007/978-981-16-7167-8_22
4. Bang, J., Kim, C., Wu, K., Sim, A., Byna, S., Kim, S., Eom, H.: HPC workload characterization using feature selection and clustering, pp. 33–40 (2020). https://doi.org/10.1145/3391812.3396270

5. Boettcher, S., Percus, A.G.: Optimization with extremal dynamics. Phys. Rev. Lett. **86**, 5211–5214 (2001)
6. Davagdorj, K., Wang, L., Li, M., Pham, V.H., Ryu, K., Theera-Umpon, N.: Discovering thematically coherent biomedical documents using contextualized bidirectional encoder representations from transformers-based clustering. Int. J. Environ. Res. Publ. Health 19(10) (2022). https://doi.org/10.3390/ijerph19105893
7. De Salvio, D., D'Orazio, D., Garai, M.: Unsupervised analysis of background noise sources in active offices. J. Acoust. Soc. Am. **149**(6), 4049–4060 (2021)
8. Dousty, M., Zariffa, J.: Towards clustering hand grasps of individuals with spinal cord injury in egocentric video, pp. 2151–2154 (2020). https://doi.org/10.1109/EMBC44109.2020.9175918
9. Dua, D., Graff, C.: UCI machine learning repository (2017). https://www.archive.ics.uci.edu/ml
10. Greenwood, D., Taverner, T., Adderley, N., Price, M., Gokhale, K., Sainsbury, C., Gallier, S., Welch, C., Sapey, E., Murray, D., Fanning, H., Ball, S., Nirantharakumar, K., Croft, W., Moss, P.: Machine learning of COVID-19 clinical data identifies population structures with therapeutic potential. iScience 25(7) (2022). https://doi.org/10.1016/j.isci.2022.104480
11. Guo, J., Chen, H., Shen, Z., Wang, Z.: Image denoising based on global image similar patches searching and HOSVD to patches tensor. EURASIP J. Adv. Signal Process. 2022(1) (2022). https://doi.org/10.1186/s13634-021-00798-4
12. He, M., Guo, W.: An integrated approach for bearing health indicator and stage division using improved gaussian mixture model and confidence value. IEEE Trans. Ind. Inform. **18**(8), 5219–5230 (2022). https://doi.org/10.1109/TII.2021.3123060
13. Kamsing, P., Torteeka, P., Yooyen, S., Yenpiem, S., Delahaye, D., Notry, P., Phisannupawong, T., Channumsin, S.: Aircraft trajectory recognition via statistical analysis clustering for Suvarnabhumi International Airport, pp. 290–297 (2020). https://doi.org/10.23919/ICACT48636.2020.9061368
14. Kwon, S., Seo, I., Noh, H., Kim, B.: Hyperspectral retrievals of suspended sediment using cluster-based machine learning regression in shallow waters. Sci. Total Environ. **833** (2022). https://doi.org/10.1016/j.scitotenv.2022.155168
15. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: 2010 IEEE International Conference on Data Mining, pp. 911–916 (2010). https://doi.org/10.1109/ICDM.2010.35
16. Lu, Y., Chen, Y., Chen, M., Chen, P., Zeng, G.: Extremal Optimization: fundamentals, Algorithms, and Applications. CRC Press (2018). https://www.books.google.ro/books?id=3jH3DwAAQBAJ
17. Malinowski, M., Povinelli, R.: Using smart meters to learn water customer behavior. IEEE Trans. Eng. Manag. **69**(3), 729–741 (2022). https://doi.org/10.1109/TEM.2020.2995529
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
19. Poggio, T., Smale, S.: The mathematics of learning: dealing with data. Not. Am. Math. Soc. **50**, 2003 (2003)
20. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Computat. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7. https://www.sciencedirect.com/science/article/pii/0377042787901257

21. Saranya, S., Poonguzhali, S., Karunakaran, S.: Gaussian mixture model based clustering of Manual muscle testing grades using surface Electromyogram signals. Physical and Engineering Sciences in Medicine **43**(3), 837–847 (2020). https://doi.org/10.1007/s13246-020-00880-5
22. Vakeel, A., Vantari, N., Reddy, S., Muthyapu, R., Chavan, A.: Machine learning models for predicting and clustering customer churn based on boosting algorithms and gaussian mixture model (2022). https://doi.org/10.1109/ICONAT53423.2022.9725957
23. Wisesty, U., Mengko, T.: Comparison of dimensionality reduction and clustering methods for SARS-CoV-2 genome. Bull. Electr. Eng. Inform. 10(4), 2170–2180 (2021). https://doi.org/10.11591/EEI.V10I4.2803
24. Zaki, M.J., Meira Jr., W.: Data Mining and Machine Learning: fundamental Concepts and Algorithms, 2 edn. Cambridge University Press (2020). https://doi.org/10.1017/9781108564175
25. Zhang, B., Yan, X., Liu, G., Fan, K.: Multi-source fault diagnosis of chiller plant sensors based on an improved ensemble empirical mode decomposition gaussian mixture model. Energy Rep. **8**, 2831–2842 (2022). https://doi.org/10.1016/j.egyr.2022.01.179
26. Zhang, J., Lu, H., Sun, J.: Improved driver clustering framework by considering the variability of driving behaviors across traffic operation conditions. J. Transp. Eng. Part A: Syst. 148(7) (2022). https://doi.org/10.1061/JTEPBS.0000686
27. Zhao, X.W., Ji, J.Z., Yao, Y.: Insula functional parcellation by searching Gaussian mixture model (GMM) using immune clonal selection (ICS) algorithm. Zhejiang Daxue Xuebao (Gongxue Ban)/J. Zhejiang Univ. (Eng Sci) 51(12), 2320–2331 (2017). https://doi.org/10.3785/j.issn.1008-973X.2017.12.003

Assessing the Performance of Hospital Waste Management in Tunisia Using a Fuzzy-Based Approach OWA and TOPSIS During COVID-19 Pandemic

Zaineb Abdellaoui(✉), Mouna Derbel, and Ahmed Ghorbel

University of Sfax, Sfax, Tunisia
Czainebabdellaoui14@gmail.com, mouna.derbel@ihecs.usf.tn, ahmed.ghorbel@fsegs.usf.tn

Abstract. Health Care Waste Management (HCWM) and integrated documentation in this hospital sector require analysis of large data collected by hospital health experts. This study presented a quantitative software index for evaluating the performance of waste management processes in healthcare by integrating Multiple Criteria Decision Making (MCDM) techniques based on ontology and fuzzy modeling combined with data mining. The HCWM index is calculated using fuzzy Ordered Weighted Average (fuzzy OWA) and the fuzzy Technique for the Order of Preference by Similarity of Ideal Solution (fuzzy TOPSIS) methods. The proposed approach is applied on a set of 16 hospitals from Tunisia. Results showed that the proposed index permit to determine weak and strong characteristics in waste management processes. A comparative analysis is made between two periods: before and during COVID-19 pandemic.

Keywords: Health care waste · Performance index · Fuzzy OWA and TOPSIS · Multiple criteria decision making · COVID-19

1 Introduction

Nowadays, as in all other organizations, the amount of waste generated in healthcare facilities is increasing due to the extent of their services. HCWM is a common problem in developing countries, including Tunisia, which are increasingly aware that healthcare waste requires special treatment. As a result, one of the most important problems encountered in Tunis is the disposal of Health Care Waste (HCW) from health facilities. The evaluation of HCW disposal alternatives, which takes into account the need to reconcile several conflicting criteria with the participation of an expert group, is a very important multi-criteria group decision-making problem. The inherent imprecision of the criteria values for HCW disposal alternatives justifies the use of fuzzy set theory. Indeed, the treatment and management of HCW is one of the fastest growing segments of the waste management industry.

Due to the rapid spread of the Human Immunodeficiency Virus (HIV) and other contagious diseases, safe and effective treatment and disposal of HCW became a major

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 786–803, 2023.
https://doi.org/10.1007/978-3-031-27409-1_72

public health and environmental problem. For a HCWM system to be sustainable, it must be environmentally efficient, economically affordable and socially acceptable [1]. The evaluation of HCW disposal alternatives, which takes into account the need to reconcile several conflicting criteria with the inherent vagueness and imprecision, is a matter of decision-making problem.

Classical MCDM methods that take into account deterministic or random processes cannot effectively deal with decision-making problems, including imprecise and linguistic information. Additionally, when a large number of performances attributes need to be considered in the assessment process, it is usually best to structure them in a multi-level hierarchy in order to conduct a more efficient analysis.

The rest of the paper is structured as follows. Sect. 2 gives an overview of the related works that treated the performance of hospital waste management using MCDM methods. Sect. 3 presents the proposed approach. An application of it on a set of hospitals is showed in Sect. 4. Then, the obtained results are analyzed and discussed in Sect. 5. Finally, Sect. 6 concludes the research.

2 Related Works

In literature, several studies focused on observing one or a few influencing criteria to describe the state of hospital waste management before COVID-19 [2–7]. For example, researchers in [5] conducted a situational analysis of the production and management of waste generated in a small hospital in the interior of the state of Ceará in Brazil. The authors checked that waste improperly disposed of in accordance with current regulations. They concluded that there is a need to educate and train professionals who handle and dispose of medical waste. Additionally, in other work [2], the authors conducted a cross-sectional comparative study to determine variations and similarities in the activities of clinical waste management practices in three district hospitals located in Johor, Perak and Kelantan. Compliance with medical waste management standards in community health care centers in Tabriz, northwestern Iran, are examined in [7] using a triangulated cross-sectional study (qualitative and quantitative). The data collection tool was a valid waste management process checklist developed based on Iranian medical waste management standards.

COVID-19 waste can play a critical role in the spread of nosocomial infections. However, several safety aspects must necessarily follow as part of the overall management of COVID-19 waste [8]. Indeed, studies conducted in Brazil, Greece, India, Iran and Pakistan have revealed that a significant prevalence of viral infection in waste collectors (biomedical/solid) can be directly attribute to pathogens in the contaminated waste [9–11].

Treatment and management of HCW is one of the fastest growing segments of the waste management industry. Due to the rapid spread of the HIV and other contagious diseases, safe and effective treatment and disposal of healthcare waste has become an important public and environmental health issue. In the literature, there are only a few analytical studies on the HCWM. Most of the time, the health facilities generating the waste are surveyed by means of prepared questionnaires, field research and interviews with staff. Some of the most common treatment and disposal methods used in the management of infectious HCW in developing countries are presented in [12]. Therefore,

classical MCDM techniques such as the Analytical Hierarchy Process (AHP) have been applied to numerous case studies to evaluate techniques used in hospital waste management [13–16]. Researchers in [13] integrated the AHP with other systemic approaches to establish first-line health care waste management systems that minimize the risk of infection in developing countries. The opinion of five Deputy Ministers is used by [17] to determine the weight of six criteria for waste management and to set out a hierarchy of methods. Hospital waste disposal methods are categorized using the fuzzy AHP and Technique for the Order of Preference by Similarity of Ideal Solution (TOPSIS) models [18]. Likewise, in [14], the AHP model is used to determine the pollution rate of hospitals in Usuzestan, Iran. They evaluated 16 hospitals with 18 criteria. The authors proposed research projects to evaluate the application of MCDM models in other scientific fields (such as water resources research).

This article presents a fuzzy multi-criteria group decision-making framework based on the principles of fuzzy measurement and fuzzy integral for the evaluation of treatment alternatives for HCWM in Tunisia, which makes it possible to incorporate imprecise data represented as linguistic variables in the analysis. For this reason, we aim to introduce two quantitative indicators over two different time periods to assess how to optimize and manage data from hospital processes in a big data environment. Any well-developed index should take into account two steps: first, select the appropriate criteria and weight them. Second, choose an appropriate algorithm by which all the evaluated information obtained from the criteria will be expressed as a unit number. We present the methodology for calculating the HCWM index before and during COVID-19 periods using the fuzzy OWA and TOPSIS methods.

3 The Proposed Approach

3.1 Fuzzy Ordered Weighted Average (OWA) Approach

This operator is in fact a weighting average where the weights of the criteria (b_j) are arranged in descending order before being multiplied by weights of order (W_j). This leads the model to become nonlinear. Indeed, the fuzzy OWA method refers to the mapping of an n-dimensional space on a one-dimensional space in which, according to Eq. 1, there is a vector of weight depending on w_j:

$$F_{OWA} : R^n \rightarrow R$$

$$F_i(r_{i1}, r_{i2}, \ldots, r_{in}) = \sum\nolimits_{j=1}^{n} W_j b_j = W_1 b_1 + W_2 b_2 + \cdots + W_n b_n \tag{1}$$

where bj is the jth large value of the input data set $\{j\}$.

In fact, the vector b indicates the decreasing ordered values of the vector a, which are indeed the weight of a criterion from the point of view of each Decision Making (DM). In this equation, n represents the number of DMs. W_j shows the order of the weights under the following condition in Eq. 2.

$$\sum\nolimits_{j=1}^{n} W_j = 1, W_j \geq 0 \tag{2}$$

The OWA method has a great variety through the different selections of order weights [19]. The order weights depend on the degree of optimism of the DM. The higher the weights at the start of the vector, the greater the degree of optimism. The degree of optimism θ is defined by [20] as presented in Eq. 3.

$$\theta = \frac{1}{n-1}\sum_{j=1}^{n}(n-j)w_j \tag{3}$$

where, n is the number of criteria. The value varies from zero to one. In addition, it can be set in three modes as shown in Fig. 1.

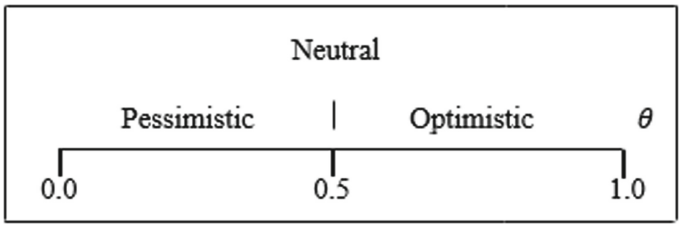


Fig. 1. Different statuses for an optimistic degree θ.

3.2 Fuzzy Technique for the Order of Preference by Similarity of Ideal Solution (TOPSIS) Approach

The TOPSIS method is proposed by [21]. The objective of this method is to choose an alternative, among a set of alternatives, which has on the one hand, the shortest distance to the ideal alternative (the best alternative on all the criteria), and, on the other hand part, which has the greatest distance to the ideal negative alternative (the one which degrades all the criteria). To do this, the TOPSIS method aims, firstly, to reduce the number of disambiguation scenarios by discarding the dominated scenarios and, secondly, to classify the effective scenarios according to their calculated overall scores.

This method measures the distance of an alternative to ideal (Si^*) and non-ideal (Si^-) solutions. In TOPSIS, a MCDM problem with m alternatives and n criteria is expressed as a matrix as follows (see Eq. (4)).

$$G = \begin{matrix} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} \begin{bmatrix} C_1 & C_2 & \cdots & C_n \\ G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ . & . & & . \\ . & . & \cdots & . \\ . & . & & . \\ G_{m1} & G_{m2} & \cdots & G_{mn} \end{bmatrix} \tag{4}$$

In this matrix, (A_1, A_2,…, A_m) are the executable alternatives, (C_1, C_2,…, C_n) are the criteria, G_{ij} is the performance of the alternative A_i from the point of view of the

criterion C_j, and W_j is the weight of the criterion C_j (see Eq. (5)):

$$W = [W_1, W_2, \ldots, W_n] \tag{5}$$

In this method, the scores of the alternatives are calculated according to the following steps:

Step 1: If the value of an alternative from a criteria point of view is defined with the matrix $x_{ij} = [X_{ij}]$, the performance matrix is first normalized in Eq. (6).

$$\mathrm{A} = [a_{ij}] = x_{ij} / \sqrt{\sum_{i=1}^{m} {x_{ij}}^2} \tag{6}$$

In the equation above, *Gij* is defined as x_{ij}. Then, the vector of the group weights of the criteria is multiplied by the matrix A of a_{ij} in order to determine the performance value (V_{ij}) of each alternative according to the following Eq. (7):

$$V_{ij} = \left[W_j . a_{ij}\right], (i = 1, 2, \ldots\ldots, m; j = 1, 2, \ldots\ldots, n) \tag{7}$$

Step 2: The distance of each alternative from the ideal and non-ideal performance values is calculated by the Eqs. 8 and 9, as follows:

$$S_{i^*} = \sqrt{\sum_{j=1}^{n} (v_{ij} - v^*_j)^2} \tag{8}$$

$$S_{i^-} = \sqrt{\sum_{j=1}^{n} (v_{ij} - v^-_j)^2} \tag{9}$$

In the equations above, v_{j^*} is the ideal performance and v_{j^-} is the non-ideal performance. Also, S_{j^*} is the ideal performance distance and S_{j^-} is the non-ideal performance distance.

Step 3: The top-down ranking of the alternatives is done on the basis of the proximity value of the *i*th alternative to the ideal solution in Eq. (10):

$$F_i = {S^-}_i / \left({S^*}_i + {S^-}_i\right), 0 < F_i < 1 \tag{10}$$

Step 4: For a better comparison between alternatives, each value F_i will be multiplied by 100.

3.3 HCWM Organization Flowchart

The flowchart for the extended waste management index is shown in Fig. 2.

Usually, several criteria are affected to assess the state of HCWM and resolve its DM issues. Moreover, each criterion possesses a specific weight and is confronted with certain ambiguities.

Foremost, as shown in Fig. 2, the appropriate criteria must be determined. These latter are taken from hospital health inspection checklists. Afterwards, a suitable number

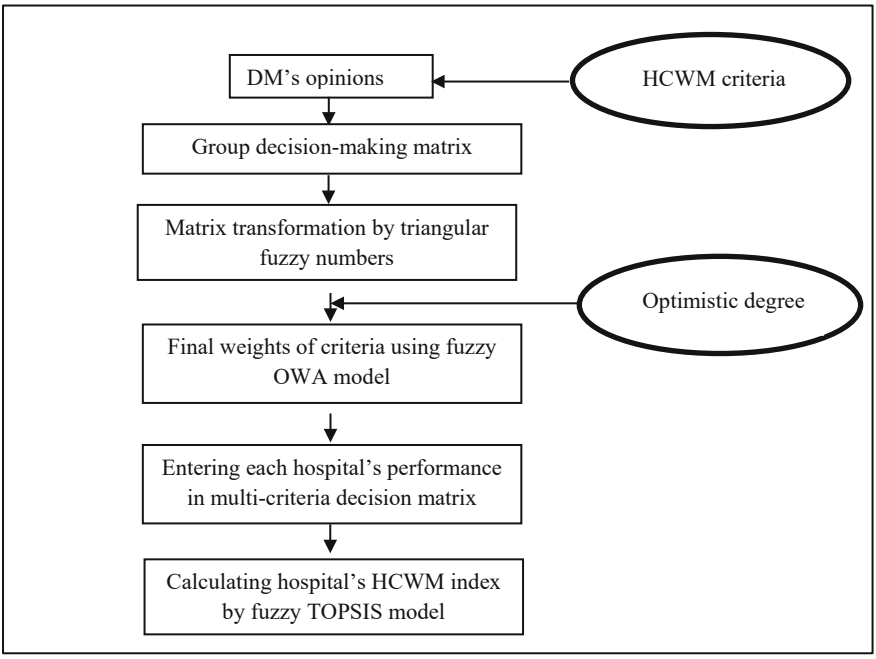


Fig. 2. Description of the procedure to be followed for calculating the HCWM index.

of stakeholders are selected as the DM of the model and experts in sanitary waste management. The position of each criterion is checked by the experts. The different linguistic terms using by the stakeholders are presented in Table 1. Indeed, this table showed the value of each stakeholder's opinion on the importance of the criteria. Since the power of DMs was defined in linguistics, Table 1 must be used to convert them into obscure numbers and use them in the model.

In the next step, since the power of DMs was defined in linguistics, Table 1 must be used to convert them into obscure numbers and use them in the model by the fuzzy OWA operator. This operator calculates the criterion weights used in the HCWM index. The fuzzy OWA is also seen as a metric in the hard process. This element is the optimistic degree of DM(θ).

In the last step, event logs are entered in the classic or fuzzy TOPSIS to calculate the value of the HCWM index for a hospital. It should be noted that these incident logs are the data collected from the observation of health experts in the waste management process and represent a hospital's performance based on each criterion.

Furthermore, we assumed that the index of 50 was the middle edge of the judgment. Nevertheless, this index number depend on specific laws and it is not a general index. Depending on the types of data in the event log, there is two folds: the first is the use of the classic TOPSIS if all performance values in the checklist are defined as exact numbers and the second is the use of the fuzzy TOPSIS as an index calculation if one or more performances cannot define as an absolute number (uncertain hospital performance). In addition, uncertain performance values can be entered into the model by means of

linguistic terms as shown in Table 2, or by triangular or trapezoidal fuzzy numbers. The linguistic weight and the equal fuzzy number are extracted from [22].

Table 1. Linguistic terms for the weight of criteria and their equivalent fuzzy number in fuzzy OWA.

Linguistic weight	Label	Equal fuzzy number
Very low	VL	(0,0,0.1)
Low	L	(0, 0.1, 0.3)
Slightly low	SL	(0.1, 0.3, 0.5)
Medium	M	(0.3, 0.5, 0.7)
Slightly high	SH	(0.5, 0.7, 0.9)
High	H	(0.7, 0.9, 1)
Very high	VH	(0.9, 1, 1)

Table 2. Linguistic terms and their equivalent fuzzy number for uncertain hospital performance in fuzzy TOPSIS.

Linguistic weight	Label	Equal fuzzy number
Very low	VL	(0,0,1)
Low	L	(0,0.1,3)
Slightly low	SL	(1,3,5)
Medium	M	(3,5,7)
Slightly high	SH	(5,7,9)
High	H	(7,9,10)
Very high	VH	(9,10,10)

4 Case Study

To illustrate the role of indicators for the treatment of medical waste, we conducted a study of 16 hospitals in Tunisia in 2020. We practically divide the work into three parts. On the one hand, we select the relevant criteria and weigh them using the OWA model. On the other hand, we calculate the HCWM index with the TOPSIS model in two different time periods.

4.1 Selection of Criteria

Thirty criteria are selected to develop the HCWM index and presented in Table 3. These criteria are taken from the hospital health inspection checklist approved by the Tunisian Ministry of Health.

Table 3. List of criteria used in the HCWM index for waste management in healthcare.

Criteria	Title
C1	Implementation of an HCWM operational program
C2	Access the list of types and locations generated by each health worker
C3	HCWM separation stations
C4	Use of yellow bags/boxes for the collection and storage of infectious waste
C5	Use of white/brown bags and boxes for the collection and storage of chemical or pharmaceutical waste
C6	Use of a safety box for needles and sharps waste
C7	Separation of radioactive waste under the supervision of a health physicist
C8	Use of black bags/boxes for the collection and storage of domestic waste in the hospital
C9	State of the bins and whether they comply with sanitary conditions
C10	Measures to get rid/release human body parts and tissues
C11	HCM Collection Frequency
C12	Labeling of bags and boxes
C13	Washing and disinfection of garbage cans after each discharge
C14	The existence of appropriate places to wash and disinfect the bins
C15	Convenient referral facilities for healthcare workers
C16	Wash and sterilize bypass facilities after each emptying
C17	Monitor prohibition of recycling of HCW
C18	The appropriate location of the temporary maintenance station
C19	Conditions for constructing a temporary maintenance station
C20	The sanitary conditions of the temporary maintenance station
C21	Development of temporary maintenance station equipment
C22	Separation of healthcare workers at a temporary maintenance station
C23	Daily weighting and documentation for HCWM
C24	Use of steam sterilization facilities
C25	Delivery of sterilized and domestic waste to the municipality
C26	Use acceptable methods to dispose of chemical waste
C27	Location of HCW Sterilization Facilities
C28	Neutralize HCWM documents and folders
C29	Terms of appointed personnel in the HCWs section
C30	Availability of equipment and facilities for personnel named in the HCWMs section

4.2 Final Weight of the Criteria

The final weight of the criteria (before and during COVID-19) was calculated by R software as shown in the Table 4. Usually there is an increase in weights indicating the awareness of decision makers of the importance of better waste management and that COVID-19 in changing attitudes towards waste.

Criterion C8 has the highest weight before COVID-19, indicating that this criterion is the best result compared to the other criteria. Although the index rose during the COVID-19 period, this criterion is placed in 5th place, which is the loss of the 4 limits. C10, C6 and C4 also have the highest weight values. However, C17, C20 and C5 are the least significant. Criteria such as C12, C9, C11, C3, C1, C26, C25, C16, C28 and C19 have medium or slightly high weights.

However, the results presented during the Corona virus are completely different from the results presented before the Corona virus as almost all criteria show an increase due to the disaster situation caused by COVID-19. The criteria C23, C6, C4, C10, C8, C11, C3, C1, C9 and C4 even have the highest weight values. In contrast, C7 and C21 have the least importance of the criteria, while the weights of criteria C19, C12, C16, C2, C13 and C29 are on average to slightly higher.

On the other hand, criterion 8 became the first with a score of 0,432144 for COVID-19. On COVID-19, on the other hand, this benchmark has been out of range, rising by 0,56922.

Furthermore, the benchmark 23 was among the last ranges before COVID-19 was in the range of 26, but during the COVID-19 its weight increased by 0,71159 and became the first on COVID-19. Indeed, there is a sharp increase in criterion 23 Daily weighing and documentation for HCWM during COVID-19 due to the growing awareness of waste management professionals in waste management.

On the one hand, the weights found for the 30 criteria will be used as follows to determine the waste mountain score for each hospital. On the other hand, the group weights of the parameters indicate the intensity of the impact of each parameter on the overall healthy waste management. Determining this measure ensures the rationality of physicians' attitudes and validates the use of each criterion of the waste management index. This is considered an aspect of the accuracy of the proposed HCWM index.

4.3 Calculation of the HOSPITAl's HCWM Index Using the Confusing TOPSIS or TOPSIS Model

Event logs were provided through in-person observations at the affected hospitals. The performance for the hospitals studied as well as the performance of the ideal ($Si*$) and non-ideal ($Si-$) hospitals constituted the multi-criteria decision matrix presented in Tables 5 and 6. These performances have been entered into the software and tacked into account the criteria weights. The HCWM index values were calculated using TOPSIS. The values of the HCWM index and the ranking of hospitals are reported in Table 7.

Table 4. Ranking of criteria according to the degree of importance (Before and during COVID-19).

Before COVID-19			During COVID-19		
Rank	Criteria	Weight	Rank	Criteria	Weight
1	C8	0,432144	1	C23	0,71159
2	C10	0,425881	2	C6	0,60523
3	C6	0,402358	3	C4	0,60058
4	C4	0,365444	4	C10	0,60007
5	C12	0,369788	5	C8	0,56922
6	C9	0,368511	6	C11	0,52598
7	C11	0,365124	7	C3	0,514823
8	C3	0,3589421	8	C1	0,513258
9	C1	0,3478222	9	C9	0,502369
10	C26	0,3476887	10	C22	0,501268
11	C25	0,3465873	11	C19	0,423652
12	C16	0,3326811	12	C12	0,422598
13	C28	0,3152222	13	C16	0,412385
14	C19	0,3025998	14	C2	0,411423
15	C29	0,2695558	15	C13	0,409583
16	C18	0,2358321	16	C29	0,408569
17	C2	0,2195238	17	C26	0,395856
18	C13	0,2036987	18	C25	0,384577
19	C24	0,2006911	19	C27	0,354871
20	C15	0,1985236	20	C17	0,345896
21	C7	0,1978563	21	C30	0,344444
22	C21	0,1965488	22	C28	0,344211
23	C30	0,1955554	23	C18	0,3369852
24	C22	0,1947222	24	C5	0,3236669
25	C27	0,1932554	25	C24	0,3215558
26	C23	0,1922211	26	C15	0,3214588
27	C14	0,1914999	27	C20	0,3201445
28	C17	0,1856999	28	C14	0,3123688
29	C20	0,1844423	29	C7	0,2785336
30	C5	0,1482369	30	C21	0,2659388

Table 5. Calculation of the performance value of ideal (Si*) and non-ideal (Si-) hospitals by the TOPSIS method (Before COVID-19).

Hospitals Criteria	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
Si*	1,764	1,766	1,796	1,769	1,674	1,032	1,746	1,807	1,682	1,779	1,823	1,775	1,745	1,689	1,588	1,523
Si-	0,298	0,245	0,241	0,285	0,489	1,511	0,478	0,214	0,487	0,602	0,699	0,301	0,578	0,478	0,783	0,9012
C1	2	4	4	4	8,25	4	4	0,75	7,25	2	0,75	6	0,75	6	4	7,25
C2	0,75	2	2	0,75	8,25	6	2	1,25	8,25	0,75	2	6	4	4	4	4
C3	2	4	4	2	8,25	6	6	4	6	1,25	2	0,75	2	4	4	7,25
C4	6	4	7,25	2	8,25	7,25	6	2	6	0,75	2	2	4	6	4	8,25
C5	7,25	6	0,75	2	8,25	7,25	2	4	8	0,75	0,75	4	0,75	6	4	8,25
C6	8,25	7,25	4	7,25	8,25	7,25	6	2	6	4	0,75	4	4	7,25	4	4
C7	4	4	0,75	4	8,25	6	4	0,75	8,25	0,75	0,75	2	0,75	4	4	2
C8	8,25	4	4	6	8,25	7,25	7,25	2	8,25	4	2	4	0,75	4	4	4
C9	6	4	4	2	8,25	4	6	2	7,75	0,75	2	4	0,75	4	4	2
C10	7,25	2	4	7,25	8,25	7,25	4	2	8,25	1,75	0,75	0,75	2	7,25	4	4
C11	8,25	2	4	2	8,25	6	4	2	4	0,75	0,75	4	4	6	4	2
C12	2	6	4	0,75	8,25	4	2	2	7,25	0,75	0,75	2	4	4	2	2
C13	0,75	6	0,75	2	8,25	4	4	2	8,25	0,75	4	2	0,75	4	4	7,25
C14	0,75	4	0,75	4	8,25	4	0,75	2	7,25	0,75	2	2	0,75	4	4	2
C15	6	6	0,75	8,25	8,25	4	2	6	8,25	0,75	2	2	0,75	4	4	2
C16	7,25	6	4	7,25	7,75	4	4	0,75	8,25	0,75	0,75	2	0,75	4	4	0,75

(*continued*)

Table 5. (*continued*)

Hospitals Criteria	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
C17	0,75	0,75	0,75	2	2	4	4	0,75	7,75	0,75	0,75	2	0,75	4	4	2
C18	2	0,75	2	4	7,75	4	4	4	6	0,75	0,75	2	0,75	4	4	0,75
C19	2	4	6	0,75	8,25	4	6	4	7,25	0,75	0,75	2	0,75	4	4	0,75
C20	4	6	6	0,75	8,25	2	2	0,75	2	0,75	2	6	2	4	4	1,75
C21	4	2	2	4	8,25	2	0,75	4	7,25	4	2	4	0,75	4	4	0,75
C22	2	4	6	4	6,75	4	4	7,25	8,25	0,75	0,75	4	0,75	4	4	0,75
C23	4	4	0,75	6	7,75	4	0,75	2	6	0,75	0,75	2	2	4	4	0,75
C24	6	2	2	6	7,75	4	0,75	7,25	8,25	4	0,75	4	4	4	4	0,75
C25	0,75	4	2	4	7,75	4	6	4	7,25	0,75	0,75	8,25	0,75	4	4	0,75
C26	4	4	7,25	2	7,75	4	7,25	0,75	8,25	0,75	0,75	8,25	0,75	4	4	0,75
C27	2	4	6	0,75	7,75	4	4	2	8,25	0,75	0,75	7,25	0,75	4	4	0,75
C28	2	4	7,25	0,75	7,75	2	0,75	7,25	8,25	4	0,75	4	0,75	4	4	0,75
C29	2	4	6	0,75	7,75	2	6	7,25	6	0,75	0,75	4	0,75	4	4	1,75
C30	2	4	0,75	4	7,75	4	2	4	7,25	4	0,75	4	0,75	4	4	0,75

Table 6. Calculation of the performance value of ideal (Si*) and non-ideal (Si-) hospitals by the TOPSIS method (During COVID-19).

Hospitals Criteria	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
Si*	0,2856	0,1958	0,4592	0,6147	0,1985	0,3258	0,4582	0.6852	0,6879	0,6177	0,6855	0,5745	0,5047	0.1875	0.3952	0,4985
Si-	0,6847	0,6625	0,6012	0,3425	0,8579	0,7785	5,856	0,3857	0,9215	0,2148	0,2798	0,3289	0,3256	0,7954	0,4852	0,4178
C1	4	7,25	6	1,25	8,75	7	8,75	0,25	9,75	5	3	3	5	8,75	5	0,25
C2	0.75	6	2	1,25	8,75	9,75	5	8,75	9,75	1,25	0,25	3	5	7	5	5
C3	4	6	4	3	9,75	9,75	8,75	8,75	9,75	0,25	1,25	1,25	5	7	5	3
C4	7,25	8,25	7,25	0,25	9,75	8,75	9,75	8,75	9,75	1,25	1,25	3	5	8,75	5	9,75
C5	7,25	7,25	0,75	0,25	8,75	8,75	5	0,25	9,75	1,25	0,25	3	5	8,75	5	5
C6	8,25	8,25	6	8,75	9,75	8,75	8,75	0,25	8,75	5	3	1,25	5	9,75	5	8,75
C7	4	6	0,75	5	9,75	8,75	5	0,25	9,75	1,25	3	3	1,25	5	5	5
C8	8,25	6	4	7	9,75	8,75	9,75	0,25	9,75	5	0,25	1,25	3	8,75	5	8,75
C9	6	6	6	1,25	9,75	9,75	7	0,25	9,75	5	3	1,25	1,25	8,75	5	8,75
C10	7,25	6	8,25	8,75	7	8,75	8,75	0,25	9,75	5	1,25	1,25	5	9,75	5	3
C11	8,25	4	6	3	5	9,75	8,75	0,25	9,75	5	0,25	3	5	9,75	5	9,75
C12	8,25	6	7,25	1,25	1,25	8,75	8,75	0,25	9,75	1,25	3	1,25	5	8,75	3	5
C13	4	8,25	0,75	3	0,75	9,75	7	0,25	9,75	0,25	1,25	3	5	7	5	7
C14	4	6	0,75	0,25	9,75	9,75	3	0,25	9,75	0,25	1,25	3	5	7	5	5
C15	6	7,25	0,75	1,25	9,75	8,75	3	0,25	9,75	0,25	1,25	3	0,25	7	5	7
C16	7,25	7,25	4	1,25	9,75	9,75	7	0,25	9,75	0,25	1,25	3	1,25	7	5	1,25

(*continued*)

Table 6. (*continued*)

Hospitals Criteria	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
C17	2	4	0,75	3	5	8,75	7	0,25	9,75	0,25	3	3	0,25	7	5	3
C18	6	4	2	1,25	9,75	7	5	0,25	9,75	0,25	3	3	1,25	7	5	0,25
C19	6	6	7,25	0.75	9,75	5	8,75	8,75	9,75	0,25	3	1,25	1,25	7	5	1,25
C20	6	4	6	0,75	8,75	5	5	0,25	5	0,25	0,25	1,25	3	7	5	1,25
C21	4	0,75	2	0,75	9,75	5	1,25	0,25	9,75	0,25	3	3	1,25	7	5	0,25
C22	6	4	6	0,75	8,75	5	7	8,75	9,75	5	0,25	1,25	1,25	8,75	5	0,25
C23	6	4	2	0,75	8,75	8,75	1,25	0,25	9,75	0,25	1,25	3	3	8,75	5	1,25
C24	6	4	6	0,75	8,75	8,75	1,25	8,75	9,75	1,25	1,25	1,25	5	8,75	5	0,25
C25	2	2	7,25	4	9,75	8,75	8,75	0,25	9,75	0,25	1,25	0,25	1,25	8,75	5	0,25
C26	4	6	7,25	2	9,75	8,75	8,75	0,25	9,75	0,25	3	1,25	3	8,75	5	0,25
C27	4	6	6	0,75	9,75	7	5	0,25	9,75	0,25	0,25	1,25	1,25	8,75	5	0,25
C28	4	6	7,25	075	9,75	7	1,25	8,75	9,75	0,25	3	1,25	0,25	8,75	5	1,25
C29	4	6	8,25	0,75	9,75	5	8,75	0,25	8,75	0,25	3	1,25	1,25	8,75	5	0,25
C30	4	2	7,25	0,75	9,75	5	3	8,75	9,75	0,25	1,25	1,25	1,25	8,75	5	0,25

5 Analysis and Discussion of Results

Based on the results presented in Table 7, we have found that some hospitals have very high characteristics, others do not. For example, the H6 hospital has very high standard weights in C4, C6 and C8 and low weights in C28 and C29.

In addition, the H16 hospital focuses on C1, C3, C4 and C5 because they have high standard weights. However, the weight of the other items is very low, such as C27, C28, C29, 30.

As there is one hospital that has a very high value in all standards, it is the H9 hospital. We have seen that hospitals with very high standards are well managed medical waste. According to the results of Table 12, it can be seen that the majority of the reference weights are high compared to the results of the previous table. This shows that experts' awareness of hospital waste management is increasing with COVID-19.

For example, the H1 hospital focuses more on C4, C5, C6, C8 and C10, and the H3 hospital also focuses on C3, C12, C28, C29 and C30. However, they focused less on C14 and C17 criteria.

The weight of the H9 hospital with all these standards, however, is very high, which shows that the hospital is in accordance with health rules.

On the contrary, during COVID-19, some standards focus on hospitals rather than other hospitals. Approved C6, C8, C10 and C19 standards are respected by H1, H3, H4, H5, H6, H9 and H14 hospitals.

However, criteria such as C21 and C22 are not well respected by healthcare institutions H4, H7, H10, H13, H16. This distribution of criteria will affect the value of the HCWM index and then the performance of waste management as a whole.

According to Table 7, the highest HCWM index values before COVID (best condition) were found in hospitals H6, H16, and H15, while the lowest values (worst condition) were found in hospitals 8, 3, and 12. However, the results presented during COVID-19 showed a sharp increase in the values of all HCWM indices during the COVID-19 period. For example, the hospital 9 index is 95,2223862, and also the index of the hospital 7 have a very important score equal to 93,5861056. Hospitals 5, 6, 2, 1, 3 and 15 have a very high HCWM index. In contrast, the lowest values were observed in hospitals 11 and 10.

Hospitals have gone beyond better waste management in the COVID-19 period due to viral transmission and diversification. As can be seen from the methodology, this study assumed that an index of 50 indicates a medium condition and could serve as a better tool for conceptual assessment. According to the results, only the index value for hospital 6 was above the median value (50) before the COVID-19. On the other hand, more than half of the hospitals were above the average level during COVID-19 period (50).

Finally, from the previous one we can conclude that the results differ from one period to another, and that the criteria have a very important influence on waste management. It is observed that hospitals before the COVID-19 do not manage medical waste, which poses very serious risks to human life and the environment. However, waste management in hospitals is improving during COVID-19 period as shown in previous tables.

Table 7. Ranking of hospitals by HCWM value (Before and during COVID-19).

Before COVID-19			During COVID-19		
Rank	Hospitals	Index value HCWM	Rank	Hospitals	Index value HCWM
1	H6	49,2879356	1	H9	95,2223862
2	H16	43,5971235	2	H7	93,5861056
3	H15	42,8888526	3	H14	84,2458622
4	H11	37,2588999	4	H5	83,5478223
5	H13	34,2002789	5	H6	79,2358222
6	H10	33,2581111	6	H2	75,4712589
7	H5	32,4823158	7	H1	72,5839412
8	H7	31,2222389	8	H3	67,8235451
9	H9	30,6523894	9	H15	64,5512111
10	H14	29,3266852	10	H16	47,5888522
11	H1	24,5833324	11	H13	41,5222228
12	H4	23,5269851	12	H8	39,8712555
13	H2	21,5223888	13	H12	37,6921358
14	H12	21,1248922	14	H4	35,5228621
15	H3	17,3625488	15	H11	30,8529832
16	H8	16,5816395	16	H10	29,8612652

6 Conclusion

In this paper, we have used the two multi-criteria methods OWA and TOPSIS to explain the methodology for calculating the HCWM index in two completely different periods (before and during COVID-19) on the basis of data obtained by survey. We are presented an ontology-based framework for decision support for multi-criteria for data optimization and conceptual data problem solving in hospitals by developing a quantitative index calculated over two different periods. The HCWM index before COVID-19 is very low, which means that the management of medical waste in Tunisian hospitals is poor. On the contrary, the HCWM during the COVID-19 index period seems to be high, which shows that healthcare institutions are very much in compliance with healthcare waste regulations. In fact, this difference is due to the catastrophic state caused by the COVID-19 pandemic.

In future work, we plan to apply these two multi-criteria methods (fuzzy OWA and TOPSIS) to other real applications, and also, we try to explore still other approaches like the Bayesian method and the Markov chain method and used in hospital waste management and see the difference.

References

1. Morissey, A.J., Browne, J.: Waste management models and their application to sustainable waste management. Waste Manage. **24**(3), 297–308 (2004)
2. Lee, B.K., et al.: Alternatives for treatment and disposal cost reduction of regulated medical wastes. Waste Manage. **24**(2), 143–151 (2004)
3. Farzadkia, M., et al.: Evaluation of waste management conditions in one of policlinics in Tehran, Iran. Iran. J. Ghazvin Univ of Med Sci. **16**(4), 107–109 (2013)
4. Miranzadeh, M.B., et al.: Study on Performance of InfectiousWaste Sterilizing Set in Kashan Shahid Beheshti Hospital and Determination of its Optimum Operating Condition. Iran. J. Health & Environ. **4**(4), 497–506 (2012)
5. Pereira, M.S., et al.: Waste management in non-hospital emergency units. Brazil. J. Rev. Latino-Am. Enfermagem. **21**, (2013)
6. Taheri, M., et al.: Enhanced breast cancer classification with automatic thresholding using SVM and Harris corner detection. In: Proceedings of the international conference on research in adaptive and convergent systems, pp. 56–60. ACM, Odense, Denmark (2016)
7. Tabrizi, J.S., et al.: A framework to assess management performance in district health systems: a qualitative and quantitative case study in Iran. Cad. Saúde Pública. **34**(4), e00071717 (2018)
8. Ouhsine, O., et al.: Impact of COVID-19 on the qualitative and quantitative aspect of household solid waste. Global J. Environ. Sci. Manag. **6**(SI), 41–52 (2020)
9. Feldstein, L.R. et al.: Multisystem inflammatory syndrome in U.S. children and adolescents. N. Engl. J. Med., 1–13 (2020)
10. Peng, M.M.J. et al.: Medical waste management practice during the 2019-2020 novel coronavirus pandemic: Experience in a general hospital. Am. J. Infect. Control. **48**(8), 918-921(2020)
11. Ilyas, S., et al.: Disinfection technology and strategies for COVID-19 hospital and bio-medical waste management. Sci. Total. Environ. **749**, (2020)
12. Diaz, L.F., et al.: Alternatives for the treatment and disposal of healthcare wastes in developing countries. Waste Manage. **25**, 626–637 (2005)

13. Brent, A.C., et al.: Application of the analytical hierarchy process to establish health care waste management systems that minimize infection risks in developing countries. Eur. J. Oper. Res. **181**(1), 403–424 (2007)
14. Karamouz, M., et al.: Developing a master plan for hospital solid waste management: A case study. Waste Manage. **27**(5), 626–638 (2007)
15. Hsu, H.J., et al.: Diet controls normal and tumorous germline stem cells via insulin-dependent and—independent mechanisms in Drosophila. Dev. Biol. **313**, 700–712 (2008)
16. Victoria Misailidou, P.T., et al.: Assessment of patients with neck pain: a review of definitions, selection criteria, and measurement tools. J. Chiropr. Med. **9**, 49–59 (2010)
17. Liu, H.C., et al.: Assessment of health-care waste disposal methods using a VIKOR-based fuzzy multi-criteria decision-making method. Waste Manage. **33**, 2744–2751 (2013)
18. Gumus, A.T.: Evaluation of hazardous waste transportation firms by using a two steps fuzzy-AHP and TOPSIS methodology. Expert Syst. Appl. **36**, 4067–4074 (2009)
19. Carlsson, B.: Technological systems and industrial dynamics. Kluwer Academic Publishers, Boston, Dordrecht, London (1997)
20. Yager, R.R.: On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans. Syst. Man Cybern. **18**, 183–190 (1988)
21. Hwang, C.L., Yoon, K.: Multiple attribute decision making: methods and applications. Springer-Verlag, New York (1981)
22. Baghapour, M.A., et al.: A computer-based approach for data analyzing in hospital's health-care waste management sector by developing an index using consensus-based fuzzy multi-criteria group decision-making models. Int. J. Med. Inform. (2018)



Applying ELECTRE TRI to Sort States According the Performance of Their Alumni in Brazilian National High School Exam (ENEM)

Helder Gomes Costa(✉), Luciano Azevedo de Souza, and Marcos Costa Roboredo

Universidade Federal Fluminense, 156 - Bloco D, Rua Passos da Pátria, Niterói 24210-240, RJ, Brazil
{heldergc,lucianos,mcroboredo}@id.uff.br

Abstract. An issue faced by governments is to design actions to rise the quality level and to empower competition skills of the public under their public management. In this scenario the Brazilian Ministry of Education applies an annual exam (Enem) that evaluates the knowledge, skills and capabilities to anyone who has completed or is about to graduate high school. In our article, we analyze the results gotten by 3,389,832 alumni that attend the a last edition (2021) of Enem. We adopted a multicriteria decision modelling method to analyse the results. The modelling was able to sort all the instances.

Keywords: Sorting · ELECTRE · Decision support · High school · Education · Higher education access

1 Introduction

The National High School Exam (Enem), established by the Brazilian Ministry of Education (MEC) in 1998 [1], is a test designed to validate high school graduates' knowledge, skills and abilities. This exam is given once a year and is eligible to anyone who has completed or is about to graduate high school.

The Enem's major objective is to assess the quality of secondary education in the country. Individuals must pay an application fee to take the exam and participate in it voluntarily. Despite this, millions of students attend it each year, likely because the scores obtained in the Enem have turned into a pipeline of access to higher education.

According to [1], Enem assesses the general skills of students who have completed or are completing high school. Unlike the traditional entrance exam, which requires specific content, the Enem analyzes students' ability of reading, comprehension, writing, as well as their ability to apply concepts.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 804–813, 2023.
https://doi.org/10.1007/978-3-031-27409-1_73

The subjects of the exam are divided into the following four areas of knowledge plus a essay:

- Languages, codes and their technologies, covering contents of Portuguese Language, Modern Foreign Language, Literature, Arts, Physical Education and Information Technology;
- Mathematics and its technologies.
- Natural Sciences and its Technologies, which covers Physics, Chemistry and Biology;
- Humanities and their technologies, covering Geography, History, Philosophy, Sociology and general knowledge.

In another perspective, the Enem results should be used to support governmental policy. In this regard, the following topic is addressed in this article: "How to classify Brazilian States based on Enem test results?"

This kind of problem is scratched in Fig. 1: given a set of States, classify them into ordered categories according to their alumni performance in Enem exam.

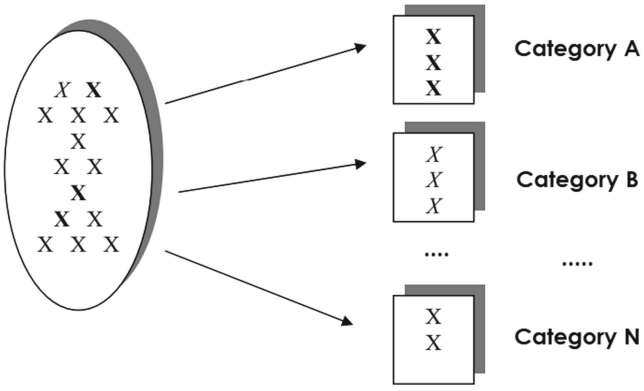


Fig. 1. The general sorting problem

Taking a reasoning similar to that describe in [2,3], one can build the Table 1, that compares the characteristics of sorting problems against those that appear in the problems issued by ELECTRE TRI method, described in [4,5].

Analysing the Table 1, we conclude that the topic covered in this article is a typical multi-criteria sorting problem that should be solved using an ELECTRE TRI based modeling. This conclusion is reinforced by recent application and advances of using ELECTRE TRI based modelling in sorting problems, as shown in [6–8], that emphazized that quality evaluation should avoid compensatory effects. To a deeper discussion about outranking fundamentals we suggest to read [9,10]. Therefore we used a modelling based on ELECTRE TRI to sort Brazilian states, according the grades their students reached in Enem.

Table 1. Comparing a ELECTRE against the problem addressed

Feature	ELECTRE TRI	The problem Addressed
Objective	To sort alternatives	To sort States according to Enem results
Criteria	Variables used to evaluate alternatives	Variables used to evaluate students graduated in a state
Grades	Performance of alternatives under each criterion/variable	Performance of alternatives under each criterion/variable
Weight	A constant of scale	A constant of scale
Category	A category or group in which the alternatives are a classified. Notice that there is a ranking relationship among the categories	A category or group in which the States are classified. Observe that there is a ranking relationship among the categories
Profiles	A vector of performances that under delimits each category	A vector of performances that under delimits each category

2 Methodology

In this section, we summarize the actions undertaken during the research. As a result, the next section describes how they are used and the outcomes obtained.

(a) To define the object of study
(b) To elicit the criteria set
(c) To define the criteria weights
(d) To define the alternatives to be sorted
(e) To evaluate the alternatives under each criterion
(f) To define the categories or groups into which the States will be sorted
(g) To define the profiles that delimits each category
(h) To run the classification algorithm.

3 Modelling and Results

In this section we apply the steps described in the previous section, justifies the modelling decisions, and, shows and discuss the results gotten.

3.1 To Define the Object of Study

The object of study is the results reached in the edition Enem 2021, by a total of 293,400 alumni from the 27 States that composes the República Federativa do Brasil—or Brazil, as this country is usually mentioned.

3.2 To Elicit the Criteria Set

The criteria set are the subjects covered in Enem exam, as it appears in Table 2.

3.3 To Define the Criteria Weights

Once we consider that there is not a criteria more relevant than other one, in this work we used the same weight or constant of scale for all criteria.

Table 2. Criteria set

Criterion code	Enem's subject	Contents covered
LC	Languages, codes and their technologies	Portuguese Language, Modern Foreign Language, Literature, Arts, Physical Education and Information Technology
MT	Mathematics and its technologies	Numbers; Geometry, Quantities; Graphs and tables; Algebric representations; and, Problem's Solving and Modeling
CN	Natural Sciences and its Technologies	Physics, Chemistry and Biology
CH	Humanities and their technologies	Geography, History, Philosophy, Sociology and general knowledge
ESSAY	Essay	Mastery of the formal writing of the Portuguese language; Comprehension of the topic and not running away from it; Organize and interpret information and build argumentation; Knowledge of the linguistic mechanisms necessary for the construction of the argument; and, Respect for human rights.

3.4 To Define the Elements to Be Sorted

The objects to be sorted are the 27 States that compose the Brazilian republic: Acre (AC), Alagoas (AL), Amazonas (AM), Amapá (AP), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Minas Gerais (MG), Mato Grosso do Sul (MS), Mato Grosso (MT), Pará (PA), Paraíba (PB), Pernambuco (PE), Piauí (PI), Paraná (PA), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rondônia (RO), Roraima (RR), Rio Grande do Sul (RS), Santa Catarina (SC), Sergipe (SE), São Paulo (SP) and Tocantins (TO).

3.5 To Evaluate the Alternatives Under Each Criterion

The mean of the grades reached by the students of each state are shown in Appendix.

3.6 To Define the Categories or Groups Into which the 27 States Should Be Sorted

We defined a set K containing five categories ($K = A, B, C, D, E$) as described bellow:

- A: So much over the median
- B: Over the median
- C: around the median
- D: Bellow the median
- E: So much bellow the median.

These categories are also aligned to the discussion about "The magical of number seven" shown in Miller [11]. According to this article, scales should have five points with symmetrical meaning.

3.7 To Define the Profiles that Delimits Each Category

The definitions of such parameters is usually based on subjective evaluations. Aiming to reduce subjective effects in d the modelling [12], choose these parameters based on standard deviation and mean of the data, while [13] were pioneer in proposing the use of triangular and rhomboid distributions to define the class of ELECTRE TRI; In our paper for each criterion we define a lower limit in such away we have a rhomboid (or two inverted symmetric triangles) based distribution of the elements in c each category. This was because the meaning of the categories mentioned in the previous subsection and c because this avoids distortions caused by eventual outliers in the data. In other words, for each criterion we have:

- 10% of States above the lower limit of class A
- 20% of States above the lower limit of class B under the lower limit of class A—it means that the lower limit of class B is 30% of States
- 40% of States above the lower limit of class C under the lower limit of class B—it means that the lower limit of class C is 70% of States
- 20% of States above the lower limit of class D under the lower limit of class C—it means that the lower limit of class D is 90% of States
- 10% of States above the lower limit of class E under the lower limit of class E—it means that the lower limit of class D is the minimum value that a State has in the criterion.

Table 3 shows the boundaries of the categories according the each criterion.

Table 3. Profiles or under boundaries

Profile	CN	CH	LC	MT	ESSAY
P_0	0.00	0.00	0.00	0.00	0.00
P_1	465.26	488.13	473.18	501.36	569.36
P_2	477.25	500.89	487.55	510.39	599.10
P_3	483.58	507.95	492.18	529.37	614.08
P_4	499.21	525.42	507.82	549.72	632.39

3.8 To Have the Brazilian States Sorted

In this step we applied the ELECTRE sorting algorithm to get the Brazilian States sorted into the categories that were described in 3.3 , To do this: for each one of the profiles P_i that appears in Table 3, the Eq. 1 is applied to calculate concordance degree with the assertive that an State has a performance at least not worse than the profile P_i, taking into account its allumni performance on Enem.

$$C(State, P_i) = \frac{1}{\sum w_j} \sum w_j * c(State, P_{ij}) \tag{1}$$

Where

- P_i is the $i - esim$ array profile that under boundaries a category K_i.
- w_i is the constant o scale or weight of $j - esim$ criterion that is shown in Table 2.
- $c(State, P_{ij})$ is the local (or at a criterion) concordance degree with the assertive that an State has a performance at least not worse than the profile P_i under the criterion j.
- P_{ij} is the value of the profile P_i under the $j - esim$ criterion.

By assuming in this problem that we are dealing with true criteria (see [10]), $c(State, P_ij)$ is calculated as shown in Eq. 2

$$c(State, P_{ij}) = \begin{cases} 1 \; if \; g_j(State) >= P_{ij} \\ 0 \; if \; g_j(State) <= P_{ij} \end{cases} \tag{2}$$

Where

- $g_j(State)$ is the performance of the State under the $j - esim$ criterion—as it appears in Table 5, in Appendix.

Applying Eqs. 1 and 2 to the data that appear in Appendix, and taking into account the values shown in 3, one can found the values that appear in Appendix B. The values express the concordance degree with the assertive that the students of a State that appears in a row have reached a performance at least as good as those the profiles that in the columns of Table that appears in Appendix.

Table 4 illustrates the final sorting obtaining by using a credibility cut-level $0,75$. This value was chosen taking into consideration:

– that it is a reference linked to the Q1 quartile.
– the values that appear in Table 5.

Table 4. Brazilian States sorted according to ENEM results

Categories	State
A	DF, ES, PR, RJ, RS, SC, SP
B	PB, MS, RN, SE
C	AL, BA, GO, MT, PE, PI
D	AC, PA, RO, RR, TO
E	AP, AM, CE, MA

4 Conclusion

This study was successful in identifying geographic regions based on the outcomes of their graduates on a nationwide proficiency exam, with a sample of 3.839.963 tests. It identified the areas that require further attention to develop alumni competencies, which should be worthy to government initiatives.

The reader should think about the fact that the test was performed during the COVID-019 pandemy, which could have impacted student performance. As an extension of this research, we intend to employ the same method to investigate the results of Enem tests conducted in prior years, with the purpose of identifying which Brazilian Geographic regions were more impacted by the pademy in terms of High School education.

Another contribution of this work is the justification to the categories' boundaries, based on statistical distributions of the data. As further work we suggest to explore the use and comparison of other data-based metrics in the definition of the boundaries ob the class used in the ELECTRE TRI methods.

5 Appendix

Tables 5 and 6 illustrate, respectively the mean of the degrees reached by the students of each one of the 27 states, and the credibility degree of the categorization.

Table 5. Brazilian States and their alumni's performances in ENEM

SG	CN	CH	LC	MT	ESSAY
AC	468.82	496.21	483.33	504.85	587.25
AL	479.55	504.94	489.20	526.85	625.70
AM	450.89	469.09	457.01	480.56	501.57
AP	465.14	491.49	473.91	494.29	576.80
BA	483.58	507.95	492.67	524.08	615.49
CE	460.83	482.22	471.73	501.17	539.62
DF	505.71	534.04	522.10	552.27	627.38
ES	502.24	525.24	508.61	553.40	637.18
GO	483.56	507.94	493.81	528.77	606.03
MA	465.33	487.56	472.09	501.49	590.74
MG	513.31	542.17	523.96	576.80	666.53
MS	487.64	510.50	497.13	536.12	601.19
MT	486.10	509.34	490.24	530.03	606.56
PA	471.70	496.74	475.97	504.38	614.08
PB	483.92	508.77	490.64	529.37	637.41

continued

Table 5. continued

SG	CN	CH	LC	MT	ESSAY
PE	482.81	504.28	493.26	533.37	612.44
PI	481.77	504.96	488.60	527.20	643.74
PR	503.84	529.24	513.07	554.10	608.28
RJ	505.70	537.06	520.41	563.57	653.74
RN	498.46	526.14	507.62	549.09	653.56
RO	473.36	492.39	478.59	510.79	575.17
RR	478.22	501.93	490.23	507.70	560.64
RS	504.32	536.29	519.43	559.24	631.19
SC	512.51	537.06	515.86	563.75	627.81
SE	487.85	510.18	492.18	532.78	650.94
SP	511.19	541.72	527.94	573.70	637.65
TO	469.42	488.50	474.59	508.78	587.42

Table 6. Credibility degree

State	'P_0'	'P_1'	'P_2'	'P_3'	'P_4'
AC	1.0	1.0	0.0	0.0	0.0
AL	1.0	1.0	1.0	0.2	0.0
AM	1.0	0.0	0.0	0.0	0.0
AP	1.0	0.6	0.0	0.0	0.0
BA	1.0	1.0	1.0	0.6	0.0
CE	1.0	0.0	0.0	0.0	0.0
DF	1.0	1.0	1.0	1.0	0.8
ES	1.0	1.0	1.0	1.0	0.8
GO	1.0	1.0	1.0	0.2	0.0
MA	1.0	0.6	0.0	0.0	0.0
MG	1.0	1.0	1.0	1.0	1.0
MS	1.0	1.0	1.0	0.8	0.0
MT	1.0	1.0	1.0	0.6	0.0
PA	1.0	1.0	0.2	0.0	0.0
PB	1.0	1.0	1.0	0.8	0.2
PE	1.0	1.0	1.0	0.4	0.0
PI	1.0	1.0	1.0	0.2	0.2
PR	1.0	1.0	1.0	0.8	0.8
RJ	1.0	1.0	1.0	1.0	1.0

continued

Table 6. continued

State	'P_0'	'P_1'	'P_2'	'P_3'	'P_4'
RN	1.0	1.0	1.0	1.0	0.4
RO	1.0	1.0	0.2	0.0	0.0
RR	1.0	0.8	0.6	0.0	0.0
RS	1.0	1.0	1.0	1.0	0.8
SC	1.0	1.0	1.0	1.0	0.8
SE	1.0	1.0	1.0	0.8	0.2
SP	1.0	1.0	1.0	1.0	1.0
TO	1.0	1.0	0.0	0.0	0.0

Acknowledgments. This study was partialy funded by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPQ)-Grants 314953/2021-3 and 421779/2021-7; and, Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro—Brasil (FAPERJ), Grant 200.974/2022.

References

1. Brasil, M.: Fazer O Exame Nacional do Ensino Medio (2022). https://gov.br/pt-br/servicos/fazer-o-exame-nacional-do-ensino-medio
2. Costa, H.G., Santafé Júnior, H.P.G., Haddad, A.N.: Uma contribuição do método ELECTRE TRI à obtenção da classificação de riscos industriais. Investigação Operacional **27**(2), 179–197 (2007)
3. Costa, H., Duarte, M.B.T.: Applying ELECTRE TRI ME for evaluating the quality of services provided by a library. In: Proceedings of the 10 applying ELECTRE TRI to sort states according results of ENEM ... ICETC 2019, pp. 278–281. Association for Computing Machinery (ACM), Amsterdam.https://doi.org/10.1145/3369255.3369313
4. Mousseau, V., Slowinski, R., Zielniewicz, P.: A user-oriented implementation of the ELECTRE-TRI method integrating preference elicitation support. Comput. Oper. Res. (2000). https://doi.org/10.1016/S0305-0548(99)00117-3
5. Greco, S., Figueira, J., Ehrgott, M.: Multiple Criteria Decision Analysis: state of Art Surveys, vol. 37. Springer, Cham (2016)
6. Da Rocha, P.M., Costa, H.G., Da Silva, G.B.: Gaps, Trends and challenges in assessing quality of service at airport terminals: a systematic review and bibliometric analysis. Sustainability **14**(7), 3796 (2022). https://doi.org/10.3390/SU14073796
7. Emamat, M.S.M.M., Mota, C.M.d.M., Mehregan, M.R., Moghadam, M.R.S., Nemery, P.: Using electre-tri and flowsort methods in a stock portfolio selection context. Financ Innov **8**, 1–35 (2022). https://doi.org/10.1186/S40854-021-00318-1/TABLES/9
8. Costa, H.G., Nepomuceno, L.D.D.O., Pereira, V.: ELECTRE ME: a proposal of an outranking modeling in situations with several evaluators. Brazilian J. Oper. Prod. Manag. **15**, 566–575 (2018). https://doi.org/10.14488/bjopm.2018.v15.n4.a10

9. Costa, H.G.: Graphical interpretation of outranking principles: avoiding misinterpretation results from ELECTRE I. J. Modell. Manag. **11**(1), 26–42 (2016). https://doi.org/10.1108/JM2-08-2013-0037
10. Roy, B.: The outranking approach and the foundations of ELECTRE methods. Theory Decis. **31**, 49–73 (1991). https://doi.org/10.1007/BF00134132
11. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**, 81–97 (1956). https://doi.org/10.1037/h0043158
12. Costa, H.G., Mansur, A.F.U., Freitas, A.L.P., Carvalho, R.A.d.: ELECTRE TRI applied to costumers satisfaction evaluation. Producao **17**(2), 230–245 (2007). https://doi.org/10.1590/S0103-65132007000200002
13. Gomes, A.R., Costa, H.G.: Potencial de consumo municipal: uma abordagem multicritério. Sistemas Gestão **3**, 233–249 (2008). https://doi.org/10.7177/sg.2008.SGV3N3A5

Consumer Acceptance of Artificial Intelligence Constructs on Brand Loyalty in Online Shopping: Evidence from India

Shivani Malhan(✉) and Shikha Agnihotri

University School of Business, Chandigarh University, Chandigarh State Hwy, Sahibzada Ajit Singh Nagar, NH-05, Ludhiana, Punjab 140413, India
shivani.e8881@cumail.in

Abstract. The main aim of this research is to study the impact of Artificial Intelligence constructs on Brand Loyalty in Online Shopping as Brand Loyalty helps in enhancing the market share as well as profitability. Three independent variables i.e., Perceived Ease of use, Experience and Trust were being studied to see their effect on dependent variable i.e., Brand loyalty. Regression analysis test was used in the research. The results pointed out that Perceived Ease of Use and Trust has an impact on brand loyalty whereas experience doesn't impact the brand loyalty. The results will help the marketers to analyse the impact and get more insights on brand loyalty.

Keywords: Brand loyalty · Artificial intelligence · Online shopping

1 Introduction

Online shopping has undergone many changes with the rapid development of digital technology (Daley, 2018). To fulfil fast changing consumer demands and to increase sales efficiency, AI has proved as an excellent tool. As a result, major online commerce is being done with the help of Artificial Intelligence (AI). According to Maynard (2019), almost 3,25,000 retailers will adopt AI technology and hence making it $12 billion by 2023.Retailers on a global level are likely to spend four times on AI services. There has been a huge increase in online product research over the last few years (Smidt and Power, 2020). The best example of integrating AI in online retail is Amazon which is USA's largest online retailer. The customers can pay in their local currency as Amazon uses pricing according to the location and messages are being sent to the customers according to their destination. This way customers can experience localised shopping journey along with fast delivery and competitive prices. (Barmada, 2020).The technology has introduced various new marketing techniques which along with use of AI, has helped in reaching target customers easier and also improves consumer experiences (Pusztahelyi, 2020).

To manage and take decisions in varied business problem, Business intelligence is a new field which helps to investigate human cognitive faculties and Artificial Intelligence

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 814–823, 2023.
https://doi.org/10.1007/978-3-031-27409-1_74

(AI) technologies (Ranjan, 2009). The magic word, Artificial Intelligence (AI), has brought many changes in personal and working life. "Though the AI is being considered important in industry 4.0, it offers lot of opportunities to various sectors but poses many challenges too. The development of AI-powered technologies has led to the growth of economy which further improves quality of life. (Dhanabalan & Sathish, 2018). In the past few years, the commonly known technologies are Artificial Intelligence, Cloud Computing, and Big Data. Understanding data is the foremost thing to be ahead of competitors. (Riley, 2018).

According to Bruckinx & Van den Poel, 2005, businesses can directly contact customers with the help of World Wide Web and other technologies. An important aim of marketing strategies is Customer Loyalty which offers various benefits (Jacob & Chestnut, 1978). Above all, for a firm´s product or service, it helps in maintaining loyal customers (Oliver, 1997).

A huge difference in earnings can be seen with a minor change in customer retention rate which further increases with the passage of time. Profound connection between profitability and loyalty helps in creating more prospects as loyal customers will be ready to buy more and pay more(Reichheld, 1993; Wright & Sparks, 1999; Zeithaml, Berry, & Parasuraman, 1996). Artificial Intelligence will influence various marketing strategies, customer service, processes related to sales, services related to customers and behaviour related to customer.

First, sales processes in various industries will get affected. Till now, making a telephone call is an important aspect of sales process for almost every salesperson. But in future, AI agent will help salesperson in monitoring tele-conversations in real time. Also, to contact sales prospects initially, AI bots can be used which will function as human salespeople. But there are negative consequences as customers may get uncomfortable on knowing that are interacting with a bot.

Second, the shopping-then-shipping model in which customers are required to place orders with online retailers and then the products are shipped is being followed currently (Agrawal, Gans & Goldfarb, 2018; Gans et al., 2017). Artificial Intelligence will help the retailers in shifting to shipping-then-shopping business model where they will predict the customers wants provided the accuracy rate of predictions is high. Thus, without an order being placed, AI will help in knowing customers preferences and ship the products accordingly where customers can return the products they don't require. (Agrawal et al., 2018; Gans et al., 2017).

In this competitive world, almost all organisations have accepted that Brand Loyalty is very important. It is a fact that it costs less to retain the existing customers than to have new ones. Hence, customer retention and profitability become dominant (Hill & Alexander, 2016).

Indian sportswear market was solely for sportspersons who were a part of the niche segment but currently it has begun developing as a customer industry. It has undergone significant development in recent years and is driven by increase in purchasing power, entry of multinational corporations due to globalization and liberalization.

Due to increase of interest in certain games like hockey, tennis and football, the people of India have started spending on sportswear which includes sport shoes. Furthermore, sportswear has been recognized as easy going wear and this approach has increased the

number of consumers for the main sports shoe brands. A few brands also offer collections to take into account the interest of the customers who wear casual footwear.

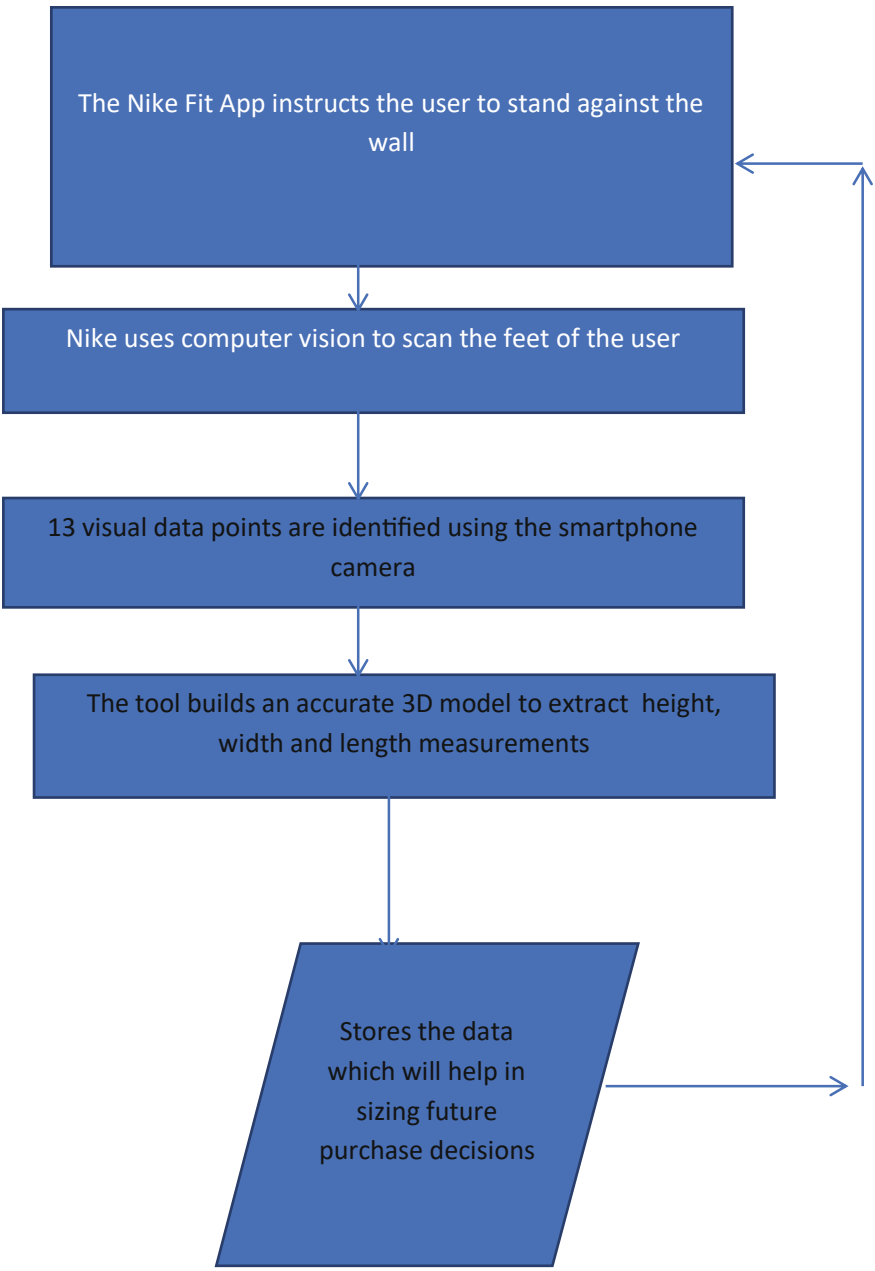
India has begun facilitating a progression of International sporting which is increasing the awareness of sports. In addition, the Indian government is additionally concentrating on advancement of sports, empowering the sportspersons, providing them with training and infrastructure facilities. It also included exports of sport goods to various countries. The athletic clothing and footwear industry is being changed from the unorganised to the unorganised segment. Many corporates have in-house exercise rooms and are supporting their workers for wellness exercises. Because of this, the offers of sportswear are increasing. The male segment has an extreme interest for basketball shoes, lightweight running shoes, warming up shoes and soccer shoes. On the other hand, the athletic footwear market has seen a strong interest for female athletic shoes, particularly from the school/college going adolescents and youthful mothers. The youngsters essentially show interest for training and casual shoes while the mothers have displayed inclination for conditioning and running shoes.

As per Euromonitor, the Indian sportswear market has developed by twenty two percent from the period 2015–2016, though the segment's worldwide increment was seven percent. By 2020, an extra twelve percent compound annual growth rate is expected. This is because of the rising attention to the individuals in the nation towards driving a sound way of life with more prominent interest in physical exercises, for example, running, strolling, training and different exercises. In addition, rising incomes of individuals in urban India have enabled them to concentrate on wellbeing and health. The representation of sports and Bollywood stars for proliferation of worldwide brands has kept sports footwear in the public eye.

Activities like running, jogging, walking and cycling have risen over three years and have become the most trending activities. There has been substantial increase in the number of gymnasiums in the nation. It is anticipated that the number of gymnasiums will increase by 7 percent till 2020. The customer observation has moved to be 'body-delightful'. The developing nearness of Bollywood stars and acclaimed competitors via web-based networking media has moved the focal point of the two guys and females towards the everyday wellness schedules and sportswear buys. Due to of the expected social and cultural shifts in the coming years, there is a probability that India is going to turn into the greatest sporting boom. The nation's sportswear category is becoming increasingly appealing to a huge group of local as well as international brands.

Artificial intelligence, augmented reality, computer vision, data science and machine learning is used by many sportshoes companies nowadays. Many customers claim that they are not able to find the right fit of sportsshoes and therefore Nike has created the Nike Fit tool which is incorporated within the Nike application. This tool helps in showcasing how the products will look like on their feet.

The Nike Fit tool scans the customer's foot and provides a read of its size by working on the consumer's friction point. The tool is used in various locations of North America, Europe and Japan.



This app helps in recommending the perfect shoe size for the users. The smartphone camera is being used to scan the feet of the user and within seconds,it collects the 13

data points mapping morphology for both feet. Profile is created for each member and the data stored can be used for shopping in the future.

Nike has also acquitted a company named Celect which is a predictive analytics and cloud computing company. The technology of Celect will help in direct selling the product to the consumer. With the help of Artificial intelligence, the direct business sales of the company increased by 12 percent which is thirty percent of Nike's total revenue.

Under Armour which is an American sportswear manufacturer employs artificial intelligence and 3D printing. In partnership with Auto Desk, it designed a unique sneaker which is printed and not stitched. For the final product, machine learning AI studies all the aspects related to the durability of the sneaker. At the end, sneaker so designed is printed in 3D.Entrupy which is a product authentication provider helps the identify the counterfeit shoes of Nike and adidas from other shoes. This is done through artificial intelligence.

Also, a data flow mechanism is used by many sports shoes companies for making sports shoes and the flow chart for the same is given below:

Data Flow Mechanism Used by Sports shoes companies.

2 Literature Review

2.1 Relation of Perceived Ease of Use, Trust and Experience on Brand Loyalty in Online Shopping

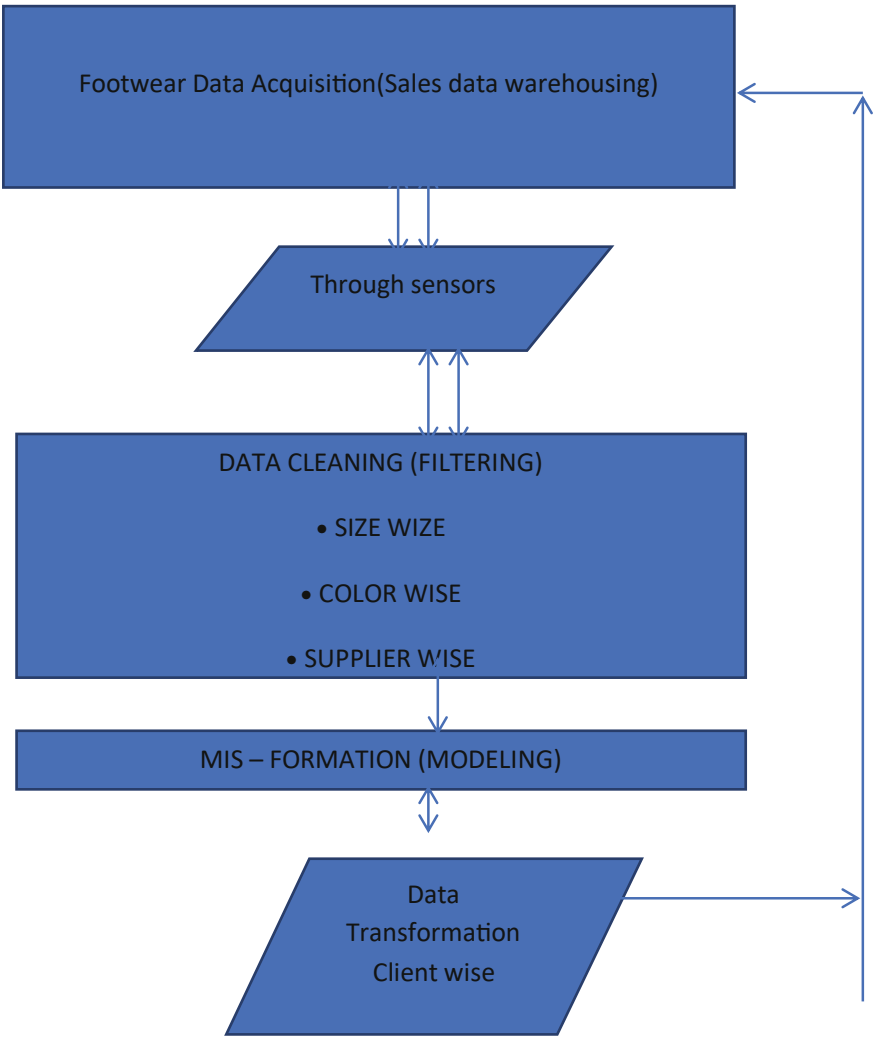
Based on the approaches to measure brand loyalty, three categories are: behavioural approach, attitudinal approach and the multi domain approach. Multi domain approach of brand loyalty is being followed by most the researchers as shown by a study of almost a year. Marketers and the researchers got benefited as this study provided good insights on brand loyalty. The results showed that brand loyalty is of great importance for every company because of which companies can increase their profits and retain the customers. Loyal customers always prefer same brand and brings along many prospects for the brand.

Perceived ease of use is the level at which a person have faith in a particular system that it will require no efforts. Positive effect of perceived usefulness on customer loyalty can be seen as it directly impacts users' loyalty in using technology but doesn't have much impact on how to use the technology.

Perceived ease of use implies the belief that people have, to do less efforts because of technology. According to Davis 1989, if application is simple, give required results, user friendly, flexible and easy to understand; the requirement of ease of use is being met.

3 H1: Perceived Ease of Use Has a Significant Positive Effect on Brand Loyalty

Brand Loyalty increases if consumer has knowledge of how to purchase from AI powered webshops. Brand loyalty is a marketing program where consumers expect personalized and unique experiences which are being matched with their values and experience related



to websites will improve the Brand Loyalty significantly [27]. So,various brands need to work on this in order to remain in competition.Theonline stores using AI provide automated assistance to its customers and experience which leads to Brand Loyalty. (Yoo, Lee and Park, 2010; Pantano and Pizzi, 2020).

4 H2: Experience Has a Significant Positive Effect on Brand Loyalty

Major role is played by trustin online commerce. Consumer's intention to buy is directly influenced by consumer confidence [19]. If the consumer trusts an online shop fully,

his interest to go for the buying process will be high. Trust becomes a critical factor in case financial risk is involved. Brand affect, brand trust and attitudinal loyalty are interrelated (Geçti & Zengin, 2013). 428 consumers filled the questionnaires. Also, brand trust has a strong connection with attitudinal loyalty and behavioural loyalty. Chaudhuri and Hoolbrook [21] discussed features of loyalty which were defined by Oliver. The two areas are purchase and attitudinal loyalty. With the increase in customer's trust, the risk acceptance and vulnerability to do a transactional activity will increase, hence they will continue buying products from the same company (Hsu, 2007; Jin et al., 2008; Zhou et al., 2009). As a result, customers will be loyal to the company and companies can retain them successfully.

5 H3: Trust Has a Significant Positive Effect on Brand Loyalty

5.1 Research Methodology

The main aim of this study was to understand the effect of the artificial intelligence constructs on brand loyalty of consumers of sports shoes. Using Online survey method, data has been collected from a sample of 1000 respondents selected on the basis of purposive sampling. The questionnaire was adapted from existing literature. The statements of perceived ease of Use, Experience, Trust and brand loyalty were taken from the study of Hu and O'Brien, 2016; Park, 2009. A 5-point Likert scale was used to measure all the items in this study. A total of 782 usable responses were collected and imputed in SPSS for the purpose of data analysis. Regression technique was used to estimate the impact of ease of use as perceived by the customer, Experience and Trust on brand loyalty.

6 Analysis

The analytical tool used in the research paper was regression analysis.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.691[a]	0.478	0.476	0.38995
[a]Predictors: (Constant), TR, PEU, EX				

The characteristics of the model are explained in the table above. Perceived Ease of Usage, Experience, Trust and brand loyalty are the main variables considered. As the R value depicts the correlation between the dependent and independent variable, a value which is greater than 0.4 is desirable. In the table, the value is 0.691, which is considered good.

In this table, the value of R square is 0.478, which is more than 0.2. So, it is acceptable. Also, the adjusted R square value is 0.476, which is considered good.

ANOVA[b]						
Model		Sum of Squares	Df	Mean Square	F	Sig
1	Regression	108.298	3	36.099	237.397	0.000[a]
	Residual	118.305	778	0.152		
	Total	226.604	781			
[a]Predictors: (Constant), TR, PEU, EX						
[b]Dependent Variable: BL						

P-value/Sig value: The p-value is 0.00 which is less than 0.05. So, it is considered appropriate and the result is considered significant. Also, the value of F ratio is 237.397, which is considered good.

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig
		B	Std. Error	Beta		
1	(Constant)	1.256	.110		11.406	0.000
	PEU	0.364	0.024	0.413	15.165	0.000
	EX	−0.012	0.019	−0.018	−0.634	0.526
	TR	0.348	0.022	0.461	15.853	0.000
[a]Dependent Variable: BL						

The significance value is 0.00 which is less than 0.05. So, there is an impact of Perceived Ease of Usage and Trust on brand loyalty. But experience doesn't have much impact on Brand Loyalty.

7 Results and Discussion

This study proved that Perceived Ease of Usage and Trust has an impact on brand loyalty. But there is no significant impact of experience on Brand Loyalty. This implies that if the customers have perceived ease of usage and Trust then greater Brand Loyalty will be achieved (Cengiz, 2016). Perceived usefulness has a significant positive effect on Brand loyalty. If the perceived usefulness increases, the loyalty of the customers also increases. The result is supported by the previous studies (Hamid et al. 2016). This is also reinforced by the theory by Davis (1989). The perceived usefulness has an effect on users' loyalty in using technology.

AI-enabled shopping apps which are easy to usehave gathered more trust from the consumers as compared to complicated apps. Higher trust helps in the formation of positive attitude towards shopping in webshops. If the customers find artificial intelligence useful in online shopping, the companies will also be motivated to do shopping in AI –powered web shops.

Trust is really important in the usage of the AI enabled webshop as its absence can lead to negative attitude towards it and can decrease the web traffic.

It is imperative that the online consumers are given tailor made offerings so that they can grab the best deal, get the products with higher values and shorten the search time of the products to increase effectiveness of shopping.

Due to COVID -19 and its effect on E-Commerce, the importance of AIin online shopping has increased manifold and is likely to increase in the future. Artificial Intelligence has been proved to be a very effective tool as it helps to provide a personalized journey to the consumer and fulfils the demands of the customer (Bloomberg, 2020).

This study has various implications. The mangers need to adapt to new technologies like Artificial Intelligence in online shopping if they want to compete with other online firms in the industry. Managers, academicians and researchers who are interested in the application and adaption of Technological acceptance model in shopping will be benefitted from this paper.

8 Conclusion

It is really important to study the impact of three variables i.e., Perceived ease of use, Experience and Trust on Brand Loyalty as it helps in enhancing the share of the market and profitability of the companies. In short, to increase brand loyalty online shoppers can improve trust and perceived ease of use in artificial intelligence. Furthermore, owners belonging to any industry could enhance the quality of management to increase customer loyalty using artificial intelligence (AI) applications.

References

1. Acharya, A., Gupta, O.P.: Influence of peer pressure on brand switching among indian college students. Int. J. Curr. Res. **6**(2), 5164–5171 (2014)
2. Ahmed, Z., Rizwan, M., Ahmad, M., Haq, M.: Effect of brand trust and customer satisfaction on brand loyalty in Bahawalpur. J. Sociol. Res. **5**(1), 306–326 (2014)
3. Akabogu, C.O.: A Theory based empirical analysis of brand loyalty to 7up. IOSR J. Bus. Manag. **16**(1), 101–108 (2014)
4. Alhaddad, A.: Perceived quality, brand image and brand trust as determinants of brand loyalty. Quest J. Res. Business Manag. **3**(4), 01–08 (2015)
5. Alhedhaif, S., Lele, P., Kaifi, B.A.: Brand loyalty and factor affecting buying behaviour of Saudi consumers. J. Business Stud. **7**, 25–38 (2016)
6. Alkhawaldeh, A., M.: Factors influencing brand loyalty in durable goods market. Int. J. Acad. Res. Bus. Social Sci. **8**(1), 326–339 (2018)
7. Alloza, A.: Brand engagement and brand experience At BBVA, the transformation of a 150 years old company. Corp. Reput. Rev. **11**(4), 371–381 (2008)
8. Tim, A., Bhattacharya C.B., Julie, E., Lane, K. K., Lemon Katherine, N., Vikas, M.: Relating brand and customer perspectives on marketing management. J. Service Res. **5** (2002)
9. Tim, A.: How Much of Brand Equity is Explained By Trust?" Manage. Decision **35/4** (1197), 283–292 (1997)
10. Awan, A.G., Rehman, A.U.: Impact of customer satisfaction on Brand Loyalty – An Empirical analysis of home appliances in Pakistan. Br. J. Mark. Stud. **2**, 18–32 (2014)

11. Aydin, S., Ozer, G.: The analysis of antecedents of customer loyalty in the Turkish Mobile telecommunications market. Eur. J. Mark. **39**, 910–925 (2005)
12. Azmat, M., Lakhani, A.S.: Impact of Brand positioning strategies on consumer standpoint (A consumer's Perception). J. Mark. Consum. Res. **14**, 109–116 (2015)
13. Bakator, M., Boric, S., Paunovic, M.: Influence of advertising on consumer-based Brand Loyalty. J. Eng. Manag. Competitiveness **7**(2), 75–83 (2017)
14. Brakus, J.J., Schmitt, B.H., Zarantonello L.: Brand experience; what is it? how is it measured? does it affect loyalty? J. Marketing 52–68 (2009)
15. Byung-Do, K., Sullivan, M.W.: "The effect of parent brand experience on line extension trial and repeat purchase", Marketing Lett. **9**(2) 181–193 (1998)
16. Chang, H., Fernando, G.D., and Tripathy, A.: an empirical study of strategic positioning and production. In: Adv. Operat. Res. 1–11 (2015)
17. Venkatesh, V.: Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. Inf. Syst. Res. **11**(4), 342–365 (2000)
18. Cengiz, H., Akdemir-Cengiz, H.: Review of brand loyalty literature: 2001–2015. J. Res. Marketing **6**(1), 407–434 (2016)
19. Kim, D.J., Ferrin, D.L., Rao, H.R.: A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decis. Support Syst. **44**(2), 544–564 (2008)
20. Gecti, F., Zengin, H.: The relationship between brand trust, brand affect, attitudinal loyalty and behavioral loyalty: A field study towards sports shoe consumers in Turkey. Int. J. Marketing Studies **5**(2), 111 (2013)
21. Chaudhuri, A., Holbrook, M.B.: The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. J. Mark. **65**(2), 81–93 (2001)
22. Parasuraman, A.: Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. J. Serv. Res. **2**(4), 307–320 (2000)
23. Legris, P., Ingham, J., Collerette, P.: Why do people use information technology? A critical review of the technology acceptance model. Inf. Manag. **40**(3), 191–220 (2003)
24. Erdoğmuş, İ, Ergun, S.: Understanding university brand loyalty: the mediating role of attitudes towards the department and university. Procedia Soc. Behav. Sci. **229**, 141–150 (2016)
25. Lee, J.E., Watkins, B.: YouTube vloggers' influence on consumer luxury brand perceptions and intentions. J. Bus. Res. **69**(12), 5753–5760 (2016)
26. Aaker, J. (1991). The negative attraction effect?A study of the attraction effect under judgment and choice. *ACR North American Advances*
27. Nagy, S., Hadjú, N.: Consumer acceptance of the use of artificial intelligence in online shopping: evidence from Hungary. Amfiteatru Econ. **23**(56), 155–173 (2021)
28. Kashiri, N., et al.: An overview on principles for energy efficient robot locomotion. Front. Robot. AI **5**, 129 (2018)

Performance Analysis of Turbo Codes for Wireless OFDM-based FSO Communication System

Ritu Gupta(✉)

Chandigarh University, Mohali 140413, India
erritugupta02@gmail.com

Abstract. This study evaluates the bit error rate (BER) performance of a turbo-coded orthogonal frequency division multiplexing (OFDM) communication connection using a terrestrial/wireless free-space optical (FSO) channel, taking atmospheric turbulence into account. The Lognormal probability density function (PDF) statistically represents the turbulence-induced intensity fading in the presence of weak turbulence. Furthermore, for enhancing BER performance, a novel low-complexity turbo code which is a channel coding scheme, and OFDM are recommended. The analysis is carried out using the bandwidth-efficient 16-QAM (quadrature amplitude modulation) modulation technology. According to the simulation results of the proposed model, the 16-QAM modulation scheme, turbo coding parameter, and connection length should be taken into account to ensure system dependability. The simulation demonstrates that over the course of a link length, a reliable communication link (10^{-9} BER) may be molded for 1 km in weak turbulence fading conditions. For easiness SISO (single input-single output) system has been studied for sustaining targeted BER (10^{-9}) in the existence of weak turbulence fading.

Keywords: Free-space optical communication · Turbo code · Atmospheric turbulence

1 Introduction

A point-to-point communication link is made possible by free space optics (FSO), also referred to as wireless optical communication (WOC) or visible light communication (VLC). FSO is now a skilled technology that can be used for both commercial and military applications. It serves as the framework for fiber optic infrastructure and can address "last-mile" or "bottleneck" issues with the current wireless technology [1–3].

The varying refractive index of the air results in channel fading, which is a random variation in the amplitude and phase of the received signal's strength. According to Andrews et al. research on atmospheric turbulence-induced fading, under clear air conditions, the attenuation related to visibility is insignificant [4]. For estimating received intensity, several statistical models have been described by Ghassemlooy et al. [5]. The

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 824–830, 2023.
https://doi.org/10.1007/978-3-031-27409-1_75

log-normal distribution model, one of these statistical models, can be used in weak turbulence situations [5, 6]. To fully exploit the accessible bandwidth, Orthogonal frequency division multiplexing (OFDM) can effectively use with FSO [7]. So far, various channel coding for error control and diversity techniques have been considered by researchers to fade the confines of the FSO communication system [8–11].

For FSO communication systems, several coding techniques have been investigated, including convolutional codes [11], turbo codes [12, 13], trellis-coded modulation (TCM) codes [14, 15], Polar codes [16] and Low-Density Parity Check (LDPC) codes [17]. Concatenated coding techniques including parallel concatenation, sequential concatenation, and serial concatenation have all been modified. Several concatenated techniques have been modified for the FSO communication system [18, 19]. For OFDM-based FSO communication systems with 16-QAM modulation, the performance of parallel concatenated coding techniques like turbo codes has been examined in this article under a weak turbulence regime. The analysis is appropriate under weak air turbulence conditions and is based on the bit error rate (BER) vs. signal-to-noise ratio (SNR) relationship. Researchers [20, 21] from many fields discuss the bit error performance of turbo codes.

The Paper is organized as in Sect. 2, the Turbo coded OFDM based FSO system model is introduced. In Sect. 3, FSO channel models for weak turbulence conditions are illustrated. Later, in Sect. 4, the performance analysis given for the OFDM-based FSO system is presented. The conclusions are presented in Sect. 5.

2 System Model

The structure of the single-input, single-output (SISO), Turbo-coded OFDM-based FSO communication system is presented in Fig. 1. In this paper, the channel coding for the said system has been taken into consideration using the 16-QAM modulation scheme. Turbo code is a type of error correction used in FSO communication systems to lessen system complexity and improve communication dependability. Contrary to previous studies, a Turbo coding technique is utilized to reduce continuing transmission defects and enhance the efficiency of information retrieval. When delivered to a typical Turbo encoder, the systematic bits of the original information sequence N is punctured and multiplexed to produce the first sequence that is Turbo-coded. Concurrent convolutional encoders are used by the Turbo encoder to encrypt the N-bit input stream. There has been discussion of a convolution encoder that uses a 16-state, 2.5-rate turbo encoder. To effectively counteract recurrent bit errors, randomization is employed to increase some redundancy.

The Turbo encoder processes the N-bit input stream, encoding the input bits using concurrent convolutional encoders. A convolution encoder with a 16-state, two-and-a-half rate turbo encoder has been considered. In FSO communication systems, error correction is done using turbo code. In contrast to past research, a Turbo coding method is used to decrease ongoing transmission faults and improve the effectiveness of information retrieval. The systematic and parity bits are punctured and multiplexed in a typical Turbo encoder, which receives the original information sequence N bits, to produce the initial Turbo-coded sequence M. Random interleaver is used to add some redundancy in order to effectively combat persistent bit errors.

The N-bit input stream is passed through the Turbo encoder, where the input bits are encoded by parallel convolutional encoders. Turbo encoder having 16-states, two ½ rate convolution encoder has been taken into consideration. Turbo code is used for error correction in FSO communication systems. By comparison with the earlier research, a Turbo coding method is adopted to reduce continuous errors in the transmission process for retrieving the information more effectively. The original information sequence N bits are transmitted to a standard Turbo encoder, where the systematic and parity bits are punctured and multiplexed to obtain an initial Turbo-coded sequence M. To overcome incessant bit errors efficiently some redundancy is added via random interleaver.

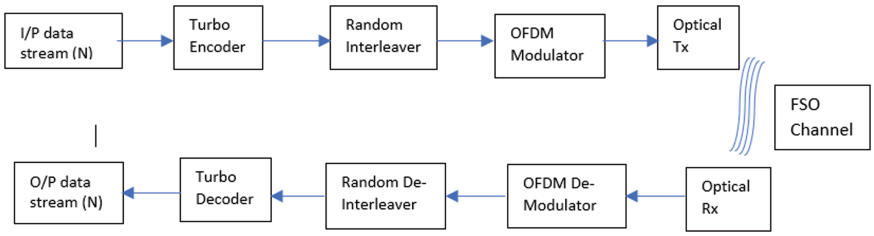


Fig. 1. Structure of the considered Turbo coded OFDM-based FSO communication system.

The first and second issues with turbo code are the size of the interleaver and the free distance and other factors for a specific length of interleaving, respectively The bit stream is then organized, and burst mistakes are eliminated by passing it via a random interleaver. as it has done The 16-QAM modulator is then applied to the sequence. For 16QAM, the QAM sequence generator that creates the in-phase (I) and quadrature signals, 4 bits per symbol is taken into consideration (Q).

The data is then transmitted by a laser-based optical transmitter after being altered by an OFDM modulator. Digital data can be encoded using the Orthogonal Frequency Division Multiplexing (OFDM) technique on various carrier frequencies. In multicarrier modulation methods, a high-rate data stream is divided into numerous lower-rate streams and sent simultaneously via several subcarriers or several narrowband channels at low data rates. The sequence is later passed through an OFDM modulator for alterations. The data is subsequently delivered using a laser-based optical transmitter. The main element of the transmitter is the Inverse Fourier Transform (IFFT) block, whereas the receiver's Fast Fourier Transform (FFT) is the opposite. A complex vector with the form Xm = [X0 X1 X2… XN1] serves as the IFFT's input, with N denoting the device's size and Xm denoting its transmit signals. The multipath fading problem that could arise while sending fast data across the atmospheric medium has a possible fix thanks to this work. A significant number of orthogonal subcarriers are used in the multicarrier digital modulation method known as OFDM. The information signal will be changed into a parallel form and arranged to overlap with distances prior to the information transfer, giving it orthogonal properties. The transmitted optical beam passes through the atmosphere. The atmosphere has been simulated with modest turbulence using the lognormal distribution model. The reverse method has been used at the receiving end.

The optical receiver initially demodulates the signal before decoding it. Table 1 lists several parameters taken into account for the mentioned system.

Table 1. System parameters used in the proposed model.

Parameters	Values
Input bit stream (N)	32400 bits
Coding scheme	Turbo code
Modulation scheme	16-QAM
Wavelength (λ)	1550 nm
Link range	1 km

3 Channel Model

An unbounded plane wave operating under mild irradiance fluctuation conditions is considered, and a straightforward and computationally efficient model for FSO communications is proposed. The obtained results are in excellent accord with theoretical estimates, demonstrating that channel coherence may be an important issue for a thorough understanding of FSO communications.

The major models are (1) the Log normal model (weak turbulence) (2) Gamma-Gamma (weak to strong turbulence (3) the Negative exponential model (strong turbulence). The major models are (1) the Log normal model (weak turbulence) (2) Gamma-Gamma (weak to strong turbulence (3) the Negative exponential model (strong turbulence). The turbulence in the atmosphere has been studied, and several channel models have been put forth [5]. The impacts of turbulence are examined in this work using the lognormal model as the channel model.

An unbounded plane wave operating under mild irradiance fluctuation conditions is described together with a straightforward and computationally effective model for free-space optical (FSO) communications. The obtained results are in excellent accord with theoretical estimates, demonstrating that channel coherence may be an important issue for a thorough understanding of FSO communications. The receiving irradiance fluctuation's probability density function is specified in Eq. 1 [5].

$$p(I) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \cdot \frac{1}{I} \cdot e^{\frac{-(In\frac{I}{I_0}-\mu)^2}{2\sigma_N^2}} \tag{1}$$

where I = signal intensity assuming the mean of I is 1 due to the normalized channel effect.

I_0 = intensity in free space (no turbulence).

μ = mean log intensity E[I].

$\sigma_N{}^2$ is the Raytov Variance or variance of light intensity and is given by Eq. 2

$$\sigma_N^2 = 1.23 C_N^2 K^{\frac{7}{6}} L^{\frac{11}{6}} \tag{2}$$

$C_N{}^2$ is the refractive-index structure parameter and varies from 10^{-17} to 10^{-13} $m^{-2/3}$ according to the atmospheric turbulence conditions.

k (wave number) $= \frac{2\pi}{\lambda}$

L is the distance between the transmitter and the receiver.

4 Simulation Results

The simulation performance has been analyzed in MATLAB and the Monte Carlo method has additionally been followed for the verification of effects through random attempts. For the considered device, simulation consequences have additionally been compared with the uncoded gadget below vulnerable turbulence situations and given in the following sections.

The log-normal distribution version is followed for characterizing susceptible turbulence, together with maritime or mild fog. The performance evaluation for faster code with 16-QAM modulation schemes has been presented in Fig. 2. The outcomes tested that the rapid code with 16-QAM has shown improved overall performance than convolutional coded and uncoded OFDM-primarily based FSO device. For accomplishing BER of 10-eight, the faster code system has shown development of 3 dB and 6 dB compared to convolutional and uncoded structures respectively.

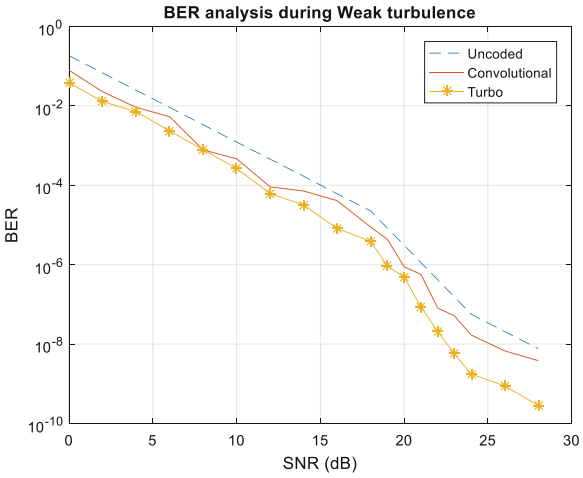


Fig. 2. System's performance during weak turbulence (light fog)

The use of coding schemes in the communication system improves its performance with a definite increase in its complexity.

5 Conclusion

This paper examines the BER performance of a wireless OFDM-primarily based FSO conversation system with a turbo coding scheme below susceptible turbulence conditions. The evaluation indicates that for the considered model, rapid code has shown higher coding benefits as compared to convolutional codes with similar parameters and uncoded systems. For modulation, a bandwidth-green 16-QAM modulation scheme has been adapted and effectively implemented to the OFDM-based FSO conversation implement. The said implement added a substantial amount of improvement to the system's overall performance.

References

1. Kedar, D., Arnon, S.: Urban optical wireless communication networks: the main challenges and possible solutions. IEEE Commun. Mag. **42**(5), S2–S7 (2004)
2. Henniger H., Wilfert, O.: An Introduction to free-space optical communications. Radio Engineering, **19** (2) (2010)
3. Duvey, D., Gupta, R.: Review paper on performance analysis of a free space optical system. Int. J. Appl. Innov. Eng. & Manag. IJAIEM **3**(6), 135–139 (2014)
4. Andrews, L. C., Philips R. L., Hopen, C.Y.: Laser Beam Scintillation with Applications. SPIE Optical Engineering Press, Bellingham, WA
5. Ghassemlooy, Z., Popoola, W., Rajbhandari, S.: Optical Wireless Communications System and Channel Modelling. CRC Press Taylor & Francis Group, pp.378 (2013)
6. Al-Habash, M. A., Andrews, L. C., Phillips, R. L.: Mathematical model for the irradiance probability density function of a laser beam propagating through turbulent media. Opt. Eng. 1554–1562 (2001)
7. Wang, Y., Wang, D., Jing, Ma.: On the performance of coherent OFDM systems in free-space optical communication. IEEE Photon. J. **7**(4) (2015)
8. Fang, X., et al.: Channel coding and time-diversity for Optical Wireless Links. Opt. Express **17**(2), 872–887 (2009)
9. Theodoros, A.T., Harilaos, G., Sandalidis, G., Karagiannidis, G.K.: Optical wireless links with spatial diversity over strong atmospheric turbulence channels. IEEE Trans. Wireless Commun. **8**(2), 951–957 (2009)
10. Zhu, X., Kahn, J.M.: Performance bounds for coded free-space optical communications through atmospheric turbulence channels. IEEE Trans. Commun. **51**(8), 1233–1239 (2003)
11. Gupta, N., Prakash, S. J., Kaushal, H., Jain, V. K., Kar, S.: Performance analysis of FSO communication using different coding schemes. AIP Conf. Proc. 387–391 (2011)
12. Pham, A T., Thang, T C., Guo, S., Cheng, Z.: Performance bounds for Turbo-coded SC-PSK/FSO communications over strong turbulence channels. IEEE ATC-(2011)
13. Hassan, MD. ZT., Bhuiyan, A., Tanzil, S.M.S., Majumder, S.P.: Turbo-Coded MC-CDMA communication link over strong turbulence fading limited FSO channel with receiver space diversity. ISRN Commun. Netw. **14** (2011)
14. Gupta, R., Kamal, T.S.: Performance analysis of OFDM based FSO communication system with TCM codes. Int. J. Light. Electron Opt., Optik **248** (2021)
15. Park, H., John, R.B.: Trellis-coded multiple-pulse-position modulation for wireless infrared communications. IEEE Trans. Commun. **52**(4), 643–652 (2004)
16. Mohan, N., Ghassemlooy, Z., Emma Li, Abadi, M.M., Zvanovec, S., Hudson, R., Hta, Z.: The BER performance of a FSO system with polar codes under weak turbulence. IET Opotelectron. **16**, 72–80 (2022)

17. Sonali, Dixit, A., Jain, V. K.: Analysis of LDPC codes in FSO communication system under fast fading channel conditions. IEEE Commun. Soc. **2**, 1663–1673 (2021)
18. Gupta, R., Kamal, T. S., Singh, P.: Concatenated LDPC-TCM codes for better performance of OFDM-FSO system using Gamma–Gamma fading model. Springer: Wireless Personal Commun. (2018)
19. Gupta, R., Kamal, T. S., Singh, P.: Performance of OFDM: FSO communication system with hybrid channel codes during weak turbulence. Hindawi: J. Comput. Netw. Commun. (2019)
20. Vucetic, B., Yuan, J.: Turbo codes: principles and applications. Kluwer Academic Publishers (2000)
21. Garello, R., Pierleoni, P., Benedetto, S.: Computing the free distance of turbo codes and serially concatenated codes with interleavers: algorithms and applications. IEEE J. Sel. Areas Commun. **19**(5), 800–812 (2001)

Optimal Sizing and Placement of Distributed Generation in Eastern Grid of Bhutan Using Genetic Algorithm

Rajesh Rai[1], Roshan Dahal[1], Kinley Wangchuk[1], Sonam Dorji[1], K. Praghash[2(✉)], and S. Chidambaram[2]

[1] Department of Electrical Engineering, Jigme Namgyel Engineering College, Dewathang, Bhutan

[2] Department of Electronics and Communication Engineering, Christ University, Bengaluru, India

prakashcospra@gmail.com

Abstract. Power system has to be stable and reliable for its users. Nevertheless, due to the aging and ignorance, it tends to be unreliable and unstable. Distributed Generation (DG) is a small-scale energy production which are usually connected towards the load. It helps in the reduction of power losses and improvement of profile of voltage in a distribution network. However, if a DG is not optimally placed and sized, it will rather lead to an increase in a power loss and also deteriorates' the voltage profile. This report exhibits the importance of DG placement and sizing in a distribution network using Genetic Algorithm (GA). Apart from the optimum DG placement and sizing, different scenarios with numbers of DGs is also being analyzed in this report. On eastern grid of Bhutan, a detailed analysis for its performance is carried out through MATLAB platform to demonstrate and study the effectiveness and reliability of a methodology that is being proposed.

Keywords: Active Power Loss · Voltage Profile · Genetic Algorithm · Optimization · Distributed Generation

1 Introduction

According to Ackermann, distribution generation (DG) is an electric power source located on the meter placed on the client side or directly connected to the distribution network [1]. While this method is relatively new in terms of electricity market economics, the concept behind is not. The use of any modular technology put throughout a utility's service area (and connected to the distribution or sub-transmission system) to lower service costs is also referred to as "distributed generation" (DG) [2]. The need for power is growing significantly. DG is one of the finest options for meeting the ever-increasing energy demand. DG units are tiny power plants that are directly connected to the distribution network or located at the client site of the meter (also known as decentralized generation, dispersed generation, and embedded generation) [2]. In addition to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 831–840, 2023.
https://doi.org/10.1007/978-3-031-27409-1_76

their positive effects on the environment, DGs also aid in the implementation of competitive energy policies, resource diversification, on-peak operational cost reduction, and the deferral of system upgrades. It also minimizes the energy loss in the system, relieves transmission congestion, improves voltage profile, increases reliability, and lowers operating costs. Installation of DG is more flexible in terms of time and expenditure because of its modest size compared to traditional generation units. Furthermore, DGs come in modular modules, which make it easier to choose locations for generators in small size, save construction time, and cut capital costs. Owners and investors make decisions on DG placement based on site and primary fuel supply, as well as meteorological circumstances. Many problems in power system such as voltage control and power loss, can be handled using distributed generation (DG). In the future power distribution system, decentralized generating will become progressively essential. Small-scale production units are now commercially available (such as fuel cells, micro-CHPs, and solar panels), and the energy market is now deregulated and it has increased this development, placing greater strain on the system [3]. Energy cost control is essential to profitability due to price fluctuations brought on by utility deregulation and system stability [4]. By ensuring that the voltage and current waveforms sent to the consumer meet several international standards, quality power is provided to various home and industrial uses [5]. Since electricity is now so essential to daily life, it is required to protect the power system from harm during faulty situations and to provide the greatest possible supply continuity [6].

2 Problem Formulation

2.1 Load Flow Equation

When simulating the power flow by thorough listing of all potential combinations of locations and DG sizes in the network, it will results in the perfect answer to the combined sizing and placement challenge. Kirchhoff's Laws are used to derive the load flow equations

$$P_i - PD_i = \sum_j |V_i||V_j||Y_{i,j}\cos(\theta_{i,j} + \delta_j - \delta_i) \tag{1}$$

$$Q_i - QD_i = \sum_j |V_i||V_j||Y_{i,j}\sin(\theta_{i,j} + \delta_j - \delta_i) \tag{2}$$

2.2 Optimal DG Allocation

Meanwhile the challenge of optimum location and size goes through "combinative explosion," and for this reason, use of heuristic techniques and artificial intelligence to solve it is justifiable. Although analytical approaches are more accurate for simple objective functions than heuristic methods, solutions for a non-simple, i.e., discrete, issue like optimum DG placement are more prone to be caught in local optima and not global optimum. Heuristic approaches, on the other hand, rely on a orderly arbitrary search of the solution space to improve the odds of discovering the global optimum.

The subject of optimal DG size has been studied extensively, with the DG often being treated as optimal active power injection. Real power loss reduction is the most prevalent goal in scientific literature. Internal resistance causes dissipation of power in systems' equipment and components which includes transmission and distribution lines, transformers, and measuring equipment, which is referred to as power loss. Real power loss is defined as PLoss = I^2 R, or while in voltages and angles, it can be also written as PLoss = V^2/R.

3 Optimization

3.1 Genetic Algorithm

It is based on a natural selection that helps in solving constrained as well as unconstrained optimization problems [21]. It will repeatedly adjust a population of peculiar and exceptional solutions. At each phase, it will select members of the present population which will serve as parents and uses them to produce the offspring of the succeeding generation. After certain number of generations, the population "develops" toward the best course of action. The evolutionary algorithm is applicable to different kinds of optimization issues, including those with nondifferentiable, stochastic, discontinuous, or highly nonlinear objective functions that are not well suited for conventional optimization techniques. When dealing with mixed integer programming problems where some of the components must have integer values, the evolutionary algorithm might be employed [22] (Fig. 1).

3.2 Optimal Allocation of Distributed Generation (Mathematical Model)

3.2.1 Objective Function

$$Min\, P_{loss}(x, u) = \sum\nolimits_{z=1}^{N_{br}} |I_z|^2 R_z \tag{3}$$

3.2.2 Equality Constraint

$$\sum\nolimits_{i=0}^{NG} P_{gi} = P_{loss} + P_d \tag{4}$$

3.2.3 Inequality Constraints

- Voltage limits

$$V_{gi}^{min} \leq V_{gi} \leq V_{gi}^{max} \tag{5}$$

- DG size limit

$$DG_{gi}^{min} \leq DG_{gi} \leq DG_{gi}^{max} \tag{6}$$

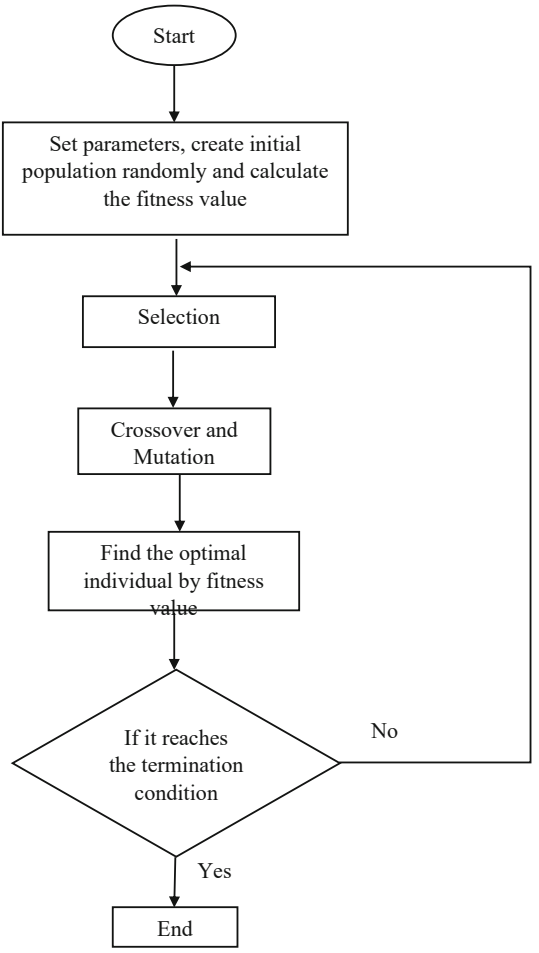


Fig. 1 Basic structure of genetic algorithm

- Active power generation limit

$$P_i^{min} \leq P_i \leq P_{gi}^{max} \tag{7}$$

where, z and Nbr are correspondingly, the branch number and the branches in total. Iz is current's absolute value flowing over the branch and Rz is the branch's' resistance. Ploss means an active power loss.

NG is a number of DG that has been installed, Pgi is a power of the DG installed and Pd is a power that is demanded by load.

Where V_{gi}^{min} and V_{gi}^{max} are the lower and upper operating voltages at each bus, DG_{gi}^{min} and DG_{gi}^{max} are the lower and upper levels of active power. Generated by DG respectively and P_i^{min} and P_{gi}^{max} are lower and upper limits of active power generation correspondingly.

3.3 Impacts of Distributed Generation on Distribution Network

- Distributed generation can impact a network's power losses; if the placement and magnitude of distributed generation are inadequate, power losses may increase, resulting in voltages that are excessively high or low on some nodes. Furthermore, certain forms of distributed generating, such as wind and solar electricity, are insecure or intermittent. As a result, dispersed generation may cause some operator uncertainty.
- Creating a distributed network often requires five to ten years of planning. Demand on the network could increase during this time. Network optimization is made more difficult by the addition of distributed generation since scheduling a distributed system is a energetic process.
- When it has access to a distributed network, it may deliver electricity to installed areas in need of extra power, lowering network investment costs. However, if the area's electrical supply is already enough, dispersed generation won't be completely utilized. Money and energy could be wasted as a result of this situation.
- Distributed generating installation helps in an improvement of a networks' power quality. The voltage profile may be improved by properly allocating distributed generation, which involves bringing the voltages nearer to the rated voltage in order to boost its effectiveness. Installing distributed generation may give electricity to the loads when the load is high, making the network more reliable. However, certain dispersed generating contributions are affected by weather or the environment. Voltage fluctuations, flicker, and droop may occur as a result of distributed generation's abrupt starts and stops [34].
- The reliability of distributed generation can be increased or decreased. To improve reliability, distributed generation can send electricity to nearby loads. If a system failure occurs, distributed generation can switch to islanded operation to provide electricity.

4 Optimization

The following simulated results are performed using MATLAB software.

4.1 Load Flow Analysis of Eastern Grid of Bhutan

See Fig. 2.

4.2 Case 1: With 4 DGs

4.2.1 Active Power Losses

See Table 1.

4.2.2 Simulated Size and Locations

See Table 2.

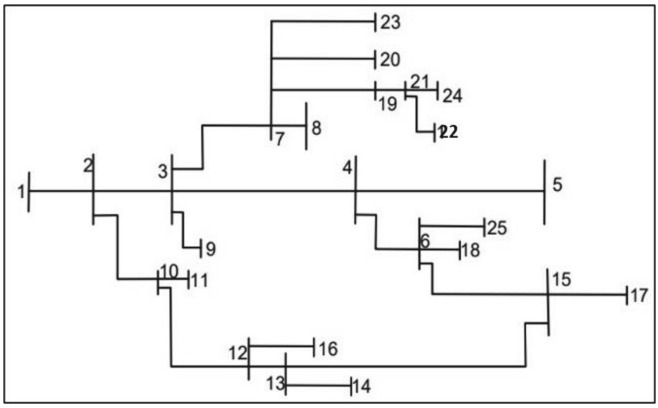


Fig. 2 Eastern grid bus bar system

Table 1 Losses obtained before and after 4 dgs

	Losses before DG (MW)	Losses after DG (MW)	Reduction in losses (%)
Active Power Losses with 4 DGs (Ploss = I^2R)	2.2127	1.20	45.8

Table 2 Optimal size and location of 4 dgs obtained

Bus location	DG size (kW)
24	3.40
22	113.90
19	0.90
6	0.80

4.2.3 Comparison of Voltage Before and After DG

4.3 Case 2: With 5 DGs

4.3.1 Active Power Losses

See Table 3.

4.3.2 Simulated Size and Location

See Table 4.

Table 3 Losses obtained before and after 5 dgs

	Losses before DG (MW)	Losses after DG (kW)	Reduction in losses (%)
Active Power Losses with 4 DGs (Ploss = I^2R)	2.2127	650	70.6

Table 4 Optimal size and location of 5 dgs obtained

Sl. No	Bus Location	DG size (kW)
1	16	17
2	22	3
3	13	2.5
4	6	23
5	24	35

4.3.3 Comparison of Voltage Before and After DG

See Fig. 3.

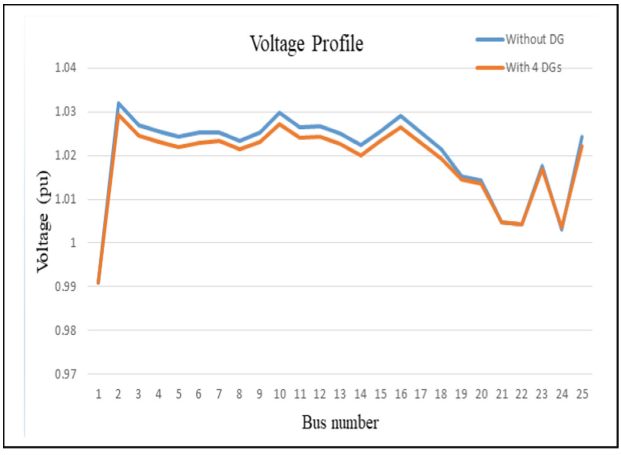


Fig. 3 Voltage profile before and after 5 DGs.

5 Conclusion

Losses of a power in a system is likely to happen because the efficiency of equipment deteriorates as it ages. However, the losses can be reduced to a great extent using various

methods out of which, DG is one of them. In order to minimize an active power loss and to improve the distribution system's voltage, DG has to be sized and placed optimally using any methods depending upon their efficiency and effectiveness. In this project, DGs have been sized and placed optimally in eastern grid of Bhutan using genetic algorithm. It was found out that the losses are reduced to a great extent as the number of DG increases, however after certain number of DGs, it becomes unproductive and increase the losses. For this very reason, the number of DGs cannot be increased arbitrarily and additionally, the size and placement of DGs cannot be chosen randomly as it increases the power losses. Thus, it can be concluded that DGs plays a vital role in reducing the losses in a distribution system, nevertheless, one must take a note of a fact that it will be effective only when it is being placed and sized optimally.

References

1. Singh, B., Mishra, D.K.: A survey on enhancement of power system performances by optimally placed DG in distribution networks. Energy Rep. **4**, 129–158 (2018). https://doi.org/10.1016/j.egyr.2018.01.004
2. Khosravi, M.: Optimal placement distributed generation by genetic algorithm to minimize losses in radial distribution systems. Bull. Env. Pharmacol. Life Sci, **3**(August), 85–91 2014, [Online]. Available: http://www.bepls.com/aug_2014/15f.pdf.
3. Peter, G., Sherine, A., Iderus, S.B.: Enhanced Z-source inverter-based voltage frequency generator to conduct induced over voltage test on power transformers. IJPELEC **12**(4), 493 (2020). https://doi.org/10.1504/IJPELEC.2020.110752
4. Iderus, S., Peter, G., Praghash, K., Vadde, A.R.: Optimization and design of a sustainable industrial grid system. Math. Probl. Eng. **2022**, 1–12 (2022). https://doi.org/10.1155/2022/4418329
5. Peter, G., Praghash, K., Sherine, A., Ganji, V.: A combined PWM and AEM-based AC voltage controller for resistive loads. Math. Probl. Eng., **2022** (2022), doi: https://doi.org/10.1155/2022/9246050.
6. Iderus, S., Peter, G., Ganji, V.: An innovative method to conduct temperature rise test on medium voltage switchgear assembly based on IEC standards in a power grid. J. Eng., **June**, 1–23, 2022, doi: https://doi.org/10.1049/tje2.12166.
7. Sattianadan, D., Sudhakaran, M., Dash, S.S., Vijayakumar, K., Biswal, B.: Power loss minimization by the placement of DG in distribution system using PSO. In: Satapathy, S.C., Udgata, S.K., Biswal, B.N. (eds.) Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA). AISC, vol. 199, pp. 497–504. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35314-7_56
8. Sattianadan, D., Sudhakaran, M., Dash, S.S., Vijayakumar, K., Ravindran, P.: Optimal Placement of DG in Distribution System Using Genetic Algorithm. In: Panigrahi, B.K., Suganthan, P.N., Das, S., Dash, S.S. (eds.) SEMCCO 2013. LNCS, vol. 8298, pp. 639–647. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-03756-1_57
9. Ayodele, T.R., Ogunjuyigbe, A.S.O., Akinola, O.O.: Optimal location, sizing, and appropriate technology selection of distributed generators for minimizing power loss using genetic algorithm. J. Renew. Energy **2015**, 1–9 (2015). https://doi.org/10.1155/2015/832917
10. Viswa Teja, R., Maheswarapu, S.: Optimal placement and sizing of distributed generators in radial distribution systems using imperialist competitive algorithm, pp. 1–6
11. Mohan, V.J., Albert, T.A.D.: Optimal sizing and sitting of distributed generation using Particle Swarm Optimization Guided Genetic Algorithm. Adv. Comput. Sci. Technol. **10**(5), 709–720 (2017)

12. Ramamoorthy, S.: Design and implementation of fuzzy logic based power system stabilizers. Middle – East J. Sci. Res. **20**(11), 1663–1666 (2014). https://doi.org/10.5829/idosi.mejsr.2014.20.11.1932
13. Tolba, M. A., Tulsky, V. N., Diab, A. A. Z.: Optimal sitting and sizing of renewable distributed generations in distribution networks using a hybrid PSOGSA optimization algorithm. In: Conf. Proc. – 2017 17th IEEE Int. Conf. Environ. Electr. Eng. 2017 1st IEEE Ind. Commer. Power Syst. Eur. EEEIC/I CPS Eur. 2017, 2017, doi: https://doi.org/10.1109/EEEIC.2017.7977441
14. Karunarathne, E., Pasupuleti, J., Ekanayake, J., Almeida, D.: The optimal placement and sizing of distributed generation in an active distribution network with several soft open points. Energies **14**(4) 2021, doi: https://doi.org/10.3390/en14041084.
15. Gidd, M. M., Mhetre, S. L., Korachagaon, I. M.: Optimum position and optimum size of the distributed generators for different bus network using genetic algorithm. In Proc. – 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi: https://doi.org/10.1109/ICCUBEA.2018.8697595
16. Sedighi, M., Igderi, A., Dankoob, A., Abedi, S. M.: Sitting and sizing of DG in distribution network to improve of several parameters by PSO algorithm. ICMET 2010 – 2010 Int. Conf. Mech. Electr. Technol. Proc., no. Icmet, pp. 533–538, 2010, doi: https://doi.org/10.1109/ICMET.2010.5598418.
17. Moradi, M.H., Abedini, M.: A combination of genetic algorithm and particle swarm optimization for optimal distributed generation location and sizing in distribution systems with fuzzy optimal theory. Int. J. Green Energy **9**(7), 641–660 (2012). https://doi.org/10.1080/15435075.2011.625590
18. Singh, D., Singh, D., Verma, K.S.: GA based energy loss minimization approach for optimal sizing & placement of distributed generation. Int. J. Knowledge-Based Intell. Eng. Syst. **12**(2), 147–156 (2008). https://doi.org/10.3233/KES-2008-12206
19. Peter, G., Livin, J., Sherine, A.: Hybrid optimization algorithm based optimal resource allocation for cooperative cognitive radio network. Array **12**, 100093 (2021). https://doi.org/10.1016/j.array.2021.100093
20. Faraji, H., Hajimirzaalian, H., Farzadpour, F., Legha, M. M.: A new hybrid particle swarm optimization approach for sizing and placement enhancement of distributed generation. Int. Conf. Power Eng. Energy Electr. Drives, (May) 1277–1281 (2013), doi: https://doi.org/10.1109/PowerEng.2013.6635796.
21. Srikanth, P., Rajendra, O., Yesuraj, A., Tilak, M., Raja, K.: Load flow analysis of IEEE 14 bus system using MATLAB. Int. J. Eng. Res. Technol. **2**(5), 149–155 (2013)
22. Husain, T., Khan, M., Ansari, M.: Power flow analysis of distribution system. pp. 4058–4065, 2016, doi: https://doi.org/10.15662/IJAREEIE.2016.0505108.
23. Martinez, J. A., Mahseredjian, J.: Load flow calculations in distribution systems with distributed resources. A review. IEEE Power Energy Soc. Gen. Meet., pp. 1–8 (2011), doi: https://doi.org/10.1109/PES.2011.6039172.
24. Geno, P.: A review about vector group connections in transformers. Int. J. Adv. Technol. **2**(2), 2011. ISSN : 0976–4860 (Online).
25. Lambora, A., Gupta, K., Chopra, K.: Genetic algorithm – a literature review. In 2019 Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. **1998**, 380–384 (2019)
26. Immanuel, S. D., Chakraborty, U. K.: Genetic algorithm: an approach on optimization. In: Proc. 4th Int. Conf. Commun. Electron. Syst. ICCES 2019, no. Icces, pp. 701–708, 2019, doi: https://doi.org/10.1109/ICCES45898.2019.9002372.
27. Drachal, K., Pawłowski, M.: A review of the applications of genetic algorithms to forecasting prices of commodities, Economies **9**(1), 2021, doi: https://doi.org/10.3390/economies9010006.

28. Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multimed. Tools Appl. **80**(5), 8091–8126 (2020). https://doi.org/10.1007/s11042-020-10139-6
29. Saleh, S. A.: A genetic algorithm for solving an optimization problem : decision making in project management, pp. 221–225 (2020)
30. Man, K. F., Tang, K. S., Kwong, S.: Genetic algorithms : concepts and applications, **43**(5) (1996)
31. Mukerji, M.: Optimal siting and sizing of solar photovoltaic distributed generation to minimize loss, present value of future asset upgrades and peak demand costs on a real distribution feeder, pp. 1–97 (2011)
32. Arief, A., Dong, Z. Y., Lumpur, K., Kong, H.: Determination of DG allocation with modal participation factor to enhance voltage.
33. Cui, H.: Optimal allocation of distributed generation in distributed network. Asia–Pacific Power Energy Eng. Conf. APPEEC, no. November (2012), doi: https://doi.org/10.1109/APPEEC.2012.6307702.
34. Peter, G., Stonier, A. A., Gupta, P., Gavilanes, D., Vergara, M. M., Lung sin, J.: Smart fault monitoring and normalizing of a power distribution system using IoT. Energies **15**(21) 8206 (2022), doi: https://doi.org/10.3390/en15218206.

ANN Based MPPT Using Boost Converter for Solar Water Pumping Using DC Motor

Tshewang Jurme[1], Thinley Phelgay[1], Pema Gyeltshen[1], Sonam Dorji[1], Thinley Tobgay[1], K. Praghash[2(✉)], and S. Chidambaram[2]

[1] Department of Electrical Engineering, Jigme Namgyel Engineering College, Dewathang, Bhutan

[2] Department of Electronics and Communication Engineering, Christ University, Bengaluru, India

prakashcospra@gmail.com

Abstract. The solar DC pump system is simple to set up and run completely on its own without the need for human intervention. Solar DC pumps require fewer solar panels to operate than AC pumps. Solar PV Arrays, a solar DC regulator, and a DC pump make up the Solar DC Pump system. The nonlinear I-V characteristics of solar cells, PV modules have average efficiency compare to other forms of energy, and output power is affected by solar isolation and ambient temperature. The prominent factor to remember is that there will be a significant power loss owing to a failure to correspond between the source and the load. In order to get the most power to load from the PV panel, MPPT is implemented in the converter circuit using PWM and a microcontroller. In order to give the most power to load from the source, the solar power system should be designed to its full potential.

Keywords: MPPT · PID · ANN · SPV · BLDC · Boost Converter

1 Introduction

As the earth's natural resources dwindle, the power sector is exploring for alternative energy sources for the rising demand of electricity. The carbon concentration in the atmosphere can be lowered by the use of renewable energy sources, so resolving the global warming issue. Due to its simple design, solar PV systems are presently the most widely used renewable energy source. The efficiency of the PV system may be raised by integrating power electronic devices with a maximum power point controller [1]. A gadget called a Maximum Power Point Tracking (MPPT) controller draws the most power possible from a solar panel. A maximum power point tracker may significantly increase a solar system's effectiveness. Since they don't need understanding of internal system characteristics, demand less processing power, and offer a compact solution for multivariable issues, artificial intelligence (AI) approaches are being employed more and more as a replacement for conventional physical modelling techniques. They have been used to solve challenging practical issues in many different domains, and they are now more frequently seen in PV systems with non-linear characteristics [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 841–851, 2023.
https://doi.org/10.1007/978-3-031-27409-1_77

2 Configuration of System

The output voltage of the PV panel is computed with the voltage of ANN MPPT. The computed voltage is tuned by the PID controller and the PWM generator generates the pulse signal which feeds the switch of the IGBT of the boost converter. Upon requirement of the load, the voltage is supplied from IGBT (Fig. 1).

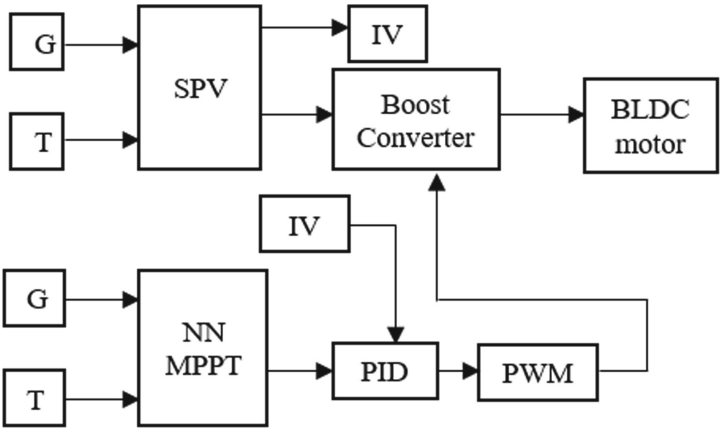


Fig. 1. Block diagram of PV generation system

The Brushless DC (BLDC) motor of 0.48 HP, 359 Watts, and discharge capacity of 19 LPM is used. Brushless motors provide longer life, more versatility, and a more noticeable operation. When compared to other motors of similar size, BLDC motors have a higher torque-to-speed ratio. It has a wider speed range than other motors, as well as more torque and speed. Because of the low rotor inertia, it improves dynamic performance while reducing the operating cycle [3]. The system has a vertical head of 30m. For the water to be pumped, 20a mm dia. (3/4') HDPE pipeline of 50m is used. The Table 1 shows the specification pf BLDC motor.

Table 1. Specification of BLDC motor

Rated power	359 Watts, 0.48 HP
Terminal voltage	90 V
Nominal current	5 A
Nominal speed	1500 RPM
Torque	2.26 Nm
Rotor inertia	21890.7 g cm^2

3 Design of SPV Array

The To have the required power to the load i.e., 359 Watts, the two parallel strings of 1Soltech 1STH-FRL-4H-250-M60-BLK PV panel with 250 Watts each is used for the Simulink. The Table 2 shows the specification of PV panel used for the simulation.

Table 2. Specification of pv panel

Maximum power (W)	250
Open circuit voltage V_{OC} (V)	38.4
The voltage at maximum power point V_{MP} (V)	30.7
Temperature coefficient of V_{OC} (%/deg.C)	−0.35599
Cells per module (Ncell)	60
Short circuit current I_{SC} (A)	8.85
Current at Maximum Power Point I_{MP} (A)	8.11
Temperature Coefficient of I_{SC} (%/deg.C)	0.07

4 PID Controller and PWM Generator

An the output voltage from the PV panel and ANN MPPT is computed. The controller tunes the computed signal. The PID controller is used and the performance of the control is detailed in Table 3.

Table 3. Performance of PID controller

Rise time	0.315 s
Settling time	0.875 s
Overshoot	4.48%
Peak	1.04
Gain margin	30.4 dB @ 33 rad/s
Phase margin	64.5 deg @ 4.2 rad/s
Close-loop stability	Stable

The PWM generator generates a pulse signal and the switching frequency of 10 kHz is selected for the boost converter for the simulation. The PWM signal is used to switch the IGBT of boost converter and as per the load requirement the required voltage is supplied [3]. The Fig. 2 shows the PMW signal used for this sudy.

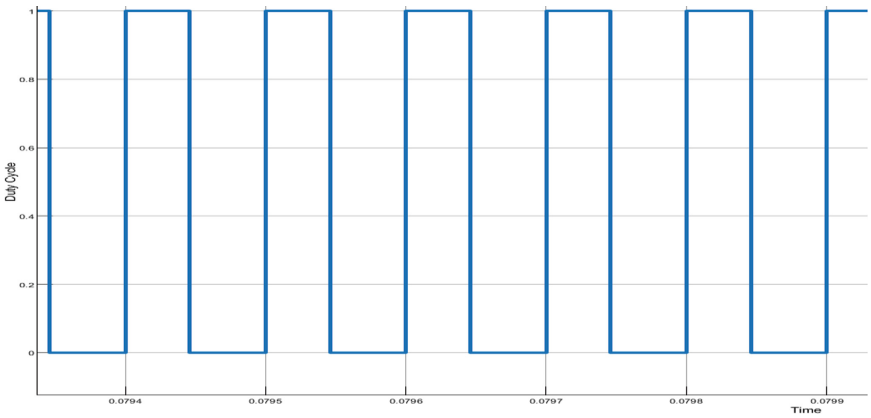


Fig. 2. PMW signal

5 Boost Converter Design

The maximum power point voltage i.e., 30.7 V is boosted to 90 V which is the input to the BLDC motor in this simulation. The value of inductance, capacitance and gain is calculated using these formulae;

$$\mathrm{L} = (\mathrm{Vin} * (\mathrm{Vout} - \mathrm{Vin}))/(\Delta \mathrm{IL} * \mathrm{fs} * \mathrm{Vout})$$

$$\mathrm{C} = (\mathrm{Ioutmax} * (1 - (\mathrm{Vin}/\mathrm{Vout})))/(\mathrm{fs} * \Delta \mathrm{Vout})$$

$$\frac{p}{\omega_r^3}$$

The value obtained is given in the design of the boost converter for the Simulink (Fig. 3). The Table 4 shows the parameter value of boost converter.

Table 4. Parameter value of boost converter

Inductor	0.0124 H
Capacitor	0.00040672 F
Gain	1.29E–04

6 Neural Nework

A depiction of linked synthetic neurons (nodes) called an artificial neural network (ANN) resembles the organization of a biological brain. An ANN typically has three layers:

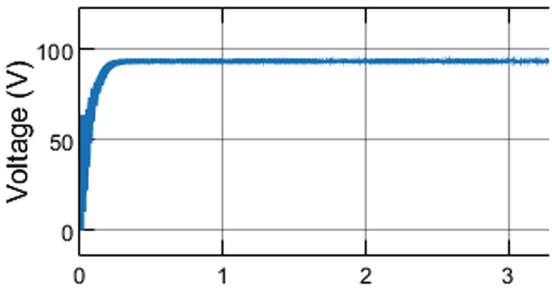


Fig. 3. Input voltage to load

input, hidden, and output. Irradiance, temperature, VOC, and ISC are frequently used as input layers, whereas voltage, duty cycle, or current are commonly used as output layers. The user chooses the number of nodes in each layer, which changes based on the requirement [4].The Fig. 4 shows the ANN network configuration.

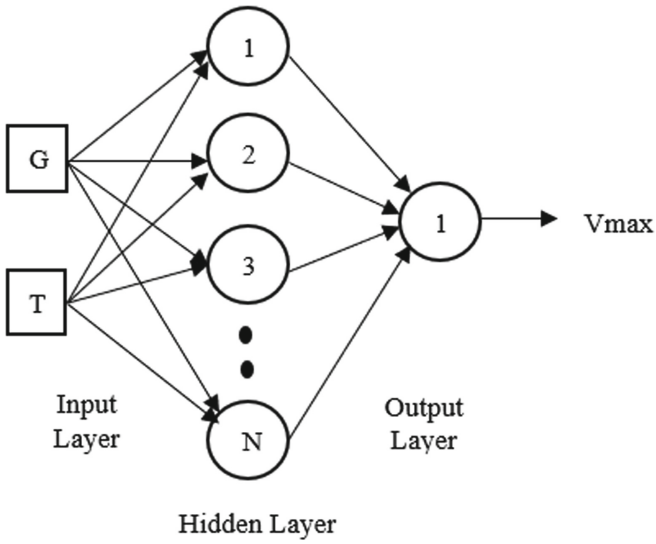


Fig. 4. ANN configuration

7 Training of Neural Network

The input to the ANN MPPT is irradiance and temperature. The data of six years (52584) has been used to train and the train network is converted into a Simulink block model using the fitnet neural network from MATLAB as shown in Fig. 5. The temperature and irradiance as an input is trained with help of neural network and later on, as an output

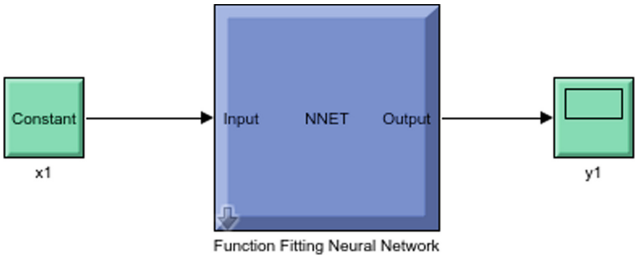


Fig. 5. Simulink block of neural network

of neural network the voltage is given. The simulink block of neural network is shown in Fig. 5.

Among various training algorithms, for the said project we have used the Levenberg–Marquardt training algorithm and the performance is based on mean square error. Fig.6 shows the training of the neural network and the training performance of the network was 4.39×10^{-9} [5].

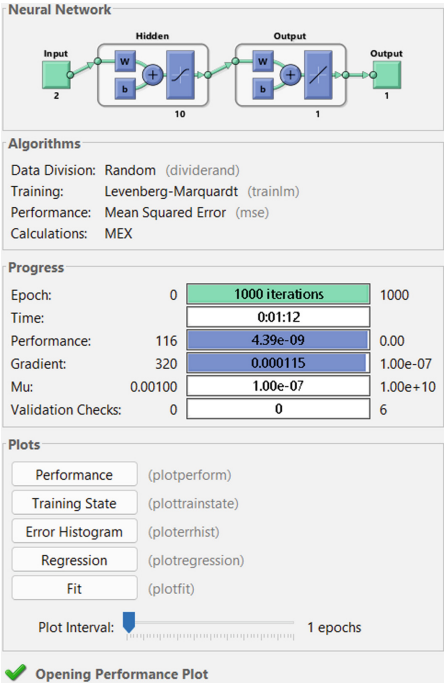


Fig. 6. Training of neural network

Fig. 7 shows the regression plot of training, validation, and testing. The data of 15 percent is used for each testing and validation and 70 percent for training. The value of

R is the relation between output and target data of network training. The value R is 1 here means that the relation between output and target is linear [5].

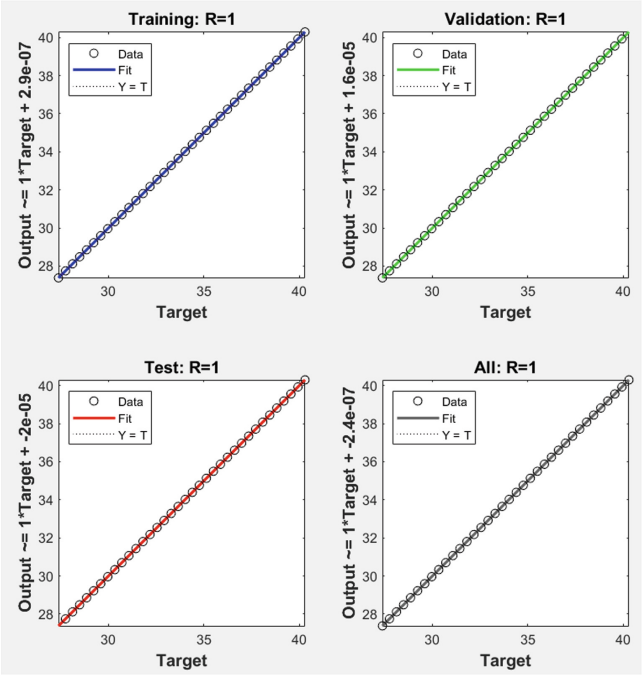


Fig. 7. Regression plot of neural network

The graph between mean squared error versus iteration is shown in the Fig. 8. For this simulation, the best performance is observed at 4.0848×10^{-9} at 1000 epoch.

8 Result and Discusssion

The simulation is done in MATLAB R2020a and the following curves are obtained. The simulation is done for two irradiance levels of 500 W/m^2 and 1000 W/m^2 in the step of 3 s at STC conditions. Figs. 9 and 10 show the PV and converter Power curve, depicting an ANN MPPT technique that can track maximum power of 500 W. The input to the load is 90 V at 1000 W/m^2 and it is observed that with a decrease in irradiance level, the voltage decreases simultaneously (73.4 V) as shown in Fig.11. The variables of the motor like speed, torque, and armature current change with solar irradiance level. In Figs. 12, 13 and 14, it is observed that with a decrease in irradiance level and power, the speed, torque, and armature current changes (decreases).

9 Conclusion

Solar DC pumps are becoming more popular for rural agricultural applications since they are simple to set up. A solar DC Pump is a part of the solar PV panels, regulator,

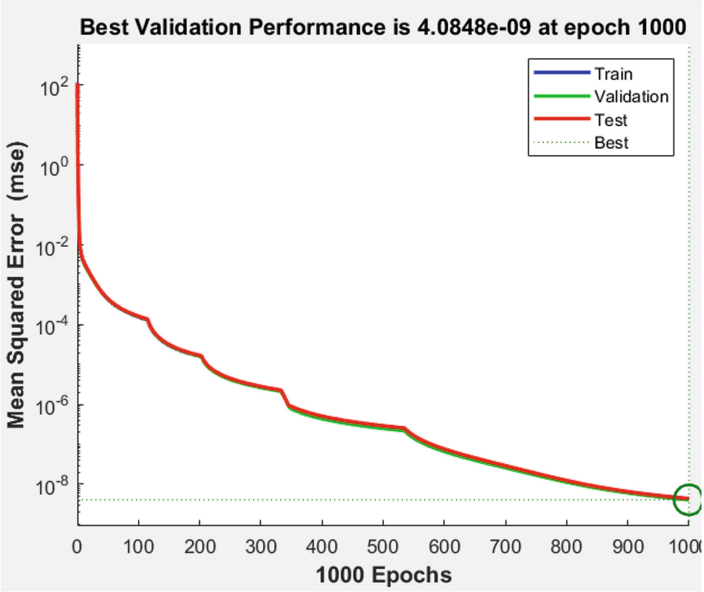


Fig. 8. Epoch graph

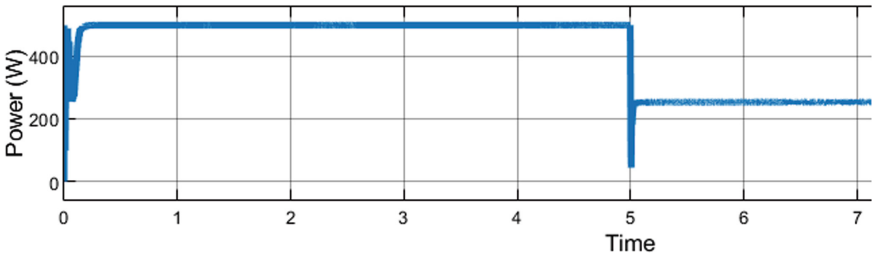


Fig. 9. PV power curve

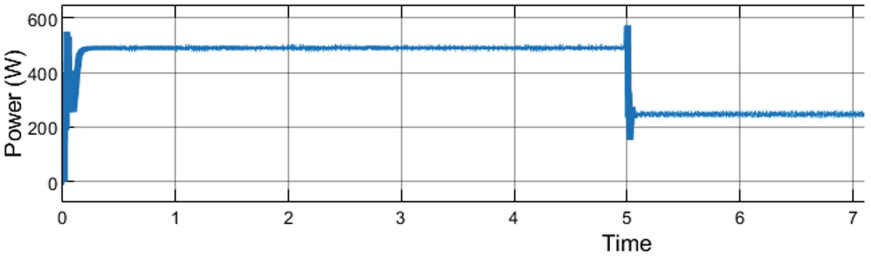


Fig. 10. Converter power curve

and submersible DC pump system. The maximum power point tracking of a PV panel was done by applying the Artificial Neural Network (ANN) approach. It is perceived

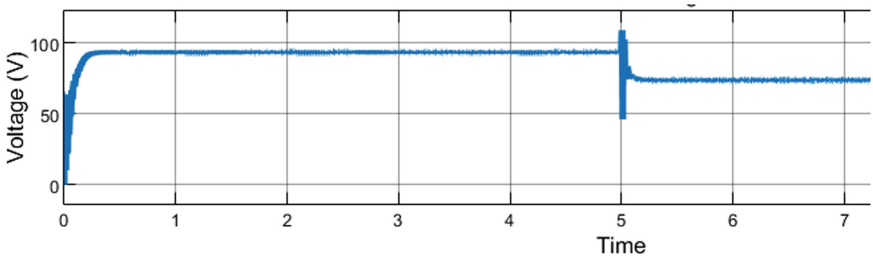


Fig. 11. Converter voltage curve

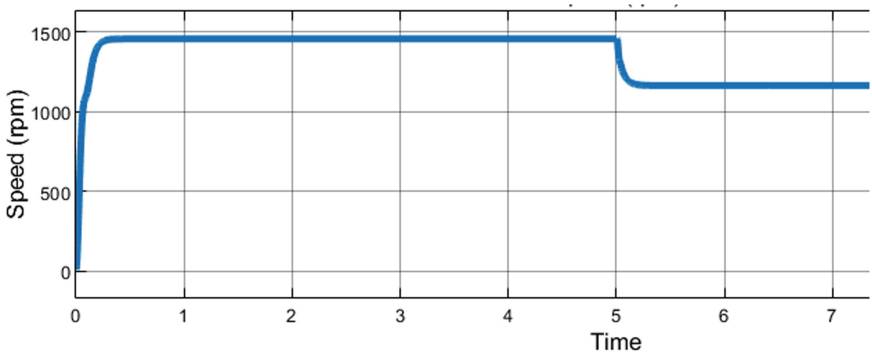


Fig. 12. Speed curve

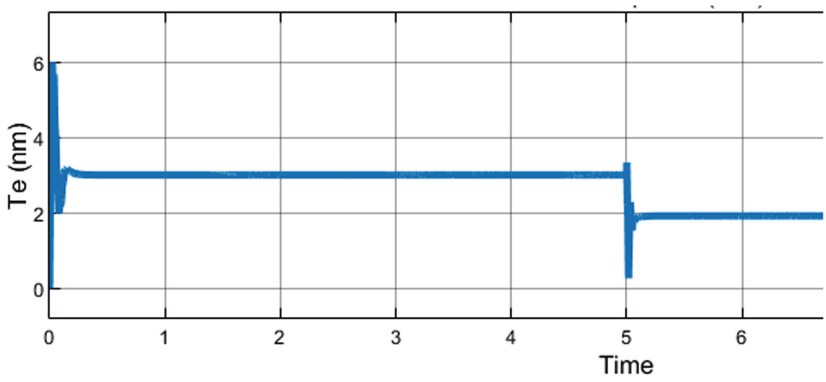


Fig. 13. Torque curve

that the maximum power of 500 W has been tracked, signifying the high efficiency of the artificial neuron network approach.

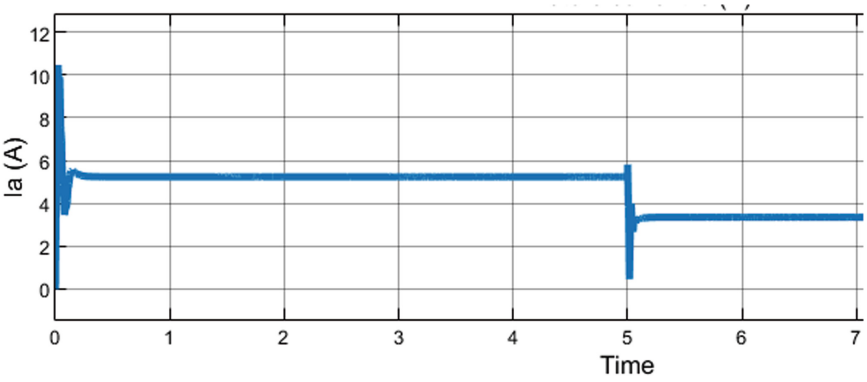


Fig. 14. Armature current curve

References

1. Paul, S., Thomas, J.: Comparison of MPPT using GA optimized ANN employing PI controller for solar PV system with MPPT using incremental conductance. In: 2014 Int. Conf. Power Signals Control Comput. EPSILON 2014, no. January, pp. 8–10 (2014). DOI: https://doi.org/10.1109/EPSCICON.2014.6887518
2. Punitha, K., Devaraj, D., Sakthivel, S.: Artificial neural network-based modified incremental conductance algorithm for maximum power point tracking in the photovoltaic system under partial shading conditions. Energy **62**, 330–340 (2013). https://doi.org/10.1016/j.energy.2013.08.022
3. Peter, G., Stonier, A. A., Gupta, P., Gavilanes, D., Vergara, M. M., Lung sin, J.: Smart fault monitoring and normalizing of a power distribution system using IoT. Energies **15**(21), 8206 (2022). doi: https://doi.org/10.3390/en15218206
4. Ahmed, J., Salam, Z.: A critical evaluation on maximum power point tracking methods for partial shading in PV systems. Renew. Sustain. Energy Rev. **47**, 933–953 (2015). https://doi.org/10.1016/j.rser.2015.03.080
5. Sunny, M. S. H., Ahmed, A. N. R., Hasan, M. K.: Design and simulation of maximum power point tracking of photovoltaic system using ANN. In: 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016 (2017), doi: https://doi.org/10.1109/CEEICT.2016.7873105
6. Peter, G., Praghash, K., Sherine, A., Ganji, V.: A combined PWM and AEM-based AC voltage controller for resistive loads. Math. Probl. Eng. **2022**, 1–11 (2022). https://doi.org/10.1155/2022/9246050
7. Diouri, O., Es-Sbai, N., Errahimi, F., Gaga, A., Alaoui, C.: Modeling and design of single-phase PV inverter with MPPT algorithm applied to the boost converter using back-stepping control in standalone mode. Int. J. Photoenergy **2019** (2019), doi: https://doi.org/10.1155/2019/7021578
8. Aashoor, F.A.O., Robinson, F.V.P.: Maximum power point tracking of PV water pumping system using artificial neural based control. IET Conf. Publ. **2014**(CP651), 1–6 (2014). https://doi.org/10.1049/cp.2014.0923
9. Bouselham, L., Hajji, M., Hajji, B., Bouali, H.: A MPPT-based ANN controller applied to PV pumping system. In: Proc. 2016 Int. Renew. Sustain. Energy Conf. IRSEC 2016, pp. 86–92 (2017), doi: https://doi.org/10.1109/IRSEC.2016.7983918

10. Kumar, R., Singh, B.: BLDC motor-driven solar PV array-fed water pumping system employing zeta converter. IEEE Trans. Ind. Appl. **52**(3), 2315–2322 (2016). https://doi.org/10.1109/TIA.2016.2522943
11. Hiyama, T., Kouzuma, S., Imakubo, T.: Identification of optimal operating point of PV modules using neural network for real time maximum power tracking control. IEEE Trans. Energy Convers. **10**(2), 360–367 (1995). https://doi.org/10.1109/60.391904
12. Peter, G., Sherine, A.: Induced over voltage test on transformers using enhanced Z-source inverter based circuit. J. Electr. Eng. **68**(5), 378–383 (2017). https://doi.org/10.1515/jee-2017-0070
13. Eltawil, M.A., Zhao, Z.: MPPT techniques for photovoltaic applications. Renew. Sustain. Energy Rev. **25**, 793–813 (2013). https://doi.org/10.1016/j.rser.2013.05.022
14. Amara, K., et al.: Improved performance of a PV solar panel with adaptive neuro fuzzy inference system ANFIS based MPPT. In: 7th Int. IEEE Conf. Renew. Energy Res. Appl. ICRERA 2018, **5**, pp. 1098–1101 (2018), doi: https://doi.org/10.1109/ICRERA.2018.8566818
15. Khan, K., Shukla, S., Sing, B.: Design and development of high efficiency induction motor for PV array fed water pumping. In: Proc. 2018 IEEE Int. Conf. Power Electron. Drives Energy Syst. PEDES 2018, pp. 1–6 (2018), doi: https://doi.org/10.1109/PEDES.2018.8707578

Sentiment Analysis from TWITTER Using NLTK

Nagendra Panini Challa[1(✉)], K. Reddy Madhavi[2], B. Naseeba[1], B. Balaji Bhanu[3], and Chandragiri Naresh[2]

[1] School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, India
paninichalla123@gmail.com
[2] Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, AP, India
[3] Department of Electronics, Andhra Loyola College, Vijayawada, India

Abstract. In the present generation, Social networking sites like Twitter, Face book play a important role for communication. Twitter is one of the Blogging platform with numerous data quantity used as an insight in various functional operations of Sentiment Analysis, elections, reviewing, market, etc. The main intension of the tweet based sentiment analysis is the ability to assess good, pessimistic, or nonpartisan opinion part in the tweet information. Tweet sentiment analysis can assist any association with tracking down individuals' assessments of their organization and items. In this paper, we have applied feeling investigation on twitter informational collection. Our model takes input tweet, feeling, and result chosen text beginning and finishing off with input twee and concluded with accuracy of the model.

Keywords: Twitter · Sentiment Analysis · Blogging · Social Networking · Reviews

1 Introduction

Now a days, the usage of the Internet is certainly changed the approach people are using to express individual views and presumption. It has been presently eminently done by writing blogs, online conversations, sites survey, focusing social media, and so on. Number of people are now days are focusing on various Social networking sites like Face book, Twitter, Google Plus, and thereby present their perception, estimation, and offer viewpoint about day-to-day life. By the web-based networking, we acquire intelligent media thereby customers spotlight and encounter others by argumentation. Online entertainment is a way of expressing tweets, blogs, audits, meetings, controversy, and predictions in form of enormous content of high quality data.

Additionally, Web-network based entertainment provides an open door opportunity to management and institutions by contributing a stage to hinge with individual clients for broadcasting. Human Beings mostly depend on client-enabled online platform mode to an unbelievable degree for navigation.

For example, if someone has a desire to purchase an item or needs for any necessity, the first and foremost perspective is to do surveys on the social network web and examine

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 852–861, 2023.
https://doi.org/10.1007/978-3-031-27409-1_78

with web-network based entertainment prior to making a choice. It is extremely enormous to examine the extent client is happy? Hence the need for computerize arises by various embedded approaches to examine distinct feeling are broadly enforced. Opinion investigation (SA) tells clients if the data related to the item is good or bad.

1.1 1.1. SENTIMENT ANALYSIS

The survey of human assumption on the level of understanding for a specific product or topic or service or elections is referred as Sentiment analysis.

Two Crucial techniques are used to examine the sentiment analysis. Which are given as:

a. Natural Language Processing
b. Machine Learning Algorithms.

Through paper survey has been done to take and asses opinion's of customers in past , it has been challenging to monitor and assemble diverse opinions. With the gradual development in usage of social media, the sentiments and opinions of customer are considered as positive or negative by creeping and crawing in a more accessible way which has become on if the easiest task (Fig. 1).

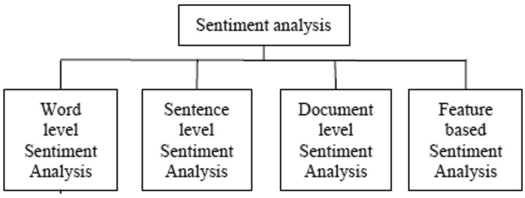


Fig. 1. Levels of sentiment analysis

1.2 Natural Language Processing Approach

Natural language processing (NLP) is the intercommunication between machine i.e. computers and human languages. In order to assess users sentiment online, annotation are used distinctly on twitter, effective sentiment. Most of the studies use the three common sentiment classifications: positive, neutral, and negative (Fig. 2).

In [11], new feature are used to productively annotate users sentiments and opinions; "Mixed Sentiment label", has two different meanings which prevail in tweets. For example, "I love movies, but I hate action movies". "movies" entity is annotated with positive sentiment label, and "action movies" entity is annotated with negative sentiment label, which means the tweet has a mixed sentiment and opinion.

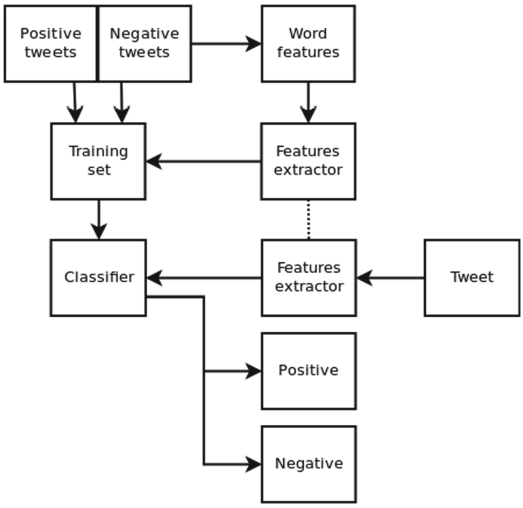


Fig. 2. Sentiment analysis architecture

2 Relevant Study

The feeling data analysis done from twitter has its tremendous appraisal in the social networking sites and in research field also with industrial applications. The central challenging issue faced in this analysis is the fluctuation of speech and multiplex datastructure when extracted. In [1], acquired information on various forms of demonetization from twitter. Its main icon tool was R language for analyzing tweets. It not only focussed tweets but also on distinct projections such as word cloud and other plots. These plots illustrated frequency of people accepting demonetization is higher than the number of people rejecting it. In [2, 3] conducted analysis on twitter information for which tweets in various formats like JSON are extracted using python lexicon dictionary towards assigning polarity to the tweets. Thereby analyzed the domain [4, 5] by a notch and implemented various learning methods in the right direction hence achieved ultimate accurate results.

Research in [7, 8] intended collaboration of two approaches corpus based and lexicon based. This has been an unique collobaration rarely. They focussed features as the on adjectives and verb, and also used corpus based techniques [8] to find the best semantic orientation of different adjectives existing in the tweets. Forecasting various emotions [9] of people watching tv shown as positive or negative. There intented towards extracting comments about various tv shows by taking data set [10] for training and testing the model [11]. The machine learning model naïve bayes classifier is used by analyzing results using a pie chart [12]. That shows polarity of tweets with respect to positive is less than negative. Automatic Detection of Diabetic Retinopathy in Retinal Images has been studied in [13], how to Detect Pneumonia Using various Deep Transfer Learning architectures was presented in [14], Geotagging approach Using Image Steganography was discussed in [15], EEG-Based Brain-Electric Activity Detection During Meditation Using Spectral Estimation Techniques in [16], relation between Correlated Documents in Distributed

Data Bases has been computed using all-conf correlation measure in [17], Expert Search on Web Using Co-Occurrence in [18], and Support Vector Machine Classification of Remote Sensing Images with the Wavelet-based Statistical Features was presented in [19]. Their analysis collected data from crypto currency and implemented with ML algorithms like naïve bayes and SVM which had further more accuracy. Other research analysis by Agarwal, Xie, Vovshaa, I., Rambow, O., and also Passonneau implemented unigram model as a baseline and competed with other models which focuses on features and kernel tree, etc.

3 Proposed Method

The implementation steps are given by:

a. Loading Twitter API
b. Loading word dictionary
c. Searching twitter feeds
d. Extracting text from feeds
e. Framing text cleaning functions
f. Preprocessing the twitter feeds
g. To Analyze and evaluating twitter feeds
h. To Plot high frequency negative and positive words.

3.1 Data Collection

An Application Programming Interface (API) is a product delegate which allows dual applications to coordinate and understand with each other for extracting the relevant information. APIs is utilized as much as for each move you initiate on your telephone, for example to send confidential information or countering the sco re of a football match. These two applications utilize an API to extract and transmit information to your telephone. An API is essentially a channel which takes solicitations, interprets, and checks the reaction. Basically, engineers boost APIs to obtain and grab the definite resources for edge clients. Obviously, to guarantee security of data and information, an API gives selected information that the functional and operational developers have unveiled. To support a solicitation, all APIs need an API Key for the crucial part. The crucial part of an API documentation consists of the important data and information needed to access guidelines and preconditions. Systematically, designers can pursue existing API documentation to formulate URL for extracting the information in a program.

a. Twitter Api

   The Twitter API is a appropriate API which empowers developers to access to Twitter in cutting edge ways. It very well may be utilized to investigate, gain from, and even collaborate with Tweets. It likewise permits cooperation's with direct messages, clients, and other Twitter assets. Twitter's API additionally permits engineers admittance to a wide range of client profile data, like client look, block records, ongoing tweets, from there, the sky is the limit. Data on API items use cases, and docs are accessible on the Developer Platform.

b. Gaining Access to the Twitter API

Prior to utilizing the Twitter API, each person as of now must have a Twitter account. It is then expected to apply for admittance to the Twitter API to acquire qualifications. The API endpoint we will take a gander at is GET/2/tweets/search/later. This profits public Tweets from the most recent seven days that match a pursuit question and is accessible to clients supported to utilize the Twitter API with the standard item track record or some other item track record.

c. Making a Basic Request with the Twitter API

As all API access keys needs to be arranged, nothing is left to be finished apart from trying out the API! The basic step is to stack authorization. One way to configure this problem is engaging the order speedy to pass by convey or token taken as climate variable and a Juptyer Notebook for making proposal and expressing reactions. Start by disclosing an order expeditious and developing a dynamic catalog to any place you wish to save the information.

d. Modifying a Request with the Twitter API

Modifying each inquiry boundaries and the edge point's offers to tease the proposal intended to send. The edge point's API reference record subtitles in the 'Question boundaries' part. An essential arrangement of administrators and could be adapted to update and change the inquiries. One can change the question, begin and peak times for the given period of time we are interested on with the maximum number of results, we can likewise cut down numerous access attributes to give supplement data about the tweet, creator, etc. The accompanying cutdowns 15 Tweets accomplishing the catchphrases like "Extreme Weather", those which are not retweet and were created on October twelfth, 2021.

4 Data Processing

A tweet consists of an enormous amount of suppositions about the data which are communicated in numerous ways by distinct clients. The dataset of utilized is classified into two classes viz. Negative and positive border. Therefore it is easy to notice the impact of different highlights from the feeling investigation of the information. The crude data having extremity is practically powerless to inconsistency and definite invariability

Preprocessing of twitter tweet integrates following features:

- Knock out all URLs (Example given as: wwwc.ab.com), hash labels
- Rewriting words, Spell check
- Rearranging rehashed characters
- Bouncing opinions for each of the emojis.
- Eliminating every accentuations, images, numbers
- Eliminating Stop Words
- Extending Acronyms
- Eliminating tweets related to Non-English

5 Data Extraction

The dataset which is pre-processed under different models which contains many attributes without null values in the dataset. in the component extraction technique we separate many angles from the handled data set. Like we will select some text or paragraph from the dataset we make some highlights from the taken text. Later this perspective is utilized to register the positive and negative margins in a sentence which is helpful for deciding the assessment of the people utilizing the models based upon the selected text. we are having various models to perform on the selected text. The well known models are unigram (one) and bigram(two). And also we can apply very smart AI strategies require addressing the highlights of the text of the selected text.

These key highlights are considered as component vectors which will be utilized for the arrangement task. Some models include that have been accounted for in writing are:

a. **Words and their frequencies**
   Whatever the words and their frequencies came from the data can be highlighted in the form of unigram and bigram and n-gram models. Actually there is more research on word presence rather than word frequencies. To better depict this problem component. Panged al.[13] stated and showed better results by utilizing presence rather than frequencies.
b. **Grammatical Forms Tags:**
   Grammatical forms like description and some words which will enhances the another words meaning (Homophones). To avoid this we can create linguistic pillar designs by parsing or reliance trees.
c. **Assessment Of Words And Phrases**
   Aside from explicit words, a few expressions and phrases which convey opinions can be utilized as elements. For example, cost somebody an arm and leg.
d. **Position Of The Terms**
   The place of a term within in a text will give one meaning but the whole paragraph will give one meaning which makes difference in general opinion of the text.
e. **Invalidation**
   Invalidation is a significant yet troublesome element to decipher. The presence of an invalidation normally changes the extremity of the assessment.

6 Experimental Results

The requirements for implementing the proposed system are Anaconda, Google Colab, and browser Anaconda for Windows (64-bit) is installed. Google Colab is used to implement the project. It offers notebooks with free GPU support. To complete this project, we created a Colab notebook and mounted the code on Google Drive.

After the implementation of machine learning techniques on the Sentiment Analysis dataset we got the accuracies as stated below. We have evaluated Sentiment Analysis on Twitter with various ML techniques and have concluded that Random Forest gives the maximum accuracy of 99.8% (Table 1).

The below figures describe the distribution of data into positive and negative and Word count. We have also experimented and shown Accuracy and Confusion matrix of various Machine Learning Algorithms (Figs. 3, 4, 5, 6, 7, 8 and 9).

Table 1. Accuracy table

S.NO	Algorithm	Accuracy
1	Naïve Bayes	94.5%
2	Support vector machines	96.5%
3	Logistic regression	76.9%
4	Random forest	99.8%

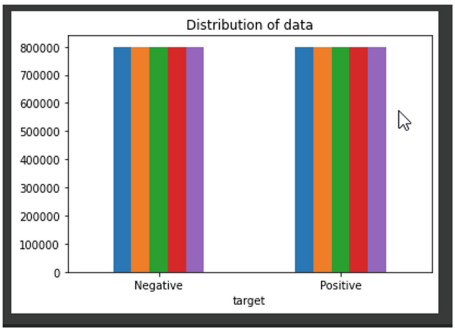


Fig. 3. Distribution of target data into positive and negative

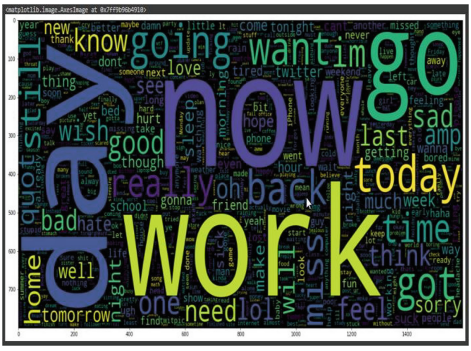


Fig. 4. Word count for positive data

7 Conclusion

In this paper, we express an outline review and relative examination of prevailing approaches used for assessing mining comprising of machine learning and dictionary-based techniques. Research results conclude that AI strategies & techniques, for example, SVM and Naive Bayes has been the most outstanding precision and can be noticed as one of the learning strategies and the significant precision is from Random Forest. At first we had obtained the data of positive and negative tweets with different model accuracies, and we had used different sentiment analysis techniques and NLP LIBRARIES

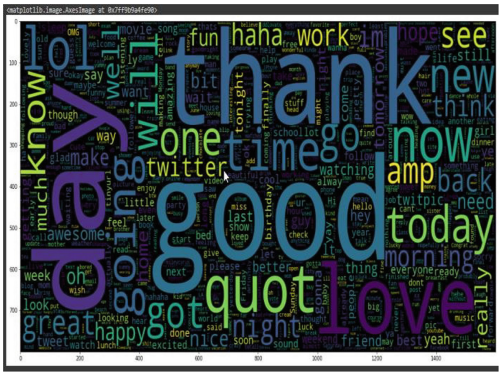


Fig. 5. Word cloud for negative data

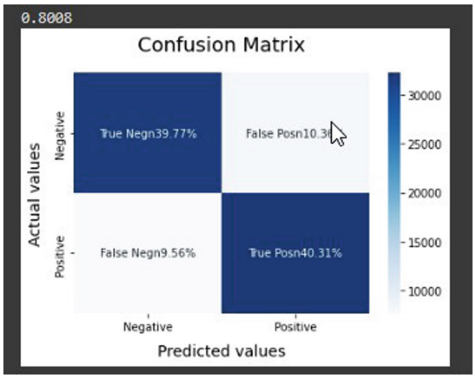


Fig. 6. Accuracy and confusion matrix for Naïve Bayes

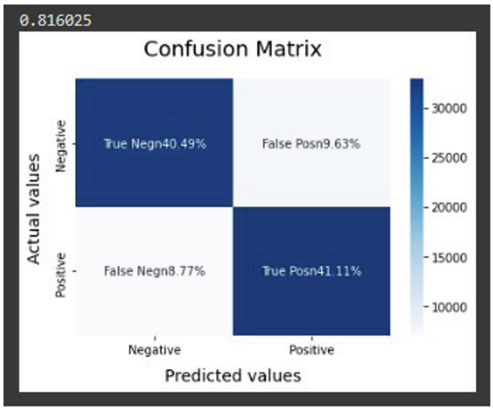


Fig. 7. Accuracy and configuration matrix for support vector machines

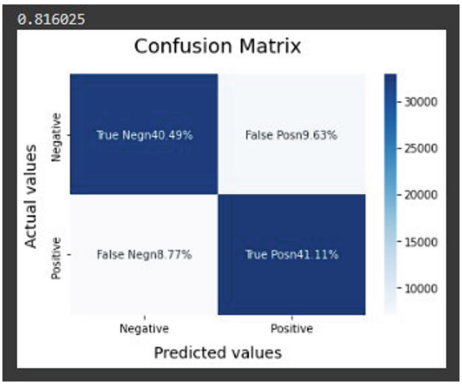


Fig. 8. Accuracy and confusion matrix for logistic regression

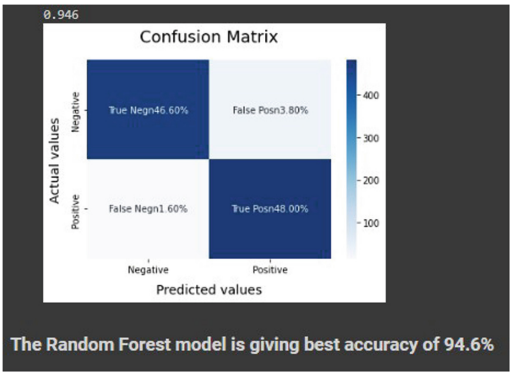


Fig. 9. Accuracy and confusion matrix for RANDOM FOREST

to preprocess the original data obtained from the API and divided that data into both testing and training data and with 2 data's, we have obtained different accuracy with different models among them we got best result from Random Forest Model with highest accuracy.

References

1. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp.1320–1326 (2010)
2. Parikh, R., Movassate, M.: Sentiment analysis of user- generated twitter updates using various classification techniques. CS224N Final Report (2009)
3. Go, Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Stanford University, Technical Paper (2009)
4. Barbosam, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. Poster Volume, pp. 36–44, COLING (2010)

5. Bifet, Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer, pp. 1–15 (2010)
6. Agarwal, Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proceedings of the ACL 2011 workshop on languages in social media, pp. 30–38 (2011)
7. Davidov, D., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. Coling 2010: Poster Volume pages 241–249, Beijing (2010)
8. Liang, P.-W., Dai, B.-R.: Opinion mining on social media data. In: IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3–6, pp 91–96. ISBN: 978-1-494673-6068-5 (2013)
9. Gamallo, P., Garcia, M.: Citius: A Naive-Bayes strategy for sentiment analysis on english tweets. In: 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23–24, pp 171–175 (2014)
10. Neethu M.S., Rajashree R.: Sentiment analysis in twitter using machine learning techniques. In: 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661
11. Turney, P. D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics (2002)
12. Statista: Most popular social networks worldwide as of July 2021, ranked by a number of active users (Accessed 04-10-21)
13. Prabhakar, T., Sunitha, G., Madhavi. G., Avanija, J., Madhavi, K.R.: Automatic detection of diabetic retinopathy in retinal images: a study of recent advances. In: Ann. Romanian Soc. Cell Biol. 25(4), 15277–15289 (2021)
14. Reddy Madhavi, K., Madhavi, G., Rupa Devi, B., Kora, P.: Detection of pneumonia using deep transfer learning architectures. Int. J. Adv. Trends Comput. Sci. Eng. **9**(5), 8934–8937 (2020), ISSN 2278–3091, https://doi.org/10.30534/ijatcse/2020/292952020
15. Chandhan, M., Reddy Madhavi, K., Ganesh Naidu, U., Kora, P.: (2021) A novel geotagging method using image steganography and GPS. Turk. J. Physiother. Rehabil. **32**(3), ISSN 2651-4451, 807–812
16. Kora, P., Rajani, A., Chinnaiah, M.C., Madhavi, R., Swaraja, K., Kollati, M. EEG-based brain-electric activity detection during meditation using spectral estimation techniques. https://doi.org/10.1007/978-981-16-1941-0_68., pp.687–693 (2021).
17. Reddy Madhavi, K., Rajani Kanth, T.V.: Finding Closed Correlated Documents in DDB using all-conf", International Journal of Engineering Science and Technology, **3**(5), 4036–4042 (2011)
18. Naime Saranya, M., Avanija, J.: Expert search on web using co-occurrence. Int. J. Appl. Eng. Res. **10**(49), 140–145 (2015)
19. Prabhakar, T., Srujan Raju, K., Reddy Madhavi, K.: (2022). Support vector machine classification of remote sensing images with the wavelet-based statistical features. fifth international conference on smart computing and informatics (SCI 2021), Smart Intelligent Computing and Applications, Volume 2. Smart Innovation, Systems and Technologies, vol 283. Springer, Singapore



Cardiac Anomaly Detection Using Machine Learning

B. Naseeba[1](✉), A. Prem Sai Haranath[1], Sasi Preetham Pamarthi[1], S. Farook[2], B. Balaji Bhanu[3], and B. Narendra Kumar Rao[4]

[1] School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, India
beebi.naseeba@vitap.ac.in
[2] Department of EEE, Sree Vidyanikethan Engineering College, Tirupati, AP, India
[3] Department of Electronics, Andhra Loyola College, Vijayawada, AP, India
[4] Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, AP, India

Abstract. The heart is an essential organ in the human body. It can be attributed to birth problems, heredity, or may even be due to our health routine. Therefore, it has become exceedingly tough for healthcare practitioners to detect and anticipate cardiovascular problems at an early stage considering several criteria such as an aberrant pulse rate or excessive BP. Which has resulted in a desperate need for an effective and reliable technique to detect them, considerably expanding with each passing day. Five Machine learning algorithms Support Vector Machine, K-Nearest Neighbour, Logistic regression, Random Forest, and Artificial Neural Network are used to operate on massive datasets gathered from the healthcare industry, anticipating, and aiding in the decision-making process. The predicted outputs are based on features like Age, Blood Pressure rate, Cholesterol, Glucose, Smoke, Alcohol, etc. The features necessary for this model are basic enough for any ordinary non-medical individual to determine whether they are at risk for any cardiac disorders with a simple blood scan rather than a medical evaluation. We achieved an accuracy of 99.8% by using the K-NN algorithm which is greater than any of the previous articles.

Keywords: Support Vector Machine (SVM) · K-Nearest Neighbour (KNN) · Supervised learning · Accuracy · Precision · Unsupervised learning · Recall · F1 Score · Semi-supervised learning

1 Introduction

The heart is a vital organ in the human body. It is divided into four compartments. These compartments are called chambers and they are the Right and Left atriums, receives the blood. Right and Left ventricles, pumps blood out. The blood is delivered and received using blood vessels. The heart has its way of delivering blood, a network made up of arteries and veins. The former supplies freshly oxygenated blood while the latter brings back all the deoxygenated blood from the organs. Since the heart is made up of such a complex system, there are many ways it could go wrong or be damaged. If any of the major blood vessels are clogged or the chambers are blocked or even if the muscles are not

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 862–871, 2023.
https://doi.org/10.1007/978-3-031-27409-1_79

strong enough, it could potentially lead to death. Cardiovascular diseases are becoming more common in our modern era. Due to present situations and food habits irrespective of their age, people are facing heart-related problems. If the heart fails, the functioning of the body stops. Therefore, it is important to predict cardiac diseases accurately with the prior small symptoms with the technology as the latest surgery machines arrived in the medical field. To save more people with an accurate prediction of heart-related diseases we came up with the latest algorithms of machine learning to predict the disease based on Precision, Accuracy, Recall and F1 Score. Random forest, K-NN, Logistic Regression, ANN, SVM, and Random Forest. These algorithms are used in the cardiac model and finding the best algorithm to predict heart disease with the grater F1 Score. We can predict them using variables like Age, Height, Blood pressure High, Blood Pressure Low, Cholesterol, Smoke, Glucose, and Alcohol.

The following are the categories of machine learning techniques.

A. Supervised learning

Supervised learning uses both factors and labels to predict or classify data. For our comparative analysis, we used Decision Tree, Linear SVM, Naïve Bayes, and K-NN algorithms.

B. Unsupervised learning

Unsupervised learning uses only factors to train and predict (or classify) data. In this paper, we used two unsupervised learning algorithms namely, Artificial neural networks (ANN), Random Forests.

C. Semi-supervised learning

Both the labeled and unlabelled data are used while training the dataset using a Semi-supervised learning model. In this paper, we used Logistic Regression and Linear Regression for our comparative analysis (Fig. 1).

2 Relevant Study

Many research publications have already employed machine learning algorithms to identify whether or not a person is predisposed to cardiac problems. Bernoulli Naive Bayes, Gaussian Naive Bayes, and Random Forest are three popular algorithms for the aforementioned investigations. We can tell if a person has cardiac illness. With an accuracy of 85% using the Cleveland dataset Bermando [1] and Kumar demonstrate how machine learning is utilised to aid in the diagnosis of a variety of ailments. He employed random forest and convolution neural networks to predict cardiac illness, with an accuracy of 80%, 78% [2]. Singh H, Navaneeth N, and Pillai G's research focuses on the percentage likelihood of developing heart disease. To handle multivariate datasets, they used Nave Bayes, KNN, Random Forest, SVM, Decision trees, and ensemble approaches [3]. Prashant Narayankar, Shantala Giraddi, Neha R. Pudakalakatti, Shreya Sulegaon, and Shrinivas D. Desai worked together to evaluate the accuracy of the classification model for cardiovascular disease prediction in the Cleveland dataset. Back-Propagation Neural Network, LR, was employed in this study [4]. Balabaeva et al. a. conducted a study in the

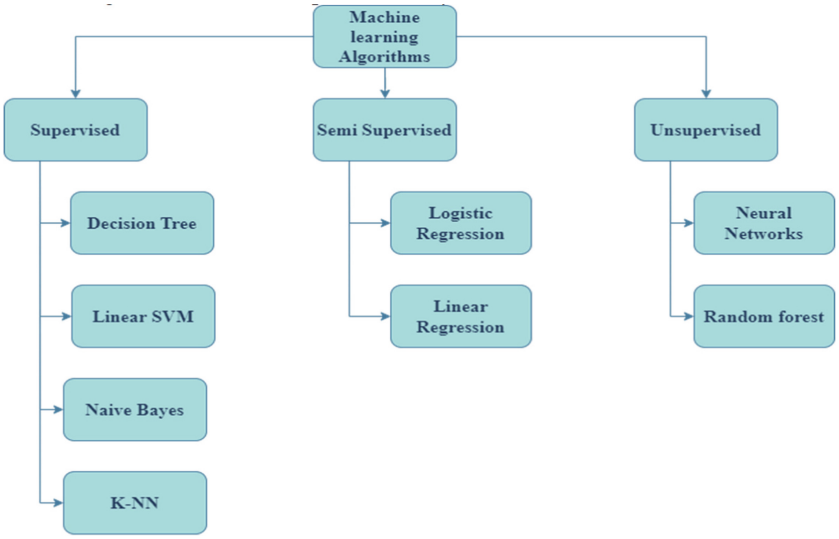


Fig. 1. Types of machine learning algorithms

year 2020. Their research employs ML algorithms including DT, LR, XGB and RF, as well as scaling approaches like Max Abs Scaler, MinMax Scaler, StandardScaler, Quantile Transformer, and Robust Scaler. Their findings indicate that general models perform better when temporal and non-temporal characteristics are combined [13]. Toward the end of 2022 A Trisal, V. Sagar, and R. Jameel collaborated to develop a viable approach for diagnosing cardiac illness using machine learning models such as Decision Tree and Support Vector Machine, and the results were evaluated using accuracy and confusion matrices [14]. Detection of COVID-19 using Deep Learning methods has been discussed in [15–17], Detection of Pneumonia Using Deep Transfer Learning architectures in [18], methods for Quality Improvement of Retinal Optical Coherence Tomography shown in [19] (Table 1).

3 Proposed methods

A. Collection of data:

The collection of data is the most crucial step in preprocessing of data. We collected data from Kaggle which is used to analyze and predict the results. A description of the Dataset is given below:

The cardiac dataset which contains seventy thousand rows and thirteen columns (Table 2).

B. Selection of Attributes:

Attribute selection must be done carefully for the prediction of heart disease—the dependent factors such as *Age, Cholesterol, Gender, Blood Pleasure*, etc. And the independent factors such as id must be deleted (Fig. 2).

C. Data Pre-Processing:

Table 1. Comparison of accuracy with the previous paper

Study	Approach	Dataset	Accuracy
1	Gaussian NB, Bernoulli NB, RF	Cleveland	85%, 85%, 75%
2	RF, CNN	Cleveland	80%, 78%
3	SVM	Cleveland	73–91%
4	Back-Propagation Neural Network, LR	Cleveland	85.074%, 92.58%
5	SVM, Cuckoo Search optimized Neural Network	Cleveland	94.4%
6	SVM	Privately owned	Specificity-78.8%, Sensitivity-62.3%, Positive predictive Value-10%, Negative predictiveValue-98.2%
7	CNN	MIT-BIH	89.07–94%
8	SVM	MIT-BIH	97.08–97.77%
9	Decision tree, SVM, KNN	UCI repository	98%, 68%, 90%

Table 2. Details of the Cardiac dataset

S. no.	Attribute name	Type
1	ID	N
2	Age	N
3	Gender	N
4	Height	N
5	BP_High	N
6	BP_Low	N
7	Cholesterol	N
8	Glucose	N
9	Smoke	N
10	Alcohol	N
11	Active	N
12	Cardio	N

Pre-processing is the preparation (cleaning and arranging) of raw data to make it suitable for training and testing Machine Learning models to acquire better results. Without proper and complete data, results are not accurate. Therefore, the data must be processed before implementing machine learning algorithms. The missing

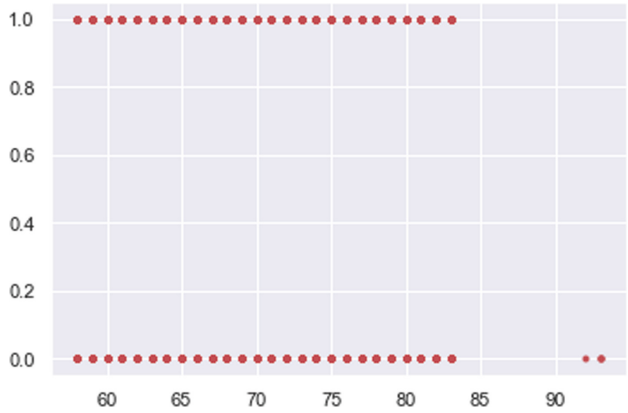


Fig. 2. Plotting of dataset

values are filled with pre-processing techniques and the null values in the data will be replaced with normalized values.

D. Building a model:

This is the most crucial step in the implementation of the model. In this step, machine learning techniques are used including Support Vector Machine, ANN, Decision Tree, Logistic Regression, K-Nearest Neighbour, Nave Bayes, Random Forest.

Cardiac Prediction using machine learning algorithms:

Step 1: Import the libraries

Step 2: Link Google Colab to the drive

Step 3: Now import the cardiac dataset from Kaggle

Step 4: Pre-processing the dataset and make the dataset unique and with no null values

Step 5: Train and Test the model

Step 6: Implementing machine learning algorithms

Step 7: Prediction of results and the accuracy, confusion matrix

E. Evaluation:

Evaluation is the last stage of the prediction of results in the working model. In this, we can predict the results of diabetes using the accuracy of ML algorithms.

4 Experimental Results

Requirements for implementing the proposed system are Anaconda, Google Collab, and the browser Anaconda for Windows (64-bit) is installed. Google Collab was used to implement the project. It offers notebooks with free GPU support. To complete this project, create a Collab notebook and use the code below to mount the Google Drive.

The algorithms have been compared using Precision, Accuracy, Confusion matrix, Recall and F1 Score.

(1) Accuracy:
It is the percentage of instances that are classified precisely.
Accuracy $= (T_n + T_p)/(T_n + F_n + F_p + T_p)$

(2) Precision:
It is the percentage of genuinely positive values.
Precision $= T_p/(F_p + T_p)$

(3) Recall:
It is the ratio of the total positive anticipated to the positive values.
Recall $= T_p/(F_n + T_p)$

(4) Confusion matrix:
It is the method to measure Accuracy, Precision, and Recall. It decides the performance of the model.
The confusion matrix is:

	Actual		
		Negative	**Positive**
Predicted	**Negative**	True Positive (T_p)	False Negative (F_n)
	Positive	False Positive (F_p)	True Negative (Tn)

(5) F1 Score:
It is the HM of both Recall and Precision.
F1 score = 2*Recall*Precision/(Recall + Precision)

- T_p True Positive: It is a value where the model predicts the +ve class precisely.
- T_n True Negative: Itis a value where the model predicts the −ve class precisely.
- F_p False Positive: It is a value where the model predicts the +ve class defectively.
- F_n False Negative: It is a value where the model predicts the −ve class defectively.

Table 3–7 shows the confusion matrix of K-Nearest Neighbour, Logistic Regression, Support Vector Machines (SVM's), Random Forest, and Artificial Neural Networks respectively.

Table 3. Matrix of Confusion for K-Nearest Neighbour

	Negative	**Positive**
Negative	10519	20
Positive	3	10458

Table 4. Matrix of confusion for Logistic Regression

	Negative	**Positive**
Negative	5116	1888
Positive	2279	4717

Table 5. Matrix of confusion for Support Vector Machines (SVMs)

	Negative	**Positive**
Negative	8756	2750
Positive	3509	8085

Table 6. Matrix of confusion for Random Forest

	Negative	**Positive**
Negative	8558	2948
Positive	3553	8041

Table 7. Matrix of confusion for Artificial Neural Networks

	Negative	**Positive**
Negative	6628	2094
Positive	2753	6025

After the implementation of machine learning techniques on the Cardiac dataset, we got the accuracies as stated below in Table 8. With all the ML techniques K-Nearest Neighbour gives the maximum accuracy of 99.8%. The performance of algorithms has been represented graphically in Fig. 3, 4.

Table 8. Accuracy table

S.NO	Algorithm	Accuracy
1	K-NN	0.998
2	Logistic Regression	0.702
3	SVM	0.729
4	Random Forest	0.718
5	ANN	0.723

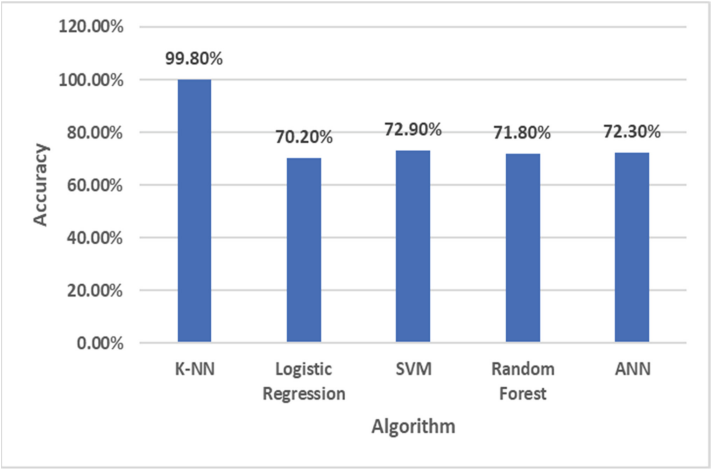


Fig. 3. Accuracy Analysis of all algorithms

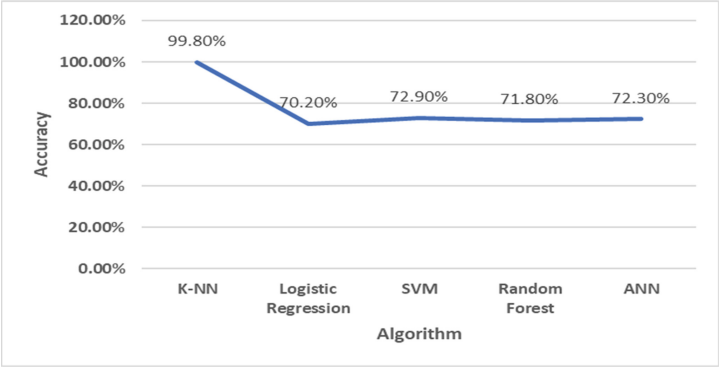


Fig. 4. Comparative analysis of accuracies

The results obtained by all the machine learning algorithms based on precision, Recall and their F1 score is stated in Table 9. K-Nearest Neighbour has a greater F1 score. Therefore, we use K-NN to predict whether a person has a cardiac disease or not.

5 Conclusion

There have been many models of cardiovascular disease detection through the ages, and the results have improved significantly. Although the primary concern has been the accuracy of the methods, better methods are discovered frequently to give better outcomes. To increase the accuracy in the detection of cardiac diseases, new methods should be implemented over the old ones. This research proposes a machine learning-based Ensemble model by using Machine learning algorithms. With this study and analysis, we have

Table 9. Comparison based on Precision, F1 score and Recall

Algorithm	Precision	Recall	F1 score
K-Nearest Neighbour	0.9997	0.9980	0.988
SVM	0.6973	0.7461	0.7208
Logistic_Regression	0.6742	0.7141	0.6935
Random Forest	0.6935	0.7317	0.6895
ANN	0.6863	0.7420	0.7130

used many ML algorithms on the targeted dataset and the implementation had been performed using many algorithms of which K-Nearest Neighbour produces the maximum accuracy of 99.8%, which is higher accuracy than the existing accuracies in the field of cardiovascular disease detection. This can be further extended by creating an application that assists doctors to predict the patient is suffering from cardiovascular disease.

References

1. Bemando, C., Miranda, E., Aryuni, M.: Machine-learning-based prediction models of coronary heart disease using naïve bayes and random forest algorithms. In: Proceedings of the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, pp. 232–237 (2021)
2. Kumar, R.R., Polepaka, S.: Performance comparison of random forest classifier and convolution neural network in predicting heart diseases. In: ICCII 2018, Proceedings of the Third International Conference on Computational Intelligence and Informatics, pp. 683–691. Springer, Singapore. (2020)
3. Singh, H., Navaneeth, N., Pillai, G.: Multisurface proximal SVM based decision trees for heart disease classification. In: Proceedings of the TENCON 2019–2019 IEEE Region 10 Conference (TENCON), Kerala, India, pp. 13–18 (2019)
4. Desai, S.D., Giraddi, S., Narayankar, P., Pudakalakatti, N.R., Sulegaon, S.: Back-propagation neural network versus logistic regression in heart disease classification. In: Advanced Computing and Communication Technologies, pp. 133–144. Springer: Berlin/Heidelberg, Germany, (2019)
5. Patil, D.D., Singh, R., Thakare, V.M., Gulve, A.K.: Analysis of ECG arrhythmia for heart disease detection using SVM and cuckoo search optimized neural network. Int. J. Eng. Technol. **7**, 27–33 (2018)
6. Liu, N., et al.: An intelligent scoring system and its application tocardiac disease prediction. IEEE Trans. Inf. Technol. Biomed. **16**, 1324–1331 (2012)
7. Acharya, U.R., et al.: A deep convolutional neural network model to classify heartbeats. Comput. Biol. Med. **89**, 389–396 (2017)
8. Yang, W., Si, Y., Wang, D., Guo, B.: Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine. Comput. Biol. Med. **101**, 22–32 (2018)
9. Ansari, A.Q., Gupta, N.K.: Automated diagnosis of coronary heart disease using neuro-fuzzy integrated system. In: Proceedings of the 2011 World Congress on Information and Communication Technologies, Mumbai, India, pp. 1379–1384 (2011)

10. Ahsan, M.M., Mahmud, M., Saha, P.K., Gupta, K.D., Siddique, Z.: Effect of data scaling methods on machine learning algorithms and model performance. Technologies **9**, 52 (2021)
11. Rubin, J., Abreu, R., Ganguli, A., Nelaturi, S., Matei, I., Sricharan, K.: Recognizing abnormal heart sounds using deep learning.arXiv 2017, arXiv:1707.04642
12. Miao, J.H., Miao, K.H.: Cardiotocographic diagnosis of fetal health based on multiclass morphologic pattern predictions using deep learning classification. Int. J. Adv. Comput. Sci. Appl. **9**, 1–11 (2018)
13. Balabaeva, k.; kovalchuk, s. comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. Procedia Comput. Sci. **156**, 87–96 (2019)
14. Trisal, A., Sagar, V., Jameel, R.: Cardiac disease prediction using machine learning algorithms. International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) **2022**, 583–589 (2022). https://doi.org/10.1109/CISES54857.2022.9844370
15. Reddy Madhavi, K., et al.: "COVID-19 detection using deep learning", In: 20th International Conference on Hybrid Intelligent Systems-HIS 2020, at Machine Intelligence Research (MIR) labs, USA, Springer AISC series (2020)
16. Abbagalla, S., Rupa Devi, B., Anjaiah, P., Reddy Madhavi, K.: "Analysis of COVID-19-impacted zone using machine learning algorithms", Springer series – Lecture Notes on Data Engineering and Communication Technology 63, pp. 621–627 (2021)
17. Reddy Madhavi, K., Madhavi, G., Rupa Devi, B., Kora, P.: "Detection of pneumonia using deep transfer learning architectures", Int. J. Advanced Trends Computer Sci. Eng. **9**(5), pp. 8934–8937 (2020). ISSN 2278-3091
18. Kora, P., Rajani, A., Chinnaiah, M.C., Madhavi, K.R., Swaraja, K., Meenakshi, K.: EEG-based brain-electric activity detection during meditation using spectral estimation techniques. In: Jyothi, S., Mamatha, D.M., Zhang, Y.-D., Raju, K.S. (eds.) Proceedings of the 2nd International Conference on Computational and Bio Engineering. LNNS, vol. 215, pp. 687–693. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1941-0_68
19. Rajani, A., Kora, P., Madhavi, R. Jangaraj, A.: Quality improvement of retinal optical coherence tomography. 1–5. (2021). doi: https://doi.org/10.1109/INCET51464.2021.9456151



Toxic Comment Classification

B. Naseeba[1,1(✉)], Pothuri Hemanth Raga Sai[1], B. Venkata Phani Karthik[1], Chengamma Chitteti[2], Katari Sai[2], and J. Avanija[2]

[1] School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, India
beebi.naseeba@vitap.ac.in

[2] Sree Vidyanikethan Engineering College, Tirupati, India

Abstract. Presently in the world social media is ruling among the people. It is used by the people to share the views and contact others. It became the powerful weapon to the people. Users nowadays leave countless comments on various social media platforms, news portals, and forums. Some of the remarks are harmful or aggressive in nature. Because manually moderating comments is impractical due to the large volume of them, most systems rely on machine learning models to detect toxicity automatically. So, to protect people from the illegal comments and fake conversations and toxic words in the comments and posts. We used the Deep Learning models like Convolution Neural Networks (CNNs), LSTMs, to detect the toxic words in the comments based on the toxic percentage. We conclude our work with a detailed list of existing research gaps and recommendations for future research topics connected to the subject of online harmful comment classification.

Keywords: Convolutional Neural Network (CNN) · Multi-layer perceptron · Long Short Term Networks (LSTMS) · Multi class classification · Binary cross entropy

1 Introduction

The main problem of the people is the toxic conversations and that is the reason why the people are worried to express their views and spread their innovative ideas. Due to fear that they get problems, and they are disturbed and are troubled by others.

The motive of this research is to utilize deep learning to detect toxicity in text, which might be used to help users avoid sending potentially harmful messages, build more respectful arguments while conversing with others, and assess the toxicity of other users' remarks.

This research will use a variety of deep learning models, including multilayer perceptron's (MLPs), and Convolutional Neural Networks (CNNs), Long Short Term Memory Networks (LSTMs)to solve the aforementioned objective, evaluating their performance on binary and multi-label classification tasks. Our initiative also investigates the uses of these models in both the public and private sectors (Fig. 1).

Individuals have been able to put forward their thoughts and freely express their opinions on a variety of subjects and situations thanks to online forums and social media platforms. These online comments may contain explicit language that is harmful to the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 872–880, 2023.
https://doi.org/10.1007/978-3-031-27409-1_80

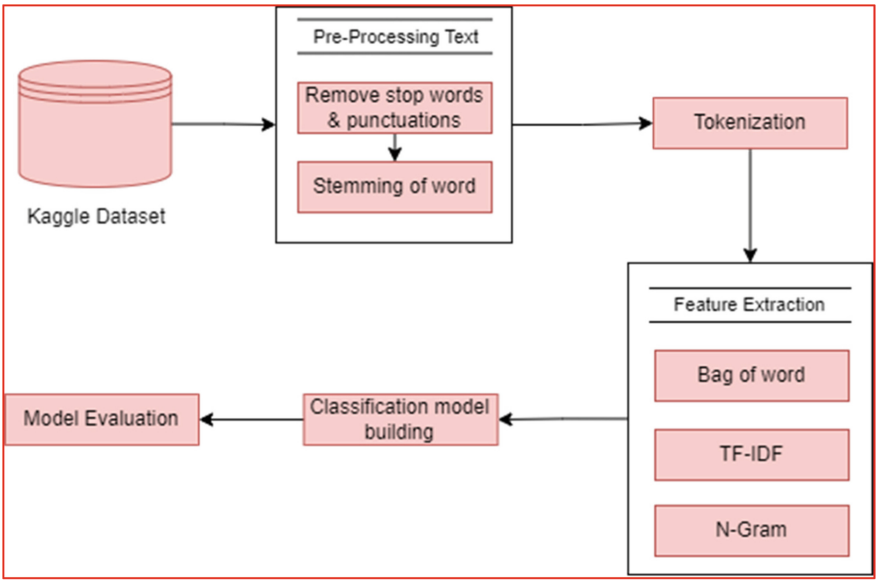


Fig. 1. Process flow of the model

readers in some situations. Severe Toxic, Toxic, Threat, obscene, Identity Hate and Insult are just few of the categories for comments that contain explicit language. Many people are afraid of being abused or harassed, so they stop expressing themselves and stop searching out diverse viewpoints.

Several Machine Learning models have been created and used to filter out the obnoxious language and protect internet users from being harassed or bullied online. In this we have predicted the toxicity level of the comment provided by the user and classified it by using the model.

2 Relevant Study

The researchers noted that model improves on the key aspects of representation, computation, statistics, learning, and these methods lowers the problem of wrong [1] data presentation. And also, they tried to minimize the risk of biased data which is employed by the training of single model. While most of the deep learning algorithms [2] search for solutions which shows the optimal solution. Wikipedia dataset is used. This model shows [3], 75% of cases by 1.5–5.4%. The Empirical Evaluation of Temporal Convolutional Network for Offensive Text Classification.

In [4], they worked on the Deep Learning based Multilabel Binary Classification of comments. This model comprises of combining of single model-based outputs for improving the accurate predictions and reliable generalization.

In [5], worked on BERT and fast Text Embeddings for Automatic Detection of Toxic Speech. The researchers proposed an automatic model for the classification of toxic speech using the deep-learning techniques and embedded words. They performed the

binary and multi-class classification using a Twitter corpus (a) Using a DNN classifier to extract the word embeddings; (b) pre-trained BERT model. We finally concluded that BERT model's performance is much better when compared with the others. The above-mentioned methodology can be used for other type of social media generated comments. The researchers stated that on a Twitter corpus BERT model performed better than feature-based approaches.

Crespi, Farahbakhsh and Mozafari (2019) worked on the LSTM and convolution neural networks in Online User Comments. The researchers described a supervised model, based on deep neural networks, for the classification of online users' comments based on their claims. And also they experimented it by using long short-term memory networks (LSTMs) and Convolution Neural Networks (CNN) to train the datasets of online user comments. They used the various types of distributional word embedding. There searchers achieved a most important improvement on one dataset in which it classifies the [6] comments as emotional or factual. The researchers measured the working of TCN network to search and classify foul language based on its criminal content with the CNN, GRU and LSTM. When compared with LSTM and GRU, TCN have the parallelism and it can get the history of long back with the residual blocks and convolutions. By these it shows that TCN's performance [7] is better when compared with the GRU, LSTM and the variants of RNN based on the f1 score of the toxic comments.

Few worked on an LSTM based approach Classification of Short Text Sentiment with Word Embeddings. For the detection of sentiment polarity from the short texts the researchers used the deep learning methods and LSTM [8] and word embeddings in the social media classification. Firstly, the words in the posts are captured into vectors with the word embeddings. Then, the word sequence in sentences is input to LSTM to learn the long-distance contextual dependency among words. The results showed that deep learning methods can perform well [9] when the social media gives the enough and sufficient training data. The performance depends on the amount of training data. The classification performance is lower for the casual comments and is higher to the movie reviews. But still LSTM is performing well when compared with ELM and NB. This shows the scope of an LSTM-based approach for short-text sentiment classification.

Modeling of Chaotic Political Optimizer for Crop Yield Prediction in [10], comparative analysis of deep neural network architectures for the dynamic diagnosis in [11], Prediction of Climate Change using SVM and Naïve Bayes Machine Learning Algorithms in [12], Analysis of COVID-19-Impacted Zone Using Machine Learning Algorithms [13], Dengue Outbreak Prediction using Regression Model [14], Detection of COVID-19 using Deep Learning methods has been discussed in [15], EEG-Based Brain-Electric Activity Detection During Meditation Using Spectral Estimation Techniques were found in [16], Support Vector Machine Classification of Remote Sensing Images with the Wavelet-based Statistical Features was presented in [17], methods for Quality Improvement of Retinal Optical Coherence Tomography shown in [18], Detection of Pneumonia Using Deep Transfer Learning architectures in [19] (Fig. 2).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1804874 entries, 0 to 1804873
Data columns (total 45 columns):
 #   Column                                 Dtype
---  ------                                 -----
 0   id                                     int64
 1   target                                 float64
 2   comment_text                           object
 3   severe_toxicity                        float64
 4   obscene                                float64
 5   identity_attack                        float64
 6   insult                                 float64
 7   threat                                 float64
 8   asian                                  float64
 9   atheist                                float64
 10  bisexual                               float64
 11  black                                  float64
 12  buddhist                               float64
 13  christian                              float64
 14  female                                 float64
 15  heterosexual                           float64
 16  hindu                                  float64
 17  homosexual_gay_or_lesbian              float64
 18  intellectual_or_learning_disability    float64
 19  jewish                                 float64
 20  latino                                 float64
 21  male                                   float64
 22  muslim                                 float64
 23  other_disability                       float64
 24  other_gender                           float64
 25  other_race_or_ethnicity                float64
 26  other_religion                         float64
 27  other_sexual_orientation               float64
 28  physical_disability                    float64
 29  psychiatric_or_mental_illness          float64
 30  transgender                            float64
 31  white                                  float64
 32  created_date                           object
 33  publication_id                         int64
 34  parent_id                              float64
 35  article_id                             int64
 36  rating                                 object
 37  funny                                  int64
 38  wow                                    int64
 39  sad                                    int64
 40  likes                                  int64
 41  disagree                               int64
 42  sexual_explicit                        float64
 43  identity_annotator_count               int64
 44  toxicity_annotator_count               int64
dtypes: float64(32), int64(10), object(3)
memory usage: 619.7+ MB
```

Fig. 2. Details about the train dataset

3 Proposed Methodology

3.1 Dataset

(1) Created date: We have a single csv file for testing and single csv file for training (Fig. 3).
(2) Comment_text: This is the data in string format is used to find the toxicity.
(3) Target: Target values which are to be predicted (has values between 0 and 1).
(4) Data also has additional toxicity subtype attributes: (Model does not have to predict these) severe_toxicity, obscene, threat, insult, identity_attack, sexual_explicit.
(5) Comment_text data also has identity attributes carved out from it, some of which are: Male, female, homosexual_gay_or_lesbian, christian, jewish, muslim, black, white, asian, latino, psychiatric_or_mental_illness.
(6) Apart from above features the train data also provides meta-data from jigsaw like: Toxicity_annotator_count, identity_anotator_count, article_id, funny, sad, wow, likes, disagree, publication_id, parent_id, article_id.
(7) Hyper parameters we used in building this model are mean squared error for finding the loss, rmsprop as an optimizer, and sigmoid function as an activation function for the output layer.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding (Embedding)        (None, 350, 100)          8700

 cu_dnnlstm (CuDNNLSTM)       (None, 350, 128)          117760

 cu_dnnlstm_1 (CuDNNLSTM)     (None, 128)               132096

 dense (Dense)                (None, 1)                 129

=================================================================
Total params: 258,685
Trainable params: 258,685
Non-trainable params: 0
_________________________________________________________________
None
```

Fig. 3. Summary of the model

3.2 Type of Machine Learning Problem

We must predict the level of toxicity (target attribute). The range of values is 0–1, inclusive. This is a case of regression. It can also be regarded as a classification problem if we consider all values below 0.5 to be non-toxic and all values above it to be harmful, giving us a binary classification problem.

3.3 Performance Metric

The competition will use ROC_AUC as the metric after converting the numeric target variable into a categorical variable by using a threshold of 0.5. Any comment above 0.5 will be assumed to be toxic and below it non-toxic. For our training and evaluation, we

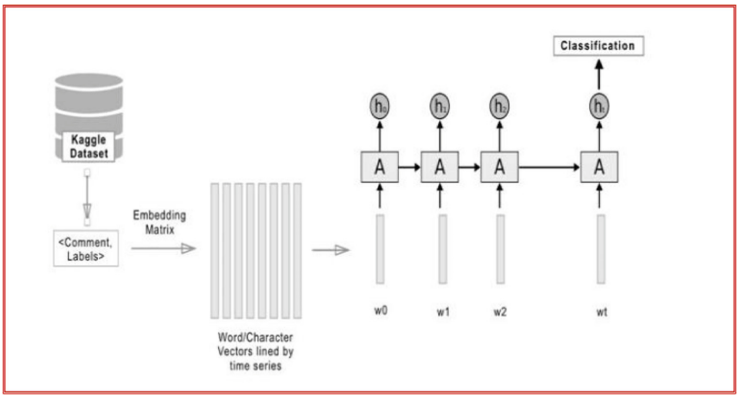


Fig. 4. Toxic comments classification

will use the MSE (Mean Squared Error) (Fig. 4).

3.4 Machine Learning Objectives and Constraints

We have used the LSTM model.

Objectives:

Predict the toxicity of a comment made by the user. (0 → not toxic, 1 → highest toxicity level).

Constraints:

- The model should be fast to predict the toxicity rating.
- Interpretability is not needed.

4 Experimental Results

Based on the training dataset we classified the toxic comments with the parameters such as insult, obscene, identity_attack, threat, and severe_toxicity.

Result of Epoch based long short term memory (LSTM) approach has higher achievement than the Ensemble Deep Learning approach (Fig. 5).

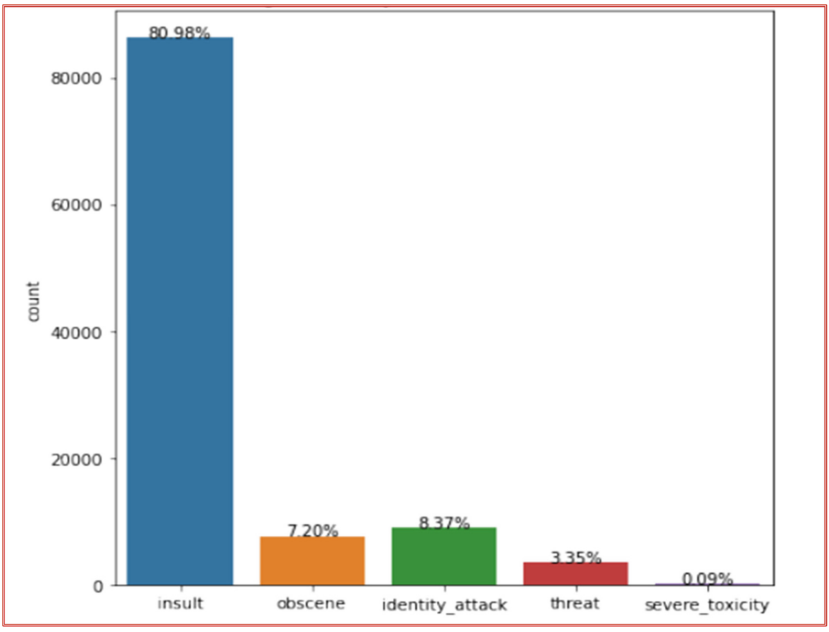


Fig. 5. Percentage of toxicity in toxic comments data.

In our train dataset only 8% of the data was toxic. Out of that 8%, 81% of the toxic comments made are insults, 8.37% are identity attacks, 7.20% are obscene, 3.35% are threats and a very small number of toxic comments are severely toxic (Fig. 6, 7).

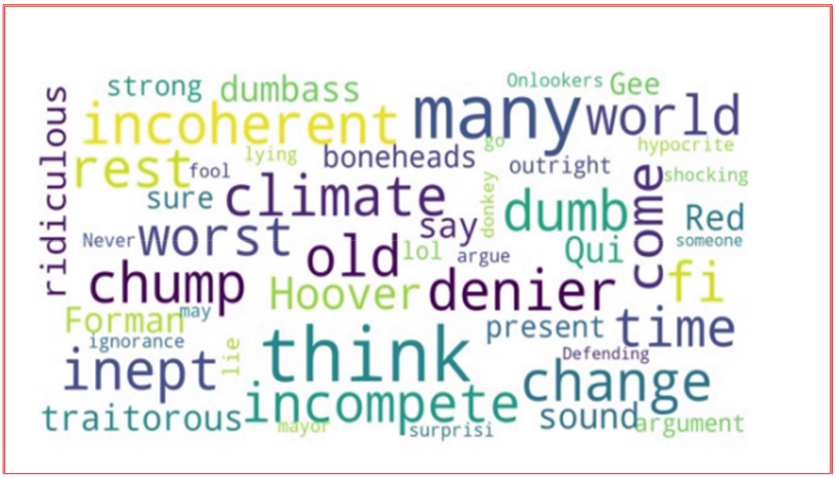


Fig. 6. Prevalent comments with insult score >0.75

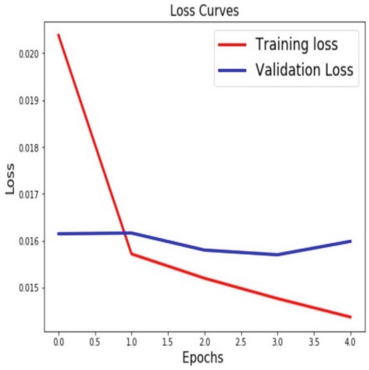


Fig. 7. Loss curves of LSTMs

Initially our model suffered from under fitting, but as the network trained in each epoch, underfitting was resolved and the model is regularly fit (Fig. 8).

5 Conclusion

The internet being a public platform, it is critical to ensure that people with diverse viewpoints are heard without fear of poisonous or hostile comments. After examining numerous techniques to solving the problem of online harmful comment classification,

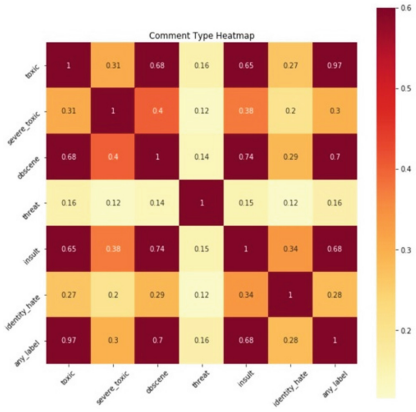


Fig. 8. Heat map

we decided to employ the LSTM strategy for greater accuracy. The future focus of this project can be to understand and give a relevant reply or help to these classified positive comments and ignore the negative comments. This can be used in social media platforms to verify whether a comment is positive or negative and if so negative, can be prevented.

References

1. Guggilla, C., Miller, T., Gurevych, I.: CNN-and LSTM-based claim classification in online user comments. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2740–2751 (2016)
2. Jabreel, M., Moreno, A.: A deep learning-based approach for multi-label emotion classification in tweets. Appl. Sci. **9**(6), 1123 (2019)
3. Haralabopoulos, Anagnostopoulos, I., & McAuley, D.: Ensemble deep learning for multilabel binary classification of user-generated content. Algorithms **13**(4), 83 (2020)
4. Sridharan, M., Swapna, T.R.: Amrita School of Engineering-CSEatSemEval-2019 Task 6: Manipulating attention with temporal convolutional neural network for offense identification and classification. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 540–546 (2019)
5. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-based transfer learning approach for hate speech detection in online social media. In: International Conference on Complex Networks and Their Applications, pp. 928–940. Springer, Cham (2019)
6. Liang, J., Meyerson, E., Hodjat, B., Fink, D., Mutch, K., Miikkulainen, R.: Evolutionary neural automl for deep learning. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 401–409 (2019)
7. Kajla, H., Hooda, J., Saini, G.: Classification of online toxic comments using machine learning algorithms. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1119–1123 (2020). IEEE
8. Feurer, M., Hutter, F.: Hyperparameter optimization. In: Automated Machine Learning (pp. 3–33). Springer, Cham. Zhang, X., Liao, Q., Kang, Z., Liu, B., Ou, Y., Du, J., ... & Fang, Z.: Self-healing originated van der Waals homojunction with strong interlayer coupling for high-performance photodiodes. ACS Nano, **13**(3), 3280–3291 (2019)

9. Tabassi, E., Burns, K.J., Hadjimichael, M., Molina-Markham, A.D., Sexton, J.T.: A Taxonomy and Terminology of Adversarial Machine Learning, (2019)
10. Sunitha, G., et al.: Modeling of chaotic political optimizer for crop yield prediction. Intelligent Automation and Soft Computing **34**(1), 423–437 (2022)
11. Sunitha, G., Arunachalam, R., Abd-Elnaby, M., Eid, M.M., Rashed, A.N.Z.: A comparative analysis of deep neural network architectures for the dynamic diagnosis of COVID-19 based on acoustic cough features. Int. J. Imaging Systems Tech. (2022)
12. Karthikeyan, C., Sunitha, G., Avanija, J., Reddy Madhavi, K., Madhan, E.S.: Prediction of climate change using SVM and naïve bayes machine learning algorithms. Turkish Journal of Computer and Mathematics Education **12**(2), 2134–2139 (2021)
13. Abbagalla, S., Rupa Devi, B., Anjaiah, P., Reddy Madhavi, K.: "Analysis of COVID-19-impacted zone using machine learning algorithms". Springer series – Lecture Notes on Data Engineering and Communication Technology, Vol.63, 621–627 (2021)
14. Avanija, J., Sunitha, G., Hittesh Sai Vittal, R.: "Dengue outbreak prediction using regression model in chittoor district, Andhra Pradesh, India." Int. J. Recent Tech. Engineer. **8**(4), 10057–10060 (2019). doi: https://doi.org/10.35940/ijrte.d9519.118419
15. Reddy Madhavi, K., et al.: "COVID-19 detection using deep learning", In: 20th International Conference on Hybrid Intelligent Systems-HIS 2020, at Machine Intelligence Research (MIR) labs, USA, Springer AISC, 1375, pp 1–7 (2020)
16. Kora, P., Rajani, A., Chinnaiah, M.C., Madhavi, R. Swaraja, K., Kollati, M.: EEG-Based brain-electric activity detection during meditation using spectral estimation techniques. pp. 687–693 (2021) doi: https://doi.org/10.1007/978-981-16-1941-0_68
17. Prabhakar, T., Srujan Raju, K., Reddy Madhavi, K.: Support vector machine classification of remote sensing images with the wavelet-based statistical features. In: Fifth International Conference on Smart Computing and Informatics (SCI 2021), Smart Intelligent Computing and Applications, Volume 2. Smart Innovation, Systems and Technologies, vol 283. Springer, Singapore (2022)
18. Rajani, A., Kora, P., Madhavi, R. Jangaraj, A.: Quality improvement of retinal optical coherence tomography. 1–5 (2021) https://doi.org/10.1109/INCET51464.2021.9456151
19. Reddy Madhavi, K., Madhavi, G., Rupa Devi, B., Kora, P.: "Detection of pneumonia using deep transfer learning architectures", Int. J. Advanced Trends Computer Sci. Engineer. **9**(5), pp. 8934- 8937 (2020). ISSN 2278-3091 https://doi.org/10.30534/ijatcse/2020/292952020

Topic Modeling Approaches—A Comparative Analysis

D. Lakshminarayana Reddy[1(✉)] and C. Shoba Bindu[2]

[1] Research Scholar, Department of Computer Science and Engineering, JNTUA, Anantapuramu, Andhra Pradesh, India
lakshmi1217@gmail.com

[2] Department of Computer Science and Engineering, JNTUACEA, Anantapuramu, Andhra Pradesh, India
shobabindhu.cse@jntua.ac.in

Abstract. Valuable information from a corpus for a specific purpose can be obtained by finding, extracting, and processing the text through text mining. A corpus is a group of documents and the documents could be anything from newspaper articles, tweets, or any kind of data that needs to study. For processing and understanding the structure of a corpus, a technique in text mining is Natural Language Processing (NLP). The study of a corpus in different fields like bioinformatics, software engineering, sentiment analysis, Education, and Linguistics (scientific research) is a challenging task as it contains a vast amount of data. Thus for latent data identification, establishing connections between data and text documents needs topic modeling. Evolutions of topic models are analyzed in this paper from 1990 to present. To better understand the topic modeling concept, a detailed evaluation of techniques is discussed in detail. In this study, we looked into highly scientific articles from 2010 to 2022 based on methods used in different areas to discover current trends, research development, and the intellectual structure of topic modeling.

Keywords: Natural Language Processing · Scientific research · Sentiment analysis · Bio-informatics · Software engineering

1 Introduction

In recent years, extracting desired and relevant information from data is tough as the size of data is growing for the analytics industry. However, technology has created several strong techniques that can be utilized to mine the data and extract the information we need. Topic Modeling is one of these text-mining techniques. A method of unsupervised machine learning is Topic modeling. As the name implies, it is a method for automatically determining the themes that exist in a text object and deriving latent patterns displayed by a text corpus facilitating wiser decision-accomplishing as a result. Due to its huge capability, topic modeling has found applications across a wide range of fields, including natural language processing, scientific literature, software engineering, bioinformatics, humanities, and more.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 881–892, 2023.
https://doi.org/10.1007/978-3-031-27409-1_81

As shown in Fig. 1, a collection of papers can be scanned using topic modeling to identify word and phrase patterns, and then the words and phrases that best characterize the collection can be automatically arranged. Especially if you work for a business that processes a whole lot, or thousands of client interactions daily, It is difficult to analyze data from virtual entertainment posts, emails, conversations, unconditional overview answers, and other sources, and it becomes even more difficult when done by people. Recognizing words from topics in a document or data corpus is known as topic modeling. Extracting words from a document is more troublesome and tedious than extracting from topics available in the content. Topic modeling helps in doing this.

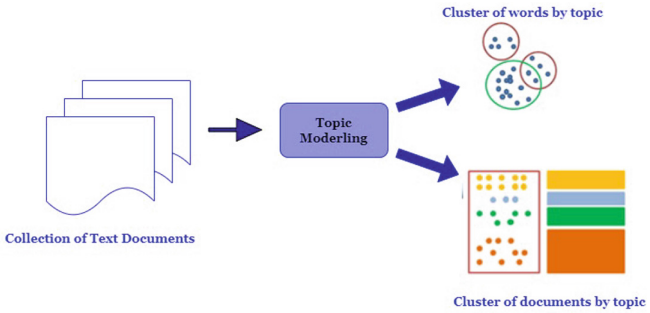


Fig. 1. Graphical representation map for topic modeling

For example, there are 5,000 documents and each document contains 600 words. To process this, 600*5.000 = 3,000,000 threads are needed. So when the document is partitioned if it contains 10 topics then the processing is just 10*600 = 6,000 threads. This looks simple than processing the entire document and this is how topic modeling has come up to solve the problem and also visualize things better.

With the evolvement of topic modeling, researchers have shown great interest across a wide range of research fields. Figs. 2 and 3 show the advancements of topic modeling from inception to present and the frequently used methods are addressed below.

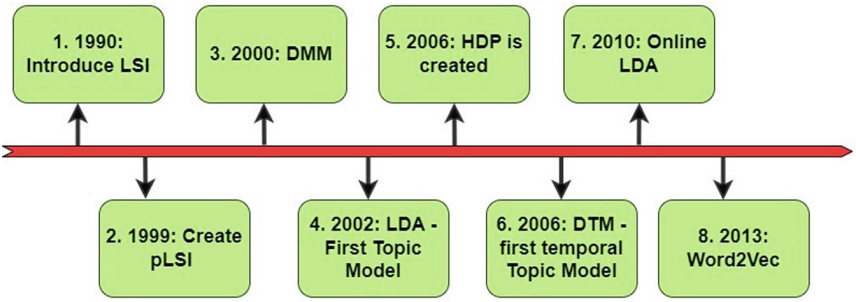


Fig. 2. Evolution of topic modeling from inception to the introduction of Word2Vec

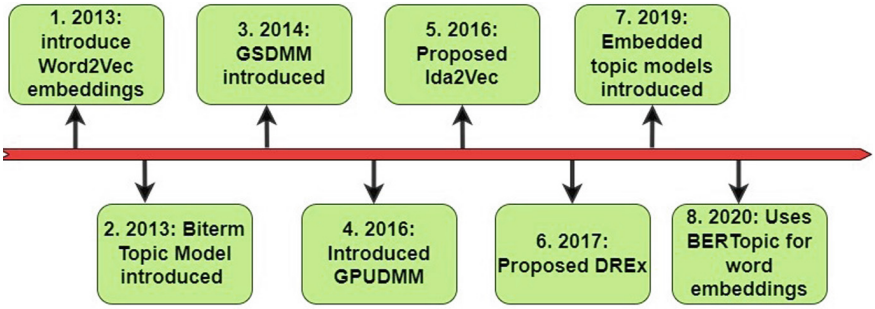


Fig. 3. Evolution of topic modeling from Word2Vec to present

The Latent semantic Index (LSI) [1] also called Latent Semantic Analysis (LSA) is described how to index automatically and retrieve files from huge databases. LSI is an unsupervised learning method that helps in choosing required documents by extracting the relationship between different words from a group of documents. LSI employs the bag of words (BoW) model, resulting in a term-document matrix. Documents and Terms are considered rows and columns in this matrix. Singular value decomposition (SVD) helps in learning latent topics from the matrix decomposition on the term-document matrix. The LSI is used as a noise reduction or document reduction technique. The probabilistic Latent Semantic Index (pLSI) [2] produces precise results more accurately than the LSI. pLSI solves the representation challenges in LSI.

Latent Dirichlet Allocation (LDA) [3] overcomes the problems in pLSI. The relationships between multiple documents can be achieved through LDA which is an analytical and graphical model. LDA finds more accurate topics not only with one topic but with probabilistically generated from many topics. (iv) Non-Negative Matrix Factorization (NMF) [4] is faster than LDA with more consistency. In NMF, the document-term matrix is taken into consideration, which was extracted from a corpus following the stop-words. The matrix will be factorized into two matrices term-topic matrix and topic-document matrix. Here, factorization is achieved by updating one column at a time while maintaining the values of the other columns.

Word2Vec [5] is the cutting edge of prediction-based word embedding. In word2vec feature vector is calculated for every word in a corpus. The word2vec model is modified by lda2vec [6] to produce document vectors in addition to word vectors. Document vectors make it possible to compare documents and compare documents to words or phrases. The algorithm for topic modeling and semantic search is called top2Vec [7]. It automatically identifies topics that are present in the text and produces embedded topics, documents, and word vectors at the same time. The latest topic model method is BERTopic [8]. For creating dense clusters BERTopic technique uses transformers and c-TF-IDF that make the topic simple to understand. BERTopic supports guided, (semi-) supervised, and dynamic topic modeling, Even LDAvis-like visuals are supported by it.

Following was the discussion of the remainder of the paper. The search criteria, search technique, and research methodology are all covered in Sect. 2 of this review. The effects of TM techniques on various fields are discussed in Sect. 3. Section 4 provides a thorough

explanation of the findings that have been reached as well as research difficulties in many application areas for future advancements. Section 5 discusses the conclusion.

2 Research Methodology

It is the exact set of steps or methods used to identify, select, process, and analyze data on a topic. It enables readers to assess the study's reliability and validity in the research document. Which helps in finding any incomplete research needs?

The PRISMA statement was followed in conducting this systematic literature review. In this research, scholarly works from 2010 to 2022 are examined about topic modeling approaches, with each article's shortcomings explained in turn, followed by suggestions to address those drawbacks.

2.1 Research Questions

Topic modeling approaches are aimed at addressing their performance in various areas. The approaches are categorized into the following questions:

- RQ1: Identification of topics in sentiment analysis needs which type of topic modeling methods.
- RQ2: Identification of topics in scientific research needs which type of topic modeling methods.
- RQ3: Identification of topics in bioinformatics needs which type of topic modeling methods.
- RQ4: Identification of topics in software engineering needs which type of topic modeling methods.

2.2 Search Strategy

It is crucial to consider pertinent keywords that can identify related articles and take out irrelevant information because index phrases act as "keys" to separate scientific papers from other articles. The keywords that have been carefully considered are "Topic Modeling", "Topic Modeling methods", "word embeddings", "clustering", "Classification", "aspect extraction" and "Natural Language Processing".

We recommend databases that regularly publish articles on the themes given our familiarity with publishing. The databases listed below were picked: Scopus, Web of Science, ArXiv, IEEE Xplore Digital Library, PubMed, and Taylor Francis.

2.3 Search Results

The selection of papers is outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Flow Chart. in Fig. 4. The number of papers considered in each approach and year-wise publications are depicted in Fig. 5.

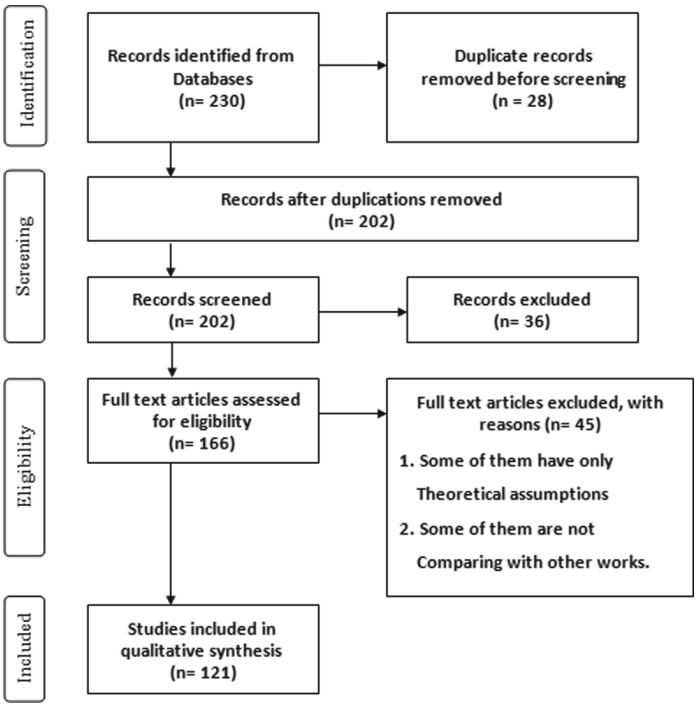


Fig. 4. PRISMA flowchart for research papers selection

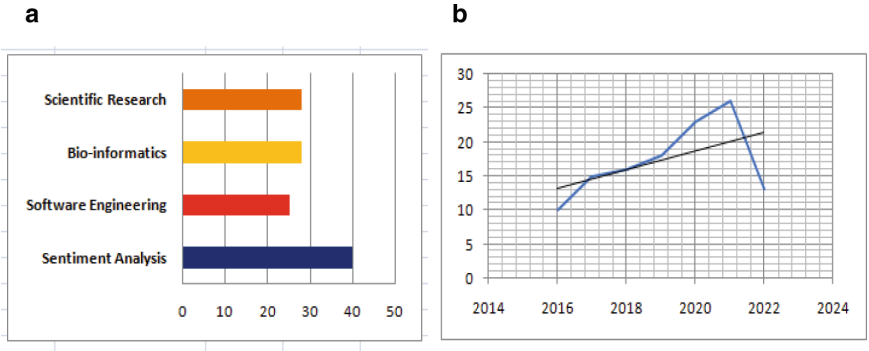


Fig. 5. **a** Number of publications considered. **b** Annual publications

3 Analysis of Topic Modeling Approach

Topic modeling is not a recent application. However, the number of papers using the strategy for classifying research papers is very low. It has mostly been utilized in various areas to locate concepts and topics in a corpus of texts. The following tables provide an overview of topic modeling approaches, including the Objective, topic modeling approach, and dataset.

3.1 Topic Modeling in Sentiment Analysis

The act of computationally recognizing and classifying opinions stated in a text, particularly to ascertain whether the writer has a positive, negative, or neutral viewpoint on a given topic, item, etc. To extract this sentiment number of researchers are going on different social media platforms like Facebook, Whatsapp, Twitter, Instagram, and Sina Weibo(Chinese social media platform), etc. (Table 1).

Table 1. Summary of topic modeling papers in sentiment analysis

References	Objective	Method used	Data set
Yin et al. [9]	Analyzing the discussions on the COVID-19 vaccine	LDA	Vaccine tweets
Amara et al. [10]	Tracking the COVID-19 pandemic trends	LDA	Facebook public posts
Zoya et al. [11]	Analyzing LDA and NMF Topic Models	LSA, PLSA, LDA, NMF	Urdu tweets
Pang et al. [12]	Detect emotions from short messages	WLTM and XETM	News headlines, Chinese blogs
Ghasiya et al. [13]	Determine and understand the critical issues and sentiments of COVID-19-related news	top2vec and RoBERTa	Covid-19 news
Ozyurt et al. [14]	Aspect extraction	Sentence Segment LDA (SS-LDA)	Blogs and websites
Wang et al. [15]	Classifying and summarizing the Public sentiment analysis during the COVID-19	BERT	COVID-19-posts
Daha et al. [16]	Mining public opinion on climate change	LDA	Geo-tagged tweets

The above table shows the research findings in the field of Sentiment Analysis. Daha et al. [16] proposed an author pooled LDA to analyze geo-tagged tweets to mine public opinion to classify sentiment, it has some limitations that are the nature of the Twitter data set because tweets are indecipherable. It makes both topic modeling and sentiment analysis ineffective on those tweets. So he proposes to combine topic modeling with sentiment analysis to produce sentiment alignment (positive, negative) associated with them. To classify sentiment categories (positive, negative, neutral) by combining topic modeling and sentiment analysis Wang et al. [15] propose an unsupervised BERT model with TF-IDF. The limitation of this study is using only a Chinese platform to classify sentiment. To overcome this Ghasiya et al. [13] propose top2ves and RoBERTa methods

to classify sentiments from different nations. Yin et al. [9] proposed the LDA-based model for analyzing discussions on COVID-19 with the tweets posted by users.

3.2 Topic Modeling in Scientific Research

In the scientific research field, topic modeling methods are classifies research papers according to topic and language. Within and across the three academic fields of linguistics, computational linguistics, and education, several illuminating statistics and correlations were discovered (Table 2).

Table 2. Summary of topic modeling papers in scientific research

References	Objective	Method used	Data set
Gencturk et al. [17]	Examining Teachers knowledge	Supervised LDA(SLDA)	Teachers Responses
Chen et al. [18]	Detecting trends in educational technologies	Structural topic model(STM)	Published papers
Chen et al. [19]	Identifying trends, explore the distribution of paper types	LDA	Publications
Yun et al. [20]	Reviewed the trends of research in the field of physics education	LDA	Newspaper Articles
Chang et al. [21]	Latent topics extracted from the dataset of different languages	A cross-lingual topic model, called Cb-CLTM	News domain
Wang et al. [22]	Automatic-related work generation	QueryTopicSum	Document set

The above table shows the research findings in the field of linguistics, computational linguistics, and education fields, etc., Gencturk et al. [17] proposed supervised LDA(SLDA) on teachers' responses to examining the teacher's knowledge. It uses a small dataset so it is difficult to understand complex problems. Chen et al. [18] introduced the structural Topic Model(STM) to find trends in educational technologies but it finds trends based on a single journal only. Yun et al. [20] reviewed the trends in education with the LDA method in AJP and PRPER journals. These two journals have the highest Coherence value for 8 topics. Chang et al. [21] compare the topics on a cross-lingual dataset with Cb-CLTM. This method generates more coherence value on US-Corpus compared with PMLDA. Wang et al. [22] proposed a new framework called ToC-RWG for generating related work and present QueryTopicSum for characterizing the process generation in scientific papers and reference papers. Here QueryTopicSum performs better than the TopicSum, LexRank, JS-Gen, and Sumbasic.

3.3 Topic Modeling in Bioinformatics

For interpreting biological information topic modeling improves the researcher's capacity. The exponential rise of biological data, such as microarray datasets has been happening recently. So extracting concealed information and relations is a challenging task. Topic models have proven to be an effective bioinformatics tool since Biological objects can be represented in terms of hidden topics (Table 3).

Table 3. Summary of topic modeling papers in bioinformatics

References	Objective	Method used	Data set
Heo et al. [23]	Investigate the bioinformatics field to analyze keyphrases, authors, and journals	ACT model	Journals
Gurcan et al. [24]	Analyzing the main topics, developmental stages, trends, and future directions	LDA	Articles
Porturas et al. [25]	Attempted to identify the most prevalent research themes in emergency medicine	LDA	Articles and abstracts
M. Gao et al. [26]	Discovering the features of topics	Neural NMF	Toy data set
Wang et al. [27]	Bioinformatics knowledge structure is being detected	Doc2vec	Journals, conferences
Zou et al. [28]	Research topics are discovered for drug safety	LDA	Titles and abstracts

The above table shows the research findings in the field of bioinformatics. In this Zou et al. [28] proposed LDA model on titles and abstracts for drug safety measures. It assumes a fixed number and known topics so the computational complexity is less. M. Gao et al. [26] proposed neural Non-Negative Matrix Factorization Method for Discovering the features in the medical dataset. The high-intensity features are not resolved in this method. Porturas et al. [25] used the LDA model for identifying research themes in emergency medicines with human interventions it is the major drawback in this study. Next Gurcan et al. [24] analyzing the trends and future studies in corpus with LDA. To accomplish wide-ranging topic analyses of key phrases, authors, and journals heo et al. [23] investigate the bioinformatics field with the ACT (Author's-Conference-Topic) model. This model was paying attention to genetics key phrases but not to subjects connected to informatics. Wang et al. [27] detect the knowledge structure in bioinformatics with the doc2vec method integrated with dimension reduction technology and clustering technology.

3.4 Topic Modeling in Software Engineering

Topic modeling plays a vital role in examining textual data in empirical research, creating new methodologies, predicting vulnerabilities, to find duplicate bug reports related to software engineering tasks. Application of topic modeling must be done based on modeling parameters and type of textual data. There is another important concept in Software Engineering is vulnerability. It is the indicator of reliability and safety in the software industry (Table 4).

Table 4. Summary of topic modeling papers in software engineering

Reference	Objective	Method used	Data
Gurcan et al. [29]	Detecting Latent Topics and Trends	LDA	Articles
Akilan et al. [30]	Detection of Duplicate Bug Reports	LDA	Eclipse dataset bug reports
Pérez et al. [31]	Locates features in software models	LDA	Models
Gül Bulu et al. [32]	Predicting the software vulnerabilities	LDA	Bug records
Johri et al. [33]	Identifying trends in technologies and programming languages	LDA	Textual data
Corley et al. [34]	Analyzing the use of streaming (online) topic models	Online LDA	Files changed information

The above table shows the research findings in the field of Software Engineering. Gurcan et al. [29] proposed the LDA model to identify the latest trends and topics in the software industry based on the articles published in various papers. After applying LDA to the dataset 24 topics were revealed. Empirical software engineering had the greatest ratio among the investigated topics (6.62%), followed by projects (6.43%) and architecture (5.74%). The lowest ratios were for the topics "Security" (1.88%) and "Mobile" (2.08%). In finding the duplicate bug reports researchers used large datasets to analyze. If the size is large there's a chance, the related master report doesn't exist in the chosen group. To overcome this Akilan et al. [30] propose LDA based clustering method. To locate the features in software models Pérez et al. [31] proposed LDA-based method on different software models. This model performs better than the baseline for interpreted models in terms of recall, precision, and F-measure. This is not the case for code-generation models. To predict software vulnerabilities Gül Bulu et al. [32] proposed LDA model along with regression and classification methods. In this, the best regression model results are 0.23, 0.30, and 0.44 MdMRE values, respectively and the best classification model result is a 74% recall score.

4 Discussion and Future Directions

In the process of finding the topics from large documents, the performance of topic models is evaluated based on Topic coherence (TC), Topic diversity (TD), Topic quality (TQ) metrics, and some standard statistical evaluation metrics like recall, precision, and F-score. On analyzing existing literature studies the following inferences were derived to ascertain the further scope of research.

- As most of the techniques are derived from existing methods in topic modeling there is a need to address optimization of the topic modeling process (sampling strategy, feature set extraction, etc.) to enhance the classification and feature extraction and reduce computational load.
- Topic modeling on Transcriptomic Data is an open research challenge in the medical domain for the analysis of breast and lung cancer.
- There has been some research on deriving inferences from the psychological dataset, but most of these works fail to achieve accuracy in clustering topics.
- Finding out the definitive topic model that is accurate and reliable is a challenging aspect since the results of classic topic modeling are unsteady in aspects of both when the model is retained on the same input documents and reformed with new documents.
- From various language families such as English–Chinese and English–Japanese extraction of corpora for common topics is done till now but not from the same language family such as Indo-European.
- Further research is required as mining unstructured texts is still at the beginning in the construction domain.
- While maintaining data privacy, more attention is required for sharing data in a cloud environment for this reason research and applications are needed in data privacy.

5 Conclusion

230 papers were analyzed regarding topic modeling in different areas. The characteristics and limitations of topic models were identified by analyzing topic modeling techniques, data inputs, data pre-processing, and the naming of topics. This study helps researchers and practitioners for the best use of topic modeling taking into account the lessons learned from other studies. In this paper, we have analyzed the topic modeling related to various areas and identified the common limitations in all areas like reliability, accuracy, and privacy in the big data era and most of the researchers use different language families for extracting common topics in scientific research. Most papers use a single nation's social media data (Twitter in India, Sina Weibo in china) for extracting sentiment in sentiment analysis. Optimization of the topic modeling process, Transcriptomic data in the medical domain, Applying the visualization to explore the aspects, building the visualization chatbots to browse tools, datasets, and research topics, and extraction of topics of the same language family, such as Indo-European languages, use positive, negative and neutral labels for finding sentiment are the research areas that can be focussed on.

References

1. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6) 391–407 (1990)
2. Hofmann, T.: Probabilistic latent semantic indexing. In SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems (NIPS), pp. 601–608 (2002)
4. Vavasis, S.A.: On the complexity of nonnegative matrix factorization. SIAM Journal on Optimization **20**(3), 1364–1377 (2010)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NIPS), pp. 3111–3119 (2013)
6. Moody, C.E.: Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. CoRR (2016)
7. Dimo, A.: (2020). Top2Vec: Distributed Representations of Topics
8. Grootendorst, Maarten. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure
9. Yin, H., Song, X., Yang, S., Li, J.: Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web. **25**, 1–17 (2022)
10. Amara, A., Taieb, H., Ali, M., Aouicha, B., Mohamed.: Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. Appl. Intell. **51**, 1–22 (2021)
11. Zoya, Latif, S., Shafait, F., Latif, R.: Analyzing LDA and NMF topic models for urdu tweets via automatic labeling. In: IEEE Access **9**, 127531–127547 (2021)
12. Pang, J., et al.: Fast supervised topic models for short text emotion detection. IEEE Trans. Cybern. **51**(2), 815–828 (2021)
13. Ghasiya, P., Okamura, K.: Investigating COVID-19 news across four nations: a topic modeling and sentiment analysis approach. IEEE Access **9**, 36645–36656 (2021)
14. Ozyurt, Baris & Akcayol, M.. (2020). A new topic modeling-based approach for aspect extraction in aspect-based sentiment analysis: SS-LDA. Expert. Syst. Appl. **168**
15. Wang, T., Lu, K., Chow, K.P., Zhu, Q.: COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. IEEE Access **8**, 138162–138169 (2020)
16. Dahal, B., Kumar, S., Li, Z.: Spatiotemporal topic modeling and sentiment analysis of global climate change tweets. social network analysis and mining (2019)
17. Copur-Gencturk, Y., Cohen, A., Choi, H.-J. (2022). Teachers' understanding through topic modeling: a promising approach to studying teachers' knowledge. J. Math. Teach. Educ.
18. Chen, X., Zou, D., Cheng, G., Xie, H.: Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. Comput. Educ. 151 (2020)
19. Chen, X., Zou, D., Xie, H.: Fifty years of British journal of educational technology: a topic modeling based bibliometric perspective. Br. J. Educ. Technol. (2020)
20. Yun, E.: Review of trends in physics education research using topic modeling. J. Balt. Sci. Educ. **19**(3), 388–400 (2020)
21. Chang, C.-H., Hwang, S.-Y.: A word embedding-based approach to cross-lingual topic modeling. Knowl. Inf. Syst. **63**(6) 1529–1555 (2021)
22. Wang, P., Li, S., Zhou, H., Tang, J., Wang, T.: ToC-RWG: explore the combination of topic model and citation information for automatic related work generation. IEEE Access **8**, 13043–13055 (2020)

23. Heo, G., Kang, K., Song, M., Lee, J.-H.: Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. BMC Bioinform. **18** (2017)
24. Gurcan, F., Cagiltay, N.E.: Exploratory analysis of topic interests and their evolution in bioinformatics research using semantic text mining and probabilistic topic modeling. IEEE Access **10**, 31480–31493 (2022)
25. Porturas, T., Taylor, R.A.: Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling. Am J Emerg Med. **45**, 213–220 (2021)
26. M. Gao, et al., Neural nonnegative matrix factorization for hierarchical multilayer topic modeling. In: 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 6–10 (2019)
27. Wang, J., Li, Z., Zhang, J. Visualizing the knowledge structure and evolution of bioinformatics. BMC Bioinformatics **23** (2022)
28. Zou, C.: Analyzing research trends on drug safety using topic modeling. Expert Opin Drug Saf. **17**(6), 629–636 (2018)
29. Gurcan, F., Dalveren, G.G.M., Cagiltay, N.E., Soylu, A.: Detecting latent topics and trends in software engineering research since 1980 using probabilistic topic modeling. IEEE Access **10**, 74638–74654 (2022)
30. Akilan, T., Shah, D., Patel, N., Mehta, R.: Fast detection of duplicate bug reports using LDA-based Topic Modeling and Classification. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1622–1629 (2020)
31. Pérez, F., Lapeña Martí, R., Marcén, A., Cetina, C.: Topic modeling for feature location in software models: studying both code generation and interpreted models. Inf. Softw. Technol. **140** (2021)
32. Bulut, F. G., Altunel, H., Tosun, A.: Predicting software vulnerabilities using topic modeling with issues. In: 2019 4th International Conference on Computer Science and Engineering (UBMK), pp. 739–744 (2019)
33. Johri, V., Bansal. S.: Identifying trends in technologies and programming languages using topic modeling. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 391–396 (2018)
34. Corley, C. S., Damevski, K., Kraft, N. A.: Changeset-based topic modeling of software repositories. In: IEEE Trans. Softw. Eng. **46**(10), 1068–1080 (2020)



Survey on Different ML Algorithms Applied on Neuroimaging for Brain Tumor Analysis (Detection, Features Selection, Segmentation and Classification)

K. R. Lavanya[1(✉)] and C. Shoba Bindu[2]

[1] Research Scholar, Dept. of CSE, JNTUA Ananthapur, Anantapuramu, India
lavanya.jntuacek@gmail.com
[2] Director of Research & Development, Dept. of CSE, JNTUA Ananthapur, Anantapuramu, India

Abstract. The Brain Tumor is one of the main causes of cancer deaths in the world. Exact reasons for brain tumor may not be specified but survival rate can be increased by detecting at early stage and well analyzing it. This papers presents analysis of many Machine Learning algorithms and approaches that are emerged and used for brain tumor detection, features selection, segmentation and classification along with the type of neuro image modalities and techniques used for brain tumor analysis, over the past three years. This presents that most of the search is being done on 2D MRI images.

Keywords: Neuroimaging · Machine learning · Brain tumor · Detection · Segmentation · Features selection

1 Introduction

1.1 A Subsection Sample

Brain tumor is the abnormal growth of tissue cells within the skull which may leads to impairment or life-threatening condition. Early diagnosis of such disease may help radiologists and oncologists to provide correct and better treatment which may increase survival rate of a patient. Brain tumors may be Low Grade tumors or High grade tumor like benign to malignant tumors. Figure 2 presents images with different grades of brain tumors observed through MRI.

It is estimated that around 3,08,102 people, worldwide, were diagnosed with primary brain or spinal cord tumor in 2020 and around 2,51,329 people, worldwide, died from primary cancerous brain and CNS (Central Nervous System) tumors in 2020 [41].

Immense new technologies are being emerged to diagnosis many diseases which lead to do great research work on neuroimaging to help radiologists and oncologists by increasing the accuracy of brain tumor analysis through Machine Learning approaches.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 893–906, 2023.
https://doi.org/10.1007/978-3-031-27409-1_82

[41] There are many ways to diagnose brain tumors like: Neuroimaging, biopsy, Cerebral angiogram, lumber puncture or spinal tap, Myelogram, EEG, and so on.

Neuroimaging of a brain helps the doctors to study the brain which in turn helps in providing treatment. [42] Neuroimaging can be a structural imaging which deals with the structure of a brain for diagnosing tumors, injuries, hemorrhages, etc., and functional imaging to measure the aspect of a brain function that defines the relationship between activity of a brain area and mental functioning, which helps in Psychological studies.

[42] Neuroimaging can be obtained through different technologies like:

(a) Computed Tomography (CT) scan that uses a series of X-ray beams to create cross-sectioned images of brain to get the structure of a brain for analysis.
(b) MRI (Magnetic Resonance Imaging): It uses echo waves to differentiate Grey matter, White Matter and cerebrospinal fluid. It is standard neuroimaging modality.
(c) Functional MRI(fMRI): Scans a series of MRIs measuring brain function. It is a functional neuroimaging technique.
(d) T1-Weighted MRI: It is standard imaging test and a part of general MRI to give clear view of brain anatomy and structure. It will be preferred only when the damage is very significant.
(e) T2-Weighted MRI: It is also a standard modality of MRI which is used to measure White Matter and Cerebrospinal fluid in the brain, as it is more suitable to measure fluid rather than soft tissues.
(f) Diffusion-Weighted MRI(DWI): It presents changes in tissue integrity. It helps in identifying stroke or ischemic injury in the brain.
(g) Fluid-Attenuated Inversion Recovery MRI (FLAIR): It is sensitive to water content in the brain tissue. Mostly FLAIR-MRI is used to visualize changes in the brain tissue.
(h) Gradient Record MRI(GRE): It is used to detect hemorrhaging in the brain tissue. Using this, micro bleeds can also be detected.
(i) Positron Emission Tomography Scan (PET): It shows how different areas of the brain use oxygen and glucose. It is also used in identifying metabolic processes.
(j) Diffusion Tensor Imaging (DTI); It is used for White Matter tracts in brain tissue. It gives information about damage to parts of CNS and also information about connections among brain regions. Figure 1 shows different neuro images acquired through different neuroimaging techniques. Figure 2 shows different grades of brain tumors at different locations of a brain.

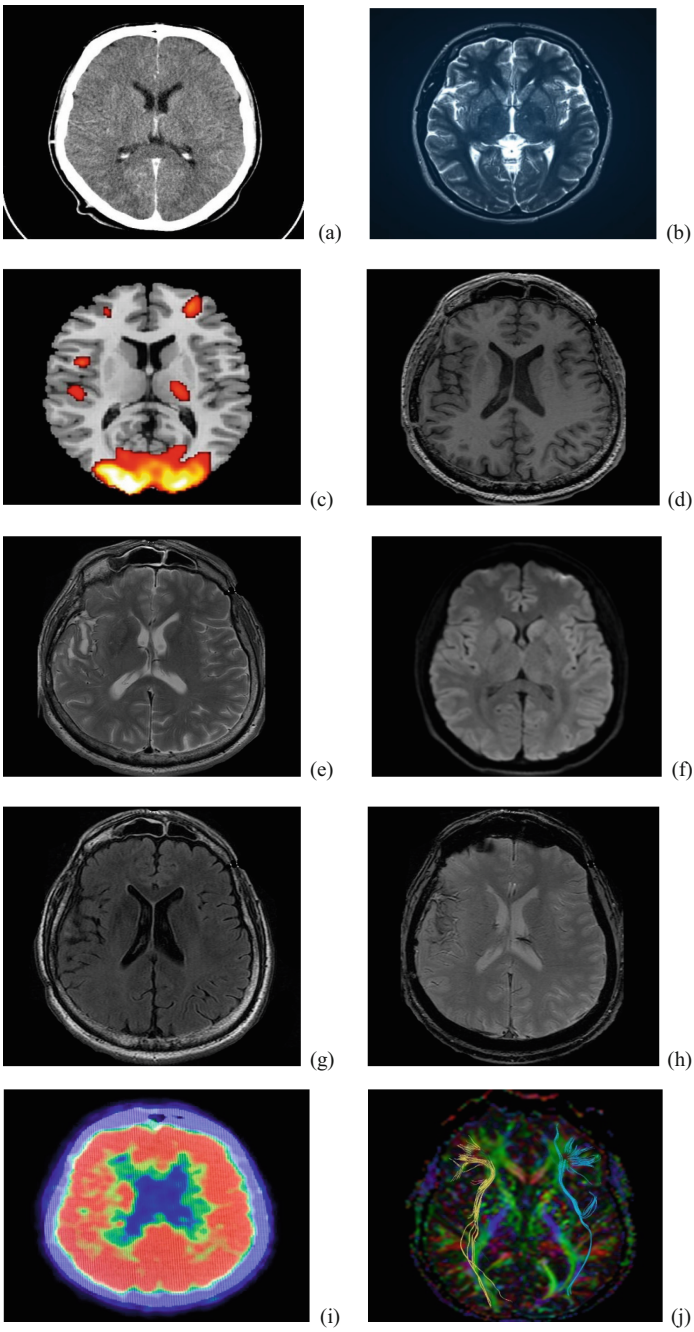


Fig. 1. Different neuro images acquired through different neuroimaging techniques. **a** CT Scan image, **b** MRI image, **c** fMRI image, **d** T1-Weighted MRI image, **e** T2-Weighted MRI image, **f** DWI image, **g** FLAIR image, **h** GRE image, **i** PET image, **j** DTI image.

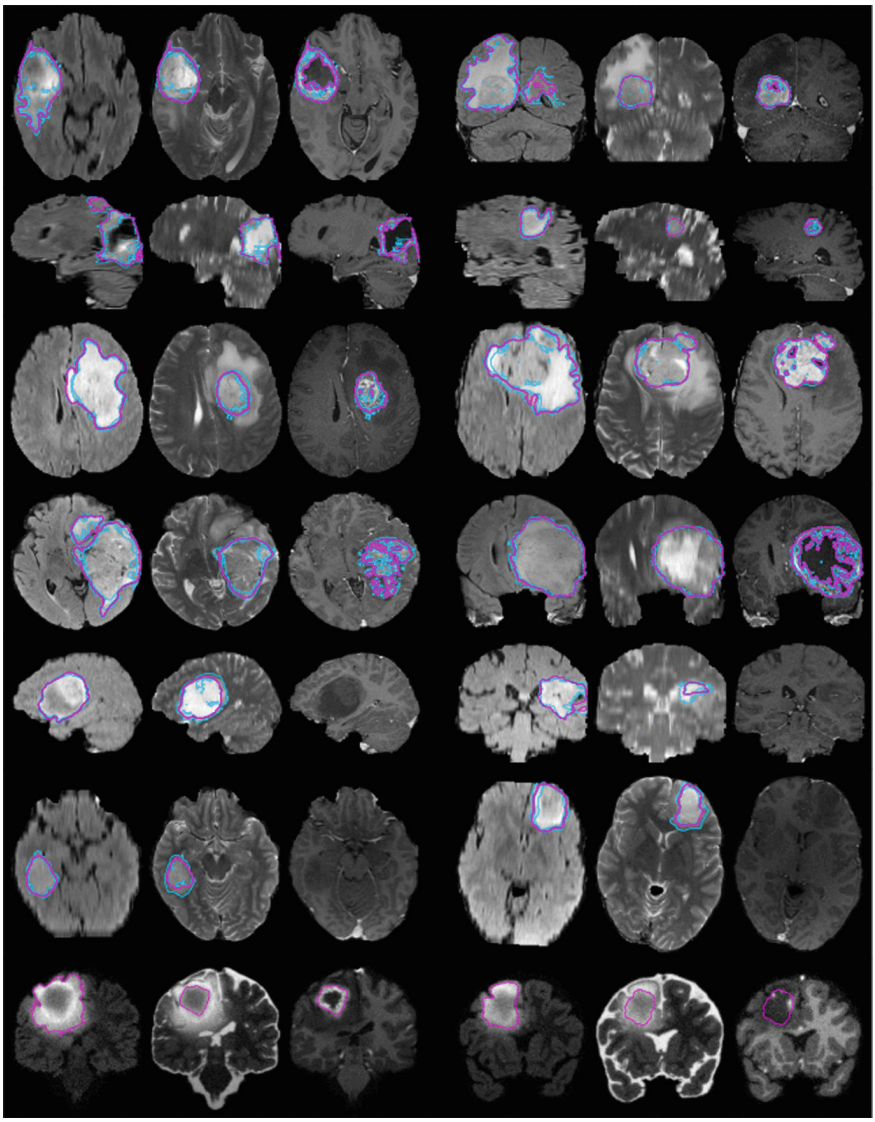


Fig. 2. Different grades of brain tumors at different locations of a brain [43].

[1] Four modalities of MRI (T1, T2, Tc, FLAIR) are used for Brain tumor analysis, which are collected from BRATS-2018 database, to increase the dice coefficient.

[11] Neuro images of Proton Magnetic Resonance Spectroscopy (H-MRS) are used for classification of brain tumor into low grade glioma and high grade glioma.

Ref. [15] presents a review on advanced imaging techniques which shows that advanced modalities of MRI, like PWI, MRS, DWI, CEST are better when compared

to conventional MRI images and also stated that radio-genomics along with ML may improve efficiency.

In Ref. [26], Normal 2D MRI images are considered for brain tumor analysis and used many features extracted like statistical, texture, curvature and fractal features for brain tumor classification, which gives impression to other researchers to select optimal features of different types for improving efficiency in brain tumor analysis.

[30] A framework has been proposed that work on multi-modality of MRI images which are acquired from BRATS-2018 database.

[33] Images from multi-neuroimaging techniques like F-FET, PET, MRI have been used for brain tumor analysis which gives the idea to other researchers that not only ensemble ML or DL methods can be used for enhancing the accuracy of classification and segmentation, fusion of multi-neuroimaging techniques may help in improving efficiency.

[34] F-FET, PET, MRS images are used as dataset for brain tumor prediction.

[38] Neuro images of MRI (BT-Small-2c, BT-large-2c, BT-large-4c) are used as dataset for brain tumor classification. A new framework with ML and CNN is used for extracting the features from the given dataset and then classification is done based on that extracted features.

[39] PET images are attenuated by using both MRI and CT images and then they are considered as PCT images for brain tumor analysis.

Even though there exist many technologies emerged to obtain neuro images, MRI is one of the best and standard technologies, as it doesn't use radiation unlike CT scan. Figure 3 shows this. Graph is drawn based on the papers considered for this review.

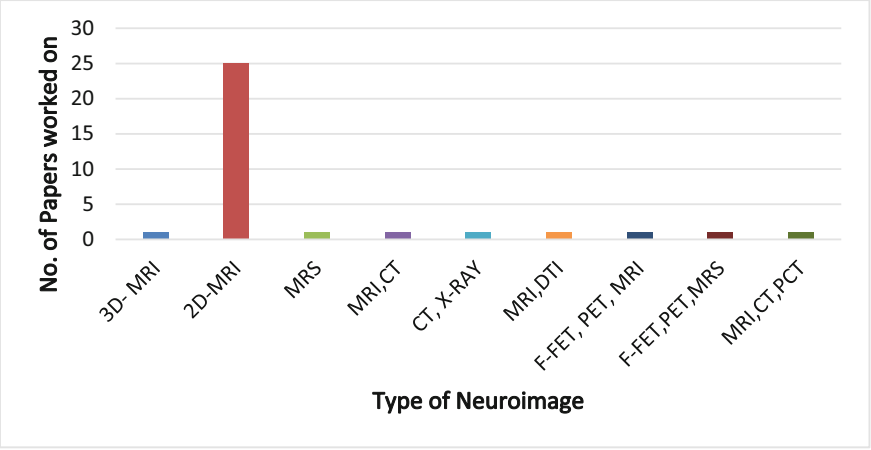


Fig. 3. Shows images acquired from different neuroimaging techniques along with the no. of papers used that images.

2 Significance of ML in Neuroimaging

In processing of neuro images for brain tumor analysis, features selection, segmentation and classification play major roles. Many researchers use different ML algorithms for this purpose to enhance the accuracy of analysis and to reduce time complexity of mathematical calculation.

Ref. [42] presents a survey report on feature selection algorithms and their application in neuro image analysis. The authors of this paper presented that features selection influences the accuracy of brain tumor detection. It also stated that research on different approaches and ML algorithms for feature selection has been started around 1960s and 1970s.

Though the research on neuro image analysis has been started decades ago, still scientists and researchers are working with new approaches and ensemble methods to meet the challenges posed by MICCAI BRATS.

The process of diagnosing brain tumor from neuroimaging may be viewed as in Fig. 4.

Image Enhancement techniques can be used in brain tumor analysis to improve accuracy of edge detection and also for better classification.

Neuro image may have numerous features to be extracted to perform statistical calculations in order to identify normal tumors and abnormal tumors, to detect tumors and to segment different grades of tumors. Features may be structural features, textural features and intensity features. So type of features and number of features using for analysis may highly influence the detection and classification accuracy.

Features may be structural features, textural features and intensity features. So type of features and number of features using for analysis may highly influence the detection and classification accuracy.

Each feature may get its own weightage in analysis. But considering all the features may be the right choice, which leads to concept of features selection and reduction. There are many techniques for features selection like Leave-one-out model, etc., and also features that doesn't hold much weightage in analysis need to be dropped out for neuro image analyzing, which helps the researchers in terms of time complexity.

There exist many Machine Learning algorithms being used in the medical field to help doctors both in analyzing the diseases and predicting the treatments and responses after giving the treatment, like predicting the possibility of reoccurrences of tumors after a surgery or any other kind of treatments. Researchers and scientists have option to select suitable Machine Learning algorithm depending on their requirement.

[5] SVD is used for feature optimization, as feature selection plays a vital role in enhancing the accuracy of detection and classification of brain tumors. The work has been done to improve the performance in terms of computational time. Computational time for training the given data using SVD is 2 min and the performance is compared with DCNN which takes 8 min and RescueNet takes 10 min.

[16] Different ML algorithms are used like LSFHS for minimizing noise in MRI, GBHS for image segmentation, TanH activation function for classification have been used. A large dataset with 25500 images are analyzed using the respective techniques.

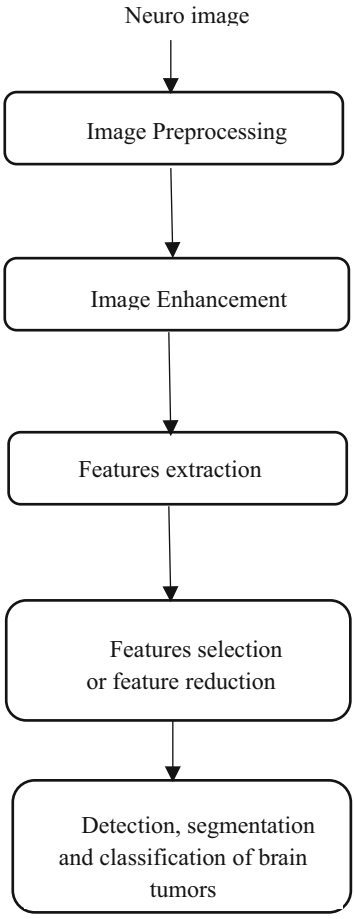


Fig. 4. The process of neuroimaging analysis.

[18] Deep Learning algorithms are enhanced for improving the accuracy of segmentation and classification. Kernel based CNN and M-SVM are used for image enhancement and SGLDM is used for feature extraction from the given MRI data.

[20] Ensemble method (DCNN-F-SVM) is used for brain tumor segmentation. Authors have stated that working with this ensemble method takes high computational time. In order to meet clinical needs, high accuracy of brain tumor analysis (detection, segmentation, classification and prediction) is required but it should be achieved with low cost and low computational time, because delay in analysis may let the patient's life in risk.

[27] Automatic brain tumor detection algorithm is proposed which worked on MRI images for brain tumor detection. Grey level intensity of an MRI images is used for brain tumor position detection.

Table 1 presents overview of papers, considered in this review, with different ML/DL algorithms used for Brain tumor analysis, based on accuracy achieved.

Figure 5 shows that CNN and SVM are mostly being used algorithms. Graph is drawn based on the papers considered for this review.

Acquiring the neuro images for brain tumor analysis is one of the complex tasks. There are many online sources that allows to use their data upon request and registration and sometimes local or clinical data, acquired from local hospitals or from radiology centers, can be used. Some researchers are using synthetic data by applying some data augmentation techniques on the available dataset. Figure 6 shows different dataset sources that are used for brain images acquisition.

[12] Presents a review on BT segmentation. According to this review, less literature has been presented on BT segmentation using BRATS dataset.

[14] CNN is used for feature extraction and classification that carried analysis on multi-modality of BRATS- 2015, 2016 datasets.

[22] They carried a research work on locally acquired data. Cerebellum area has been cropped in BT detection process, because of this they were unable to detect cerebellum tumors, which stands as a limitation of this research.

[28] SVM and DNN algorithms are used for brain tumor classification of MRI images acquired from different sources like Figshare, Brainweb and Radiopaedia. They considered gender and age as additional and significant features for brain tumor classification.

[31] Elastic Regression and PCA methods are used for M-Score detection. A large dataset has been used collected from different sources: 2365 samples are from 15 Glioma datasets like GEO, TCGA, CCGS and so on. 5842 pan-cancer samples are collected for BT analysis.

[32] SPORT algorithm is applied for BT analysis on MRS sequence data acquired through 3T MR Magnetom Prisma Scanner, at University of Freiburg. Acquired images are placed in TCGA, TCIA websites to help other researchers who work on brain tumor analysis.

[35] 3D-Multi-model segmentation algorithm used and RnD has been proposed for feature selection which impacts the efficiency of brain tumor analysis. Normal images are acquired from Medical Segmentation Decathlon and LGG images are acquired from BRATS-2018 dataset.

[36] SSC has been used by introducing some percent of Gaussian noise to the MRI data and experiment results were shown that SCC performs better even with some noise in image when compared to some other ML algorithms. The images acquired from BRATS-2015 database.

Table 1. Presents overview of papers, considered in this review, with different ML/DL algorithms used for Brain tumor analysis, based on accuracy achieved

Author & Reference No	Ml/DL method used	Objective of that method	Accuracy achieved	Limitations
[2]	CNN	Classification & Segmentation	0.971	Small dataset is used
[3]	CNN	Classification	0.973	FCN can be used for classification of brain tumors
[4]	CNN	Segmentation	0.9849 ± 0.0009	Synthetic images are used
[6]	CNN, SVM, RBF	Brain tumor analysis	98.3% on Brainweb data and 98.0% on Figshare data	High computational time
[7]	Fuzzy + BSO	Classification & Segmentation	93.85%	FBSO can be applied for detection
[8]	CNN,SVM	Feature extraction & classification	95%	Better SR can be used
[9]	CNN,SVM, KNN	Feature extraction & prediction	99.70%	Multi-model images can be used
[10]	(VGG19, MobileNetV2) CNN architectures	Feature extraction & prediction	91% (using python) and 97%(using Google Colab)	MRI data can also be used along with CT and X-ray images
[13]	RELM and Hybrid PCA-NGIST	Image enhancement & classification	94.23%	Other classifiers like SVM,RF can be applied
[17]	MSCNN and FSNLM	Classification & Noise removing	91.20%	High computational cost

(*continued*)

Table 1. (*continued*)

Author & Reference No	Ml/DL method used	Objective of that method	Accuracy achieved	Limitations
[19]	SR-FCM-CNN	Detection & Segmentation	98.33%	Performance varies depending on training dataset
[21]	DBFS-EC, CNN	Brain tumor detection & feature extraction	99.56%	Only static features are considered
[23]	SVM	Classification	95.70%	ROI has to be manually selected and also unable to detect LGG
[25]	U-Net, 2D-Mask-R-CNN, 3D-ConvNet, 3D-volumetric CNN	Segmentation, tumor grading and classification	0.963(for tumor grading) & 0.971(for classification)	3D MRI or Multi-neuro images can used
[37]	LCS,DNN, MobileNetV2,M-SVM	Edge detection, feature extraction, feature selection, segmentation and classification	97.47%(on BRATS-2018) & 98.92% (on Figshare data)	Computational time for feature selection is high
[40]	k-Means clustering, FCM, DWT, BPNN	feature extraction and classification	93.28%	Multi-neuro images can used

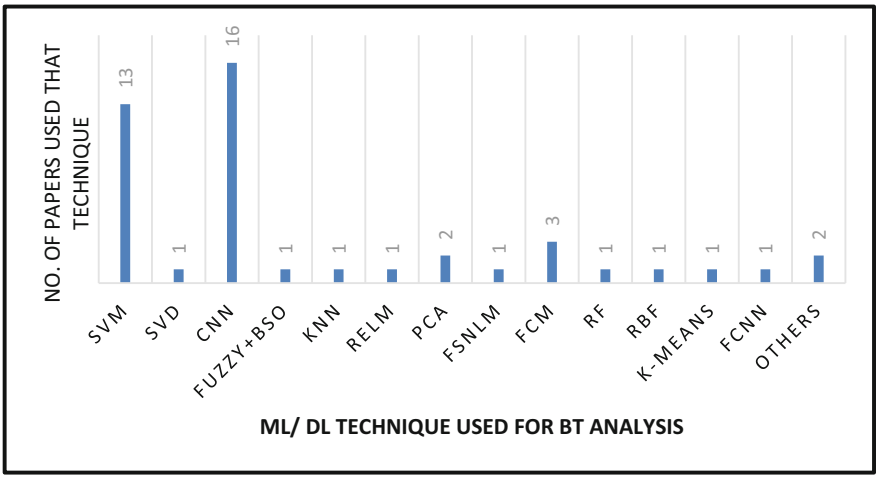


Fig. 5. Different ML/DL algorithms used for Brain tumor analysis.

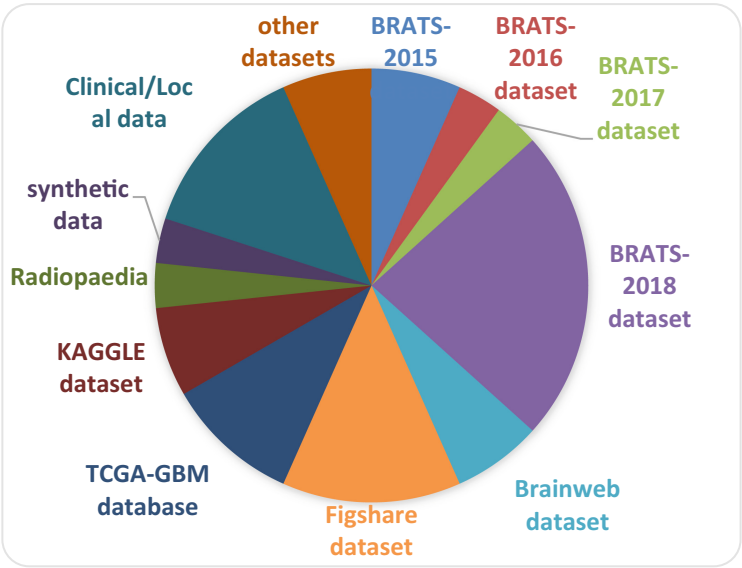


Fig. 6. Different dataset for brain images acquisition.

3 Conclusion

It is observed that most of the research work has been carried out using MRI images (mostly using T1- Weighted, T2- Weighted and FLAIR MRI images), even though there exist some other imaging techniques like PET, MRS, etc., Some researchers used two or more modalities to increase accuracy in detection and segmentation. Neuro images acquired using different imaging techniques like: FET-PET, PET-CT, PET-MRI images

can be combined to improve efficiency. In order to meet clinical needs, high accuracy of brain tumor analysis is required but it should be achieved with low cost and low computational time. As future work, some other Machine Learning algorithms may be ensemble to increase the accuracy and to achieve the challenges posed by BRATS 2021 & BRATS 2022. Researchers may work on Machine learning algorithms ensemble with Deep Learning methods to get better accuracy in less computational time when compared to just ML algorithms.

References

1. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop, pp. 311–320. Springer, Cham. (2018)
2. Özcan, H., Emiroğlu, B. G., Sabuncuoğlu, H., Özdoğan, S., Soyer, A., & Saygı, T.: A comparative study for glioma classification using deep convolutional neural networks (2021)
3. Díaz-Pernas, F. J., Martínez-Zarzuela, M., Antón-Rodríguez, M., & González-Ortega, D.: A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. In Healthcare, Vol. 9, No. 2, p. 153. MDPI. (2021)
4. Islam, K.T., Wijewickrema, S., O'Leary, S.: A deep learning framework for segmenting brain tumors using MRI and synthetically generated CT images. Sensors **22**(2), 523 (2022)
5. Aswani, K., Menaka, D.: A dual autoencoder and singular value decomposition based feature optimization for the segmentation of brain tumor from MRI images. BMC Med. Imaging **21**(1), 1–11 (2021)
6. Haq, E. U., Jianjun, H., Huarong, X., Li, K., & Weng, L.: A Hybrid Approach Based on Deep CNN and Machine Learning Classifiers for the Tumor Segmentation and Classification in Brain MRI. Comput. Math. Methods Med. (2022)
7. Narmatha, C., Eljack, S. M., Tuka, A. A. R. M., Manimurugan, S., & Mustafa, M. A hybrid fuzzy brain-storm optimization algorithm for the classification of brain tumor MRI images. J. Ambient. Intell. Hum.Ized Comput. 1–9 (2020)
8. Sert, E., Özyurt, F., & Doğantekin, A.A.: New approach for brain tumor diagnosis system: single image super resolution based maximum fuzzy entropy segmentation and convolutional neural network. Med. hypotheses **133**, 109413 (2019)
9. Kibriya, H., Amin, R., Alshehri, A. H., Masood, M., Alshamrani, S. S., & Alshehri, A.: A novel and effective brain tumor classification model using deep feature fusion and famous machine learning classifiers. Comput. Intell. Neurosci. (2022)
10. Khan, M. M., Omee, A. S., Tazin, T., Almalki, F. A., Aljohani, M., & Algethami, H.: A novel approach to predict brain cancerous tumor using transfer learning. Comput. Math. Methods Med. (2022)
11. Qi, C., Li, Y., Fan, X., Jiang, Y., Wang, R., Yang, S., Li, S.: A quantitative SVM approach potentially improves the accuracy of magnetic resonance spectroscopy in the preoperative evaluation of the grades of diffuse gliomas. NeuroImage: Clinical **23**, 101835 (2019)
12. Gumaei, A., Hassan, M.M., Hassan, M.R., Alelaiwi, A., Fortino, G.: A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. IEEE Access **7**, 36266–36273 (2019)
13. Hoseini, F., Shahbahrami, A., Bayat, P.: AdaptAhead optimization algorithm for learning deep CNN applied to MRI segmentation. J. Digit. Imaging **32**(1), 105–115 (2019)
14. Overcast, W.B., et al.: Advanced imaging techniques for neuro-oncologic tumor diagnosis, with an emphasis on PET-MRI imaging of malignant brain tumors. Curr. Oncol. Rep. **23**(3), 1–15 (2021). https://doi.org/10.1007/s11912-021-01020-2

15. Kurian, S. M., Juliet, S.: An automatic and intelligent brain tumor detection using Lee sigma filtered histogram segmentation model. Soft Comput. 1–15 (2022)
16. Yazdan, S.A., Ahmad, R., Iqbal, N., Rizwan, A., Khan, A.N., Kim, D.H.: An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD. Tomography **8**(4), 1905–1927 (2022)
17. Thillaikkarasi, R., Saravanan, S.: An enhancement of deep learning algorithm for brain tumor segmentation using kernel based CNN with M-SVM. J. Med. Syst. **43**(4), 1–7 (2019)
18. Özyurt, F., Sert, E., Avcı, D.: An expert system for brain tumor detection: Fuzzy C-means with super resolution and convolutional neural network with extreme learning machine. Med. Hypotheses **134**, 109433 (2020)
19. Wu, W., Li, D., Du, J., Gao, X., Gu, W., Zhao, F., Yan, H.: An intelligent diagnosis method of brain MRI tumor segmentation using deep convolutional neural network and SVM algorithm. Comput. Math. Methods Med. (2020)
20. Zahoor, M.M., et al.: A new deep hybrid boosted and ensemble learning-based brain tumor analysis using MRI. Sensors **22**(7), 2726 (2022)
21. Di Ieva, A., et al.: Application of deep learning for automatic segmentation of brain tumors on magnetic resonance imaging: a heuristic approach in the clinical scenario. Neuroradiology **63**(8), 1253–1262 (2021). https://doi.org/10.1007/s00234-021-02649-3
22. Shrot, S., Salhov, M., Dvorski, N., Konen, E., Averbuch, A., Hoffmann, C.: Application of MR morphologic, diffusion tensor, and perfusion imaging in the classification of brain tumors using machine learning scheme. Neuroradiology **61**(7), 757–765 (2019). https://doi.org/10.1007/s00234-019-02195-z
23. Pflüger, I., Wald, T., Isensee, F., Schell, M., Meredig, H., Schlamp, K., Vollmuth, P.: Automated detection and quantification of brain metastases on clinical MRI data using artificial neural networks. Neuro-oncol. Adv. **4**(1), vdac138 (2022)
24. Zhuge, Y., et al.: Automated glioma grading on conventional MRI images using deep convolutional neural networks. Med. Phys. **47**(7), 3044–3053 (2020)
25. Alam, M. S., Rahman, M. M., Hossain, M. A., Islam, M. K., Ahmed, K. M., Ahmed, K. T., Miah, M. S.: Automatic human brain tumor detection in MRI image using template-based K means and improved fuzzy C means clustering algorithm. Big Data Cogn. Comput. **3**(2), 27 (2019)
26. Wahlang, I., et al.: Brain magnetic resonance imaging classification using deep learning architectures with gender and age. Sensors **22**(5), 1766 (2022)
27. Nadeem, M.W., et al.: Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges. Brain Sci. **10**(2), 118 (2020)
28. Liu, X., Yoo, C., Xing, F., Kuo, C. C. J., El Fakhri, G., Kang, J. W., & Woo, J.: Unsupervised black-box model domain adaptation for brain tumor segmentation. Front. Neurosci. 341 (2022)
29. Zhang, H., Luo, Y. B., Wu, W., Zhang, L., Wang, Z., Dai, Z., Liu, Z.: The molecular feature of macrophages in tumor immune microenvironment of glioma patients. Comput. Struct. Biotechnol. J. **19**, 4603–4618 (2021)
30. Franco, P., Würtemberger, U., Dacca, K., Hübschle, I., Beck, J., Schnell, O., Heiland, D. H.: SPectroscOpic prediction of bRain Tumours (SPORT): study protocol of a prospective imaging trial. BMC Med. Imaging **20**(1), 1–7 (2020)
31. Haubold, J., Demircioglu, A., Gratz, M., Glas, M., Wrede, K., Sure, U., ... & Umutlu, L. Non-invasive tumor decoding and phenotyping of cerebral gliomas utilizing multiparametric 18F-FET PET-MRI and MR Fingerprinting. Eur. J. Nucl. Med. Mol. Imaging 47(6), 1435–1445 (2020)
32. Bumes, E., Wirtz, F. P., Fellner, C., Grosse, J., Hellwig, D., Oefner, P. J., Hutterer, M.: Non-invasive prediction of IDH mutation in patients with glioma WHO II/III/IV based on

F-18-FET PET-guided in vivo 1H-magnetic resonance spectroscopy and machine learning. Cancers **12**(11), 3406 (2020)
33. Wang, L., et al.: Nested dilation networks for brain tumor segmentation based on magnetic resonance imaging. Front. Neurosci. **13**, 285 (2019)
34. Liu, L., Kuang, L., Ji, Y.: Multimodal MRI brain tumor image segmentation using sparse subspace clustering algorithm. Comput. Math. Methods Med. (2020)
35. Maqsood, S., Damaševičius, R., Maskeliūnas, R.: Multi-modal brain tumor detection using deep neural network and multiclass SVM. Medicina **58**(8), 1090 (2022)
36. Kang, J., Ullah, Z., Gwak, J.: Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers. Sensors **21**(6), 2222 (2021)
37. Yang, X., Wang, T., Lei, Y., Higgins, K., Liu, T., Shim, H., Nye, J. A.: MRI-based attenuation correction for brain PET/MRI based on anatomic signature and machine learning. Phys. Med. & Biol. **64**(2), 025001 (2019)
38. Malathi, M., Sinthia, P.: MRI brain tumour segmentation using hybrid clustering and classification by back propagation algorithm. Asian Pac. J. Cancer Prev.: APJCP **19**(11), 3257 (2018)
39. https://www.cancer.net/cancer-types/brain-tumor/introduction
40. https://www.brainline.org/
41. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
42. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis **1**(1–4), 131–156 (1997)
43. Kuraparthi, S., Reddy, M.K., Sujatha, C.N., Valiveti, H., Duggineni, C., Kollati, M., Kora, P., V, S.: Brain tumor classification of MRI images using deep convolutional neural network. Traitement du Signal **38**(4), 1171–1179 (2021)



Visual OutDecK: A Web APP for Supporting Multicriteria Decision Modelling of Outranking Choice Problems

Helder Gomes Costa(✉)

Universidade Federal Fluminense, Niterói, Rua Passos da Pátria, 156, Bloco D 24210-240, RJ, Brazil
heldergc@id.uff.br

Abstract. To choose options or alternatives to compose a subset from a whole set of replacements is still a problem faced by Decision Makers (DM). The Multicriteria Decision Aid/Making (MCDA/M) community has being employing efforts to contribute to solve the problems in this subject. In the MCDA/M field there are two mainstreams of development: Multi Attribute Utility Theory (MaUT) and outranking methods modelling. An unusual difficult in outranking modelling is measure the effects of cut-level parameters and criteria weights on the results. In this article we describe a web app tool to support DM in evaluating how much the results are sensible to these modelling parameters.

Keywords: Decision · Decision analysis · Multicriteria · MCDA · MCDM · Outranking · ELECTRE · Web app

1 Introduction

According to [6], multicriteria decision situations can be categorized into:

- Choice: to chose at least one option from a set of alternatives.
- Ranking: to rank objects from a set.
- Sorting: to sort objects from a set into categories that are ranked.
- Descriptive: to describe a decision situation aiming to support decision making.

This list was extended in [2] that included two other types of problems:

- Clustering: to assign objects into categories that have not preference among then
- Sharing: to distribute or share resources among a set of targets—like it occurs in portfolio problems.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 907–916, 2023.
https://doi.org/10.1007/978-3-031-27409-1_83

Another classification is according the interactions among alternatives intracriterion or even intercriteria. In this case the decision situations can be classified as having a behaviour based either on MultiAttribute Utility Theory (MAUT [5]) or on Outranking principles [7].

Multicriteria decision problems can also be classified according the number of decision units they are designed to address. In this stream they are classified either as mono decisor/evaluator (if one criteria can accept only one evaluation by alternative) or as multiple evaluators/decisors (if the modelling takes into account evaluations from more than one evaluator for each criterion).

In this paper we describe an app designed to deal with multicriteria choice problems based on ELECTRE method: the visual OutDecK (Outranking Decision & Knowledge). Since its simplicity we hope our contribution will be worthy for introducing those from no-coding areas into the outranking decision world.

2 Background

2.1 The Choice Outranking Problem

In the choice problem the Decision Maker (DM) selects a subset composed by one or more s from a set composed by n options, as it appears in Fig. 1.

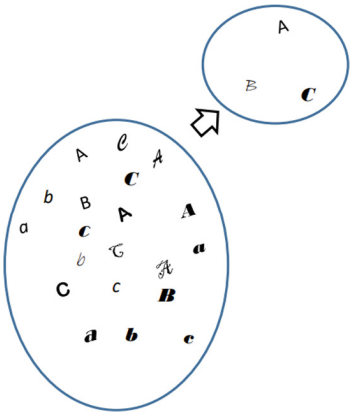


Fig. 1. A general choice problem

It is usual to adopt a ranking algorithm to choose a set composed by the best options instead of selecting a subset that provides the best performance. It is not a problem if the DM is selecting only one alternative from the whole set of options. But, it can be a problem if one is choosing a subset composed by more than one alternative, once the set with the best options should not be the set that provides the best performance, as shown in [2]. According to [2], in outranking methods there is not any interaction among alternatives. Therefore,

the performance's value of an alternative under a criterion perspective can not be added to the performance of another one in such criterion, as it occurs when we apply a MAUT based method.

As an example of this kind of problem, we mention the evaluation of a computer in which the functionality of a microphone can not be substituted by the added of a keyboard. It is a outranking situation.

If we have a set composed by more than one microphone and more than one keyboard, and if the performances of microphones are greater than the performance of the keyboards (values gotten to a unified scale). On a hypothetical situation, one could choose to select two microphones and none keyboards, when using a MAUT based decision algorithm, instead of an outranking one.

Notice that this is a particular situation where the problem is a typical outranking one. There are other situations there are typically additive and in which MAUT is more suitable than outranking. As this paper focuses the outranking choice problems, no example of a MAUT situation is provided here.

2.2 The ELECTRE I

Based on [2,4,7] in outranking modelling one assumes that:

- $A = \{a, b, c, \ldots, m\}$ is a set of alternatives not mutually exclusive, so that one could choose one or more options from A.
- $F = \{c_1, c_2, c_3, \ldots, c_k, \ldots, c_n\}$ is a family or set of n independent criteria that one could take into account while evaluating the alternatives in A.
- $g(a) = \{g_1(a), g_2(a), g_3(a), \ldots, g_k(a), \ldots, g_n(a)\}$ is a vector that records the performance or grade of an alternative a under the set of criteria F.

Based on these assumptions, the following metrics are defined:

- The local concordance degree $(c_j(a, b))$ calculated as it appears e in Eq 1, that means the concordance degree with the assertive that "the performance of alternative a is not worse than the performance of b under the jth criterion. Or, in other words, the agreement degree with the assertive that a is not outranked by b under the jth criterion

$$c(a, b) = \begin{cases} 1, iff\ g_j(a) >= g_j(b) \\ 0, iff\ g_j(a) < g_j(b) \end{cases} \quad (1)$$

The overall concordance degree $C_j(a, b)$ is calculated as it appears in Eq. 2, that means the overall concordanceFigueiredo2022 degree with the assertive that "the performance of a is not worse than b that " a is not outranked by b taking into account all the criteria. In this equation, n is the number o criteria and w^j is the weight or relevance of the jth criterion.

$$C(a, b) = \frac{1}{\sum w_j} \sum w_j * c(a, b) \quad (2)$$

The discordance $D(a, b)$ calculated as it appears e in Eq. 3. It means disagreement degree with the assertive that "the performance of a is not worse

than b" or that "a is not outranked by b taking into account all the criteria in F"

$$D(a,b) = max[\frac{g_j(a) - g_j(b)}{\gamma_{max_j}}] \quad (3)$$

Where,

$$\gamma_{max_j} = max[g_j(a) - g_j(b)] \quad (4)$$

By comparing the values of the metrics calculated by Eqs. 2 and 3, against cut levels cd and dd, respectively, one can build a outranking relation S so that aSb means "a outranks b". It is usual to represent the outranking relations by a graphs as it appears in the Fig. 2.

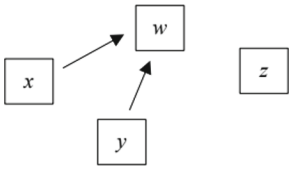


Fig. 2. Example of graph representation of outranking relationships

In this figure, one can observe that xSw and ySw. One also can notice that X, Y and Z are incomparable under the criteria set and other parameters used to evaluate and compare them. The incompatibilities relationships are represented as xRy, xRz and yRz. Once the outranking relations are defined, **A** is partitioned into two subsets **N** and **D**, according the following two rules:

- *Rule 1*: the alternatives in **N** have not outranking relationships among them, at all. In other words they are incomparable under the criteria set and modelling parameters. This subset **N** is named as Kernel or as Non-dominated.
- *Rule 2*: each alternative in **D** is outranked by at least one in **N**. Therefore, this subset **D** is called dominated.

One can conclude that the subset **D** outranks the subset **D** Notice that, this is a conclusion related to subsets, which does neans a relationship among alternatives. In other words, this relationship does not imply in all alternatives in **D** being outranked by all alternatives **N**.

For example, if someone had applied the ELECTRE partitioning taking into account the graph that appears in Fig. 2, it would result in:

- $\mathbf{A} = \{x, y, z, w\}$
- $\mathbf{N} = \{x, y, z\}$
- $\mathbf{D} = \{w\}$.

The solution pointed out by the ELECTRE is to choose the subset $\mathbf{N} = \{x, y, z\}$. One should observe that $\{x, y, z\}S\{w\}$ does not imply in that zSw.

3 The Visual Outdeck Interface

This section describes the visual OutDecK, designed to support DM in using the outranking principles for choosing a set of options that best feet the DM targets, from a whole set of alternatives.

At this time, it is fully supporting ELECTRE I modelling and the true criterion versions of ELECTRE III and PROMETHE [1] . It also make it easier to analyze the results' sensibility to variations in criteria weights and even to the concordance and discordance cutting levels.

3.1 Example's Parameters

This description approaches an example of the selection of a team composed by two collaborators to work in a multidisciplinary project. In this example, the project manager desires the following skills to be shown by the team: Geophysics, Chemistry, Ecology, Computer Sciences, Negotiation, Finances, Transports, Culture, Gastronomy, and, Law.

Table 1 shows the performance of a set of collaborators available under the ten criteria mentioned above. As a constraint, there will be not additive or multiplicative interaction among the members of **A** which means that outranking approach should be used in the modelling.

Table 1. Example data

	Antony	John	Phi	Fontaine	Bill
Geophysics	14	11	2	10	8
Chemistry	14	11	2	10	8
Ecology	14	11	2	10	8
Computer Sciences	14	11	2	10	8
Negotiation	14	11	2	10	5
Finances	14	11	2	10	5
Tranports	14	11	2	10	5
Culture	14	6	2	7	5
Gastronomys	6	2	16	4	5
Law	6	2	16	4	5

3.2 The Initial Screen of OutDecK

Observe in Fig. 3 that the Visual OutDecK is loaded with a sample model, which title, description, and summary are shown in the top of the right side screen.

If one rolls down the screen through pulling down the bar in the right side of the right screen, he/she can see:

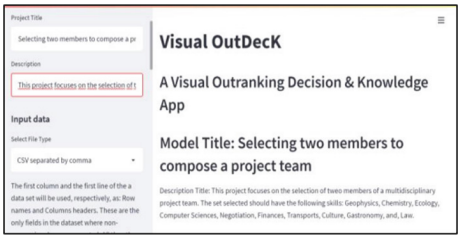


Fig. 3. Initial screen of VisualOutdecK

– A summary of the sample model's data and also the results from the modelling
– The concordance matrix
– The results from applying ELECTRE I: graph and Kernel **N** and dominated **D** sets.

In the left side of the screen, the DM can set or configurates the model by:

– Updating the models' title and description.
– Upload a csv file containing the data to be used in the model.
– Change concordance and discordance cut-levels.
– Change the criteria weight.

Uploading the Dataset After updating the title and description of the model, the user must input the model's data, by importing the dataset file. As it appears in Fig. 5, the user may first specify the file format, choosing one from the following: CSV separated by comma, CSV separated by semicolon or Excel xlsx. After it, he/she can drag the file into the field that appears at the left bottom of Fig. 4. Or, alternatively, the user can perform a search through browsing the files.do a browse the file.

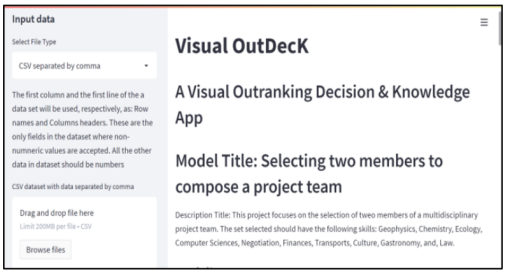


Fig. 4. Loading the dataset

Notice that the he first column and the first line of the data set will be used, respectively, as: Row names and Columns headers. These are the only fields in

the dataset where non-numeric values are accepted. All the other data in dataset should be numbers. The user may be careful to avoid "blank spaces" from the data set before inputting uploading it, otherwise, if there is any blank space in the data, it will cause an error. It is a frequent error, mainly when using data in excel format in which is more difficult to differ cells filled with a blank space from cells there are not active in the sheet. Or, alternatively, the user can perform a search through browsing the files.

Viewing the Summary of the Model Just after uploading the dataset file, the right side of the screen reacts and updates the summary of the model, as it appears in Fig. 5.

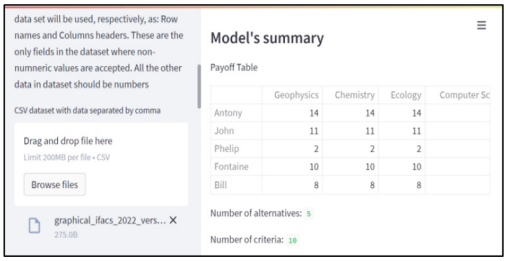


Fig. 5. Summary of the model

Viewing the Results The right side of the screen changes just after the upload ends, showing the values of the concordance matrix, outranking graph and the partition composed by the subsets N and D, as it is shown in Figs. 6 and 7. These results means that the best subset composed by two collaborators is $S = N = \{Antony, Phil\}$, which provides the best performance $g(N) = \{14, 14, 14, 14, 14, 14, 14, 14, 16, 16\}$ along the whole set of criteria.

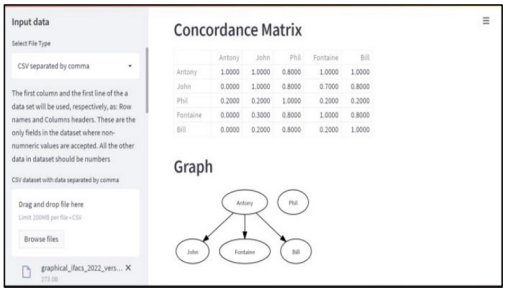


Fig. 6. Concordance matrix and outranking graph

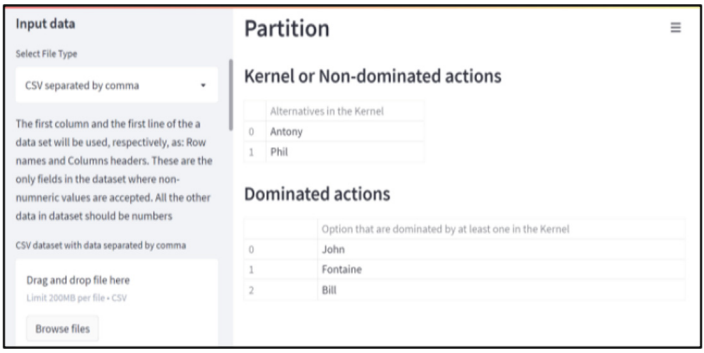


Fig. 7. Partition composed by the subsets **N** and **C**

Observe that $S = \{Antony, Phil\}$ is the choice selected, despite John has an overall performance greater than Phil (John is in the second position, while Phil has the worst overall performance). This is because the performance of $g(\{Antony, John\}) = \{14, 14, 14, 14, 14, 14, 14, 14, 6, 6\}$. In other words, Phil better complements the other options even is not a good option alone.

Sensibility Analysis The OutDecK web based app allows sensibility analysis in an easy way through out a set of sliders that appears at the bottom of the left side of the app screen. As one can see in Fig. 8, one can change the concordance and discordance cult-levels (cd and dd , respectively), and, criteria weights.

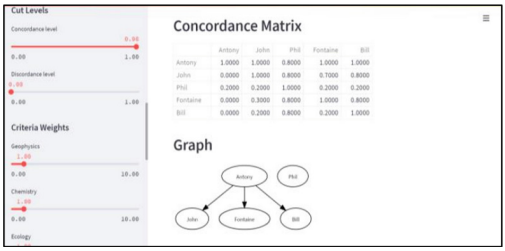


Fig. 8. Sliders to facilitate sensibility analysis

For example, taking a look in the graph that appears in Fig. 8, one can conclude that, as a matter of the facts, the performance of Antony outranks or covers the performance of John, Fontaine and Bill. But, one can also observe that there is not a complete agreement that Antony performance covers Phil performance, or vice-versa. So it, for any reason it is necessary to contract only one alternative or collaborator, how to choose the more suitable option. Well, looking for the concordance matrix that appears in Fig. 8, one should conclude that no changes inn the outranking occurs for values of $c0.8$.

Based on this conclusion, one could change the value of concordance cut-level to 0.8 which would cause a just in time change in the graph, in the Kernel (**N**) and in the dominated subset (**D**)—as one can see in Figs. 9. As can be concluded, if a relaxation in the level of concordance exigence is done, the option Antony can be contracted.

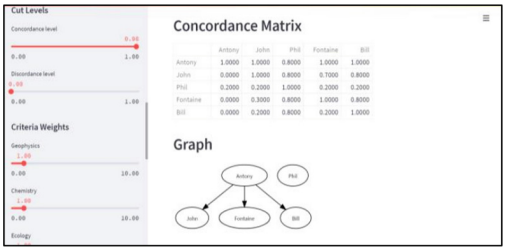


Fig. 9. Sliders to facilitate sensibility analysis

4 Conclusion

The OutDecK app fulfills a relevant lack in MCDA modelling, by providing support to DM in facing two issues they usually have not support: to evaluate the influence of weights, and, concordance and discordance cut-levels, on the modelling results. This is done through an easy-to-apply visual sensibility analysis supported by the OutDecK app.

The user an also easily vary the value of the weights of the criteria by moving the slider tat appear in the left side of the screen. If reader wants more information about weights assignment, we suggest to read the classical [5] that discuss the weights as constant of scales (that converts scales using different metrics), [8] that performed an interesting survey to elicit weights, and, [3] which provides a deep and recent review in methods already used for criteria weighting in multicriteria decision environments. AS further works we suggest to improve the app by including other algorithms and providing comparisons among results from different methods.

Acknowledgments. This study was partialy funded by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPQ)-Grants 314953/2021-3 and 421779/2021-7; and, Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro—Brasil (FAPERJ), Grant 200.974/2022.

References

1. Brans, J.P., Mareschal, B., Vincke, P.: PROMETHEE: a new family of outranking methods in multicriteria analysis, pp. 477–490. North-Holland, Amsterdam, Neth, Washington, DC, USA (1984)
2. Costa, H.G.: Graphical interpretation of outranking principles: avoiding misinterpretation results from ELECTRE I. J. Modell. Manag. 11(1), 26–42 (2016). https://doi.org/10.1108/JM2-08-2013-0037
3. da Silva, F.F., Souza, C.L.M., Silva, F.F., Costa, H.G., da Hora, H.R.M., Erthal Junior, M.: Elicitation of criteria weights for multicriteria models: bibliometrics, typologies, characteristics and applications. Brazilian J. Oper. Prod. Manag. **18**(4), 1–28 (2021). https://doi.org/10.14488/BJOPM.2021.014
4. Greco, S., Figueira, J., Ehrgott, M.: Multiple Criteria Decision Analysis: state of Art Surveys, vol. 37. Springer, Cham (2016)
5. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: preferences and Value Tradeoffs, p. 569. Cambridge University Press, London (1993)
6. Roy, B.: The outranking approach and the foundations of ELECTRE methods. Theory Decis. **31**, 49–73 (1991). https://doi.org/10.1007/BF00134132
7. Roy, B.: Classement et choix en presence de points de vue multiples. Revue francaise de matematique et de recherche operationnelle **2**(8), 57–75 (1968). https://doi.org/10.1051/ro/196802v100571
8. de Castro, J.F.T., Costa, H.G., Mexas, M.P., de Campos Lima, C.B., Ribeiro, W.R.: Weight assignment for resource sharing in the context of a project and operation portfolio. Eng. Manag. J. **34**(3), 406–419 (2022). https://doi.org/10.1080/10429247.2021.1940044

Concepts for Energy Management in the Evolution of Smart Grids

Ritu Ritu(✉)

Department of Computer Science & Engineering, APEX Institute of Technology, Chandigarh University, Mohali, Punjab, India
ritu.e12475@cumail.in

Abstract. In the aim of a more sustainable and intelligent growth of Distributed Electric Systems, this paper presents an overview of fundamental power management ideas and technical difficulties for smart grid applications. Several possibilities are outlined, with an emphasis on the potential technical and economic benefits. The paper's third section looks at the fundamental issues of integrating electric cars into smart networks, which is one of the most important impediments to energy management in smart grid growth.

1 Introduction

The use of non-polluting renewables is known as green energy that can produce a limitless and clean supply for our planet's long-term growth. The most broadly utilized Renewable Energy Systems (RES), like breeze, sun powered, hydrogen, biomass, and geothermal advancements, have a fundamental impact in gathering the world's extending energy interest around here. Subsequently, creative and elite execution environmentally friendly power energy advances and procedures are overwhelmingly popular steadily rising in order to reduce Greenhouse Gas emissions and address issues such as Climate change, global warming, and pollution will be addressed in accordance with the International Energy Agency's (IEA) objective of reducing global emissions by 80% by 2050 [1]. Furthermore, to meet the ever-increasing electrical energy demand, RES innovative and flexible economic methods for network energy management must be incorporated.

In order to generate new emerging ideas and studies, many challenges must be addressed in this context for the From production to implementation viewpoints, the development and utilisation of green energy technologies should be optimised.

The purpose of a Smart Grid (SG) is to combine digital technology, distributed energy systems (DES), and information and communication technology for the reduction of energy usage, which improves the existing power grid'sFlexibility, dependability, and safety are all important factors to consider. These aspects increases entire system's efficiency while also benefiting users financially [2]. In terms of Singapore's ICT growth, End users can acquire power price and incentive signals with this technology, allowing them to choose whether to sell or consume of the energy to the grid. This attribute [3] emphasises the fact that end users have become energy generation resources.

As a result, in recent years, the evolution of SG has run against a number of roadblocks [4, 5]. The present smart grids' development procedures and roadblocks will

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 917–928, 2023.
https://doi.org/10.1007/978-3-031-27409-1_84

face in the upcomming years are depicted in Figure 1. Specifically, several factors must be considered, Each of them is connected to a specific area of interest, including such technological elements of the growth of equipment/software machinery, technical concerns of new power system development and planning approaches, or social challenges of enhancing service quality while decreasing customer prices. By using real-time data analysis discovered from the network, ICT technology may also prevent outages. CO_2 emissions are also decreased when end consumers utilise less energy [6].

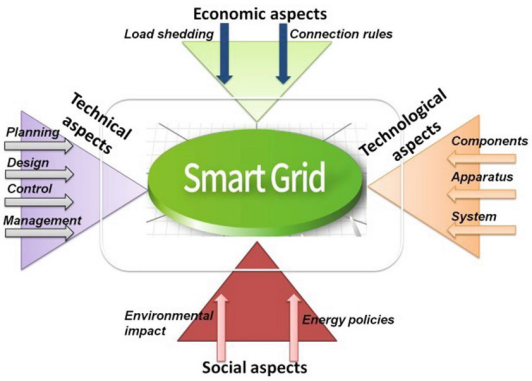


Fig. 1. Challenges for the smart grid evolution process

SGs might possibly operate as systems for regulating the energy use of domestic appliances by continually monitoring the grid frequency [7]. Many residential appliances may be turned off for particular periods of time, reducing peak-hour spikes and reducing congestion difficulties, which may be detected by a drop in frequency. Real-time surveillance systems capable of dynamically answering to the separation of specific appliances as needed can execute this function [8].

On the way to a more sustainable and wiser DES development in this area, this article provides an overview of the fundamental power management concepts and technical obstacles for smart grid applications. The paper is divided into two sections for further information: Section II identifies and investigates potential smart grid evolution scenarios and roadblocks, while Section III focuses on the basic difficulties of the integration of electric vehicles into smart grids.

2 Smart Grids Concepts and Challenges

Traditional electric energy distribution systems and design standards have experienced considerable changes in recent years, mostly as a result of new worldwide directives aimed at reducing pollution and assuring the world's sustainability. As a result, Distributed Generation has become widely adopted in electrical distribution networks, drastically changing the traditional structure and function of these systems.

Indeed, their prior passive job of delivering electrical power from power stations to customers has given way to a more active one that includes Demand-Side Control

(DSM), energy conservation, load shifting, and other operations. In order to optimise the voltage stability factor and support functions, this smart energy distribution system must be correctly connected with additional automation functions and high-performance ICTs [9, 10].

The major structural distinctions between conventional distribution systems and smart grids are shown in Figure 2. As seen in this diagram, vertical integration with centralised generation has developed into distributed energy resources with cross power flows, enhancing consumer contact and addressing user needs for dependability and economies.

To maximise the benefits afforded by this application, many problems must be solved in this scenario, including as adaptability, control, load peak covers, exploitation of renewable energy, energy loss reduction, safety, and fault prediction.

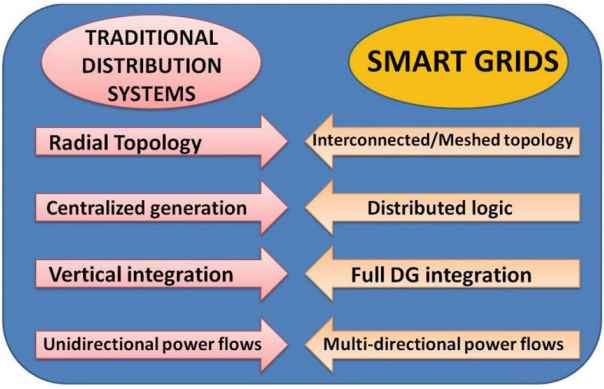


Fig. 2. The smart grid evolution

These objectives can be accomplished by executing progressed robotization frameworks, like Supervisory-Control-And-Data-Analysis, or SCADA [11, 12], which can work on the capacity for the turn of events and execution of sensors, microcontrollers, ICTs, choice frameworks, and controllers for information obtaining and process observing continuously. Thus, this shrewd electric framework will be fit for incorporating the general action got from clients, buyers, and makers for an exceptionally effective, solid, and maintainable energy conveyance, guaranteeing power quality and security of the energy supply, as indicated by the administration orders.

In general, three different grid models may be used to illustrate a Smart Grid, which are described in the following subsections.

A. Active Grids

Active grids are networks that can control and manage the electric energy required by loads, as well as the electric power generated by generators, power flows, and bus voltages.

In active grids, the control approach may be divided into three categories:

- The most basic control level is Level I, which is based on local control at the point of connection for energy generation.
- As shown in Fig. 3, Level II achieves voltage profile optimization and coordinated dispatching. This is accomplished by total control of all dispersed ERs inside the regulated region.
- Level III is the most advanced level of control, and it is implemented using a solid structure based on the connectivity of local regions (cells), as illustrated in Fig. 4. Local areas are in charge of their own management at this level of governance.

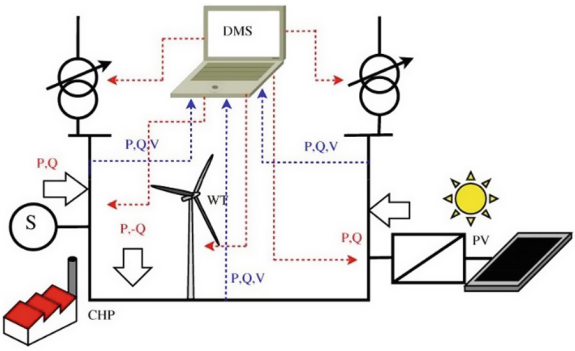


Fig. 3. Principle of the decentralized control

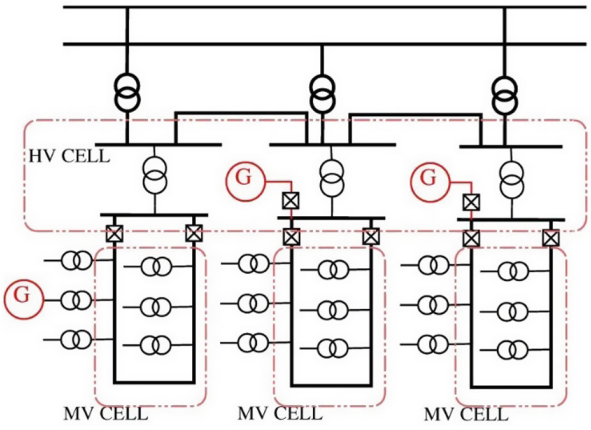


Fig. 4. Cell-organized distribution system

B. Microgrids

Microgrids, as shown in Figure 5, are made up of interconnected generators, energy storage devices, and loads that may function independently of the electric grid and as a regulated and flexible unit.

A microgrid, which may likewise work in an islanding design, is alluded to as a phone of a functioning framework since it is privately overseen by a control framework for every one of the exercises expected for the electric energy moves through generators, burdens, and outer organizations.

The three types of microgrids that may be classified based on their outlet are AC microgrids, DC microgrids, and hybrid microgrids. Combining the first two structures allows for the latter.

For both AC and DC microgrids, controllability is one of the most pressing concerns. A variety of control techniques have been devised in recent literature [13, 14] to address concerns about the unpredictability of microgenerator energy output.

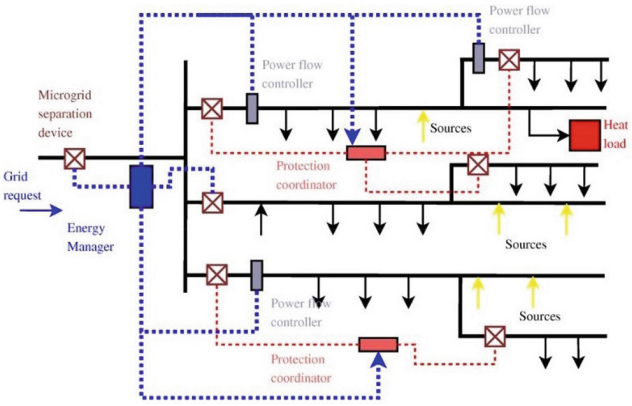


Fig. 5. Microgrid schematic

These techniques are often categorised as follows [15]:

- Concentrated Control (CC), in which terminals are associated with a focal mind, for example, an expert slave control, which might change the chose procedure in light of the working state. This characteristic gives incredible adaptability and constancy, even as the intricacy of these frameworks develops.
- Decentralized Control (DC) is a control technique in which the best control strategy is chosen based on local data, resulting in control ease. As a result, a lack of communication between terminals limits control strategy selection flexibility, potentially resulting in power quality degradation.

Power converters which are connected to the generators and loads can be used to physically regulate microgrids. The following approaches can be used to regulate power converters:

(1) Voltage-Frequency Control (VFC), which is based on maintaining a steady voltage and frequency.

(2) Droop Control, which is based on differential control and is triggered by a change in active or reactive output converter power, resulting in voltage and frequency variations.
(3) Power Quality Control, which comprises maintaining a constant active and reactive output power of the converter.
(4) Consistent Voltage Control: This maintains a constant output voltage from the DC Bus.
(5) Constant Current Control, which aims to maintain a constant output current from the microterminal.

In AC microgrids, the first three procedures appear to be employed, but in DC microgrids, the last two approaches appear to be used.

C. Power Plants on the Cloud

A Virtual Power Plant (VPP), otherwise called a Virtual Utility (VU), is a stage that considers the ideal and compelling administration of a heterogeneous arrangement of conveyed energy assets (either dispatchable or non-dispatchable), otherwise called DER, by planning all frameworks, including dispersed generators, stockpiling, and burden, to partake in the power market. In order to increase power generation, VPPs do actually process signals from the electrical market [16–18].

The main difference between a microgrid and a VPP is that the former is designed to optimise and balance power systems, whilst the latter is designed to optimise for the applicability and demands of the electrical market [19].

As depicted in Fig. 6, which depicts a schematic example of the VPP working concept, the virtual utility is essentially made up of the following components:

(1) Wind turbines, solar plants, and biomass generators are examples of distributed generators and hydroelectric generators that minimise transmission network losses by being closer to clients and thus improve electric energy transmission power quality.
(2) Batteries and super capacitors, for example, store the energy generated by distributed generators, allowing demand and supply to be adjusted while eliminating the eccentricity associated with energy generated by PVs and wind turbines.
(3) Data and Communication Technologies, which coordinate all parts of a VPP structure and oversee information from capacity frameworks, disseminated generators, and burdens.

Furthermore, Virtual Utilities may be divided into two groups based on their functionality:

- CVPP, or Commercial VPP, is a company that administers the difficulties pertaining to dispersed bilateral contracts Generators and consumption units are two types of units.
- TVPP, or Technical VPP, is responsible for ensuring that the previously described VPP pieces are functioning properly, as well as handling and processing bidirectional

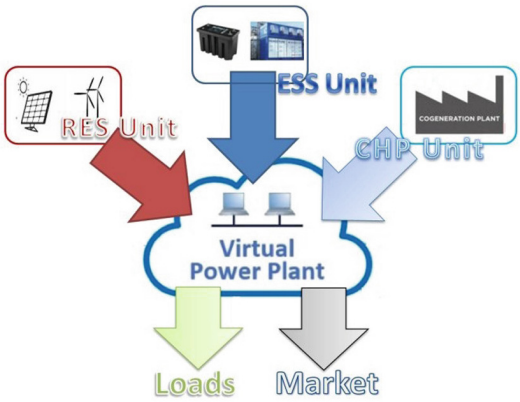


Fig. 6. Schematic principle of the VPP

data from dispersed generators to load units. TVPP also offers a variety of services, including asset monitoring, maintaining proper energy flow, and identifying potential system issues.

In this case, various problems must be overcome in order to provide effective VPP power control for the aforementioned goals. In further detail, the EMS (Energy Management System) is one of the most significant VPP elements, since it collects and analyses data from all VPP elements in order to deliver the best energy management operating plan. As a result, the EMS should provide a reliable and intelligent energy management system capable of efficiently coordinating and regulating VPP activities.

Perhaps the most difficult issue is that the Virtual Utility is delayed to respond to showcase signals, bringing about changing advantages because of market evaluating precariousness. Subsequently, late writing [35–38] proposes different methodologies and strong techniques for creating models for ideal dispatching, including the commitment of both breeze power and electric vehicles at times, and the VPP procedure can be streamlined from either a market or client request reaction point of view in different cases [20].

As a result, new models and trading approaches should be investigated in order to improve VPP performance and maximise the economic gains associated with VPP integration. In terms of sustainable development, successful VPP systems give benefits such as reduced global warming, new business opportunities, reduced economic risk for suppliers and aggregators, enhanced network efficiency and quality factor, and so on. Independent vehicles, these state of the art ideal models will likewise deal with signal handling and human-machine communication advancements [41–53].

The electric automobile is used as a storage system in the Vehicle-to-Grid (V2G) concept, providing energy to the grid while it is not in use by the users during their grid connection. Many challenges remain unsolved in achieving an optimal V2G connection, ranging from technical challenges relating to battery performance and charging time in comparison to traditional ICE vehicles to social acceptance from the V2G innovation's driver, due to the uncertainty of this flow of EV power and grid loading. As a result, one

of the most critical challenges for V2G integration is optimising the electric charging profile. PEVs (Plug-in Vehicles) may, in fact, act as both generators and loads, supplying energy to the grid via on-board charging management systems. In reality, a number of things impact their charging habits, including the charging type (conventional or quick). The problem with the grid connection has exacerbated in the most recent circumstances as power absorption has grown. The charging location and timing must also be considered in order to reduce the likelihood of load peaks.

One method for beating the previously mentioned troubles is to reinforce the EG framework with the goal that it can deal with any future combination of EV frameworks completely. On the opposite side, this element prompts impractical expenses. Another technique is to drive Interest Side Management innovation for charging electric vehicles to satisfy the framework's energy requests, subsequently decentralizing the DG idea in savvy lattices. Moreover, compelling EV intercommunication matched with the reconciliation of SCADA frameworks equipped for checking the vehicle's condition of-charge for brilliant metering could get unquestionable advantages terms of V2G usefulness, as well as composed and shrewd charging methods that keep away from power tops.

3 Smart Grid Integration with Electric Vehicles

It might be argued that transportation electrification is a helpful solution to the global climate change problem since it decreases fossil fuel-related greenhouse gas (GHG) emissions.

EVs, on the other hand, offer a lot of promise for serving the electric grid as a distributed energy source, transferring the energy stored in their batteries to provide auxiliary services like renewable energy integration and peak-shaving power. As a consequence of the rising interest in coordinated charging and discharging of electric automobiles in recent years, the concepts of Vehicle-to-Grid (V2G) and Grid-to-Vehicle (G2V) [21–28] have evolved.

For a total information on the key worries, a multidisciplinary exploration of the specialized, financial, and strategy components of EVs' impact on power frameworks is required.

In such manner, a huge level of ebb and flow research is centered around the genuine mechanical development of electric vehicles concerning fitted sensors, control calculations, and actuators, determined to further develop information examination and the executives for savvy network reconciliation [29–40]. To deliver safe and semi-

4 Outcome

This examination has given a general talk on the standards and challenges for the improvement of savvy frameworks from a specialized, mechanical, social, and financial position. Various basic factors, including the right coordination and equilibrium of V2G and G2V ideas, give off an impression of being basic to the eventual fate of brilliant networks.

References

1. EC Directive, 2010/31/EU of the European Parliament and of the Council of 19May 2010 on the Energy Performance of Buildings (2010)
2. Miceli, R.: Energy management and smart grids. Energies **6**(4), 2262–2290 (2013)
3. Moslehi, K., Kumar, R.: A reliability perspective of the smart grid. In: IEEE transactions on smart grid, vol. 1, no. 1 (2010)
4. Pilz, M., Al-Fagih, L.: Recent advances in local energy trading in the smart grid based on game-theoretic approaches. IEEE Trans. Smart Grid **10**(2), 1363–1371 (2019)
5. Singla, A., Chauhan, S.: A review paper on impact on the decentralization of the smart grid. In: 2018 2nd international conference on inventive systems and control (ICISC), Coimbatore (2018)
6. Tcheou, M.P., et al.: The compression of electric signal waveforms for smart grids: state of the art and future trends. IEEE Trans. Smart Grid **5**(1), 291–302 (2014)
7. Liu, J., Xiao, Y., Gao, J.: Achieving accountability in smart grid. IEEE Syst. J. **8**(2), 493–508 (2014)
8. Zhang, K., et al.: Incentive-driven energy trading in the smart grid. IEEE Access **4**, 1243–1257 (2016)
9. Musleh, S., Yao, G., Muyeen, S. M.: Blockchain applications in smart grid–review and frameworks. In IEEE Access, vol. 7
10. Wang, Y., Chen, Q., Hong, T., Kang, C.: Review of smart meter data analytics: applications, methodologies, and challenges. IEEE Trans. Smart Grid **10**(3), 3125–3148 (2019)
11. Almeida, B., Louro, M., Queiroz, M., Neves, A., Nunes, H.: Improving smart SCADA data analysis with alternative data sources. In: CIRED – Open Access Proceedings Journal, vol. 2017, no. 1
12. Albu, M.M., Sănduleac, M., Stănescu, C.: Syncretic use of smart meters for power quality monitoring in emerging networks. IEEE Trans. Smart Grid **8**(1), 485–492 (2017)
13. Liu, Y., Qu, Z., Xin, H., et al.: Distributed real-time optimal power flow control in smart grid[J]. IEEE Trans. Power Systems (2016)
14. Strasser, T., et al.: A review of architectures and concepts for intelligence in future electric energy systems. IEEE Trans. Industr. Electron. **62**(4), 2424–2438 (2015)
15. Kumar, S., Saket, R. K., Dheer, D. K., Holm-Nielsen, J. B., Sanjeevikumar, P.: Reliability enhancement of electrical power system including impacts of renewable energy sources: a comprehensive review. In: IET Generation, Transmission & Distribution **14**
16. Francés, Asensi, R., García, Ó., Prieto, R., Uceda, J.: Modeling electronic power converters in smart DC microgrids—an overview. In: IEEE Trans. Smart Grid, **9**(6) (2018)
17. Yang, Y., Wei, B., Qin, Z.: Sequence-based differential evolution for solving economic dispatch considering virtual power plant. In: IET Generation, Transmission & Distribution **13**(15) (2019)
18. Wu, H., Liu, X., Ye, B., Xu, B.: Optimal dispatch and bidding strategy of a virtual power plant based on a Stackelberg game. In IET Generation, Transmission & Distribution **14**(4) (2020)
19. Huang, C., Yue, D., Xie, J., et al.: Economic dispatch of power systems with virtual power plant based interval optimization method. CSEE J. Power Energy Syst. **2**(1), 74–80 (2016)
20. Mnatsakanyan, A., Kennedy, S.W.: A novel demand response model with an application for a virtual power plant. IEEE Trans. Smart Grid **6**(1), 230–237 (2015)
21. Vaya, M.G., Andersson, G.: Self scheduling of plug-in electric vehicle aggregator to provide balancing services for wind power. IEEE Trans. Sustain. Energy **7**(2), 1–14 (2016)
22. Shahmohammadi, A., Sioshansi, R., Conejo, A.J., et al.: Market equilibria and interactions between strategic generation, wind, and storage. Appl. Energy **220**(C), 876–892 (2018)

23. Kardakos, E.G., Simoglou, C.K., Bakirtzis, A.G.: Optimal offering strategy of a virtual power plant: a stochastic bi-level approach. IEEE Trans. Smart Grid **7**(2), 794–806 (2016)
24. Viola, F., Romano, P., Miceli, R., Spataro, C., Schettino, G.: Technical and economical evaluation on the use of reconfiguration systems in some EU countries for PV plants. IEEE Trans. Ind. Appl. **53**(2), art. no. 7736973, 1308–1315 (2017)
25. Pellitteri, F., Ala, G., Caruso, M., Ganci, S., Miceli, R.: Physiological compatibility of wireless chargers for electric bicycles. In: 2015 International Conference on Renewable Energy Research and Applications, ICRERA 2015, art. no. 7418629, pp. 1354–1359 (2015)
26. Di Tommaso, A.O., Miceli, R., Galluzzo, G.R., Trapanese, M.: Efficiency maximization of permanent magnet synchronous generators coupled to wind turbines. In: PESC Record – IEEE Annual Power Electronics Specialists Conference, art. no. 4342175, pp. 1267–1272 (2007)
27. Di Dio, V., Cipriani, G., Miceli, R., Rizzo, R.: Design criteria of tubular linear induction motors and generators: A prototype realization and its characterization. In: Leonardo Electronic Journal of Practices and Technologies **12**(23), 19–40 (2013)
28. Cipriani, G., Di Dio, V., La Cascia, D., Miceli, R., Rizzo, R.: A novel approach for parameters determination in four lumped PV parametric model with operative range evaluations. In: Int. Rev. Electr. Eng. **8**(3), 1008–1017 (2013)
29. Di Tommaso, A.O., Genduso, F., Miceli, R., Galluzzo, G.R.: Computer aided optimization via simulation tools of energy generation systems with universal small wind turbines. In: Proceedings - 2012 3rd IEEE International Symposium on Power Electronics for Distributed Generation Systems, PEDG 2012, art. no. 6254059, pp. 570–577 (2012)
30. Di Tommaso, A.O., Genduso, F., Miceli, R.: Analytical investigation and control system set-up of medium scale PV plants for power flow management. Energies **5**(11), 4399–4416 (2012)
31. Di Dio, V., La Cascia, D., Liga, R., Miceli, R.: Integrated mathematical model of proton exchange membrane fuel cell stack (PEMFC) with automotive synchronous electrical power drive. In: Proceedings of the 2008 International Conference on Electrical Machines, ICEM'08 (2008)
32. Di Dio, V., Favuzza, S., La Caseia, D., Miceli, R.: Economical incentives and systems of certification for the production of electrical energy from renewable energy resources. In: 2007 International Conference on Clean Electrical Power, ICCEP '07, art. no. 4272394 (2007)
33. Schettino, G., Benanti, S., Buccella, C., Caruso, M., Castiglia, V., Cecati, C., Di Tommaso, A.O., Miceli, R., Romano, P., Viola, F.: Simulation and experimental validation of multicarrier PWM techniques for three-phase five-level cascaded H-bridge with FPGA controller. Int. J. Renew. Energy Res. **7** (2017)
34. Acciari, G., Caruso, M., Miceli, R., Riggi, L., Romano, P., Schettino, G., Viola, F.: Piezoelectric rainfall energy harvester performance by an advanced arduino-based measuring system. IEEE Trans. Ind. Appl. **54**(1), art. no. 8036268 (2018)
35. Caruso, M., Cecconi, V., Di Tommaso, A.O., Rocha, R.: Sensorless variable speed single-phase induction motor drive system based on direct rotor flux orientation. In: Proceedings – 2012 20th International Conference on Electrical Machines, ICEM 2012 (2012)
36. Imburgia, A., Romano, P., Caruso, M., Viola, F., Miceli, R., Riva Sanseverino, E., Madonia, A., Schettino, G.: Contributed review: Review of thermal methods for space charge measurement. Rev. Sci. Instrum. **87**(11), art. no. 111501 (2016)
37. Busacca, A.C., Rocca, V., Curcio, L., Parisi, A., Cino, A.C., Pernice, R., Ando, A., Adamo, G., Tomasino, A., Palmisano, G., Stivala, S., Caruso, M., Cipriani, G., La Cascia, D., Di Dio, V., Ricco Galluzzo, G., Miceli, R.: Parametrical study of multilayer structures for CIGS solar cells. In: 3rd International Conference on Renewable Energy Research and Applications, ICRERA 2014, art. no. 7016528 (2014)

38. Caruso, M., Cecconi, V., Di Tommaso, A.O., Rocha, R.: Sensorless variable speed single-phase induction motor drive system (2012). In: IEEE International Conference on Industrial Technology, ICIT 2012, Proceedings, art. no. 6210025, pp. 731–736 (2012)
39. Caruso, M., Di Tommaso, A.O., Miceli, R., Ognibene, P., Galluzzo, G.R.: An IPMSM torque/weight and torque/moment of inertia ratio optimization. In: 2014 International Symposium on Power Electronics, Electrical Drives, Automation and Motion, SPEEDAM 2014, art. no. 6871997, pp. 31–36 (2014)
40. Caruso, M., Di Tommaso, A.O., Miceli, R., Galluzzo, G.R., Romano, P., Schettino, G., Viola, F.: Design and experimental characterization of a low-cost, real-time, wireless AC monitoring system based on ATmega 328P-PU microcontroller. In: 2015 AEIT International Annual Conference, AEIT 2015, art. no. 7415267 (2015)
41. Caruso, M., Di Tommaso, A.O., Marignetti, F., Miceli, R., Galluzzo, G.R.: A general mathematical formulation for winding layout arrangement of electrical machines. Energies 11, art. no. 446 (2018)
42. Caruso, M., Di Tommaso, A.O., Imburgia, A., Longo, M., Miceli, R., Romano, P., Salvo, G., Schettino, G., Spataro, C., Viola, F.: Economic evaluation of PV system for EV charging stations: Comparison between matching maximum orientation and storage system employment. In: 2016 IEEE International Conference on Renewable Energy Research and Applications, ICRERA 2016, art. no. 7884519, pp. 1179–1184 (2017)
43. Schettino, G., Buccella, C., Caruso, M., Cecati, C., Castiglia, V., Miceli, R., Viola, F.: Overview and experimental analysis of MCSPWM techniques for single-phase five level cascaded H-bridge FPGA controller-based. In: IECON Proceedings (Industrial Electronics Conference), art. no. 7793351, pp. 4529–4534 (2016)
44. Caruso, M., Di Tommaso, A.O., Genduso, F., Miceli, R., Galluzzo, G.R.: A general mathematical formulation for the determination of differential leakage factors in electrical machines with symmetrical and asymmetrical full or dead-coil multiphase windings. In: IEEE Trans. Ind. Appl. **54**(6), art. no. 8413120 (2018)
45. Caruso, M., Cipriani, G., Di Dio, V., Miceli, R., Nevoloso, C.: Experimental characterization and comparison of TLIM performances with different primary winding connections. Electr. Power Syst. Res. **146**, 198–205 (2017)
46. Caruso, M., Di Tommaso, A.O., Imburgia, A., Longo, M., Miceli, R., Romano, P., Salvo, G., Schettino, G., Spataro, C., Viola, F.: Economic evaluation of PV system for EV charging stations: Comparison between matching maximum orientation and storage system employment. In: 2016 IEEE International Conference on Renewable Energy Research and Applications, ICRERA 2016, art. no. 7884519, pp. 1179–1184 (2017)
47. Schettino, G., Buccella, C., Caruso, M., Cecati, C., Castiglia, V., Miceli, R., Viola, F.: Overview and experimental analysis of MC SPWM techniques for single-phase five level cascaded H-bridge FPGA controller-based. In: IECON Proceedings (Industrial Electronics Conference), art. no. 7793351, pp. 4529–4534 (2016)
48. Viola, F., Romano, P., Miceli, R., Spataro, C., Schettino, G.: Survey on power increase of power by employment of PV reconfigurator. In: 2015 International Conference on Renewable Energy Research and Applications, ICRERA 2015, art. no. 7418689, pp. 1665–1668 (2015)
49. Livreri, P., Caruso, M., Castiglia, V., Pellitteri, F., Schettino, G.: Dynamic reconfiguration of electrical connections for partially shaded PV modules: technical and economical performances of an Arduino-based prototype. Int. J. Renew. Energy Res. **8**(1), 336–344 (2018)
50. Ko, H., Pack, S., Leung, V.C.M.: Mobility-aware vehicle-to-grid control algorithm in microgrids. IEEE Trans. Intell. Transp. Syst. **19**(7), 2165–2174 (2018)
51. Ala, G., Caruso, M., Miceli, R., Pellitteri, F., Schettino, G., Trapanese, M., Viola, F.: Experimental investigation on the performances of a multilevel inverter using a field programmable gate array-based control system. Energies **12**(6), art. no. en12061016 (2019)

52. Di Tommaso, A.O., Livreri, P., Miceli, R., Schettino, G., Viola, F.: A novel method for harmonic mitigation for single-phase five-level cascaded H-Bridge inverter. In: 2018 13th International Conference on Ecological Vehicles and Renewable Energies, EVER 2018, pp. 1–7 (2018)
53. Yilmaz, M., Krein, P. T.: Review of the impact of vehicle-to-grid technologies on distribution systems and utility interfaces. In: IEEE Trans. Power Electron. **28**(12) (2013)

Optimized Load Balancing and Routing Using Machine Learning Approach in Intelligent Transportation Systems: A Survey

M. Saravanan(✉), R. Devipriya, K. Sakthivel, J. G. Sujith, A. Saminathan, and S. Vijesh

Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore 641407, India
sarvan148@yahoo.com

Abstract. Mobile Adhoc Networks (MANET) evolves towards high mobility and provides better provision for connected nodes in Vehicular Adhoc Networks (VANET) and that faces different challenges due to high dynamicity in vehicular environment, which encourages reconsidering of outdated wireless design tools. Several applications like traveler evidence system, traffic management and public transportation systems are supported by intelligent transportation systems (ITS). In order to improve traffic safety, public transportation and compact eco-friendly contamination ITS supports well using smart city urban planning scheme. In this survey we reviewed more number of papers and extracted various insights about the high mobility node and its environment. Parameters like packet delivery ratio, traffic security, traffic density, transmission rate etc. are considered and measured its contribution towards the attainment of parameter in the scale of high, medium and low.

Keywords: Vehicular adhoc network · Mobile adhoc network · Intelligent transportation system · Traffic security and traffic density

1 Introduction

Due to the high dynamics in wireless networks that come from their evolution of high mobility, connected cars are also improved, a number of new issues. Considering established ideas with regard to transportation environments techniques for wireless design. Future smart automobiles, which are becoming crucial components have larger mobility networks. This encourage the use of Machine Learning (ML) to solve the problems that result [1]. Cities, enhancing transportation, transit, and road and traffic safety increased energy efficiency, decreased environmental impact, and increased cost-effectiveness pollution. We investigate the application of machine learning (ML) in this survey, which has recently seen tremendous growth in its ability to support ITS [2]. Having a secure network of communication for not just vehicle to vehicle but also for vehicle to infrastructure, which includes vehicle to vehicle communication, is a critical component of transportation in this day and age. The correspondence with RSU (Road Side Units).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 929–939, 2023.
https://doi.org/10.1007/978-3-031-27409-1_85

The use of machine learning an important part of offering solutions for Secure V2V and Secure V2I connection. This article discusses the fundamentals. Ideas, difficulties, and recent work by researchers in the subject [3].

Vehicle communication is now a common occurrence, which will result in a shortage of spectrum. Communication between vehicles can be effectively solved by using cognitive radio in vehicular communication. For effective use, a robust sensing model is needed. Vehicles therefore sense the spectrum and transmit their sensed data to the eNodeB.In our proposal, a revolutionary clustering method a technique to improve the effectiveness of vehicle communication. The proposed approach used artificial intelligence to create the clusters intelligence. The best possible group of cluster heads is formed using our suggested procedure to obtain the optimum performance [4]. Internet of vehicle (IoV) was first from the union of the Internet of Things (IoT) with vehicular ad hoc networks (VANET), one of the biggest obstacles to the on-road (IoV) implementation is security. Existing security measures fall short of the extremely high (IoV) requirements are dynamic and constantly evolving. The importance of trust in assuring safety, particularly when communicating between vehicles. Among other wireless ad hoc networks, vehicular networks stand out for their distinctive nature [5].

2 Related Works

IOV (Internet of Vehicle) has an advantage of reducing the traffic and improving the high traffic, then enhance the safety for people. The main challenge is to achieve Vehicle to Everything (V2X) which mean fast and efficient communication between different vehicles and smart devices. But it is very hard to maintain the privacy data about the vehicles in IOV system. AI is a smart tool to solve the issue while driving the vehicle [6]. At some time's there may be lost or drop packets while sharing the information. Due to some unusual traffic in the cities by observing the current as well as past behavior of the vehicles in the surrounding environment.The four phases of instruction detection scheme 1.Bayesian learner,2.node's history dadabase,3.Dempster-Shafer adder and rule based security [7].

This project SerIoT has the possible solution and providing the useful and then helpful for reference framework to monitor the real time traffic through IoT heterogeneous platform. The goal is to share the reliable communication safe and secure among the Connected Intelligent Transportation System (C-ITS).According to different scenarios it has been tested during the threats situations. The SerIoT system enhancing the ensuring safety and managing the traffic [8]. A unique intrusion detection system (IDS) utilizing deep neural networks has been developed to improve the security of in-vehicle networks (DNN). The vehicle network packets are used to train and extract the DNN parameters. The vehicles get normal and attack packets from DNN that are class-discriminating. The suggested methods result in deep learning advancements including encouraging and initializing parameters through unsupervised learning of deep neural networks (DBN). As a result, it can improve deduction in the controller area network (CAN) and increase people's safety [9]. In this research, Li Fi technology is used to create a smart vehicular communication system that protects against vehicle crashes on the roads. As with LEDS, the application is inexpensive. Simple and affordable methods for signal processing transmission and generation. a simple transceiver [10].

In order to provide solution for this problem we use Light-Fidelity by which we can transfer huge amount of data at dynamic state at low cost. The vehicular network has tested on different scenario and it have provided better results. And we use machine learning algorithms to provide solution to the road accidents [11]. LiFi is a technology which transfer data or signal from one place to another using light as a medium. The full form of LiFi is Light fidelity. Before the invention of LiFi we use cable communication in which transfer of data is very complicated. For the alternative of cable communication we use visible light communication (VLC). Here we are going to explain the possibility of using LiFi, the possibility of using LiFi over cable communication and explaining the advantages and disadvantages of using LiFi over cable communication in this paper [12]. To complete the desire work or to finish the work between the person communications is the key thing which has to be notified. Over the period of time the communication between the person and the machine has evolved. Consider the machine is very near to us in this case we use switches which is connected to the machine which act as a communication medium. We have modern remote switches in which we can communicate with the remote devices. Here we discuss about the advantages and disadvantages of using LiFi technology in the vehicles to avoid accidents [13].

There are many options to avoid accidents but here we use the direct interaction between the machine (vehicle) and the driver. Over the period of years the communication between the driver and the vehicle has evolved. Each has its own advantages and disadvantages. But the Visible Light Communication (VLC) method has less disadvantage. In the VLC communication we can transfer data at high speed and in this model we have high security. The process of using Visible Light Communication is known as Light Fidelity (Li-Fi). The cost of using this communication is very low compare to the other communication system [14].

3 Research Articles for Study

There are various communication channels between in which the signal can transmit. But the Vehicular Communication System can chose the best path for the security and reliability. In this project we are going to use machine learning algorithm to choose the correct path for the data transmission and security reasons. With the training of Back-Propagation neural Network (BPNN) we can get scenario identification model. With the scenario identification we can improve the communication performance. Hence the model we used here is good and provide good performance. The analysis is done in four areas or done in different places this is because of the reliability [15].

In this paper first we are going to see the outline or frame work to arrange the resources. Second, during the frame work of the algorithms that the author designed. In this part, we are seeing how the process has been made with these limitations, which the author designed. At last, to allocate resource in vehicular networks with the help of machine learning is very challenging. This has been identified in this survey. In this study how the vehicular networks is benefitted by machine learning algorithms have been provided [16]. Intelligent transportation system helps to develop roads and traffic safety. Now a days pollutions are increasing very drastically, to control this environmental pollution this intelligent transportation system can be used, which decreases the pollution

level. For future cities intelligent transportation provides safe traffic system. Intelligent transportation system has lots of quality of services and it will generate high amount of data. In this study we gave a thorough information about the use of Machine Learning (ML) technology in intelligent transportation systems services. It is analyzed by studying the services such as cooperative driving and road hazard warning [2].

The aim of this paper is to discuss about the implementation of THz system and their advantages and the disadvantages and the problems which occurs in the implementation of terahertz is labelled as AI problems. There are many external factor that causes the problem with the AI algorithm we can provide solution to this problem [17]. The collected data from the each node is created as lower copy of comprehension by using the collective design, this copy can be stored easily. In this slowly we can see three section. In first case, the people suffering a lot due to transportation. So, those problems are discussed in this session. In second case, According to transport services the cities solved their problem that solution is taken as survey. At more extreme problem appears on transport system the solutions are in machine learning darning. These are the things discussed in this section. In third case, the survey has been taken on the records for success rate attained for the same researches. In four to six we can see the accuracy of vehicular detection exactly 99% by the experiment has been takes place [18].

The latest research is introducing 6th generation network into the vehicular network with machine learning algorithms to enhance the vehicular application services. To provide solution for the vehicular communication issues we use two algorithms. One is integrated reinforcement and the other one is deep reinforcement algorithms. The device or the vehicle has various use cases so vehicular network has an important research area [19]. Although numerous unmanned aerial vehicle (UAV)assisted routing protocols have been developed for vehicular ad hoc networks, some research have examined load balancing algorithms to support upcoming traffic increase rate and look with complicated dynamic network settings concurrently [20]. Assuming varying quantities of energy our study defines various parameters for transmission in each car. Upper bounds on the distance between two consecutive RSUs for Routing that is approximately load balanced the issue has been identified. 1-D linear network with uniform vehicle distribution road. Simulation simulations demonstrate that the suggested strategy boosts the network performance substantially in terms of energy utilization, Average packet delay and network load [21].

VANETs are direct offshoots of MANETs with special properties such as dynamic changing topology, high speed etc., Because of these distinguishing characteristics, routing in automotive networks been a difficult problem, but aside from effective routing, relatively little notice has been dedicated to load balancing. So, in this research, the focus is on load management in VANET, and a protocol is presented with a new metric which employs interface's queue length. The new protocol is an extension of standard AODV, and it is changed to account for VANET factors [22]. Data distribution utilizing Road Side Units in VANETs is becoming increasingly crucial in order to aid inter-vehicle communication and overcome the distance between two vehicle frequent disconnection issues. In this research, we present a co-operative multiple-Road Side Units model that allows RSUs with large volume workloads to transmit part of their overloaded requests to other Road Side Units with small workloads and situated in same direction as the vehicle runs [23].

In the eyes of researchers, data distribution in VANET has a broad range of vision, assuring its reliability and effectiveness in both V2V and V21 communication models. Our intended research focuses on effective data distribution on V21 communication model. The suggested approach takes into account real-world temporal delays without enforcing delay tolerance. CSIM19 was used to create a real-time simulation environment for this proposal, and the results are summarized [24]. Each vehicle in a VANETs is capable of connecting with adjacent cars and obtaining network information. In VANETs, there are two primary communication models: V2V and V21. Vehicles having wireless transceivers can connect with other vehicles or roadside units. RSUs serving as gateways provide cars with Internet access. Naturally. Vehicles frequently select adjacent RSUs as serving gateways. The first method divides the whole network into sub-regions based on RSU placements. According to simulation findings, the suggested approaches can enhance RSU packet delivery ratio, packet latency, and load balancing [25].

The Cluster-on-Demand VANET clustering (CDVC) method is proposed. Urban cars are distinguished by unpredictability of movement. These problems are addressed by CDVC. The initial state of grouping cars establishes the boundaries of each cluster. Self-Organizing Maps (SOMs) are used in cluster merging to re-cluster clusters based on node similarity, ensuring cluster stability. It eventually leads to load balancing. The information of location and mobility are merged in cluster head selection [26]. However, systems based on AP signal strength neglect the loading conditions of multiple APs and so cannot efficiently utilize the bandwidth. When some APs are overcrowded, the QoS suffers. To address this issue, we may pre-configure the APs and limit their number based on the kind of traffic. We provide a quick method in this paper. Quality Scan is a VANET handoff strategy that reduces handoff latency while also taking into account the loading conditions of regional APs. Our approach collects the loading statuses of the APs on a regular basis and anticipates network traffic for the following instant using the pre-established AP Controller [27].

Various ways have been proposed to increase the efficiency of routing in VANET, but relatively little attention has been dedicated to the issue of load balancing. Load balancing can affect network speed and performance. We suggested a unique load balancing routing scheme in this study. Mechanism based on the Ant colony optimization method which is a meta-heuristic algorithm inspired by ant behavior [28].

The implementation of Mobile Edge Computing (MEC) in automotive networks has shown to be a potential paradigm for improving Vehicular Services by outsourcing computation intensive jobs to the Mobile Edge Computing server. To prevent this, the large idle assets of stop vehicles which are parked may be efficiently used to ease the server's computing strain. Furthermore, unequal load distribution may result in increased delay and energy usage. The multiple parked vehicle-assisted edge computing (MPVEC) paradigm is described for the first time in this study. It is designed to reduce System costs under time constraints [29].

Vehicle-to-Vehicle (V2V) ecosystems that are dynamic. It is impossible to accurately assess V2V channels that change rapidly using the IEEE 802.11p design without a satisfactory figure of pilot carriers in the frequency domain and training symbols in the time domain. Even for larger data packets, the preamble-based station valuation

of IEEE 802.11p cannot guarantee proper equalization in urban and highway contexts. Research has looked into this restriction in various cases works, which indicate that choosing an accurate method of channel update is a significant challenge estimation of packet length that complies with the norm. Regarding bit error rate (BER) and root-mean-square error, the results demonstrate that the suggested system outperforms earlier schemes (RMSE) [30]. The primary benefit of converting to reduce congestion due to 20 MHz channel space will result in Congestion control algorithms may become unnecessary, or even unnecessary. The tutorial sections of the article will go through fundamental OFDM design and describe the stated values of critical V2X channel rules parameters, such as path loss, delay spread, and Doppler spread, and explain the current frequency distribution throughout the US and European continents. The story portions of the study will test the validity of the OFDM design guidelines. Evaluate and measure the effectiveness of computer simulations of 10 MHz and 20 MHz systems. It has been determined that of 20 MHz [31].

This paper includes two equations explaining the link between for each vehicle on the road. Received signal strength versus concurrent transmission, Nodes that are simultaneously transmitting and those that are not density of nodes. The availability of these equations allows for nodes to determine the current node density around them. The solution is designed to perform in the difficult conditions in which nodes lack topological knowledge of the network, and the results demonstrate that the system's accuracy and reliability are enough. As a result of this work can be utilized in a variety of situations. Node density has an impact on the protocol [32]. The maximal likelihood estimator (MLE), which may reach higher estimation accuracy, was introduced for the system for better estimation of accuracy when N is greater than M greater precision than DFE. Its precision will, however, be significantly diminished in the enormous. When $N > M$ in a MIMO-OFDM system, the estimate error increased proportionally due to the growth in NT transmit antennas. In order to address these issues, this study suggests a preamble symbol plus scattered-pilot in the direct time-domain estimator preamp-when $N > M$ for the large MIMO-OFDM system. When compared to MLE, the proposed technique has three significant benefits that stand out: improved estimation accuracy while preserving almost the same computation cost, Bit Error Rate (BER) with simple detection of MIMO data, and higher transmission data rate [33].

In this study, numerous ML-based methods for WSN and VANETs are described along with a brief review of the key ML principles. Then, in connection to ML models and methodologies, diverse algorithms, outstanding topics, challenges of quickly changing networks, and ML-based WSN and VANET applications are examined. In order to give this developing topic more consideration, we have listed a few ML approaches. An overview of the use of ML approaches is provided, along with a breakdown of their intricacies to address any remaining questions and serve as a springboard for more study. With its comparative study, this article offers great coverage of the most cutting-edge ML applications employed in WSNs and VANETs [34]. VANET security, and shifting focus on the underutilization of network capacity in the literature Several Input Several Outputs (MIMO). The analysis has revealed MU.As a superior option than SU-MIMO in commercial and VANETs' safety applications, which double throughput, PDR is greatly increased, and end-to-end delay is decreased to almost half [35].

The regardless of the number of, capacity is constrained by a constant C both the type of routing systems and the nodes. This essay seeks to: CSMA/CA is used to assess a VANET's spatial reuse. Identify the maximum capacity. The suggested model is a lengthy of a standard packing puzzle. We explicitly establish that the maximum intensity of transmitters operating at once (Maximum spatial usage) converging to a constant and suggesting a easy estimation of this constant Practical simulations demonstrate that very close bound is provided by the theoretical capacity. Authentic simulations demonstrate that a very precise bound on the practical capacity is provided by the theoretical capacity [36]. In this study, we assess the effectiveness of RF jamming assaults on 802.11p-based vehicle communications. In particular, we describe the trans-a car-to-car link's mission success rate subject to a constant, RF jamming that is reactive and periodic. First, we carry out in-depth measures in an acoustical environment, where we investigate the advantages of integrated interference mitigation methods. In ad-we note that the recurrent transmission of preamble-Signal jamming can prevent effective communication. Despite being about five times weaker than the interest signal. Finally, we perform outside measurements simulating an automobile platoon and research the dangers that RF jamming presents to this VANET application. We notice that the jammer is reactive, periodic, and constant can obstruct communication across broad propagation areas, which would put traffic safety at risk [37].Ad hoc network directional antennas have greater advantages than conventional universally directed antennas. It is feasible to increase spatial reuse accompanied by directional antennas Wi-Fi channel. An increase in directional antenna gain enables terminals to transmit over longer distances with fewer hops the location. Numerical outcomes demonstrate that our methodology outperforms the current multi-channel protocols in a mobile setting [38].

Some VANET safety applications exchange a lot of data, necessitating a significant amount of network capacity. In this essay, we emphasize applications for enhanced perception maps that incorporate data from nearby and far-off sensors to provide help when driving (Collision avoidance, autonomous driving, etc.) This paper demonstrates using a mathematical exemplary and a great number of simulations showing a considerable increase in network capacity increased [39].

In the above table various MANET and VANET algorithms are compared with its result performance. Table 1 clearly gives the involvement of algorithms and its impact on parameters, how it supports for improving the performance of connected vehicles in mobile environment [1]. Majorly supports for the high mobility and efficiency and [2] mobility, throughput, bandwidth and data transmission rate. Likewise all the referred papers supports in various aspects in different environments. In order to strengthen the untouched parameter we shall refer other algorithms supported by another reference paper and we shall attain the parameters.

Table 1. Comparison of various parameters versus various MANET and VANET algorithms

Paper no	High mobility	Throughput	Packet delivery ratio	Bandwidth	Traffic safety	Energy efficiency	Data transmission rate	Traffic density
[1]	High					High		
[2]	low	High		High			Low	
[3]			High			High		
[4]		Low					High	
[5]	High			High				High
[6]					Low		High	
[7]		High			High			
[8]			Medium					
[9]			Low				Medium	
[10]						High		
[11]	Low					Low		
[12]		Low		High				
[13]			High			High		
[14]					Low		Medium	
[15]	High			High				High
[16]			Low		High			
[2]		High			High			
[17]			Low				Medium	
[18]			High	Medium		High		
[19]	Low							Low
[20]	High			High			High	High
[21]			Low					
[22]		High			High		Low	
[23]			High			High		High
[24]			High			High		
[25]				High				
[26]	High			High				High
[27]					High			
[28]		High	Medium		High		Medium	
[29]				Medium				
[30]			High			High		

(*continued*)

Table 1. (*continued*)

Paper no	High mobility	Throughput	Packet delivery ratio	Bandwidth	Traffic safety	Energy efficiency	Data transmission rate	Traffic density
[31]		High						
[32]	High			High				High
[33]				Low				
[34]		High			High			
[35]						High		
[36]	High			High				High
[37]			High					
[38]		High			High			
[39]				Medium			Low	

4 Conclusion

In this article we have gone through various papers related to vehicular adhoc network and its applications. Various routing algorithms are used in vehicular environment respect to the scenario like high way or smart city communication, it performs well in specific parameters. This survey gives knowledge about various algorithms used in VANET and in which level (like High, Medium or Low) it supports for explicit restriction. Node mobility, Throughput, Packet Delivery Ratio, Bandwidth, Traffic Safety, Energy Efficiency, Data transmission Rate and Traffic Density are the major parameters we concentrated and several algorithms supported for this parameters in various scale and it is observed in Table 1. By this study we identified various research problems and it can be solved in future with appropriate schemes and implementations.

References

1. Liang, L., Ye, H., Li, G.Y.: Toward intelligent vehicular networks: a machine learning framework. IEEE Internet Things J. **6**(1), 124–135 (2018)
2. Yuan, T., da Rocha Neto, W., Rothenberg, C.E., Obraczka, K., Barakat, C., Turletti, T.: Machine learning for next-generation intelligent transportation systems: a survey. Trans. Emerg. Telecommun. Technol. **33**(4), e4427 (2022)
3. Sharma, M., Khanna, H.: Intelligent and secure vehicular network using machine learning. JETIR-Int. J. Emerg. Technol. Innov. Res. (www. jetir. org), ISSN 2349-5162 (2018)
4. Bhatti, D.M.S., Rehman, Y., Rajput, P.S., Ahmed, S., Kumar, P., Kumar, D.: Machine learning based cluster formation in vehicular communication. Telecommun. Syst. **78**(1), 39–47 (2021). https://doi.org/10.1007/s11235-021-00798-7
5. Rehman, A., et al.: Context and machine learning based trust management framework for Internet of vehicles. Comput. Mater. Contin. **68**(3), 4125–4142 (2021)

6. Ali, E.S., Hasan, M.K., Hassan, R., Saeed, R.A., Hassan, M.B., Islam, S., Bevinakoppa, S.: Machine learning technologies for secure vehicular communication in internet of vehicles: recent advances and applications. Secur. Commun. Netw. (2021)
7. Alsarhan, A., Al-Ghuwairi, A.R., Almalkawi, I.T., Alauthman, M., Al-Dubai, A.: Machine learning-driven optimization for intrusion detection in smart vehicular networks. Wireless Pers. Commun. **117**(4), 3129–3152 (2021)
8. Hidalgo, C., Vaca, M., Nowak, M.P., Frölich, P., Reed, M., Al-Naday, M., Tzovaras, D.: Detection, control and mitigation system for secure vehicular communication. Veh. Commun. **34**, 100425 (2022)
9. Kang, M.J., Kang, J.W.: Intrusion detection system using deep neural network for in-vehicle network security. PLoS ONE **11**(6), e0155781 (2016)
10. Bhateley, P., Mohindra, R., Balaji, S.: Smart vehicular communication system using Li Fi technology. In: 2016 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC), pp. 222–226. IEEE (2016)
11. Hernandez-Oregon, G., Rivero-Angeles, M.E., Chimal-Eguía, J.C., Campos-Fentanes, A., Jimenez-Gallardo, J.G., Estevez-Alva, U.O., Menchaca-Mendez, R.: Performance analysis of V2V and V2I LiFi communication systems in traffic lights. Wirel. Commun. Mob. Comput. (2019)
12. George, R., Vaidyanathan, S., Rajput, A.S., Deepa, K.: LiFi for vehicle to vehicle communication–a review. Procedia Comput. Sci. **165**, 25–31 (2019)
13. Mugunthan, S.R.: Concept of Li-Fi on smart communication between vehicles and traffic signals. J.: J. Ubiquitous Comput. Commun. Technol. **2**, 59–69 (2020)
14. Mansingh, P.B., Sekar, G., Titus, T.J.: Vehicle collision avoidance system using Li-Fi (2021)
15. Yang, M., Ai, B., He, R., Shen, C., Wen, M., Huang, C., Zhong, Z.: Machine-learning-based scenario identification using channel characteristics in intelligent vehicular communications. IEEE Trans. Intell. Transp. Syst. **22**(7), 3961–3974 (2020)
16. Nurcahyani, I., Lee, J.W.: Role of machine learning in resource allocation strategy over vehicular networks: a survey. Sensors **21**(19), 6542 (2021)
17. Boulogeorgos, A.A.A., Yaqub, E., di Renzo, M., Alexiou, A., Desai, R., Klinkenberg, R.: Machine learning: a catalyst for THz wireless networks. Front. Commun. Netw. **2**, 704546 (2021)
18. Reid, A.R., Pérez, C.R.C., Rodríguez, D.M.: Inference of vehicular traffic in smart cities using machine learning with the internet of things. Int. J. Interact. Des. Manuf. (IJIDeM) **12**(2), 459–472 (2017). https://doi.org/10.1007/s12008-017-0404-1
19. Mekrache, A., Bradai, A., Moulay, E., Dawaliby, S.: Deep reinforcement learning techniques for vehicular networks: recent advances and future trends towards 6G. Veh. Commun. 100398 (2021)
20. Roh, B.S., Han, M.H., Ham, J.H., Kim, K.I.: Q-LBR: Q-learning based load balancing routing for UAV-assisted VANET. Sensors **20**(19), 5685 (2020)
21. Agarwal, S., Das, A., Das, N.: An efficient approach for load balancing in vehicular ad-hoc networks. In: 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1–6. IEEE (2016)
22. Chauhan, R.K., Dahiya, A.: Performance of new load balancing protocol for VANET using AODV [LBV_AODV]. Int. J. Comput. Appl. **78**(12) (2013)
23. Ali, G.M.N., Chan, E.: Co-operative load balancing in vehicular ad hoc networks (VANETs). Int. J. Wirel. Netw. Broadband Technol. (IJWNBT) **1**(4), 1–21 (2011)
24. Vijayakumar, V., Joseph, K.S.: Adaptive load balancing schema for efficient data dissemination in Vehicular Ad-Hoc Network VANET. Alex. Eng. J. **58**(4), 1157–1166 (2019)
25. Huang, C.F., Jhang, J.H.: Efficient RSU selection approaches for load balancing in vehicular ad hoc networks. Adv. Technol. Innov **5**(1), 56–63 (2020)

26. Zheng, Y., Wu, Y., Xu, Z., Lin, X.: A cluster–on–demand algorithm with load balancing for VANET. In: International Conference on Internet of Vehicles, pp. 120–127. Springer, Cham (2016)
27. Wu, T.Y., Obaidat, M.S., Chan, H.L.: QualityScan scheme for load balancing efficiency in vehicular ad hoc networks (VANETs). J. Syst. Softw. **104**, 60–68 (2015)
28. محمدنیای, رحمانی, مبزور, محمدنیای, بکوری, & مسعود.: A load balancing routing mechanism based on ant colony optimization algorithm for vehicular adhoc network. Int. J. Netw. Comput. Eng. **7**(1), 1–10 (2016)
29. Hu, X., Tang, X., Yu, Y., Qiu, S., Chen, S.: Joint load balancing and offloading optimization in multiple parked vehicle-assisted edge computing. Wirel. Commun. Mob. Comput. (2021)
30. Wang, T., Hussain, A., Cao, Y., Gulomjon, S.: An improved channel estimation technique for IEEE 802.11 p standard in vehicular communications. Sensors **19**(1), 98 (2018)
31. Ström, E.G.: On 20 MHz channel spacing for V2X communication based on 802.11 OFDM. In IECON 2013–39th Annual Conference of the IEEE Industrial Electronics Society, pp. 6891–6896. IEEE (2013)
32. Khomami, G., Veeraraghavan, P., Fontan, F.: Node density estimation in VANETs using received signal power. Radioengineering **24**(2), 489–498 (2015)
33. Mata, T., Boonsrimuang, P.: An effective channel estimation for massive MIMO–OFDM system. Wireless Pers. Commun. **114**(1), 209–226 (2020). https://doi.org/10.1007/s11277-020-07359-2
34. Gillani, M., Niaz, H.A., Tayyab, M.: Role of machine learning in WSN and VANETs. Int. J. Electr. Comput. Eng. Res. **1**(1), 15–20 (2021)
35. Khurana, M., Ramakrishna, C., Panda, S.N.: Capacity enhancement using MU-MIMO in vehicular ad hoc network. Int. J. Appl. Eng. Res. **12**(16), 5872–5883 (2017)
36. Giang, A.T., Busson, A., Gruyer, D., Lambert, A.: A packing model to estimate VANET capacity. In: 2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 1119–1124. IEEE (2012)
37. Punal, O., Pereira, C., Aguiar, A., Gross, J.: Experimental characterization and modeling of RF jamming attacks on VANETs. IEEE Trans. Veh. Technol. **64**(2), 524–540 (2014)
38. Xie, X., Huang, B., Yang, S., Lv, T.: Adaptive multi-channel MAC protocol for dense VANET with directional antennas. In: 2009 6th IEEE Consumer Communications and Networking Conference, pp. 1–5. IEEE (2009)
39. Giang, A.T., Lambert, A., Busson, A., Gruyer, D. Topology control in VANET and capacity estimation. In: 2013 IEEE Vehicular Networking Conference, pp. 135–142. IEEE (2013)

Outlier Detection from Mixed Attribute Space Using Hybrid Model

Lingam Sunitha[1(✉)], M. Bal Raju[2], Shanthi Makka[1], and Shravya Ramasahayam[3]

[1] Department of CSE, Vardhman College of Engineering, Hyderabad, India
sunithavvit@gmail.com
[2] CSE Department, Pallavi Engineering College, Hyderabad, India
[3] Software Development Engineer, Flipkart, Banglore, India

Abstract. Modern times have seen a rise in the amount of study being done on outlier detection(OD). Considering as setting the appropriate parameters for the majority of the existing procedures needed the guidance of a domain expert, and that the methods now in only use handle categorical or numerical data. Therefore, there is a requirement for both generalized algorithms for mixed datasets and current algorithm that can operate without domain interaction (i.e. automatically). The created system can automatically differentiate outliers from inliers in data having only one data type and data with mixed-type properties, such as data with either quantitative or categorical characteristics. The main objective of the described work is to removing outliers automatically. The current study makes use of the hybrid model called Hybrid Inter Quartile Range (HIQR) outlier detection technique.

Keywords: IQR · Outlier detection · Mixed attributes · AOMAD · HIQR

1 Introduction

Recent developments in information technology have revolutionized a numerous industry. Complex Approaches have been put forth to automatically extract useful data from databases and generate new knowledge. Methods for outlier mining are necessary for classifying, examining, and interpreting data Outliers are nearly always present in a practical dataset because of problems with the equipment, processing problems, and a non - representative sample Outliers have the potential to distort summary statistics just like mean and variance Outliers can lead to a poor fit and less accurate predictive model performance in a classification or regression dataset SVM and other similar algorithms are sensitive to outliers present in the training dataset. Most machine learning algorithms may be impacted by training data with outliers. Making a broad model while ignoring extreme findings is the goal the outcomes of classification tasks may be skewed if outliers are included. Accurate classification is essential in real-time scenarios. Many works already in existence do not handle mixed properties. A domain expert is often required to choose the hyper parameters for outlier detection in many previous studies. With little to no user interaction, a mixed attribute dataset is handled in this study endeavor.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 940–947, 2023.
https://doi.org/10.1007/978-3-031-27409-1_86

2 Inter Quarter Range (IQR)

Finding outliers from given dataset is not possible with Gaussian distribution always. Data set not follow any underlying distribution. Another statistical solution which will suitable for any dataset is IQR. The inter quartile range (IQR) is a useful measure for describing a sample of data with a non-Gaussian distribution. The box plot is defined by the IQR, which is determined as the deviation between 25% and 75% percentiles. Keep in mind that by sorting the data and choosing values at particular indices, percentiles can be determined. For an even number of cases, the 50th percentile is the middle value, or the average of the two middle values. The average of the 50th and 51st values would represent the 50th percentile if we had 100 samples. Since the data is separated into four groups by the 25th, 50th, and 75th values, we refer to the percentiles as quartiles (quartile means four). The middle 50% of the data is defined by the IQR. The typical data points would appear in the high probability regions of a stochastic model, whereas outliers would emerge in the low probability parts of a stochastic model, according to statistics-based outlier detection approaches. A difficult problem, outlier detection in mixed-attribute space has just a few proposed solutions. However, the fact that there isn't an automatic technique to officially discern between outliers and inliers makes such existing systems suffer (Fig. 1).

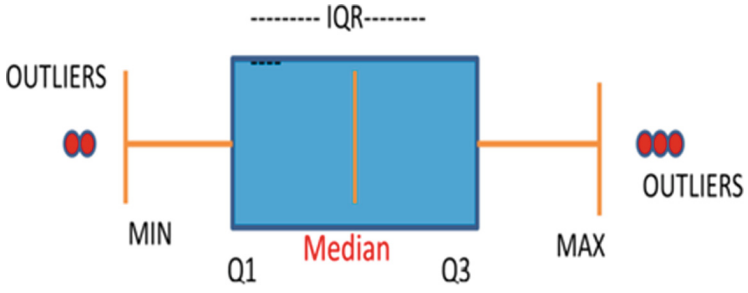


Fig. 1: Box plot

2.1 Related Work

Hawkin's Definition [1]: A sample that considerably deviate from the rest of the observations are considered as an "outlier" and it is generated by a unique mechanism. Mahalanobi's distance is inferior to the model presented by Herdiani et al. [2]. All observations identified as outliers in the original data by the MVV technique however, were also identified in the outlier contaminated data [3]. Using Surface-Mounted Device (SMD) machine sound, evaluate designed model. Yamanishi et al. (2000) [4], The learnt model assigns a score to the item, with a high score suggesting a maximum likelihood of being a statistical outlier. Health insurance pathologic data, it is adaptive to non- stationary sources of data; it is affordable and it can support both categorical and numerical variables. Outlier Detection Algorithm Based On [5] by Liu et al., 2019. Each Gaussian component

now incorporates the three-time standard deviation concept, which reduces accuracy when complicated source data and big data samples are included. Koufakou and Georgiopoulos' (2010) [6] distance-based methodology considers the dataset's sparseness. Accelerated by its distribution. The approach used by Koufakou et al. (2011) [7] estimates an outlier value for every data point using the notion of frequent item set mining. Inliers are points with groups of elements that commonly appear together in data sets. Outliers are rarely occurring to identify outliers in categorical data, Yang et al. (2010) [8]. A survey on various methods to detect outliers from wireless sensor networks, offers a thorough analysis of the current outlier detection methods created especially for wireless sensor networks. Due to the nature of sensor data, as well as particular requirements and restrictions, traditional outlier detection approaches are not directly applicable to wireless sensor networks. The higher a data pattern's outlier score, according to Zhang and Jin (2010) [9], the better it is at describing data objects and capturing relationships between various sorts of attributes. The outlier scores for objects with mixed attribute are then estimated using these patterns. Outliers are defined as the top n points with the highest score values. POD is unable to handle categorical variables directly. Automatic detection of Outliers for mixed attribute data space (AOMAD) by Mohamed Bouguessa (2015) [10] proposed method that will work to mixed-type attribute top 10% objects that are fixed outliers and can automatically distinguish outliers from inliers. Kovács et al. (2019) [11] employed evaluation metrics for time-series datasets to test anomaly detection systems. For anomaly detection, new performance measurements are developed. uCBA [12] is an associative classifier that can categorize both certain and uncertain data. This method, which reshapes the measures support as well as confidence, rule pruning, and classification technique, performs well and acceptable even with unclear data. In [13] Aggarwal discussed database operations like join processing, query, OLAP queries and indexing and mining techniques outlier detection, classification, and clustering. Methodologies to process and in case of uncertain data. Finding ST-Outliers may reveal surprising and fascinating information like local instability and deflections [14]. Here are some instances of such spatial and temporal datasets: meteorological data, traffic data, earth scientific data, and data on disease outbreaks. A data point can be considered an outlier if it does not belong to any of the groupings [15]. A density-based approach for problems involving unsupervised anomaly identification in noisy geographical databases was developed by combining DBSCAN with LOF. Cluster analysis is the basis of a more well presence outlier detection method. Bartosz Krawczyk [16] discussed issues as well as challenges that must be resolved if the field of unbalanced training. The whole range of learning from unbalanced data is covered by a variety of crucial study fields that are listed in this domain. The approach [17] integrates the identification of frequent execution patterns with a cluster-based anomaly detection procedure; in particular, this procedure is well-suited to handle categorical data and is thus interesting by itself, given that outlier detection has primarily been researched on statistical domains inside the literary works. Thudumu et al. [18] is a fundamental research issue with several practical applications, anomaly detection in high-dimensional data is becoming ever more important. Due to so-called "big data," which consists of high-volume, high-velocity data generated by a number of sources, many current abnormal detection mechanisms, however, are unable to maintain acceptable accuracy. Aggarwal CC [19] More applications now have access

to the availability of sensor data as a result of the growing developments in mobile and hardware technology for sensor processing. Classify dimensionality reduction methods and the underlying mathematical intuitions, other surveys, like those in the list, can also be seen by Pathasarathy [20] raise focus on the problems with either high-dimensional data or anomaly detection.

3 Algorithm Hybrid Inter Quartile Range (Hybrid IQR)

Input: Consider Dataset X consists of n' objects and 'm' attributes.
Output: X[target] = O if outlier N if not outlier.

1. Scan the dataset X
2. //Every object in X, find outlier measure TS
3. repeat
4. // Every attribute of the i^{th} object find an outlier score S
5. repeat
6. If (j is numerical attribute) then

$$S(X[i][j]) = (X[i][j] - \mu j)/\sigma_j$$

7. Else if (J is Categorical Attribute) then

$$S(X[i][j]) = P(y[i]/X[i][j])$$

where S(Xi[j]) is an outlier score of i^{th} object and j^{th} attribute of D.

8. End if
9. Until (i = m)
10. TS[i] = S(X
11. Until i = n
12. Use the outlier score (TS) computed to remove Outliers
13. Q3 = 3rdQuartileTS
14. Q1 = 1stQuartileTS
15. IQR = Q3-Q1
16. For i: 1 to n
17. If TS[i] > Q3 or TS[i] < Q1:
18. X [target] = O
19. Else:
20. X[target] = N
21. End if
22. End For

4 Experimental Results

See Figs. 2, 3 and Tables 1, 2 and 3.

Table 1: Dataset description and outliers using Hybrid IQR

Data Set	Number of objects	Numerical Attributes	Categorical Attributes	Number ofoutliers	%outliers
Cylinder Bands	540	20	20	11	2 .03
Credit Approval	690	6	10	35	5.09
German	1000	7	14	5	0.5
Australian	690	6	9	31	4.49
Heart	303	5	9	7	2.31

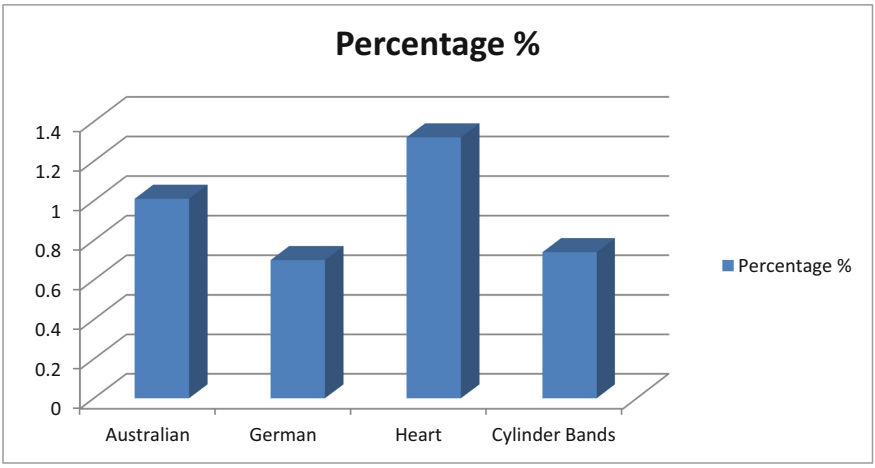


Fig. 2: Bar graph for comparison of percentage of outliers

Table 2: Hybrid IQR algorithm performance measures

Data set	Accuracy	Sensitivity	F1 Score	FPR
Credit Approval	97.20	96.95	97.20	2.55
Australian	98.23	100	98.32	3.66
Heart	98.87	100	98.92	2.32
Cylinder Bands	98.42	100	98.40	3.20
German	99.16	100	99.12	1.59

5 Conclusion

In General outliers are very few in number in every dataset. Achieving accuracy has been difficult because of the rarity of the ground truth in real-world situations. Another challenge is finding outliers dynamic data. There is huge scope for outlier detection, so

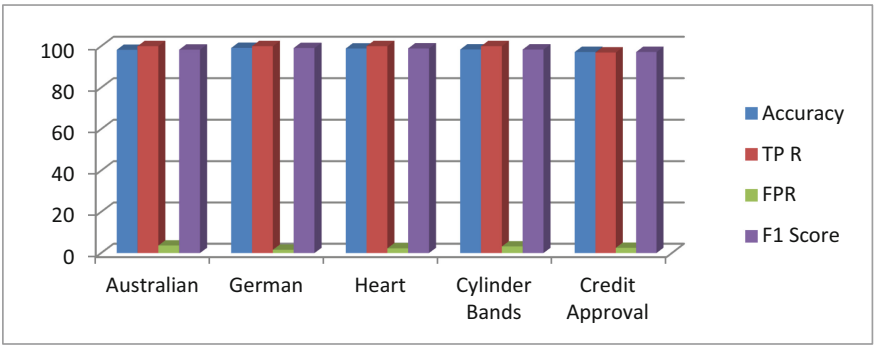


Fig. 3: Bar Graph for Performance measures of HIQR

Table 3: Comparison of HIQR (proposed)and Existing (AOMAD) Algorithms

Data sets	Accuracy		TPR		FPR		F1 Score	
	HIQR	AOMAD	HIQR	AOMAD	HIQR	AOMAD	HIQR	AOMAD
Australian	98.23	98.77	100	98.60	3.66	0.28	98.32	0.972
German	99.16	98.72	100	100	1.59	1.40	99.12	0.934
Heart	98.87	98.74	100	98.46	2.32	1.22	98.92	0.934
Cylinder Bands	98.42	97.60	100	88.80	3.2	1.48	98.4	0. 872
Credit Approval	97.2	93.34	96.95	100	2.55	0.72	97.20	0.964

new models and algorithms are needed to more reliably detect outliers when challenging scenarios, like outlier detection in IoT devices with dynamic sensor data. The cost of using deep learning approaches to address issues with outlier identification is high. Therefore, there is still need for future research on the application of deep learning algorithms for outlier detection methodology. Further research is required to understand how to effectively and appropriately update the current models in order to discover the outlying trends.

6 Future Scope

Learning from unbalanced data is the one key field of research despite the progress over the past 20 years. Identification of outliers falls under imbalanced classification. The issue, which initially arose as a result of outlier detection of binary tasks, has well surpassed this original understanding. We have developed a greater understanding of the nature of imbalanced learning while also facing new obstacles as a result of the development of machine learning and deep learning, as well as the advent of the big data era. Methods at the algorithmic and data-level are constantly being developed,

and proposed schemes are becoming more and more common. Recent developments concentrate on examining not only the disparity across classes but also other challenges posed by the nature of data.

The need for real-time, adaptive, and computationally efficient solutions is driving academics to focus on new problems in the real world. There are two more directions, first one is influence of outliers on classification, next we can focus on performance metrics for outlier's classification.

References

1. Hawkins, D.M.: Identification of Outliers. Springer , vol. 11 (1980)
2. Herdiani, E.T., Sari, P., Sunusi, N.: Detection of outliers in multivariate data using minimum vector variance method. J. Phys.: Conf. Ser. IOP Publ. 1341(9), 1–6
3. Oh, D.Y., Yun, I.D.: Residual error based anomaly detection using auto-encoder in SMD machine sound. Sensors (Basel, Switzerland) **18**(5) (2018)
4. Yamanishi, K., Takeuchi, J., Williams, G. et al.: On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Min. Knowl. Discov. **8**, 275–300 (2004)
5. Liu, W., Cui, D., Peng, Z., Zhong, J.: Outlier detection algorithm based on gaussian mixture model. In: 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2019, pp. 488–492
6. Koufakou, A. · Georgiopoulos, M.: A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes 259–289
7. Koufakou, A., Secretan, J., Georgiopoulos, M.: Non-derivable item sets for fast outlier detection in large high-dimensional categorical data. Knowl. Inf. Syst. **29**, 697–725 (2011)
8. Zhang, Y., Meratnia, N., Havinga, P.: Outlier detection techniques for wireless sensor networks: a survey. IEEE Commun. Surv. Tutor. **12**(2), 159–170
9. Zhang, K., Jin, H.: An effective pattern based outlier detection approach for mixed attribute data. In: Li, J. (ed.) AI 2010: Advances in Artificial Intelligence. AI 2010. Lecture Notes in Computer Science, vol. 6464. Springer (2010)
10. Bouguessa, M.: A practical outlier detection approach for mixed-attribute data. Expert Syst. Appl. **42**(22), 8637–8649 (2015)
11. Kovács, G., Sebestyen, G.: A Hangan Evaluation metrics for anomaly detection algorithms in time-series. Acta Univ. Sapientiae Inform. **11**(2), 113–130 (2019)
12. Qin, X., Zhang, Y., Li, X., Wang, Y.: Associative classifier for uncertain data. In: Proceedings, Web-Age Information Management. Springer, Berlin, pp. 692–703 (2010)
13. Aggarwal, C.C., Yu, P.S.: A survey of uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng. **21**(5), 609–623 (2009)
14. Cheng, T., Li, Z.: A multiscale approach for spatio-temporal outlier detection. Trans. GIS **10**(2), 253–263 (2006)
15. Aggarwal, C.C.: Proximity-based outlier detection. In: Outlier Analysis, New York, NY, USA:Springer Nature, pp. 111–148 (2017)
16. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. **5**(4), 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0
17. An, A., Matwin, S., Raś, Z.W., Ślęzak, D. (eds.): LNCS (LNAI), vol. 4994. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68123-6
18. Thudumu, S., Branch, P., Jin, J. et al.: A comprehensive survey of anomaly detection techniques for high dimensional big data, springer. J. Big Data **7**, 42 (2020)

19. Aggarwal, C.C.: Managing and Mining Sensor Data. Springer Science & Business Media, Berlin (2013)
20. Parthasarathy, S., Ghoting, A., Otey, M.E.: A survey of distributed mining of data streams. In: Data Streams. Springer, pp. 289–307 (2007)

An ERP Implementation Case Study in the South African Retail Sector

Oluwasegun Julius Aroba[1,2(✉)], Kameshni K. Chinsamy[3], and Tsepo G. Makwakwa[3]

[1] ICT and Society Research Group; Information Systems, Durban University of Technology, Durban 4001, South Africa
Oluwaseguna@dut.ac.za

[2] Honorary Research Associate, Department of Operations and Quality Management, Faculty of Management Sciences, Durban University of Technology, Durban 4001, South Africa

[3] Auditing and Taxation; Auditing and Taxation Department, Durban University of Technology, Durban 4001, South Africa

Abstract. The enterprise resource planning (ERP) is an ever-growing software used globally and in all sectors of business to increase productivity and efficiency, however, the south African market does not show any symptoms that it needs such facilities as we tangle the whys and how's on this case study. We use previous studies from the literatures that show an ever-thriving sector such as the South African retail can continue to thrive in the absence of ERP and remain relevant and the biggest market contributors as they have been for the past decades. We focus our sources from year 2020 to 2022 to further influence our case to openly clarify the question of the implementation of ERP system. Our studies settle the unanswered question of the implement ability of an ERP system in the retail sector by exploring both functioning and failed installations and how those were resolved, the effectiveness, efficiency and productivity in the absence and presence of ERP system in place in similar economies such as the South African retail sector, both in the past and present times. The south African retail sector has adopted expensive and difficult to maintain ERP systems, which has a drastic increased improvement in the productivity together with the risks of failure. Such risks were witnessed with Shoprite closing doors in Botswana, Nigeria, and Namibia, this has been proof in failure of expensive and fully paid enterprise resource planning that still failed in more than one country. Our solutions consist of methodology contributed an easy to implement solutions to the retail sectors and can be adapted for different purpose, the integration between large retailers and our system would save millions, time and resources.

Keywords: Enterprise Resource Planning (ERP) implementation · Retail sector · South African market · National GDP · ERP Prototype

1 Introduction

The Enterprise Resource Management is defined as a platform companies use to manage and integrate the essential parts of their businesses; the ERP software applications are

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 948–958, 2023.
https://doi.org/10.1007/978-3-031-27409-1_87

critical to companies because they help them implement resource planning by integrating all the processes needed to run their companies with a single system [1]. Back in the days, organizations would organize their data manually and spend a lot of time searching for what to use when needed [2]. Unlike what the modern world, where everything can be accessed within seconds and be made available for use. The race of global market improvement is endless, with that in mind, we hope to identify new or at least endorse the best use of ERP as the ultimate solution to the South African Retail sector. To answer that question, it is necessary to go through brief history of the ERP and an in dept research of its capabilities in comparison to its rivals and possible best solutions to the retail sector in the retail sector, that will best benefit the market in the recent age. In the late 90's, organizations saw the need to introduce a method that would better assist them integrate their material stock processes without having to literally walk around searching for it, specifically in 1960, a system that would later give birth to ERP was introduced and named Manufacturing Resource Planning (MRP) [3–5]. The change brought about by the MRP gave an idea for the creation of an ERP which would not only manage or help integrate the manufacturing process in the manufacturing sector, but to be as effective for the whole organization.

Firstly, introduced by the Gartner Group, a company founded in 1979 for the purpose of technological research and consulting for the public and the private sector. The group having numerous employees at their headquarters in Stamford, Connecticut, United State, they needed a method to keep their date accessible to the members of the organization and public, at the same time to keep their ongoing research from leaking to the public before it is ready for publishing [15]. Their first public shining of a system that could link the finance department to the manufacturing and to the human resources within itself, came about in 1990 and they named it, an Enterprise Resource Planning because it would save them resources, time, and money, considering how expensive it is to install an effective ERP, this method becomes a question that for ages has been ignored. The system kept on growing and developed by many organizations over the years, and today, organizations all over the world grew dependent of these systems as well as they became more effective because of the ERP systems [6].

1.1 Problem Statement

The Enterprise Resource Planning has become the corporate savior of businesses once established and well maintained, however, many organizations with access to such luxury especially in the retail sector in South Africa have decreased the number employees that would be responsible for date capturing, safe keeping and suppliers, causing an eruption of unemployment and dependency of virtual data than proof and full control of its access, contributing to the recent 44,1% unemployment rate (Quarterly Labour Force Survey, 2022).

Problems caused by Enterprise resource planning software can only be blamed on the system not people, taking away the privilege of accountability and direct implementable solutions. Whatever that is lost due to failure of the system cannot be recovered because no one had the actual data except the system itself. We fully rely on keeping data on the cloud that never fills up in the name of privacy, however, we do not take into consideration the fact that the creators and maintainers of the clouds have access to everything in it,

increasing the risks of corporate espionage and unauthorized data access by bidders of data [7]. The truth of the matter is, the absence of ERP does not completely erase all these risks, but it does leave traces of information and who to hold accountable and recover from loss.

As it turns out, installing an effective and fully functional Enterprise Resource Planning system in an organization, would cost between R2 550 000 to R12 750 000 in sector that probably makes less than a million rands per annum. According to (Peatfield), 2022 ERP report showed that the average budget per user for an ERP project is $9,000 US dollars. When you consider the fact of how many users your system may have especially for larger businesses, and added costs, you'll find an ERP implementation can cost anything between $150,000 US dollars and $750,000 US dollars for a middle-sized business. This emphasizes the fact that apart from increasing unemployment rates, it costs a lot of money to install while it still comes with unavoidable risks, which again uncovers that fact that ERP systems are not as necessary as we deem them to be. Their essentiality comes at cost and are valuable to some extent [8]. As costly as it is, an ERP system saves time, money resources over time to any functional retail sector in any level. Even though there are some factors affecting the costs of an ERP such as size, operations, departments, services provided and the complexity of an organization, nevertheless, the problem lies upon costs and the reality of implementable an ERP in south African retails sector.

The paper arrangement is as follows: section one: Introduction, section two is the literature review while section three is the methodology section and the paper concluded with section four which is the conclusion segment.

2 Literature Review

Almost all major enterprises have adopted one or another enterprise resource planning to boost their business activity. Although, implementing an enterprise resource planning system can be a difficult path as implementing the system takes steps and the cooperations of management to make it work. To implement an Enterprise resource planning system has and will always be a complex process which is one of the challenges the retail sector experienced.it has been very difficult seen from the high rate of failure rates of implementing an enterprise resource planning system. It has been found that at least 65% of Enterprise resource planning system are classified as failures and in turn failure of Enterprise resource planning can lead to the collapse of a business resulting to bankruptcy. According to Babin et al. [9], at least 67% of ERP implementations are not able to meet the customer expectations [9]. The purpose of this paper is to focus on the process of pointing out, arranging, and examining the failure factors of ERP implementation using analytical methods.

In retail-based business, consolidation of several business functions is a necessary condition, [4]. Many retail chains in South Africa have already invested in ERP system to enhance their businesses. Retail chain in South Africa rely on ERP to track the supply chain, financial process, inventory, sales and distribution, overall visibility of the consumer across every channel and take customer centricity to a new level. However, many retailers in South Africa are still using various islands of automation which are

not integrated with each other to manage their core business functions. This strategy can result in somewhat lower levels of effectiveness and efficiency. Implementation of ERP systems is a highly complex process which is influenced not only by technical, but also by many other factors. Hence, to safeguard the success of ERP implementation it becomes imperative for the retailers to get a deeper insight into the factors which influence the ERP implementation.

According to Polka, the efficiency of your ERP will depend on how the end-users adapt and utilise your system. Thus, it's condemnatory to make sure that the users are properly trained to interact with the system without any assistance. This will not only save money and time, but it will improve organisation's processes and ERP solutions need to include consumer-oriented functionality in addition to standard ERP features [10]. Some of these solutions are made specifically for things such as clothing items, food, and cleaning supplies and can supply features that can benefit the company greatly. A resource has been developed to assist buyers to adopt the best ERP solutions for retail to fit the needs of their organisation.

The ERP system will integrate all business functions at the lower cost after the initial installation coats have been covered, covering all the different sectors of the organization. An effective ERP cost's is dependent upon the size for the organization. Similarly, according to Noris SAP (2021), states that the complexity of the organization or business and the degree of its vertical integration has a major influence on the costs and package to be selected when purchasing an ERP package. Seconded by the revenue that the business already generates or plans to bring into the business. The scope of the functions to be covered by the system also has a major influence on the costs, this incudes amongst others if the system will be required to be integrate different business models or one that deals with a single product. Businesses dealing with manufacturing, distribution, sales, and human resources would require a more complex system because it would be integrating several departments into one central source of information where the next department would consult for the next processes [11]. Smaller companies use smaller systems, therefore less costs would be accumulated. The integrated systems would require less resources or at least focus on one a single department such as manufacturing alone that would only communicate information between the supplier of the material and the company responsible for manufacturing. Ian Write on his version of the SAP Comprehensive guide on Write (2020), states that the degree of sophistication and unique requirements in the company's future business process, are there unique customer information requirements or ways that are needed to cut and present information, this is referred to how much of a custom solution is needed [12]. The budget in placed for the system and as well as the hardware that would be installed to get the system operational. Some of the challenges and methods used in proffering solution to ERP challenges are listed in Table 1.

2.1 Research Methodology

In this study Analytical research method is used to understand the ERP implementation in the South African retail sector. Researchers frequently do this type of research to find supporting data that strengthens and authenticates their earlier findings. It is also done to come up with new concepts related to the subject of the investigation.

Table 1: Research Gaps on Enterprise Resource Planning Solutions

Year	Author	Challenges	Method or Systems	Solution
2022	Bill Baumann [13]	Weak Management for projects related to ERP Systems	Panaroma consulting group systems (ERP Problems and solutions to consider before implementation)	Shifting of roles, sharing responsibilities and outsourcing manpower has been the mist effective solutions ever implementable
2020	TEC team [14]	Business philosophy changes	Product Lifecycle management	Using a flexible system, able to be updated and upgraded for the current and present purpose of the organization
2022	TEC team [14]	Overprized expenses in installations of an ERP	ERP Software lifecycle	Make use of an EAM (Enterprise Asset Management) they are cheaper and useful to all organizational sizes

According to Lea et al. [15] For a business owner in South Africa, an ERP system like SAP Business One is the ideal way to improve productivity and manage your company's operations across all functional areas, from accounting and financials, purchasing, inventory, sales, and customer relationships, to reporting and analytics, helping you stay competitive in this economic age. In an enterprise, different business functions are making decisions that have an impact on the organization at any time. ERP enables centralized management of all corporate units and operations. We use historical cost as per figure one and current costs to determine a possible future cost of an ERP, that is, to minimize costs and determine whether an ERP is a solution we seek, or an alternative should be introduced.

Table 2 shows analysis that are used to understand the ERP implementation in the South African retail sector. Researchers frequently do this type of research to find supporting data that strengthens and authenticates their earlier findings. It is also done to come up with new concepts related to the subject of the investigation. According to Niekerk (2021) For a business owner in South Africa, an ERP system like SAP Business One is the ideal way to improve productivity and manage your company's operations across all functional areas, from accounting and financials, purchasing, inventory, sales, and customer relationships, to reporting and analytics, helping you stay competitive in this economic age [14].

Table 2. Likelihood and impacts of the implementation of the ERP system

Problem	In the present (before 2022)	In the present (2022)	Future hypothesis	Solutions
The use of an ERP in the retail industry	26%	53%	The reliance of the systems is growing while there are still questions of who has access to the information stored on the systems outside the organization [13].	The system requires constant update with compatible with the new ERP features [13].
Cause of an employment rate	26.91%	33.9%	The level of dependency on system is rapidly increasing, by year 2050 it is possible that human effort will not be required in the retail sector. As it is, the estimated gross estimate is $117.69 billion [13]	Keeping a team fully involved and updated in every step of the way, allowing them to interact with the system. (Organizational change management) [13]
ERP system failures	50%+	70%+	From the previous results, it is possible that reliable systems will cost more than businesses can make [13].	Continuous establishment and measuring of KPI's to ensure that the system is delivering as expected and the needs are met. Implementing the continuous improvement system [13]

In an enterprise, different business functions are making decisions that have an impact on the organization at any time. ERP enables centralized management of all corporate units and operations. As shown in Fig. 1, it is a typical ERP implementation strategy organized into six phases, each with its own set of objectives shown in the diagram below:



Fig. 1. The 6 basic phases of an ERP implementation plan

Implementing an ERP system offers many benefits for various businesses. It enables departments to operate simultaneously and assists in the storage of data in a single database. The implementation of ERP integrates all the departments, including customer service, human resources, supply chain, accounting, finance, and inventory management, and enables them to collaborate [15].

2.2 SAP ERP Modern Business Process Model

In the above figure, Step 1. An order is placed into the system. The other is not complete or processed or considered a sale until one requirement is met by the customer or client. The second step validates the order by paying, the system notifies the user of the system about the payment and the product is immediately made available for the client. The third step happens simultaneously as the second one, a sale is validated as soon as the payment is received and confirmed for all online sales.

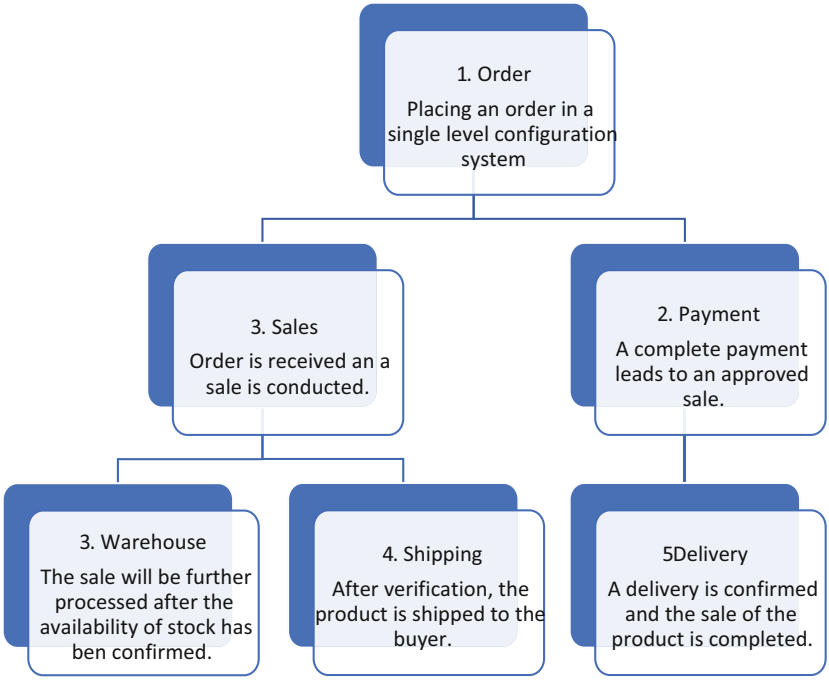


Fig. 2. SAP ERP modern business process model

The warehouse confirms the availability of the ordered items, in case they are available, they are outsourced and made available. As soon the item is available, shipping is arranged to the address stated on the clients' order and the next and final step is processed. A delivery concluded and finalized a successful sale.

2.3 Prototype of an ERP System: Using Java Scrip Web Responses

SA Luxury Clothing Pty Ltd.

Using a java script web responses as an ERP system to process sales and update stock.

An order is placed by a customer, online or offline, the initial process takes the process to the regional server, where all stocks for that region is stored and constantly updated after every sale. The system is used by Facebook market as you mark the number of items you have in stock, after every sale update, the number goes down to indicate the amount of stock left. This happens automatically. Orders conducted offline are quicker and done of the sale point (Till) [16]. The system is connected to the regional data server so that it is updated every time a physical sale is conducted. When an order is done online, the systems check the ordered item on the system, if it is not available on the nearest shop from the order is received, it requests for an order to the next nearer shop. If the item is not available on the whole system, the order is cancelled, and no further process occurs. If the item is found within the system, regardless of the distance from the point of order, the system proceeds with the sale and requests for payment. Depending on the distance

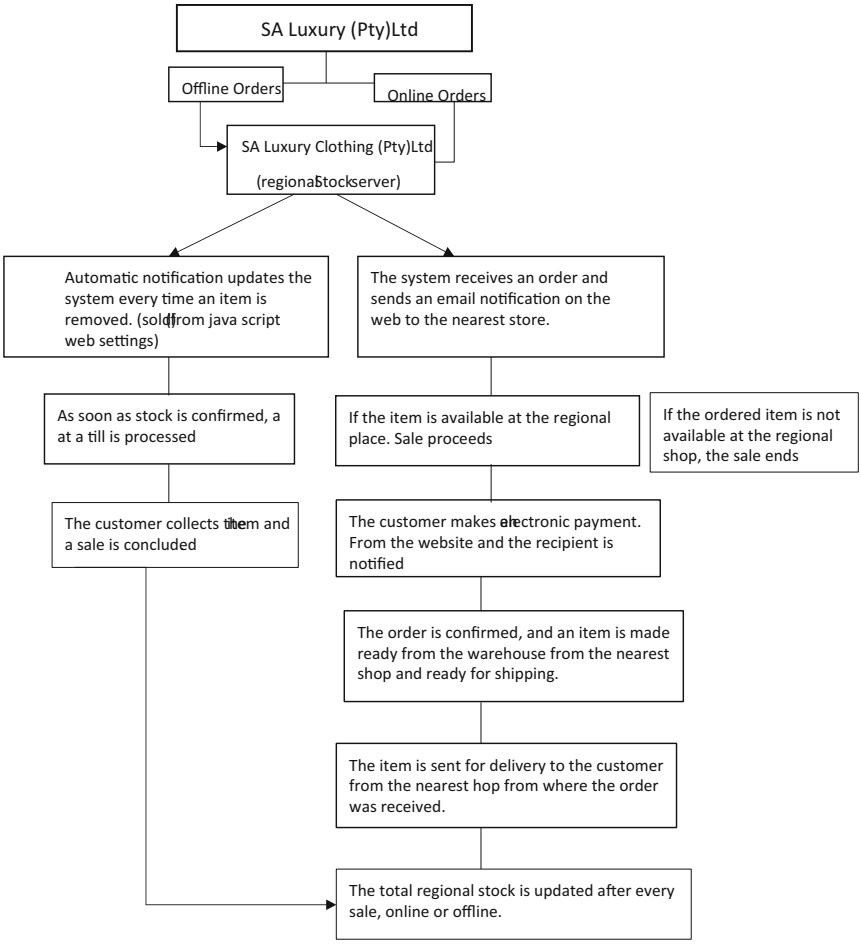


Fig. 3. Prototype of an ERP system

of the available stock from the order, the system will be manually updated for a delivery and the system will update the client/customer of the delivery date and possible time [17–25]. As soon as the item is confirmed, the system will send a notification to the sever as soon the item is sent for shipping, updating the remainder of the stock on the regional server without manual assistance. At the end of the delivery, collection the sale would have been completed and the amount of stock will be updated and ready for the next sale. This is the simplest process which would require a minimum subscription of R250 pm and more depending on the complexity of the system. It would save money and no further installing, or updates are required on the actuals system. The few disadvantages would be vulnerability to hacking which no ERP is full proof from.

3 Conclusion

From solving a problem to establishing a multibillion-dollar enterprise used by many organizations all over the world, the ERP system has proved to be the most effective system for small, medium, and major enterprises. With costs above expectations and being the major costs of unemployment, the above study has proved beyond reasonable doubt that the ERP system has been the reason behind the thriving strategy of the retail sector. The web Java scrip used by take a lot, Facebook market and many other online stores, proves to be the next phase and game changer as indicated on our prototype above. It is efficient, timely and costs close to nothing. Apart from saving time and money, it allows sales to me conducted offline, in store and online simultaneously. Our system would prove to be the next solution to efficiency and cost management problems that the mostly used ERPs are unable to solve and manage. The integration between java script, which is mostly used for free, with an additional financial management software would be the ultimate software solution for the south African retail sector, with less costs, less time management, effective and limitless in logs. Our estimated cost for an advanced java script ran software with background financial management cost nothing more than R10 000 a month, depending on the organization. This makes our prototype suitable for big a small enterprise without financial suffocation.

References

1. Karagiorgos, A.l.: Complexity of costing systems, integrated information technology and retail industry performance. J. Account. Tax. **14**(1), 102–111 (2022)
2. Grandhi, R.B.: The role of IT in automating the business processes in retail sector with reference to enterprise resource planning. Int. J. Bus. Manag. Res. (IJBMR) **9**(2), 190–193 (2021)
3. Subarjah, V.A., Ari Purno, W.: Analysis and design of user interface and user experience of regional tax enterprise resources planning system with design thinking method. Inform: Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi **7**(2), 96–106 (2022)
4. Hove-Sibanda, P., Motshidisi, M., Igwe, P.A.: Supply chain risks, technological and digital challenges facing grocery retailers in South Africa. J. Enterprising Communities: People Places Glob. Econ. **15**(2), 228–245 (2021)
5. Schoeman, F., Seymour, L.F.: Understanding the low adoption of AI in South African medium sized organisations. S. Afr. Inst. Comput. Sci. Inf. Technol. **85**, 257–269 (2022)
6. MunyakaI, J.B., YadavalliII, V.S.S.: Inventory management concepts and implementations: a systematic review. S. Afr. J. Ind. Eng. **33**(2) (2022)
7. Kimani, C.W.: Developing A Multifactor Authentication Prototype for Improved Security Of Enterprise Resource Planning Systems For Kenyan Universities (Published master's thesis). Africa Nazarene University Nairobi, Kenya (2022)
8. Khaleel, H.: ERP Trends: Future of Enterprise Resource Planning. SelectHub (2022). Accessed 30 Sept 2022
9. Babin, R., Li, Y.: Digital Transformation of Grocery Retail: Loblaw (Teaching Case). Available at SSRN 4138488 (2022)
10. Jepma, W.: 14 of the best ERP solutions for retail oriented businesses in 2022. Solut. Rev. (2022). Accessed 1 Jan 2022
11. Teuteberg, S.: Retail Sector Report 2021. Labour Research Services (2021)

12. Mushayi, P., Mayayise, T.: Factors affecting intelligent enterprise resource planning system migrations: the South African customer's perspective In: Yang, X.S., Sherratt, S., Dey, N., Joshi, A. (eds.), Proceedings of Seventh International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol. 447. Springer, Singapore (2022)
13. Bill Baumann: "The panorama approach" The world-leading independent ERP Consultants and Business Transformation. Panorama Consulting Group 2023 (2022)
14. Chethana, S.R.: A study on ERP implementation process, risks and challenges; unpublished master's thesis, Department of Management Studies New Horizon College of Engineering, Outer Ring Road, Marathalli, Bengaluru (2022)
15. Lea, B.R., Gupta, M.C., Yu, W.B.: A prototype multi-agent ERP system: an integrated architecture and a conceptual framework. Technovation **25**(4), 433–441 (2005)
16. Pitso, T.E.: "Exploring the challenges in implementing enterprise resource planning systems in small and medium-sized enterprises" (Unpublished master's thesis). North-West University, Province of North-West (2022)
17. Gartner: "Inc. 2021 Annual Report (Form 10-K)". U.S. Securities and Exchange Commission (2022)
18. Jepma, W.: What is endpoint detection, and how can it help your company? Solut. Rev. (2022)
19. Kimberling, E.: What is SAP S/4HANA? | Introduction to SAP | Overview of SAP ERP. In: Third Stage Consulting Group (2021)
20. Kimberling, E.: "Independent Review of Unit4 ERP Software", Third Stage Consulting Group. (2022)
21. Rankinen, J.: ERP System Implementation. University of Oulu, Faculty of Technology, Mechanical Engineering (2022)
22. Grigoleit, U., Musilhy, K.: RISE with SAP for modular cloud ERP: a new way of working. SAP News Centre (2021)
23. Aroba, O.J., Naicker, N., Adeliyi, T., Ogunsakin, R.E.: Meta-analysis of heuristic approaches for optimizing node localization and energy efficiency in wireless sensor networks. Int. J. Eng. Adv. Tech. (IJEAT) **10**(1), 73–87 (2020)
24. Aroba, O.J., Naicker, N., Adeliyi, T.: An innovative hyperheuristic, Gaussian clustering scheme for energy-efficient optimization in wireless sensor networks. J. Sens. 1–12 (2021)
25. Aroba, O.J., Xulu, T., Msani, N.N., Mohlakoana, T.T., Ndlovu, E.E., Mthethwa, S.M.: The adoption of an intelligent waste collection system in a smart city. In: 2023 Conference on Information Communications Technology and Society (ICTAS), pp. 1–6. IEEE (2023)



Analysis of SARIMA-BiLSTM-BiGRU in Furniture Time Series Forecasting

K. Mouthami[1(✉)], N. Yuvaraj[2], and R. I. Pooja[2]

[1] Department of Artificial Intelligent and DataScience, KPR Institute of Engineering and Technology, Coimbatore, India
mouthamik@gmail.com

[2] Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India

Abstract. Due to the non-stationary nature of furniture sales, forecasting was highly challenging. The cost of maintaining inventory, placing investments at risk, and other expenses could all increase due to unexpected furniture sales in forecasts. To accurately predict furniture sales in the future market, the forecasting framework can extract the core components and patterns within the movements for furniture sales and detect market changes. Exiting ARIMA (Auto-Regressive Integrated Moving Average), LSTM (Long Short-Term Memory) and other algorithm have lesser levels of accuracy. The proposed work employs forecasting techniques such as SARIMA (Seasonal Auto-Regressive Integrated Moving Average), Bi-LSTM (Bidirectional Long Short-Term Memory), and Bi-GRU (Bidirectional gated recurrent unit). This model would estimate and predict the future prices of a furniture stock based on its recent performance and the organization's earnings based on previously stored historical data. The results of the experiments suggest that using multiple models can enhance prediction accuracy greatly. The proposed strategy ensures high consistency regarding positive returns and performance.

Keywords: Sales Prediction · Forecasting · Deep learning · Customized estimation

1 Introduction

Customization of items has become a challenging trend in recent years. Competitive pressure, sophisticated client requirements, and customer expectations trigger additional requirements for manufacturers and products [1]. Forecasting and prediction techniques have advanced substantially in the last ten years, with a constantly increasing trend. The three methods of predicting are machine learning, time series, and deep learning [2]. Deep understanding aims to assess data and classify feature data. In terms of a time series analysis, predicting behavior is a means of determining sales value over a specific time horizon.

A time series is a set of data collected over time and used to track business and economic movements. It helps in understanding current successes and forecasting furniture

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 959–970, 2023.
https://doi.org/10.1007/978-3-031-27409-1_88

sales. Forecasting is a valuable method for planning and managing furniture resources such as stock and inventory. Forecasting demand for specific seasons and periods is known as demand forecasting. As a result, decision support tools are critical for a company to maintain response times and individual orders, as well as appropriate manufacturing techniques, costs, and timeframes [3]. Customers are increasingly seeking out unusual furnishings to make a statement. It has an impact on the price and sales of the particular furniture. Pricing and quality are essential factors in the furniture sales process. The components utilized and the production process's complexity impact the furniture price. The furniture cost is calculated before manufacturing, but the low sales cost affects the profit. The accuracy and assessment of costs can significantly impact a company's earnings. Profits are lowered when costs are cut, and consumers are reduced when costs are increased. Cost estimation is a method of determining furniture price before all stages of the manufacturing process are completed. Data is crucial for a company's success in today's competitive world [4]. For the new customer to become a stronger relationship, it is imperative to understand what data to collect, how to analyze and use it, and how to apply it. The goal of every business, online or off, is to offer services or furniture.

On-time fulfilment of customer expectations for arrival date and other requirements can increase customer happiness, increase competition, streamline production, and aid businesses in making more educated pricing and promotion choices. The forecasting of sales will have an impact on transportation management. E-commerce companies react to consumer and market demands more quickly than traditional retail organizations to obtain a competitive edge. As a result, e-commerce companies must be able to forecast the amount of furniture they will sell in the future. Regression is a common occurrence in machine learning algorithms. The model is iteratively adjusted using a metric of prediction error. Sales, inventory management, and other parts of the business can all benefit from these forecasts. According to previous research, linear, machine learning, and deep learning are frequently used to estimate sales volume [5].

2 Literature survey

The section briefly describes previous research works and their approaches to sales prediction actions. Approaches like the decision Tree provide conditions on values with specific attributes that are used to predict sales with its accuracy. Furniture manufacturing offers various items and prices, from simple holders to large, expensive furniture sets. Early furniture cost estimation is advantageous for accelerating product introduction, cutting costs, and improving quality while preserving market competitiveness. The rapid rise of the e-commerce industry is felled by fierce rivalry among different businesses. Convolution Neural Network architecture takes the input, and importance is assigned to various aspects that can differentiate from one another. The sentiment category used a single one-dimensional convolution layer with cross filters, a max pooling layer for identifying prominent features, and a final completely connected layer [6]. In the ARIMA model, an autoregressive and a moving average element are both supported. The main disadvantage of ARIMA is that it needs to help with seasonal data. That is an iterative cycle time series. In forecasting, the above approach has a less specified accuracy level [7].

It takes longer to train LSTM (Long Short-Term Memory). To train, LSTMs demand additional memory. It's simple to overfit LSTMs. In LSTMs, dropout is far more challenging to implement. Different random weight initializations affect the performance of LSTMs [8]. The three gates of the LSTM unit cell that update and control the neural network's cell state are the input gate, forget gate and output gate. When new data enters the network, the forget gate selects which it should erase information in the cell state. LSTM and RNNs (recurrent neural networks) can manage enormous amounts of sequential attention data. The RNN encode and decode technology is effectively used in language transformation. The performance of each child node is increased by adding one LSTM layer to the RNN. GRU (Gated Recurrent Unit) is a recurrent neural network that can retain a longer-term information dependency and is commonly utilized in business. Conversely, GRU still suffers from delayed convergence and poor learning efficiency [9]. Making a lot more exact price correction at a particular time is made possible using supply chain models and machine learning, which allow for previous data and the consequences of various aspects of revenues.

3 Proposed work

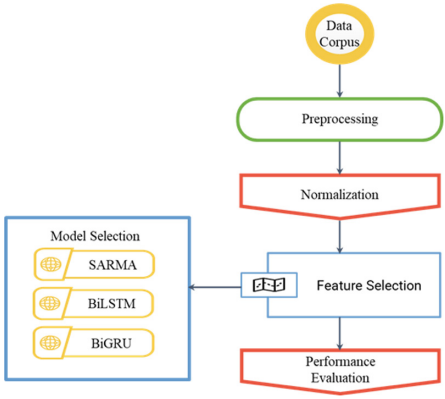


Fig. 1. Framework for proposed work

In particular, we Generally utilized the Support Vector (SV) and Machine Learning (ML) models in a range of prediction models to produce good results in many prior models, and the hyperplane is the closest to these models. They have been effectively shown in various situations. The issue of sales forecasting and prediction has been the subject of numerous studies. The suggested methods have been used in furniture sales; we compare three distinct algorithms to achieve high accuracy. If a platform wishes to keep its competitive advantage, it must better match user needs and perform well in all aspects of coordination and management [10]. The precise forecasting of e-commerce platform sales volume is critical now. Hence, we propose a method to forecast furniture sales using different algorithms, namely SARIMA, BiLSTM and BiGRU, shown in

Fig. 1. The central part of the novelty lies in comparing these three algorithms to achieve higher prediction accuracy.

3.1 SARIMA

Seasonal ARIMA, or Seasonal Auto-regressive Integrated Moving Average (SARIMA), is an ARIMA extension that explicitly supports seasonal uni-variate time series data [11]. It introduces three new top-level model parameters for the seasonal component of the series, including auto-regression (AR), differencing (I), and moving average (MA), as well as a fourth parameter for the seasonality period.

Configure SARIMA. To select top-level model parameters for the series trend and seasonal elements used in Eqs. (1) to (4). Configuration is required for three trend items. They are identical to the ARIMA model, in particular: X_t: Auto regression order Trend, α, ϕ, β: Trend moving average order, ε_t and $e(t)$: encoder value, y_t: Trend difference order four seasonal elements must be configured that are not part of ARIMA same as non-seasonal

$$X_t = \alpha_0 + \alpha_1 y_t - 1 + \alpha_2 y_t - 24 + \alpha_3 y_t - 25 + \varepsilon_t \tag{1}$$

SARIMA (25,0,0) model (with a few coefficients set to zero). SARIMA (1,0,0) (1,0,0)24 Actually, they are identical as much as a limit at the coefficients. SARIMA(1,0,0)(1,0,0)24, the subsequent have to hold:

$$\alpha_1 = \phi_1,\ \alpha_2 = \phi_{24},\ \alpha_3 = -\phi_1\phi_{24} \tag{2}$$

Hence, for a given pair (β_1,β_2)(β_1,β_2), the remaining coefficient $\beta 1_1 \beta_3$ is fixed:

$$\alpha_3 = -\alpha_1\alpha_2\alpha_3 \tag{3}$$

In SARIMA (25,0,0) rather than a SARIMA (1,0,0) (1,0,0)24, if this constraint does not hold 25, take a look at the speculation H0:$\alpha 3 = -\alpha 1\alpha 2$. If we cannot reject it, it will move for SARIMA (1,0,0) (1,0,0)24;

$$x(t) = \alpha 1 * Y(t - 1) + e(t) \tag{4}$$

BI-LSTM. A softmax output layer with three neurons per word is coupled to the bi-LSTM through a connected hidden layer. To avoid over-fitting, we practice dropout among the Bi-LSTM layer and the hidden layer, in addition to among the hidden layer and the output layer. The act of creating any neural community collection data in each approach backwards (destiny to beyond) or forwards (in advance to future) is referred to as bidirectional long-time period memory (BI-LSTM) [11] (beyond to destiny).

Bidirectional LSTMs range from traditional LSTMs in that their entry flows in directions. We could make enter glide in a single path using a traditional LSTM, both backwards or forwards. However, with bi-directional access, we will have the data glide in each direction, keeping each the destiny and the beyond data [12]. Let's take observe an instance for a higher understanding. Many collections processing jobs gain from analyzing each the future and the beyond at a given factor within the series. However,

maximum RNNs are most straightforwardly designed to observe information in a single path: backward. A partial treatment to this flaw consists of a put-off among inputs and their corresponding targets, supplying the internet with some time steps of destiny context. But this is mainly similar to the set time-home windows hired using MLPs, which RNNs had been created to replace.

LSTM networks structure turned into in the beginning advanced with the aid of using Hochreiter and Schmidhuber More formally, an enter series vector $x = (x_1, x_2, \ldots\ldots, x_n)$ is given, wherein n suggests the duration of the enter sentence. Three manipulate gates modify a reminiscence mobiliary activation vector, that's the LSTMs primary structure. The first overlook gate determines how lots of the mobiliary country C_{t-1} on the preceding time is retained till the cutting-edge mobiliary country C_t; the second one enters gate determines the quantity to which the center of the community is stored to the cutting-edge mobiliary country C_t; the 0.33 output gate determines how lots of the mo-biliary country C_t stands for cutting edge transmit output price H_t of the LSTM networks. Input, forget and output gates are linked with LSTM architecture is proven in the following Eqs. (5) to (9):

$$Input\ Gates: i_{gt} = \sigma\left(W_{gx}X_t + W_{gh}h_{t-1} + a_g\right) \tag{5}$$

$$Forget\ Gates: f_{st} = \sigma\left(W_{sx}X_t + W_{sh}h_{t-1} + a_s\right) \tag{6}$$

$$Output\ Gates: o_{rt} = \sigma\left(W_{rx}X_t + W_{rh}h_{t-1} + a_r\right) \tag{7}$$

$$CellStates: c_{bt} = f_t * b_{t-1} + i_t * tanh(W_{bx}X_t + W_{bh}h_{t-1} + a_b) \tag{8}$$

$$Cell\ Outputs: c_{kt} = o_{rt} * tanh_{bt} \tag{9}$$

where σ stands for sigmoid function, x_t is the word vector in t^{th} sentence, kt is the hidden layer, W is the weight matrices of the terms; likewise, W_{xf} forget gate weight matrix, W_{bx} backward gate weight matrix) and bt stands for bias vectors for three gates, respectively [13]. The activation function can use linked data from past and future contexts thanks to this structure. Unlike a hidden forward sequence and a backward hidden sequence, a Bi-LSTM determines the input sequence from $x = (x_1, x_2, \ldots, x_n)$. The encoded vector is created by concatenating the last forward and backward outputs, where $y = (y_1, y_2, \ldots y_t \ldots, y_n)$ represents the first hidden layer's output sequence

BI-GRU. A time-collection forecasting approach uses beyond records to predict the object running country found within the destiny period [14]. Over time, the found time-collection records changes. GRUs utilizes Gating Mechanisms to modify the statistics that the community maintains, figuring out whether or not it must ship the statistics to the subsequent layer or overlook it. A GRU most effectively has gates replace gate (u_t)and reset gate (r_t). It uses less matrix multiplication to grow the version's schooling speed. The replacement gate is used to modify the subsequent occasion after the preceding event. Meanwhile, the reset gate is hired to save you the previous occasion's country statistics from being forgotten.

In all transmission states, GRU is a unidirectional neural community version this is forwarded in a single course. Bi-GRU is a bi-directional neural community version that takes entry from one system and forgets the country inside the contrary path. The result of the contemporary time is tied to the statuses of preceding and destiny occurrences. That's how Bi-GRU is introduced. The Bi-GRU neural community version has GRUs that is unidirectional [15]. Bi-GRU has a complete series of statistics in a given sequence at any time. The equation has also been defined using the Bi-GRU as follows:

$$BiGRU(q_t,\ s_{t-1}) = s_t \oplus s_t \tag{10}$$

$$s_{ft} = GRU(q_t,\ s_{t-1}) \tag{11}$$

$$s_{bt} = GRU(q_t,\ s_{t+1}) \tag{12}$$

Equation (10) denotes that the hidden layer of the Bi-GRU at time t is obtained from the input Ext, *sft* where forward hidden layer output is Eq. (11), and *sbt* stands for Backward hidden layer output is Eq. (12) which is concatenated to get Bi-GRU.

4 Dataset

Data gathering is a critical constraint in deep learning and a topic of intense debate in many communities. Data collecting has recently become a key concern for two reasons. To begin with, as machine learning becomes more extensively employed, we're seeing new applications that don't always have enough tagged data. In contrast to regular machine learning, deep learning methods automatically produce features, minimizing feature engineering costs but sometimes requiring more classification models. Dataset ratio is shown in Table1.

Table 1. Data Process

Training Data	80%
Testing Data	20%

After gathering the data, it may be necessary to pre-process it to make it appropriate for deep learning. While there have been many proposed crowd operations, the relevant ones are data curation, entity resolution, and dataset joining. Data Tamer is a full-featured data curation solution that can clean, convert, and semantically integrate datasets. Data Tamer includes a crowd-sourcing component (Data Tamer Exchange) that allocates tasks to employees.

4.1 Training Phase

During this phase, the dataset is pre-processed using a specialized technique called SARIMA. It must train numerical data to the scheme; initially, the source data was

separated by criteria such as prior sales records. The training process is processed using the training labels in the pre-processing and feature extraction state based on the features. The model is then used to turn the data into a dataset for the model architecture after being fed.

4.2 Testing Phase

This testing step evaluates the model, which measures output correctness and improves accuracy by increasing the number of training stages. When the numerical test data is fed into the model, the past sales records are examined and the features extracted using the BI-LSTM and BIGRU [16] algorithms, which are then compared to the learned model. A histogram comprises boxes that are adjacent (bordering). It is divided into two axes, one horizontal and the other vertical. It can show the data on the horizontal axis, which is labelled [17]. On the vertical axis, the frequency or relative frequency vector is labelled. The graph will have the same shape regardless of the label. Like the stem plot, the histogram can show you the data form and its centre and spread.

When provided data from a time series, the succeeding values in the series usually correlate. Persistence, sometimes known as inertia, is a series correlation in which the lower frequencies of the frequency spectrum have more strength. Persistence can significantly minimize the degrees of freedom in time series modelling (AR, MA, ARMA models). Because persistence reduces the number of independent observations, it complicates the statistical significance test [18].

5 Results and Discussion

5.1 Settings

The models are implemented with Google Colab, the deep learning package. The dataset with 12121 numerical records is used. Use the model training time for evaluation metrics. The objective is to forecast sales for the upcoming seasonal days. The stats model package implements the Tri Simple Exponential model. The SARIMA model is adjusted with auto-ARIMA, and the ideal model is ARIMA (p, q, d), which considers the number of instances the raw data are differentiable and the dimensions of the time series window. The model parameters can predict the results.

Use lagged data from past years as input to the deep learning models. The deep learning models are then implemented. A total of 28 unique deep learning models are generated for each type of deep learning model. Because deep learning models are sensitive to input data dispersion, we pre-processed the sales history data using normalization. Deep learning models can also forecast a sequence for the following few days and years simultaneously.

5.2 Evaluation Parameters

We used Python3, Anaconda, Google Colab and Virtual Environments to implement the proposed module. Python libraries are used to display all the results and graphs. In this

work, the standard furniture sales dataset consists of comprehensive 12121 data where 11194 data are tagged as positive, and 1927 are negative. Based on the quantity and discount, the sales average is gradually increased. When the quantity is increased, the sales are also increased.

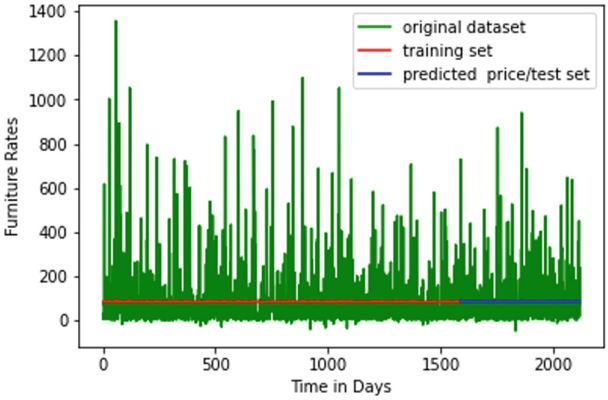


Fig. 2. BI-LSTM Observed Forecast

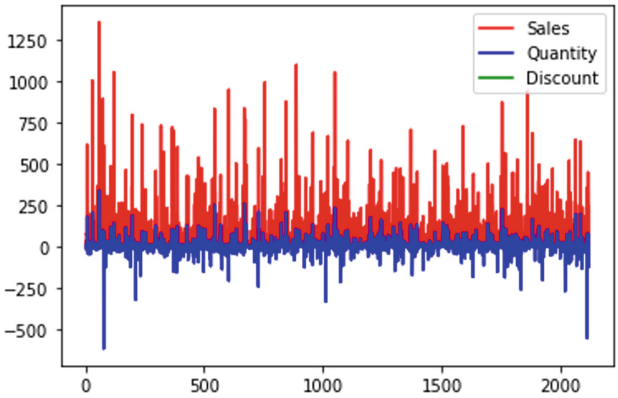


Fig. 3. BI-LSTM Sales Forecast

Considering the performance and validation of the variables Fig. 2 shows the sales forecasted using the BI-LSTM algorithm. Whereas Fig. 3 provides the forecast that is observed frequently depending upon the sales. Figure 4 displays the forecast utilizing the BI-GRU algorithm while considering the factors' performance and validation. Figure 5 contrasts this by showing the forecast observed regarding furniture sales.

$$\text{Precision} = \frac{\text{True Positive(TP)}}{\text{True Positive(TP)} + \text{False Positive(FP)}} \tag{13}$$

$$\text{Recall} = \frac{\text{True Positive(TP)}}{\text{True Positive(TP)} + \text{False Negative(FN)}} \tag{14}$$

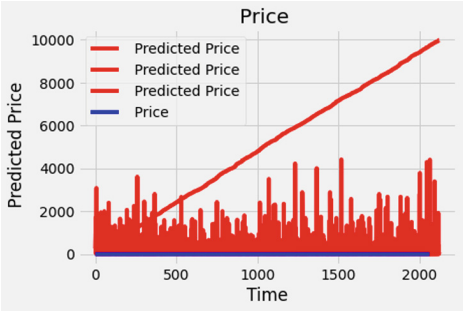


Fig. 4. BI-GRU sales forecast

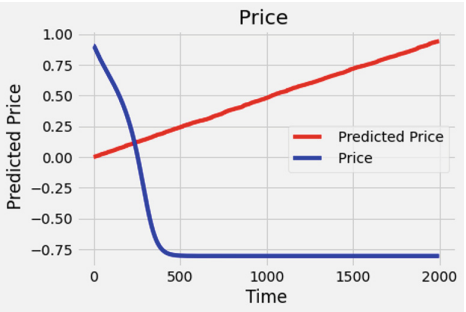


Fig. 5. BI-GRU observed forecast

$$\mathrm{F-Measure} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{15}$$

Additionally, we performed several analyses on the input sales data for furniture while projecting future prices week-wise sale prediction by hybrid (SARM-BiLSTM-BiGRU) as seen in Eqs. (13) to (15).

Table2. Analysis Parameters of furniture Dataset

Algorithm	Precision	Recall	F-Measure	Accuracy
SARMA- BiLSTM	88.27	88.89	89.19	89.01
SARMA- BiGRU	88.29	89.31	89.11	89.05
SARMA	81.21	82.71	82.31	82.11
BiLSTM	83.54	83.69	84.20	84.18
BiGRU	84.31	84.23	84.11	84.01

Table 2 shows that the hybrid approach can better sales more accurately than the traditional model. The performance analysis of our models is precision, recall and F-measure for the furniture datasets as shown in Fig. 6, the numerical classification, i.e., positive and negative, which means feelings are assessed to determine whether furniture sales are positive or negative.

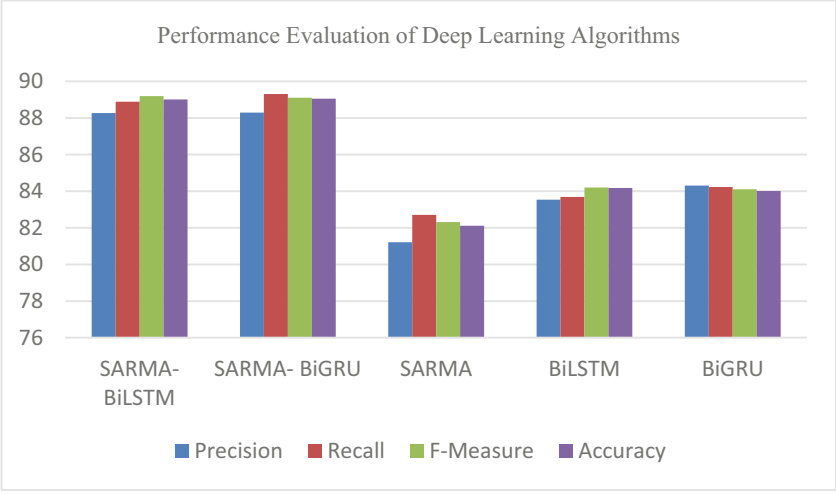


Fig. 6. Performance Metrics on Multi-model

6 Conclusion and Future Work

Predicting future developments in the marketplace is critical in deep learning approaches to maintaining profitable company activity. It can reduce the time it takes to predict with greater precision, allowing for faster furniture manufacturing. Experiments with sales forecasting can be used as a baseline for determining how well furniture sells. State space models, SARIMA, Bi-LSTM, and Bi-GRU models were used to anticipate sales for a multinational furniture retailer operating in Turkey in this study. In addition, it examined the performance of specific commonly used combining methods by comparing them to weekly sales. Furniture datasets are used to evaluate the proposed approaches, and the results demonstrate the superiority of our method over the standard procedures. We'll look at the time component in the future because one of Deep Learning's key limitations is that it takes a long time to process data due to the high number of layers. Deep learning will always be the preferred approach for fine-grained sentiment analysis prediction and classification. We can solve time consumption issues; new learning research and the creation of a new self-attention mechanism are expected to improve quality further.

References

1. Pliszczuk, D., Lesiak, P., Zuk, K., Cieplak, T.: Forecasting sales in the supply chain based on the LSTM network: the case of furniture industry. Eur. Res. Stud. J. 0(2), 627–636 (2021)
2. Ensafi, Y., Amin, S.H., Zhang, G., Shah, B.: Time-series forecasting of seasonal items sales using machine learning – a comparative analysis. Int. J. Inf. Manag. Data Insights **2**, 2667–0968 (2021)
3. Mitra, A., Jain, A., Kishore, A., et al.: A comparative study of demand forecasting models for a multi-channel retail company: a novel hybrid machine learning approach. Oper. Res. Forum **3**, 58 (2022)
4. Ungureanu, S., Topa, V., Cziker, A.C.: Deep Learning for Short-Term Load Forecasting—Industrial Consumer Case Study, vol. 21, p. 10126 (2021)
5. Haselbeck, F., Killinger, J., Menrad, K., Hannus, T., Grimm, D.G.: Machine learning outperforms classical forecasting on horticultural sales predictions. Mach. Learn. Appl. **7**, 2666–8270 (2022)
6. Rosado, R., Abreu, A.J., Arencibia, J.C., Gonzalez, H., Hernandez, Y.: Consumer price index forecasting based on univariate time series and a deep neural network. Lect. Notes Comput. Sci. **2**, 13055 (2021)
7. Falatouri,, T., Darbanian, F., Brandtner, P., Udokwu, C.: Predictive analytics for demand forecasting – a comparison of SARIMA and LSTM in retail SCM. Procedia Comput. Sci. **200**, 993–1003 (2022)
8. Ang, J.-S., Chua, F.-F.: Modeling Time Series Data with Deep Learning: A Review, Analysis, Evaluation and Future Trend (2020)
9. Kim, J., Moon, N.: CNN-GRU-based feature extraction model of multivariate time-series data for regional clustering. In: Park, J.J., Fong, S.J., Pan, Y., Sung, Y. (eds.) Advances in Computer Science and Ubiquitous Computing. Lecture Notes in Electrical Engineering, vol. 715 (2021)
10. Ibrahim, T., Omar, Y., Maghraby, F.A.: Water demand forecasting using machine learning and time series algorithms. In: IEEE International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 325–329 (2020)
11. Buxton, E., Kriz, K., Cremeens, M., Jay, K.: An auto regressive deep learning model for sales tax forecasting from multiple short time series. In: 18th IEEE International Conference on Machine Learning And Applications (ICMLA), pp. 1359–1364 (2019)
12. Ferretti, M., Fiore, U., Perla, F., Risitano, M., Scognamiglio, S.: Deep learning forecasting for supporting terminal operators in port business development. Futur. Internet **14**, 221 (2022)
13. Júnior, S.E.R., de Oliveira Serra, G.L.: An approach for evolving neuro-fuzzy forecasting of time series based on parallel recursive singular spectrum analysis. Fuzzy Sets Syst. **443**, 1–29 (2022)
14. Li, X., Ma, X., Xiao, F., Xiao, C., Wang, F., Zhang, S.: Multistep Ahead Multiphase Production Prediction of Fractured Wells Using Bidirectional Gated Recurrent Unit and Multitask Learning, pp. 1–20 (2022)
15. Li, Y., Wang, S., Wei, Y., Zhu, Q.: A new hybrid VMD-ICSS-BiGRU approach for gold futures price forecasting and algorithmic trading. IEEE Trans. Comput. Soc. Syst. **8**(6), 1357–1368 (2021)
16. Kadli, P., Vidyavathi, B.M.: Deep-Learned Cross-Domain Sentiment Classification Using Integrated Polarity Score Pattern Embedding on Tri Model Attention Network, vol. 12, pp. 1910–1924 (2021)
17. Kurasova, O., Medvedev, V., Mikulskienė, B.: Early cost estimation in customized furniture manufacturing using machine learning. Int. J. Mach. Learn. Comput. **11**, 28–33 (2021)

18. Sivaparvathi, V., Lavanya Devi, G., Rao, K.S.: A deep learning sentiment primarily based intelligent product recommendation system. In: Kumar, A., Paprzycki, M., Gunjan, V.K. (eds.) ICDSMLA 2019. LNEE, vol. 601, pp. 1847–1856. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-1420-3_188

VANET Handoff from IEEE 80.11p to Cellular Network Based on Discharging with Handover Pronouncement Based on Software Defined Network (DHP-SDN)

M. Sarvavnan[1(✉)], R. Lakshmi Narayanan[2], and K. Kavitha[3]

[1] Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India
sarvan148@yahoo.com

[2] Department of Networking and Communications, SRM Institute of Science and Technology, Chennai, India
lakshmir4@srmist.edu.in

[3] Department of Computer Science and Engineering, Kongu Engineering College, Erode, India

Abstract. In Vehicular Adhoc Network (VANET) is an emerging domain with high dynamic mobility and disrupted network in Discharging with Handover Pronouncement based on Software Defined Network (DHP-SDN) with advance predictive management tool described in this work for discharging of vehicle-to-infrastructure communication with Software Defined Network (SDN) architecture. Various parameters were addressed in research articles like maintaining connectivity, identifying appropriate intermediate node for carrying signal and transferring of data but in the proposed model SDN monitors the discharging signal based on the speediness, terrestrial location, nearby vehicle RSU from the vehicle which has cellular and IEEE 802.11p network boundary. SDN regulator computes when the due time is to settle on a choice, chooses whether it is appropriate or not have the vehicle to handoff from cell organization to the ahead IEEE 802.11p organization. The simulation results that using DHP-SDN method maintains the networking quality by reducing the load and traffic congestion.

Keywords: Cellular Network · Software Defined Network · Vehicular Communication · VANET Routing

1 Introduction

In heterogeneous VANET an IPv_6 protocol based architecture is introduced for the cloud to smart vehicle convergence that covers three parameters location manager, cloud server and mobility management [1]. In paper [2] it is elucidated about the VANET communications using the short range communication along with the defies and developments. Complete study about the VANET design, features and solicitations are given and it advises to use the appropriate simulation tool which supports for effectiveness communication [3]. Paper [4] how much amount of data off loaded using Wi-Fi in 3G and also

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 971–980, 2023.
https://doi.org/10.1007/978-3-031-27409-1_89

indicates how much amount of battery power can be saved in real time using the given data traffic. An analysis about the vehicular opportunistic offloading is elaborated and it provides the offloading strategies to the grid operator and vehicular manipulators [5]. Concepts of Network SDN and its architectures, challenges and features are described and both configuring the network in accordance with specified policies and modifying it to handle faults, load, and changes are challenging. The necessary flexibility is dependent on the introduction of a concern separation between the development of network policies, their implementation in switching hardware, and the forwarding of traffic [6]. Rapid increment of smart phones, cellular phones and laptops usage in the recent years it increases the data traffic in networking. In order handle the data traffic optimistically various methods where introduced in VANET, [7] gives the complete reviews of technologies that supports for offloading.

Study about the WiFi offloading and how it creates a significant impact for congestion avoidance and overhead issues in the heterogeneous network is elaborated. In order to prevent overload and congestion on cellular networks and to ensure user happiness, network operators must perform. In [8] it is described that a number of WiFi offloading strategies that are currently in use and talk about how different types of heterogeneous networks' properties affect the choice of offloading. In [9] two types of mobile offloading like opportunistic and delayed offloading technique is analyzed with the parameters, residence time (both WiFi and Cellular), delay, and data rate and session duration efficiency. A framework for data service analysis in VANET using queuing analysis is expounded. To be more specific, we take into account a generic vehicle user with Poisson data service arrivals to download/upload data from/to the Internet using the affordable WiFi network or the cellular network giving complete service coverage [5]. By using an M/G/1/K queuing model, it is established an explicit relationship between offloading effectiveness and average service delay, and then look at the tradeoff between the two models.

2 Related Works

In [10] novel method is introduced for roaming decision and intelligent path selection for the optimistic utilization and balanced data traffic in VANET. Method that helps mobile nodes choose the best time to determine whether to ramble and the preferred point of service based on the operator's policies and the current state of the network. Additionally, it introduces the 3GPP ANDSF, TS24.312 simulation model of a heterogeneous network with WiFi and cellular interworking. This technique improves throughput dynamically by directing the mobile nodes from the node to access point. With the help of delayed offloading scheme in networking, it is possible to handle in data flare-up issue suffered by both user and provider. In a market with two benefactors, we take into account the possibility that one of the providers, let's say A, introduces a delayed Wi-Fi offloading service as a stand-alone service from the primary cellular service.

This would enable users of the other provider, let's say B, to sign up for the offloading service from A, though some would have to pay a switching fee [11]. Considering the link availability, connectivity, and quality between the node and Road Side Unit (RSU) an analysis is done based on improved issue to maximize the data flow vehicular network

[12]. Increased number of nodes in cyber physical system leads to overloaded data traffic and this problem is controlled by using mixed-integer solution and the QoS is 70% guaranteed [13]. In [14] two types of game based offloading (auction and congestion) mechanism is used and which provides good performance for vehicle users and fairness. Introducing big data analysis in Internet of Vehicle (IoV) in VANET is an unavoidable scheme to handle the enormous data in traffic and providing a preference to access the network appropriately. That is done [15] using big data analysis through the traffic model and diagnostic structure. The quality of service for automobiles in a cellular and VANET-based vehicular heterogeneous network is contingent on efficient network selection. We create an intelligent network recommendation system backed by traffic big data analysis to address this issue.

First, big data analysis is used to construct the network recommendation traffic model. Second, by using the analytical system that takes traffic status, user preferences, service applications, and network circumstances into account, vehicles are advised to access an appropriate network. Additionally, an Android application is created that allows each vehicle to automatically access the network based on the access recommender. Finally, thorough simulation results demonstrate that our concept may efficiently choose the best network for vehicles while simultaneously utilizing all network resources.

In [16] offloading is performed by considering the parameter like, link quality between node and road side unit, link quality between two nodes in the same direction and channel contention. Due to the more usage of smart phones and greedy applications in cellular networks it leads to data overloaded. This is handled with optimization problem technique by enhancing offloading mechanism that considers various parameters like link quality, channel capacity and road side unit [17]. VANET offloading is done with help of heuristic algorithm with the parameters like link quality, channel capacity and bandwidth efficiency. Utilization patterns of both wireless and backhaul links are reviewed for the intention of enhancing resource exploitation, taking into account practical concerns such as link quality variety, fairness, and caching. A two-phase resource allocation process is used [18]. Bravo-Torres et al. [19] Debates about the VANET mobile offloading in urban area with the implementation of virtualization layer which deals about the vehicle mobility and virtual nodes.

By this both the routing (topological and geographical) shows the better performance than any other conventional routing methodology. In V2V (Vehicle to Vehicle) communication a virtual road side unit is introduced in [20] to reduce the problem of local minima. For replacing femtos and WiFi in cellular communication by offloading, to handle data overflow and maximizing the downloader route VOPP is done as an analytical study [21].

In vehicular communication for uploading data from vehicle to centralized remote center, it faces many challenges and that is solved by implementing WAVE or IEEE 802.11p routing in [22]. In this case, the goal is to offload cellular networks. In this work, we suggest and talk about implementing the WAVE/IEEE 802.11p protocols, which the most recent technology for short-range vehicle-to-vehicle and vehicle-to-roadside communications Avoiding the occurrence of congestion in VANET, DIVERT mechanism is implemented to re-route the data in alternative path [23]. SDN architecture based rerouting in done for mobile off-loading with data loss detection. This study

proposes an architecture for "Automatic Re-routing with Loss Detection" that uses the Openflow protocol's queue stat message to identify packet loss.

The Re-routing module then attempts to discover a workaround and applies to flow tables [24]. For taking the critical decisions for reducing the routing overhead data aggregation method is used in [25]. Network quality like life time, bandwidth resource utilization may get affected due to the injection of false data, which is analyzed in [26]. In [27] for multimedia data transmission, encryption system is introduced for providing secured data. Secured data sharing in machine learning and Internet of things are done by implementing filmic cryptographic method [28]. In [29] reinforcement learning is used for optimized routing in VANET. In [30] data are securely transmitted using secured protocol for effective communication.

3 SDN-VANET Architecture and Issues

3.1 SDN-VANET Based Architecture

Conventional VANET architecture communication range is very short, since it uses short range communication protocol. Integrating Software Defined Network with VANET gives global view of the communication. Here SDN controller is introduced and it contains apt regulation and symptom for all nodes. This mechanism concentrates mainly on data transmission and maximum possible control verdicts.

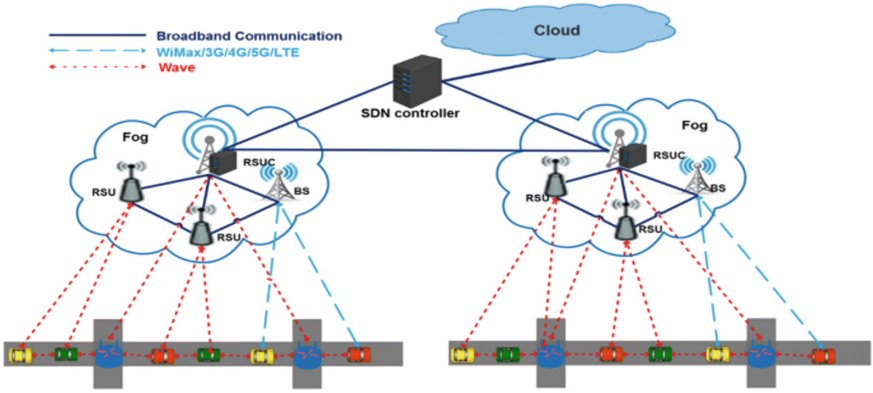


Fig. 1. SDN based VANET architecture

Figure 1 shows the SDN-VANET architecture, here SDN controller manages the overall communication without any interruption by providing mobile offloading. This architecture supports for vehicle to vehicle communication through road side unit. Here controller receives all the messages forwarded by participating nodes and perform calculation to take appropriate decisions to avoid data traffic. After the decision apt vehicle will be identified for data transmission using mobile offloading. Each node participating in communication should transmit the data periodically to the controller through

an interface. Sometimes controllers gives privileges for vehicles to check the suitability about nearby RSU, for enhancing handoff. Then data transmission takes place via suitable RSU.

3.2 SDN-VANET Architectural Issues

Main anxiety about offloading between cellular network to WiFi and vice versa are, whether handoff decision can be made before vehicle A sensing the signal. Because if the decision made before enters in to the network then handoff overhead shall be avoided. When decision is about positive handoff, it takes long time in local RSU and when decision is about negative handoff, maintain in the cellular network. First vehicle enters into the RSU signal area and followed by that, if handoff session is made then it takes more computing time. In the RSU coverage area if the driving path is very short and velocity of the vehicle is fast means such a node cannot stay in that coverage area for long time.

4 Regulator Contrivances

In order to control the above mentioned issues SDN based handoff mechanism is proposed in this work. As per the instruction vehicle and IEEE 802.11p participating in communication will share its information to the SDN controller and issues will be handled by controller. Periodically information about the vehicle's direction, velocity and vehicle ID will be updated.

4.1 Offloading Decision Making and Evaluation of Stay Time

Based on the information like direction, velocity and ID provided by vehicle and RSU in VANET communication, decision for handoff is done by SDN controller that estimates the distance and time taken between vehicle boundary and RSU. Stay time of every node in the RSU signal area are estimated by SDN controller based on its velocity. Using Cartesian coordinates (C_x, C_y) considering length of the path and node velocity it is calculated. CSMA with collision avoidance method (back-off algorithm) is used for improving network quality. This algorithm uses counters for balancing transmission and when the collision occurs retransmission takes place accordingly. Contention window of the channel get doubled when it is occupied by other vehicle. This window helps for sustaining network quality by doubling and retransmission.

5 Proposed Handoff Scheme

DHP-SDN based Handoff relies on Software Defined Network. It carries three stages like offload decision, selection of RSU and applying function. SDN controller estimates time of the vehicle and controls entire network. Algorithm will be triggered by SDN controller that depicts the control scheme process. RSU's highest score will be estimated by looking the recently updated database by SDN controller. Then the RSU ID will be returned to the consistent vehicle else if it returns NULL vehicle should stay in the network. Handoff migration from IEEE 802.11p to cellular network is done, while signal strength of the VANET is getting weak for data transmission.

Table 1. Simulation Configuration

Parameter	Value
RSU Coverage Range	300 m
Number of Vehicles	25
Velocity	11.10 m/s
Duration	250 s
Packet Payload	1496 Bytes
RTS/CTS	Off
RSU Number	6
Cellular Network	LTE
RSU Bandwidth	8 Mbps
Cellular Bandwidth	24 Mbps
Data Sending Rate	1 Mbps/2 Mbps

6 Performance Analysis

Performance is done with the existing algorithm while the IEEE 802.11p vehicle drives through the VANET coverage range. RSU coverage range is about 300m is used for 25 vehicles with the velocity of 11.10 m/s. Duration taken for the entire calculation is about 250 s, packet payload is about 1496 bytes. Total number of RSU taken for estimation is 6 with LTE network. RSU and Cellular bandwidth is 8 Mbps and 24 Mbps respectively. Total data sending rate is with 1 Mbps. When the number of data packets moves from one end to another with small amount of intermediate nodes it takes more time to reach destination for carrying complete payload. In case of VANET more number of nodes are participating to carry signal from one end to another with minimum time. Our performance is measured respect to definite time, payload and participating nodes with the above data mentioned Table 1.

6.1 Result Analysis

Following section presents the performance analysis of four major parameters like throughput, RSU throughput, delay and RSU coverage ratio. All the parameters additionally increases the network capacity. This results focuses on the energy efficiency, throughput and space road side unit coverage (Fig. 2).

Compared our algorithm with already existing algorithms, respect to the other existing algorithms our DHP-SDN algorithm coverage and delivery ratio is comparatively high. Algorithm gives better performance. When the number of vehicles are initially coverage of space is comparatively low and irrespective of existing and proposed algorithms.

But still in this scenario our proposed algorithm performs well and its coverage increased gradually for the number of vehicles from 50, 60,70,80,90. Figure 3 shows

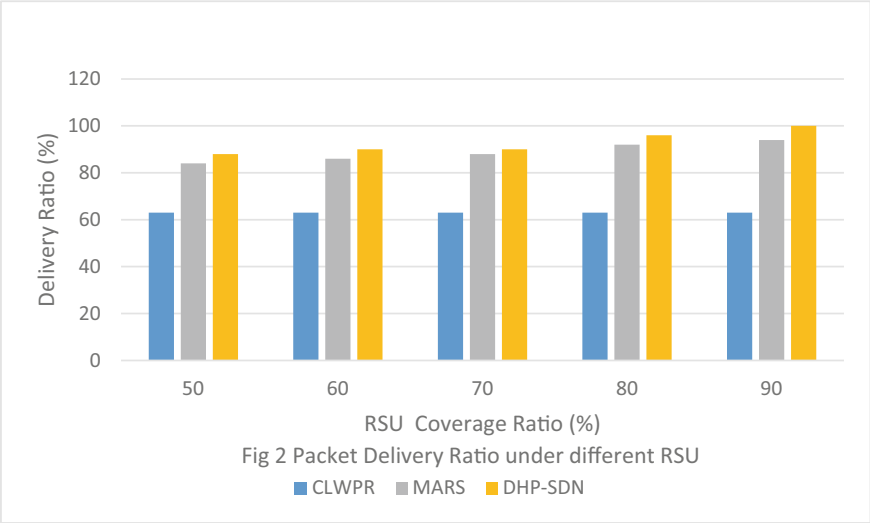


Fig. 2. RSU Coverage Ratio vs Delivery Ratio

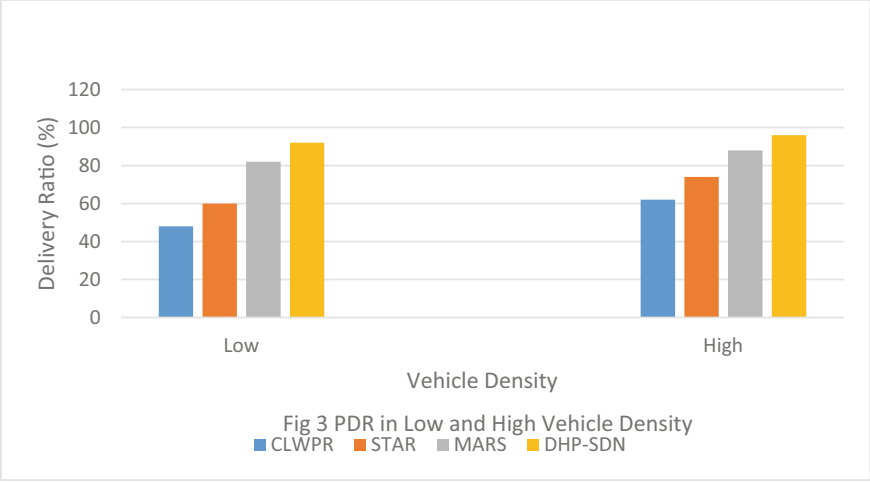


Fig. 3. Vehicle Density vs Delivery Ratio

the vehicle density versus delivery ratio and in this too even in low and higher density delivery ratio is comparatively good by DHP-SDN network. Both the cases our algorithm performs effectively. In all the networking concepts when the density of the participating vehicles are high due to hand hovering of data from one node to another it takes time in VANET this helps a lot for reaching destination as soon as possible (Fig. 4).

Energy efficiency is estimated based on participated vehicles, where energy is improved when the number of vehicles get increased. Because when number of vehicles increased it gives more chances for sharing its energy with its all participating nodes.

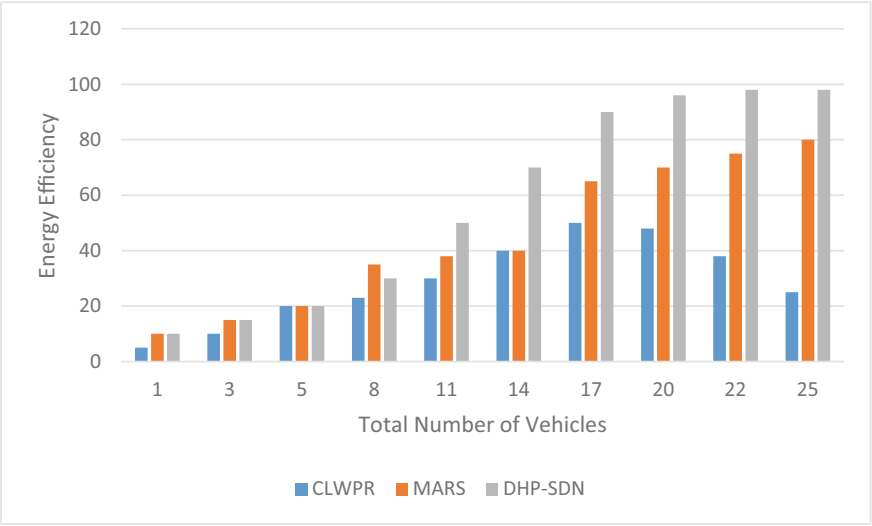


Fig. 4. Total number of Vehicles vs Energy Efficiency

Comparing to other wireless networks and mobile adhoc networks providing energy for all the nodes in VANET is not that much complicated issue due to auto generation of its power.

7 Conclusion

In VANET for the purpose of smooth hand-over from IEEE 802.11p to Cellular network and vice versa, a predictive management tool is introduced and that works with the support of SDN-Controller. Idea behind this implementation is collects all the all the participating vehicles information and passes to the RSU comes in the coverage area. Depends on the signal strength of the node participating for communication, smart decision will be taken by tool using the SDN controller. Comparing to the existing algorithm our proposed DHP-SDN algorithm works efficiently in various parameters like vehicle density, coverage ratio and network capacity etc. In future it is planned to implement the same concept for urban and rural area and try to identify its significant benefits and research issues.

References

1. Matzakos, P., Härri, J., Villeforceix, B., Bonnet, C.: An IPv6 architecture for cloud-to-vehicle smart mobility services over heterogeneous vehicular networks. In: 2014 International Conference on Connected Vehicles and Expo (ICCVE), pp. 767–772. IEEE (2014)
2. Wu, X., et al.: Vehicular communications using DSRC: challenges, enhancements, and evolution. IEEE J. Sel. Areas Commun. **31**(9), 399–408 (2013)
3. Al-Sultan, S., Al-Doori, M.M., Al-Bayatti, A.H., Zedan, H.: A comprehensive survey on vehicular ad hoc network. J. Netw. Comput. Appl. **37**, 380–392 (2014)

4. Lee, K., Lee, J., Yi, Y., Rhee, I., Chong, S.: Mobile data offloading: how much can WiFi deliver? IEEE/ACM Trans. Netw. **21**(2), 536–550 (2012)
5. Cheng, N., Lu, N., Zhang, N., Shen, X. S., Mark, J.W.: Opportunistic WiFi offloading in vehicular environment: a queueing analysis. In: 2014 IEEE Global Communications Conference, pp. 211–216. IEEE (2014)
6. Kreutz, D., Ramos, F.M., Verissimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-defined networking: a comprehensive survey. Proc. IEEE **103**(1), 14–76 (2014)
7. Aijaz, A., Aghvami, H., Amani, M.: A survey on mobile data offloading: technical and business perspectives. IEEE Wirel. Commun. **20**(2), 104–112 (2013)
8. He, Y., Chen, M., Ge, B., Guizani, M.: On WiFi offloading in heterogeneous networks: various incentives and trade-off strategies. IEEE Commun. Surv. Tutor. **18**(4), 2345–2385 (2016)
9. Suh, D., Ko, H., Pack, S.: Efficiency analysis of WiFi offloading techniques. IEEE Trans. Veh. Technol. **65**(5), 3813–3817 (2015)
10. Nguyen, N., Arifuzzaman, M., Sato, T.: A novel WLAN roaming decision and selection scheme for mobile data offloading. J. Electr. Comput. Eng. (2015)
11. Park, H., Jin, Y., Yoon, J., Yi, Y.: On the economic effects of user-oriented delayed Wi-Fi offloading. IEEE Trans. Wireless Commun. **15**(4), 2684–2697 (2015)
12. el Mouna Zhioua, G., Labiod, H., Tabbane, N., Tabbane, S.: VANET inherent capacity for offloading wireless cellular infrastructure: an analytical study. In: 2014 6th International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. IEEE (2014)
13. Wang, S., Lei, T., Zhang, L., Hsu, C.H., Yang, F.: Offloading mobile data traffic for QoS-aware service provision in vehicular cyber-physical systems. Futur. Gener. Comput. Syst. **61**, 118–127 (2016)
14. Cheng, N., Lu, N., Zhang, N., Zhang, X., Shen, X.S., Mark, J.W.: Opportunistic WiFi offloading in vehicular environment: a game-theory approach. IEEE Trans. Intell. Transp. Syst. **17**(7), 1944–1955 (2016)
15. Liu, Y., Chen, X., Chen, C., Guan, X.: Traffic big data analysis supporting vehicular network access recommendation. In: 2016 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2016)
16. el mouna Zhioua, G., Labiod, H., Tabbane, N., Tabbane, S.: A traffic QoS aware approach for cellular infrastructure offloading using VANETs. In: 2014 IEEE 22nd International Symposium of Quality of Service (IWQoS), pp. 278–283. IEEE (2014)
17. Zhioua, G.E.M., Labiod, H., Tabbane, N., Tabbane, S.: Cellular content download through a vehicular network: I2V link estimation. In: 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), pp. 1–6. IEEE (2015)
18. Chen, J., Liu, B., Gui, L., Sun, F., Zhou, H.: Engineering link utilization in cellular offloading oriented VANETs. In: 2015 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE (2015)
19. Bravo-Torres, J.F., Saians-Vazquez, J.V., Lopez-Nores, M., Blanco-Fernandez, Y., Pazos-Arias, J.J.: Mobile data offloading in urban VANETs on top of a virtualization layer. In: 2015 International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 291–296. IEEE (2015)
20. Bazzi, A., Masini, B. M., Zanella, A., Pasolini, G.: Virtual road side units for geo-routing in VANETs. In: 2014 International Conference on Connected Vehicles and Expo (ICCVE), pp. 234–239. IEEE (2014)
21. el Mouna Zhioua, G., Zhang, J., Labiod, H., Tabbane, N., Tabbane, S.: VOPP: a VANET offloading potential prediction model. In: 2014 IEEE Wireless Communications and Networking Conference (WCNC), pp. 2408–2413. IEEE (2014)
22. Bazzi, A., Masini, B.M., Zanella, A., Pasolini, G.: IEEE 802.11 p for cellular offloading in vehicular sensor networks. Comput. Commun. **60**, 97–108 (2015)

23. Pan, J., Popa, I.S., Borcea, C.: Divert: A distributed vehicular traffic re-routing system for congestion avoidance. IEEE Trans. Mob. Comput. **16**(1), 58–72 (2016)
24. Park, S.M., Ju, S., Lee, J.: Efficient routing for traffic offloading in software-defined network. Procedia Comput. Sci. **34**, 674–679 (2014)
25. Kumar, S.M., Rajkumar, N.: SCT based adaptive data aggregation for wireless sensor networks. Wireless Pers. Commun. **75**(4), 2121–2133 (2014)
26. Kumar, S.M., Rajkumar, N., Mary, W.C.C.: Dropping false packet to increase the network lifetime of wireless sensor network using EFDD protocol. Wireless Pers. Commun. **70**(4), 1697–1709 (2013)
27. Mary, G.S., Kumar, S.M.: A self-verifiable computational visual cryptographic protocol for secure two-dimensional image communication. Meas. Sci. Technol. **30**(12), 125404 (2019)
28. Selva Mary, G., Manoj Kumar, S.: Secure grayscale image communication using significant visual cryptography scheme in real time applications. Multimed. Tools Appl. **79**(15–16), 10363–10382 (2019). https://doi.org/10.1007/s11042-019-7202-7
29. Saravanan, M., Ganeshkumar, P.: Routing using reinforcement learning in vehicular ad hoc networks. Comput. Intell. **36**(2), 682–697 (2020)
30. Saravanan, M., kumar, S.M.: Improved authentication in vanets using a connected dominating set-based privacy preservation protocol. J. Supercomput. **77**(12), 14630–14651 (2021). https://doi.org/10.1007/s11227-021-03911-4

An Automatic Detection of Heart Block from ECG Images Using YOLOv4

Samar Das[1], Omlan Hasan[1], Anupam Chowdhury[2], Sultan Md Aslam[1], and Syed Md. Minhaz Hossain[1(✉)]

[1] Premier University, 4000 Chattogram, Bangladesh
samardas.cca@gmail.com, omlanhasan@gmail.com, smaslam199320@gmail.com, minhazpuccse@gmail.com

[2] International Islamic University Chittagong, Chattogram, Bangladesh
anu8chy@gmail.com

Abstract. Cardiovascular diseases are one of the world's significant health issues. It is becoming a major health issue in Bangladesh and other poor nations, particularly heart block. It is a condition in which the heart beats too slowly (bradycardia). The electrical impulses that command the heart to contract are partly or completely blocked between the top chambers (atria) and the lower chambers in this situation (ventricles).Therefore, computer-assisted diagnosis techniques are urgently needed to aid doctors in making more informed decisions. In this study, a deep learning model, you only look once (YOLOv4) backbone with CSPDarkNet53 is proposed to detect four classes including three types of heart blocks, such as, 1st degree block (A-V block), left bundle branch block (LBBB) and right bundle branch block (RBBB), and no block. We prepared a novel dataset of the patient's electrocardiogram (ECG) images. This dataset contains 271 images of Bangladeshi patients. The model's mAP@0.5 on test data was 97.65%. This study may also find application in the diagnosis and classification of block and heart diseases in ECG images.

Keywords: YOLOv4 · ECG · Heart block · Cardiovascular diseases

1 Introduction

Heart disease is a major reason for death worldwide. The term "heart disease" refers to a multitude of cardiac problems (CVDs) According to the World Health Organization (WHO), 17.9 million individuals died from CVD during 2019, responsible for 32% of all deaths worldwide. Heart attacks were responsible for 85% of these fatalities [12]. Low and middle-income nations account for over three-quarters of CVD mortality. Low- and medium-income countries accounted for 82% of the Seventeen million deaths occurring (before the age of 70) due to noncommunicable illnesses in 2015, with cardiovascular disease accounting for 37% [1]. A single component causes the majority of CVDs: heart block. Patients

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 981–990, 2023.
https://doi.org/10.1007/978-3-031-27409-1_90

with heart block are more prone to have heart attacks and other CVDs, both of which may be fatal. A "heart block" is an obstruction in the normal conduction of electrical impulses in the heart. Heart block is caused by natural or artificial degeneration or scarring of the electrical channels in the heart muscle [2].

In the medical industry, it is challenging to collect real-time data. Furthermore, although collecting the genuine ECG signal is difficult, collecting the scanned ECG picture and reprinted ECG image is quite easier. As there is no standard and authentic digital ECG record for Bangladeshi patients, it is one of our contributions to prepare a novel dataset on Bangladeshi patients. However, there has been little study on ECG data [11]. Several solutions to these problems are being explored. One option is to process mammography images using different computer-aided detection (CAD) technologies. Image processing methods based on deep learning and machine-learning are now among the most promising CAD design techniques [8,10].

Deep learning has already shown to be a useful approach for a variety of applications, including image classification [15,16], object identification [8,11] and segmentation, and natural-language processing [12–14]. Deep learning has also shown potential in medical image analysis for object recognition and segmentation, such as radiology image analysis for examining anatomical or pathological human body features [4,9,13,14]. Deep learning algorithms can extract comprehensive, multi-scaled data and integrate it to help specialists make final decisions. As a consequence, its applicability in a variety of applications for object recognition and classification tasks has been proven [5]. This resulted in a plethora of cutting-edge models that performed well on natural and medical imagery. These models progressed from basic Convolutional Neural Networks (CNNs) to R-CNNs, Fast CNNs, and Faster R-CNNs [6]. CNN-based CAD systems outperform traditional machine learning techniques for x-ray image identification and recognition on the examined datasets [3]. These well-known strategies have solved many of deep learning's problems. Most of these models, however, need a large amount of time and computational power to train and implement. However, training and implementing most of these models requires a significant amount of time and computer memory. As a consequence, You-Only-Look-Once (YOLO) has been identified as a quick object recognition model suited for CAD systems. YOLOv4 is a CNN-based one-stage detector that identifies lesions on images as well [5], with an accuracy of 80–95%. In this paper, we provide a YOLO-based model for an end-to-end system that can detect and categorize heart blockages. Our suggested model's key contributions, such as, are noted below.

(i) Prepare a novel dataset consisting of ECG images of Bangladeshi patients.
(ii) Utilize a deep learning model, YOLOv4 in order to increase precision in detecting heart blocks.

2 Related Researches

To get an understanding into a population's pattern of probability for a chronic disease-related adverse outcome, Song et al. [15] presented a hybrid clustering-ARM technique. The Framingham heart study dataset was utilized, and the

adverse event was Myocardial Infarction (MI, sometimes known as a "heart attack"). This approach was shown by displaying some of the generated participant sets, clustering procedures, and cluster numbers. The authors of [16] provided an overview of current data exploration strategies in databases that use data mining techniques being used in medical research, most notably in heart disease prediction. In this research, they experimented with Neural Networks (NN), K-Nearest Neighbor (KNN), Bayesian classification, classification using clustering, and Decision Tree (DT). They do admirably in DT (99.2%), clustering classification (88.3%) and Naive Bayes (NB) (96.5%). While classifying ECG data, Hammad et al. [8] compared KNN, NNs and Support Vector Machine (SVM) classifiers, to the suggested classifier. The suggested approach makes use of 13 different characteristics collected from every ECG signal. According on the results of the experiments, the suggested classifier outperforms existing classifiers and achieves the greatest average classification precision of 99%. Three normalization types, four Hamming window widths, four classifier types, genetic feature (frequency component) selection, layered learning, genetic optimization of classifier parameters, stratified tenfold cross-validation and new genetic layered training were combined by the authors of [13] to create a new system (expert votes selection). They created the DGEC system, which has a detection intensity of 94.62% (40 errors/744 classifications), a precision of 99.37%, a specificity of 99.66%, and a classification time of 0.8736 s. Authors in [11] investigated a variety of artificial intelligence technologies for forecasting coronary artery disease. The following computational intelligence methods were used in a comparative analysis: Logistic Regression (LR), SVM, Deep NN, DT, NB, Random Forest (RF), and KNN. The performance of each approach was assessed using Statlog and the Cleveland heart disease datasets, which were obtained from the UCI database and investigated using a variety of methods. According to the research, deep NN have a highest accuracy of 98.15%, with the precision and sensitivity of 98.67% and 98.01%, respectively.

3 Materials and Methods

In this work, the detection phase of ECG images is considered as an object detection problem. The methodology of our proposed method is as shown in Fig. 1.

3.1 Dataset

The ECG image dataset from the medical center is used in this work for training. This dataset includes 271 images of four different classes. There are approximately 6000 augmented images. From the Cardiology Departments of Chittagong Medical College and Dhaka Medical College Hospital, we have gathered information on a total of 271 patients. The samples of ECG images are shown in Fig. 2.

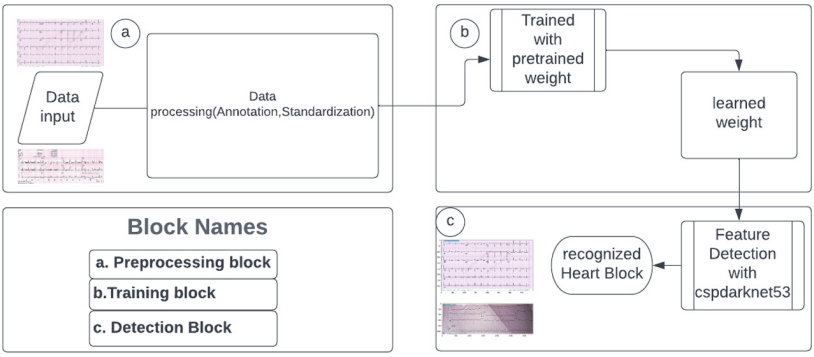


Fig. 1. Proposed method for detecting heart block from ECG images using YOLO.

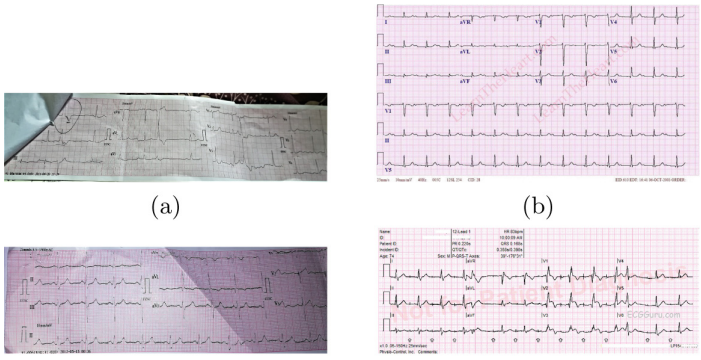


Fig. 2. Samples of four classes: **a** No block. **b** First degree block. **c** Left bundle branch block. **d** Right bundle branch block.

Data Annotation The annotating of data is required by many deep learning algorithms. Data annotation is finished with the aid of specialists. The ECG images have all been converted to YOLO format. Using data from the csv file provided with the dataset, the labeling software generates a text file for each image. A bounding box in the YOLO format is defined with the features associated to a class.

3.2 Pre-processing

Machine learning and deep learning algorithms both call for data cleaning. Scaling problems may occur if the data's pre-processing is done incorrectly. It also permits us to work within a set of limitations. The pre-processing methods are as follows:

(i) Normalization
(ii) Data augmentation
(iii) Image standardization.

Before using the raw images, the dataset has to be pre-processed to make it suitable for training. The raw images are resized to specific size (608 × 608), the image format are converted to a suitable format (as the dataset in DICOM image format), the contrast and brightness of images are adjusted, and noise filters are utilized to reduce noise in the dataset. Besides, re-scales the images to have a values of pixels in between 0 and 1. The augmentation techniques with parameters are shown in Table 1.

Table 1. Augmentation on our dataset

Augmentation technique	Factors
Contrast	0–4
Brightness	0–4
Saturation	1.5
Hue	0.1
Angle	0

3.3 Train-Test Split

The dataset is separated into two sections: training and testing. Training uses 80% of the pre-processed data, and testing uses 20% of the pre-processed data. Train and test image samples are shown in Table 2.

Table 2. Training and test datasets

Class	#Training images	#Test images
No block	51	13
1st Degree Block	55	14
Left Bundle Branch Block	61	16
Left Bundle Branch Block	49	13
Total	216	55

3.4 Detection Using YOLOv4

As majority of the obtained datasets have few samples and frequently have an unbalanced distribution, two approaches: data augmentation and transfer learning are utilized in our study to address this issue. The augmentation techniques with parameters are shown in Table 1. Another one is transfer learning based model YOLOv4 using CSPDarkNet53.

Model Dimension and Architecture The YOLO technique is a one-stage detector that predicts the coordinates of a specific number of bounding boxes with various properties, such as classification outcomes and confidence levels, rather than using a separate algorithm to construct regions. Then moves the boxes into position. The fully convolutional neural network (FCNN) construction is the foundation of the YOLO architecture. This method divides each entire image into N $\times$ N nets and returns B limiting frames for each net along with an evaluation of the class C's importance and probability [7]. The implemented YOLOv4 design, seen in Fig. 3, places the CSPDarkNet53 architecture at the input level. The CSPDarkNet53 is a CUDA- and C-based open source neural network framework.



Fig. 3. Backbone model architecture for detecting heart block from ECG images using YOLO.

Pre-trained Learning Transfer learning is a contemporary method used to accelerate convergence while deep learning algorithm training. It entails using a model that has already been trained on a separate dataset (for MS COCO). In our situation, only the layers responsible for low-level feature identification (first layer) were loaded with the pre-trained weights.

Model Selection In-depth and thorough testing with YOLOv4 for ECG image has been conducted in order to discover heart block. In our work, we utilize a variety of combinations and changes to achieve the best outcome for YOLOv4 network resolution. Models along with mean average precision (mAP) and F1-score at various iterations, were examined during testing in order to determine the combination that performed the best.

Training Setup The number of batches after which the learning rate grows from 0 to the learning rate in epoch 0 was defined as burn in, and it was set to 1000. The learning rate was set to 0.001 and burn in was set to 1000. At 0.949 and 0.0005, respectively, momentum and weight decay are set. Due to restrictions imposed by the GPU RAM available, batch size and mini-batch size were both set to 64. As a result, one epoch for our training set of 216 is equal to 216/64, which, when rounded to the next whole integer, results in 4 iterations. The loss and mean average precision (mAP) have stabilized after 6000 cycles of training the model. The hyper-parameters of our model are as shown in Table 3.

Table 3. Hyper-parameters of our model

Hyper-parameters	Factors
Epoch	20,000
Batch	64
mini-batch	64
Learning rate	0.001
Momentum	0.949
Weight decay	0.0005

4 Result and Observation

The most popular hosted Jupyter notebook service is Google Colaboratory. In comparison to the ordinary version, Colab Pro features faster GPUs, longer sessions, fewer interruptions, terminal access, and more RAM. On a Colab Pro with two virtual CPUs, an NVIDIA P100 or T4 GPU, and 32 GB of RAM, the experiment is conducted. The suggested model was created in Python and heavily utilizes Python modules.

4.1 Model Selection

We trained 216 ECG images of four classes. Figures 4 and 5 represent the effectiveness of various training iterations as measured by mean average precision (mAP) and F1-scores of test dataset. For 6000 iterations, it was discovered that the model produced the highest AP and F1 scores for test dataset.

4.2 Model Evaluation

We test 55 ECG images for detecting heart blocks. The performance of our model is evaluated on F1-score, Intersection over union (IoU) and mean average precision (mAP). The F1-score, and mean average precision (mAP) are calculated

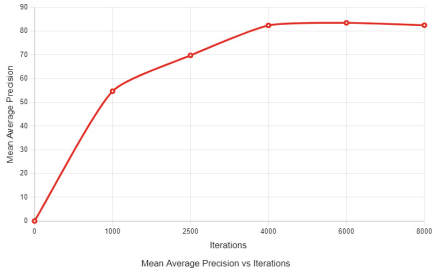


Fig. 4. Mean average precision versus iterations.

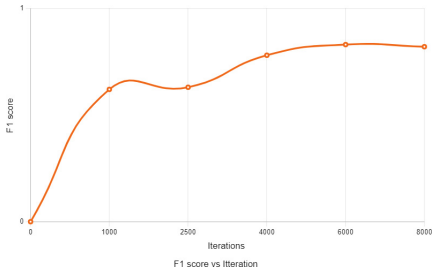


Fig. 5. F1-score versus iterations.

Table 4. Performance evaluation

Iteration	F1-score (%)	IoU	mAP (%)
1000	62	0.49	54.67
2500	63	0.63	69.76
4000	78	0.82	82.38
6000	84	0.83	83.45
8000	83	0.82	82.75

for each item class at 0.5 IoU. The performance of our proposed YOLOv4 model is as shown in Table 4.

Table 5 represents the F1-score, and mean average precision (mAP) of each class.

Table 5. Performance evaluation of each class

Class	F1-score (%)	mAP (%)
No Block	66	85.23
1st Degree Block	65	80.00
Left Bundle Branch Block	78	87.08
Right Bundle Branch Block	79	73.48

5 Conclusion and Future Work

The leading cause of death in the world is heart disease. Heart attacks and other CVDs, both of which have a high mortality rate, are more common in patients with heart block. One of the most promising CAD design methodologies nowadays is image processing based on deep learning. As there is no standard and authentic digital ECG record for Bangladeshi patients, it is one of our contributions to prepare a novel dataset on Bangladeshi patients. The mean average accuracy (mAP) and F1-scores are used to assess how well test data performed on different training iterations. It was found that the model produced the highest mAP and F1 scores across 6000 iterations. In future, we will increase the high volume of dataset of Bangladeshi patients and investigate the different YOLO versions for detecting the heart blocks accurately.

References

1. Statistics of CVD (2022). https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseasess
2. What is heart block (2022). https://www.webmd.com/heart-disease/what-is-heart-block
3. Al-antari, M.A., Al-masni, M.A., Park, S.U., Park, J., Metwally, M.K., Kadah, Y.M., Han, S.M., Kim, T.S.: An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network. J. Med. Biol. Eng. **38**, 443–456 (2018)
4. Alarsan, F.I., Younes, M.: Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. J. Big Data **6**(1), 1–15 (2019). https://doi.org/10.1186/s40537-019-0244-x
5. Baccouche, A., Zapirain, B., Elmaghraby, A., Castillo, C.: Breast lesions detection and classification via yolo-based fusion models **69**, 1407–1425 (2021) (CMC Tech Science Press). https://doi.org/10.32604/cmc.2021.018461

6. Baccouche, A., Zapirain, B., Elmaghraby, A., Castillo, C.: Breast lesions detection and classification via yolo-based fusion models. Cmc -Tech Science Press- **69**, 1407–1425 (06 2021). 10.32604/cmc.2021.018461
7. Bochkovskiy, A., Wang, C., Liao, H.M.: Yolov4: optimal speed and accuracy of object detection. CoRR (2020). arxiv:2004.10934
8. Hammad, M., Maher, A., Wang, K., Jiang, F., Amrani, M.: Detection of abnormal heart conditions based on characteristics of ECG signals. Measurements **125**, 634–644 (2018). https://doi.org/10.1016/j.measurement.2018.05.033
9. Hasan, N.I., Bhattacharjee, A.: Deep learning approach to cardiovascular disease classification employing modified ECG signal from empirical mode decomposition. Biomed. Signal Process. Control **52**, 128–140 (2019)
10. Li, R., Xiao, C., Huang, Y., Hassan, H., Huang, B.: Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: a review. Diagnostics **12**(2) (2022). https://doi.org/10.3390/diagnostics12020298, https://www.mdpi.com/2075-4418/12/2/298
11. N, J., A, A.L.: SSDMNV2-FPN: A cardiac disorder classification from 12 lead ECG images using deep neural network. Microprocess. Microsyst. **93**, 104627 (2022). https://doi.org/10.1016/j.micpro.2022.104627, https://www.sciencedirect.com/science/article/pii/S0141933122001648
12. Nahar, J., Imam, T., Tickle, K., Chen, Y.P.P.: Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst. Appl. **40**, 1086–1093 (2013). https://doi.org/10.1016/j.eswa.2012.08.028
13. Pławiak, P., Acharya, U.R.: Novel deep genetic ensemble of classifiers for arrhythmia detection using ECG signals. Neural Comput Appl **32**(15), 11137–11161 (2019). https://doi.org/10.1007/s00521-018-03980-2
14. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 520–527. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_65
15. Song, S., Warren, J., Riddle, P.: Developing high risk clusters for chronic disease events with classification association rule mining. In: Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management, vol. 153, pp. 69–78 (2014)
16. Soni, J., Ansari, U., Sharma, D., Soni, S.: Predictive data mining for medical diagnosis: an overview of heart disease prediction. Int. J. Comput. Appl. **17**(8), 43–48 (2011)



Attendance Automation System with Facial Authorization and Body Temperature Using Cloud Based Viola-Jones Face Recognition Algorithm

R. Devi Priya[1(✉)], P. Kirupa[2], S. Manoj Kumar[2], and K. Mouthami[2]

[1] Department of Computer Science and Engineering, Centre for IoT and Artificial Intelligence, KPR Institute of Engineering and Technology, Coimbatore, India
scrpriya@gmail.com

[2] Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India

Abstract. Face recognition works all over the world from various perspectives. In the field of attendance systems, many methodologies have faced various drawbacks. In the headset of attendance monitoring, the attendance is registered immediately by scanning the face and compares with pattern in the database and marking the student attendance along with the detection of their body temperature automatically by using object vision algorithms. In the proposed system, facial feature recognition and detection is performed based on the Viola-Jones face detection algorithm. The system is for school or university students with any large strength, and even a single student with at least 5–6 images with different angle of their face to be stored. Hence, these automation systems need a large amount of storage space. Hence, cloud storage server is also used to store any number of images. The student's attendance is recorded by the camera installed in the class entrance. Students have to record their faces one by one before entering the classes and the camera can create snapshots for a particular set of defined timings. Then, it detects faces in snapshot images, compares them with the cloud database and then attendance is marked. The experimental results show that the proposed Viola-Jones face detection algorithm is better than many existing algorithms.

Keywords: Face recognition · Automatic Attendance · Attendance Monitoring · Cloud server · Viola Jones algorithm

1 Introduction

The whole world became automated, and everything in the world became easier and faster. As automation has taken objects or things and made them interact with one another by connecting through the internet, even almost all manual processes have become automated, so the attendance-taking system in schools, colleges, or other institutions is being customized by face recognition technology.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 991–1001, 2023.
https://doi.org/10.1007/978-3-031-27409-1_91

The students have to face the camera, then their faces can be snapped and compared with the previous test image, and then their attendance has been marked along with their body temperature, and the respective time, date, department can be recorded. The attendance rate can be recorded and calculated by the system for every individual student. Later, according to the schedule, the admin and staff can generate a report for every student and mark the percentage of the academics who have the ability to attend the exams or do not, according to the university norms.

Practically, it is difficult for the lecturer to check the student's presence in each and every period by taking attendance. In this manual register, a lot of errors can occur, resulting in inaccurate or wrong records. A survey investigation of human posture estimation and movement acknowledgment late improvements utilizing multi-view data [1]. There are different fields where automatic face recognition has been implemented, like the Raspberry Pi, RFID, NFC, SVM, MLP, and CNN.

Smart phones are used in education in many applications. iOS or Android platform is used to support students in learning from lectures in many studies [2]. It is very commonly used for monitoring attendance. In the biometric system, the recorded fingerprints are compared with the stored patterns. But, there are a lot of reasons the system can fail if it is very sensitive and records sweat, dust and even some minor wounds. Many machine learning and deep learning algorithms are used for pattern recognition and data processing in many applications [3, 4]. With this inspiration, they are attempted in attendance recognition systems and they are successful in face recognition.

In the RFID system, which has a large number of students, purchasing tags for everyone is costly. In the LDA combined with SVM, the system has to take the images from the video recorded in the classroom. LDA is used for extracting the features from the face to decrease the inter-class dispersion by finding the linear transformation of that image. But it causes some problems when it causes facial extraction and SVM is usually used for face recognition.

The installation and use of RFID and Raspberry Pi are really costly, and the sensors are sometimes too sensitive, so the results can be less accurate. Manually taking attendance needs a lot of records and papers, and reviewing the recording process needs a long time. Hence, these attendance record documents need large storage spaces in physical form. There is a lack of fingerprint attendance systems.

Hence, it is understood that there are many practical difficulties in implementing automatic attendance monitoring and the paper proposes a novel method for addressing the issue. The paper suggests implementation of Viola-Jones face recognition algorithm and unlike other existing methods, it also suggests cloud based storage which is very efficient.

2 Literature Survey

Most of the existing systems record the students' face in the classroom and forwards them for further image processing processes. Later, after the image has been enhanced, it is moved forward to the Face detection and recognition modules. After the recognition, the attendance is recorded on the database server. The students have to enroll their individual face templates for storage in the Face database [5]. The suggested methods have tried

recognizing face expressions with Field Programmable Gate Array [6]. The camera in the classroom takes a continuous shot of images of students in the class for detection and recognition [7].

Using the face detection method proposed [8] manual work can be completely reduced. It is found that the proposed process showed performance enhancement and gave higher detection accuracy (95%) and doubled the detection speed as well.

In the other system, [9] LBPH is particularly used for the conversion of the color image to grayscale image to generate a histogram. Then the noise is removed from the image, and they use ROI to reshape the picture. In that system, the GSM is connected to that feature to send messages to the students who are absent, and the message is also forwarded to the student's parents mobile number. But the accuracy of the LBPH is only 89%.

In the previous system, SURF (Speed Up Robust Feature) algorithm [10] is used to recognize face. The systems make use of simple ideas in face recognition. In this system, the SURF algorithm is used for comparing the train images with the images that are taken during class time. Then it searches for the features that are the same in both images. Then it filters out and leaves a few points. If the points with the minimum distances are matched according to Euclidean distance, the trained data are stored in the database, which is created using MS-Excel. The process starts after the test image is taken during the class time. It can take students who are in front of camera in a defined time gap. Then it initialize after function call then it places the interest point in the test image and then it locates filter interest points if the is recognized it can make a record in the Excel. Then the loop is continued. SURF is performing well in face recognition. The limitations of other techniques like LDA, SIFT, and PCA are beaten by these SURF methods. Special properties like scale invariant and orientation of image are very practically used in real-time face recognition.

Some literatures have used face recognition for monitoring attendance with the help of deep transfer learning algorithm [11]. Since, it contains pretrained models, it was performing better than other existing methods and the results were found to be satisfactory.

Principal Component Analysis for detecting features and Opencv framework for recording electronic attendance have been implemented with high quality and information assessibility [12].

The other system have implemented attendance tracking system using mobile phones enabled with GPS and Near Field Communication (NFC) [13]. But, In the NFC technique, it works only over 10–20 cm or fewer distances, and very low data transfer rates.

By analyzing literature, it is understood that still there are many unresolved issues in existing methods and there is a large scope for introduction of more efficient methods for face detection and recognition.

3 Proposed Work

The paper proposes to achieve effective face recognition and attendance by applying Viola-Jones face detection algorithm. The system capture the faces of students who

are all entering into the classroom. These systems then take individual images of each student to be stored in the database, where these images can be used for comparison.

Here, cloud storage is used for storing large amounts of data; therefore, there is no file compression. Once the application is installed, the admin has to enter the student's details along with the access privileges for the staff who want to generate the report for each individual. The flow diagram of the proposed systems is shown in Fig. 1. They also monitor the body temperature of each student and let them know if they have any medical expenses.

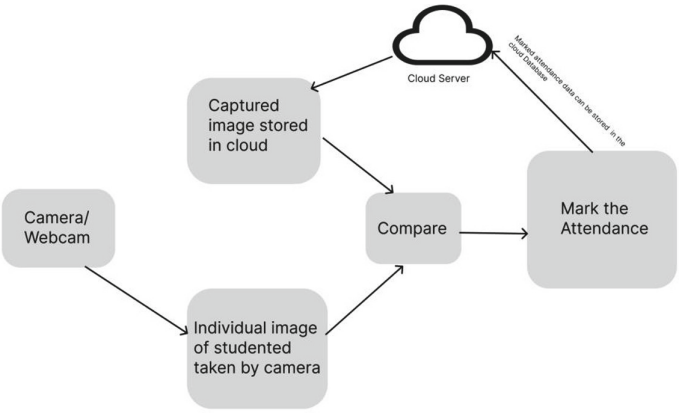


Fig. 1. Block diagram of processes involved in attendance monitoring system.

It can also provide the time and date of each student's entry. The staff can easily calculate the attendance percentage for each student. The staff can also monitor the student temperature. The information can be stored in the cloud server.

MODEL PROTOTYPE

This system process can start with collection of data from the camera and followed by various processes for recognizing the faces from the stored database and the image captured by the camera. If the system finds the face, then the system can mark that individual student's attendance. The use case diagram of these systems has been described in Fig. 2.

3.1 Data Collection

This system proposes that when the student enters the class, his or her face and body temperature can be automatically recorded. The webcam/camera can record the face of the student, then make snapshots with it to detect the face along with face thermal recognition. Likewise, whenever the student enters, the system can record the face to recognize the dataset.

3.2 Data Preparation

Preparation of the dataset is the main and initial step in these systems. The admin/staff has to save the dataset for each and every student image along with their respective

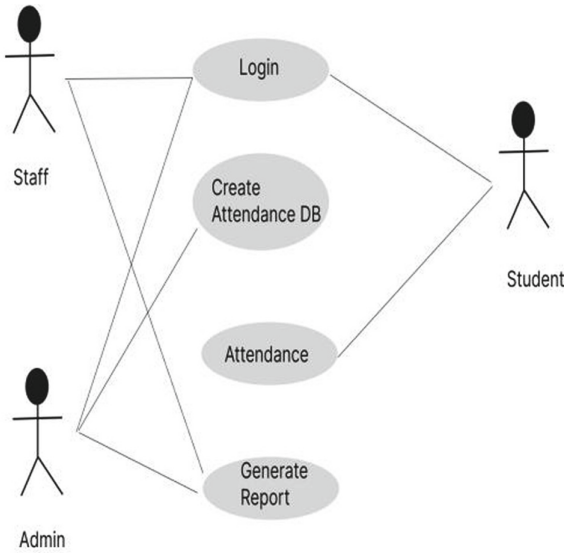


Fig. 2. Use case Diagram for these System

name, roll number, department and course. Each student have at least 3–4 images with their different face angles. Hence, these have to be stored in individual folders for each student as shown in Fig. 3. Here, this huge amount of data storage can be less problematic because we are using the cloud for storage. Each and every time a student enters the system, it has to check for the faces that are already in the dataset and mark attendance if the faces are matched.

4 Face Recognition

There are several steps to recognize face in the database, also called "Trained Image," with the image that is newly captured "Test Image". The steps that have been used in the proposed system are described below:

i. **Face Detection**

Here, the image that is snapshotted by the webcam or other external camera when the student enters the class is recognized and detected to mark the faces and to locate the bounding boxes and the coordinates of these pixels and mark them.

ii. **Face Alignments**

In this step, the face image will be normalized. Because the images captured by the camera have different tones, they are changed into grayscale images for image enhancement. In these systems, Histogram normalization enhances the contrast in the image. This can

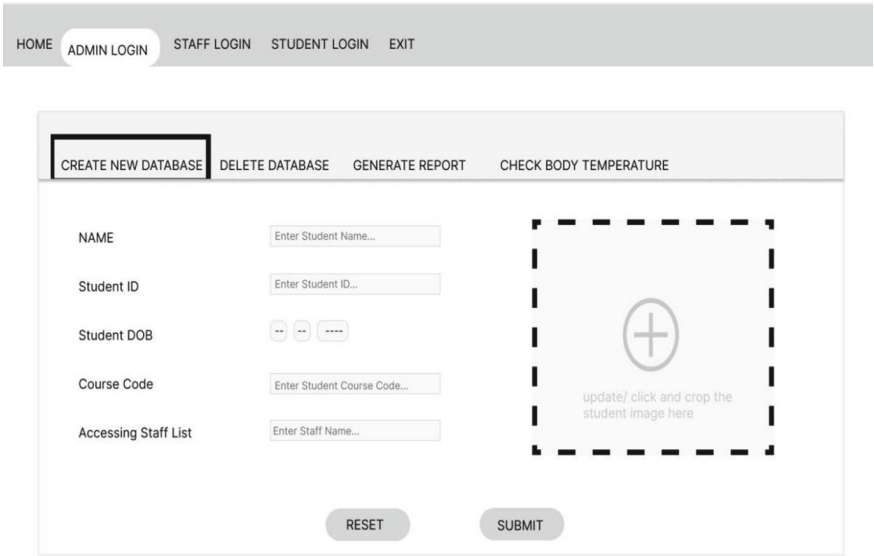


Fig. 3. Admin Login Page (the admin can create the new student dataset here)

be done for removing noise and smoothing the image, like FFT, low-pass filtering or by the median filter.

Here Local Binary Pattern Histogram (LBPH) algorithm is combined with the HOG descriptor, which defines the face of the image as one data vector; later, it is used for the face recognition process.

iii. **Face Extraction**

The Insert Object Annotation, which returns a rectangular shape image annotation with shape and label. The face can be recognized by its distinctive facial features, including parts like mouth, nose, left eye, and right eye.

Hence, the system can find the face by comparing these facial features with the trained and test images. The Step function performs multi-scale Object detection and returns “B boxes” with 4 matrices.

iv. **Face Recognition**

This step is used to recognize various unique facial structures like nose, mouth and eyes by using the Viola-Jones face detection algorithm. The below algorithm defines how the face recognition is done step by step for unique facial features using Viola-Jones face detection algorithm. The below figure shows how the individual features have been broken down step by step for face detection Using object vision.CascadeObjectDetector System. The steps in identifying the features of face are given in Fig. 4.

Processes like feature detection, matching and extraction are completed in sequence. Matching of the face is done with one or more known faces in a cloud database. The system can recognize the images in the trained images dataset and new images with the specified facial features. Features of face are detected by using two different methods namely Histogram of Oriented Gradients (HOG) and Local Binary Pattern Histogram (LBPH). The flow diagram of the Viola-Jones face detection algorithm is diagrammatically shown in Fig. 5.

```
Input: original test image
Output: image with face indicators as rectangles
for i ← 1 to n images do
  Downsample image to create image_i
  Develop integral image, image_ii
  for j ← 1 to n of shift steps of sub-window do
      for k ← 1 to n of stages in cascade classifier do
          for l ← 1 to n of filters of k stages do
              Filter the detection sub-window of images
              Accumulate filter outputs
          end for
          if accumulation does not satisfy per-stage threshold then
             Reject sub-window's face
             Break
          end if
      end for
      if sub-window pass all per-stage checks then
         Accept the sub-window as a face
      end if
  end for
end for
```

Fig. 4. Steps in feature detection

v. **Body Temperature Detection**

To detect body temperature, a thermal image is used. By analyzing the grayscale image that was obtained after the Histogram normalization in order to find the pixel value with high intensity in the thermal region.

4.1 Attendance Marking

PCA algorithm is used for marking the students' individual attendance [10]. The system finds the matched face pattern in the database and then updates the new information in the log table and makes the individual student attendance along with the system time, which we have considered the entry time of that particular student.

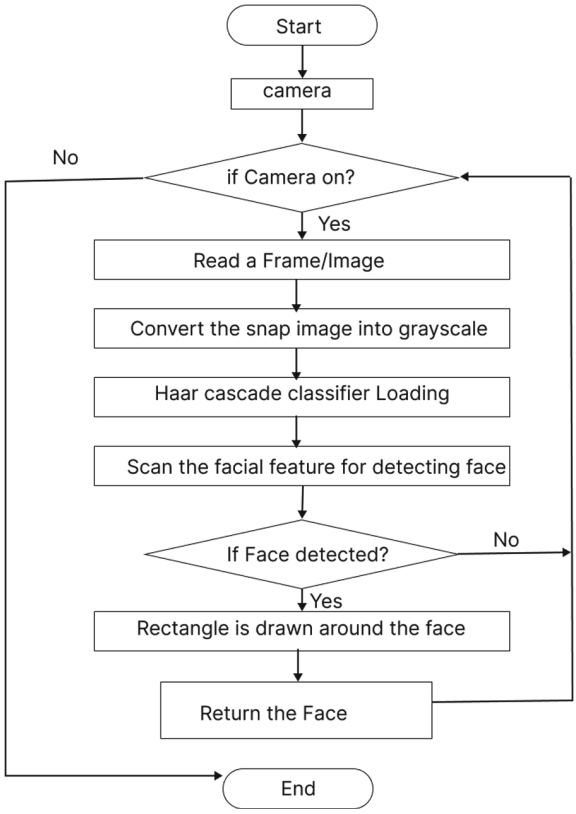


Fig. 5. The flow diagram of Viola-Jones face detection algorithm

4.2 User Interface

The admin or staff has to login into the system. Then they can view the report of the each student.

Admin: who can create or delete student information and give staff access to the student report.

Staff: who can view the report for whole or individual student attendance data in their class and can also calculate the attendance percentage for them.

The database with the school student images which are stored in the cloud storage server the images can be compared. Then by the thermal analysis the body temperature of the student can be identified. In the User Interface, the attendance marked details can be displayed. The system can describe how the staff can view the reports of students. The output can be displayed according to the filters and other options. Similarly, if the staff wants the list of students who are all present in the particular class, attendance can be registered.

Even the staff can get a list of particular students who are all attending their classes with the entry time, date, time, body temperature, and attendance percentage for that particular class.

In which the single student attendance details for a particular subject have been entered in the filter tab and the results are displayed according to the requirements.

5 Experimental Results and Discussion

The proposed algorithm is validated with experiments using the collected images of 500 students from an educational institution. By using this algorithm, the system can recognize the images with more accuracy and speed. The data for these images is stored in the cloud and hence the storage problem is reduced. The admin or the staff have to enter the student details for the first and they can update whenever they need. The performance results of the method proposed are compared with that of the existing methods like SURF, CNN and SVM as given in Table 1.

Table 1: Performance comparison of methods

Algorithm Used	Facial Feature Recognition Technique	Classification Accuracy (%)
SURF	–	90.2
CNN	–	88.4
SVM	PCA	55.9
	LDA	57.7
VIOLA-JONES	HOG	94.2
	LBPH	95.3

In addition to classification accuracy, other measures like precision, recall and F1 measure are also recorded and the results are given in Table 2. The results show that implementation of Support vector machine classifier using feature selection methods like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) show the leas performance. CNN, the Deep learning algorithm and SURF algorithm show comparatively better performance than the proposed Viola Jones face recognition algorithm. Feature selection using HOG and LBPH are contributing more for the improved performance of the proposed method.

The overall working status of proposed system along with the image snapped time is described in the Table 3.

The proposed Viola-Jones face detection algorithm is applied for live facial recognition in digital cameras and it is very faster at face detection as compared with other techniques with better accuracy.

In this algorithm, the CascadeObjectDetector detector in the computer vision system toolbox uses only the quick and efficient features that are marked in the rectangular region of an image, whereas the SURF algorithm has thousands of features that are much more complicated.

Table 4 compares the execution time of these algorithms to detect and recognize the input faces. The experimental results show that when compared to all algorithms,

Table 2. Comparison of Precision, Recall and F1 measure

Algorithm Used	Facial Feature Recognition Technique	Precision	Recall	F1 Measure
SURF	–	0.85	0.74	0.71
CNN	–	0.87	0.71	0.74
SVM	PCA	0.67	0.59	0.66
	LDA	0.73	0.65	0.69
VIOLA-JONES	HOG	0.94	0.93	0.91
	LBPH	0.95	0.96	0.93

Table 3: Sample output matching using Viola-Jones algorithm

Image snapshot time	Image detected	Image segmented	Image matching	Matched student ID	Attendance marked
09:27:20:94	yes	yes	yes	CS005	yes
09:27:21:19	yes	yes	yes	CS025	yes
09:27:21:56	yes	yes	yes	IT016	yes
09:27:21:88	yes	yes	yes	CS007	yes
09:27:22:64	yes	yes	yes	CE005	yes
09:27:23:43	yes	yes	yes	EC009	yes
09:27:23:90	yes	yes	yes	CS012	yes
09:27:24:20	yes	yes	yes	ME045	yes
09:27:24:90	yes	yes	yes	CS001	yes

the proposed Viola Jones algorithm completes the task faster than other algorithms. It is because of the fact that the feature selection algorithms used selects the significant features faster and hence the classification process is better than other methods.

Table 4: Comparison of Execution time of all algorithms

Algorithm Used	Facial Feature Recognition Technique	Execution Time (ms)
SURF	–	97
CNN	–	86
SVM	PCA	112
	LDA	149
VIOLA-JONES	HOG	230
	LBPH	74

6 Conclusion

The proposed face recognition system used for generating attendance records has better performance for marking the presence/absence of each student who enters the class along with their leaving time. This proposed system can minimize the time and effort of the staff who is entering and maintaining individual student attendance.

Here, by using the database stored in the cloud, addition and deletion of students can be done easily and does not need a large number of hardware disks for storage. The system implements feature selection methods like HOG and LBPH thereby improving accuracy of the result. In this automated system, the student's attendance can be marked by the institutions without any human error while recording attendance. It can save time and work for students, staff and the institution.

References

1. Holte, M B.: Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. IEEE J. Sel. Top. Signal Process. **6**(5) (2012)
2. Douglas, A., Mazzuchi, T., Sarkani, S.: A stakeholder framework for evaluating the utilities of autonomous behaviors in complex adaptive systems. Syst. Eng. **23**(5), 100–122 (2020)
3. Sivaraj, R., Ravichandran, T., Devi Priya, R.: Solving travelling salesman problem using clustering genetic algorithm. Int. J. Comput. Sci. Eng. **4**(7), 1310–1317 (2012)
4. DeviPriya, R., Sivaraj, R.: Estimation of incomplete values in heterogeneous attribute large datasets using discretized Bayesian max-min ant colony optimization. Knowl. Inf. Syst. **56**(309), 309–334 (2018)
5. Duan, L., Cui, G., Gao, W., Zhang, H.: Adult image detection method base-on skin color model and support vector machine. In: ACCV2002: The 5th Asian Conference on Computer Vision, 23–25 January (2002), Melbourne, Australia
6. Lin, J., Liou, S., Hsieh, W., Liao, Y., Wang, H., Lan, Q.: Facial expression recognition based on field programmable gate array. In: Fifth International Conference on Information Assurance and Security, Xi'an, (2009), pp. 547–550
7. Xu, X., Wang, Z., Zhang, X., Yan, W., Deng, W., Lu, L.: Human face recognition using multi-class projection extreme learning machine. IEEK Trans. Smart Process. Comput. **2**(6), 323–331 (2013)
8. Godara S.: Face detection & recognition using machine learning. Int. J. Electron. Eng. Int. J. Electron. Eng. **11**(1), 959–964 (2019)
9. Arjun Raj A., Shoheb, M., Arvind, K., Chethan, K.S.: Face recognition based smart attendance system. In: 2020 International Conference on Intelligent Engineering and Management (ICIEM), pp. 354–357 (2020)
10. Mohana, H.S., Mahanthesha, U.: Smart digital monitoring for attendance system. In: International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering, pp. 612–616 (2020)
11. Alhanaee, K., Alhammadi, M., Almenhali, N., Shatnawi, M.: Face recognition smart attendance system using deep transfer learning. Procedia Comput. Sci. **192**, 4093–4102 (2021)
12. Muhammad, A., Usman, M.O., Wamapana, A.P.: A generic face detection algorithm in electronic attendance system for educational institute. World J. Adv. Res. Rev. **15**(02), 541–551 (2022)
13. Chiang, T.-W., et al.: Development and evaluation of an attendance tracking system using smartphones with GPS and NFC. Appl. Artif. Intell. **36**, 1 (2022)



Accident Prediction in Smart Vehicle Urban City Communication Using Machine Learning Algorithm

M. Saravanan(✉), K. Sakthivel, J. G. Sujith, A. Saminathan, and S. Vijesh

Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore 641407, India
sarvan148@yahoo.com

Abstract. The severity of traffic accidents is a serious global concern, particularly in developing nations. Recognizing the main and supporting variables may diminish the severity of traffic collision. This analysis identified the most insightful significant target specific causes for the severity of traffic accidents. The issue of road accidents has had an effect on both the nation's economy and the general welfare of the populace. Creating accurate models to pinpoint accident causes and offer driving safety advice is a vital task for road transportation systems. Models are developed in this research effort based on the variables that affect accidents, such as the weather, causes, characteristics of the roads, conditions of the roads, and accident types.This analysis identified the most insightful significant target specific causes for the severity of traffic accidents.By analysing through datasets which contains a massive amount of data,the process is made.VANET has been used to go through the vehicle communication process where comparing to other existing algorithms, our algorithm ensures high level vehicle communication.Both the number of vehicles and the number of individuals using them have grown throughout the course of the year. As a result, there are more accidents and more mortality. Supported machine learning methods for predicting the route discovery and implementing effective vehicular communication. Our method also aids in the forecasting of traffic accidents and works to prevent their occurrence in urban locations.We use various machine learning algorithms for different road accident causes. In this paper Random forest algorithm is proposed based on machine learning approaches that is expected to estimate the probability for vehicle get accident in urban area communication. This algorithm compares its result with existing conventional algorithm and gives improved throughput and reduced latency.

Keywords: Machine learning · Routing algorithm and random forest

1 Introduction

Over a period of time the mode of transport have evolved in the recent times we use vehicles to move from one place to another. Companies produce vehicles at low budget so all set of people can buy vehicles even though the roads are good the road accident get

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1002–1012, 2023.
https://doi.org/10.1007/978-3-031-27409-1_92

increased. To avoid this we have to find the reasons for the road the road accidents.This paper aims to give solution for the road accidents. To give the solution the same type of accidents are grouped together. With the similar type of accident (like analyzing road, external factors, driver) we can avoid this incident by using Machine Learning algorithms. We use logistic regression to produce solution to the accident. The purpose of using logistic regression is, it will analyze all the reasons for the accident. In the result we will know that different types of accident has different type of solution [1].Over the past few years the road accident cause major problem to the society. Due to accident more number of people die and it stands in ninth place for causing death to the human-kind.Now it becomes the major society problem which government need to take action for the citizen's happiness. We use Machine learning algorithms because it will analyze the accident more deeply. With the help of decision Tree, K – Nearest Neighbors (KNN), Naive Bayer's and adaboost we can find the reasons for the road accident. The accidents are categorized into four and they are fatal, Greivous, simple injury and motor collision, with the help of this four learning technique we can classify the seriousness of the accidents. Among the four learning techniques Adaboast is the best which can analyze the accident more deep [2].

This reason for the research is to find the reason how the accident happens on the black spot. In the result we can increase the safety of the drivers, which means we can decrease the number of accidents which is yet to happened. The reach which is going to done in the PalangkaRanya – Tangkiling national road, over the period of years the number of vechicles using Palangka Raya – Tangkinling national road is increased and the accidents in the national road also increased. The black-spot is located in this road.The main cause of accidents are traffic volume and the drivers characteristics. There is a two major type of accidents which happen in the Palangka Raya – Tangkiling national road and they are rear-head collision and motor cycle [3].One of the main social problem is the road safety. Over the period of years the road accidents have increased. To provide road safety we have to analysis the factors which are influencing the road accident by which we can provide solution to the road safety. Some of the major influencing factors are the external condition of weather condition driver etc. By grouping the similar type of accident we can easy provide solution to the problem. Here we use logistic regression to analysis the accident in depth with the help of the logistic regression we can find the which factor is the major reason for the accident.In the result we will know that each accident type has different solution with the help of Machine Learning algorithms we can provide solution to the same type of accidents [4].

Road accidents are the major communal issue for every countries in the world. This paper shows the purpose to estimate the number of traffic and road accidents that is mainly depending on a series of variables in the country Romania from 2012–2016. Which includes collision mode, road configuration,condition of occurrence, road category and type of vehicle which is involved, personal factors like in experience, lack of skills etc. and length of the time of driving license. It helps to identify the major cause of the accidents, road safety performance measures and risk indicators from the analysis of road accidents. A framework is suggest for improvising the road safety system and to reduce accidents by having these identified data. The data used for this analysis are provided by the Romanian Police. The National Institute of Statistics(NIS) in Romania

and the European commission [5]. Over the past few years road accidents are the major issue all over the world. The reason for the research is to analyse the factors which influence the road accidents such as road condition, external factors which causes the accident, effect of driving license duration, vehicles which involves in the accident,road conditions and the influence of the weather.By the analysis of the above factors we can provide solution for the road accidents or we can provide the framework.By which we can save the number of lives. To build the framework we have to analysis the previous accidents. The previous accidents data are taken from the police department. In the given data the above mentioned factors are analysed. This paper will provide solution for the road accidents and will save many lives. As a result the common factor will influence the many number of accidents [6].

The WHO said that over 1.35 million people die in the road accident every year. Accidents are the one of the root issue in many countries. In the worlds death people who dies on the accident is in ninth place. As a whole we can see that the accidents will cause economy problem too. By analysing the road accident we can build the framework by which we can give solution to the road accident problem. Here we use logistic regression to analyse the road accidents in depth. Some of the external factor which influence the road accidents are weather conditions, Road conditions.In this framework we use Logistic Regression (LR), K-Nearest Neighbor(KNN), Naïve bayes (NB), Decision Tree(DT) and Random forest(RF). The above algorithms analysis the accident in-depth and helps to avoid the accidents. The Decision Tree (DT) is the best algorithms which has the occurance of 99.4%.By analyzing all the factors with the above algorithm decision tree provided that it has the better occurancy with 99.4% [7]. In the recent year's road accidents becomes the major problem all over the world. Many peopledied through the road accidents. Now it become a social problem all over the world. The presice analysis is done to avoid the road accidents. Here we use Machine Learning algorithms to givesocial to this social problem. By analyzing the road accidents with Machine Learning algorithms we can provide solution to the road accidents. Here we use two supervised learning neural network and Adaboas. The supervised learning techniques classify the accidents into four categories and they are Fatal, Grievous, Simple injury and motor collision [8]. Accident Prediction model can be used to identify the factors which contributes largely.Using Artificial Neural Network,main accident factors are successfully analysed and Measurements are taken accordingly to prevent the crashes.The performance of the Artificial Neural Network are validated by the MAPE-Mean Absolute Percentage Error [9]. Accident Prediction targets to predict the probabilities of crashes within a short-time duration.This proposes a long and short term memory which are LSTM and CNN.LSTM gathers the long term data whereas CNN features the time invariance.Various types of data are used to predict the crash risk.Different data techniques are also applied.This proposal indicates the Power shot performance of using LSTM and CNN to predict the accidents [10].

2 Related Work

The current method of calculating the benefit of accident reduction uses predetermined accident rates for each level of road. Therefore, when assessing the effect of accident reduction, the variable road geometry and traffic characteristics are not taken into

account. Models were created taking into account the peculiarities of traffic and road alignments in order to overcome the challenges outlined above. The accident rates on new or improved roads can be estimated using the developed models. The first step was to choose the elements that affect accident rates. At the planning stage of roadways, features including traffic volumes, intersections, linking roads, pedestrian traffic signals, the presence of median barriers, and lanes are also chosen depending on their ability to be obtained. Based on the number of lanes, the elevation of the road, and the presence of median barriers, roads were divided into 4 groups for this study. For each group, the regression analysis was carried out using actual data related to traffic, roads, and accidents [11].

Accidents are the primary problems facing the world today since they frequently result in numerous injuries, fatalities, and financial losses. Road accidents are a problem that has impacted the general public's well-being and the economy of the country. A fundamental task for road transportation systems is to develop precise models to identify the cause of accidents and provide recommendations for safe driving. This research effort creates models based on the factors that cause accidents, such as weather, causes, road characteristics, road conditions, and accident type. Likewise, select a number of significant elements from the best model in order to build a model for describing the cause of accidents. Different Supervised Machine Learning techniques, such as Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), and Random Forests (RF), are used to analyse accident data in order to understand how each factor affects the variables involved in accidents. This results in recommendations for safe driving practises that aim to reduce accidents. The results of this inquiry show that the Decision Tree can be a useful model for determining why accidents happen. Weather, Causes, Road Features, Road Condition, and Type of Accident were all areas where Decision Tree performed better than the other components, with a 99.4% accuracy rate [12].

The ninth most common cause of mortality worldwide and a major problem today is traffic accidents. It has turned into a serious issue in our nation due to the staggering number of road accidents that occur every year. Leading its citizens to be killed in traffic accidents are completely unacceptable and saddening. As a result, a thorough investigation is needed to manage this chaotic situation. In this, a deeper analysis of traffic accidents will be conducted in order to quantify the severity of accidents in our nation using machine learning techniques. We also identify the key elements that clearly influence traffic accidents and offer some helpful recommendations on this subject. Deep Learning Neural Network has been used to conduct the analysis [13].

Accidents involving vehicles in foggy weather have increased over time. We are witnessing a dynamic difference in the atmosphere irrespective of seasons due to the expansion of the earth's contamination rate. Street accidents on roads frequently have fog as a contributing element since it reduces visibility. As a result, there has been an increase in interest in developing a smart system that can prevent accidents or the rear-end collision of vehicles by using a visibility go estimate system. If there is a barrier in front of the car, the framework would alert the driver. We provide a brief summary of the industry-leading approach to evaluating visibility separately under foggy weather situations in this document. Then, using a camera that may be positioned locally on a moving vehicle

or long separation Sensors or anchored to a street side unit (RSU), we describe a neural system approach for analysing visibility separations. The suggested technique can be developed into an intrinsic component for four-wheelers or other vehicles, giving the car intelligence [14].

In the US, motor vehicle accidents result in an average of over 100 fatalities and over 8000 injuries daily. We offer a machine learning-powered risk profiler for road segments utilising geospatial data to give drivers a safer travel route. In order to extract static road elements from map data and mix them with additional data, such as weather and traffic patterns, we created an end-to-end pipeline. Our strategy suggests cutting-edge techniques for feature engineering and data pre-processing that make use of statistical and clustering techniques. Hyper-parameter optimization (HPO) and the free and open-source Auto Gluon library are used to significantly increase the performance of our model for risk prediction. Finally, interactive maps are constructed as an end-user visualisation interface. The results show a 31% increase in model performance when applied to a new geo location compared to baseline. On six significant US cities, we put our strategy to the test. The results of this study will give users a tool to objectively evaluate accident risk at the level of road segments [15]. In the year 2030 car crashes going to be the 5th causing death for the humankind.There are many reason for the car crash, Some of them are very complex reason like the drivers mindset, the road vehicle is going on, and the climate in which the vehicle is going on. To give solution to the road accidents we use Machine Learning methods to analysis the causefor the accident. In Machine learning there are different algorithm models have we use logistic regression to analysis the accident in depth. In the end will know that each accident group has different with the Machine learning algorithms we can provide solution to the road accidents and save's many lives. Machine learning models takes up a deep analysis of the details gathered from the accidents. A deep study also should be made about the road accidents like identifying the speed of the vehicle and identifying the type of vehicle. The data requirements for the Machine learning model may vary about the algorithms which we use [16].

In India, road accidents causes the innocent lives to the major loss of events.To prevent the road accidents and making it in control has been a crucial task.So,to prevent this we majorly focus on accident prone areas.This model targets the causes of accident prone areas considering the factors. To solve this, here we use the Data Mining and Machine Learning concept-K to identify the causes and to take the resolution for them. Data mining techniques analyses the parameters such as number of occurences of accidents in the accident prone areas, time zone when major accidents occur, the regularity of accidents in that particular area. These data mining techniques may be used to developing the Machine learning models for road accidents prediction [17]. Over 1.32 lakh people died in the accident in the year of 2020 an it is the lowes count in the last decade. Even though the driver drives the vehicle very carefully there is a high probability that accident could happen. So we use Machine learning to reduce the accidents.First we analysis the reason for the accident with the logistic regression algorithm. Because it analysis in-depth. Finally we know that each group has different solution. So here we can save number of lives from the road accidents Machine learning models also collects information about the accidents and gives out reason for the accidents like weather and road situations using the decision tree algorithm. Decision tree algorithm is of neural network technology that

it considers all the possibilities and analyses every parameters of details gathered in the accidents. Decision tree model can be a accurate model to predict the reason and causes for the accidents [18].

Road accidents are one of the major problems faced among countries. The Romanian Cops, the National Institute of Statistics (NIS) in Romania, and the European Commission has given the data that to be used for the analysis. These data are analysed and evaluated considering the constraints.This paper will provide an informationof road accidents in the form of image and will implement a tool or framework for decreasing effects in road transport. As a outcome of analysis, we have concluded that the combination of vehicles and personal factors are the constraints that influences the number of traffic and road accidents Also this provides outline about guidelines to road accidents effects in the road transport. This framework helped out the Romanian cops to identify the cause of the road accidents and found out the solution to reduce the effects of road accidents [19].

Now a day's road accidents are the major causing deaths. In India many innocent people were died because of the road accident. It is very complex to control road accidents. The aim of this paper to predict the reason behind the accident causing factors by using the data mining technique-apriori and machine learning concept. The use of apriori technique will result in predicting time zone where the accidents occurs and peak accident time in that particular area using the Machine learning concepts. This model also tries to provide recommendations to minimalize the number of accidents. Machine learning also predicts the severity of accidents using different data mining techniques to predict cause for the accidents [20].

3 Proposed System

Machine learning approaches supported for predicting routing path and successful communication between vehicles in vehicular communication. This proposed system also helps to predict road accident and tries to avoid occurrence accident in urban environment. A cluster's points are all closer together. More far from any other than they are from their centroid. The K-means technique's main objective is to reduce the $D(C_i,E_j)$ between each object's Euclidean distancein relation to the centroid.Intra-cluster variance as a resultcan be decreased, and the similarity between clusters can rise. The squared error function was represented in Eq. 1.

$$f(x) = \sum_{i=1}^{k} \sum_{j=1}^{n} D\left(C_i,\ E_j\right) \tag{1}$$

Over the period of year's vehicle has increased and number of people using the vehicle also increased. This cause more accidents and many people die due to the accident. Now the accident have become major social problem all around the globe which cause many dead across the globe. In the intention to increase the saftey of the driver's We use ML (Machine Learning) approach to give solution to this problem we can use different algorithms Here we can use LSTM-CNN, it is a combination of LSTM and CNN many factor's influences the crash or accident and same of them are whether condition, signal

timing's and other external condition's. LSTM combine with CNN will analyse all this factor in-depth. By this we can predict that the accident that is going to happen and we can provide solution to the accident.

Development of Rap

The current road accident prediction is for only the particular road conditions. We cannot apply if for all the road conditions we use various Machine Learning algorithm to provide solution to this problem. In this particular case we use algorithms to analyse the road conditions like the alignment, traffic on the road, road condition etc. By this model we can provide solution to the accidents on the all road conditions. To provide solution we analyse the external factor which influences the accident like damaged road, whether the road in village or city in village the rate of traffic is less compare to the city, signals located, turning point, connection between the roads. By analyzing this external condition we can find the solution for the all accident types. By grouping the roads into different groups. We analyse all the groups which the regression method. Regression is used because it will analyze the groups in-depth. By the help of regression algorithms we can avoid accident on all type of roads. In the year of 2030 the car accidents going to cause the major death across the world, and it is going to stand in the place of fifth of causing the death. It is one of the major social problem there are many factors causing the accident such as the psychological factors of the driver or the drivers mindset, and the others external factors like condition of the road and the environment condition like weather, raining etc. To avoid accident we use machine learning algorithms to analyze the factors causing the accident. Here we use linear regression to analyze the factors. The linear regression is used because it analyze the factors in-depth. By the help of the linear regression we could provide solution for the road accident and save many lives.

3.1 Random Forest Approach

Breiman and Adele Cutler's ensemble classification technique, known as random forest, focuses mostly on creatingTo create uncorrelated decision trees, use numerous trees. One of the reliable algorithms to forecast a large number is this one, datasets, etc.decision trees are mostly prone to overstating, howeverto minimize overstating, random forest employs numerous tresses.The random forest produces numerous superficial, randomsubgroup trees, then aggregate or merge subtrees todo not overfit.

Additionally, when used with huge datasetsdelivers more accurate forecasts and is unable to give up itsaccuracy even when there are numerous missing data. Multiple Decision Trees are combined using Random Forest during training.Takes the sum of it to construct a model. Consequently, weakCombining estimators results in better estimates. Despite someif the decision trees weaken, there is a general desire. The output findings are typically precise.

3.2 Proposed Approach

Road accident data are now kept in a sizabledatabase storage. The datasets are made up of a lot of data. Complexity of the training and testing phase increases andestimating effectiveness consequently, it requires a strong model.

To get around or reduce the complexity of a massive numberdata set. We created a K-Means and random forest hybrid. To improve the effectiveness and accuracy of the prediction model, use the forest model to obtain a better efficient one. Typically, K-means is an unsupervised machine.Finding related groupings is the major use of the Learning method.Throughout the dataset. Despite the fact that this is an unattended with the k-means technique, the performance of the classifier can be enhanced by adding additional features to the training set. Clusteringa cluster feature is made, then it is included to the training set. Thenusing a random forest on clustered training data, Assess the RTA's severity. The result of that combo will bestrong model for making predictions in terms of generalization ability and predictive precision.

4 Results and Discussions

Throughput and latency are the two parameters that we considered for measuring the performance of network and concluding decision for predicting the accident. When accident is avoided it increases throughput and reduces latency (Fig. 1).

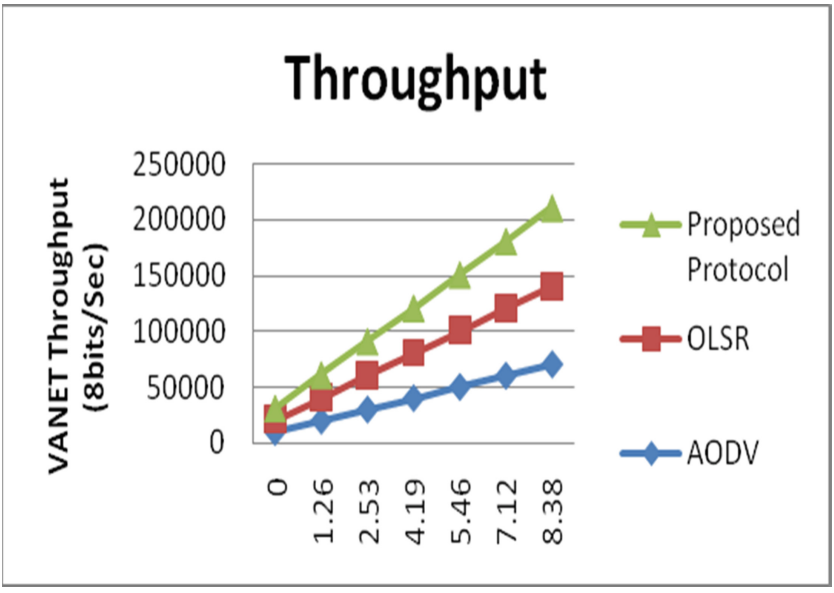


Fig. 1. Throughput vs time

Throughput ensures the number of packet received at receiver end within the stipulated time. Our proposed algorithm is compared with various conventional algorithms like AODV and OLSR. Comparing to existing algorithms our proposed algorithms ensures high throughput in vehicular communication. Indirectly this quality of throughput supports for predicting occurrence of accident in urban communication. This shows proposed algorithm is better than the algorithms like OLSR and AODV considering

the VANET throughput. In this proposed algorithm, the receiver receives more number of packets with less time compared to other algorithms. This throughput increases the number of communication in the VANET and this helps in predicting number of occurrences of accidents in the urban cities. This throughput determines the frequency in the communication of VANET. Also, by using supervised transmitters and receivers in the VANET communication, the throughput can be increased.

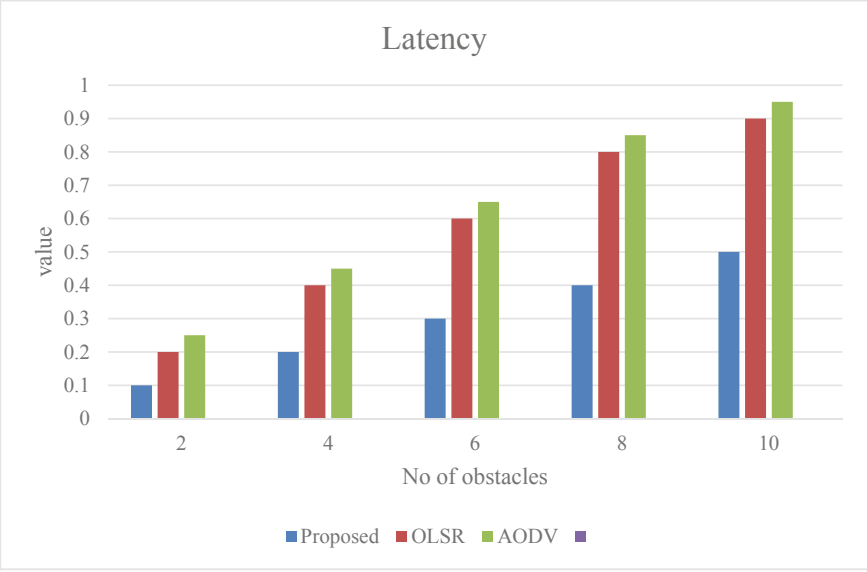


Fig. 2. Latency vs No of obstacles

Figure 2. Gives graphical representation for estimation of latency and our proposed algorithm ensures that, our algorithm gives reduced latency. Several obstacles may disturb communication in VANET and it brings increased latency as result. In our scenario, proposed algorithm reduces latency very well. Reduced latency will improve the performance of communication in VANET. So, by this graphical representation it is identified that if number of obstacles increases, the latency also decreases that affects the VANET communication. These obstacles must be controlled by establishing increased frequency in the communication network. It can be improved by using increased throughput routers and transmitters which transmits supervised signals between vehicles. So that the communication level in the VANET even if the obstacles are present.

5 Correlation of Diver and Drivers Age

According to the data number of accidents is inversely proportion to the driver age. This shows that the teenage drivers make more accident compare to the old age people. This speaks about the psychology of the human behavior. The accident cause by the low age

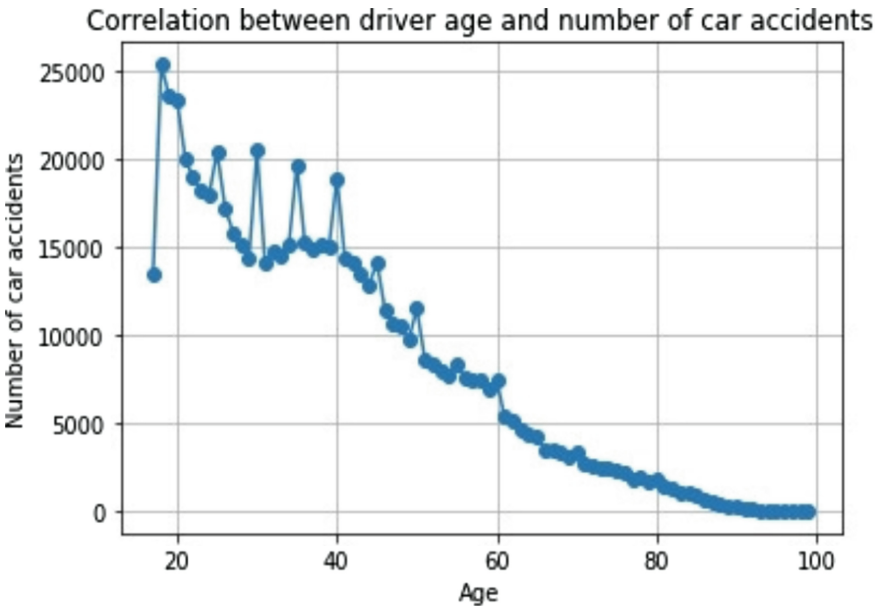


Fig. 3. Drivers age vs car accidents

or teenage people is because the lack of concentration on the road and there are many factor influencing the accidents (Fig. 3).

In the above figure it shows that the correlation between the driver age and the number of accidents. As the age increases the number accidents decreases so the age is inversely proportional to the number of accidents. By this we can understand the phychological factors affecting the accident.

6 Conclusions

In vehicular communication routing is an important parameter to ensure quality of vehicular network. Addition to that predicting accident in urban area becomes very important for making effective communication. In this paper we addressed accident prediction as vital constraint added with that throughput and latency also considered. Various existing algorithms are compared with our proposed algorithms and our work ensures improved throughput and reduced latency. This safeguards or predicting accident in urban vehicular communications. Traffic safety in urban areas can be enhanced by this accident prediction. Also, this study can help and support the government by providing information about the road accidents to the cops. So that the road safety measures can be taken to decrease the road accidents in the urban areas and to provide higher safety to the traffic. These models are dependent on input data set and it should be considered the influence of details of traffic accidents in urban cities. In future, the accuracy of the model are planned to increase by integrating more relevant traffic accident parameters such as road conditions, traffic flow and other related constraints. In addition to that the alert signals can also be established in the accident prone areas through the result obtained by the model.

References

1. Li, P.: A Deep Learning Approach for Real-time Crash Risk Prediction at Urban Arterials (2020)
2. Hong, D., Lee, Y., Kim, J., Yang, H.C., Kim, W.: Development of traffic accident prediction models by traffic and road characteristics in urban areas. In: Proceedings of the Eastern Asia Society for Transportation Studies, vol. 5, pp. 2046–2061 (2005)
3. Eboli, L., Forciniti, C., Mazzulla, G.: Factors influencing accident severity: an analysis by road accident type. Transp. Res. Procedia **47**, 449–456 (2020)
4. Ketha, T., Imambi, S.S.: Analysis of road accidents to indentify major causes and influencing factors of accidents-a machine learning approach. Int. J. Adv. Trends Comput. Sci. Eng. **8**(6), 3492–3497 (2019)
5. Jadhav, A., Jadhav, S., Jalke, A., Suryavanshi, K.: Road accident analysis and prediction of accident severity using machine learning. Int. J. Eng. Res. Technol. (IJERT) **7**(12), 740–747 (2020)
6. Shi, Y., Biswas, R., Noori, M., Kilberry, M., Oram, J., Mays, J., Chen, X.: Predicting road accident risk using geospatial data and machine learning (Demo Paper). In: Proceedings of the 29th International Conference on Advances in Geographic Information Systems, pp. 512–515 (2021)
7. Reddy, A.P., Shekhar, R., Babu, S.: Machine Learning Approach to Predict the Accident Risk during Foggy Weather Conditions
8. Ballamudi, K.R.: Road accident analysis and prediction using machine learning algorithmic approaches. Asian J. Humanit. Art Lit. **6**(2), 185–192 (2019)
9. Rana, V., Joshi, H., Parmar, D., Jadhav, P., Kanojiya, M.: Road accident prediction using machine learning algorithm. IRJET-2019 (2019)
10. Dabhade, S., Mahale, S., Chitalkar, A., Gawhad, P., Pagare, V.: Road accident analysis and prediction using machine learning. Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) **8**, 100–103 (2020)
11. Chong, M., Abraham, A., Paprzycki, M.: Traffic accident analysis using machine learning paradigms. Informatica **29**(1) (2005)
12. Labib, M.F., Rifat, A.S., Hossain, M.M., Das, A.K., Nawrine, F.: Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC), pp. 1–5. IEEE (2019)
13. Yannis, G., Papadimitriou, E., Chaziris, A., Broughton, J.: Modeling road accident injury under-reporting in Europe. Eur. Transp. Res. Rev. **6**(4), 425–438 (2014). https://doi.org/10.1007/s12544-014-0142-4
14. Soemitro, R.A.A., Bahat, Y.S.: Accident analysis assessment to the accident influence factors on traffic safety improvement. In: Proceedings of the Eastern Asia Society for Transportation Studies, vol. 5, pp. 2091–2105 (2005)
15. Liu, M., Chen, Y.: Predicting real-time crash risk for urban expressways in China. Math. Probl. Eng. (2017)
16. Ahammad Sharif, M.: Real-time crash prediction of urban highways using machine learning algorithms (Doctoral dissertation) (2020)
17. Ramli, M.Z.: Development of accident prediction model by using artificial neural network (ANN) (Doctoral dissertation, UniversitiTun Hussein Onn Malaysia) (2011)
18. Cioca, L.I., Ivascu, L.: Risk indicators and road accident analysis for the period 2012–2016. Sustainability **9**(9), 1530 (2017)

Analytical Study of Starbucks Using Clustering

Surya Nandan Panwar, Saliya Goyal, and Prafulla Bafna(✉)

Symbiosis International (Deemed University), Symbiosis Institute of Computer Studies and Research, Pune, Maharashtra, India
{sap2022115,srg2022104,prafulla.bafna}@sicsr.ac.in

Abstract. Customer experience is having more significance than the product itself. Placing Company at the right location is a critical decision and impacts the sales of that company. Clustering technique is used which generates decision regarding location/place for new store. The study is associated with analysis that has been performed on Starbucks corporation dataset using hierarchical clustering and k-means clustering model. Various clustering evaluation parameters like entropy and silhouette plot are used. Hiearachical and k-means clustering is applied to get the details of locationwise records of starbucks store. Due to location based cluster analysis, it will help to decide the location of newly introduced store. of K-means shows better performance with average silhouette width and purity as 0.7., entropy 0.

Keywords: Analysis · Cluster · entropy · Silhouette Coefficient Recession

1 Introduction

Jerry Baldwin was an English teacher, Zev Siegl was an history teacher and Gordon Bowker was a writer and all of them wanted to sell high quality coffee beans as they were inspired by Alfred Peet, a coffee roasting entrepreneur who taught them his style of roasting coffees. After a time, span of ten years, Howard Schultz visited their store and started planning to build a strong company and expanding high quality coffee business with the name of Starbucks. There are various strategies of the one of the biggest corporations of America: STARBUCKS that is Growth Strategy, Corporate Social Responsibility, Customer Relationship, Management, Financial Aspect, Marketing Strategy. To study these strategies to produce decisions like where to place next Starbucks store to gain maximum profit, needs data mining techniques to be implemented. There are several steps in data mining. It starts from Data preparation to algorithm.

Data Preparation

It is a very first step which occur as soon as the data is inputted by the user. In the process the raw data is prepared for the further processing step. Data preparation includes various steps like collecting, labelling etc.

Collecting Data: These steps include the collection of all the data that is required for the further processing. It is an important step because data is collected through different

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1013–1021, 2023.
https://doi.org/10.1007/978-3-031-27409-1_93

data sources which includes laptops data warehouses and inside application on devices. Now connecting to each such type can be very difficult.

Cleaning Data: Raw data contains various errors, blank spaces and incorrect values. Now, all this can be corrected in this very step. This step includes correcting of errors, filling of missing space and Esurance of data quality.

After the data has been cleaned, and now can be transformed into a readable format. This includes various steps like Change in field formats, Modifying naming conventions etc. Clustering refers to an unsupervised learning in which the entire dataset that is provided by the user is distributed into various groups based on the similarities. This helps to differentiate one common group with similar characteristics from the other group. Clustering can be classified into two categories.

- Hard Clustering: In such a type of clustering each datapoint belongs to a group either completely or incompletely.
- Soft Clustering: Now talking about soft clustering a datapoint can belong to more than one group with similar characteristics.

Hierarchical Clustering

This method is used to find similar clusters(groups) based on certain parameters(characteristics). It will form tree structures as figure shown below. These tree structures are formed based on data similarities. Further sub-clusters that are related to each other based on distance between data points are formed. It generates a dendrogram plot of the hierarchical binary cluster tree. It consists of many U- lines that connect data points in a hierarchical tree. It can be performed with either a distance matrix or raw data. Whenever raw data is provided by the user the system starts creating a distance matrix in the background. This distance matrix will show the distance between the objects that are present in the data. This type of clustering is mainly done by using a formula named as Euclidean distance. A straight line is drawn from one cluster to the other one and the distance is calculated based on the length of the straight line. Entropy is said to be a measure of the contaminant or ambiguity which is present in the dataset. Basically, in simple words entropy measures the impurity that might be present in the system. This will help the user to evaluate the quality of clusters. It tells the similarity of each observation with the cluster it has been assigned in comparison to the other clusters. If we take out the mean of all the silhouette values the result will help us know the number of clusters present in the dataset. Its main use is to study the distance that separates the two clusters. It will further tell us how close one point in one cluster is to the neighbouring cluster and thus provides a way to assess parameters like number of clusters. K-means clustering algorithm is required when we have to locate groups which are not labelled in the dataset. Its main use is to make sure about the business assumptions as in what types of groups exist in an organization. It can also be used to identify certain unknown groups in complex datasets.

2 Literature Review

Starbucks carries a very brilliant marketing strategy where it identifies its potential customers and further uses marketing mix to target them. It is a four-way process which includes segmentation, targeting, positioning and differentiation. Starbucks Logo plays a significant role as well as it is present on every product and the brand is recognised by it. There were many changes in the logo over years but the brand keeps on simplifying it to make it more recognizable. Talking in terms of Corporate Social Responsibility, Starbucks is a highly devoted member of it. Customers are the main focus point of any business; therefore, Starbucks maintain a very healthy relationship with its customers [1]. Starbucks, which competes in the retail coffee and snacks store industry operates in sixty-two countries with over 19,767 stores and 182,000 employees. Starbucks also had a time when they experienced a major slowdown which was in 2009 due to the economic crisis and changing consumer tastes. Talking about the market share Starbucks has 36.7% marke60share in United States and has operations in over 60 countries and these are also the strengths of Starbucks which are included in SWOT Analysis [2]. How to respond to efficiently and effectively to a change it has always been a constant question. And to get an answer to this question Starbucks went through research. Around 2006 Starbucks performance started decreasing. What factors led such a strong MNC to fail? So, this paper aims on the dynamics capabilities concept and to apply them on Starbucks case. Taking step back in time it is very necessary to understand the basis of the concept of dynamic capabilities. The view of dynamic capabilities has always been evolving since its first appearance. Starbucks made a difference through its unique identity, focusing its strategies of providing a distinctive coffee tasting experience. After seeing a downfall in 2008 they made some numerous changes and they were back on their feet [3]. Starbucks is considered as one of the leaders in the coffee industry which operates from five different regions of the world. Americas with includes United States, Canada and Latin America, China and Asia Pacific (CAP), Europe, Middle East, and Africa (EMEA) and Channel Development. Starbucks history shows various variations in its development. Its first store got opened in Seattle, Washington and the name was inspired by the character Starbuck of the book Moby Dick [4]. One of the methods that is considered important in data mining is Clustering analysis. The clustering results are influenced directly by the clustering algorithm. The standard k-means algorithm has been discussed in this paper and the shortcomings of the same are analyzed. The basic role of k-means clustering algorithm is to evaluate the distance between each data object and all cluster centers. This helps in lowering the efficiency of clustering. A simple data structure is required to store information in every iteration [5]. To solve a problem a k-means clustering a set of n data points are required in d-dimensional and an integer k. The problem that is given is to determine a set of k points in Rd which are known as centers. Lloyd's algorithm is one of the examples, which is quite easy to implement [6].

3 Research Methodology

This study talks about a dataset that is related to Starbucks corporation. [https://www.kaggle.com/datasets/starbucks/store-locations]. Clustering has been performed and an analysis has been provided about its location. A preferred place has been chosen which was done through various functions. It was concluded that a place which has few stores is preferable as compared to others. Figure 1 shows the Diagrammatic Representation of Research Methodology.

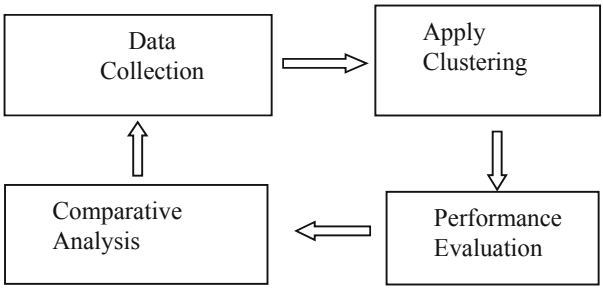


Fig. 1. Steps in research methodology

1. *Data Collection*

The dataset is related to Starbucks corporation and consists of 5 columns and 30 rows. It has fields like store no, street address, city, state and country. These stores are located in three different which are of 3 types. These three cities are present in two different states and two countries. The data is converted into numerical form.

2. *Algorithm execution*

Hierarchical clustering model, k-means clustering has been used. In a sample of 30 values 3 clusters were formed of sizes 10,10,10.

3. *Performance evaluation of algorithm*

In a sample dataset of 10 values entropy is 4.902341 and average silhouette width is 0.63. In a sample dataset of 30 values entropy is 5.902351 and average silhouette width is 0.64.In a sample dataset of 50 values entropy is 6.902361 and average silhouette width is 0.71.

4 Results and Discussions

Table 1 shows the Comparative analysis of different algorithms on variant datset size.

Through Table 1 we come to a conclusion about scalability of algorithm.Even for that data set of size 50, it shows better performance.

Table 1. Comparative Analysis of Clustering Algorithm

Dataset size	K-means			HAC		
	Entropy	Purity	Silhouette Coefficient	Entropy	Purity	Silhouette Coefficient
10	0.11	0.8	0.74	0.12	0.7	0.64
20	0.13	0.8	0.75	0.14	0.71	0.65
35	0.15	0.8	0.76	0.15	0.73	0.66
50	0.1	0.7	0.72	0.11	0.73	0.62

The sample dataset is shown in Table 2. It has attributes like Store numberStreet Address City State/ province Country.

Table 2. Sample Dataset

Store number	Street Address	City	State/province	Country
1	1	1	1	1
–	–	–	–	–
2	2	2	1	2
…	–	–	–	–
30	3	1	1	2

Figure 2 shows the dendogram cluster of the 30 stores of Starbucks. It has been divided into three clusters and it can be interpreted that cluster 1 contains the maximum records. Hierarchical clustering has been applied on the dataset and a dendogram plot has been generated. It generates a dendogram plot of the hierarchical binary cluster tree. It consists of U- lines that connect data points in a hierarchical tree. Entropy function has been applied which gives us a value of 5.902351.

In Fig. 3 Sk2 value has been interpreted using the silhouette plot. Various km functions have been applied.

Figure 4 shows the complete silhouette plot that has been generated and average silhouette width has been calculated as 0.64 (Figs. 5 and 6).

Shows the distance matrix

Figure 6 represents the dendogram cluster of 50 values which have been divided into 3 clusters containing 17, 16, 17 records respectively.

Figure 7 shows the silhouette plot for 50 records and average silhouette width has been calculated as 0.71.

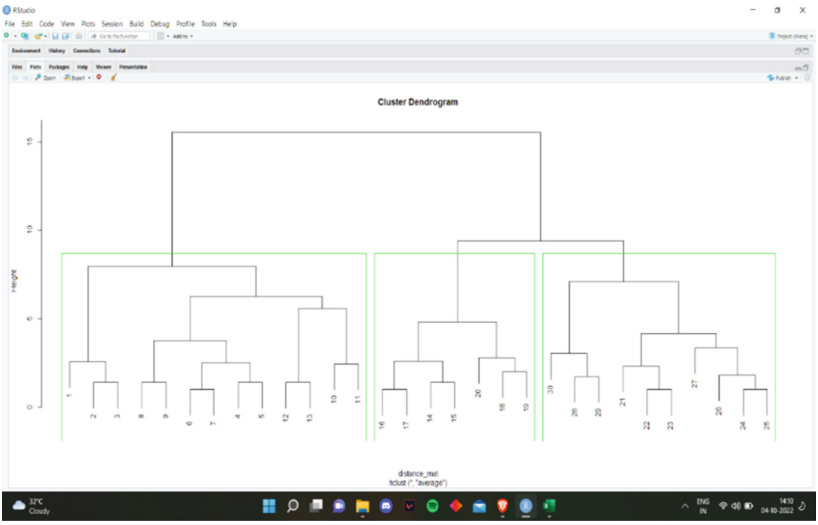


Fig. 2. Dendogram Clustering for a dataset of size 30 value

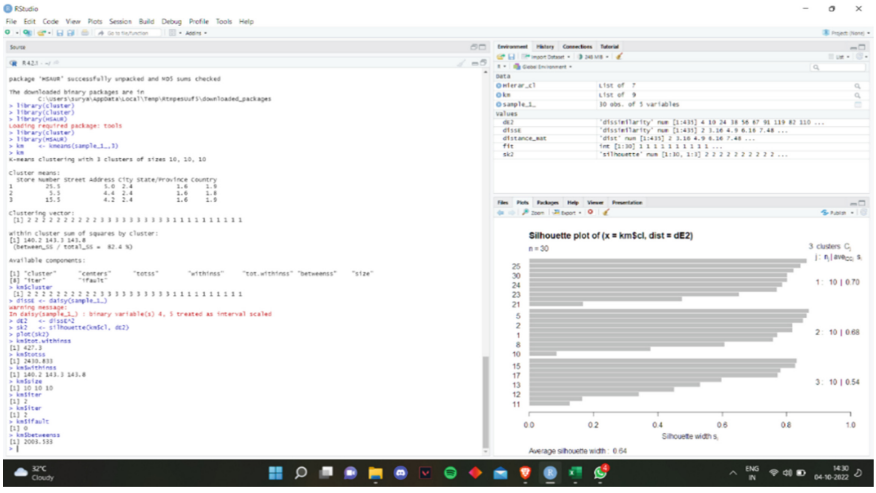


Fig. 3. Application of km functions

It can be concluded that k- means clustering shows best performance on a data of size 50 values. Average silhouette width can be calculated as 0.71. Figure 8 shows different type of parameters used for experiments.

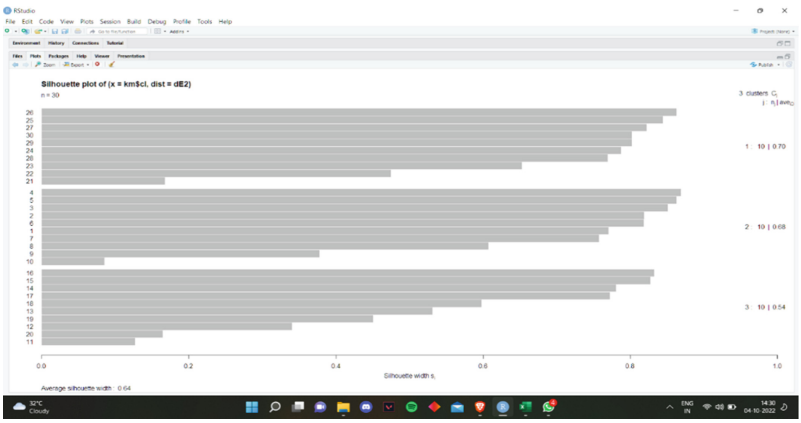


Fig. 4. Silhouette plot for dataset of 30 values

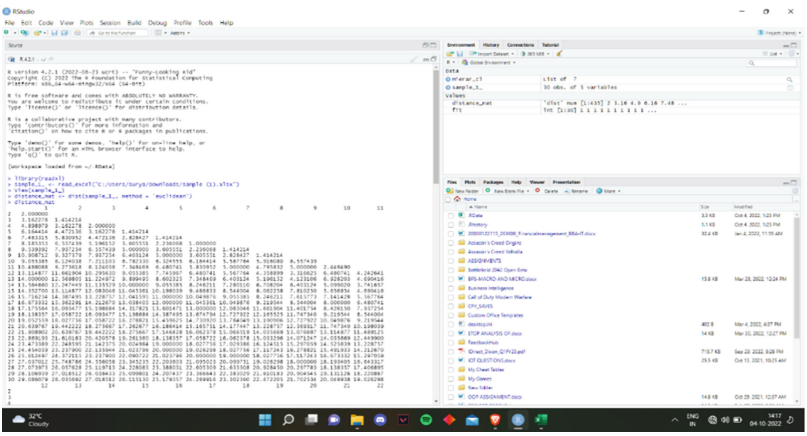


Fig. 5. Distance matrix

5 Conclusions

Clustering techniques are used for decision making to decide the location/place for new store.

The study is associated with analysis that has been performed on Starbucks corporation. Hierarchical clustering and k-means clustering are used and suitability of clustering technique is suggested. Various clustering evaluation parameters like entropy and silhouette plot are used. Hieararchical and k-means clustering is applied to get the details of locationwise records of starbucks store. Due to location based cluster analysis, it will help to decide the location of newly introduced store. of K- means shows better performance with average silhouette width and purity as 0.7., entropy 0.1, The future work focuses on increasing size of the dataset as well as different types of algorithms.

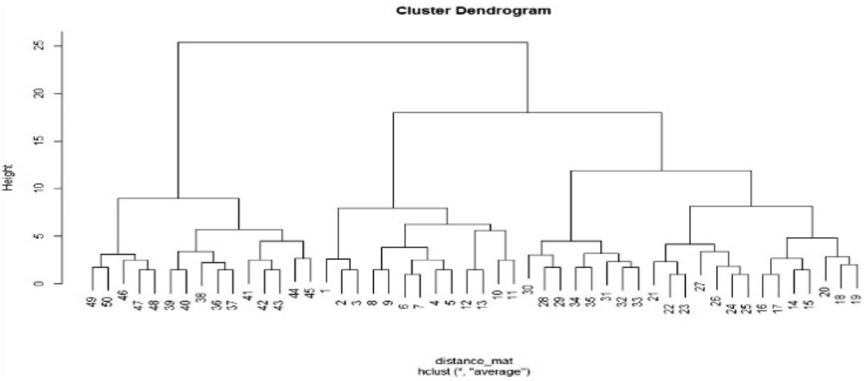


Fig. 6. Dendogram cluster of data set of 50 values

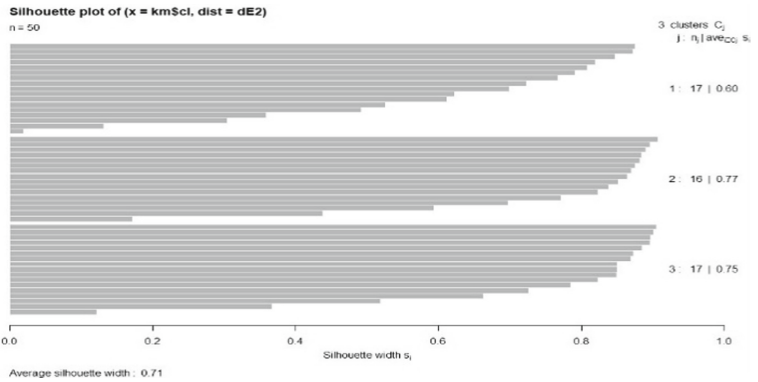


Fig. 7. Silhouette plot for 50 values dataset

In a sample dataset of 10 values available components are "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"
Width: 0.63
Totts: 111.1111
Withinss; 4.000000 5.333333 4.00000
Betweenss: 97.77778
Tot.withinss: 13.33333
Size: 3 3 3
Iter: 2
Ifault: 0

Fig. 8. Parameter setting for clustering

References

1. Haskova, K.: Starbucks marketing analysis. CRIS-Bull. Cent. Res. Interdiscip. Study **1**, 11–29 (2015)
2. Geereddy, N.: Strategic analysis of Starbucks corporation. Harvard [Електронний ресурс].– Режим доступу: http://scholar.harvard.edu/files/nithingeereddy/files/starbucks_case_analysis.pdf (2013)
3. Vaz, J.I.S.D.S.: Starbucks: the growth trap (Doctoral dissertation) (2011)
4. Rodrigues, M.A.: Equity Research-Starbucks (Doctoral dissertation, Universidade de Lisboa (Portugal)) (2019)
5. Na, S., Xumin, L., Yong, G.: Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 63–67. IEEE (2010)
6. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
7. Sarthy, P., Choudhary, P.: Analysis of smart and sustainable cities through K-means clustering. In: 2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control PARC), pp. 1–6. IEEE (2022)
8. Wu, C., Peng, Q., Lee, J., Leibnitz, K., Xia, Y.: Effective hierarchical clustering based on structural similarities in nearest neighbour graphs. Knowl.-Based Syst. **228**, 107295 (2021)
9. Shetty, P., Singh, S.: Hierarchical clustering: a survey. Int. J. Appl. Res. **7**(4), 178–181 (2021)
10. Batool, F., Hennig, C.: Clustering with the average silhouette width. Comput. Stat. Data Anal. **158**, 107190 (2021)



Analytical Study of Effects on Business Sectors During Pandemic-Data Mining Approach

Samruddhi Pawar, Shubham Agarwal, and Prafulla Bafna(✉)

Symbiosis International (Deemed University), Symbiosis Institute of Computer Studies and Research, Pune, Maharashtra, India
prafulla.bafna@sicsr.ac.in

Abstract. In this article there is an analysis of different business sectors during pandemic. It explains about how to find a pivot and capitalize in best possible way. It will also be backed by real dataset of 200 listed companies to give an in depth understanding on how to make the best out of available data and predict the future outcomes according to it. It includes predictions of the indexes of September 2022 with the help of historical data by performing least square method, Classification, Hierarchical clustering. Classification gives input data based on the attributes gathered. Clustering helps to group similar stocks together based on their characteristics. K-Means Clustering and OLS beta provided results with the best accuracy for the dataset as can be seen from the Confusion Matrix. Sectors like FMCG & Utility tend to possess a lower beta < 0.85 whereas discretionary and automobile possess beta > 1. K-Means Clustering has fared well over a longer period of timeline with an accuracy of 78% throughout the dataset. Clustering and classification together results in dynamism of experiment.

Keywords: Recession · Classification · Prediction · Least square method · Clustering

1 Introduction

While most people think that doing business in recession is bad idea, well it's the best time to be in business. The best examples could be how Disney survived the Great Depression in 1929, Netflix from Dot Com bubble in 2000, Airbnb from Financial Crisis in 2008, Paytm from Demonetization in 2016. Following is the case study to understand this better. E.g. - Evolution of Ice Industry [5].

It evolved in 3 stages: Ice 1 0, 2.0 & 3. 0.

Ice 1.0 was in early 1900s where people used to wait for winters went to Alps to get ice, come here and sell it [7].

Then 30 years later i.e., Ice 2.0 wherein ice was produced in a factory & then the iceman use to sell ice nearby the factory.

And then 30 years later i.e., Ice 3.0 there was another paradigm shift wherein ice was available in our house that is modern day refrigerator.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1022–1030, 2023.
https://doi.org/10.1007/978-3-031-27409-1_94

The point to be noted over here is that neither the Ice 1.0 companies made it to 2.0 nor 2.0 companies made it to 3.0. There was a pivot in 1.0 & 2.0 which these companies couldn't understand and thus didn't make.

it to 3.0 & thus went out of business. As an entrepreneur if you are able to identify this pivot you will be able to get ahead of your competition. The highlight over here is during the pivot the customer's need for the product, the demand and the supply didn't change, what changes is the medium of supply which results into this paradigm shift. E.g. -: Before demonetization people used to make payments via cash & and after demonization payment is being done digitally [6].

Every time paradigm shift happens the entire business ecosystem divides in four segments & the company will fall into either one of these 4 segments.

Type 1: Best category of all-Perfect product + Perfect supply chain. E.g. The sanitization industry during Covid times.

Type 2: Product is perfect but supply chain needs to be change.

E.g., MTR Foods. In 1975 India went through socio-economic crisis and inflation rise up to 15% & Government told the restaurants to drop down their prices, so it was almost impossible for a person to run a business. P Sadanand Maiya, founder of MTR realized that the customer's need & demand for the product is same but it needs to change its supply chain. So, he started packaging the dry mix of idlis and dosas & sold it to customers. The sales shot up & started making profits. Even after the crises got over it was making crores out of this & the business was in boom.

Type 3: Perfect supply chain but product needs to be change.

E.g., The textile industry. In Covid times, the textile companies were using same people, machines, and supply chain but rather than manufacturing clothes they were making PPE Kits thus becoming a beneficiary of pivot [8–10].

Type 4: Toughest of all- Change both product & supply chain.

E.g. -: Many local bhel-puri stalls for e.g., Ganesh/ Kalyan Bhelpuri Wala. Before pandemic they used to sell bhel-puri on stalls but as soon as Pandemic hit, it got shut. But the need and demand of customers were same it's just that people are reluctant to buy it from street vendors. So, they started packing bhel & chutneys & now sell to grocery stores and have also expanded their radius by 10 kms and thus have increased their customer base by 4x.

To explain this better, we have taken Covid 19 phase of March-April 2020 as Recession period. The Covid 19 was one of the worst recessions ever known to mankind which shook the entire world economy. It accelerated the economic downturn, affected the lives of poor the most which resulted to increase in extreme poverty, decrease in production and supply, company started mass layoffs to reduce their expenses which made the situation even worse for a common man and finally so as to cope with all these economic turmoil Government started to hike prices of basic commodities which resulted to increase in inflation.

In this research paper, authors have taken the datasets of listed companies of India. To give reader a taste about what an impact does a recession has, authors have drawn comparative analysis of companies in different sectors like FMCG, Utility, Automobile, Pharmaceutical and Steel Industry in which authors have.

(1) Compared the indexes of their stocks.
(2) Based on that calculated the risk factor (beta) of these listed companies.
(3) And based on the data of FY 20–21 authors predicted what will be the situation of these companies in September 2022 [12, 12–14].

Beta Formula:

Under the CAPM model, there are two types of risks: systematic and unsystematic. While the former is related to the market wide movement and affects all firms and investments, the latter is firm-specific and affects only one firm or stock. Thus, while there is no way to eliminate systematic risk with portfolio diversification, it is possible to remove unsystematic risk with proper diversification. In CAPM, there is no need to take and reward a risk being eliminated by diversification. Systematic risk (or no diversifiable risk) is the only risk rewarding (Simonoff, 2011: 1). βİ is conceived as a measure of systematic risk and can be calculated as [11].

2 Background

This research paper has conducted a thorough market research on how Public Listed Companies react to recession. It has break down the data of 4700 companies which include Fortune 500 companies, companies going from public to private and also some of them filing for bankruptcy. It highlights that what few great companies have done better and was able to analyze the situation where the threat was right at their doorstep which other companies didn't understand and react hastily. The reason why these companies have different results was because of their approach towards the situation. In this paper, they say there are three ways a company reacts to situations like recession. Prevention focus: They implement policies like reduction of operating costs, layoffs of employees, preserving the cash, there sole focus is to reduce working capital expenditure. They also delay things like investments in R&D, buying assets, etc. This too defensive approach leads the organization to aim low which hammers the innovation and overall enthusiasm of a company's work culture. Promotion focus: In this scenario a company invests heavily in almost every sector hoping that once it gets bounce back it will have largest market share. It ignores the fact that post recessions, does the end consumer has that appetite to buy the product or even need for that specific product as he/she has also felt the heat of the recession. So, a company should also focus on what consumer want on that specific duration. Progressive focus: In this situation, a company maintains a perfect balance between cutting of operational costs at the same time get the best out of what we have. It doesn't really focus on mass layoff because it understands the employee's perspective and want to retain their trust with company. They spend considerable amount of money in R&D so as to stay ahead of the game. This research paper tells us that how companies' approach to pivoting times and how it affects the momentum of the organization [1].

In this article suggest the preventive measures when a company knows that there is going to be certain volatility in the market and later in recession period what actions or steps should be taken so that the recession doesn't become a threat but rather an opportunity to stay ahead of the curve.

First step is to don't burn out on cash because that's the only source of fuel that keeps the business's engine going. The more debt a company possess, more difficult it becomes to bounce back and it just becomes a matter of survival rather than creating something unique. Next step is decision making i.e., to allocate right amount of funds in each sector and have a perfect balance of working capital expenditure and investments in buying assets. A company should also look beyond layoffs and find a way to retain them as employees are considered to be asset. And after situation comes back to normal the rehiring process also becomes very hefty for an organization. Invest in new technology so as to give best customer experience but also considered an important factor of what the customer needs are and what are currents trends in market [2].

The research paper focuses on better management and review-based system that focuses on internal management of a firm. There are organizational capabilities and resources which are used to create an edge over competition. There are assumptions which crucial to its successful implementation, ranging from resources being distributed across divisions differently in a sustainable time period. This ensures that companies have precious, irreplaceable and suitable resources which enable them to create leverage and unique strategies ensuring that they outlive their competitors. When economic conditions change, a organization should be able to iterate, hold, eliminate and adapt to stakeholders' requirements. Hence, companies require the agility to change resources into combination of resources that pave the path to survival [3].

During economic crisis, consumers experience a shift in their preferences which requires businesses to adapt and strategize to retain its customer's pool. Three ways in which companies try to maintain their market share are lowering prices to retain sales, reducing costs to maintain profits, not making any changes. These recessions usually happen due to rapid growth in credit debt and it ballooning up out of control which leads to a sharp drop in demand for goods and services, leading to a recession [4].

Gulati studied more than 4700 publicly listed companies during 1980 crisis. He found out that 17% of the companies were either bankrupted or taken by competition and 80% could not reach their sales and profit figures of prevailing three years and only 9% were able to beat their crisis prevailing numbers by at least 10%. Jeffery Fox identified that companies surpassing their competitors in innovation tend to do better over a longer time period. Thus, new age business models are perhaps the best way to deal with downside in economics in order to stay in market and become the beneficiary of the pivot. And for those who fail to see this pivot by default don't see it as opportunity but as a threat [5].

3 Research Methodology

This section depicts what steps are followed so as to get the desired output for future predictions. It tells from where the data is been gathered and what kind of algorithms were performed during the entire process. And lastly which classifier predicts the value closest to actual value.

The paper is organized as follows. The work done by other researchers on the topic is presented as a background in the next section. The third section presents the methodology; the fourth section depicts Results and discussions. The paper ends with a conclusion and future directions. Table 1 shows steps in the proposed approach.

Table 1: Steps and Packages Used in The Proposed Approach

Step	Library/Pack
Dataset Collection	pandas, opencv, google finance
Apply & Evaluate Classifiers	numpy, linear-reg, matplot, knn
Selecting the best classifier and predicting the future prices of securities	K-Means, Hierarchical Clustering

Predicting price action of various commodities and securities help analysts and derivatives traders make a better choice during periods of recession. As a novice, a regular Joe can to make an informed choice when focusing on sectors and growth areas during such times. Figure 1 below depicts our research flow and methodology. The research uses hierarchical and KNN clustering in order to predict similar sectored based companies across the Nifty 200 Index of India.

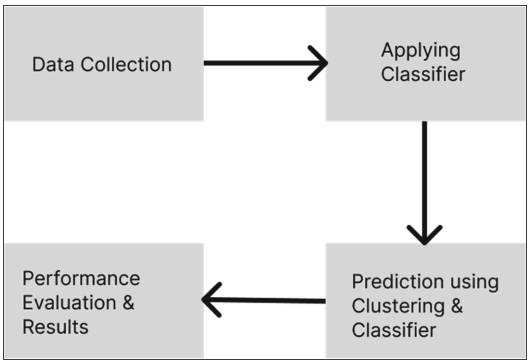


Fig. 1. Steps in research methodology

Data Collection:

The data includes the indexes of listed companies across different sectors like FMCG, Utility, Automobile, Pharmaceutical and Steel Industry which is having 200 records.

Data Training and Preparation:

Model Training using Ordinary Least Square Method and using Clustering for predictions: The labeled dataset of around 20 listed Nifty 50 companies across sectors is used to find the ordinary lease square with respect to the index movement which provides us with a variable, beta.

Prediction:

K-Means Clustering and Hierarchical Clustering techniques are used to classify over 200 listed corporations across the nation and it was observed that K Means Clustering had performed better over Hierarchical Clustering.

Performance Evaluation:

CAPM is a simple model but includes a strong assumption. It implies that the expected return of stock depends on a single factor (index). According to the model, the beta is a relative risk measure of securities as a part of a well-diversified portfolio.

4 Results and Discussion

			Actual Values	
			Actual up	Actual down
Predicted Values	30-day	Predicted up	926	416
		Predicted down	241	392
	60-day	Predicted up	1023	423
		Predicted down	207	322
	90-day	Predicted up	1050	424
		Predicted down	198	302

Fig. 2. Confusion matrix by OLS beta

Figure 2 shows the hypothesis that we have applied to derive OLS beta and sectored clustering method to map the market price movement for a period of 30, 60 and 90, i.e., one month, 2 months and 3 months' time frame.

The findings fared better on a longer time period rather than short term market price predictions mainly due to volatile nature and quarterly results affecting the price movement which was not an issue on a longer time frame.

The aim of a confusion matrix is to test the findings on real market data with predicted ups and downs and how many actual ups and downs occurred during the same period.

K-Means Clustering and OLS beta provided us results with the best accuracy for the dataset as can be seen from the Confusion Matrix of the tested data for the period of March 2020 - May 2020. Figure 3 shows High Volatility, Medium Volatility points on a plot.

Figure 4 shows a clustering map of 200 + odd listed corporations across India over the Nifty Index where they are classified in three groups of High Volatility (red), Medium Volatility (Blue) and Low Volatility (green). The map considers OLS beta of corporations over a larger period of time and Low Volatility does not just signify lower volatility but also how it has performed over a longer period of time based on accuracy and growth.

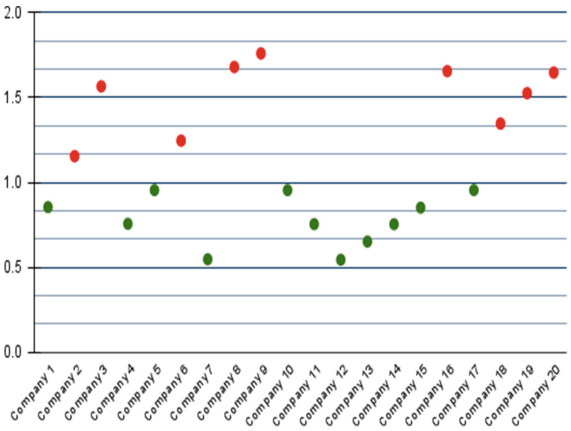


Fig. 3. Volatile plot



Fig. 4. Kmeans cluster plot

Figure 5 shows how these 20 Nifty 50 corporations have spanned out based on Least Square Beta and broadly defined into three groups based on volatility.

5 Conclusions

The current findings and study achieve a prediction of price movement of the selected 5 sectors across the Indian Markets and group's unlabeled data based on their sectored volatility successfully. The formation of the classes of achieved through least square method and beta formula of CAPM model. Experiment was conducted on over 200 + listed Indian Entities and the current version manages to label the given entity based on their performance, volatility and price action. Stock price is a series of different patterns based on the historical data. Classification and clustering are both the central concepts of pattern recognition.

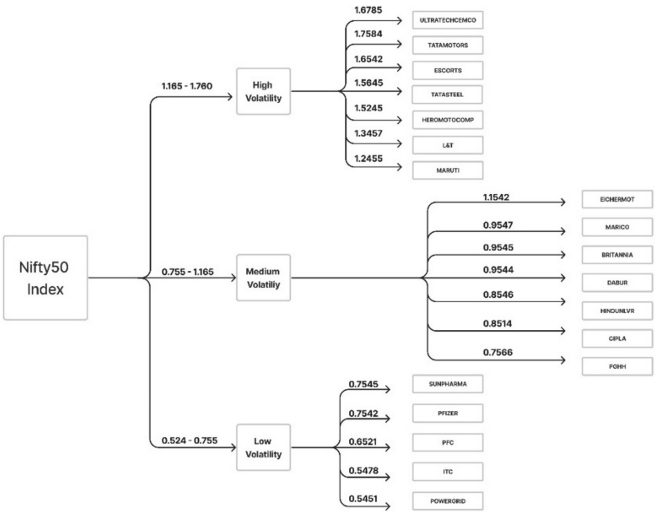


Fig. 5. Hierarchical Clustering

Classification gives input data into one or more pre-specified classes based on the attributes gathered. Clustering helps to group similar stocks together based on their characteristics. Thus, the proposed clustering and classification framework is very beneficial to predict stock prices in a multi-dimensional factors-oriented environment. K-Means Clustering, with its accuracy of over 78% tends to function better on longer duration of database since volatility tends to be lower whereas hierarchical clustering creates a tree like formation for better clustering of dataset based on their OLS beta. High Volatility is defined as greater than 1.165 beta, Medium is in the range of 0.755 and 1.165 whereas lower volatility listings fall under 0.524 and 0.755 based on clustering results. The future work focuses on increasing size of the dataset as well as different types of algorithms.

References

1. Auerbach, A., Gorodnichenko, Y., Murphy, D., & McCrory, P. B. (2022). Fiscal multipliers in the covid19 recession. J. Int. Money Financ. 102669 (2022)
2. Domini, G., Moschella, D.: Reallocation and productivity during the Great Recession: evidence from French manufacturing firms. Ind. Corp. Chang. **31**(3), 783–810.3 (2022)
3. Goldberg, S.R., Phillips, M.J., Williams, H.J.: Survive the recession by managing cash. J. Corp. Account. Financ. **21**(1), 3–9.4 (2009)
4. Vafin, A.: Should firms lower product price in recession? A review on pricing challenges for firms in economic downturn. ResearchBerg Rev. Sci. Technol. **2**(3), 1–24.6 (2018)
5. Friga, P.N.: The great recession was bad for higher education. Coronavirus could be worse. Chron. High. Educ. **24**(7) (2020)
6. Patel, J., Patel, M., Darji, M.: Stock Price prediction using clustering and regression: a (2018)
7. Gandhmal, D.P., Kumar, K.: Systematic analysis and review of stock market prediction techniques. Comput. Sci. Rev. **34**, 100190 (2019)
8. Shah, D., Isah, H., Zulkernine, F.: Stock market analysis: A review and taxonomy of prediction techniques. Int. J. Financ. Stud. **7**(2), 26 (2019)

9. Xing, F.Z., Cambria, E., Welsch, R E.: Intelligent asset allocation via market sentiment views. IEEE ComputatioNal iNtelligeNCe magazine **13**(4), 25–34 (2018)
10. Gandhmal, D.P., Kumar, K.: Systematic analysis and review of stock market prediction techniques. Comput. Sci. Rev. **34**, 100190 (2019);. 14. https://documents1.worldbank.org/curated/en/185391583249079464/pdf/Global-Recessions.pdf
11. Mendoza-Velázquez, A., Rendón-Rojas, L.: Identifying resilient industries in Mexico's automotive cluster: policy lessons from the great recession to surmount the crisis caused by COVID 19. Growth Change **52**(3), 1552–1575 (2021)
12. Jofre-Bonet, M., Serra-Sastre, V., Vandoros, S.: The impact of the Great Recession on health-related risk factors, behaviour and outcomes in England. Soc. Sci. Med. **197**, 213–225 (2018)
13. McAlpine, D.D., Alang, S.M.: Employment and economic outcomes of persons with mental illness and disability: the impact of the Great Recession in the United States. Psychiatr. Rehabil. J. **44**(2), 132 (2021)
14. Zhai, P., Wu, F., Ji, Q., Nguyen, D.K.: From fears to recession? Time-frequency risk contagion among stock and credit default swap markets during the COVID pandemic. Int. J. Financ. Econ. (2022)

Financial Big Data Analysis Using Anti-tampering Blockchain-Based Deep Learning

K. Praghash[1], N. Yuvaraj[2], Geno Peter[3(✉)], Albert Alexander Stonier[4], and R. Devi Priya[5]

[1] Department of Electronics and Communication Engineering, Christ University, Bengaluru, India

[2] Department of Research and Publications, ICT Academy, IIT Madras Research Park, ManagerChennai, India

[3] CRISD, School of Engineering and Technology, University of Technology Sarawak, Sibu, Malaysia

drgeno.peter@uts.edu.my

[4] School of Electrical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

[5] Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

Abstract. This study recommends using blockchains to track and verify data in financial service chains. The financial industry may increase its core competitiveness and value by using a deep learning-based blockchain network to improve financial transaction security and capital flow stability. Future trading processes will benefit from blockchain knowledge. In this paper, we develop a blockchain model with a deep learning framework to prevent tampering with distributed databases by considering the limitations of current supply-chain finance research methodologies. The proposed model had 90.2% accuracy, 89.6% precision, 91.8% recall, 90.5% F1 Score, and 29% MAPE. Choosing distributed data properties and minimizing the process can improve accuracy. Using code merging and monitoring encryption, critical blockchain data can be obtained.

Keywords: Anti-Tampering Model · Blockchain · Financial Big Data · Deep Learning

1 Introduction

Traditional financial systems support trust and confidence with formal and relational contracts and courts [1]. Scale economies lead to concentration. Increased concentration raises transaction fees, entry barriers, and innovation, but boosts efficiency [2]. By distributing economic infrastructure and governance control to trusted intermediaries, concentration weakens a financial system's ability to withstand third-party interference and failure [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1031–1040, 2023.
https://doi.org/10.1007/978-3-031-27409-1_95

In modern economic systems, a third party manages financial transactions [4]. Blockchain technology can ease decentralized transactions without a central regulator. It's used for many things, but mostly in finance. Blockchain is encrypted data blocks. Blockchain technology organizes economic activity without relying on trusted middlemen by encrypting a transaction layer [5]. Cryptography and technology limit blockchain systems, but real performance depends on market design [6].

Proof of work links participants' computational ability to their impact on transaction flow and history to prevent Sybil attacks [7]. When proof of work is used, miners must invest in specialized infrastructure to do costly computations. This adds security because it's difficult to gather enough processing power to corrupt the network. Inefficient calculation [8]. Member power is tied to their ability to prove they own cash or other system stakes, reducing the need for it [9].

Despite the growth of blockchain technology and economic analysis, there is limited research on whether blockchain-based market designs are scalable. Our idea of blockchain-based long-run equilibrium is a good one. Long-term market design elements are different for proof-of-work and proof-of-stake [10].

Blockchain has gained popularity due to its ability to secure data. Many non-financial and financial industries are interested in blockchain technology. Many businesses are developing, evaluating, and using blockchain because of its potential and costs. Blockchain can improve services and save money. Blockchain technology helps fight fraud and money laundering while speeding up multi-entity transactions. This paper improves data system security. Most financial sectors struggle to protect e-commerce customer data. This paper illustrates economic sector issues and offers a blockchain-based model solution.

2 Background

Blockchain is a decentralized payment method that simulates virtual consumption. Distribution is random. This device sends and receives data from multiple network locations. Mutual justice networks are decentralized. Decentralized networks have this advantage. When all network nodes are destroyed, it's a threat. Figure 1 shows a decentralized network.

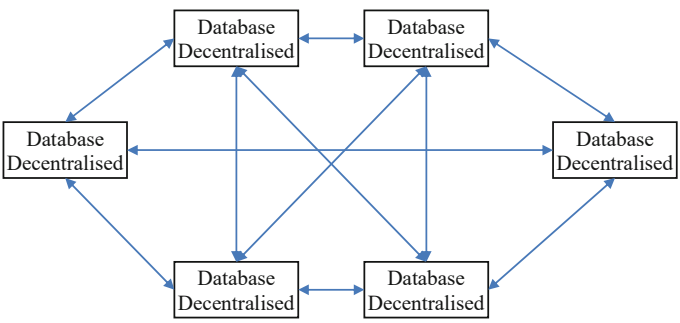


Fig. 1. Decentralized network architecture

Centralized networks have centralized servers. This hub connected all devices. Every device can talk. It's an end-user communication network.

The hash value [7] limits the time between each data item. Bitcoin [8] eliminates the second transaction. Digital capital circulation needs oversight and control, so secondary payment is needed before blockchain payment. Blockchain payment allows confidential face-to-face resource sharing. Public and private keys improve blockchain payment security. Only users with an agreement can access the two-way encrypted blockchain fundholders. Users can use the block browser to see how each dispersed point is connected.

Blockchain's core is task-checking reaction chains or indestructible accounting books. Blockchains can be used for more than just financial transactions and recording user-generated content [11–14]. Symbolizing your data is enough. These notebooks contain transaction records and vouchers. Next, blockchains evaluate these credentials. The block link found credentials can't be changed.

The diagram below shows blockchain data flow. When a profit is made, the payment is distributed to each network. Each decentralized point store accepted payment content. It's possible to simultaneously evaluate data and generate blockchains at each distributed point [15, 16]. Once qualified blockchains are established, the data is distributed. Individual blockchains are linked into one long chain. This procedure does not require process confirmation or third-party oversight. Only a large reputation network and consensus are needed.

When a new customer joins a bank, KYC and KYB begin. Customer identity is verified per regulations. A first customer profile helps tailor services to retail or corporate customers. The KYC/KYB process is dynamic, making it difficult to keep profiles and documents up to date as consumer information and regulations change [17]. A financial institution usually requests several documents to get to know a customer. Centralized client documentation can help. Data leaks and cyberattacks can compromise this system.

Blockchain technology can help by decentralizing and protecting KYC. Blockchain has many benefits in this situation, including:

- **Decentralization**: Customer records are recorded in a decentralized manner, which decreases the data protection of centralized storage. In addition to enhancing security, decentralization improves KYC data consistency.
- **Improved Privacy Control**: Decentralized apps and smart contracts handle this. Financial contracts protect client data. KYC (or other) access to customer info requires permission.
- **Immutability**: Saved blockchain data cannot be changed. This ensures that all blockchain-using financial institutions have accurate consumer data. When an account is closed, the GDPR's right to be forgotten may require that the customer's personal information be removed from the company database. Stakeholders' solutions diverge on how blockchain data can support this premise.

Financial companies collaborate across the value chain. Fast transactions require two or more banks. Cybercriminals target the transaction infrastructure. Recent attacks on financial services infrastructure show that critical infrastructures of financial organisations remain vulnerable.

Financial institutions should share information to prevent supply chain attacks. Sharing security information throughout the economic chain may spur future supply chain security collaboration.

Blockchain can share physical and cyber-security data more efficiently. Distributed ledgers allow security experts to share data securely, easing collaboration.

Financial institutions can collect, process, and share physical and cyber security data with value chain parties. This data isn't just about attacks and threats; it may also include asset and service data.

First, track and verify data. Security and data stability can be maintained while maintaining logistical data integrity. Second, prevent tampering during logistics. The blockchain technology used to protect cargo data is integrated into the data flow process, allowing it to understand each link's outputs and inputs in real-time. Consider users carrying mobile phones. To avoid cargo, recall problems and be more versatile, the user should master and record all cargo code circulation data. Customers can use blockchain to get supply chain info anytime.

3 Proposed Method

This section presents how blockchain enables significant data transactions via big learning using an anti-tampering model.

3.1 Feature Extraction:

A financial chain database data score extraction method can evaluate the classic combination and highlight significant database data. Model matcher matches financial predictions. These are the details provided by economic sector predictors to predict and compare this year's finances.

When a split equation is applied to supply chain data features, the features are filtered and extracted using size score values. When data is sparse, the sparse Score is a major factor in selecting actual data characteristics.

This can ensure that the collected data is sparse. X represents aggregated data, and Y represents dispersed accurate data. Using the L1 model coefficient matrix, you can recover the coefficient matrix when computing data vector dilution. A formula explanation follows in Eq. (1).

$$Y = min||S(i)||; \tag{1}$$

The X' matrix is the data vector-free matrix. s represents the data vector during reconstruction. After collecting dilution data, a reconstruction matrix and transformation coefficients can be expressed. When the coefficient of reconstruction is defined correctly, the difference between a data set's reconstruction and the original feature can be quantified. Example: We can compare data sample reconstruction and dataset characteristics. The gap narrows as a game's features and performance are maintained. S(r), an objective function, meets the following criteria in Eq. (2):

$$S(r) = \frac{\sum_{i=1}^{n} \left(x(ir) - X^{'}(si, r) \right)^2}{Var(X(r))} \tag{2}$$

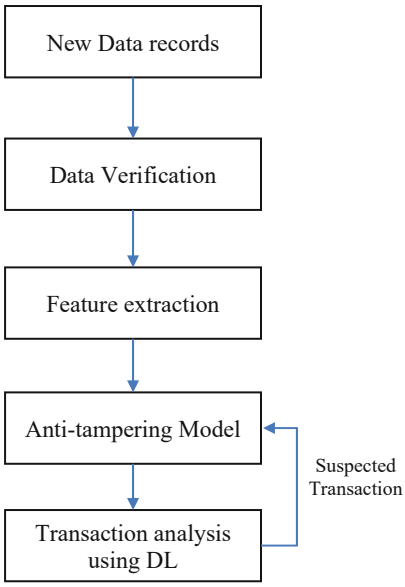


Fig. 2: Anti-Tamper BC DL model

Divide the difference between the dataset size and reconstruction features to get the data set dimensionality feature dispersion. The data set data function uses the concept score to extract outliers. Using correct feature suggestions, expression performance can be maintained if feature variation is less than reconstruction [5].

3.2 Anti-tampering Model

Figure 2 shows the algorithm's two steps. Section 2 describes the environmental model's global and local scenarios. Using the former, you can compare light-levelled alternatives to transaction C. Second paragraph describes the set's objects. Using the current transaction C condition, the algorithm computes the probability of each possible outcome.

Examine the light source's general qualities. Finally, the camera's current image is compared to a set of targets for each scenario. Probabilistic methods are used this time. If too many targets aren't visible, the system alerts. A rule-based decision module makes decisions and notifies on alert. Model matcher matches financial predictions. These are the details provided by financial sector predictors to predict and compare this year's finances. Figure 3 shows the feature extraction.

Decision module considers time of day and alert duration. Some alerts are triggered when an alert condition lasts for a long time, raising the possibility that a camera has been permanently damaged. Others are triggered when the alert condition lasts for seconds, raising the same possibility. Figure 4 shows anti-tampering architecture.

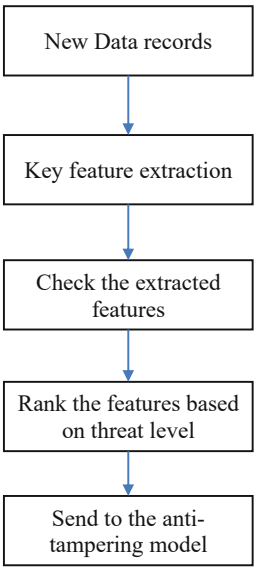


Fig. 3: Feature Extraction

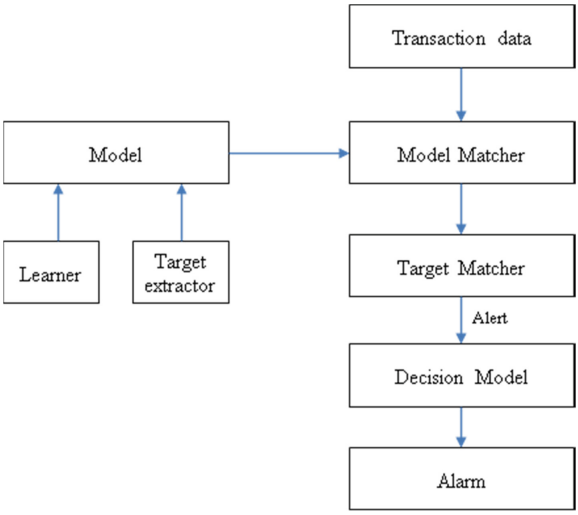


Fig. 4: Anti-tampering algorithm architecture

3.3 Deep Learning-Based BC Transaction

Blockchains and AI are emerging technologies being studied. Deep learning blockchains are updated by selecting the best computing methods to meet transactional needs. Deep learning models predict financial futures. The prediction level model must use deep

learning efficiently. The output results get financial predictions. Self-input and updates improve the database. Internal learning needs W1 and W2.

The following results Eq. (3) are if the data adheres to the W_1 learning principles:

$$\{x_1, y_1\}\{x_2, y_2\}...\{x_n, y_n\} \tag{3}$$

It is possible to write it as follows in Eq. (4):

$$qQV_{1,1}^{l} = qQV_{1,1}^{l-1} - \sigma\left(a^l - qQV_{1,1}^{l-1}\right) \tag{4}$$

where,

$QV_{1,1}^{l}$- weight matrix,

l- Input data,

q– total weight matrices.

To calculate the result, the following formula in Eq. (5) is used:

$$min\left(\frac{e_q}{e_g}, \frac{e_g}{e_q}\right) > \frac{1-\theta}{1+\theta} \tag{5}$$

where

eq $-fQV_{1,1}^{l-1}$ measurements

e.g. - $gQV_{1,1}^{l-1}$ measurements

When it comes to altering the weight matrices, the following is the equation Eq. (6) to be used:

$$qQV_{1,1}^{l} = qQV_{1,1}^{l-1} - \sigma\left(a^l - qQV_{1,1}^{l-1}\right)gqQV_{1,1}^{l} \tag{6}$$

In this case, W1 and W2 have the wrong data column, even though the imported data matches both. Unless otherwise specified, imported data will be mirrored in adjacent data at once. W1 will ensure a more precise first analysis, leading to more consistent and reliable results.

4 Results and Discussions

Information and data from distributed networks must be examined for ant tampering performance and improved through supply chain trials to prevent data forging.

In the experiment, two PCs have IIS5.0 and SWS databases. Most of the experiment data is simulated. Second, companies can test RDTP using MATLAB. Protocol experimental operations must be conducted with the same data to understand each case's safety performance. Generally, the finance network can keep adequate information.

RDTP article protocol limits the amount of data that can be updated, unlike PCI and ECDG. This proves that the article strategy can improve anti-attack for a distributed

supply chain finance network. Distributed blockchain data manipulation over financing is eliminated.

Most company receivable accounts can be turned into financing instruments and payment settlements, allowing us to engage in financing operations or external payments. The blockchain can link upstream core companies with downstream suppliers without capital transactions.

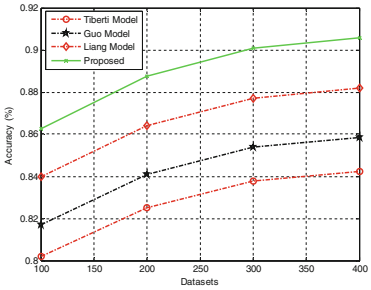


Fig. 5: Accuracy of Anti-Tampering Model

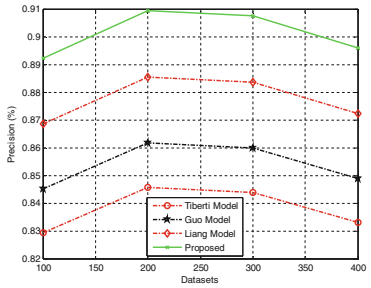


Fig. 6: Precision of Anti-Tampering Model

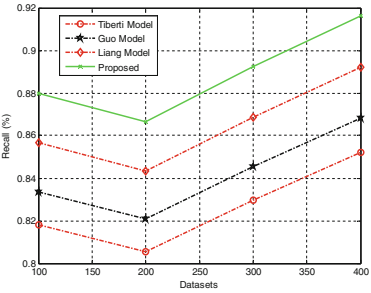


Fig. 7: Recall of Anti-Tampering Model

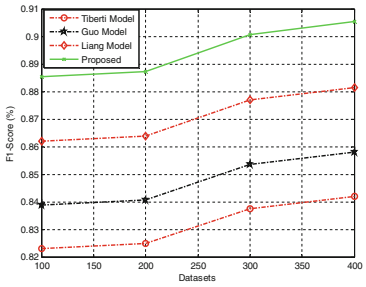


Fig. 8: F-measure of Anti-Tampering Model

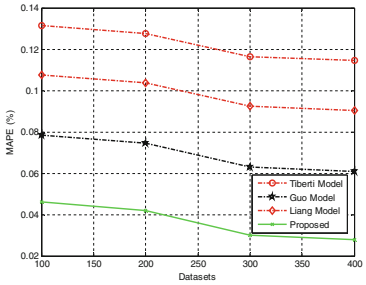


Fig. 9: MAPE of Anti-Tampering Model

Figures 5, 6, 7, 8 and 9 shows the graphical representation of accuracy, precision, recall, F-measure and MAPE of the anti-tampering model respectively. Increasing financial transparency reduces supply chain costs and speeds up financing. The MAPE reduction shows the reduction of the required flow. Once the values of the MAPE are increased, security threats will be available. So, need to supply some more unique features to reduce the weight of MAPE. It creates cost and time duration values that are higher. So, reducing MAPE improves results. By eliminating offline verification of accounts receivable legality, a bank can establish an electronic office for development.

5 Conclusions

Deep learning-based blockchains could improve capital flow stability, transaction security, and the value and competitiveness of the financial industry. Future financial trading will be impacted by blockchain's deep learning capabilities. Given current supply-chain finance research approaches, researchers must propose an encrypted blockchain-based method for protecting massive, distributed databases. The proposed model had 90.2% accuracy, 89.6% precision, 91.8% recall, 90.5% F1 Score, and 29% MAPE. This improves accuracy by setting up scattered data attributes and minimising the procedure. The proposed model's blockchain provides better protection than existing models. This proposed model will improve data security and access with enhanced code protection and storage. Big financial data is stored in cloud-based databases.

References

1. Liang, X., Xu, S.: Student performance protection based on blockchain technology. J. Phys.: Conf. Ser. **1748**(2), 022006) (2021). IOP Publishing
2. Li, X.: An anti-tampering model of sensitive data in link network based on blockchain technology. In: Web Intelligence (No. Preprint, pp. 1–11). IOS Press
3. Liu, W., Li, Y., Wang, X., Peng, Y., She, W., Tian, Z.: A donation is tracing blockchain model using improved DPoS consensus algorithm. Peer-To-Peer Netw. Appl. **14**(5), 2789–2800 (2021)
4. Zhang, F., Ding, Y.: Research on anti-tampering simulation algorithm of block chain-based supply chain financial big data. In: 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), pp. 63–66. IEEE (2021)
5. Zhang, Y., Zhang, L., Liu, Y., Luo, X.: Proof of service power: a blockchain consensus for cloud manufacturing. J. Manuf. Syst. **59**, 1–11 (2021)
6. Haoyu, G., Leixiao, L., Hao, L., Jie, L.I., Dan, D., Shaoxu, L.I.: Research and application progress of blockchain in area of data integrity protection. J. Comput. Appl. **41**(3), 745 (2021)
7. Jia, Q.: Research on medical system based on blockchain technology. Medicine, 100(16) (2021)
8. Yumin, S.H.E.N., Jinlong, W.A.N.G., Diankai, H.U., Xingyu, L.I.U.: Multi-person collaborative creation system of building information modeling drawings based on blockchain. J. Comput. Appl. **41**(8), 2338 (2021)
9. Li, F., Sun, X., Liu, P., Li, X., Cui, Y., Wang, X.: A traceable privacy-aware data publishing platform on permissioned blockchain. Trans. Emerg. Telecommun. Technol. e4455
10. Kuo, C.C., Shyu, J.Z.: A cross-national comparative policy analysis of the blockchain technology between the USA and China. Sustainability **13**(12), 6893 (2021)
11. Zhang, Z., Zhong, Y., Yu, X.: Blockchain storage middleware based on external database. In: 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), pp. 1301–1304. IEEE (2021)
12. Gong-Guo, Z., Zuo, O.: Personal health data identity authentication matching scheme based on blockchain. In: 2021 International Conference on Computer, Blockchain and Financial Development (CBFD), pp. 419–425. IEEE (2021)
13. Pang, Y., Wang, D., Wang, X., Li, J., Zhang, M.: Blockchain-based reliable traceability system for telecom big data transactions. IEEE Internet Things J. (2021)
14. Ma, J., Li, T., Cui, J., Ying, Z., Cheng, J.: Attribute-based secure announcement sharing among vehicles using blockchain. IEEE Internet Things J. **8**(13), 10873–10883 (2021)
15. Peter, G., Livin, J., Sherine, A.: Hybrid optimization algorithm based optimal resource allocation for cooperative cognitive radio network. Array **12**, 100093 (2021). https://doi.org/10.1016/j.array.2021.100093
16. Das, S.P., Padhy, S.: A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. Int. J. Mach. Learn. Cybern. **9**(1), 97–111 (2015). https://doi.org/10.1007/s13042-015-0359-0
17. Kumar, N.A., Shyni, G., Peter, G., Stonier, A.A., Ganji, V.: Architecture of network-on-chip (NoC) for secure data routing using 4-H function of improved TACIT security algorithm. Wirel. Commun. Mob. Comput. **2022**, 1–9 (2022). https://doi.org/10.1155/2022/4737569

A Handy Diagnostic Tool for Early Congestive Heart Failure Prediction Using Catboost Classifier

S. Mythili[1(✉)], S. Pousia[1], M. Kalamani[2], V. Hindhuja[3], C. Nimisha[3], and C. Jayabharathi[4]

[1] Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, India
mythilikarthikeyan911@gmail.com
[2] Department of ECE, KPR Institute of Engineering and Technology, Coimbatore, India
[3] UG scholar, Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, India
[4] Department of E&I, Erode Sengunthar Engineering College, Perundurai, India

Abstract. In this world 33% of deaths are due to the Cardiovascular diseases (CVDs) which affects the people globally irrespective of ages. Based on the saying "Prevention is better than cure" there is a necessity for early detection of heart failures. By addressing the behavioural risk factors such as tobacco use, obesity and harmful use of alcohol there is a possibility to circumvent it. But the people with disorders or more risk factors like hypertension, diabetes, hyperlipidaemia etc., need early detection and management wherein a machine learning model is of great help. The analyzed dataset contains 12 features that may be accustomed to predict the heart failure. Various algorithms such as SVM, KNN, LR, DT, Cat boost algorithms are taken into consideration in the aspect of accurate heart failure prediction. Analysed Cat boost classifier algorithm proves that it suits for earlier heart failure prediction with high accuracy level. It is further deployed in real time environment as a handy tool by integrating the trained accurate model with a User Interface for the heart failure prediction.

Keywords: Heart failure prediction · Machine learning model · Accuracy · Cat boost Classifier · User Interface

1 Introduction

Heart failure occurs when the heart cannot pump enough blood to meet the body's needs (HF) [1, 2]. Narrowing or blockage of the coronary arteries is the most frequent cause of heart failure. Coronary arteries are the blood vessels that deliver blood to the gut. Shortness of breath, swelling legs, and general weakness are some of the most typical heart failure signs and symptoms. Due to a shortage of trustworthy diagnostic equipment and inspectors, problem diagnosis might be challenging. Similar to other medical conditions, heart failure is typically diagnosed using a variety of tests suggested by Croakers, a patient's medical history, and an examination of associated symptoms. A

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1041–11, 2023.
https://doi.org/10.1007/978-3-031-27409-1_96

significant one of them is angiography, which is used to diagnose heart failure. This is considered to be an approach that could be helpful for identifying heart failure (HF). This diagnostic method is used to look for cardiovascular disease. Its high cost and related adverse effects have some restrictions. Advanced skills are also needed. Expert systems based on machine learning can reduce the health hazards connected to physical tests [3]. This permits quicker diagnosis [4].

Among them, angiography is acknowledged as a key tool for diagnosing HF. It is seen as a potentially useful technique for detecting cardiac failure (HF). This type of diagnostic seeks to establish cardiovascular disease. Because of its high cost and related side effects, it has some limitations. Additionally, it calls for a high level of competence. Expert systems based on machine learning can reduce the health hazards related to medical tests [3, 5, 6]. Additionally, it enables quicker diagnosis [4].

2 Literature Survey

Their main objective, according to a recent study article [7], is to create robust systems that can overcome challenges, perform well, and accurately foresee potential failures. The study uses data from the UCI repository and has 13 essential components. SVM, Naive Bayes, Logistic Regression, Decision Trees, and ANN were among the methods employed in this study. Up to 85.2% accuracy has been shown to be the best performance of SVM. Some applications of the work additionally involve a comparison of each technique. In this study, we also employ model validation techniques to construct the best correct model in a certain context.

According to a study [8, 9] that examined information from medical records, serum creatinine and ejection fraction alone are sufficient to predict longevity in individuals with coronary artery failure. Revealed by the model. It also demonstrates that utilizing the first dataset's function as a whole produces more accurate results. According to studies that included months of follow-up for each patient, serum creatinine and ejection fraction are the main clinical indications in the dataset that predict survival in these circumstances. When given a variety of data inputs, including clinical variables, machine learning models frequently produce incorrect predictions.

Solving typical machine learning challenges for heart disease prediction utilizing z-scores, min-max normalization, and artificial minority oversampling (SMOTE) techniques is the prevalent unbalanced class problem in this area examined in relation to the model [10, 11]. The findings demonstrate the widespread application of SMOTE and z-score normalization in error prediction.

Research [12–14] indicates that the subject of anticipating cardiac disease is still relatively new and that data are only recently becoming accessible. Numerous researchers have examined it using a range of strategies and techniques. To locate and forecast disease patients, data analytics is frequently used [15]. Three data analysis approaches (neural networks, SVM, and ANN) are used to datasets of various sizes to increase their relative accuracy and stability, starting with a preprocessing stage that uses matrices to choose the most crucial features. The neural network discovered is simple to set up and produces significantly superior results (93% accuracy).

3 Proposed Methodology

Machine learning models were able to predict heart failure with 70% to 80% accuracy using a variety of classification and clustering techniques, including k-means clustering, random forest regressors, logistic regression, and support vector machines. But CatBoost uses a decision tree method to improve gradients. App development and machine learning are the two key areas of interest for this project. The creation of models that are more accurate than current models is one of the main objectives of machine learning. In order to do this, many machine learning models have been investigated, with supervised machine learning models receiving special consideration because the dataset contains labeled data. The proposed flow is shown in Fig. 1. But keep in mind that in this case, unsupervised techniques like clustering are still applicable. This is due to the assertion being incorrect and the Matter statement treating the outcome as binary (anticipated illness/unexpected illness).

3.1 Dataset Collection

The first and most important step in the process is data collection. The platforms on which information is provided are numerous. The patient's confidentiality has been upheld. The clinical data set extracted from 919 patients is in the open source repository Kaggle. The dataset features considered are Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, Maximum Heart Rate, Exercise Angina, Old Peak, ST_Slope. As shown in Fig. 2, in this study, 70% of the data is trained which contains the count of 644, and 30% of the data that is 275 samples are considered for testing.

3.2 Data Analyzing

The facts must be understood in order to proceed with Data analysis. To handle the missing data, negative values, and undesirable strings, and to convert given values to integers, pre-process the input the data analysis is important and done here in these aspects.

3.3 Feature Selection and Exploratory Data Analysis

To reduce computational complexity and enhance model performance the significant parameters are chosen that influence model correctness. Following the completion of the link between features, the major characteristics are chosen in this step. Datasets and traits that affect findings are studied and graphically displayed as EDA in Fig. 3. Its corresponding features correlation matrix plot is shown in Fig. 4.

3.4 Fitting into the Model

The model can now process the data because it has undergone preprocessing. Several models, including Decision Trees (DT), Support Vector Machine (SVM), Logistic

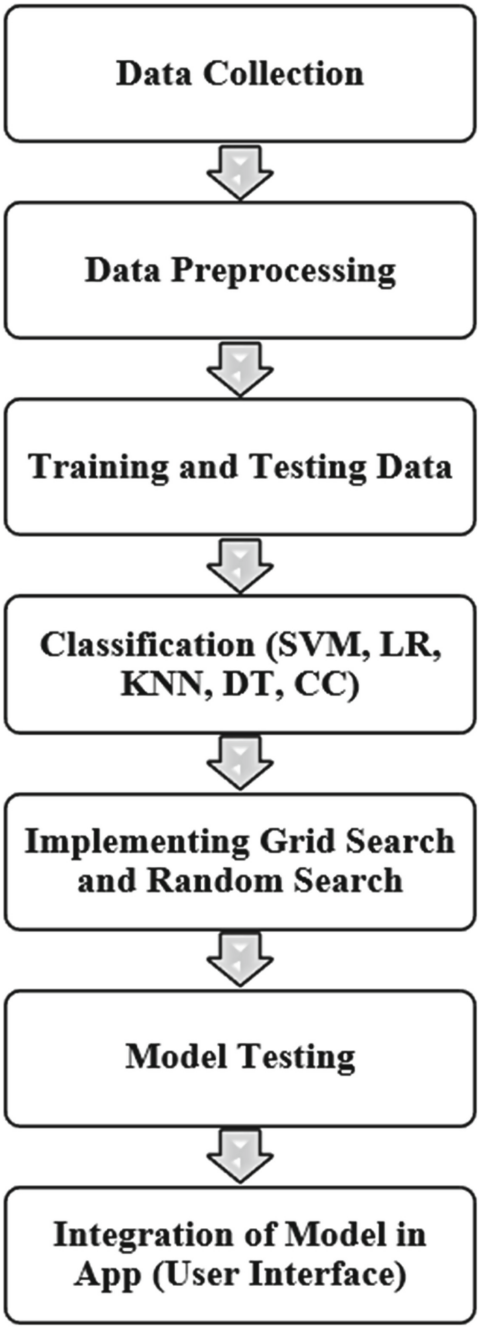


Fig. 1. Work flow of the proposed methodology

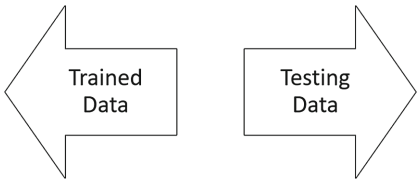


Fig. 2. Data Collection

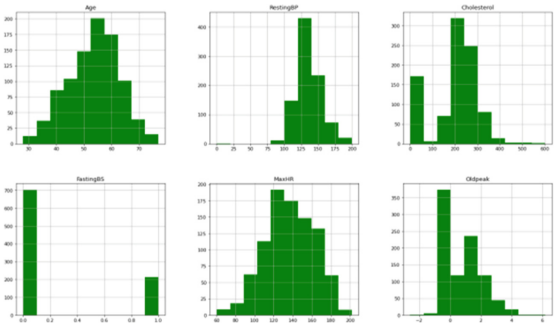


Fig. 3. Exploratory Data Analysis (EDA)

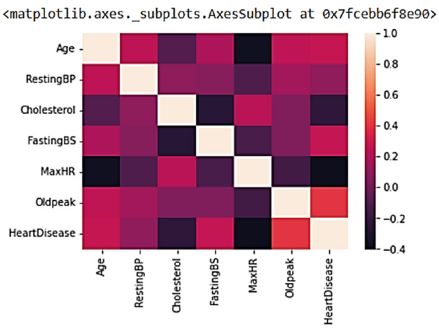


Fig. 4. Correlation Matrix

Regression (LR), K-Nearest Neighbour (KNN) and Catboost Classifier are built using this dataset.

Decision Tree: Every leaf node in a decision tree may be a category label, similar to a flowchart, and every interior node could be a "test" for an attribute (e.g., whether a coin lands heads or tails). By arranging the instances in a 10-tree from a base node to leaf nodes that provide a categorization of the instance, a decision tree may categorize an instance. Starting with the root node of the tree, evaluating the attribute represented by that node, and then following the branches that approximate the value of the attribute are the steps used to classify instances.

Support Vector Machine: Regression and classification issues are resolved with it, which supports Vector Machines or SVMs. However, its main application is to machine learning classification issues. The SVM technique aims to provide appropriate decision boundaries, or lines, that categorise the dimensional space in order to facilitate the addition and arrangement of new data in the future with the least amount of disruption. This ideal decision boundary is called a hyperplane. The SVM chooses strong point vectors that are then utilised to build the hyperplane. Support vector machines is the name given to this approach for this reason. "Support vectors" are used to describe these atypical circumstances.

Logistic Regression: Supervised classification is the primary characteristic of logistic regression. In a classification task, only discrete values of X, target variables, and Y are possible for a given set of features (or inputs). Regression models may include logistic regression, contrary to popular opinion. This model generates a regression model that forecasts the likelihood that a particular piece of input data falls into the position denoted by the number "1." In a manner similar to how linear regression assumes that the data are distributed linearly, logistic regression models the data using a sigmoid function.

K Nearest Neighbour(KNN): The K-nearest-neighbor algorithm, sometimes referred to as KNN or k-NN, is a nonparametric supervised learning classification that makes use of proximity to anticipate or categorize how specific data points would be grouped together. It is an object. Because it depends on the likelihood that analogous points would be located nearby, it can be used to solve classification or regression issues, but it is most frequently employed as a classification algorithm.

Suggested model: Cat boost Classifier: Catboost Classifier is the suggested method for the Congestive Failure Prediction. CatBoost is a technique for decision trees that uses gradient boosting. It was created by Yandex engineers and researchers and is used for a variety of activities including weather forecasting, self-driving cars, personal assistants, and search grade. An ensemble machine learning approach called boosting is typically applied to classification and regression issues. It is easy to use, handles heterogeneous data well, and even handles relatively tiny data. In essence, it builds a strong learner out of a collection of weak ones. Numerous strategies exist to handle categorical characteristics in boosted trees, which are typically present in datasets. CatBoost automatically handles categorical features in contrast to other gradient boosting techniques (which need numeric input).One of the most popular methods for processing categorical data is one-hot coding, however it is not viable for many tasks. To overcome this, traits are categorized using goal statistics (assumed target values for each class). Target stats are typically determined using a variety of strategies, including greedy, holdout, leave-one-out, and ordered. CatBoost provides a summary of the target stats.

Features of the Cat Boost Classifier:

- Cat boosting will not function if the primary cat trait boost classifier trait column index is not recognised as a cat trait and the categorical trait is manually coded. Without it, categorical features cannot be subjected to Catboost preprocessing.
- Catboost employs one-hot encoding for all functions at the highest one-hot-max-size-unique value. In this instance, hot coding wasn't used. This is caused by how many

distinctive values there are for the categorical traits. The value, though, is based on the data you gather.

- Learning rate and N-estimators: The learning rate decreases as the number of n estimators needed to use the model increases. Typically, this method starts with a learning rate that is quite high, adjusts other parameters, and subsequently lowers learning rate while increasing the number of estimators.
- *max depth:* Base tree depth; this value greatly affects training time.
- *Subsample:* Sample rate of rows; incompatible with Bayesian boosting.
- Column sample rates include colsample by level, colsample by the tree, and colsample by the node.
- L2 regularization coefficient (l2 leaf)
- Every split is given a score, and by adding some randomness to the score with random strength, overfitting is lessened.

The cat boost classifier requires proper tuning of hyper parameters for its greatest performance. Optimizing the hyper parameter tuning is a great challenge while working with the cat boost classifier algorithm as its performance can be very bad if the variables are not properly tuned. To overcome the tuning issues, two optimization techniques can be implemented to the algorithm hyper parameters to automatically tune the variables, thus increasing the performance of the cat boost classifier. The optimization technique implemented using grid search that works on brute force method by creating a grid of all possible hyper parameter combinations and other techniques is random search, that does not include all combinations but random combinations of hyper parameters. This automatically navigates the hyper parameter space. Thus combining both the optimization techniques will lead to better performance of the cat-boost algorithm.

4 Results and Discussion

The confusion matrix of LR, KNN, DT and Catboost classifier is shown in Fig. 5. The confusion matrix values include True negatives, True positives, False negatives and False positives as represented in Fig. 6.

It can be used to calculate the accuracy using the formula given as follows: Accuracy = TN + TP/ [TN + TP + FP + FN].

The machine learning models' varying degrees of accuracy based on prediction is shown in Table 1. A comparison of different classifiers (SVM, LR, KNN, DT, Catboost) can be seen in Fig. 7.

As compared to all other classifiers, Cat boost is more accurate. So the cat boost classifier is suited to be the best in-terms of accuracy. But with a reasonable factor, the prediction model suitability cannot be judged. For a detailed view, the model evaluation of the preferred Cat boost classifier is done and it is shown in Fig. 8 by taking many more parameters into consideration for analysis.

The results of model evaluation also proved that it is the most suitable for Congestive heart disease prediction.

Figure 9 indicates the accuracy, AUC, Recall, Precision, F1 Score, Kappa, and MCC for the suggested cat boost classifier.

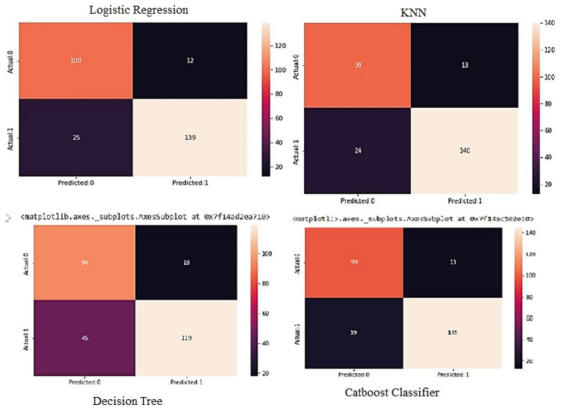


Fig. 5. Confusion matrix of the models

TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

Fig. 6. Confusion matrix of the Cat Boost Classifier

Table 1: Tested algorithms with its accuracy

Tested Algorithms	**Accuracy**
Support Vector Machine	88.40%
Logistic Regression	86.59%
K Nearest Neighbor	85.54%
Decision Tree	77.17%
Catboost Classifier	88.59%

Using the Cat boost classifier's model, the tuned accuracy level is 88%. The accuracy shows that it performs well and is quicker for prediction. The prediction process using Cat boost requires less time and has proven that with its accuracy. The hyper parameter tuning achieves a higher F1 score of 0.9014. The joint effect of larger and smaller values indicates that it supports for an optimal interpretation. It improves the performance and this indicates perfect precision and recall are possible. And it is proved with the cat boost hyper parameter tuning algorithm. The higher the recall more the positive test sample detection. Based on the Grid search and Random search the recall measure is high for the

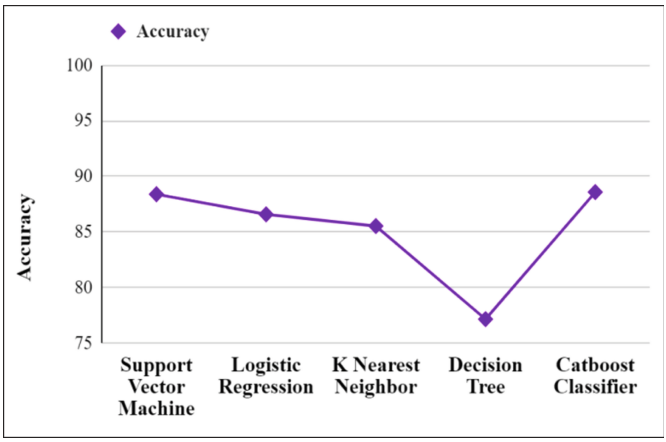


Fig. 7. Comparison of different Classifiers with Cat boost classifier

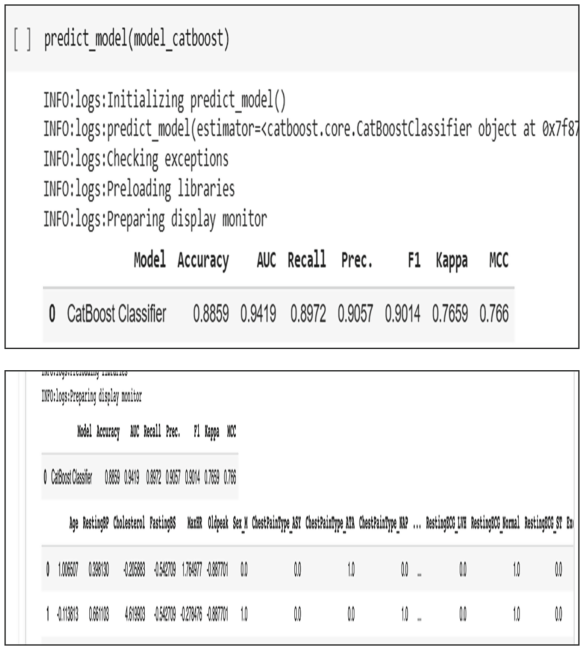


Fig. 8. Model Evaluation of Cat boost classifier

actual prediction. It is 0.8972 for the proposed model. The kappa range of the proposed algorithm shows better agreement and a good rating for the patient data evaluation. The system is achieved as reliable with a kappa score of 0.7659.

The Mathews correlation coefficient is obtained by using the formula.

MCC = [(TP*TN) – (FP*FN)]/ sqrt [(TP + FP)(TP + FN)(TN + FP)(TN + FN)].

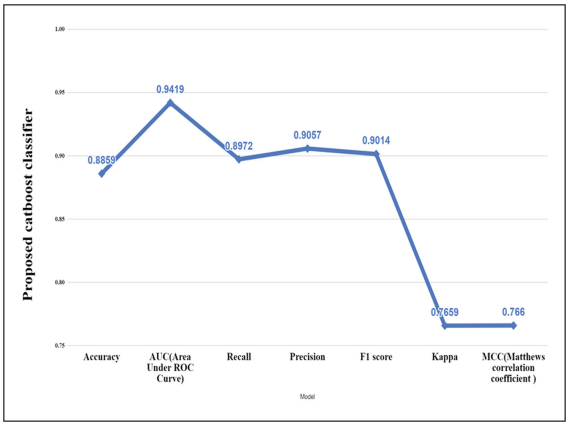


Fig. 9. Analysis of Cat boost classifier

With this equation for the proposed model the obtained MCC value is 0.766. As a whole the overfitting issues will not exist in the run model for congestive heart failure prediction using Cat boost classifier. By this validation it is further taken into the development phase of web application with a user interface for easy visibility of the cardiac status at the doctor's side.

5 User Interface

The identified best model is the Cat boost classifier which is converted to pickle file using the Pickle Library in Python. This pickle file is used to develop an API to pass the input data in the form of json format and get the output. The output will display that whether the person will experience heart failure in the future or not based on the model trained and the file fed using the API.

The user interface has been designed using Flutter and the input from the user is then passed to the API and the response is obtained as shown in Fig. 10. By this way the handy early stage prediction of heart failure is done.

6 Conclusion

Heart failure could be a regular event caused by CVDs and it needs wider attention on the early stage itself. As per the study, one of the solutions is machine learning model deployment for its prediction. The analysis is done on different models with a kaggle dataset. Based on the trained and tested datasets the accurate model is predicted with various parameters analysis such as AUC, Recall, Precision, Kappa value and MCC. Among SVM, LR, KNN, Decision Tree algorithms, the Catboost Classifier's output has the highest accuracy, 88.59%. Therefore, this model is implemented in a mobile application with an effective user interface. Further doctors can use it to forecast a patient's likelihood of experiencing heart failure and to make an early diagnosis in order to save the patient's life.



Fig. 10. User Interface and Model Deployment

References

1. Huang, H., Huang, B., Li, Y., Huang, Y., Li, J., Yao, H., Jing, X., Chen, J., Wang, J.: Uric acid and risk of heart failure: a systematic review and meta-analysis. Eur. J. Heart Fail. **16**(1), 15–24 (2014). https://doi.org/10.1093/eurjhf/hft132.Epub. 2013 Dec 3. PMID: 23933579
2. Ford, I., Robertson, M., Komajda, M., Böhm, M., Borer, J.S., Tavazzi, L., Swedberg, K.: Top ten risk factors for morbidity and mortality in patients with chronic systolic heart failure and elevated heart rate: the SHIFT Risk Model. © 2015 Elsevier Ireland Ltd. All rights reserved. Int. J. Cardiol. 184C (2015). https://doi.org/10.1016/j.ijcard.2015.02.001
3. Olsen, C.R., Mentz, R.J., Anstrom, K.J., Page, D., Patel, P.A.: Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. Am. Heart J. (IF 5.099) Pub Date: 2020–07–16. https://doi.org/10.1016/j.ahj.2020.07.009
4. Olsen, C.R., Mentz, R.J., Anstrom, K.J., Page, D., Patel, P.A.: Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. Am. Heart J. **229**, 1–17 (2020). https://doi.org/10.1016/j.ahj.2020.07.009. Epub 2020 Jul 16. PMID: 32905873
5. Held, C., Gerstein, H.C., Yusuf, S., Zhao, F., Hilbrich, L., Anderson, C., Sleight, P., Teo, K.: ONTARGET/TRANSCEND investigators. Glucose levels predict hospitalization for congestive heart failure in patients at high cardiovascular risk. Circulation. **115**(11), 1371–1375 (2007). https://doi.org/10.1161/CIRCULATIONAHA.106.661405. Epub 2007 Mar 5. PMID: 17339550
6. Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jr, Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T., Jr, Roccella, E.J.: Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension **42**(6), 1206–1252 (2003). https://doi.org/10.1161/01.HYP.0000107251.49515.c2. Epub 2003 Dec 1. PMID: 14656957
7. Sahoo, P.K., Jeripothula, P.: Heart Failure Prediction Using Machine Learning Techniques (December 15, 2020). http://dx.doi.org/https://doi.org/10.2139/ssrn.3759562
8. Chicco, D., German, N.G.: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med. Inform. Decis. Mak. **20**, 16 (2020). ISSN: 1472-6947, https://doi.org/10.1186/s12911-020-1023-5

9. Wang, J.: Heart failure prediction with machine learning: a comparative study. J. Phys.: Conf. Ser. **2031**, 012068 (2021). https://doi.org/10.1088/1742-6596/2031/1/012068
10. Wang, J.: Heart failure prediction with machine learning: a comparative study. J. Phys: Conf. Ser. **2031**, 012068 (2021). https://doi.org/10.1088/1742-6596/2031/1/012068
11. Ali, L., Bukhari, S.A.C.: An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. IRBM **42**(5), 345–352 (2021). ISSN 1959-0318. https://doi.org/10.1016/j.irbm.2020.04.003
12. Salhi, D.E., Tari, A., Kechadi, M.-T.: Using machine learning for heart disease prediction. In: Senouci, M.R., Boudaren, M.E.Y., Sebbak, F., Mataoui, M. (eds.) CSA 2020. LNNS, vol. 199, pp. 70–81. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69418-0_7
13. J. Am. Coll. Cardiol. 2005 **46**(6), e1–82 (2005). https://doi.org/10.1016/j.jacc.2005.08.022
14. Fang, H., Shi, C., Chen, C.-H.: BioExpDNN: bioinformatic explainable deep neural network. IEEE Int. Conf. Bioinform. Biomed. (BIBM) **2020**, 2461–2467 (2020). https://doi.org/10.1109/BIBM49941.2020.9313113
15. Dangare, C.S., Apte, S.S.: Improved study of heart disease prediction system using data mining classification techniques. Int. J. Comput. Appl. **47**(10), (2012). https://doi.org/10.5120/7228-0076

Hybrid Convolutional Multilayer Perceptron for Cyber Physical Systems (HCMP-CPS)

S. Pousia[1], S. Mythili[1](✉), M. Kalamani[2], R. Manjith[3], J. P. Shri Tharanyaa[4], and C. Jayabharathi[5]

[1] Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, India
mythilikarthikeyan911@gmail.com
[2] Department of ECE, KPR Institute of Engineering and Technology, Coimbatore, India
[3] Department of ECE, Dr. Sivanthi Aditanar College of Engineering, Tiruchendur, India
[4] Department of ECE, VIT University, Bhopal, India
[5] Department of E&I, Erode Sengunthar Engineering College, Perundurai, India

Abstract. Due to the rapid growth of cyber-security challenges via sophisticated attacks such as data injection attacks, replay attacks, etc., cyber-attack detection and avoidance system has become a significant area of research in Cyber-Physical Systems (CPSs). It is possible for different attacks to cause system failures, malfunctions. In the future, CPSs may require a cyber-defense system for improving its security. The different deep learning algorithms based on cyber-attack detection techniques have been considered for the detection and mitigation of different types of cyber-attacks. In this paper, the newly suggested deep learning algorithms for cyber-attack detection are studied and a hybrid deep learning model is proposed. The proposed Hybrid Convolutional Multilayer Perceptron for Cyber Physical Systems (HCMP-CPS) model is based on Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP). The HCMP-CPS model helps to detect and classify attacks more accurately than the conventional models.

Keywords: Cyber Physical System · Deep Learning algorithm · Cyber Attack

1 Introduction

Technology advancements and the accessibility of everything online have significantly expanded the attack surface. Despite ongoing advancements in cyber security, attackers continue to employ sophisticated tools and methods to obtain quick access to systems and networks. To combat all the risks, we confront in this digital era, cyber security is essential. Hybrid deep learning models should be used in new cyber-attack detection in order to secure sensitive data from attackers and hackers and to address these issues [1].

In DDoS attacks, numerous dispersed sources are flooded with traffic of overwhelming volume in order to render an online service inaccessible [9]. News websites and banks are targeted by these attacks, which pose an important barrier to the free sharing and obtaining of vital information [2]. DDoS attacks mimic browser requests that load a web

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1053–1063, 2023.
https://doi.org/10.1007/978-3-031-27409-1_97

page by making it appear as if web pages are being attacked. An individual website could be accessed and viewed by hundreds of people at once. The website hosts are unable to offer service due to the enormous volume of calls, which results in notifications. This prevents the public from accessing the site. The afflicted server will get a lot of information quickly in the event of a DDoS assault [10]. This information is not same but share same features and divided into packets. It can take some time to recognize each of these applications as a part of an adversarial network. In contrast, each packet is a piece of a wider sequence that spans through time that may assess them all at once to ascertain their underlying significance [11]. In essence, time-series data provides a "big picture" that enables us to ascertain whether your server is being attacked. It therefore draws the conclusion that it is always advisable to consider the time each data point is in crucial information [6].

The main objectives of the suggested approach are identifying network invaders and safeguarding computer networks from unauthorized users, including insiders. Creating a prediction model (hybrid model) that can differentiate between "good" regular connections and "bad" connections is the aim of the Intrusion Detect Learning Challenge (often called intrusions or attacks). Hybrid deep learning models along with dataset properties are used to detect the cyber dangers. This technology's goal is to identify the most prevalent cyber threats in order to protect computer networks. In order to prevent data loss or erasure, cybersecurity is crucial. This includes private information, Personally Identifiable Information (PII), Protected Health Information (PHI), information pertaining to intellectual property, as well as systems and enterprises that use information that are used by governments.

2 Literature Survey

The proposed method can be used to create and maintain systems, gather security data regarding intricate IoT setups, and spot dangers, weaknesses, and related attack vectors. Basically smart cities rely heavily on the services offered by a huge number of IoT devices and IoT backbone systems to maintain secure and dependable services. It needs to implement a fault detection system that can identify the disruptive and retaliatory behavior of the IoT network in order to deliver a safe and reliable level of support. The Keras Deep Learning Library is used to propose a structure for spotting suspicious behavior in an IoT backbone network.

The suggested system employs four distinct deep learning models, including the multi-layer perceptron (MLP), convolutional neural network (CNN), deep neural network (DNN), and autoencoder, to anticipate hostile attacks. Two main datasets, UNSW-NB15 and NSLKDD99, are used to execute a performance evaluation of the suggested structure, and the resulting studies are examined for accuracy, RMSE, and F1 score. The Internet of Things (IoT), particularly in the modern Internet world, is one of the most prevalent technical breakthroughs. The Internet of Things (IoT) is a technology that gathers and manages data, including data that is sent between devices via protocols [3]. Digital attacks on smart components occur as a result of networked IoT devices being connected to the Internet. The consequences of these hacks highlight the significance of IoT data security. In this study, they examine ZigBee, one of his most well-known

Internet of Things innovations. We provide an alternative model to address ZigBee's weakness and assess its performance [5]. Deep neural networks, which can naturally learn fundamental cases from a multitude of data, are one of the most intriguing AI models. It can therefore be applied in an increasing variety of Internet of Things (IoT) applications [6].

In any event, while developing deep models, there are issues with vanishing gradients and overfitting. Additionally, because of the numerous parameters and growth activities, the majority of deep learning models cannot be used lawfully on truck equipment. In this paper, we offer a method to incrementally trim the weakly related loadings, which can be used to increase the slope of conventional stochastic gradients. Due to their remarkable capacity to adapt under stochastic and non-stationary circumstances, assisted learning techniques known as learning automata are also accessible for locating weakly relevant loads. The suggested approach starts with a developing neural system that is completely connected and gradually adapts to designs with sparse associations [7, 8, 12–14].

3 Hybrid Convolutional Multilayer Perceptron for Cyber Physical Systems (HCMP-CPS) Model

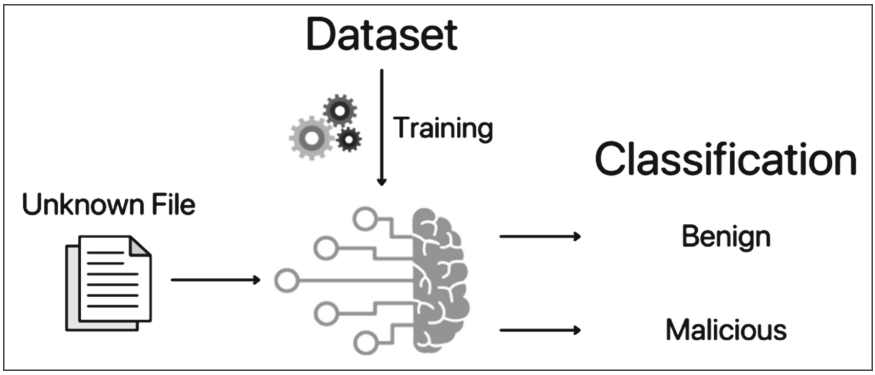


Fig. 1. Cyber-attack Prediction

Figure 1 depicts the structure of a cyber-detection system. Modern cyber-attack detection systems, as shown in Fig. 2, use hybrid deep learning models to identify cyber-attacks based on numerous traits gathered from datasets with four distinct attack classifications. To more precisely identify and categorize different sorts of attacks, a hybrid approach integrating CNN, MLP, and LSTM is applied.

The following stages describe cyber-attack identification using deep learning:

Step-1: First, import every library that will be used for next implementations ie. Matplot lib, Pandas, and Numpy.

Step-2: Import the NSL-KDD dataset and divide it.

Step-3: Dividing the dataset used to create the model (training and testing).

Step-4: To choose the most pertinent features for the model, the top features from the dataset were chosen.

Step-5: Analyze the features from the dataset such as protocol type, service, flag and attack distributions by using the EDA process.

Step-6: Create classification models with various layers and related activation functions using LSTM, MLP, and CNN.

Step-7: Create a high-fidelity hybrid model by combining multiple layers of one model.

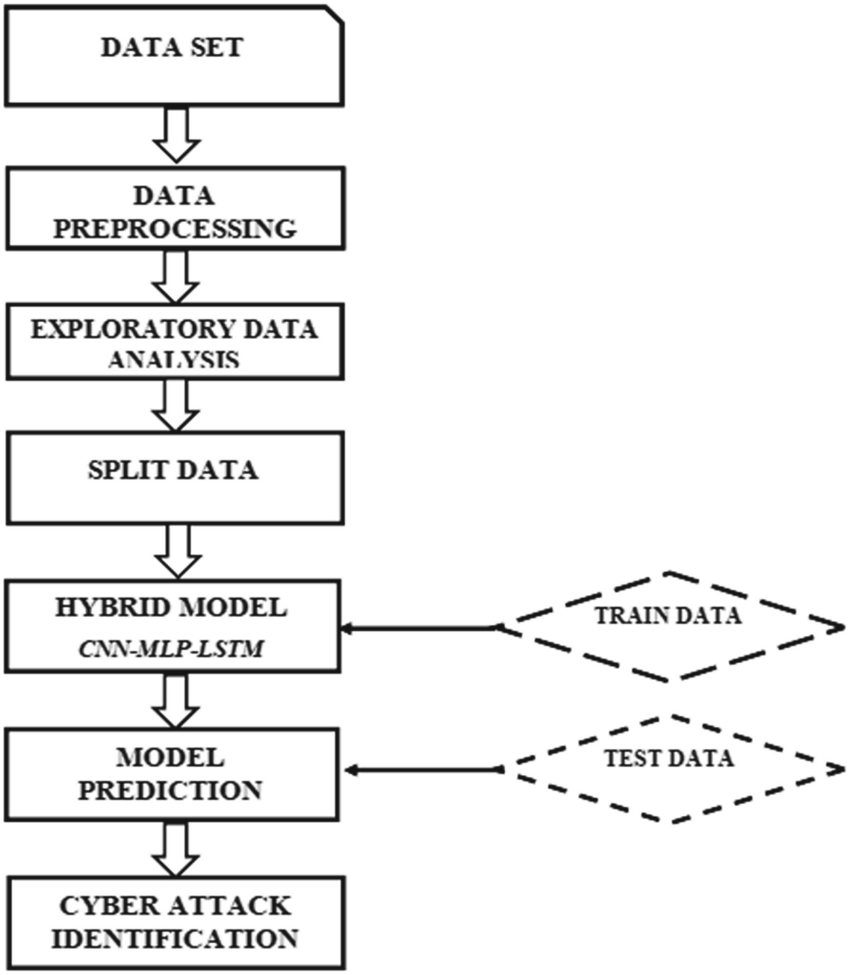


Fig. 2. Cyber-attack identification using deep learning

3.1 Dataset

There are 120000 records in the NSL-KDD data collection overall (80% training records and 20% testing records). The epoch in this situation denotes how many times the loop has finished. An entire data collection cannot be given to a neural network at once. The training data set is then used to build a stack.

- ***DOS:*** Attacks that cause denial of service restrict targets from valid inquiries like: Syn Flooding from resource depletion. Different attack types include Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, UDP Storm, Process table, and Worm.
- ***Probing:*** To gather more information about the distant victim, surveillance and other probing attacks, such as port scanning are used. Source Bytes and "Connection Time" are significant factors. Attack kinds include Satan, Ipsweep, Nmap, Portsweep, Mscan, and Saint.
- ***U2R:*** In order to get access to root or administrator credentials, an attacker could try to log in to a victim's machine using a regular account. If an attacker has the ability to access the user's local super user without permission, then this will happen (root). There is a connection between the attributes "number of files produced" and "number of shell prompts executed." A few examples of diverse assaults include buffer overflows, load modules, rootkits, Perl, SQL attacks, Xterms, and Ps.
- ***R2L:*** An attacker can enter a victim's computer without authorization and get local access by using a remote computer. At the network level, connection time and service request characteristics, as well as host-level information related to the number of failed login attempts. Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmp Guess, Snmp GetAttack, HTTP Tunnel, and Password Guessing are some examples of attack types.

3.2 Data Pre-processing

Data processing was necessary once the information was gathered from the dataset depicted in Fig. 3. Here, the model-best features are utilized. By using trait selection, the most important traits are chosen and included in the model. It aids in identifying optional features as well. Performance may be enhanced, and overfitting may be decreased. Hybrid models work well with organized data.

3.3 Exploratory Data Analysis

The EDA technique is used to perform data analysis. By carefully inspecting the dataset, it is able to draw conclusions about potential trends and outliers. EDA is a technique for investigating the implications of data for modeling. Distribution, protocol types, services, flags, and attack distribution all require EDA.

3.4 Data Splitting

The processed data were split into training and test sets using data partitioning. Analyzing model hyper parameters and generalization performance is possible using this strategy. Figure 4 illustrates hybrid models for cyber-attack prediction.

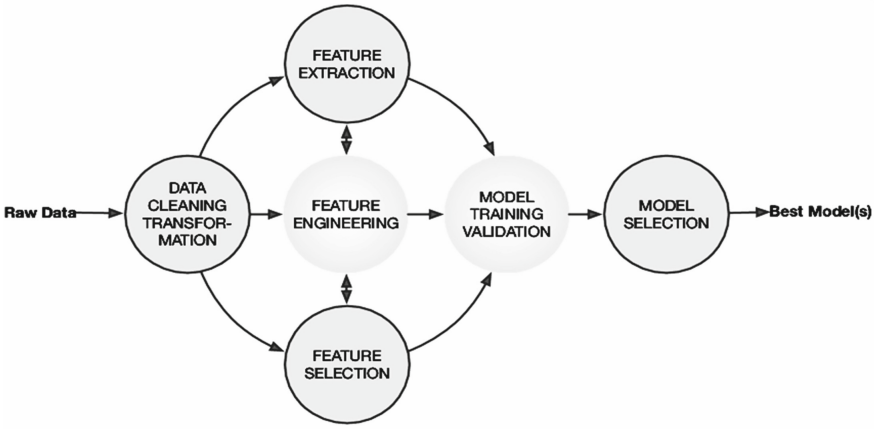


Fig. 3. Data Preprocessing

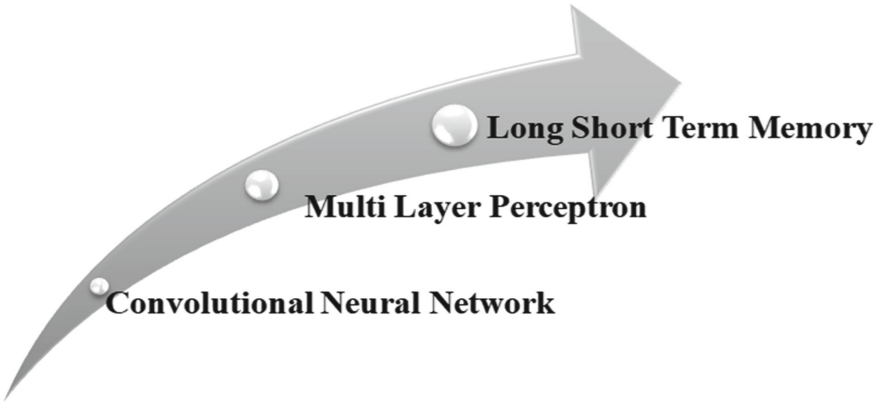


Fig. 4. Hybrid Model for Cyber-attack prediction

3.5 Hybrid Model Creation

To more precisely identify and categorize different kinds of assaults, hybrid algorithms including CNN, MLP, and LSTM were developed. Results for TCP, UDP, and ICMP protocols will differ depending on attacks such as DoS, probing, R2L, and U2R.

Convolutional Neural Network (CNN): As seen in Fig. 5, the CNN automatically classifies the data in this instance and offers a better classification. A second neural network classifies the features that a CNN pulls from the input dataset. A feature extraction network makes use of several input data sets. The received feature signal is used for categorization by a neural network. The network's output layer has a fully linked soft max as well as three average pooling layers and a convolutional layer. In this instance, the output tensor is produced by convolving the convolution kernel with the input layer in one spatial dimension using the CNN's Conv1D layer. Each layer of a neural network

contains neurons that calculate a weighted average of the inputs in order to send them via nonlinear functions.

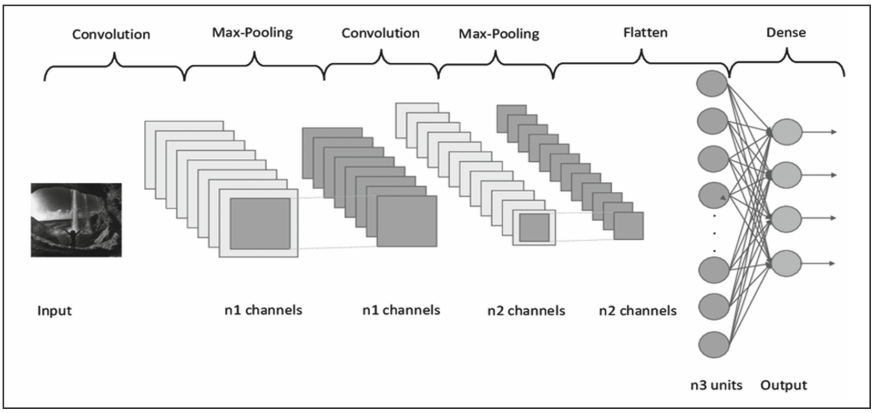


Fig. 5. CNN

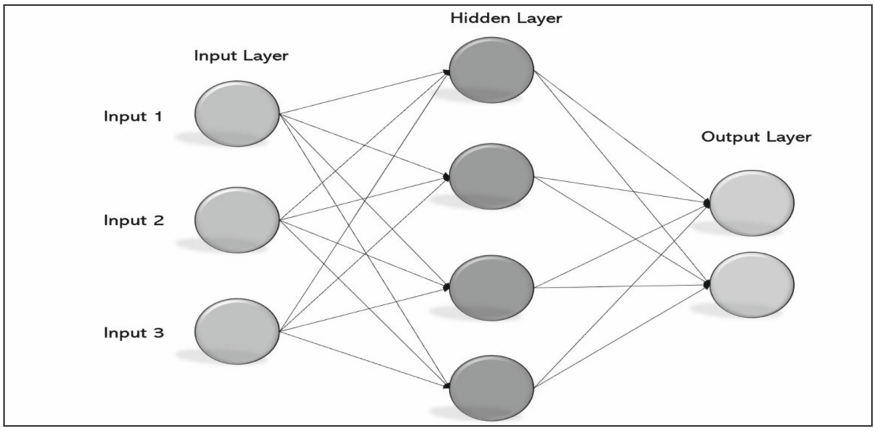


Fig. 6. MLP

Multi-layer Perceptron (MLP): Fig. 6 illustrates how MLP is used to recognize attacks as a successful method of thwarting cyberattacks. Since there are more levels in this algorithm, it is less vulnerable to hacking. MLP employs hidden layers to nonlinearly adjust the network's input.

LSTM: LSTM is capable of learning the qualities from the data collection that the training phase's data extraction was required to offer. This feature enables the model to discriminate between security hazards and regular network traffic with accuracy. Long Short-Term Memory is a type of artificial recurrent neural network (RNN) architecture used in deep learning [4]. LSTM networks can be used to analyze sequence data and

produce predictions [15] based on various sequence data time steps as illustrated in Fig. 7.

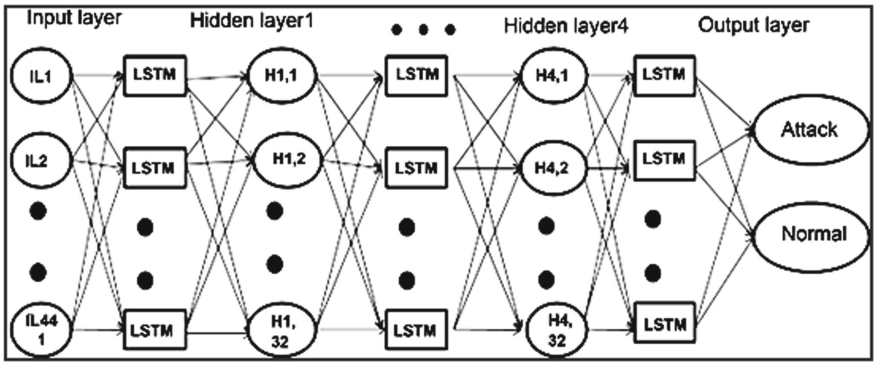


Fig. 7. LSTM

4 Results and Discussion

To assess how well hybrid deep learning models, work at spotting cyberattacks, confusion matrices are utilized. Four dimensions are shown in Table 1. There are both positive and negative categories in the classes that are positive and negative. They are called TP, FP, TN, and FN. TCP, UDP, and ICMP protocols are impacted by DoS attacks, probes, R2L, and U2R. A score that satisfies both the anticipated and actual criteria for a positive score in detecting cyberattacks is known as a true positive (TP) score. A value that is nonnegative but actually ought to be negative is referred to as a false negative value (FN). A value is considered a true negative if it is both lower than expected and lower than reality (TN).

Table 1: Confusion matrix

Model	TP	TN	FP	FN
CNN	**52**	**49**	**2**	**4**
LSTM	**53**	**52**	**2**	**5**
MLP	**51**	**46**	**4**	**6**
Hybrid	**57**	**55**	**1**	**2**

Figures 8,9,10 presents an analysis of the accuracy, precision and F1 score for various deep learning models in comparison with Hybrid model and it demonstrates that the proposed HCMP-CPS outperforms other conventional methods in identifying cyber threats.

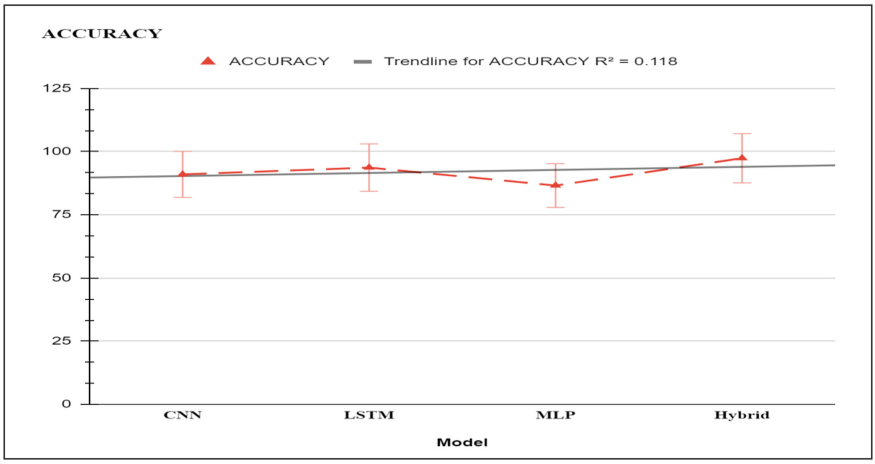


Fig. 8. Performance comparison of Accuracy in various Deep Learning model

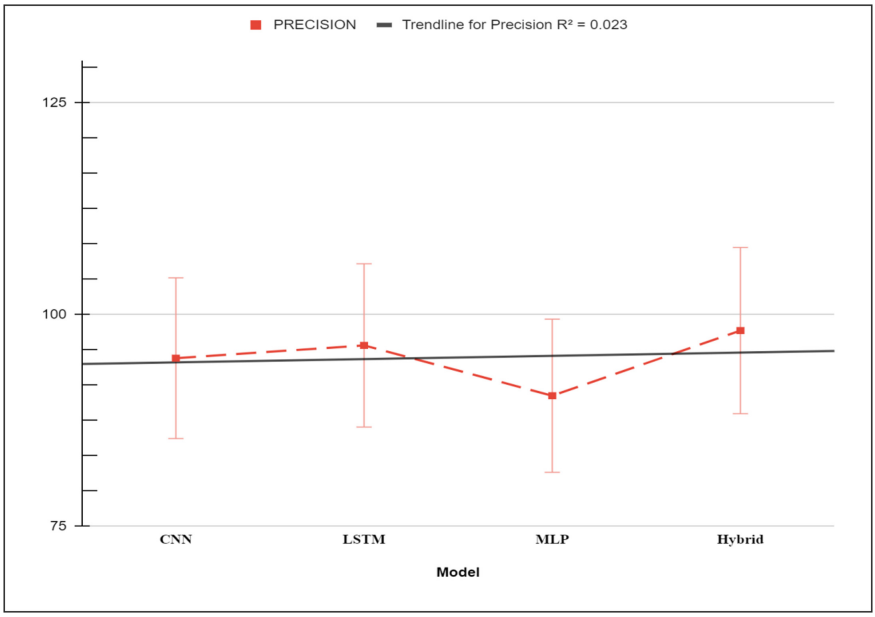


Fig. 9. Performance comparison of Precision in various Deep Learning model

5 Conclusion

The suggested system's primary function is to detect network intruders and protect against unauthorized users. Cyberattack detection using the proposed Hybrid Convolutional Multilayer Perceptron for Cyber Physical Systems (HCMP-CPS) model analyses the features from various datasets to detect intrusions or cyberattacks using features

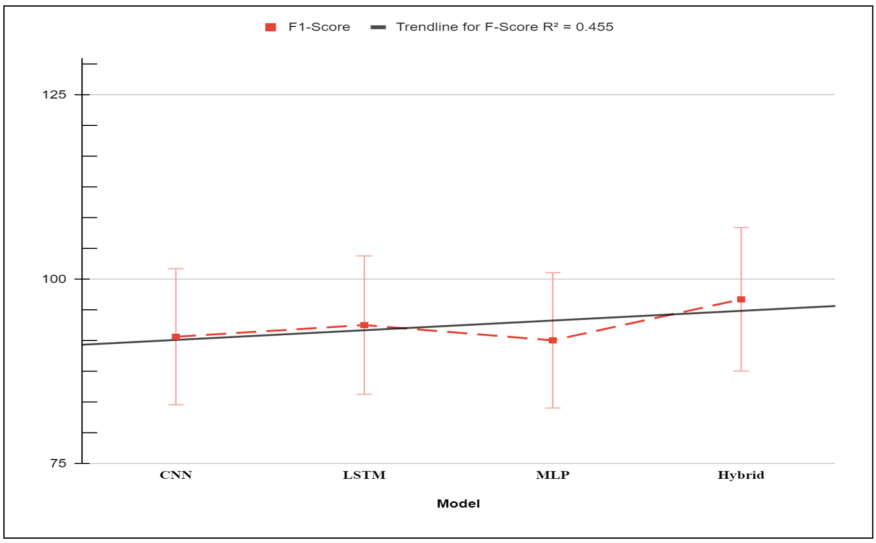


Fig. 10. Performance comparison of F1-Score in various Deep Learning model

gleaned from datasets. To evaluate the model's ability to detect and rate the cyberattacks, the NSL-KDD dataset is ideal. Attack predictions using HCMP-CPS improves detection accuracy to an average of 96%.

References

1. Barati, M., Abdullah, A., Udzir, N.I., Mahmod, R., Mustapha, N.: Distributed denial of service detection using hybrid machine learning technique. In: Proceedings of the 2014 International Symposium on Biometrics and Security Technologies (ISBAST), pp. 268–273, Kuala Lumpur, Malaysia, August. 2014
2. Chong, B.Y., Salam, I.: Investigating Deep Learning Approaches on the Security Analysis of Cryptographic Algorithms. Cryptography, vol. 5, p. 30 2021. https:// doi.org/https://doi.org/10.3390/cryptography5040030
3. Ghanbari, M., Kinsner, W., Ferens, K.: Detecting a distributed denial of service attack using a preprocessed convolutional neural network. In: Electrical Power and Energy Conference, pp. 1–6. IEEE (2017)
4. Goh, J., Adepu, S., Tan, M., Lee, Z.S.: Anomaly detection in cyber physical systems using recurrent neural networks. In: International Symposium on High Assurance Systems Engineering, pp. 140–145. IEEE (2017)
5. He, Y., Mendis, G.J., Wei, J.: Real-time detection of false data injection attacks in smart grid: a deep learning based intelligent mechanism. IEEE Trans. Smart Grid **8**(5), 2505–2516 (2017)
6. Hodo, E., Bellekens, X., Hamilton, A., Dubouilh, P.L., Iorkyase, E., Tachtatzis, C., et al.: Threat analysis of IoT networks using artificial neural network intrusion detection system. In: International Symposium on Networks, Computers and Communications, pp. 1–6. IEEE (2016)
7. Hosseini, S., Azizi, M.: The hybrid technique for DDoS detection with supervised learning algorithms. Comput. Netw. **158**, 35–45 (2019)

8. Wang, F., Sang, J., Liu, Q., Huang, C., Tan, J.: A deep learning based known plaintext attack method for chaotic cryptosystem (2021). https://doi.org/10.48550/ARXIV.2103.05242
9. Kreimel, P., Eigner, O., Tavolato, P.: Anomaly-based detection and classification of attacks in cyberphysical systems. In: Proceedings of the International Conference on Availability, Reliability and Security 2017. ACM (2017)
10. Wang, X., Ren, L., Yuan, R.,. Yang, L.T., Deen, M.J.: QTT-DLSTM: a cloud-edge-aided distributed LSTM for cyber-physical-social big data.: IEEE Trans. Neural Netw. Learn. Syst. https://doi.org/10.1109/TNNLS.2022.3140238
11. Thiruloga, S.V., Kukkala, V.K., Pasricha, S.: TENET: temporal CNN with attention for anomaly detection in automotive cyber-physical systems. In: 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022, pp. 326–331. https://doi.org/10.1109/ASP-DAC52403.2022.9712524
12. Alassery, F.: Predictive maintenance for cyber physical systems using neural network based on deep soft sensor and industrial internet of things. Comput. Electr. Eng. **101**, 108062 (2022). ISSN 0045-7906. https://doi.org/10.1016/j.compeleceng.2022.108062
13. Shin, J., Baek, Y., Eun, Y., Son, S.H.: Intelligent sensor attack detection and identification for automotive cyber-physical systems. In: IEEE Symposium Series on Computational Intelligence, pp. 1–8 (2017)
14. Teyou, D., Kamdem, G., Ziazet, J.: Convolutional neural network for intrusion detection system in cyber physical systems (2019). https://doi.org/10.48550/ARXIV.1905.03168
15. Hossain, M.D., Ochiai, H., Doudou, F., Kadobayashi, Y.: SSH and FTP brute-force attacks detection in computer networks: LSTM and machine learning approaches. In: 2020 5th International Conference on Computer and Communication Systems (ICCCS) (2020). https://doi.org/10.1109/ICCCS49078.2020.9118459

Information Assurance and Security



Deployment of Co-operative Farming Ecosystems Using Blockchain

Aishwarya Mahapatra, Pranav Gupta, Latika Swarnkar, Deeya Gupta, and Jayaprakash Kar(✉)

Centre for Cryptography, Cyber Security and Digital Forensics, Department of Computer Science & Engineering Department of Communication & Computer Engineering, The LNM Institute of Information Technology, Jaipur, India
jayaprakashkar@lnmiit.ac.in

Abstract. Blockchain has helped us in designing and developing decentralised distributed systems. This, in turn, has proved to be quite beneficial for various industries grappling with problems regarding a centralised system. So, we thought of exploring blockchain's feasibility in the agricultural industry. India is a country where a large part of the population is still dependent on agriculture, however, there's no proper system in use as yet that can help the farmers get the right price for their farm products and help the consumers get an affordable price for their needs.Thus, we propose a blockchain based decentralized marketplace where we will implement a collaborative model between farmers and consumers. This model will allow the farmers to record their potential crops and the expected output on decentralised ledger besides, enabling them to showcase their integrity and credibility to consumers. The consumers, on the other hand, can actually check everything about the farmers with the help of their information based on the previous supplies. This open and full proof digital market framework will thus reduce black marketing, hoarding, adulteration, etc. In this research paper, we have explored one possible blockchain model by creating a breach proof ledger of records. We have used solidity and ethereum frameworks for the working of the model.

Keywords: Blockchain · Ethereum · Agriculture industry · Decentralized

1 Introduction

This paper provides a model for the implementation of Blockchain in the Agriculture market.We have read about many farmers' suicide incidents due to heavy debts and bad yield from farming. The suicide rate among farmers is around 17.6%. For a nation like India, with a continuous rise in population, the dependency on the land is increasing rapidly. Thus, the fertile land for farming is being occupied for fulfilling the other requirements. However, because of the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1067–1081, 2023.
https://doi.org/10.1007/978-3-031-27409-1_98

high population, more yield is equally required to satisfy the needs of the country. Agriculture in India contributes about 16.5% to the GDP. But every year a farmer faces a huge debt which can go up to as much as 5 lakhs and inability to handle this debt leaves the farmer desolate. One of the reasons is definitely the middlemen. A farmer gets only 3$ for his products while the customers at the retailer market get it at around 26 times the original price. Besides, because of no proper infrastructure for warehouses and pest infestation, a huge amount of yield is wasted. And for this degraded quality of crops, farmers receive an even less price. The major issues are:

- Difficulty in collecting the initial investment money due to high interest rates of banks.
- Not being able to get a reasonable price for their produce due to middlemen's intervention in the market.
- Inability to analyse the modern market trends and customer needs. Currently, the farmers and customers are in no contact with each other because of the middlemen.
- Issues in storage and transportation that may lead to deterioration of crops.

Similarly, the customers are also suffering because of the high price they have to pay for commodities with undesirable quality of produce. They are forced to purchase whatever is available at whatever price set by the seller in the market. Moreover, various ill practices by the middlemen like black marketing, hoarding, adulteration, etc., further increases the prices for the farm products. All in all, the biggest challenge in the agro-markets today, is the farmers and consumers being separated by the middlemen. So, the solution to the above problems is by the use of a Decentralized agricultural market with micro-functionality that helps farmers pay back their debts and connects them to the consumers.

1.1 Blockchain

Blockchain is a distributed and decentralised ledger in which each transaction is recorded and maintained in sequential order to gain a permanent and tamper-proof record [12]. It's a peer-to-peer network that keeps track of time-stamped transactions between multiple computers. This network prevents any tampering with the records in the future, allowing users to transparently and independently verify the transactions. It is a chain of immutable blocks, and these blocks are a growing stack of records linked together by cryptographic processes [3,15]. Every block consists of the previous block's hash code, a sequence of confirmed transactions, and a time-stamp. A unique hash value identifies these blocks [13] (Fig. 1).

Every block has two components: the header and its block body. All of the block's confirmed and validated transactions are there in the block's body, whereas the block header mainly contains a Time-stamp, a Merkle Tree Root Hash, Previous Block's Hash Code, and a Nonce [4,14] (Figs. 2 and 3).

The time-stamp is used to keep track of when blocks are created and when they are updated.The hash code that identifies every block transaction is verified

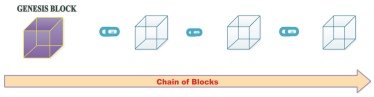


Fig. 1. Blockchain representation as a chain of blocks

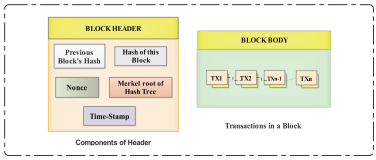


Fig. 2. Components of blocks in blockchain

using the Markle Tree Root Hash. It's a recursively defined binary tree of hash codes, aiming to provide secure and efficient verification of transactions.Previous Block Hash Code: It is usually a hash value of SHA-256 bit that references to the previous block. The chronology and connection between different blocks of a blockchain are established through this component. Genesis Block, which is the starting block, does not have this component.A nonce is a one-time number that is used for cryptographic communication. It is modified with each hash computation and generally starts with zeros [5, 10].

1.2 Consensus Algorithms

Before learning about two main consensus algorithms,let us understand first about miners and mining. Miners are the special nodes that can create new blocks in blockchain by solving a computational puzzle. These miners receive all the pending transactions, verifies all the transactions and solve the complex cryptographic puzzles. And, the one who solves the puzzle will create a new block, append all transactions and broadcast to all other peers. And the first Block creator will get rewarded. The rewards can be either bitcoin or transaction fees. Bitcoin is given as a reward in case of Bitcoin Cryptocurrency, and

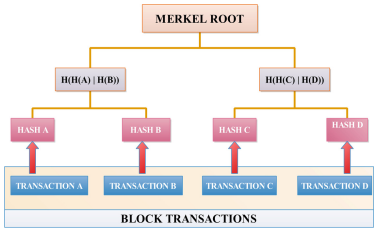


Fig. 3. Merkel tree for hash generation of a block

transaction fees is for ethereum. This entire process is called the Mining process. Mining is necessary as it helps to maintain the ledger of transactions [8,9].

1. **Proof of Work (PoW)** In the bitcoin network,The first consensus protocol to achieve consistency and security was PoW. Miners, here, compete to solve a complex mathematical problem and the solution so found is called the Proof-of-Work [15]. Miners keep adjusting the value of nonce (which is a one time number) to get the correct answer, which requires much computational power [18]. They use a complex machinery to speed up these mining operations. Bitcoin, Litecoin, ZCash and many others uses PoW as their Consensus protocol [7].
2. **Proof of Stake (PoS)** PoS is the most basic and environmentally-friendly alternative of PoW consensus protocol. to overcome the disadvantages like excessive power consumption by POW in bitcoin, PoS was proposed. Here, the miners are called validators. Instead of solving crypto puzzles, these validators deposit stake into the network in return for the right to validate. The more the stake, the more the possibility of getting chance to create new block [18]. The block validator is not predetermined and randomly selected to reach the consensus. The nodes which produce valid blocks get incentives but they also lose some amount of their stake, if the block is not included in the existing chain [7].

1.3 Related Work

With blockchain rapidly growing in agriculture industry some platforms are already developed are currently used for different agricultural activities [4]. This subsection gives a glimpse about such agriculture related platforms developed with the help of blockchain: FTSCON (Food Trading System with Consortium blockchain) It involves transaction mechanism which is automatic in nature for merchants and supply chain in agri-food. FTSCON upgrades privacy protection as well as it also improves transaction security with the help of smart contracts. It also uses consortium blockchain generally more efficient and friendly than a public blockchain when we are talking about computational power and financial cost [16].

Harvest Network: It is a blueprint which is a traceability applications. In this Ethereum blockchain along with various IoT devices were combined with GS1 standards. Al last the network developed in result i.e. Harvest Network gave them the idea to tokenize the smart contracts, with the help of which the underlying contract is generally not subject to any global consensus and as a result it need not to be validated by whole network present over there. This network is be processed only by node clusters which are of dynamic size so as a result efficiency is improves up to a great extent [16].

Provenance: It was founded by Jessi Baker in year 2013 and it is first developed platform which supports various supply chain activities. It also allows producers, consumers and retailers to keep an eye on their products during various stages

and during entire life cycle of their product. It authenticates and enables each and every single physical product with the help of "a digital passport" which not only confirms confirms its authenticity but also keeps track of the origin so as to prevent selling of fake goods [17]. With the help o Provenance's trust engine, producers and consumers are now easily able to substantiate ongoing supply transactions so as to get a much better integrity throughout supply chain network. Moreover in this they can turn certifications which are in digital formats easily to data marks so that customer can review and use it and can forward it to blockchain so that there it can be stored in a genuine and secured way. Provenance also allows various stakeholders share and tell their truthful stories related to their products and goods in a reliable mode. Producers as well as consumers can trace their items and products with the help of this tracking tool. Moreover using provenance we can issue a digital asset for physical products and as a result it can be connected with the help of protected tag such as NFC, which will reduce the time taken to trace to a great extent from days to a few seconds which results in reduction of frauds, improves transparency, provides recalls at faster rate and also protects brand values.

OriginTrail is also a much similar type of platform which was developed using blockchain so as to provide validation and data integrity in supply chain activities [16].

AppliFarm: This is a very wide and vast blockchain platform which was founded int 2017 by Neovia. Most commonly it is used while providing digital proof related to animal welfare, livestock gazing etc. [16]. In animal production sector so as to identify areas in which cow and other animals are gazing a tracker is linked by linking tags around the neck of cow and in this way sufficient amount of data can be gathered and thus we can make sure high-quality grazing, moreover it can also be used to track livestock data [4].

AgriDigital: This is a blockchain platform based on cloud which was founded in year 2015 by a group of farmers from Australia and some professionals based on agribusiness.AgriDigital as a result makes supply chain easy to use and is secure w.r.t. farmers and consumers. contracts, deliveries, orders and payments all can be easily managed by the farmers as well as all stakeholders all in real-time [17]. Basically this platform have five main subsystems. (1) Transactions: In this stakeholders and farmers are able to buy and at the same time sell various goods very easily by the help of this system. (2) Storage: In this sensitive information like the accounts, payments, orders, delivers are digitized and then stored [4], (3) Communications: In this farmers can build connection patterns for consumers. (4) Finance: Using this Farmers can have all virtual currency transfer and financial transactions with consumers. (5) Remit: Can be used to transfer real time remittances issues to various farmers. Main feature of this platform is that it can create digital assets in the form of tokens which represent the agricultural goods (e.g. tons of grains) which are in physical form [16]. An immutable data and physical asset is formed using proof of concept protocol because of the asset transfer from farmer to the consumer in digital form. And

this is formed along supply chain. And at last Once digital asset is created and issued then producers and consumers can at last use this application layer to send/receive data [16].

Blockchain-based agricultural systems: There are various blockchain-based systems which are used in agriculture. These are:
(1) **Walmart Traceability Pilot**-The main aim of this application was to easily trace the production and origin of mangoes and pork in Walmart. It was implemented using Hyperledger Fabric platform. It was the first known project in blockchain used to track shrimp exports from Indian farmers to overseas retailers.
(2) **Use case of egg distribution**-The main aim of this application was to Trace distribution of egg from farm to consumer. It was implemented using Hyperledger Sawtooth platform.
(3) **Brazilian Grain Exporter**-This applications Helps the producer in Brazil to track grains to trade with global exports. It was implemented on Hyperledger Fabric with platform.
(4) **Agrifood Use Case**-This application is used to Verify the certificates of table grape shipped from Africa and sold in Europe Platform used for implementation was Hyperledger Fabric and.
(5) **E-Commerce food chain**-This application is used to Design a tracking and certificate system for e-commerce food supply chain. Platform used is Hyperledger Fabric.
(6) **Food safety traceability**. This application Combines blockchain with EPCIS standard for reliable traceability system.Platform used is Hyperledger ethereum.
(7) **Product transaction traceability**-This application Implements product traceability system with evaluations of deployment costs and security analysis. Platform used is Hyperledger etherum.
(8) **OriginChain**-This application uses Blockchain to trace the origin of products Traceability Ethereum.
(9) **RFID traceability**-Use RFID tags to trace cold chain food in the entire supply chain.
(10) **AgriBlockIoT**-Traceability of all the IoT sensor data in an entire supply chain.
(11) **Water control system**-This application is used in Smart agriculture scenario for irrigation system of plants to reduce water waste.
Smart watering system: This application integrates a fuzzy logic decision system with blockchain storage for data privacy and reliability.
Fish farm monitoring: Secure all the monitoring and control data in a fish farm.
IoT-Information: Information sharing system for accumulated timeline of hoe acceleration data.
Business transactions on soybean: Track and complete business transactions in soybean supply chain.

1.4 Motivation and Contribution

Our sole motivation behind working on this is to help the farmers get the right price for their product. Besides, the costumers too are getting the product at prices way higher than they can get. The major reason behind this problem are the middlemen. Due to the middlemen and their unfair means, the prices skyrocket for the consumers whereas, the farmers suffer from a very poor deal of their crops. It's a loss for both the parties i.e., farmers as well as the consumers. However, our proposed transparent system completely removes the middlemen and gets both, the farmers and the consumers directly in contact to set the deal. This, in turn, proves to be a win-win for both. It maximises the profit of the farmers while the consumers can buy it at the most affordable prices. This work may improve the relationship between the farmer and the final consumer. Transactions will be secured by Blockchain technology. Right now there is very less work done in this field. So we are trying to contribute. By the success of the idea we ensure that the food we eat will have much less cost. And the quality will get improved. It will significantly improve the lives of the farmers as they will no longer stay in debt for the whole year. The farmers will get the complete reward for their hard work.

2 The Proposed System

A Blockchain-based network is proposed where the farmers and consumers work cooperatively to sell and buy the farm's yield or produce.In this way, a decentralized, transparent and tamper-proof cooperative environment is established without any intermediaries (Fig. 4).

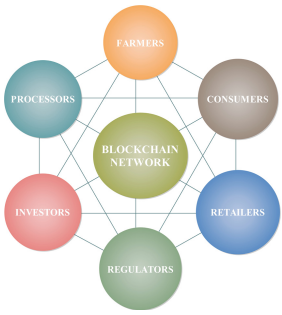


Fig. 4. Blockchain network

Above figure shows a network of blockchain where the main participating or controlling entities include Farmers, Investors, Retailers, Processors, Regulators and the end Customer. All these stakeholders have access to the transaction records of all the transactions. The Ethereum Virtual Machine executes the

smart contract on this blockchain network. Because the timestamp of each transaction is recorded, counterfeiting anywhere in the supply chain can be quickly detected. The product's total traceability to the customer is ensured in this manner [11]. Therefore, a consensus can be made between the farmers and consumers, allowing the consumers to fund fields or specific crops of their choice at no interest and receive farm yield and all the profit made by its market value. The farmer does not need to rely on any other lending system for loans or financing to fund his initial investment, eliminating the middlemen. The proposed flow of this solution:

1. The farmer must first give details of all prospective crops and the estimated yield on the decentralized public ledger.
2. Farmers can then sell their agricultural produce in the market to the processor.
3. The quality tester checks the crop quality. This quality report is saved on the blockchain network, which is added to the Blockchain at each step. This report is utilized by the processor to verify whether the raw material is of good quality or not.
4. After that, the processor can sell the product to a retailer. Now when the product reaches the customer, then the entire report from the farmer to the retailer can be made available
5. Customers can view all these details and assess farmers' credibility with the help of their farm's previous cultivation and delivery. In this manner, the consumers can ensure good quality products at a low cost by investing early in the crops.

So, the best farmer will make the most profit from the product's production, and the best investor or customer will be able to provide his family with high-quality food. Thus, both the farmers and consumers can build a reliable and cooperative environment where both of these can obtain profits.

3 Implementation Using Smart Contract

Smart contract is nothing but a self-implementing contract in which there are terms of agreement between the sellers and buyers and these are written into lines of code. These lines of code and the agreements are kept across a blockchain network which is distributed and decentralized [1]. Moreover, these codes control the implementation and the transactions are irreversible plus, traceable for that matter.

3.1 Solidity

Solidity is the smart contract programming language used on the Ethereum blockchain to create smart contracts. It is a high-level programming language just like C++ and python. It is a contact-oriented programming language, which means smart contracts are responsible for storing all of the logics that interacts

with the block-chain. The Solidity programming language is operated by the EVM (Ethereum Virtual Machine), which is hosted on nodes of Ethereum linked with the Blockchain. It's statically typed, with inheritance, libraries, and other features [2].

3.2 Truffle and Ganache

Truffle Suite is build on Ethereum Blockchain. It is basically a development environment and used to develop Distributed Applications(DApps). There are three parts of truffle suite: (1) Truffle: Development Environment, Used as a testing framework and also Assets pipeline in Ethereum Blokchains, (2) Ganache: Personal Ethereum Blockchain and is also used to test smart contracts and Drizzle: Collection of libraries. Ganache provides virtual accounts which have crypto-currency of pre-defined amount. And after each transaction, there is a deduction in crypto-currency from the main account on which transaction is performed. Each account has its own private key in Ganache and also has a unique address [6].

3.3 Code Analysis

Below is the Smart Contract written in Solidity Language.We have tried to create a ecosystem in which the customer and farmer will directly interact with each other without any middle-men in between.

```
pragma solidity ^0.5.0;

contract MyContract
{

    uint256 public no_of_farmer_entries = 0;
    uint256 public no_of_lots=0;
    mapping (address => uint) balances;

    function fundaddr(address _account) public
    {
        balances[_account] = 2000;
    }


    function sendMoney(address receiver, uint amount, address
     sender) public returns(bool sufficient)
    {
        if (balances[sender] < amount)
        return false;
        balances[sender] -= amount;
        balances[receiver] += amount;
        return true;

```

```
    }

    function getBalance(address addr) view public returns(uint)
    {
        return balances[addr];
    }

    struct farmer
    {

        uint256 fid;
        string fname;
        string location;
        string crop;
        uint256 contact_no;
        uint quantity;
        uint expected_price;
    }

    struct lot
    {
        uint256 lotno;
        uint mrp;
        string grade_of_crop;
        string testdate;
        string expected_date;
    }

    mapping (uint256 => farmer) farmer_map;
    farmer[] public farmer_array;

    mapping (uint256 => lot) lot_map;
    lot[] public lot_array;



    function Register(uint256  id, string memory name, string memory loc, string memory _crop, uint256 phone, uint _quantity, uint price) public{

            MyContract.farmer memory fnew = farmer(id,name,loc,_crop,phone,_quantity,price);
            farmer_map[id] = fnew;
            farmer_array.push(fnew);
            no_of_farmer_entries++;

    }

```

```
69      function get_farmer_detail(uint256  j) public view
        returns(uint256 ,string memory,string memory,string
        memory,uint256,uint,uint)
70      {
71              return (farmer_map[j].fid,farmer_map[j].fname,
        farmer_map[j].location,farmer_map[j].crop,farmer_map[j].
        contact_no,farmer_map[j].quantity,farmer_map[j].
        expected_price);
72      }
73
74      function quality(uint256  lot_no, uint256 _mrp,string
        memory _grade, string memory _testdate, string memory
        _expected_date) public{
75
76              MyContract.lot memory lnew=lot(lot_no,_mrp,_grade
        ,_testdate,_expected_date);
77              lot_map[lot_no]=lnew;
78              lot_array.push(lnew);
79              no_of_lots++;
80
81      }
82
83      function getquality(uint256 k) public view returns(
        uint256 ,uint,string memory,string memory,string memory)
84      {
85          return( lot_map[k].lotno, lot_map[k].mrp, lot_map[k].
        grade_of_crop, lot_map[k].testdate, lot_map[k].
        expected_date);
86
87      }
88  }
```

Here are the functionalities of our code:

- **balances** variable stores the money at a particular address. It is just like a bank account where at each index(here address) a sum of money is stored. We use **fundaddr()** function to store the amount at a particular address. It can be the account of both the farmers and customers.
- **sendMoney()** function is used to send money from sender to receiver. **getBalance()** will be used to keep track of updated balance at a particular address.
- We have two struct type variables, namely, **farmer** and **lot** which stores all the details of farmer and lot allotted to them.
- **Register()** function will register all the farmer details like id, name, location, crop he/she want to sell, contact no., quantity of produce and expected price. All these details are stored in farmer_array which is an array of type farmer. Then we use mapping farmer_map to get all the farmer details using the **get_farmer_detail()** function.
- After registering, the quality of crop will be checked and assigned a lot number which helps in locating the specific type of crop from different types he/she

produce, MRP, grade based on quality of crop, test_date and expected_date of the product.This is what is done by function **quality()**.
– Now the customer can enter the farmer_id and lot no. to get the details of desired produce (**get_farmer_detail()** and **get_quality()** function) and directly pay him the required money (**sendMoney()** function).

3.4 Results

For the deployment of this smart contract through Ganache and Truffle, we have used '2_deploy_contracts.js' and 'migrations.sol' files (Figs. 5 and 6).

```
2_deploy_contracts.js
=====================

   Replacing 'MyContract'
   ----------------------
   > transaction hash:    0x6a30e0a26c3fb9f865ec04cd8f08dbb4ac38a94451e795596889
e6e67d4213b1
- Blocks: 0            Seconds: 0
   > Blocks: 0            Seconds: 0
   > contract address:    0x6EAdf7C96Dd3DF9eF4621B796ecc8baba1F00495
   > block number:        8
   > block timestamp:     1640432442
   > account:             0xa10D5deD9553c8c390DFe06F828Ce8e264D06603
   > balance:             99.76848696
   > gas used:            1399699 (0x155b93)
   > gas price:           20 gwei
   > value sent:          0 ETH
   > total cost:          0.02799398 ETH

   > Saving artifacts
   -------------------------------------
   > Total cost:          0.02799398 ETH


Summary
=======
> Total deployments:   1
> Final cost:          0.02799398 ETH
```

Fig. 5. Deploying MyContract.sol

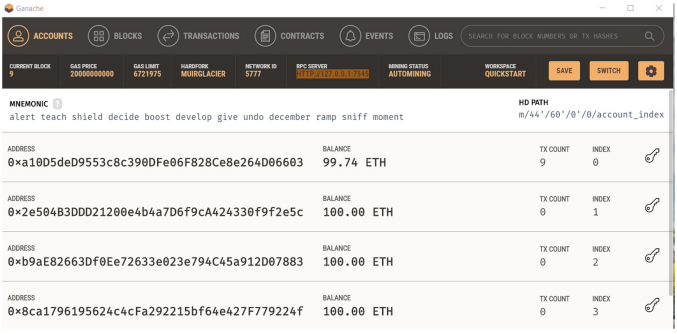


Fig. 6. Ganache account and remaining balance out of 100 ethers

As we can see from above that on deployment of the above Smart contract, the **Transaction Hash**, **Contract address** (address on which this smart contract is deployed), **Block Number**, **Block Time**, **account** (one of the account

from Ganache), remaining **balance** (after transaction), amount of **gas used**, **gas price** and **total cost** of this transaction are updated.

4 Advantages of Our Solution

- Bank loans and the other money lending mechanisms are too time taking. Our proposed solution will make it simple and straight forward because in our solution we let the consumers to fund for the crops of which they want the end product for their use. As soon as the deal is done from both farmer and consumer side, our money lending mechanism will directly transfer the fund to the farmer's bank. In this way farmers will not have to re-pay in the form of money, so they won't have a burden to pay the interest to the banks, or we can say this mechanism will lead to a zero interest funds to the farmer, they just have to do their work of growing crops.
- Now a days we have to pay higher because of middle men and they give very nominal price to farmer for their crop, our system will provide a good quality product to consumer in a lesser price and the farmer will get much higher profits for their work.
- There are many farmers having small land farms and some are household farmers, since our system is based on crop on demand so this will help those farmers to grow for profit and provide good quality product and this will lead to a good profit to them.
- Our system is kind of supply chain, which can enable point to point update over immutable chains. Customer will have a transparent system with which they can choose a particular farmer for a particular product.
- By our system we let consumer and farmer to interact with each other so that consumer can rate a farmer on his service in this way he can get a reputation in urban areas and indirectly this will increase a farmer's profits, because of middle men we don't know which farmer is growing the crop for us, today's scenario is really bad for both consumer and farmer because of middle men who hide everything from both farmer and consumer so our system will cutout middle men and build a transparency between farmer and consumer.
- In case of discrepancies like natural calamities, climate change or any other situation of crop loss, now only the farmer will not have to suffer, blockchain's smart contracts can handle these type of problems and settle these situations.

5 Conclusion

A distributed system of food supply chain based on blockchain helps both the farmers and the buyers to create a cooperative atmosphere. This help farmers analyse the market and customer needs.In our proposed model,the farmer first lists the expected yield of the potential crops on the decentralized public ledger.

Then the customers checks for the respective details of the his/her desired crop and also checks the credibility of farmer based on the quality he/she is

assigned while quality testing. This way, the consumer is ensured of a tamper-proof and transparent digital market system.

Thus, a kind of consensus or agreement can be formed between the buyers and the farmer, such that the buyer can fund the crops he/she wants to buy on a prior basis and then, acquire the crops once ready.

This helps farmer to have customers before actual crop is ready for market and avoid wastage of food in warehouses. This will ultimately help us resolve the grave agrarian crisis India is facing. Developing countries would see less suicides in the sector.

In a nutshell,blockchain technology can help us to curb a crisis India is heading to.

References

1. Introduction to smart contracts.: (2016–2021). https://docs.soliditylang.org/en/v0.8.11/introduction-to-smart-contracts.html
2. Solidity.: (2016–2021). https://docs.soliditylang.org/en/v0.8.11/
3. Albarqi, A., Alzaid, E., Al Ghamdi, F., Asiri, S., Kar, J., et al.: Public key infrastructure: a survey. J. Inf. Secur. **6**(01), 31 (2014)
4. Bach, L.M., Mihaljevic, B., Zagar, M.: Comparative analysis of blockchain consensus algorithms. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1545–1550. IEEE (2018)
5. Bermeo-Almeida, O., Cardenas-Rodriguez, M., Samaniego-Cobo, T., Ferruzola-Gómez, E., Cabezas-Cabezas, R., Bazán-Vera, W.: Blockchain in agriculture: a systematic literature review. In: International Conference on Technologies and Innovation, pp. 44–56. Springer (2018)
6. Ganache-Cli.: https://trufflframework.com/docs/ganache/overview
7. Hazari, S.S., Mahmoud, Q.H.: Comparative evaluation of consensus mechanisms in cryptocurrencies. Internet Technol. Lett. **2**(3), e100 (2019)
8. Kar, J., Mishra, M.R.: Mitigating threats and security metrics in cloud computing. J. Inf. Process. Syst. **12**(2), 226–233 (2016)
9. Kaur, S., Chaturvedi, S., Sharma, A., Kar, J.: A research survey on applications of consensus protocols in blockchain. Secur. Commun. Netw. **2021** (2021)
10. Kumari, N., Kar, J., Naik, K.: Pua-ke: practical user authentication with key establishment and its application in implantable medical devices. J. Syst. Arch. **120**, 102307 (2021)
11. Leduc, G., Kubler, S., Georges, J.P.: Innovative blockchain-based farming marketplace and smart contract performance evaluation. J. Clean. Prod. **306**, 127055 (2021)
12. Moubarak, J., Filiol, E., Chamoun, M.: On blockchain security and relevant attacks. In: 2018 IEEE middle East and North Africa communications conference (MENACOMM), pp. 1–6. IEEE (2018)
13. Nofer, M., Gomber, P., Hinz, O., Schiereck, D.: Blockchain. Business & information. Syst. Eng. **59**(3), 183–187 (2017)
14. Puthal, D., Malik, N., Mohanty, S.P., Kougianos, E., Das, G.: Everything you wanted to know about the blockchain: Its promise, components, processes, and problems. IEEE Consum. Electron. Mag. **7**(4), 6–14 (2018)

15. Thirumurugan, G.: Blockchain technology in healthcare: applications of blockchain. Gunasekaran Thirumurugan (2020)
16. Torky, M., Hassanein, A.E.: Integrating blockchain and the internet of things in precision agriculture: analysis, opportunities, and challenges. Comput. Electron. Agric. 105476 (2020)
17. Xu, J., Guo, S., Xie, D., Yan, Y.: Blockchain: a new safeguard for agri-foods. Artif. Intell. Agric. **4**, 153–161 (2020)
18. Zhang, S., Lee, J.H.: Analysis of the main consensus protocols of blockchain. ICT Express **6**(2), 93–97 (2020)



Bayesian Consideration for Influencing a Consumer's Intention to Purchase a COVID-19 Test Stick

Nguyen Thi Ngan and Bui Huy Khoi(✉)

Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
buihuykhoi@iuh.edu.vn

Abstract. This study identifies the variables influencing customers' propensity to purchase a COVID-19 test stick. The 250 survey-responding consumers are working in HCMC, Vietnam. According to the findings of this study, five variables influence customers' intentions to purchase COVID-19 test sticks: Perceived usefulness (PU), Price Expectations (PE), Satisfaction (SAT), Global Pandemic Impact (GPI), and Perceived Risk (PR). The findings also show that the intention to purchase and use test sticks is positively and significantly influenced by knowledge of the COVID-19 outbreak, subjective indicators, and perceived benefits. The paper uses the optimum selection by Bayesian consideration for influencing a consumer's intention to purchase a COVID-19 test stick.

Keywords: BIC Algorithm · COVID-19 test stick · Perceived usefulness of the product · Price expectations · Satisfaction · Global pandemic impact · And perceived risk

1 Introduction

The COVID-19 pandemic has exaggerated people's lives and health. This pandemic is more dangerous than the diseases we have ever experienced, such as the H1N1 flu, or the severe acute respiratory syndrome (SARS) outbreak. In the situation of complicated epidemics, COVID-19 test strips are one measure to help detect pathogens as early as possible and play an effective role in the disease's prevention. Up to now, COVID-19 test strips are widely sold around the world. Besides, many factors make consumers wonder and decide to buy a product. Therefore, understanding the wants and intentions of consumers to buy products is the key point in this research. This study determines the factors affecting consumers' intention to buy COVID-19 test strips in Ho Chi Minh City. From there, we can provide useful information for COVID-19 test strip businesses. Can listen to and understand the thoughts of consumers, to improve - improve product quality. And contribute to proposing some solutions to continue to hold on to the market to serve consumers in the best way, and to satisfy customers' needs more during the epidemic period, along with the difficult economic situation nowadays. The article uses the optimum selection by Bayesian consideration for influencing a consumer's intention to purchase a COVID-19 test stick.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1082–1092, 2023.
https://doi.org/10.1007/978-3-031-27409-1_99

2 Literature Review

2.1 Perceived usefulness (PU)

According to Larcker and Lessi [1], usefulness is an important construct in examining existing measures of perceived usefulness, showing that the tools developed are neither validated nor validated by their trust. In this paper, a new tool for the two-way measurement of perceived usefulness has been developed. The results of an empirical study tested the reliability and validity of this instrument. According to Ramli and Rahmawati [2], perceived usefulness has a positive and significant impact on purchase intention, perceived usefulness has the most important effect on the intention to purchase and expenditure compared to perceived ease of using its ease. According to Li et al. [3], the global COVID-19 epidemic is very dangerous. This real-time PCR test kit has many limitations. Because PCR testing requires a lot of money, requires a professional, and needs a test site. Therefore, it is necessary to have a precise and fast testing method to promptly recognize many infected patients and transporters of Covid-19. We have developed a quick and simple test technique that can detect patients at various stages of infection. With this test kit, we performed clinical studies to confirm its clinically effective uses. The overall sensitivity of the rapid test is 88.66% and the specificity is 90.63% [3]. A quick and accurate self-test tool for diagnosing COVID-19 has become a prerequisite to knowing the exact number of cases worldwide and in Vietnam in particular. And take health actions that the government deems appropriate [4]. An analysis of Vietnam's COVID-19 pandemic policy responses from the beginning of the outbreak in January 2020 to July 24, 2020 (with 413 confirmed cases and 99 days with no recent cases of infection from the resident community). The results show that Vietnam's policy has responded promptly, proactively and effectively in meeting product sources in the essential period [5]. The hypothesis is built as follows..

H1: Perceived usefulness (PU) has an impact on the intention to purchase (IP) a COVID-19 test stick

2.2 Price Expectations (PE)

According to Février and Wilner [6], consumers feel completely justified and keep their price expectations. It is testable provided market-level data on prices and purchases is available. Realize that consumers have simple expectations about price. The predictive effect, due to strategically delaying a purchase, accounts for one-fifth of normal-time purchase decisions. These results have implications for demand estimation, optimal price, and welfare calculation of that product. The common psychology of people is most afraid of spending money and buying fake test strips. Because the SARS-CoV-2 rapid test kit and COVID-19 treatment drugs are conditional business items, must be licensed by health authorities, ensure quality, and have clear origins, the business of these items is currently in turmoil. Facing the erratic situation in the price of COVID-19 test strips, the Ministry of Health has now sent a document to businesses selling COVID-19 test strips to ensure supply to cope with the situation of the current epidemic and make the price as listed. The market management inspection agency will also regularly inspect and strictly handle the places that sell test strips that take advantage of the scarcity of

goods to take advantage of consumers [7]. According to Essoussi and Zahaf [8], it is recommended that when the product is of good quality and has certificates that guarantee the origin and safety, it will increase the interest and purchase intention of consumers. The consumer perceives that the product has value and benefits and feels it is appropriate for the income level. That's why they will pay to buy a product. Therefore, the following hypothesis is proposed.

H2: Price Expectations (PE) have an impact on the intention to purchase (IP) a COVID-19 test stick.

2.3 Satisfaction (SAT)

According to Veenhoven [9], when talking about satisfaction, there are six questions to be considered: (1) What is the point of studying satisfaction? (2) What is satisfaction? (3) Can satisfaction be measured? (4) How to be satisfied? (5) What causes us to be satisfied or unsatisfied? (6) Is it possible to increase the level of satisfaction? These questions are considered at the individual level and the societal level. Consumer satisfaction is not only an important performance outcome but also a major predictor of customer loyalty, as well as a retailer's persistence and success. There are many types of COVID-19 test strips on the market, but most are very easy to use, with no qualified person. It is possible to test quickly. Information, as well as instructions on how to use the COVID-19 test strips, are widely disseminated on e-commerce sites such as the internet, television, radio, etc., or there are instructions for use directly on the packaging. COVID-19 test strips products can be found and purchased widely at drugstore systems, and reputable online trading establishments on e-commerce platforms [10]. According to Essoussi and Zahaf [8] t is recommended that when the product is of good quality and has certificates that guarantee the origin and safety, it will increase the interest and purchase intention of consumers. Consumers perceive the product is worth more than expected for the price, so they will pay for it. In this study, the perception of ease of use of COVID-19 test strips is the perception of consumers that it is completely easy-to-use test strips to detect the disease. You don't need to have too much medical knowledge or expertise to use it. From here, this hypothesis is proposed as follows.

H3: Satisfaction (SAT) has a positive effect on the intention to purchase (IP) a COVID-19 test stick

2.4 Global Pandemic Impact (GPI)

The COVID-19 pandemic has become one of the most serious health crises in human history, spreading extremely rapidly globally from January 2020 to the present. With quick and drastic measures, Vietnam is one of the few countries that has controlled the outbreak [5]. It has recently been documented in the literature that humidity, temperature, and air pollution may all contribute to the COVID-19 epidemic's respiratory and contact transmission. The number of instances was unaffected by temperature, air humidity, the number of sunny days, or air pollution. Additionally, the effect of wind speed (9%) on the number of COVID-19 cases is moderated by population density. The discovery that the invisible COVID-19 virus spreads more during windy conditions shows that airborne viruses are one threat to people, with wind speeds enhancing air circulation [11].

H4: Global Pandemic Impact (GPI) has a positive effect on the intention to purchase (IP) a COVID-19 test stick

2.5 Perceived Risk (PR)

According to Peters et al. [12], risk characteristics such as fear and product likelihood, negative reaction, and vulnerability to medical errors to emerge, etc. have fueled anxiety about purchase intention. Worrying about medical errors is a factor in consumer intention to buy, as well as a perception of risk. An understanding of the anxiety effect on the product. It is established that psychological variables have a significant impact on how people respond to the risk of infection and the harm that infection can inflict, as well as how they comply with public health interventions like immunizations. Any infectious disease, including COVID-19, should be managed to take these factors into account. The present COVID-19 pandemic clearly shows each of these characteristics. 54% of respondents in a study of 1210 people from 194 Chinese cities in January and February 2020 classified the psychological effects of the COVID-19 outbreak as moderate or severe; 29% of respondents experienced moderate-to-severe anxiety symptoms, and 17% reported moderate-to-severe depression symptoms. Although answer bias is possible, this is a very high incidence and some people are likely at higher risk [13]. So for each of us, we need to protect our health, as well as the community. By following the 5K rule, and getting tested together when you have symptoms or have been in contact with a patient or suspected person before. Use the rapid test method with COVID-19 test strips to detect the disease as soon as possible and isolate and treat it. Post-COVID-19 patients leave a lot of sequelae for the body, so please detect and not get sick in time. Responses to a pandemic like COVID-19 are concerned with up-to-date health information, pandemic information, and information on methods to help detect diseases as quickly as rapid test strips. COVID-19. Most people are afraid of buying poor quality test strips, not at the right price, pirated products, or even afraid of not having products during a stressful epidemic [7]. This hypothesis is proposed to build as follows.

H5: Perceived Risk (PR) affects the intention to purchase (IP) a COVID-19 test stick

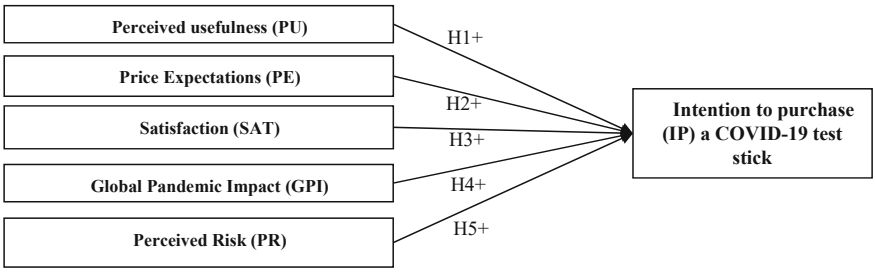


Fig. 1. Research model

All hypotheses and factors are shown in Fig. 1.

3 Methodology

3.1 Sample Size

Tabachnick and Fidell [14] claim that N 8m + 50 should be the minimum sample size for the optimal regression analysis. Where m is the number of independent variables and N is the sample size. According to the formula N > = 8m + 50, there are 8 * 6 + 50 samples total in the survey. The authors investigated consumers living in Ho Chi Minh City, Vietnam in 2022. Research information is collected by the author by submitting google forms and distributing survey forms directly to consumers. Respondents were selected by convenience method with an official sample size of 240 people. Table 1 shows the sample characteristics and statistics.

Table 1. Statistics of Sample

Characteristics		**Amount**	**Percent (%)**
Sex and Age	Male	105	43.8
	Female	135	56.3
	< 18	13	5.4
	18 – 25	80	33.3
	26 – 35	80	33.3
	> 35	67	27.9
Income	< 5 VND mils	64	26.7
	5 - 10 VND mills	117	48.8
	> 10 VND mills	59	24.6
Job	Student	52	21.7
	Working People	163	67.9
	Retirement	6	2.5
	Other	19	7.9

3.2 Bayesian Information Criteria

In Bayesian statistics, prior knowledge serves as the theoretical underpinning, and the conclusions drawn from it are mixed with the data that have been seen [15–17, 19]. According to the Bayesian approach, probability is information about uncertainty; probability measures the information's level of uncertainty [20]. The Bayesian approach is becoming more and more well-liked, especially in the social sciences. With the rapid advancement of data science, big data, and computer computation, Bayesian statistics became a well-liked technique [21]. The BIC is an important and useful metric for choosing a full and straightforward model. A lower BIC model is chosen based on the BIC

information standard. When the minimum BIC value is reached, the best model will end [22].

First, the posterior probability $P(\beta_j \neq 0|D)$ given by variable X_j with *(j = 1, 2,..., p)* indicates the possibility that the independent variable affects the occurrence of the event (or a non-zero effect).

$$P(\beta_j \neq 0|D) = \sum_{M_k \in A} P(M_k|D) * I_k(\beta_j \neq 0) \tag{1}$$

where A is a set of models selected in Occam's Window described in Eqs. 1 and $I_k(\beta_j \neq 0)$ is 1 when β_j in the model M_k and 0 if otherwise. The term $P(M_k|D) * I_k(\beta_j \neq 0)$ in the above equation means the posterior probability of the model M_k not included $X_j = 0$. The rules for explaining this posterior probability are as follows [18]: Less than 50%: evidence against impact; Between 50% and 75%: weak evidence for impact; Between 75% and 95%: positive evidence; Between 95% and 99%: strong evidence; From 99%: very strong evidence;

Second, the formula provides an estimate of the Bayes score and standard error.

$$E(\beta_j|D) = \sum_{M_k \in A} \widehat{\beta}_j P(M_k|D) \tag{2}$$

$$SE(\beta_j|D) = \sqrt{\sum_{M_k \in A} \left\{\left[var(\beta_j|D, M_k) + \widehat{\beta}_j^2\right] P(M_k|D)\right\} - E(\beta_j|D)^2} \tag{3}$$

with $\widehat{\beta}_j$ is the posterior mean of βj in the Mk model. Inference about βj is inferred from Eqs. (1), (2) and (3).

4 Results

4.1 Reliability Test

The Cronbach's Alpha test is a method that the author can use to determine the reliability and quality of the observed variables for the important factor. This test determines whether there is a close relationship between the requirements for compatibility and concordance among dependent variables in the same major factor. The reliability of the factor increases with Cronbach's Alpha coefficient. The values listed below are included in Cronbach's Alpha value coefficient: Very good scale: 0.8 to 1, good use scale: 0.7 to 0.8, qualified scale: 0.6 and above. A measure is considered meeting the requirements if the Corrected item-total correlation (CITC) is greater than 0.3 [23].

Table 2 displays the Cronbach's Alpha coefficient of Perceived usefulness (PU), Price Expectations (PE), Satisfaction (SAT), Global Pandemic Impact (GPI), Perceived Risk (PR) for Intention to purchase (IP) a COVID-19 test stick is all greater than 0.7. Table 2 shows that Some CITCs is greater than 0.3. CITC of PR5 equal to 0.181 shows that this factor is not reliable, so it is rejected. That shows that the items are correlated in the factor and they contribute to the correct assessment of the concept and properties of

Table 2. Reliability

Factor	A	Item	Code	CITC
Perceived usefulness (PU)	0.828	Test sticks are easy to use without expertise	PU1	0.616
		Wide network of product supply locations and convenience for buying and selling	PU2	0.753
		Test stick meet the needs of customers	PU3	0.630
		Test stick products give quick results	PU4	0.620
Price Expectations (PE)	0.775	The cost of test sticks is always public and sold at the listed price	PE1	0.586
		Some many types and prices make it easy to choose	PE2	0.617
		The price of a test stick is suitable for the average income of Vietnamese people	PE3	0.629
Satisfaction (SAT)	0.789	Good product quality and price	SAT1	0.653
		Shops provide test sticks exactly as advertised	SAT2	0.594
		The product provides complete information and instructions for use	SAT3	0.644
Global Pandemic Impact (GPI)	0.769	The danger of a global pandemic that spreads quickly	GPI1	0.628
		The virus can spread easily through the respiratory tract	GPI2	0.576
		Rapid testing is required after exposure to F0 or symptoms of infection	GPI3	0.605
Perceived Risk (PR)	0.769	Worried about buying products of unknown origin, poor quality	PR1	0.644
		Afraid the product is difficult to use	PR2	0.699
		Fear of lack of supply at the peak of the pandemic	PR3	0.620
		The product price varies from the listed price	PR4	0.608

(*continued*)

Table 2. (*continued*)

Factor	A	Item	Code	CITC
		Worried about test sticks giving quick and inaccurate results	PR5	0.181
Intention to purchase (IP) a COVID-19 test stick	0.806	Continue to buy Covid-19 test sticks during the coming pandemic period	IP1	0.639
		Trust in the product of the Covid-19 test stick	IP2	0.674
		I will recommend to others to buy a Covid-19 test stick	IP3	0.646
$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma^2(x_i)}{\sigma_x^2}\right]$				

each factor. Therefore, in testing the reliability of Cronbach's Alpha for each scale, the author found that all the observed variables satisfy the set conditions that the Cronbach's Alpha coefficient is greater than 0.6 and the Corrected item coefficient - Total Correlation is greater than 0.3, so all items are used for the next test step.

4.2 BIC Algorithm

To find association rules in trans-action databases, many algorithms have been developed and inspected. More mining capabilities were offered by the presentation of additional mining algorithms, including incremental updating, generalized and multilevel rule mining, quantitative rule mining, multidimensional rule mining, constraint-based rule mining, mining with multiple minimum supports, mining associations among correlated or infrequent items, and mining of temporal associations [24]. Two data science subfields that are attracting a lot of attention are big data analytics and deep learning. Big Data has grown in importance as an increasing number of people and organizations have been gathering massive amounts of Deep Learning algorithms intending to purchase (IP) a COVID-19 test stick [25]. R program used BIC (Bayesian Information Criteria) to determine which model was the best. BIC has been employed in the theoretical environment to select models. To estimate one or more dependent variables from one or more independent variables, BIC can be employed as a regression model [26]. For determining a complete and simple model, the BIC is a significant and helpful metric [27–29]. Based on the BIC information standard, a model with a lower BIC is selected [18, 22, 26, 30]. R report displays each stage of the search for the ideal model. Table 3 lists BIC's choice of the top 2 models.

There are five independent and one dependent variable in the models in Table 3. Perceived usefulness (PU), Price Expectations (PE), Satisfaction (SAT), and Perceived Risk (PR) have a probability of 100%. Global Pandemic Impact (GPI) has a probability of 80.9%.

Table 3. BIC model selection

IP	Probability (%)	SD	model 1	model 2
Intercept	100.0	0.42294	1.5989	1.9920
PU	100.0	0.05054	0.2799	0.3022
PE	100.0	0.04780	0.2205	0.2402
SAT	100.0	0.04882	0.2117	0.2348
GPI	80.9	0.06691	0.1333	
PR	100.0	0.05052	-0.2842	-0.3081

Table 4. Model Test

Model	nVar	R^2	BIC	post prob
model 1	5	0.615	-201.4242	0.809
model 2	4	0.601	-198.5399	0.191
BIC = -2 * LL + log (N) * k				

4.3 Model Evaluation

Table 4's findings show that Model 1 is the best option, and BIC (−201.4242) demonstrates this is the minimum. Perceived usefulness (PU), Price Expectations (PE), Satisfaction (SAT), Global Pandemic Impact (GPI), Perceived Risk (PR) impact Intention to purchase (IP) a COVID-19 test stick is 61.5% (R^2=0.615) in table 4. BIC finds model 1 is the optimal choice and two variables have a probability of 80.9% (post prob=0.809). The above analysis shows the regression equation below is statistically significant.

$$\mathbf{IP = 1.5989 + 0.2799\,PU + 0.2205\,PE + 0.2117\,SAT + 0.1333\,GPI - 0.2842\,PR}$$

Code: Intention to purchase (IP) a COVID-19 test stick, Perceived usefulness (PU), Price Expectations (PE), Satisfaction (SAT), Global Pandemic Impact (GPI), Perceived Risk (PR).

5 Conclusions

The BIC Algorithm for the Intention is the best option for this investigation to purchase (IP) a COVID-19 test stick. Results of BIC Algorithm analysis on 5 factors of consumers' intention to buy COVID-19 test sticks have the following results: Perceived Risk (−0.2842), Perceived usefulness (0.2799), Price Expectations (0.2205), Satisfaction (0.2117), and Global Pandemic Impact (0.1333), in which Perceived Risk (−0.2842) has the strongest impact. This can be appropriate, because the consumers in this survey are mostly young people, working, and the prime needs to learn deeply and thoroughly about the benefits of the product. They care whether the product is good, has any benefits,

is convenient or not, and is necessary. And they are also afraid of the risk of buying fake COVID-19 test sticks, imitations, poor-quality products, etc.

Implications

Antigen test kits have been widely used as a screening tool during the worldwide coronavirus (SARS-CoV-2) pandemic. The 2019 coronavirus (COVID-19) pandemic has highlighted the requirement for different diagnostics, comparative validation of new tests, faster approval by federal agencies, and rapid production of test kits to meet global needs. Rapid antigen testing can diagnose SARS-CoV-2 contamination and is commonly used by people after the onset of symptoms. Rapid test kits for diagnostic testing are one of the important tools in the ongoing epidemiological process. Early diagnosis is still as important as the early stages of the COVID-19 pandemic. Because PCR testing is sometimes not workable in developing countries or rural areas, health professionals can use rapid antigen testing with the COVID-19 rapid test kit for diagnosis. The COVID-19 pandemic has exaggerated people's lives and health. This pandemic is more dangerous than the diseases we have ever experienced, such as the H1N1 flu, or the severe acute respiratory syndrome (SARS) outbreak. In the situation of complicated illness developments, COVID-19 test strips are one measure to help detect pathogens as early as possible and play an effective role in the disease's prevention. Up to now, COVID-19 test strips are widely sold around the world. Besides, many factors make consumers wonder and decide to buy a product. Therefore, understanding the wants and intentions of consumers to buy products is the key point in this research. From there, they can provide useful information for COVID-19 test strip business units. Can listen to and understand the thoughts of consumers, to improve - improve product quality. And contribute to proposing some solutions to continue to hold on to the market to serve consumers in the best way, and to satisfy customers' needs more during the epidemic period.

References

1. Larcker, D.F., Lessig, V.P.: Perceived usefulness of information: a psychometric examination. Decis. Sci. **11**(1), 121–134 (1980)
2. Ramli, Y., Rahmawati, M.: The effect of perceived ease of use and perceived usefulness that influence customer's intention to use mobile banking application. IOSR J. Bus. Manag. **22**(6), 33–42 (2020)
3. Li, Z., et al.: Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. J. Med. Virol. **92**(9), 1518–1524 (2020)
4. Merkoçi, A., Li, C.-Z., Lechuga, L.M., Ozcan, A.: COVID-19 biosensing technologies. Biosens. Bioelectron. **178**, 113046 (2021)
5. Le, T.-A.T., Vodden, K., Wu, J., Atiwesh, G.: 'Policy responses to the COVID-19 pandemic in Vietnam'. Int. J. Environ. Res. Public Health **18**(2), 559 (2021)
6. Février, P., Wilner, L.: Do consumers correctly expect price reductions? Testing dynamic behavior. Int. J. Ind. Organ. **44**, 25–40 (2016)
7. http://soytetuyenquang.gov.vn/tin-tuc-su-kien/tin-tuc-ve-y-te/tin-y-te-trong-nuoc/danh-sach-cac-loai-test-nhanh-duoc-bo-y-te-cap-phep.html
8. Essoussi, L.H., Zahaf, M.: 'Decision making process of community organic food consumers: an exploratory study'. J. Consum. Mark. (2008)

9. Veenhoven, R.: 'The study of life-satisfaction', Erasmus University Rotterdam (1996)
10. https://hcdc.vn/category/van-de-suc-khoe/covid19/tin-tuc-moi-nhat/cap-nhat-thong-tin-test-nhanh-d4a19c00e2d7eb23e10141e1a1569d3d.html
11. Coşkun, H., Yıldırım, N., Gündüz, S.: The spread of COVID-19 virus through population density and wind in Turkey cities. Sci. Total Environ. **751**, 141663 (2021)
12. Peters, E., Slovic, P., Hibbard, J.H., Tusler, M.: 'Why worry? Worry, risk perceptions, and willingness to act to reduce medical errors'. Health Psychol. **25**(2), 144 (2006)
13. Cullen, W., Gulati, G., Kelly, B.D.: 'Mental health in the COVID-19 pandemic'. QJM: Int. J. Med. **113**(5), 311–312 (2020)
14. Tabachnick, B., Fidell, L.:Using Multivariate Statistics, 4th edn., pp.. 139–179. HarperCollins, New York (2001)
15. Bayes, T.: LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philos. Trans. R. Soc. Lond. **1763**(53), 370–418
16. Thang, L.D.: The Bayesian statistical application research analyzes the willingness to join in area yield index coffee insurance of farmers in Dak Lak province, University of Economics Ho Chi Minh City (2021)
17. Gelman, A., Shalizi, C.R.: Philosophy and the practice of Bayesian statistics. Br. J. Math. Stat. Psychol. **66**(1), 8–38 (2013)
18. Raftery, A.E.: Bayesian model selection in social research. Sociological Methodology, pp. 111–163 (1995)
19. Thach, N.N.: How to explain when the ES is lower than one? A Bayesian nonlinear mixed-effects approach. J. Risk Financ. Manag. **13**,(2), 21 (2020)
20. Kubsch, M., Stamer, I., Steiner, M., Neumann, K., Parchmann, I.: Beyond p-values: using Bayesian data analysis in science education research. Pract. Assess Res. Eval. **26**(1), 4 (2021)
21. Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D.: Beyond Traditional Probabilistic Methods in Economics. Springer (2018)
22. Kaplan, D.: On the quantification of model uncertainty: a Bayesian perspective. Psychometrika **86**(1), 215–238 (2021). https://doi.org/10.1007/s11336-021-09754-5
23. Nunnally, J.C.: Psychometric theory 3E (Tata McGraw-hill education, 1994. 1994)
24. Gharib, T.F., Nassar, H., Taha, M., Abraham, A.: An efficient algorithm for incremental mining of temporal association rules. Data Knowl. Eng. **69**(8), 800–815 (2010)
25. Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E.: Deep learning applications and challenges in big data analytics. J. Big Data **2**(1), 1–21 (2015). https://doi.org/10.1186/s40537-014-0007-7
26. Raftery, A.E., Madigan, D., Hoeting, J.A.: Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. **92**(437), 179–191 (1997)
27. Ngan, N.T., Khoi, B.H., Van Tuan, N.: BIC algorithm for word of mouth in fast food: case study of Ho Chi Minh City, Vietnam. In: Book BIC Algorithm for Word of Mouth in Fast Food: Case Study of Ho Chi Minh City, Vietnam, pp. 311–321. Springer (2022)
28. Thi Ngan, N., Huy Khoi, B.: BIC Algorithm for Exercise Behavior at Customers' Fitness Center in Ho Chi Minh City, Vietnam': 'Applications of Artificial Intelligence and Machine Learning'. Springer, pp. 181–191 (2022)
29. Lam, N.V., Khoi, B.H.: Bayesian model average for student learning location. J. ICT Stand. 305–318–305–318 (2022)
30. Ngan, N.T., Khoi, B.H.: Using behavior of social network: Bayesian consideration. In: Book Using Behavior of Social Network: Bayesian Consideration, pp. 1–5. IEEE (2022)



Analysis and Risk Consideration of Worldwide Cyber Incidents Related to Cryptoassets

Kazumasa Omote(✉), Yuto Tsuzuki, Keisho Ito, Ryohei Kishibuchi, Cao Yan, and Shohei Yada

University of Tsukuba,Tennoudai 1-1-1, Tsukuba, Ibaraki 305-8573, Japan
omote@risk.tsukuba.ac.jp

Abstract. Cryptoassets are exposed to a variety of cyber attacks, including exploit vulnerabilities in blockchain technology and transaction systems, in addition to traditional cyber attacks. To mitigate incidents related to cryptoassets, it is important to identify the risk of incidents involving cryptoassets based on actual cases that have occurred. In this study, we investigate and summarize past incidents involving cryptoassets one by one using news articles and other sources. Then, each incident is classified by the "target of damage" and the "cause of damage", and the changing incident risk was discussed by analyzing the trends and characteristics of the time series of incidents. Our results show that the number of incidents and the amount of economic damage involving cryptoassets are on the increase. In terms of the classification by the target of damage, the damage related to cryptoasset exchanges is very large among all incidents. In terms of the classification by cause of damage, it was revealed that many decentralized exchanges were affected in 2020.

1 Introduction

Cryptoassets, an electronic asset using cryptography-based blockchain, have attracted attention since the Bitcoin price spike in 2017. According to Coinmarketcap [1], there are 9,917 cryptoasset stocks as of June 18, 2022, and their size is growing every year, as shown in Fig. 1. In recent years, expectations for cryptoassets have been increasing due to the decrease in out-of-home opportunities caused by the spread of COVID-19, and the increasing use of online financial instruments.

Cryptoassets have unique advantages such as decentralized management and difficulty in falsifying records due to the use of blockchain technology. In contrast, cryptoassets are vulnerable to a variety of cyberattacks, including exploit vulnerabilities in blockchain technology and transaction systems, in addition to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1093–1101, 2023.
https://doi.org/10.1007/978-3-031-27409-1_100

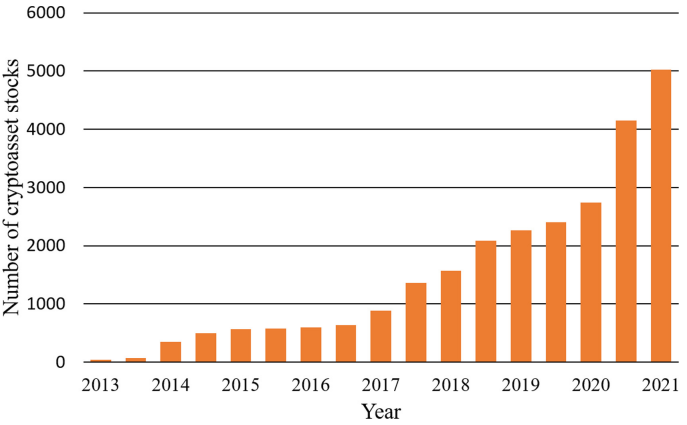


Fig. 1. Number of cryptoasset stocks

traditional cyberattacks. In fact, many cryptoasset incidents have occurred: in 2016, the cryptoasset exchange Bitfinex [2] suffered a hacking of approximately 70 million dollars in Bitcoin, which led to a temporary drop in the price of Bitcoin. Price fluctuations caused by incidents are detrimental to the stable operation of cryptoassets, and measures to deal with incidents are necessary.

Li et al. [3] and Wang et al. [4] summarize the major risks and attack methods of blockchain technology, but does not deal with actual incident cases. Grobys et al. [5] and Biais et al. [6] investigate the impact of major incidents on the price volatility of cryptoassets, but they do not mention the analysis of individual incidents or countermeasures, and the incidents discussed are only some incidents related to price volatility.

In this study, we investigate and summarize past incidents involving cryptoassets one by one using news articles and other sources. Then, each incident is classified by the "target of damage" and the "cause of damage", and the changing incident risk was discussed by analyzing the trends and characteristics of the time series of incidents. Our results show that the number of incidents and the amount of economic damage involving cryptoassets are on the increase. In terms of the classification by the target of damage, the damage related to cryptoasset exchanges is very large among all incidents. Our results also show that the risk of incidents related to cryptoassets increases due to the spread of altcoins in recent years. In terms of classification by cause of damage, incident risk due to blockchain and smart contract vulnerabilities is on the rise in recent years, and hence it was revealed that many decentralized exchanges were affected in 2020.

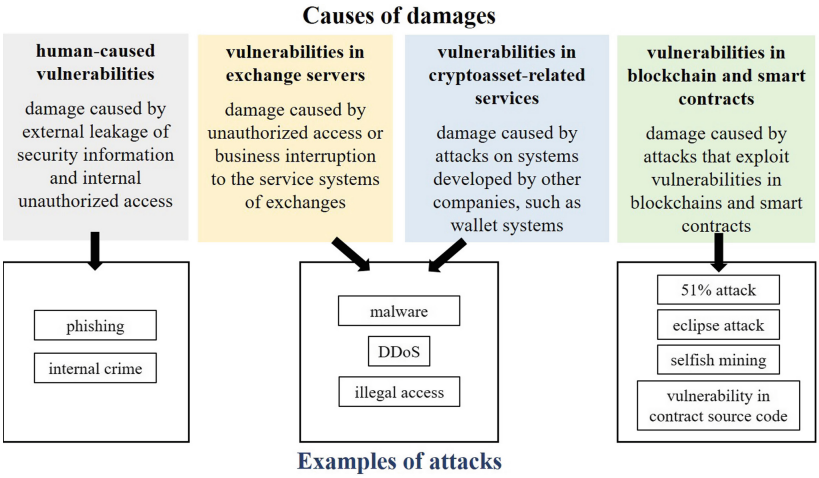


Fig. 2. Classification of attacks

2 Analysis

2.1 Classification Methodology

To understand the overall characteristics and chronological trends of incidents, we investigated incidents that actually occurred from 2009, the beginning of Bitcoin, to 2020. We refer to the official websites of cryptoasset exchanges and overseas news articles (the number of articles is 109) for incident cases in which actual financial damage is reported. In order to clarify the incident risk in detail, we categorize each incident according to the "target of damage" and the "cause of damage" to understand the overall characteristics and time-series trends of the incidents.

2.2 Classification of Incidents

2.2.1 Classification by the Target of Damage The classification is made based on the target of damage, and classified into three types: "cryptoasset exchanges", "cryptoasset-related services", and "cryptoasset stocks". Cryptoasset exchanges generally act as an agent for users to trade cryptoassets. In this case, cryptoasset exchanges store a large number of assets and signature keys for users, which are likely to be the target of attacks. The most commonly known "cryptoasset-related services" are related to cryptoassets other than exchanges, such as wallet services, decentralized finance (DeFi), and initial coin offerings (ICOs), as these services can cause a lot of damage if they are attacked. When ordinary users want to handle cryptoassets, they usually use at least one of the "cryptoasset exchanges" or "cryptoasset-related services". Cryptoassets themselves, including BTC or ETH, which are classified as "cryptoasset stocks", may

be vulnerable to attacks because they can have software and hardware vulnerabilities.

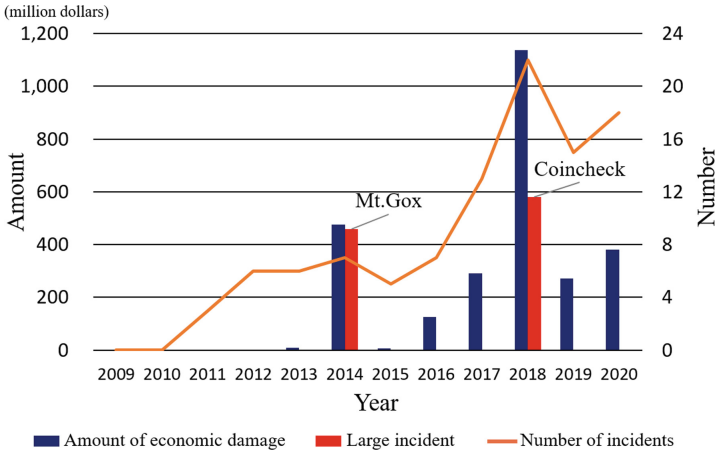


Fig. 3. The amount of economic damage and the number of incidents

2.2.2 Classification by the Cause of Damage Figure 2 shows the classification of attacks by the cause of damage. There are four types of causes: human-caused vulnerabilities, vulnerabilities in exchange servers, vulnerabilities in cryptoasset-related services, and vulnerabilities in blockchain and smart contracts. "Human-caused vulnerabilities" represents the damage caused by external leakage of security information and internal unauthorized access, such as phishing and insider trading, which already existed before the advent of cryptoassets. It is difficult to improve this situation unless users' security awareness is raised. "Vulnerability of exchange servers" represents the damage caused by unauthorized access or business interruption to the service systems of exchanges that handle transactions of cryptoassets on behalf of users. "Vulnerabilities in cryptoasset-related services" represent the damage caused by attacks on systems developed by other companies, such as wallet systems. Such systems, along with exchange servers, may be vulnerable to malware, DDoS, unauthorized access, and other attacks. "Vulnerabilities in blockchain and smart contract" represents the damage caused by attacks that exploit vulnerabilities in blockchains and smart contracts, including 51% attack, eclipse attack, selfish mining, and vulnerability in contract source code.

We use an example of decentralized exchanges to illustrate our classification. Decentralized exchanges allow users to manage their own wallets and secret keys, rather than having them managed by the cryptoasset exchange, and to conduct transactions directly with other users. This type of exchange avoids the risk of assets being concentrated in a single location, as is the case with

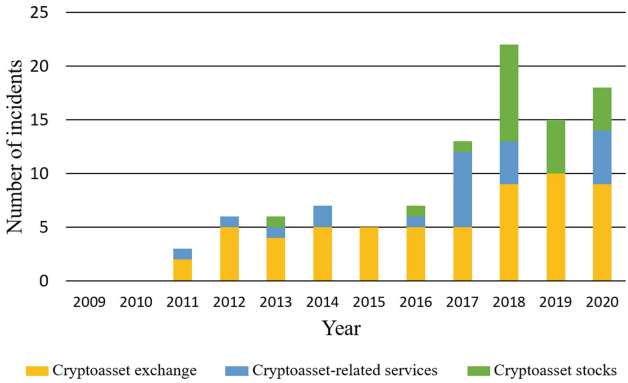


Fig. 4. Classification by object of damage (mumber of incidents)

traditional "centralized exchanges", but it is subject to the vulnerability of the smart contracts that conduct transactions. When an incident occurs, the target of damage is classified as a "cryptoasset exchange" and the cause of damage is classified as a "vulnerabilities in blockchain and smart contract".

3 Results

3.1 Total Number of Incidents and Total Economic Damage

Figure 3 shows the results of our analysis of actual incidents from 2009 to 2020. The total number of incidents is 102 and the total amount of economic dam-

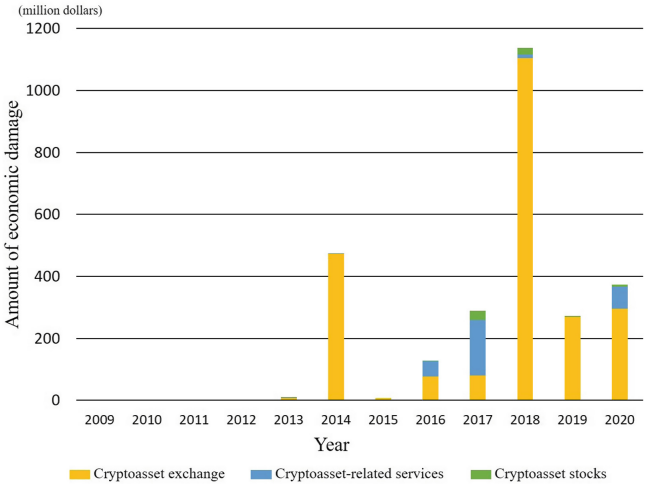


Fig. 5. Classification by object of damage (amount of economic damage)

age is 2.69 billion dollars. The total amount of economic damage is prominent in 2014 and 2018 due to the occurrence of large incidents. Excluding the Mt. Gox [7] incident in 2014 and the Coincheck [8] incident in 2018, the amount of economic damage and the number of incidents have been increasing every year. The cause is thought to be influenced by the increase in the value and attention of cryptoassets.

3.2 Classification by the Object of Damage

Figures 4 and 5 show the results of classifying incidents by the number of incidents and the amount of economic damage, respectively. Figure 4 shows that cryptoasset exchanges have the largest number of incidents, while cryptoasset-related services and cryptoasset stocks have almost the same number of incidents. Figure 5 shows that incidents against cryptoasset exchanges are by far the largest in terms of the amount of economic damage. Cryptoasset exchanges manage the wallets of a large number of users, and this is thought to be the reason why damage tends to be large. In addition, the number of incidents involving cryptoasset-related services has been increasing since around 2017, and the number of incidents involving cryptoasset stocks has been increasing since around 2018. This is due to the increase in smart contract-related cryptoasset services such as DeFi and ICOs, as well as the increase in altcoins.

3.3 Classification by the Cause of Damage

Figure 6 shows the number of incidents classified by the cause of damage, and Fig. 7 shows the amount of economic damage. Figure 6 shows that the number of incidents caused by vulnerabilities in blockchain and smart contract has

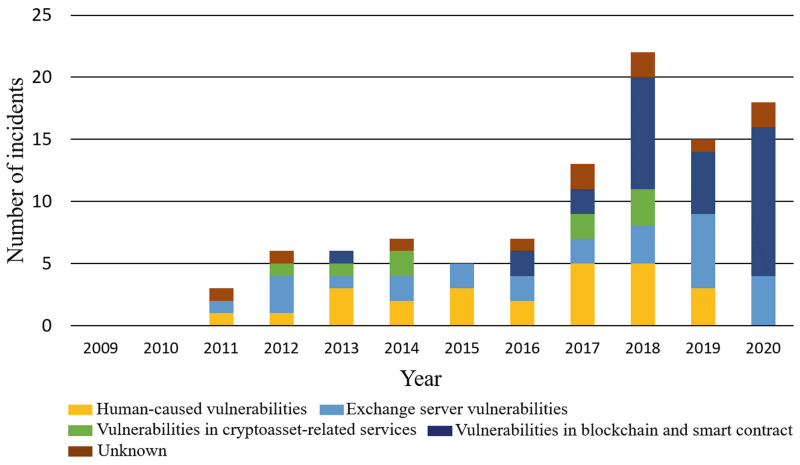


Fig. 6. Classification by cause of damage (number of incidents)

increased. This can be caused by the increase in the number of altcoins, services using smart contracts, and cryptoasset exchanges. Exchange server vulnerabilities continue to occur and are on the rise. In recent years, a relatively large amount of economic damage is caused by exchange server vulnerabilities. Incidents of human-caused vulnerabilities occur almost every year, and in recent years, frauds using cryptoasset-related services have also occurred.

To understand the chronological trends of incidents, the actual incidents from 2011 to 2020 are divided into five-year periods, and the number of incidents and the amount of economic damage are shown in Fig. 8. The number of human-caused vulnerability incidents is always large and the amount of economic damage increases significantly, and in some cases, a single incident can cause a huge amount of economic damage. Therefore, it is necessary for users to have a high level of information literacy when handling cryptoassets. The number of blockchain and smart contract vulnerabilities has increased, but the amount of economic damage has not. The number of exchange server vulnerabilities has increased both in the number of incidents and the amount of economic damage, and both of them are relatively larger than other causes of damage.

4 Discussion

Our results show that the number of incidents and the amount of economic damage involving cryptoassets is increasing every year. In addition, the probability of being the target of an attack and the types of attack methods have increased as the attention to cryptoassets has grown, and the incident risk has increased.

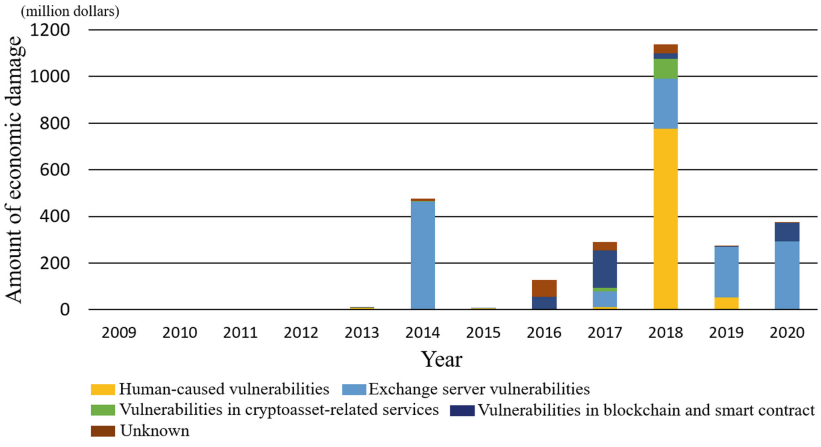


Fig. 7. Classification by cause of damage (amount of economic damage)

There are several findings from the classification of incidents into target and cause of damage. First, there are a number of incidents in which cryptoasset exchanges are the target of damage and the cause of damage. Cryptoasset

exchanges manage large amounts of cryptoassets and are easy target because a successful attack can lead to large profits. Because of this, risk countermeasures for exchanges are very important. Incidents related to blockchain and smart contracts have also increased in recent years, likely due to the increase in new altcoins and services related to cryptoassets using smart contract technology. These altcoins are relatively susceptible to 51% attacks, and their services often have high security risks, such as inadequate security. In fact, many decentralized exchanges that were considered revolutionary and highly secure in 2020 have suffered from incidents, requiring countermeasures for future operations. Furthermore, incidents caused by "human-caused vulnerability" have been occurring every year, indicating that lack of knowledge and understanding of information held by people is always an issue, and suggesting the need for users to have high information literacy when handling cryptoassets.

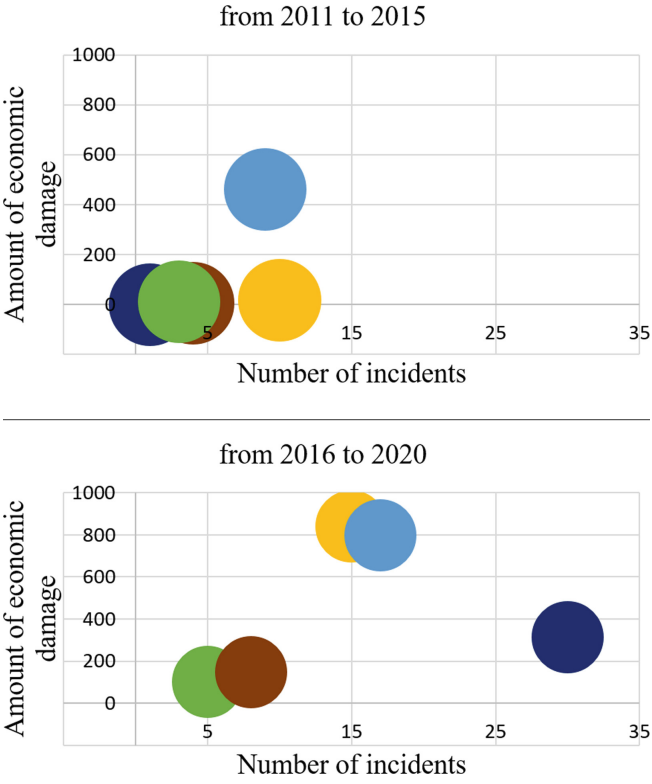


Fig. 8. Classification by object of damage

5 Conclusion

The purpose of this study is to clarify the incident risks surrounding cryptoassets, and a analysis of incidents that have occurred worldwide in the past was conducted. Our analysis shows that the number of incidents has been increasing worldwide, and that there is a diverse mix of incident risks, including exchange-related risks, which have remained a major issue since the early days, and blockchain-related risks, which have emerged in recent years with the development of cryptoassets.

As a result of our analysis, we believe that cryptoasset users and cryptoasset providers need to take measures to ensure the stable management of cryptoassets in the future. First, it is most important for users to understand the risks involved in using cryptoassets exchanges and cryptoasset-related services. Then, it is important for users to be cautious with cryptoassets by diversifying their investments and avoiding new services unnecessarily in order to reduce the damage caused by incidents. In contrast, service providers should not only make conscious improvements, but also establish a framework for providing secure services by setting uniform security standards for providing services. In addition, while research on the risks of cryptoassets has so far focused only on blockchain, which is the fundamental technology for cryptoassets, we believe that research focusing on the risks of services that handle cryptoassets, such as cryptoassets exchanges, will become more important in reducing actual incidents in the future.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number JP22H03588.

References

1. CoinMarketCap: cryptocurrency historical data snapshot. https://coinmarketcap.com/historical. Last viewed: 4 Sep. 2022
2. Coindesk: the bitfinex Bitcoin hack: what we know (and don't know) (2016). https://www.coindesk.com/bitfinex-bitcoin-hack-know-dont-know. Last viewed: 11 Oct 2021
3. Li, X., Jiang, P., Chen, T., Luo, X., Wen, Q.: A survey on the security of blockchain systems. Future Gener. Comput. Syst. **107**, 841–853 (2020)
4. Wang, Z., Jin, H., Dai, W., Choo, K.-K.R., Zou, D.: Ethereum smart contract security research: survey and future research opportunities. Front. Comput. Sci. **15**(2), 1–18 (2020). https://doi.org/10.1007/s11704-020-9284-9
5. Klaus, G., Niranjan, S.: Contagion of uncertainty: transmission of risk from the cryptocurrency market to the foreign exchange market. SSRN Electron. J. (2019)
6. Biais, B., Bisiere, C., Bouvard, M., Casamatta, C., Menkveld, A.J.: Equilibrium Bitcoin pricing. SSRN Electron. J., 74 (2018)
7. WIRED: the inside story of Mt. Gox. Bitcoin's $460 Million Disaster (2014). https://www.wired.com/2014/03/bitcoin-exchange/. Last viewed: 11 Oct 2021
8. Trend Micro: Coincheck suffers biggest hack in cryptocurrency history; Experty users tricked into buying false ICO (2018). https://www.trendmicro.com/vinfo/fr/security/news/cybercrime-and-digital-threats/coincheck-suffers-biggest-hack-in-cryptocurrency-experty-users-buy-false-ico. Accessed 11 Oct 2021



Authenticated Encryption Engine for IoT Application

Heera Wali, B. H. Shraddha(✉), and Nalini C. Iyer

KLE Technological University, Hubballi, Karnataka, India
{heerawali,shraddha_h,nalinic}@kletech.ac.in

Abstract. With the increase of connected devices in IoT paradigm for various domains that include wireless sensor networks, edge computing, embedded systems. Hence the cryptographic primitives deployed on these devices have to be lightweight as the devices used are low cost and low energy. The cryptographic techniques and algorithms for data confidentiality only aim at providing data privacy, but the authenticity of the data is not being addressed. Hence Authenticated Encryption (AE) is used to provide a higher level of security. Authenticated encryption is a scheme used to provide authenticity along with confidentiality of the data. In this pper, AE is implemented using a lightweight PRESENT encryption algorithm and hashing SPONGENT algorithm. These algorithms have the smallest foot-print compared to other lightweight algorithms. The proposed design uses a PRESENT block cipher for the key size variants of 80 bits and 128 bits for a block size of 64-bit and SPONGENT variants of 88, 128, and 256 bits for authentication. Simulation, analysis, inference, and synthesis of the proposed architectures i.e. Encryption then MAC (EtM) is implemented on the target platform ARTY A7 100T. A comparative analysis states that the combination of the PRESENT 80 block cipher and SPONGENT 88 variant is best suited for the resource-constrained Internet of Things applications as the world is slowly approaching the brink of mankind's next technological revolution.

Keywords: Cryptography · Cyber-security · Symmetric block cipher · Authentication · Internet of Things

1 Introduction

In crucial applications, the need for security in IoT applications is dramatically increasing. These applications require efficient and more secure implementable cryptographic primitives including ciphers and hash functions. In such applications of constrained resources, the area and power consumption have major importance. These resource constraints devices are connected over the internet to transmit and receive data from one end to the other end. Hence it is necessary to protect the data which is transmitted from the intervention of the third party. The conventional cryptographic primitives cannot be used for resource-constrained devices to achieve the tasks of protecting the data as they are expensive to implement. To overcome this situation, some significant research has been

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1102–1113, 2023.
https://doi.org/10.1007/978-3-031-27409-1_101

performed. The lightweight cryptographic primitive designs have closely approached the minimal hardware footprint and the designs for lightweight. Due to this, the scope for the design and effective implementation of lightweight cryptographic primitives arises. IoT is a network of sensors, controlling units, and software that exchange data with other systems over the internet. Hence to provide both confidentiality and authenticity of data in resource-constrained environments like IoT, authenticated encryption should be implemented using a lightweight encryption algorithm and a lightweight hashing algorithm. The encryption algorithm helps to maintain the confidentiality of the data or message. To find whether the data received is genuine or not, hashing algorithm or message authentication code (MAC) is used. The hash value/message digest (output of hashing algorithm which is computationally irreversible) is sent to the receiver along with cipher text. Due to the same reason authenticated encryption is used. This paper implements architecture of Encrypt then MAC (EtM) which is designed and has been implemented using a modular approach.

2 Related Work

Due to the environmental changes over the last decade, green innovation is gaining importance. Green innovation in the field of technology consists of green computing and networking. The trend aims at the selection of the methodologies with energy-efficient computation and minimal resource utilization wherever possible [1]. The lightweight cryptography project was started by NIST (National Institute of Standards and Technology) in 2013. The increase in deployment of small computing devices that are interconnected to perform the assigned task with resource constraints led to the integration of cryptographic primitives. The security of the data in these devices is an important factor as they are concerned with the areas like sensor networks, healthcare, the Internet of Things (IoT), etc. The current NIST- approved cryptographic algorithms are not acceptable as they were designed for desktops/servers. Hence, the main objective of the NIST lightweight cryptography project was to develop a strategy for the standardization of lightweight cryptographic algorithms [2]. Naru et al. [3] describes the need for security and lightweight algorithms for data protection in IoT devices. The conventional cryptographic primitives cannot be used in these applications because of the large key size as in the case of RSA and high processing requirements. The lightweight cryptography on Field Programmable Gate Array (FPGA) has become a research area with the introduction of FPGAs to battery powered devices [4, 13]. The re-configurability feature of FPGA is an advantage along with the low cost and low power. The cryptographic primitives should be lightweight for application in the field of resource-constrained environments. PRESENT is an ultra-lightweight block cipher with Substitution Permutation Network (SPN). The hardware requirements of PRESENT are less in comparison with other lightweight encryption algorithms like MCRYPTON and HIGHT. The PRESENT algorithm is designed especially for the area and power-constrained environments without compromising in security aspects. The algorithm is designed by looking at the work of DES and AES finalist Serpent [5]. PRESENT has a good performance and implementation size based on the results as described in the paper [6]. As per the discussions and analysis from [7, 12], SPONGENT has the round function with a smaller logic size

than QUARK (a lightweight hashing algorithm). SPONGENT is a lightweight hashing algorithm with sponge-based construction and SPN. The SPONGENT algorithm has a smaller footprint than other lightweight algorithms like QUARK and PRESENT in hashing mode. Jungk et al. [8] illustrates that the SPONGENT implementations are most efficient in terms of throughput per area and can be the smallest or the fastest in the field, depending on the parameters. The paper [9] describes the need of using authentication algorithm with the encryption algorithm. It states that the security of the data with authenticated encryption is more compared with the only encryption scheme. Authenticated Encryption has 3 different compositions/modes. They are as follows, (1) Encrypt and MAC, (2) MAC then Encrypt, and (3) Encrypt then MAC.

The security aspects of all three modes of AE are tabulated in Table 1. The security of encryption algorithm considering in-distinguishability under Chosen Plaintext Attack (IND-CPA) and Chosen Cipher text Attack (IND-CCA) and authentication algorithm considering the integrity of plaintext (INT-PTXT) and integrity of cipher text (INT-CTXT). As per the discussion of Lara-Nino et al. [10], the Encrypt and MAC mode is secure compared to the other two modes.

Table 1. Security aspects of three modes of AE

AE mode	Confidentiality		Authentication	
	IND-CPA	IND-CCA	INT-PTXT	INT-CTXT
Encrypt and MAC	Insecure	Insecure	Secure	Insecure
MAC then Encrypt	Secure	Insecure	Secure	Insecure
Encrypt then MAC	Secure	Secure	Secure	Secure

3 Overview of Lightweight Encryption and Authentication Algorithm

The authenticated encryption proposed in this paper makes use of a lightweight encryption block cipher i.e., PRESENT algorithm, and lightweight hashing function i.e., SPONGENT algorithm to produce a cipher text and hash value as the output respectively. The flow of AE in Encrypt then MAC (EtM) is described in Sect. 3.1.

3.1 Authenticated Encryption in Encrypt then MAC Mode

Encrypt then MAC mode follows the below steps and provides the hash value of cipher text as the result.

1. The message (plaintext) and key are given as input to an encryption algorithm.
2. The output of the encryption algorithm, i.e., the cipher- text is provided as input to the MAC algorithm.
3. The output of the MAC algorithm is the output of this AE mode.

3.2 PRESENT Lightweight Encryption Algorithm

Encryption algorithms take two inputs plaintext and key to obtain cipher text. The PRESENT algorithm is a 32 round Substitution-Permutation network (SPN) based block cipher with a block size of 64-bits and a key with a length of 80-bits or 128-bits. The algorithm is further divided into two sections,

1. To update the block of 64-bits to produce cipher text of 64-bits in 32 rounds.
2. The key scheduling, where the key (80-bits or 128-bits) is updated for every round.

The top-level description of the PRESENT algorithm is as shown in Fig. 1. The three operations are carried out for each round. The three operations are,

1. addRoundKey: The MSB 64-bits of roundKeyi (which is updated at each round using the Key scheduling section) is XORed with the 64 bits of the block.
2. sBoxLayer: It takes the 4-bit input from the previous stage and provides 4-bit output by following the rule as described in Table 2. The value from each group of 4- bits combined in the order which was divided gives the updated value of the block of 64-bits.
3. pLayer: It is a rearrangement of the bits of the block. The ith bit of state is moved to the P(i)th position of the block of 64-bits. The order of rearrangements of the bits is formulated as provided in the mathematical expression below. The updated value of the block of 64- bits from the previous step is taken as input to this step and updated according to the Eq. (1).

$$P(i) = \begin{matrix} i(16)\, mod\, (63); & \text{if } 0 \leq i < 63 \\ 63; & \text{if } i = 63 \end{matrix} \tag{1}$$

These operations are carried out for 32 rounds. Key scheduling: The user provided key is updated by performing some bit manipulation operations. The bit manipulation operations carried for a key size of 80-bits and 128- bits are described below. The Key value provided as input is assigned to roundKey_i (i represents the round of PRESENT algorithm) initially for 1st round. The following steps are performed for 80-bit key scheduling. $k_{79}, k_{78} \ldots k_1, k_0 = k_{18}, k_{17} \ldots k_1, k_0, k_{79}, k_{78}, k_{20}, k_{19}$.

First 4-bit is passed through S-box as, First 4-bit is passed through S-box as, $[k_{79}\ k_{78}\ k_{77}\ k_{76}] = S\ [k_{79}\ k_{78}\ k_{77}\ k_{76}]$ The value of counter is XORed with 5-bit of key as shown, $k_{19}\ k_{18}\ k_{17}\ k_{16}\ k_{15} = k_{19}\ k_{18}\ k_{17}\ k_{16}\ k_{15} \oplus$ roundKeyi. Similarly, with 128-bit key, the following three steps is performed for key scheduling as shown below. $k_{127}, k_{126},., k_1, k_0 = k_{66}, k_{65},., k_1, k_0, k_{127}, k_{126},., k_{68}, k_{67}$ $[k_{127}\ k_{126}\ k_{125}\ k_{124}] = S\ [k_{127}\ k_{126}\ k_{125}\ k_{124}]$ and $[k_{123}\ k_{122}\ k_{121}\ k_{120}] = S\ [k_{123}\ k_{122}\ k_{121}\ k_{120}]$ $k_{66}\ k_{65}\ k_{64}\ k_{63}\ k_{62} = k_{66}\ k_{65}\ k_{64}\ k_{63}\ k_{62} \oplus$ roundKeyi.

3.3 SPONGENT Lightweight Hashing Algorithm

SPONGENT is a hashing algorithm used to produce the message digest of the given input message. The construction of the algorithm explores an iterative design to produce

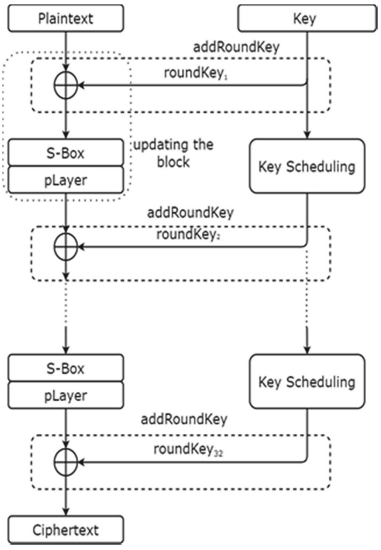


Fig. 1. The top-level description of PRESENT algorithm

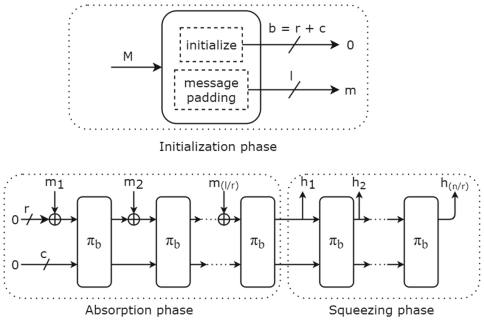


Fig. 2. The top-level description of SPONGENT algorithm

Table 2. S-Box of Present

x	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
S(x)	C	5	6	B	9	0	A	D	3	E	F	8	4	7	1	2

the hash value of n-bits based on permutation block π_b, operating on a fixed number of bits 'b', where 'b' is block-size. The SPONGENT hashing algorithm mainly consists of sponge construction blocks as shown in Fig. 2. The SPONGENT algorithm has three phases of operation i.e., initialization phase, absorption phase, and squeezing phase. The data of block size 'b' is padded with the multiple of 'r' bits that define the bitrate such that $b = r + c$ where 'c' defines the capacity and 'b' defines the state. In the later stage,

the padded message of length 'l' bit is divided into 'r' bit message blocks m1, m2, and $m_{(l/r)}$, which are XORed into the first 'r' bits of the state of 'b' bits, this is known as absorption of the message. The state value is passed to the permutation blocks π_b. The following operations are carried in the order given below in one round of a permutation block π_b on the state of b-bits. The value of state after one round of permutation block operation is the input value of the state to the next round of permutation block operation. In each permutation block, the following operation is carried out for 'R' rounds in a sequential manner [11]. The rounds 'R' for 3 different SPONGENT variants are listed in Table 3.

Table 3. SPONGENT Variants

SPONGENT variant	Bits				Rounds
	n	c	r	b	R
SPONGENT–88/80/08	88	80	8	88	45
SPONGENT–128/128/08	128	128	8	136	70
SPONGENT–256/256/16	256	256	16	272	140

Once the operation of permutation block π_b is completed, the next r-bits of the message are XORed with the first r-bits of the state, which is the output of the previous permutation block. This is carried until all bits of the padded message is absorbed and operated in permutation block π_b. Further, when all the blocks are absorbed the first 'r' bits of the state are returned which is represented as h1 in Fig. 3. This is the first MSB r-bits of the hash value. These 'r' bits after every permutation block π_b i.e., h2, h3, and up to $h_{(n/r)}$ are combined in MSB to LSB fashion to produce the hash value (output), until n-bits of the hash value are generated, where 'n' is the hash-size.

4 Proposed Design for Implementation of AE

The proposed design is described in the subsections below. The subsections include the details of the design of the AE module and its FSM, PRESENT algorithm with its FSM, and SPONGENT algorithm with its FSM.

4.1 Proposed Design for Authenticated Encryption (AE)

The architecture consists of a top module for the operation of AE, consisting of two sub-modules PRESENT and SPONGENT, as shown in Fig. 3. The FSM of the proposed design is as shown in Fig. 4 it alters the value of present_reset and spongent_reset depending on the current state, which triggers the respective sub-modules to carry out the functions.

When the reset is '1', the FSM is in State 0. In this state, the present_reset and spongent_reset values are set to '1'. When reset becomes '0' it moves to State 1, the

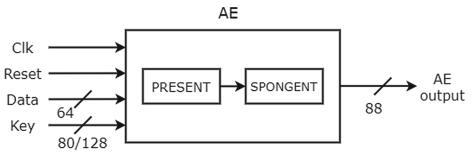


Fig. 3. Top module, Authenticated Encryption

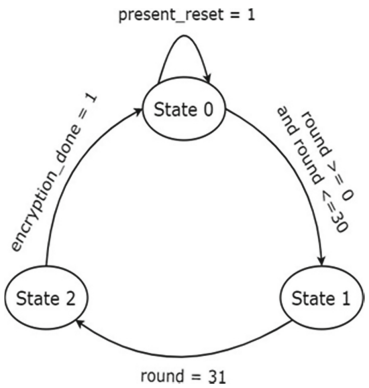


Fig. 4. State Diagram of Authenticated Encryption module

PRESENT algorithm block is triggered (present reset value is set to '0') in this state. When the output of the PRESENT algorithm is obtained, the encryption done value is set to '1' by the PRESENT algorithm block. If encryption done is '1' the state will transit to State 2. In-State 2, the SPONGENT algorithm block is triggered (spongent_reset value is set to '0') to carry out its operation. After the completion of the SPONGENT algorithm's operation, the output hash value is obtained. The state will transit to State 0 when the hash length is '0' and encryption_done is '1'. Authenticated Encryption is implemented for variants of PRESENT (encryption algorithm) and SPONGENT (hashing algorithm) as tabulated in Table 4.

Table 4. Implemented AE with Variants of PRESENT and SPONGENT

Authenticated Encryption (Encrypt then MAC)	PRESENT variant	SPONGENT variant
	PRESENT (80 bit key)	SPONGENT–88/80/08
		SPONGENT–128/128/08
		SPONGENT–256/256/16
	PRESENT (128 bit key)	SPONGENT–88/80/08
		SPONGENT–128/128/08
		SPONGENT–256/256/16

4.2 Description of PRESENT Encryption Algorithm FSM

The PRESENT algorithm takes the plaintext of length 64-bits and a key of 80-bits or 128-bits to produce the cipher-text of length 64-bits output. The state diagram (FSM) for the implementation of the PRESENT module is shown in Fig. 5. The transition from one state to another state in this FSM is based on the values of present_reset, round, and encryption done at the negative edge of the clock signal. When the value of the present_reset is '1' the state remains in State 0. The state moves to State 1 from State 0 when the present_ reset value is '0' and remains in that state till the round value is less than or equal to 30.

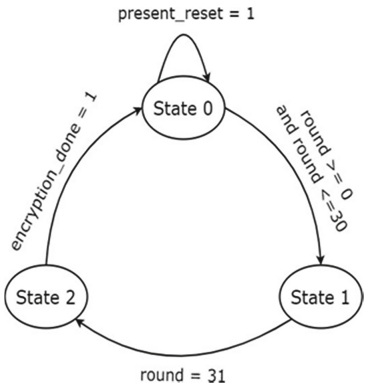


Fig. 5. State Diagram of module PRESENT

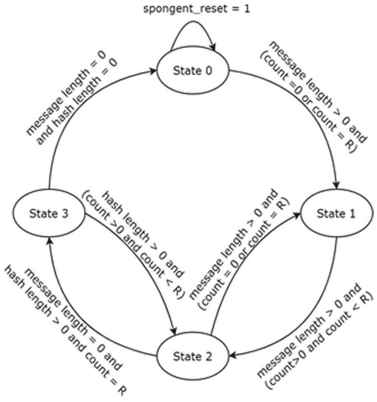


Fig. 6. FSM for top module SPONGENT

When the round value is 31, the state transits from State 1 to State 2. When encryption done is '1', the control shifts to the SPONGENT FSM for generating the hash value of the obtained cipher text.

4.3 Description of SPONGENT Lightweight Hashing Algorithm FSM

The state diagram for the Spongent hashing algorithm is shown in Fig. 6. The finite state machine of the Spongent hashing algorithm consists of four states from State 0 to State 3 where State 0 represents the idle state [8]. The state diagram represents the transitions of one state to another state depending on the values of spongent_reset, Message length, Hash length, and count.

5 Results

The proposed design is been implemented on the target platform Arty A7 100T FPGA board using Vivado Design Suite 2018.1. Figure 7 shows the hardware setup. The output waveforms for variants of PRESENT and SPON- GENT using EtM mode is captured on Vivado Simulation tool. The following variant combinations PRESENT-80 with variants of SPONGENT 88, 128 and 256, PRESENT-128 with variants of SPONGENT 88, 128 and 256. The resource utilization, throughput and logic delay for the above mentioned variants of PRESENT and SPONGENT is tabulated in Table 5. For the inputs, Plaintext = 0x0000000000000000 (64 bits) and Key = 0xffffffffffffffffffff (80 bits) or Key = 0xffffffffffffffffffffffffffffffff (128 bits), the output of simulation is obtained as shown in Fig. 8. The output waveforms for variants of PRESENT and SPONGENT using EtM mode is captured on Vivado Simulation tool. The following variant combinations PRESENT-80 with variants of SPONGENT 88, 128 and 256, PRESENT-128 with variants of SPONGENT 88, 128 and 256. The resource utilization, throughput and logic delay for the above mentioned variants of PRESENT and SPONGENT is tabulated in Table 5. For the inputs, Plaintext = 0x0000000000000000 (64 bits) and Key = 0xffffffffffffffffffff (80 bits) or Key = 0xffffffffffffffffffffffffffffffff (128 bits).

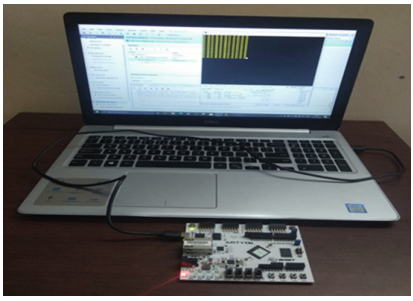


Fig. 7. Hardware Setup

The proposed method has an increased throughput even though the area might increase in terms of the number of slices when compared to paper [1] and is shown in able 7. The overall logical delay is reduced. Table 6 describes where the proposed method stands among the other FPGA implementations of the lightweight algorithms.

Table 5. Resource Utilization of AE with Different Combinations of PRESENT and SPONGENT

PRESENT variant (key size)	SPONGENT variant	Resource utilization				Throughput (Mbps)	Throughput/slice (Kbps/ slice)	Logic delay (ns)
		LUT	Slice	FF	Max Freq (MHz)			
PRESENT (80-bit)	SPONGENT 88/80/08	827	207	673	203.890	171.696	829.449	2.452
	SPONGENT 128/128/08	1639	410	854	217.732	183.353	447.202	2.29
	SPONGENT 256/256/16	1910	478	1368	196.757	165.690	346.632	2.451
PRESENT (128-bit)	SPONGENT 88/80/08	1264	316	721	206.432	173.83	550.095	2.422
	SPONGENT 128/128/08	1726	432	941	211.762	178.32	412.778	2.31
	SPONGENT 256/256/16	2003	501	1510	182.62	153.78	306.946	2.49

6 Conclusion

The proposed work gives an efficient implementation of the authenticated encryption (AE) using the lightweight crypto- graphic algorithms and hashing algorithm on a target board ARTY A7 100T. The design is being validated for different test cases. PRESENT algorithm is chosen for encryption and SPONGENT algorithm for authentication. All the variants of SPONGENT 88, 128 and 256 is been realized with the PRESENT for the key size of 80 bits and 128 bits respectively as AE paradigm. The different combinations of the authenticated encryption with PRESENT and SPONGENT implementations have been tabulated of which PRESENT-80 with block size of 64-bits and SPONGENT-88 having a smaller footprint and hence efficient in terms of flip flop utilization and throughput (Table 5).

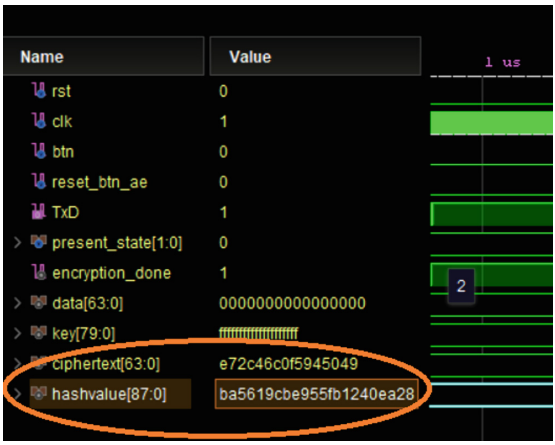


Fig. 8. AE (PRESENT 80 and SPONGENT 88)

Table 6. Performance Comparison of FPGA Implementations oF Lightweight Cryptographic Algorithms

Parameters	Proposed Design	George Hatzivasilis et al.
FPGA	Arty A7 100T	Virtex 5
Algorithm	PRESENT 80	PRESENT 80
Flip flops	150	–
Slice	68	162
Throughput	883.61 Mbps	206.5 Kbps
Throughput/slice	12.994 Mbps/Slice	1.28 Kbps/Slice
Algorithm	SPONGENT 88	SPONGENT 88
Flip flops	304	–
Slice	231	95
Throughput	–	195.5 Kbps
Throughput/slice	–	2.06 Kbps/Slice
Algorithm	ETM with PRESENT 80 and SPONGENT 88	ETM with PRESENT 80 and SPONGENT 88
Flip flops	673	149
Slice	207	174
Throughput	171.696 Mbps	82.64 Kbps
Throughput/slice	829.449 Kbps/Slice	0.47 Kbps/Slice

References

1. Hatzivasilis, G., Floros, G., Papaefstathiou, I., Manifavas, C.: Lightweight authenticated encryption for embedded on-chip systems. Inf. Secur. J.: A Global Perspect. **25** (2016)
2. McKay, K.A., Bassham, L., So¨nmez Turan, M., Mouha, N.: Report on Lightweight Cryptography. National Institute of Standards and Technology (2016)
3. Naru, E.R., Saini, H., Sharma, M.: A recent review on lightweight cryptography in IoT. In: International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (2017)
4. Yalla, P., Kaps, J.-P.: Lightweight cryptography for FPGAs. In: 2009 International Conference on Reconfigurable Computing and FPGAs (2009)
5. Bogdanov, A., Knudsen, L.R. Leander, G., Paar, C., Poschmann, A., Robshaw, M.J.B., Seurin, Y., Vikkelsoe, C.: PRESENT: An Ultra-Lightweight Block Cipher
6. Lara-Nino, C.A., Morales-Sandoval, M., Diaz-Perez, A.: Novel FPGA-based low-cost hardware architecture for the PRESENT Block Cipher. In: Proceedings of the 19th Euromicro Conference on Digital System Design, DSD 2016, pp. 646–650, Cyprus, September 2016
7. Bogdanov, A., Knezevic, M., Leander, G., Toz, D., Varıcı, K., Verbauwhede, I.: SPONGENT: The Design Space of Lightweight Cryptographic Hashing
8. Jungk, B., Rodrigues Lima, L., Hiller, M.: A Systematic Study of Lightweight Hash Functions on FPGAs. IEEE (2014)
9. Andres Lara-Nino, C., Diaz-Perez, A., Morales-Sandova, M.: Energy and Area Costs of Lightweight Cryptographic Algorithms for Authenticated Encryption in WSN, September (2018)
10. Bellare, M., Namprempre, C.: Authenticated encryption: relations among notions and analysis of the generic composition paradigm. J. Cryptol. J. Int. Assoc. Cryptol. Res. **21**(4), 469–491 (2008)
11. Lara-Nino, C.A., Morales-Sandoval, M., Diaz-Perez, A.: Small lightweight hash functions in FPGA. In: Proceedings of the 2018 IEEE 9th Latin American Symposium on Circuits & Systems (LASCAS), pp. 1–4, Puerto Vallarta, Feburary (2018)
12. Buchanan, W.J., Li, S., Asif, R.: Lightweight cryptography methods. J. Cyber Secur. Technol. (2017), vol. 1, March (2018)
13. Shraddha, B.H., Kinnal, B., Wali, H., Iyer, N.C., Vishal, P.: Lightweight cryptography for resource constrained devices. In: Hybrid Intelligent Systems. HIS 2021. Lecture Notes in Networks and Systems, Vol. 420. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96305-7_51



Multi-layer Intrusion Detection on the USB-IDS-1 Dataset

Quang-Vinh Dang(✉)

Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
dangquangvinh@iuh.edu.vn

Abstract. Intrusion detection plays a key role in a modern cyber security system. In recent years, several research studies have utilized state-of-the-art machine learning algorithms to perform the task of intrusion detection. However, most of the published works focus on the problem of binary classification. In this work, we extend the intrusion detection system to multi-class classification. We use the recent intrusion dataset that reflects the modern attacks on computer systems. We show that we can efficiently classify the attacks to attack groups.

Keywords: Fraud Detection · Machine learning · Classification · Intrusion Detection

1 Introduction

Cyber-security is an important research topic in the modern computer science domain, as the risk of being attacked has been increasing over the years. One of the most important tasks in cyber-security is intrusion detection, in that an intrusion detection system (IDS) must recognize attacks from outside and prevent them to get into the computer systems.

Traditional IDSs rely on the domain expertise of security experts. The experts need to define some properties and characteristics, known as "signature", of the attacks. The IDS then will try to detect if incoming traffic has these signatures or not.

Indeed, the manual definition approach can not deal with the high number of new attacks that appear every day. The researchers started to use machine learning algorithms [5,8] to classify intrusion.

The machine learning-based IDSs have achieved a lot of success in recent years [8]. However, most of the published research works focus only on binary classification [7], i.e. they focus on predicting whether incoming traffic is malicious or not.

Detecting malicious traffic is definitely a very important task. However, in practice, we need to know what type of attack is [16]. By that, we can plan an effective defensive strategy according to the attack type [4,12].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1114–1121, 2023.
https://doi.org/10.1007/978-3-031-27409-1_102

In this paper, we address the problem of multi-attack type classification. We show that we can effectively recognize the attack class. We evaluate the algorithms using the dataset USB-IDS-1 [3], a state-of-the-art public intrusion dataset. We show that we can effectively classify attack class but not yet attack type.

2 Related Works

In this section, we review some related studies.

The authors of [5] studied and compared extensively some traditional machine learning approaches for tabular data to detect intrusion in the networks. They concluded that boosting algorithms perform the best.

Several authors argue that we don't need the entire feature set introduced with the open dataset like CICIDS2018 to perform the intrusion detection task. The authors of [11] presented a method to select relevant features and produce a lightweight classifier. The authors of [1] suggested a statistical-based method for feature extraction. The authors of [6] argued that explainability is an important indicator to determine good features.

Supervised learning usually achieved the highest results compared to other methods but required a huge labeled dataset. Several researchers have explored the usage of reinforcement learning to overcome the limitation of traditional supervised learning [9]. The methods presented in the work of [9] are extended in [15].

Deep learning has been studied extensively for the problem of intrusion detection [14]. The authors of [13] used collaborative neural networks and Agile training to detect the intrusion.

3 Methods

3.1 Logistic Regression

Logistic regression is a linear classification algorithm. The idea of the logistic regression is visualized in Fig. 1.

3.2 Random Forest

The random forest algorithm belongs to the family of bagging algorithms. In the random forest algorithm, multiple decision trees are built. Each tree will give an individual prediction, then these predictions are combined into the final prediction. The algorithm is visualized in Fig. 2.

3.3 Catboost

The catboost algorithm [10] belongs to the family of boosting algorithms. In the algorithm, multiple trees are built consequently. The following tree will try to recover the prediction error made by the previous tree. The details comparison of catboost versus other popular boosting libraries such as lightgbm and xgboost are presented in [2].

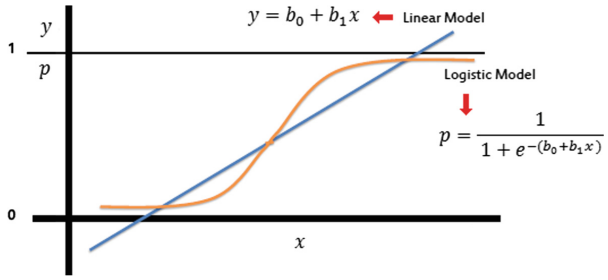


Fig. 1. Logistic regression

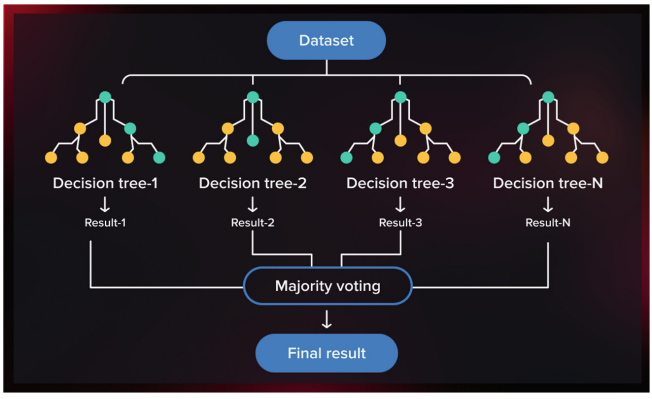


Fig. 2. Random forest

4 Datasets and Experimental Results

4.1 Dataset

We evaluated the algorithms against the dataset USB-IDS-1 [3].

Even though many public datasets about intrusion detection have been published [4], most of them do not consider the defense methods of the victim hosts. It makes the dataset for realistic.

Table 1. Class distribution in the dataset USB-IDS-1

Name of the `csv` file	Total	Attack	Benign
Hulk-NoDefense	870485	870156	329
Hulk-Reqtimeout	874382	874039	343
Hulk-Evasive	1478961	770984	707977
Hulk-Security2	1461541	762070	69471
TCPFlood-NoDefense	330543	48189	282354
TCPFlood-Reqtimeout	341483	59102	282381
TCPFlood-Evasive	341493	59113	282380
TCPFlood-Security2	341089	58716	282373
Slowloris-NoDefense	2179	1787	392
Slowloris-Reqtimeout	13610	13191	419
Slowloris-Evasive	2176	1784	392
Slowloris-Security2	2181	1790	391
Slowhttptest-NoDefense	7094	6695	399
Slowhttptest-Reqtimeout	7851	7751	100
Slowhttptest-Evasive	7087	6694	393
Slowhttptest-Security2	7090	6700	390

The distribution of benign and attack classes in the dataset is visualized in Table 1.

4.2 Experimental Results

Table 2. Performance of binary classifier

Method	Accuracy	F1	AUC
LR	0.86	0.84	0.89
RF	0.92	0.93	0.92
CB	0.99	0.99	0.99

We presented the performance of binary classifiers in Table 2. We see that the CatBoost algorithm achieved the highest performance.

We show the confusion matrix of CatBoost algorithm in Fig. 3. CatBoost can accurately classify all the instances in the test set.

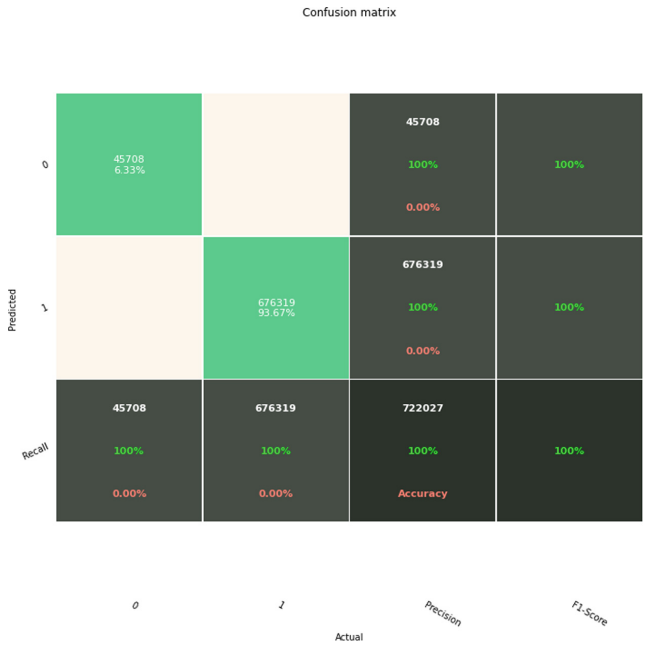


Fig. 3. Binary classification: benign vs malicious

We show the attack class classification in Fig. 4 and attack type classification in Fig. 5.

We can see that CatBoost can classify attack class, but it misclassifies when it needs to detect the attack type.



Fig. 4. Attack class classification

5 Conclusions

Detecting attack class is an important task in practice. In this paper, we study the problem of attack class classification. We can classify attack classes, but there is still room for improvement to detect attack types. We will investigate this problem in the future.

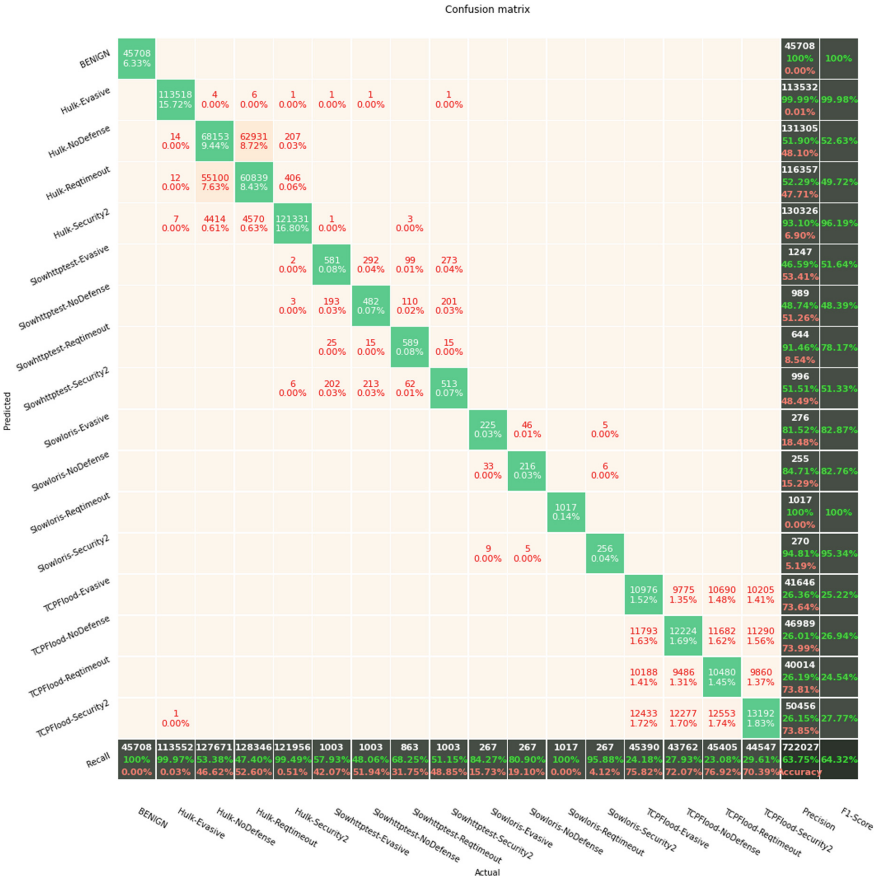


Fig. 5. All class classification

References

1. Al-Bakaa, A., Al-Musawi, B.: A new intrusion detection system based on using non-linear statistical analysis and features selection techniques. Comput. Secur., 102906 (2022)
2. Al Daoud, E.: Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. Int. J. Comput. Inf. Eng. **13**(1), 6–10 (2019)
3. Catillo, M., Del Vecchio, A., Ocone, L., Pecchia, A., Villano, U.: Usb-ids-1: a public multilayer dataset of labeled network flows for ids evaluation. In: 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), pp. 1–6. IEEE (2021)
4. Catillo, M., Pecchia, A., Rak, M., Villano, U.: Demystifying the role of public intrusion datasets: a replication study of dos network traffic data. Comput. Secur. **108**, 102341 (2021)

5. Dang, Q.V.: Studying machine learning techniques for intrusion detection systems. In: International Conference on Future Data and Security Engineering, pp. 411–426. Springer (2019)
6. Dang, Q.V.: Improving the performance of the intrusion detection systems by the machine learning explainability. Int. J. Web Inf. Syst. (2021)
7. Dang, Q.V.: Intrusion detection in software-defined networks. In: International Conference on Future Data and Security Engineering, pp. 356–371. Springer (2021)
8. Dang, Q.V.: Machine learning for intrusion detection systems: recent developments and future challenges. In: Real-Time Applications of Machine Learning in Cyber-Physical Systems, pp. 93–118 (2022)
9. Dang, Q.V., Vo, T.H.: Studying the reinforcement learning techniques for the problem of intrusion detection. In: 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 87–91. IEEE (2021)
10. Dorogush, A.V., Ershov, V., Gulin, A.: Catboost: gradient boosting with categorical features support (2018). arXiv:1810.11363
11. Kaushik, S., Bhardwaj, A., Alomari, A., Bharany, S., Alsirhani, A., Mujib Alshahrani, M.: Efficient, lightweight cyber intrusion detection system for IoT ecosystems using mi2g algorithm. Computers **11**(10), 142 (2022)
12. Kizza, J.M., Kizza, W., Wheeler: Guide to Computer Network Security. Springer (2013)
13. Lee, J.S., Chen, Y.C., Chew, C.J., Chen, C.L., Huynh, T.N., Kuo, C.W.: Conn-ids: intrusion detection system based on collaborative neural networks and agile training. Comput. Secur., 102908 (2022)
14. Malaiya, R.K., Kwon, D., Kim, J., Suh, S.C., Kim, H., Kim, I.: An empirical evaluation of deep learning for network anomaly detection. In: ICNC, pp. 893–898. IEEE (2018)
15. Pashaei, A., Akbari, M.E., Lighvan, M.Z., Charmin, A.: Early intrusion detection system using honeypot for industrial control networks. Results Eng., 100576 (2022)
16. Van Heerden, R.P., Irwin, B., Burke, I.: Classifying network attack scenarios using an ontology. In: Proceedings of the 7th International Conference on Information-Warfare & Security (ICIW 2012), pp. 311–324 (2012)



Predictive Anomaly Detection

Wassim Berriche[1] and Francoise Sailhan[2(✉)]

[1] SQUAD and Cedric Laboratory, CNAM, Paris, France
[2] IMT Atlantique, LAB-STICC Laboratory,Brest, France
francoise.sailhan@cnam.fr

Abstract. Cyber attacks are a significant risk for cloud service providers and to mitigate this risk, near real-time anomaly detection and mitigation plays a critical role. To this end, we introduce a statistical anomaly detection system that includes several auto-regressive models tuned to detect complex patterns (e.g. seasonal and multi-dimensional patterns) based on the gathered observations to deal with an evolving spectrum of attacks and a different behaviours of the monitored cloud. In addition, our system adapts the observation period and makes predictions based on a controlled set of observations, i.e. over several expanding time windows that capture some complex patterns, which span different time scales (e.g. long term versus short terms patterns). We evaluate the proposed solution using a public dataset and we show that our anomaly detection system increases the accuracy of the detection while reducing the overall resource usage.

Keywords: Anomaly detection · ARIMA · Time series · Forecasting

1 Introduction

In the midst of the recent cloudification, cloud providers remain ill-equipped to cope with security and cloud is thereby highly vulnerable to anomalies and misbehaviours. It hence becomes critical to monitor today's softwarised cloud, looking for unusual states, potential signs of faults or security breaches. Currently, the vast majority of anomaly detectors are based on supervised techniques and thereby require significant human involvement to manually interpret, label the observed data and then train the model. Meanwhile, very few labelled datasets are publicly available for training and the results obtained on a particular controlled cloud (e.g. based on a labelled dataset) do not always translate well to another setting. In the following, we thus introduce an automated and unsupervised solution that detects anomalies occurring in the cloud environment, using statistical techniques. In particular, our anomaly detector relies on a family of statistical models referring to AutoRegressive Integrated Moving Average (ARIMA) and its variants [1], that model and predict the behaviour of the softwarised networking system. This approach consists in building a predictive

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1122–1131, 2023.
https://doi.org/10.1007/978-3-031-27409-1_103

model to provide an explainable anomaly detection. Any observation that is not following the collective trend of the time series is refereed as an anomaly. Still, building a predictive model based on the historical data with the aims of forecasting future values and further detect anomalies, remains a resource-intensive process that entails analysing the cloud behaviour as a whole and typically over a long period of time, on the basis of multiple indicators, such as CPU load, network memory usage, packet loss collected as time series to name a few. It is therefore impractical to study all the historical data, covering all parameters and possible patterns over time, as this approach hardly scales. Furthermore, the performance of such approach tends to deteriorate when the statistical properties of the underlying dataset (a.k.a. cloud behaviour) changes/evolves over time. To tackle this issue, some research studies, e.g., [2], determine a small set of features that accurately capture the cloud behavior so as to provide a light detection. An orthogonal direction of research [3] devises sophisticated metrics (e.g., novel window-based or range-based metrics) that operate over local region. Differently, we propose an adaptive forecasting approach that addresses these issues by leveraging expanding window: once started, an expanding window is made of consecutive observations that grow with time, counting backwards from the most recent observations. The key design rational is to make predictions based on a controlled set of observations, i.e. over several expanding time windows, to capture some complex patterns that may span different time scales and to deal with changes in the cloud behaviour. Overall, our contributions includes:

- an unsupervised anomaly detection system that incorporates a family of autoregressive models (Sect. 3.2) supporting both univariate, seasonal and multivariate time series forecasting. Using several models – in opposition to a single model that is not necessarily the best for any future uses – increases the chance to capture seasonal patterns and complex patterns. The system decomposes the observations (i.e., time series) and attempts to forecast the subsequent behaviour. Then, any deviation from the model-driven forecast is defined as an anomaly that is ultimately reported.
- Our system uses expanding windows and therefore avoids the tendency of the model to deteriorate over time when the statistical properties of the observations change at some points. When a significant behavioural change is observed, a new expanding window is started. This way, the observations depicting this novel behaviour are processed separately. Thus, the resulting forecast better fits and the anomaly detection is robust to behaviour changes.
- Following, we assess the performances associated with our anomaly detector (Sect. 4) considering a cloud native streaming service.

2 Adaptive Anomaly Detection

Auto-regressive algorithms are commonly used to predict the behaviour of a system. As illustration, network operator attempt to predict the future bandwidth/application needs [4] so as to provision in advance sufficient resources. In the same way, we propose to monitor and predict the behaviour of the softwarised

network. Then, anomalies are detected by comparing the expected/predicted behaviour with the actual behaviour observed in the network; the more deviant this behaviour is, the greater the chance that an attack is underway. In practice, the problem is that detection accuracy tends to degrade when there is (even a small) change in behaviour. We thus introduce an anomaly detection system that relies on several auto-regressive models capable of capturing seasonal and correlated patterns for which traditional methods, including the small body of works ([5] leveraging univariate methods, fail on this aspect. In addition, our anomaly detection system uses several expanding windows to deal with a wider range of behavioural patterns that span different time scales and may change over time. From the moment a noticeable change of behaviour is observed, a new windows that runs over an underlying collection is triggered. Our anomaly detector consists in studying past behaviour based on some key indicators (e.g. CPU usage, amount of disk read) that are expressed as a set of K time series $\mathcal{Y} = \{y_t^1\}, \{y_t^2\}, \cdots, \{y_t^K\}$, with $K \geq 1$ and t is the time, with $t \geq T_0$ where T_0 denotes the start time. The behaviour forecasting is performed at equally spaced points in time, denoted $T_0, T_1, \cdots T_i, \cdots$. At time T_i, the resulting forecast model $\hat{\mathcal{M}}_i$ is established accordingly for the next period of time $\Delta T = [T_i, T_{i+1}]$. In particular, we rely on 3 regressive models (as detailed in Sect. 3.1) so as to establish in advance the expected behaviour of the softwarised network and compare them with that observed, at any time $t \geq T_1$. Rather than exploiting the whole historical dataset $\mathcal{Y}$, the analysis is focused on several time windows (i.e. time frames) to achieve some accurate predictions. Time window has the advantage of not having to conveniently deal with the never-ending stream of historical data that are collected. Small window typically accommodates short-term behaviour whilst allowing real-time anomaly detection at low cost. As a complement, larger window covers a wider variety of behaviours and ensures that long term behaviour are considered. Any expanding window W_j (with $1 \leq j \leq J$) is populated with the most recent data points and moves step-wise along the time axis as new observations are received: as time goes, the window grows. Let $\{1, \cdots, w_j\}$ denote the time stamps sequence of observations that are collected during any given time window W_j. This rolling strategy implies that observations are considered for further data analysis as long as they are located in the current window W_j. At time T_i (with $T_i \geq T_1$), all the windowed times series $\{\mathcal{Y}\}_{t=T_i-W_1}^{T_i}, \{\mathcal{Y}\}_{t=T_i-W_2}^{T_i}, \cdots, \{\mathcal{Y}\}_{t=T_i-W_j}^{T_i}, \cdots$ are analysed by a data processing unit that performs the forecasting and produces the predictive model $\hat{\mathcal{M}}^i = \hat{\mathcal{M}}^i(\{\mathcal{Y}\}_{t=T_i-W_1}^{T_i}), \cdots, \hat{\mathcal{M}}^i(\{\mathcal{Y}\}_{t=T_i-W_n}^{T_i})$. For this purpose, a family of predictive models denoted $\hat{\mathcal{M}}^i = \hat{\mathcal{M}}^i_{ARIMA}, \hat{\mathcal{M}}^i_{SARIMA}, \hat{\mathcal{M}}^i_{VARMA}$ is used. Based on $\hat{\mathcal{M}}_i$, the aim is to detect some anomalies $\mathcal{A}_i$, with $\mathcal{A}_i \subset \{\mathcal{Y}\}_{T_i}^{T_{i+1}}$.

3 Anomaly Detection Based on Time Series Forecasting

We introduce an anomaly detection system that continuously detects anomalies and supports time series forecasting, which corresponds to the action of predicting the next values of the time series, leveraging the family of predictive models (Sect. 3.1) and making use of expanding windows (Sect. 3.2) to detect anomalies (Sect. 3.3).

3.1 Time Series Forecasting

Time series forecasting is performed by a general class of extrapolating models based on the frequently used AutoRegressive Integrated Moving Average (ARIMA) whose popularity is mainly due to its ability to represent a time series with simplicity. Advanced variants, including Seasonal ARIMA and Vector ARIMA are further considered to deal with the seasonality in the time series and multidimensional (a.k.a multivariate) time series.

Autoregressive Integrated Moving Average (ARIMA) process for univariate time series combines Auto Regressive (AR) process and Moving Average (MA) process to build a composite model of the time series. During the auto regressive process that periodically takes place at time $T_i = t_O + i\Delta T$ for any expanding windows W_j (with $1 \leq j \leq J$), the variable of interest y_t^k (with $T_i \leq t \leq T_{i+1}$ and $1 \leq k \leq K$) is predicted using a linear combination of past values of the variable $y_{t-1}^k, y_{t-2}^k, \cdots, y_{t-w_j}^k$ that have been collected during w_j:

$$y_t^k = \mu^k + \sum_{i=1}^{w_j} \phi_i^k y_{t-i}^k + \varepsilon_t^k \tag{1}$$

where μ^k is a constant, ϕ_i^k is a model parameter and y_{t-i}^k (with i= 1, $\cdots$, w_i) is a lagged value of y_t^k. ε_t^k is the white noise a time t, i.e., a variable assumed to be independently and identically distributed, with a zero mean and a constant variance. Then, the Moving Average (MA) term y_t^k is expressed based on the past forecast errors:

$$y_t^k = c^k + \sum_{j=1}^{q} \theta_j^k \varepsilon_{t-j}^k + \varepsilon_t^k = \theta(B)\ \varepsilon_t^k \tag{2}$$

where θ_i^k and respectively ε_{t-i}^k (with i= 1, $\cdots$, ΔT) are the model parameter and respectively the random shock at time $t-j$. ε_t^k is the white noise at time t, B stands for backshift operator and $\theta(B) = 1 + \sum_{j=1}^{w_j} \theta_j^k B^j$. Overall, the effective combination of Auto Regressive (AR) and moving average (MA) processes forms a class of time series model, called ARIMA, whose differentiated time series $y't$ is expressed as: $\phi^k(B)(1-B)^d y_t^k = \mu^k + \theta^k(B)$ with $\phi^k(B) = 1 - \sum_{i=1}^{p} \phi_i^k B^i$ and $y_t'^k = (1-B)^d y_t^k$ and d represent the number of differentiation. When seasonality is present in a time series, the Seasonal ARIMA model is of interest.

Seasonal ARIMA (SARIMA) process deals with the effect of seasonality in univariate time series, leveraging the non seasonal component, and also an extra set of parameters P, Q, D, π to account for time series seasonality: P is the order of the seasonal AR term, D the order of the seasonal Integration term, Q the order of the seasonal MA term and π the time span of the seasonal term. Overall, the SARIMA model, denoted SARIMA(p,d,q)(P,D,Q)π, has the following form:

$$\phi_p^k(B)\Phi(B^S)(1-B)^d(1-B^\pi)^D y_t^k = \theta_q(B)\Theta_Q(B^\pi)\varepsilon_t^k \tag{3}$$

where B is the backward shift operator, π is the season length, ε_t^k is the estimated residual at time t and with:

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$
$\Phi(B^\pi) = 1 - \Phi_s B^\pi - \Phi_{2\pi} B^{2\pi} - \cdots - \Phi_{P\pi} B^{P\pi}$
$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$
$\Theta_Q(B^\pi) = 1 + \Theta_s B^\pi + \Theta_{2s} B^{2\pi} + \cdots + \Theta_{Q\pi} B^{Q\pi}$

Vector ARIMA (VARMA) process - Contrary to the (S)ARIMA model, which is fitted for univariate time series, VARMA deals with multiple time series that may influence each other. For each time series, we regress a variable on w_i lags of itself and all the other variables and so on for the q parameter. Given k time series $y_{1t}, y_{2t}, ..., y_{kt}$ expressed as a vector $Vt = [y_{1t}, y_{2t}, ..., y_{kt}]^T$, VARMA(p,q) models is defined by the following Var and Ma models:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \cdot \\ \cdot \\ \cdot \\ y_{Kt} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_k \end{bmatrix} + \begin{bmatrix} \phi^1_{1,1} & \cdots & \phi^1_{1,k} \\ \phi^1_{2,1} & \cdots & \phi^1_{2,k} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \phi^1_{k,1} & \cdots & \phi^1_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \cdot \\ \cdot \\ \cdot \\ y_{k,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} \phi^{w_i}_{1,1} & \cdots & \phi^{w_i}_{1,k} \\ \phi^{w_i}_{2,1} & \cdots & \phi^{w_i}_{2,k} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \phi^{w_i}_{k,1} & \cdots & \phi^{w_i}_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-w_i} \\ y_{2,t-w_i} \\ \cdot \\ \cdot \\ \cdot \\ y_{k,t-w_i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{kt} \end{bmatrix}$$

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \cdot \\ \cdot \\ \cdot \\ y_{Kt} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_k \end{bmatrix} + \begin{bmatrix} \theta^1_{1,1} & \cdots & \theta^1_{1,k} \\ \theta^1_{2,1} & \cdots & \theta^1_{2,k} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \theta^1_{k,1} & \cdots & \theta^1_{k,k} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{k,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} \theta^q_{1,1} & \cdots & \theta^q_{1,k} \\ \theta^q_{2,1} & \cdots & \theta^q_{2,k} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \theta^q_{k,1} & \cdots & \theta^q_{k,k} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-q} \\ \varepsilon_{2,t-q} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{k,t-q} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{kt} \end{bmatrix} \quad (4)$$

where μ_i is a constant vector, the $k \times k$ matrices, denoted $\phi^r_{i,j}$ and respectively $\theta^r_{i,j}$ (with $i, j = 1, \cdots, k$ and $r = 1, \cdots, p$) are the model parameters, the vector $y_{k,t-i}$ (with $i = 1, \cdots, p$) correspond to the lagged values, vector $\varepsilon_{i,t-u}$ (with $i = 1, \cdots, k$ and $u = 1, \cdots, q$) represents random shocks and ε_{it} (with $i = 1, 2, ..., k$) is the white noise vector.

In summary, the proposed anomaly detection system relies on ARIMA, SARIMA and VARIMA that predict the future behaviour on a regular basis, i.e., during the consecutive time periods $[T_1, T_2], \cdots [T_i, T_{i+1}], \cdots$. In particular, the prediction method further utilises several expanding windows to support anomaly detection at different resolutions. At T_i (with $i > 0$), the resulting predictive models $\hat{\mathcal{M}}^i = \hat{\mathcal{M}}^i(\{\mathcal{Y}\}^{T_i}_{t=T_i-W_1}), \cdots, \hat{\mathcal{M}}^i(\{\mathcal{Y}\}^{T_i}_{t=T_i-W_j}), \cdots$. makes a prediction of the behaviour over the next period of time $[T_i, T_{i+1}]$. For each iteration step T_i (with $i \geq 1$), the complexity[1] associated with forecasting the values with ARIMA, SARIMA and VARIMA for all the spanning windows corresponds to:

$$\sum_{j=1}^{J} (w_j + q_j)^2 (1 + K^2) + (w_j + q_j + P_j + Q_j)^2 \quad (5)$$

[1] Complexity can be reduced by distributing and paralleling [6].

As a forecast is performed for each window, this implies that the more windows there are, the more expensive the forecast becomes.

3.2 Expanding Windows

In order to control the forecasting cost associated with handling several expanding windows, the windows management problem then amounts to (i) determine when a new expanding windows needs to be added and (ii) suppress an existing expanding window if needed. The design of the expanding windows management is such that it favours the forecasting with expanding windows that produce the fewest forecast errors while privileging the less computationally demanding ones in the case of an error tie. A novel expanding window starts if an existing expanding windows provides erroneous predictions (i.e. the prediction error is greater than a given threshold). If required (i.e. the number of windows is too large and reaches the desired limit), this addition leads to the deletion of another window.

3.3 Threshold-Based Anomaly Detection

The anomaly detection process is periodically triggered at time T_i (with $T_i \geq T_1$) considering the three predictive models $\hat{\mathcal{M}}^i = \hat{\mathcal{M}}^i_{ARIMA}, \hat{\mathcal{M}}^i_{SARIMA}, \hat{\mathcal{M}}^i_{VARMA}$. In particular, a subset of values $\mathcal{A}_i \subset y_{t_{t=T_i}}^{T_{i+1}}$ is defined as anomalous if there exists a noticeable difference between the observed value y_t^k and one of the forecast values at time t in $\hat{\mathcal{M}}^i = \hat{\mathcal{M}}^i(\{\mathcal{Y}\}_{t=T_i-W_1}^{T_i}), \cdots, \hat{\mathcal{M}}^i(\{\mathcal{Y}\}_{t=T_i-W_J}^{T_i})$. In one of the given models, a noticeable difference between the observation value y_t^k and the forecasted values $\hat{y^k}_t$ at time t (with $T_i \leq t \leq T_{i+1}$) is greater than a threshold. The threshold is calculated using to the so-called three sigma rule [7], which is a simple and widely used heuristic that detects outlier [8]. Other metrics such as the one indicated in [3] could be easily exploited. Based on all the prediction errors $\{\epsilon_t^k\}_{t=T_i-w_i}^{T_i}$ observed during $[T_i - w_i, T_i]$ where $\epsilon_t^k = |y_t^k - \hat{y}_t^k|$, the threshold is defined as:

$$\delta_{w_i}(T_i) = \alpha\ \sigma(\epsilon_t^k) + \mu(\epsilon_t^k) \tag{6}$$

α is a coefficient that can be parameterised based on the rate of false positive/negative observed/expected and $\sigma(\epsilon_t^k)$ and resp. $\mu(\epsilon_t^k)$ correspond to the standard deviation and resp. mean of the prediction error.

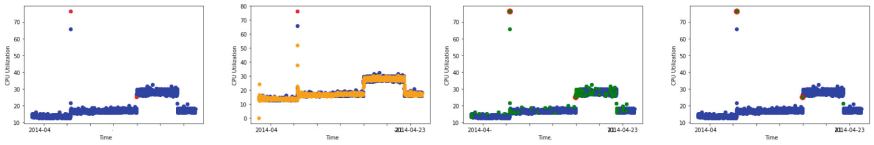
4 Assessment

Our solution supports the forecasting along with anomaly detection, provided relevant measurements (a.k.a time series). The proposed solution is evaluated relying on a public dataset, which contains data provided by a monitored cloud-native streaming service. The Numenta Anomaly Benchmark (NAB) dataset[2]

[2] https://www.kaggle.com/boltzmannbrain/nab.

corresponds to a labelled dataset, i.e. the dataset contains anomalies for which the causes are known. The dataset depicts the operation of some streaming and online applications running on Amazone cloud. The dataset reported by the Amazon Cloud watch service includes various metrics, e.g., CPU usage, incoming communication (Bytes), amount of disk read (Bytes), ect. Our prototype implementation is focused on the preprocessing of the monitored data, forecasting and detection of anomalies. The prototype requires a Python environment as well panda[3], a third-party packages handling times series and data analytic. Our detector proceeds as follows. Filtered and converted into an appropriate format. Then, measurements are properly scaled using Min-Max normalisation [9] of the features. As suggested by Box and Jenkins, the ARIMA model along with their respective (hyper)parameters are established. Finally, anomalies are detected.
Relying on the dataset and our prototype, we evaluate the performances associated with the proposed anomaly detector. We consider two time frames lasting 23 days (Figs. 1 and 3) and one month (Fig. 2) during which labelled anomalies (red points) are detected (green points in Figs. 1c, 1d, 2c, 2d, 3c and 3d) or not. As expected, forecast values (orange points in Figs. 1b and 2b) are conveniently close to the normal observations (blue points). In both cases, anomalies are not always distant from both the normal values (blue points), which makes anomaly detection challenging even if in both cases they are adequately detected. With a dynamic threshold (Figs. 1d and 2d), the number of false positives (green points not circled in red in Fig. 1d and 1c) is negligible comparing to a static threshold (Fig. 1c and 2c) that involves a very high false positive rate.

When we focus on a multivariate prediction and detection (Fig. 3), we see that the parameterization of the threshold plays a significant role in the detection accuracy and in the rate of false positives and false negatives. Comparing to a static threshold, a dynamic threshold constitutes a fair compromise between a accurate detection and an acceptable false positive rate.



(a) normal observations (blue), known anomalies (red)

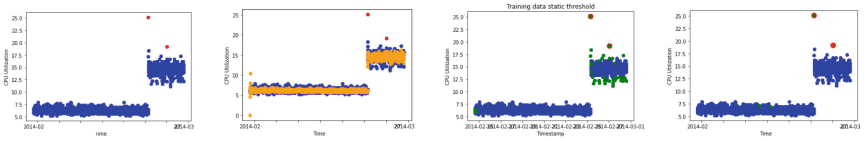
(b) forecast values (orange), known anomalies (red)

(c) detected anomalies (green) with static threshold

(d) detected anomalies (green) with dynamic threshold

Fig. 1. Observations versus forecast measurement - CPU utilisation of cloud native streaming service during 23 days.

[3] https://pandas.pydata.org.



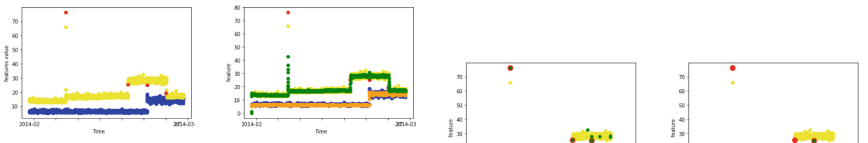
(a) Observations (blue), known anomalies (red)

(b) Forecasted values (orange), known anomalies (red)

(c) Detected anomalies (green) with static threshold

(d) Detected anomalies (green) with dynamic threshold

Fig. 2. Observations versus forecast values - CPU utilisation of cloud native streaming service during 1 month.



(a) 2 time series with normal observations (blue, yellow) and anomalies (red)

(b) Observations (blue and yellow) versus forecast observations (orange and green)

(c) Anomaly detection (green) with static threshold

(d) Anomaly detection (green) with dynamic threshold

Fig. 3. Multivariate Forecast

5 Related Work

Anomaly detection is an long-standing research area that has continuously attracted the attention of the research community in various fields. The resulting research on anomaly detection is primarily distinguished by the type of data processed for anomaly detection and the algorithms applied to the data to perform the detection. The majority of the works deals with temporal data, i.e., typically discrete sequential data, which are univariate rather than multivariate: in practice, several time series (concerning e.g. CPU usage, memory usage, traffic load, etc.) are considered and processed individually. Based on each time series, traditional approaches [10] typically apply supervised or unsupervised techniques to solve the classification problem related to anomaly detection. They construct a model using (un)supervised algorithms, e.g., random forests, Support Vector Machine (SVM), Recurrent Neural Networks (RNNs) and its variants including Long Short-Term Memory (LSTMs) [11–13] and deep neural network (DNN).

Recently, another line of research that have been helpful in several domains, is to analyse time-series to predict their respective behaviour. Then, an anomaly is detected by comparing the predicted time series and the observed ones. To model non-linear time series, Recurrent Neural Networks (RNNs) and some variants, e.g. Gated Recurrent Units (GRUs), Long Short-Term Memory (LSTMs) [14] have been studied. Filinov et al. [14] use a LSTM model to forecast values and detect anomalies with a threshold applied on the MSE. Candielieri [15]

combines a clustering approach and support vector regression to forecast and detect anomalies. The forecasted data are clustered; then anomaly is detected using Mean Absolute Percentage Error. In [5], Vector Auto Regression (VAR) is combined with RNNs to handle linear and non-linear problems with aviation and climate datasets. In addition, a hybrid methodology called MTAD-GAT [16] uses forecasting and reconstruction methods in a shared model. The anomaly detection is done by means of a Graph Attention Network. The works mentionned above rely on RNNs that are non-linear models capable of modelling long-term dependencies without the need to explicitly specify the exact lag/order. In counterpart, they may involve a significant learning curve for large and complex models. Furthermore, they are difficult to train well and may suffer from local minima problems [17] even after carefully tuning the backpropagation algorithm. The second issue is that RNNs might actually produce worse results than linear models if the data has a significant linear component [31]. Alternatively, autoregressive models, e.g. ARIMA, Vector Autoregression (VAR) [5] and latent state based models like Kalman Filters(KF) have been studied. Time series forecasting problems addressed in the literature, however, are often conceptually simpler than many tasks already solved by LSTM.

For multivariate time series, anomaly is detected by comparing the predicted time series and the observed ones. To predict the future values, several algorithms are employed. Filinov et al. [14] use a LSTM-based model to forecast values and detect anomalies with a threshold on the MSE. Candielieri [15] combines clustering and support vector regression to forecast and detect anomalies: forecasted values are mapped into clusters and anomalies are detected using Mean Absolute Percentage Error. R2N2 [5] combines the traditional Vector Auto Regression (VAR) and RNNs to deal with both linear and non linear problems in the aviation and climate datasets. An hybrid methodology [16] uses forecasting and reconstruction methods in a shared model while anomaly detection is done with Graph Attention Network.

6 Conclusion

Anomaly detection plays a crucial role on account of its ability to detect any inappropriate behaviour so as to protect every device in a cloud including equipment, hardware and software, by forming a digital perimeter that partially or fully guards a cloud. In this article, we have approached the problem of anomaly detection and introduced an unsupervised anomaly detection system that leverages a family of statistical models to predict the behaviour of the softwarised networking system and identify deviations from normal behaviour based on past observations. Existing solutions mostly exploit the whole set of historical data for model training so as to cover all possible patterns spanning time. Nonetheless, such a detection approach may not scale and performance of these models tend to deteriorate as the statistical properties of the underlying data change across time. We address this challenge through the use of expanding windows with the aim of making predictions based on a controlled set of observations. In

particular, several expanding time windows capture some complex patterns that may span different time scales (e.g. long term versus short terms patterns), and, deal with changes in the cloud behaviour. Following, we have implemented and experimented our solution. Our prototype contributes to enhancing the accuracy of the detection at a small computational cost.

References

1. Box, G.E., Reinsel, G.M.J.R.: Time Series Analysis: Forecasting and Control. Wiley (2015)
2. Hammi, B., Doyen, G., Khatoun, R.: Toward a source detection of botclouds: a PCA-based approach. In: IFIP International Conference on Autonomous Infrastructure, Management and Security (AIMS) (2014)
3. Huet, A., Navarro, J.-M., Rossi, D.: Local evaluation of time series anomaly detection algorithms. In: Conference on Knowledge Discovery & Data mining (2022)
4. Yoo, W., Sim, A.: Time-series forecast modeling on high-bandwidth network measurements. J. Grid Comput. **14** (2016)
5. Goel, H., Melnyk, I., Banerjee, A.: R2n2: residual recurrent neural networks for multivariate time series forecasting (2017). https://arxiv.org/abs/1709.03159
6. Wanga, X., Kanga, Y., Hyndmanb, R., et al.: Distributed Arima models for ultra-long time series. Int. J. Forecast. (June 2022)
7. Pukelsheim, F.: The three sigma rule. Am. Stat. **48** (1994)
8. Rüdiger, L.: 3sigma-rule for outlier detection from the viewpoint of geodetic adjustment. J. Surv. Eng., 157–165 (2013)
9. Zheng, A., Casari, A.: Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc. (2018)
10. Gümüsbas, D., Yildirim, T., Genovese, A., Scotti, F.: A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. IEEE Syst. J., 15(2) (2020)
11. Kim, T., Cho, S.: Web traffic anomaly detection using C-LSTM neural networks. Exp. Syst. Appl. **106**, 66–76 (2018)
12. Su, Y., Zhao, Y., Niu, C., et al.: Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: 25th ACM International Conference on Knowledge Discovery & Data Mining, pp. 2828–2837 (2019)
13. Diamanti, A., Vilchez, J., Secci, S.: LSTM-based radiography for anomaly detection in softwarized infrastructures. In: International Teletraffic Congress (2020)
14. Filonov, P., Lavrentyev, A., Vorontsov, A.: Multivariate industrial time series with cyber-attack simulation: fault detection using an LSTM-based predictive data model (2016). https://arxiv.org/abs/1612.06676
15. Candelieri, A.: Clustering and support vector regression for water demand forecasting and anomaly detection. Water **9**(3) (2017)
16. Zhao, H., Wang, Y., Duan, J., et al.: Multivariate time-series anomaly detection via graph attention network. In: International Conference on Data Mining (2020)
17. Uddin, M.Y.S., Benson, A., Wang, G., et al.: The scale2 multi-network architecture for IoT-based resilient communities. In: IEEE SMARTCOMP (2016)



Quantum-Defended Lattice-Based Anonymous Mutual Authentication and Key-Exchange Scheme for the Smart-Grid System

Hema Shekhawat(✉) and Daya Sagar Gupta

Rajiv Gandhi Institute of Petroleum Technology, Amethi, UP, India
hemashekhawatrgipt@gmail.com

Abstract. The Smart-grid (SG) systems are capable of empowering information sharing between smart-metres (SMs) and service providers via Internet protocol. This may lead to various security and privacy concerns because the communication happens in the open wireless channel. Several cryptographic schemes have been designed to help secure communications between SMs and neighbourhood-area-network gateways (NAN-GWs) in the SG systems. Prior works, on the other hand, do not maintain conditional identity anonymity, and compliant key management for SMs and NAN-GWs in general. Therefore, we introduce a quantum-defended lattice-based anonymous mutual authentication and key-exchange (MAKE) protocol for SG systems. Specifically, the proposed scheme can allow robust conditional identity anonymity and key management by exploiting small integer solutions and inhomogeneous small integer solutions lattice hard assumptions, eliminating the demand for other complicated cryptographic primitives. The security analysis demonstrates that the scheme offers an adequate security assurance against various existing as well as quantum attacks and has the potential to be used in the SG implementation.

Keywords: Lattice-based cryptography · mutual authentication (MA) · smart-grid systems · post-quantum cryptosystems · key-exchange scheme

1 Introduction

The SG system is a new bidirectional electricity network that employs digital communication and control technologies [14]. It offers energy reliability, self-healing, high fidelity, energy security, and power-flow control. Recently, the power industry has been merging the power distribution system with information and communication technology (ICT). The traditional unidirectional power-grid system only transmits energy by adjusting voltage degrees, and it is incapable of meeting the growing demand for renewable energy generation origins like tide,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1132–1142, 2023.
https://doi.org/10.1007/978-3-031-27409-1_104

wind, and solar-energy. Smart metres (SMs) are the most critical components of SG systems. SM collects data on consumer consumption and transmits it to a control centre (CC) via neighbourhood-area network gateways (NAN-GWs). Because of its bidirectional communication link, the CC may create a realistic power-supply methods for SG to optimize the electricity consumption in high-peak and off-peak intervals. However, the security and privacy of communicating users in SG systems continue to be critical issues that must be addressed before using the SG for a variety of purposes. The inherent characteristics of the SG systems, like mobility, location awareness, heterogeneity, and geo-distribution, may be misused by attackers. Therefore, mutual authentication (MA) is an efficient approach to guarantee trust as well as secure connections by validating the identities of connected components without transferring critical information over an open wireless channel.

Because of their resource-constrained SMs and other internet-of-things (IoT) devices, traditional public-key cryptosystems are unquestionably incompatible with SG systems. P. Shor in [10] showed that several traditional cryptographic schemes, resulting in public-key cryptographic advances, are vulnerable to the reality of quantum technology. In its application to traditional and emerging security interests like encryption, digital signatures, and key-exchange [9], lattice-based cryptosystems offer a promising post-quantum method. Due to their intrinsic traits, lattice-based cryptosystems deliver improved security characteristics against quantum cryptanalysis while being easily implementable. Besides, the dependence on the registration authority (RA) to issue key-pairs continually and for new devices suffers from high communications costs and asynchronous problems. Hence, to achieve properties efficiently and effectively, such as MA, key exchange, key update, and conditional anonymity, we designed a lattice-based anonymous MAKE scheme for the SG systems. The scheme also provides conditional traceability and linkability.

Most of the authentication protocols so far can be implemented using classical cryptographic methods based on integer factorization and discrete logarithmic assumptions, which are prone to evolving quantum threats. It is also concluded that the classical schemes are incompatible with resource-constrained (battery-power, processing, memory, or computation) SMs. Therefore, the paper presents a lattice-based anonymous MAKE scheme for the SG. The proposed scheme utilises lattice-based computations, which are resilient to quantum attacks and inexpensive to implement. It allows efficient conditional anonymity to protect SMs' privacy and key management by exploiting small integer solutions and inhomogeneous small integer solutions hard assumptions in the lattice.

The body of the article is laid out in the following. Section 2 presents some MA related research articles for the SG systems. Section 3 explains the preliminary information that will be applied throughout the proposed work along with the system model. Section 4 explains the proposed work, while Sect. 5 provides a security analysis that satisfies the security standards. Finally, in Sect. 6, the paper concluded with certain remarks.

2 Related Work

The authenticated key-agreement schemes have recently gained popularity, with a focus on reliable and secure communications in SG systems. In recent years, numerous authentication solutions for SG systems have been presented [2–4,6–8,11–13,15]. In [3], the authors introduced a lightweight message authentication protocol for SG communications using the Diffie-Hellman key-agreement protocol. However, they conduct simulations to show the effectiveness of their work in terms of few signal message conversations and low latency. The authors of [2] introduce LaCSys, a lattice-based cryptosystem for secure communication in SG environments, to address the privacy and security issues of software-defined networking (SDN), considering the quantum computing era. Despite the complexity and latency sensitivity of SG, the work in [8] proposes a lightweight elliptic curve cryptography (ECC)-based authentication mechanism. The proposed methodology not only enables MA with reduced computing and communication costs, but it also withstands all existing security attacks. Unfortunately, the work in [7] demonstrated that, using the work in [8], the attacker can impersonate the user using an ephemeral secret leaking attack. In [4], a safe and lightweight authentication strategy for resource-constrained SMs is provided, with minimum energy, communication, and computing overheads. However, rigorous performance evaluation verifies the proposed protocol's efficiency over the state-of-the-art in providing enhanced security with minimum communicational and computing overheads. In traditional cloud-based SG systems, achieving low latency and offering real-time services are two of the major obstacles. Therefore, there has been an increasing trend toward switching to edge computing. The authors of [11] designed a blockchain-based MAKE strategy for edge-computing-based SG systems, that provides efficient conditional anonymity and key-management. In [15], a decentralised safe keyless signature protocol based on a consortium consensus approach that transforms a blockchain into an autonomous access-control manager without the use of a trusted third party. The authors of [12] introduce BlockSLAP, which uses cutting-edge blockchain technology and smart contracts to decentralise the RA and reduce the interaction process to two stages. The authors of [13] primarily address several identity authentication concerns that have persisted in the SG. Therefore, a trustworthy and efficient authentication approach for SMs and utility centres is presented using blockchain, ECC, a dynamic Join-and-Exit mechanism, and batch verification. Furthermore, the schemes in [11–13] demonstrate that their work is safe under both computational hard assumptions and informal security analysis. In [6], the authors proposed a SG MAKE scheme based on lattices that enables secure communication between the service provider and the SMs. They claimed their work was resilient to quantum attacks. In contrast to other schemes studied in the literature, we designed a lattice-based anonymous MA protocol which not only provides conditional identity anonymity and key-management but also withstands quantum attacks with easy implementation.

3 Preliminaries

In this section, we summarise lattice-based cryptography and its hard problems, as well as the proposed work's system model.

3.1 Lattice-Based Cryptography

Lattice-based cryptography is a promising tool to construct very strong security algorithms for the post-quantum era. The security proofs of lattice-based cryptography are centred on worst-case hardness, comparatively inexpensive implementations, and reasonable simplicity. Therefore, lattices are considered a robust mathematical structure to strengthen cryptographic schemes against quantum attacks [1,5]. A lattice is an m-dimensional set of points with a regular structure defined as follows.

Definition 1 The lattice is the set of vectors formed by n-linearly independent vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_n \in \mathbb{R}^m$, represented as $\mathbf{L}$ can be defined as:

$$\mathbf{L}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_n) = \left\{ \sum_{k=1}^{n} d_k \boldsymbol{y}_k : d_k \in \mathbb{Z}^+ \right\} \tag{1}$$

The basis vectors are the n-linearly independent vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_n$. The integer n and m are rank and dimension of $\mathbf{L}$, respectively.

The shortest minimum distance of $\mathbf{L}$, is the shortest non-zero vector in $\mathbf{L}$ can be computed by the given formula.

$$\mathcal{D}(\mathbf{L}) = \min_{\boldsymbol{y} \in \mathbf{L}/\{0\}} ||\boldsymbol{y}|| \tag{2}$$

Definition 2 Assume that the basis of a $\mathbf{L}$ is defined as $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_n\} \in \mathbb{Z}^{m \times n}$, where columns of $\boldsymbol{Y}$ are the basis vectors. The expression $\mathbf{L}(\boldsymbol{Y}) = \{\boldsymbol{Y}\boldsymbol{d} : \boldsymbol{d} \in \mathbb{Z}^n\}$ is defined as a lattice in m-dimensional Euclidean space $\mathbb{R}^m$, where $\boldsymbol{Y}\boldsymbol{d}$ is a matrix-vector multiplication.

Definition 3 (*Shortest vector problem (SVP)*): Given a basis $\boldsymbol{Y} \in \mathbb{Z}^{m \times n}$ of a lattice $\mathbf{L}$, finding a non-zero vector $d \in \mathbf{L}$ is computationally hard whose Euclidean norm such as $||\boldsymbol{d}|| = \mathcal{D}(\mathbf{L})$ is minimum.

Definition 4 (*Closest vector problem (CVP)*): Given a basis $\boldsymbol{Y} \in \mathbb{Z}^{m \times n}$ of a lattice $\mathbf{L}$ and a vector $\boldsymbol{e} \notin \mathbf{L}$, finding a non-zero vector $d \in \mathbf{L}$ is computationally hard whose Euclidean norm such as $||\boldsymbol{t} - \boldsymbol{d}|| = \mathcal{D}(\mathbf{L})$ is minimum.

***q*-ary Lattice** The q-ary lattice ($\mathbb{Z}_q^n \subseteq \mathbf{L} \subseteq \mathbb{Z}^n$) supports modular arithmetic for an integer q. The proposed scheme utilizes the notion of q-ary lattices hard assumptions which are defined as follows.

Definition 5 (*Small integer solution (SIS)*): Given an integral modular matrix $\boldsymbol{Y} \in \mathbb{Z}^{n \times m}$ and an constant $\beta \in \mathbb{Z}^+$, finding a non-zero vector $\boldsymbol{d} \in \mathbf{L}$ is computationally hard, where $||\boldsymbol{d}|| < \beta$ and $\{\boldsymbol{Y}\boldsymbol{t} = 0 \bmod q\}$.

Definition 6 (*Inhomogeneous small integer solution (ISIS)*): Given an integral modular matrix $\boldsymbol{Y} \in \mathbb{Z}^{n \times m}$, an constant $\beta \in \mathbb{Z}^{+}$ and a vector $\boldsymbol{e} \in \mathbb{Z}_q^m$, finding a non-zero vector $\boldsymbol{d} \in \mathbf{L}$ is computationally hard, where $||\boldsymbol{d}|| < \beta$ and $\{\boldsymbol{Yt} = \boldsymbol{e} \bmod q\}$.

3.2 System Model

Here, we consider the system model and network key assumptions related to the network model of the SG metering infrastructure.

Smart-Grid (SG) Network Model The SG metering infrastructure comprises of registration authority (RA), smart-meter (SM), and NAN-gateways (NAN-GW). The RA is a trusted service provider in the SG system. The role of RA is to distribute public key parameters for each SM and NAN-GW in the SG system. In the network model, each SM can connect with their nearby NAN-GW. The SM and NAN-GW register in the RA. After verifying the authenticity of the keys released by RA, SM uses the keys to pass NAN-GW's authentication. Finally, for the forthcoming communication, SM can use the exchanged session key to interact with NAN-GW. Similar steps are also performed by NAN-GW to pass SM's authentication and key-exchange.

Network Assumptions The network key assumptions for the proposed work are illustrated as follows.

1. The public keys and hashed identities of SMs are known as NAN-GWs. Hence, in the SG system, NAN-GWs function as the relay-nodes which serve timely service, thus it is not required to preserve identity anonymity for NAN-GWs.
2. NAN-GW key resources should not be repeatedly revoked or updated unless they are alleged to be compromised. If NAN-GW is suspected, RA will suspend the server, reject every service request from SMs, and close the connections.
3. Some sensitive data can only be retrieved by the authorised user.

4 Proposed Work

The stages of the proposed scheme comprise of *system setup*, *registration*, and *MAKE*.

4.1 System Setup

RA inputs 1^t for the security parameter t and executes system setup stage for system deployment, which are explained in the following.

1. RA chooses a prime modulo q, an integer m, and a square modular matrix $\boldsymbol{M} \in \mathbb{Z}_q^{m \times m}$.

2. RA takes five one-way hash functions such as $H_1 : \{0,1\}^* \rightarrow \{0,1\}^t$, $H_2 : \{0,1\}^t \times \mathbb{Z}_q^m \rightarrow \{0,1\}^t$, $\boldsymbol{H}_3 : \mathbb{Z}_q^m \times \mathbb{Z}_q^* \rightarrow \mathbb{Z}_q^m$, $H_4 : \mathbb{Z}_q^m \times \mathbb{Z}_q^m \times \mathbb{Z}_q^m \times \{0,1\}^* \rightarrow \{0,1\}^t$, $H_5 : \mathbb{Z}_q^m \times \{0,1\}^t \times \mathbb{Z}_q^* \times \mathbb{Z}_q^* \rightarrow \{0,1\}^t$, and $H_6 : \mathbb{Z}_q^* \times \mathbb{Z}_q^* \times \mathbb{Z}_q^* \times \{0,1\}^* \rightarrow \{0,1\}^t$.
3. RA selects random vector $\boldsymbol{s} \in \mathbb{Z}_q^{m \times 1}$ as its master private key.
4. For SM, RA computes corresponding master public key $\boldsymbol{P}_s = \boldsymbol{M} \cdot \boldsymbol{s}$.
5. Similarly, for NAN-GW, RA computes corresponding master public key $\boldsymbol{P}_n = \boldsymbol{s}^T \cdot \boldsymbol{M}$.

The RA conceals $\boldsymbol{s}$ while publishing public system parameters:

$$\Delta = (\boldsymbol{M}, m, q, \boldsymbol{P}_s, \boldsymbol{P}_n, H_1, H_2, \boldsymbol{H}_3, H_4, H_5, H_6).$$

4.2 Registration

The RA, SMs, and NAN-GWs interactively execute the registration stage. Here, the registration steps of SM and NAN-GW are provided as follows. The RA communicates with both SM and NAN-GW in secure and private channels. Initially, SM and NAN-GW send their hashed identities to RA. Then, RA generates the communication keys and sends them back to them. The registration steps are illustrated in the following.

1. SM computes $HID_s = H_1(ID_s)$ (or NAN-GW computes $HID_n = H_1(ID_n)$) and sends them to the RA via secure channel.
2. The RA receives registration requests from SM (or NAN-GW), then verifies whether the SM (or NAN-GW) is registered. If so, it will terminate the registration request. Otherwise, it will calculate SM/NAN-GW's key-pair.
   (a) For SM, RA selects a random vector $\boldsymbol{r}_s \in \mathbb{Z}_q^{m \times 1}$ and computes $\boldsymbol{R}_s = \boldsymbol{M} \cdot \boldsymbol{r}_s$, $\boldsymbol{x}_s = \boldsymbol{r}_s + \boldsymbol{s} \cdot H_2(HID_s, \boldsymbol{R}_s)$ and $\boldsymbol{Pk}_s = \boldsymbol{M} \cdot \boldsymbol{x}_s$.
   (b) RA sends the secret message tuple $(\boldsymbol{x}_s, \boldsymbol{Pk}_s)$ to SM.
   (c) Similarly, for NAN-GW, RA selects a random vector $\boldsymbol{r}_n \in \mathbb{Z}_q^{m \times 1}$ and computes $\boldsymbol{R}_n = \boldsymbol{r}_n^T \cdot \boldsymbol{M}$, $\boldsymbol{x}_n = \boldsymbol{r}_n + \boldsymbol{s} \cdot H_2(HID_n, \boldsymbol{R}_n)$, and $\boldsymbol{Pk}_n = \boldsymbol{x}_n^T \cdot \boldsymbol{M}$.
   (d) RA sends the secret message tuple $(\boldsymbol{x}_n, \boldsymbol{Pk}_n)$ to NAN-GW.
3. The following processes are used by SM to validate the received key-pair.
   (a) SM examines the validity of $\boldsymbol{M} \cdot \boldsymbol{x}_s = \boldsymbol{R}_s + \boldsymbol{P}_s \cdot H_2(HID_s, \boldsymbol{R}_s) = \boldsymbol{Pk}_s$.
   (b) If the above verification is successful, then SM securely stores the secret key $\boldsymbol{x}_s$. SM requests RA to start the registration process again.
4. Similarly, NAN-GW also verify the validity of the received key-pair by following steps.
   (a) NAN-GW examines the validity of $\boldsymbol{x}_n^T \cdot \boldsymbol{M} = \boldsymbol{R}_n + \boldsymbol{P}_n \cdot H_2(HID_n, \boldsymbol{R}_n) = \boldsymbol{Pk}_n$.
   (b) If the above verification is successful, then NAN-GW securely stores the secret key $\boldsymbol{x}_n$. NAN-GW requests RA to start the registration process again.

4.3 Mutual Authentication and Key-Exchange (MAKE)

The registered SMs and NAN-GWs can only execute the MAKE stage, as illustrated in the following.

1. SM → NAN-GW: $M_1 = (\boldsymbol{A}, \boldsymbol{pid}_s, \boldsymbol{k}_s, t_s)$
   (a) SM selects a random vector $\boldsymbol{a} \in \mathbb{Z}_q^{1\times m}$, and computes $\boldsymbol{A} = \boldsymbol{M} \cdot \boldsymbol{a}^T$, $\boldsymbol{pid}_s = \boldsymbol{Pk}_s + \boldsymbol{H}_3(\boldsymbol{A}, \boldsymbol{Pk}_n \cdot \boldsymbol{a}^T)$.
   (b) SM computes $\boldsymbol{k}_s = \boldsymbol{a}^T + \boldsymbol{x}_s \cdot H_4(\boldsymbol{Pk}_s, \boldsymbol{pid}_s, \boldsymbol{A}, t_s)$, where t_s is the prevailing timestamp.
   (c) SM sends $(\boldsymbol{A}, \boldsymbol{pid}_s, \boldsymbol{k}_s, t_s)$ to NAN-GW.
2. NAN-GW → SM: $M_2 = (\boldsymbol{B}, \boldsymbol{pid}_n, \boldsymbol{k}_n, t_n)$ and $M_3 = (w)$
   (a) NAN-GW obtains SMs' public key by computing $\boldsymbol{Pk}_s^{'} = \boldsymbol{pid}_s - \boldsymbol{H}_3(\boldsymbol{A}, \boldsymbol{x}_n^T \cdot \boldsymbol{A})$ if $|t_s - t_s^{'}| \leqslant \bigtriangledown t$, where $t_s^{'}$ is the current timestamp and $\bigtriangledown t$ is a agreed threshold value.
   (b) NAN-GW verifies the SM, if $\boldsymbol{M} \cdot \boldsymbol{k}_s = \boldsymbol{A} + \boldsymbol{Pk}_s{}' \cdot H_4(\boldsymbol{Pk}_s^{'}, \boldsymbol{pid}_s, \boldsymbol{A}, t_s)$ holds.
   (c) NAN-GW picks a random vector $\boldsymbol{b}$ and computes $\boldsymbol{B} = \boldsymbol{b} \cdot \boldsymbol{M}$ and $\boldsymbol{pid}_s = \boldsymbol{Pk}_n + \boldsymbol{H}_3(\boldsymbol{B}, \boldsymbol{b} \cdot \boldsymbol{Pk}_s)$.
   (d) NAN-GW computes $\boldsymbol{k}_n = \boldsymbol{b} + \boldsymbol{x}_n^T \cdot H_4(\boldsymbol{Pk}_n, \boldsymbol{pid}_n, \boldsymbol{B}, t_n)$, where t_n is the current timestamp.
   (e) NAN-GW sends $(\boldsymbol{B}, \boldsymbol{pid}_n, \boldsymbol{k}_n, t_n)$ to SM.
   (f) It computes $K_1 = \boldsymbol{x}_n^T \cdot \boldsymbol{A} + \boldsymbol{b} \cdot \boldsymbol{Pk}_s$ and $K_2 = \boldsymbol{b} \cdot \boldsymbol{A}$, if the above verification satisfied. Then, it computes $Sk_{ns} = H_5(\boldsymbol{Pk}_s^{'}, HID_n, K_1, K_2)$ and $w = H_6(Sk_{ns}||K_1||K_2||t_n)$.
   (g) NAN-GW sends (w) to SM.
3. SM: Sk_{sn}
   (a) SM obtains NAN-GWs' public key by computing $\boldsymbol{Pk}_n^{'} = \boldsymbol{pid}_n - \boldsymbol{H}_3(\boldsymbol{B}, \boldsymbol{B}.\boldsymbol{x}_s)$, if $|t_n - t_n^{'}| \leqslant \bigtriangledown t$, where $t_n^{'}$ is the current timestamp.
   (b) SM verifies NAN-GW, if $\boldsymbol{k}_n \cdot \boldsymbol{M} = \boldsymbol{B} + \boldsymbol{Pk}_n^{'} \cdot H_4(\boldsymbol{Pk}_n^{'}, \boldsymbol{pid}_n, \boldsymbol{B}, t_n)$ holds.
   (c) It computes $K_3 = \boldsymbol{Pk}_n \cdot \boldsymbol{a}^T + \boldsymbol{B} \cdot \boldsymbol{x}_s$ and $K_4 = \boldsymbol{B} \cdot \boldsymbol{a}^T$, if the above verification satisfied.
   (d) Then, SM computes the session key $Sk_{sn} = H_5(\boldsymbol{Pk}_s, HID_n, K_3, K_4)$, if $H_6(Sk_{sn}||K_3||K_4||t_n) = w$ holds.

Correctness of the Protocol From the given equations, we can verify the correctness of the proposed scheme.

$$K_1 = \boldsymbol{x}_n^T \cdot \boldsymbol{A} + \boldsymbol{b} \cdot \boldsymbol{Pk}_s = \boldsymbol{x}_n^T \cdot \boldsymbol{M} \cdot \boldsymbol{a}^T + \boldsymbol{b} \cdot \boldsymbol{M} \cdot \boldsymbol{x}_s = \boldsymbol{Pk}_n \cdot \boldsymbol{a}^T + \boldsymbol{B} \cdot \boldsymbol{x}_s = K_3 \quad (3)$$

$$K_2 = \boldsymbol{b} \cdot \boldsymbol{A} = \boldsymbol{b} \cdot \boldsymbol{M} \cdot \boldsymbol{a}^T = \boldsymbol{B} \cdot \boldsymbol{a}^T = K_4 \quad (4)$$

Thus, we obtain $Sk_{sn} = Sk_{ns}$. Both SM and NAN-GW establish a secure and common session key between them.

5 Security Analysis

Here, we explain how the proposed method meets the security standards outlined below.

1. **Mutual authentication (MA):** The registered SM and NAN-GW are only authorized to use the designed scheme for verifying the communicator's identity preceding to message interchange in the SG system. There are two possible cases where our scheme proves its relevance.
   (a) Firstly, for SM $\rightarrow$ NAN-GW authentication, assume that an intruder can fabricate real message, then we have $\boldsymbol{M} \cdot \boldsymbol{k}_s = \boldsymbol{A} + \boldsymbol{Pk}_s{}' \cdot H_4(\boldsymbol{Pk}_s^{'}, \boldsymbol{pid}_s, \boldsymbol{A}, t_s)$. The attacker attempts to repeat the operation with the identical input randomness but will obtain mismatched hashed values. Hence, a legitimate authentication message cannot be forged by any attacker.
   (b) Secondly, suppose that an intruder outputs a valid messages $M_3 = (w)$ and $M_2 = (\boldsymbol{B}, \boldsymbol{pid}_n, \boldsymbol{k}_n, t_n)$ to run the verification of the SM, then there will be a solution to break SIS/ISIS problems. If $w = H_6(Sk_{ns}||K_1||K_2||t_n)$ is a valid authentication message, then attacker has to compute $K_3^* = K_3 = \boldsymbol{Pk}_n \cdot \boldsymbol{a}^T + \boldsymbol{B} \cdot \boldsymbol{x}_s$. Then it can obtain valid authentication message. Hence, a legitimate authentication message cannot be forged by any attacker.
2. **Session key exchange:** During execution of the designed scheme, a session key should be produced for further confidential message exchange between SM and NAN-GW. Even RA is clueless about the session key. An intruder must have the values of $K_1 = \boldsymbol{x}_n^T \cdot \boldsymbol{A} + \boldsymbol{b} \cdot \boldsymbol{Pk}_s$ and $K_2 = \boldsymbol{b} \cdot \boldsymbol{A}$ to calculate session key, even if the intruder knows the public key $\boldsymbol{Pk}_s$ and hashed identity HID_n. While the private keys and random vectors ($\boldsymbol{a}$ and $\boldsymbol{b}$) are not sent on the open wireless channel. The intruder can gain a authentic session-key only if the SIS and ISIS assumptions are violated.
3. **Conditional identity anonymity:** It should ensure the privacy of the SM's identification so that no possible intruder can get the SM's true identity during authentication. To ensure identity anonymity, RA generates the key-pairs for SM and NAN-GW using their hashed identities (HID_s and HID_n). Additionally, the SM authenticates with the NAN-GW using its public key rather than its true identity. To provide unlinkability, we conceal the public key $\boldsymbol{Pk}_s$ using $\boldsymbol{pid}_s = \boldsymbol{Pk}_s + \boldsymbol{H}_3(\boldsymbol{A}, \boldsymbol{Pk}_n \cdot \boldsymbol{a}^T)$.
4. **Conditional traceability:** To monitor the identification of fraudulent or disruptive clients, the scheme should ensure that there is only one entity capable of revealing the user's true identity. Because the proposed scheme offers identity anonymity and unlinkability for SM, any intruder cannot track SM activities. Even the trusted authority RA cannot access the true identities of SM and NAN-GW because hashed identities (HID_s and HID_n) are sent to RA.
5. **Perfect forward secrecy:** To safeguard lastly transmitted messages, the scheme should ensure that any intruder, even if it has communicators' private

keys, is unable to retrieve previous session keys. Assume that both SM/NAN-GW's private keys are exposed, and the messages $M_1 = (\boldsymbol{A}, \boldsymbol{pid}_s, \boldsymbol{k}_s, t_s)$, $M_2 = (\boldsymbol{B}, \boldsymbol{pid}_n, \boldsymbol{k}_n, t_n)$, and $M_3 = (w)$ are interrupted by an adversary. To gain a prior session-key Sk_{sn}, the intruder can simply retrieve the other parts of the session-key but cannot obtain $K_2 = \boldsymbol{b} \cdot \boldsymbol{A}$ (or $K_4 = \boldsymbol{B} \cdot \boldsymbol{a}^T$), because $\boldsymbol{a}$ and $\boldsymbol{b}$ are chosen at random and are not communicated over the public channel. Hence, the intruder cannot calculate the session key.

6. **Malicious registration:** The RA keeps a database of visitors' IP addresses and access times. When a DDoS attack occurs, the RA has the option to deny the request for the first time. Additionally, every RA that initiates the registration will check whether the same address already exists. A dual protection approach supports the proposed scheme of preventing DDoS attacks.
7. **Resilience against other attacks:** To improve security, the scheme should be resistant to other frequent attacks as well, which are illustrated in the following.
   (a) **Man in the middle (MITM) attack:** Both the SM and the NAN-GW seek signatures in the proposed scheme for mutual authentication. SM and NAN-GW exchange messages $M_1 = (\boldsymbol{A}, \boldsymbol{pid}_s, \boldsymbol{k}_s, t_s)$, $M_2 = (\boldsymbol{B}, \boldsymbol{pid}_n, \boldsymbol{k}_n, t_n)$, and $M_3 = (w)$ for verification. The messages transferred by SM and NAN-GW were easily verified by $\boldsymbol{M} \cdot \boldsymbol{k}_s = \boldsymbol{A} + \boldsymbol{Pk}_s{}' \cdot H_4(\boldsymbol{Pk}_s^{'}, \boldsymbol{pid}_s, \boldsymbol{A}, t_s)$ and $\boldsymbol{k}_n \cdot \boldsymbol{M} = \boldsymbol{B} + \boldsymbol{Pk}_n^{'} \cdot H_4(\boldsymbol{Pk}_n^{'}, \boldsymbol{pid}_n, \boldsymbol{B}, t_n)$ from NAN-GW and SM, respectively. This verification demonstrates the generation of the correct session key Sk_{sn} between SM and NAN-GW. Assume an intruder desires to initiate an MITM attack against the proposed protocol. The invader must first solve the SIS/ISIS hard assumptions to compute the values K_1, K_2, K_3, and K_4 based on the communicated tuples M_1, M_2, and M_3.
   (b) **Impersonation attack:** In order to impersonate an authenticated user, the intruder must obtain the corresponding private keys, which are $\boldsymbol{s}$, $\boldsymbol{x}_s$, $\boldsymbol{x}_n$ of RA, SM and NAN-GW, respectively. Therefore, the intruder must first solve the SIS/ISIS hard assumptions to compute these private keys which is computationally impossible.
   (c) **Replay attack:** To provide replay attack robustness, we use both timestamps and randomness in the proposed protocol. Following the collaborative authentication phase standard, both SM and NAN-GW will examine the timeliness ($|t_s - t_s^{'}| \leqslant \bigtriangledown t$) of messages before authentication.

6 Conclusion

Secure and private communication between SM and NAN-GW is an important concern in the SG systems. Therefore, the article presents a quantum-defended lattice-based anonymous MAKE for the SM and NAN-GW. The inclusion of SIS/ISIS lattice hard assumptions ensures that the proposed scheme will withstand several known and quantum attacks. Because it only employs matrix operations, the scheme is simple and fast to implement in SG systems. Moreover, the

security analysis demonstrates that the designed methodology is secure against a variety of security threats and is also capable of satisfying various security requirements. In the future, to address the single-point failure due to the centralised working of RA, we are going to implement the proposed work using blockchain technology.

References

1. Ajtai, M.: Generating hard instances of lattice problems. In: Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing, pp. 99–108 (1996)
2. Chaudhary, R., Aujla, G.S., Kumar, N., Das, A.K., Saxena, N., Rodrigues, J.J.: LaCSys: lattice-based cryptosystem for secure communication in smart grid environment. In: 2018 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2018)
3. Fouda, M.M., Fadlullah, Z.M., Kato, N., Lu, R., Shen, X.S.: A lightweight message authentication scheme for smart grid communications. IEEE Trans. Smart Grid **2**(4), 675–685 (2011)
4. Garg, S., Kaur, K., Kaddoum, G., Rodrigues, J.J., Guizani, M.: Secure and lightweight authentication scheme for smart metering infrastructure in smart grid. IEEE Trans. Ind. Inf. **16**(5), 3548–3557 (2019)
5. Gupta, D.S., Karati, A., Saad, W., da Costa, D.B.: Quantum-defended blockchain-assisted data authentication protocol for internet of vehicles. IEEE Trans. Veh. Technol. **71**(3), 3255–3266 (2022)
6. Gupta, D.S.: A mutual authentication and key agreement protocol for smart grid environment using lattice. In: Proceedings of the International Conference on Computational Intelligence and Sustainable Technologies, pp. 239–248. Springer (2022)
7. Liang, X.C., Wu, T.Y., Lee, Y.Q., Chen, C.M., Yeh, J.H.: Cryptanalysis of a pairing-based anonymous key agreement scheme for smart grid. In: Advances in Intelligent Information Hiding and Multimedia Signal Processing, pp. 125–131. Springer (2020)
8. Mahmood, K., Chaudhry, S.A., Naqvi, H., Kumari, S., Li, X., Sangaiah, A.K.: An elliptic curve cryptography based lightweight authentication scheme for smart grid communication. Future Gener. Comput. Syst. **81**, 557–565 (2018)
9. Shekhawat, H., Sharma, S., Koli, R.: Privacy-preserving techniques for big data analysis in cloud. In: 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), pp. 1–6 (2019)
10. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Rev. **41**(2), 303–332 (1999)
11. Wang, J., Wu, L., Choo, K.K.R., He, D.: Blockchain-based anonymous authentication with key management for smart grid edge computing infrastructure. IEEE Trans. Ind. Inf. **16**(3), 1984–1992 (2019)
12. Wang, W., Huang, H., Zhang, L., Han, Z., Qiu, C., Su, C.: Blockslap: blockchain-based secure and lightweight authentication protocol for smart grid. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1332–1338. IEEE (2020)
13. Wang, W., Huang, H., Zhang, L., Su, C.: Secure and efficient mutual authentication protocol for smart grid under blockchain. Peer-to-Peer Netw. Appl. **14**(5), 2681–2693 (2021)

14. Wang, W., Lu, Z.: Cyber security in the smart grid: survey and challenges. Comput. Netw. **57**(5), 1344–1371 (2013)
15. Zhang, H., Wang, J., Ding, Y.: Blockchain-based decentralized and secure keyless signature scheme for smart grid. Energy **180**, 955–967 (2019)

Intelligent Cybersecurity Awareness and Assessment System (ICAAS)

Sumitra Biswal(✉)

Bosch Global Software Technologies (BGSW), Bosch, Bangalore, India
Sumitra.Biswal@in.bosch.com

Abstract. With increasing use of sophisticated technologies, attack surface has widened leading to many known and unknown cybersecurity threats. Interestingly, lack of cybersecurity awareness among product manufacturers continues to be the major challenge. Cybersecurity is considered as non-functional requirement and even today, the criticality of cybersecurity is undermined. Several researches have been made to improve cybersecurity models and awareness among developers, however, there is limited to no research on engaging interventions that can help in preliminary education of product manufacturers regarding security-by-design. Besides, poor cybersecurity practices followed by suppliers and ignorance of the same among product manufacturers leads to additional cybersecurity challenges. This study suggests an innovative and convenient approach to help product manufacturers become more holistically aware of cybersecurity issues and help them make more successful and cost-effective decisions on cybersecurity plans for their products.

Keywords: Cybersecurity awareness · Supply chain management · Artificial intelligence · Risk assessment · Manufacturers

1 Introduction

A cybersecurity professional's responsibility includes assessing security relevance of a component and performing its threat and risk assessment. However, the process of cybersecurity does not begin at this stage. Cybersecurity issues are prevalent despite existing cybersecurity practices. This is owing to the fact that, while product manufacturers or Original Equipment Manufacturers (OEMs) (terms used interchangeably in this paper) are consumed at building products with emergent technologies, their awareness of impending danger and attack surfaces associated with these technologies is negligible. There are several OEMs who consider cybersecurity as overrated and do not want to invest time, effort, and finance in implementing cybersecurity measures for their product, let alone security-by-design. Additionally, large supply chains follow poor security practices. OEMs usually do not consider secure supply chain management while dealing with such suppliers. This further contributes towards insecure product design and development. Lack of such mindfulness among OEMs makes it a primary and major cybersecurity challenge. It is realized that, addressing this challenge needs to be the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1143–1153, 2023.
https://doi.org/10.1007/978-3-031-27409-1_105

preliminary step in any cybersecurity engineering process. While researches have been oriented towards improving cybersecurity processes such as threat and risk assessments and upgrading cybersecurity skills among developers, there has been limited research on holistic cybersecurity educational awareness for OEMs.

This study suggests an innovative automated intelligent intervention that would help not only OEMs but also help cybersecurity experts become better at assisting OEMs in effectively understanding cybersecurity requirements. It can aid OEMs in making educated decisions about secure supply chain management and encourage the inclusion of cybersecurity in the design phase. Section 2 provides information on related researches, Sect. 3 highlights observed challenges, Sect. 4 describes the proposal and mechanism, and finally Sect. 5 summarizes the proposal with future directions.

2 Related Work

Several researches have been made to improve cybersecurity threat and risk assessment. For instance, automated risk assessment architectures for Cloud [1] and smart home environment [2] inform about relevant vulnerabilities, threats, and countermeasures. Researchers [3, 4] have proposed games for software developers to educate them about security threats. Cybersecurity awareness and training program lab using Hardware-In-The-Loop simulation system has been developed that replicates Advanced Persistent Threat (APT) attacks and demonstrates known vulnerabilities [5]. Digital labels have also been introduced to inform consumers and passengers about cybersecurity status of automated vehicle being used [6]. Besides, there have been researches [7], briefings [8], framework [9], and articles [10] to highlight the importance of cybersecurity in supply chain management.

However, these studies have some drawbacks, including a lack of understanding of vulnerabilities and automations designed only to improve secure development skills or for a specific domain. Research is required to determine if it is possible to identify important assets and threats (both known and undiscovered) by automated analysis of product design and specifications. A consistent platform for educating OEMs on the various facets of secure product development, including secure supply recommendations based on product requirements, does not exist. Such a platform might inform OEMs about the value of secure architecture and assist them in choosing secure components that are both affordable and effective for their products. This in turn could help in saving a lot of time, finance, and effort invested in mitigating security issues which is believed to be primary and crucial step towards effective cybersecurity.

3 Challenges

Various scenarios in cybersecurity procedures exhibit challenges, but are not limited to, the following:

1. Sub-standard level of cybersecurity awareness among suppliers leads to development of insecure software and hardware.

2. Limited to no cybersecurity awareness among OEMs leads to procurement of compromised hardware and software from suppliers.
3. Despite several security guidelines and best practices, the challenge persists in understanding right amount of security necessary for a product [11]. This is due to insufficient supportive evidence for concerned threats leading to disregard for cybersecurity among stakeholders.
4. The management of cybersecurity engineering processes is resource- and time-intensive due to the lack of a unified platform for cybersecurity awareness and assessment across different domains among associated stakeholders.

4 ICAAS: Proposal and Mechanism

Intelligent Cybersecurity Awareness and Assessment System (ICAAS) intends to mitigate aforementioned challenges by provisioning an intelligent one-stop platform. ICAAS is proposed to incorporate features for following objectives:

1. To assess security requirements and identify assets efficiently from customer designs and specifications.
2. To identify and map known and unknown potential threats to identified assets. This includes, investigating case studies, history, and statistics of relevant attacks as supportive evidence for identified threats, predicting secondary threats from primary threats, and assessing their likelihood using different metrics.
3. To identify and define right security controls by providing security recommendations and references, recommending best practices with sufficient case studies, and elucidating test cases for security controls with foundation.
4. To recommend security compliant supplies available in market.

ICAAS includes four modules—Data acquisition, Psychology oriented cybersecurity (Psyber-security), Security recommender, and Supplies recommender. The high-level architecture of ICAAS is represented in Fig. 1.

4.1 Data Acquisition

An OEM can have diverse specifications defining the functionality of various components and features of a product. However, not all specifications are security specific. It is imperative to identify components that contribute towards such specifications otherwise, realizing security controls will be impossible. Identifying such security relevant specifications using conventional manual methods becomes time-consuming and challenging when OEM lacks basic cybersecurity knowledge with respect to the concerned product and associated components. Besides, time and quality trade-offs are common in manual assessment. Quick processes result in omission of potential security-relevant components whereas, deriving security components from vast specifications takes longer time.

The Data Acquisition module of ICAAS helps in averting these challenges. The inputs to this module are voluminous stacks of product specifications and architectural

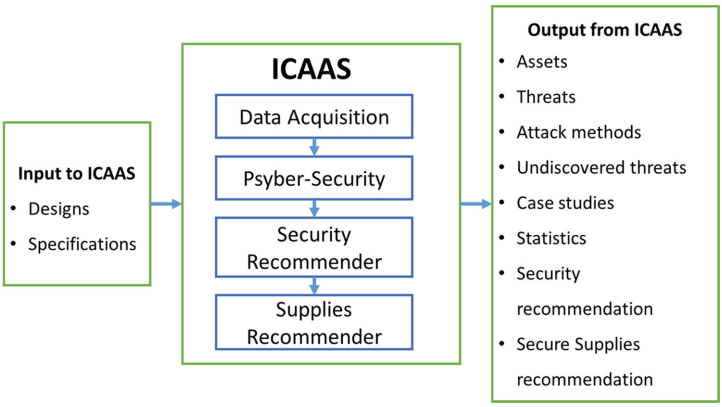


Fig. 1. ICAAS Architecture

design. This module involves Natural Language Processing (NLP) with n-gram models and image interpretation techniques. These techniques derive features and components after extracting security relevant specifications from the inputs. If overall interpretation results are inadequate, then the module generates security specifications by using training rules and repository of specifications relevant to functionalities. Such rules and specifications can be derived from existing product documents, datasheets, and records curated by subject matter experts of different domains. Post refinement, the module identifies security assets in the product that is further used as input for the next module of ICAAS. The high-level architecture of Data Acquisition module is represented in Fig. 2.

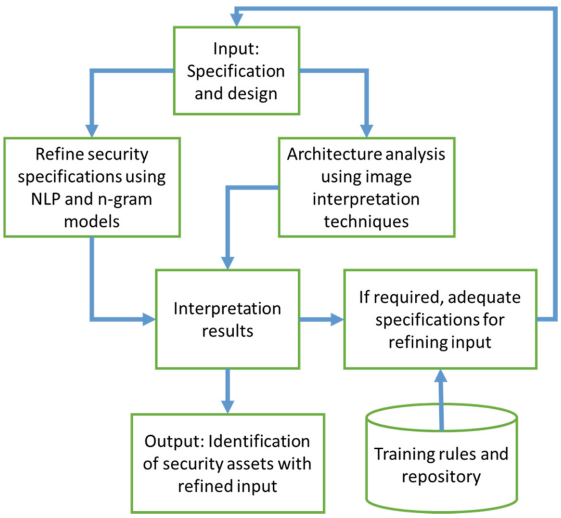


Fig. 2. Data Acquisition Module Architecture

4.2 Psyber-Security

Identification of assets alone does not fulfil the objective. An iterative and logical connection needs to be established between assets and threats, such that the psychological perception of ICAAS user is sufficiently trained to visualize the feasibility of an attack. This module helps in improving decision-making process towards security of assets. The mechanism includes identifying threats' case studies, histories, and attack statistics for a given asset and its functionality. Based on these, relevant components or sub-assets are identified such as, logical and physical interfaces. Relevant threats are selected from threat repository. These threat repositories can be constructed from knowledge bases available on the Internet such as MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) [15]. AI mapping technique [12] is used to map assets and sub-assets with relevant threats and to predict advanced threats using graphs. These graphs associate primary threats to potential secondary threats that will occur if primary threats are not addressed. This helps user perceive the significance of mitigating primary threats. Each threat exploitation is elaborated with Tactic, Techniques, and Procedures (TTPs) to help user realize the attacks' feasibility. Severity of each threat is mapped to corresponding likelihood, attack tree values, weakness (CWE), and vulnerability (CVE) scores.

The output of this module such as threats, exploitation mechanisms, ratings, along with corresponding assets are used as input for the next module of ICAAS. The high-level architecture of Psyber-Security module is represented in Fig. 3.

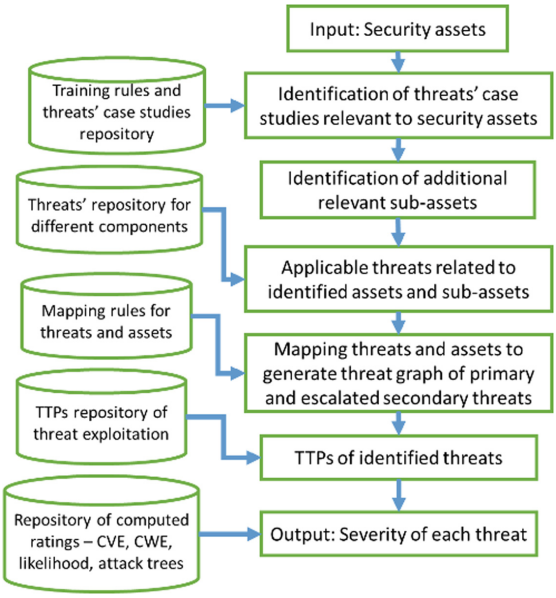


Fig. 3. Psyber-Security Architecture

4.3 Security Recommender

In several cases, due to lack of knowledge, redundant measures are adopted to ensure security of assets that in turn affect usability of the product. Similarly, certain security measures are considered sufficient to combat multiple threats. Due to such misconceptions, it gets difficult to incorporate effective security measures. For instance, in cases of threats related to manipulation of vehicle bus messages, often Cyclic Redundancy Checksum (CRC) is considered adequate for ensuring integrity. However, CRC detects manipulation due to transmission errors and not the manipulated data injected by malicious parties into vehicle bus messages [13]. In such cases, Secure On-board Communication (SecOC) with MAC verification ensures integrity. Hence, it is crucial to incorporate right security that allows usability of product in secure manner.

The Security Recommender module in ICAAS serves the purpose by identifying security best practices, recommendations, and guidelines from repositories. Such repositories can be derived from sources available on the Internet such as National Institute of Standards and Technology (NIST) guidelines [17]. The security recommendations are mapped to identified security threats and assets using AI mapping technique. Appropriate security controls are mapped that fulfil the security recommendations. Finally, security test cases are selected from test cases repository for corresponding security controls. Test cases repository can also be derived from sources available on the Internet such as Open Web Application Security Project (OWASP) [16] and other such publicly available reliable and relevant knowledge bases. The output of the module is used as input for next module. The high-level architecture of Security Recommender module of ICAAS is represented in Fig. 4.

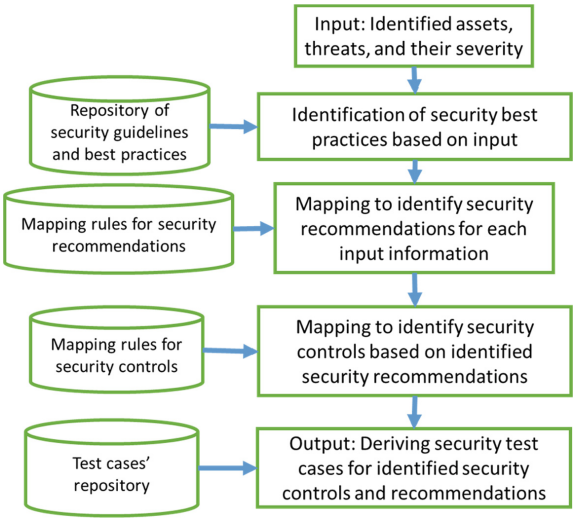


Fig. 4. Security Recommender Architecture

4.4 Supplies Recommender

Awareness of cybersecurity is incomplete with knowledge of assets, security threats, and controls alone. OEMs need to ensure they procure right secure supplies. It is essential to bear this knowledge in advance. In conventional methods, such information is limited and refining right secure supplies is cumbersome.

Supplies Recommender module in ICAAS helps in assisting OEMs with supplies recommendation by identifying supplies list from repository. Such supplies repository can be derived from various sources on the Internet. The identified supplies are mapped to security controls and relevant information. Further, these supplies are recommended based on various factors, not limiting to, supporting security features and specifications of supplies, security test reports of supplies and test case reports based on their use, reviews, and ratings of supplies, identified vulnerabilities, security fixes, versions, and market analysis such as cost and availability along with alternate similar supplies. The high-level architecture of Supplies Recommender module of ICAAS is represented in Fig. 5.

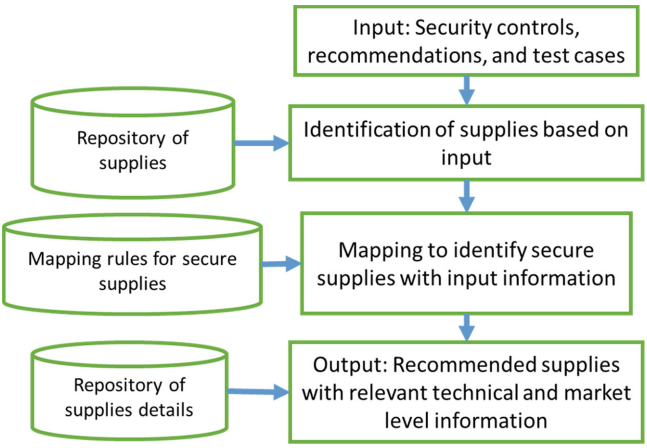


Fig. 5. Supplies Recommender Architecture

5 Experimental Evaluation

In this research, preliminary experimentation was conducted on data acquisition module wherein, publicly available customer specifications were collected and pre-processed to filter security relevant specifications using n-gram based Stochastic Gradient Descent (SGD) classifier. These specifications were further processed to identify the assets using contextual analysis. The significance score based on Bidirectional Encoder Representations from Transformers (BERT) for each of the identified assets were recorded. Owing to time and resource constraint, the total training size for the experimentation was taken as 1000 and test size of 200. A 10-fold based cross-validation was set for the training of

the data acquisition model. Table 1 depicts the n-gram based results of identified security relevant specifications and Table 2 depicts a sample of identified assets along with their BERT based significance scores.

Table 1. Comparison matrix of classifiers

Classification metrics	Classifiers and metrics as per n-grams (1, 2, 3 and 4)		
	SVM	Logistic Regression	Perceptron
Accuracy	TF: 0.48, 0.45, 0.46, 0.48 TF-IDF: 0.41, 0.45, 0.50, 0.51	TF: 0.48, 0.45, 0.46, 0.48 TF-IDF: 0.41, 0.45, 0.50, 0.51	TF: 0.38, 0.46, 0.50, 0.51 TF-IDF: 0.43, 0.45, 0.50, 0.51
Precision (Security relevant)	TF: 0.46, 0.45, 0.47, 0.48 TF-IDF: 0.36, 0.30, 0.00, 0.00	TF: 0.46, 0.45, 0.47, 0.48 TF-IDF: 0.38, 0.30, 0.00, 0.00	TF: 0.35, 0.47, 0.00, 0.00 TF-IDF: 0.38, 0.30, 0.00, 0.00
Precision (Not security relevant)	TF: 0.50, 0.45, 0.46, 0.50 TF-IDF: 0.45, 0.48, 0.51, 0.52	TF: 0.50, 0.45, 0.46, 0.50 TF-IDF: 0.44, 0.48, 0.51, 0.52	TF: 0.41, 0.47, 0.51, 0.52 TF-IDF: 0.46, 0.48, 0.51, 0.52
Recall (Security relevant)	TF: 0.38, 0.62, 0.76, 0.79 TF-IDF: 0.28, 0.10, 0.00, 0.00	TF: 0.38, 0.62, 0.76, 0.79 TF-IDF: 0.31, 0.10, 0.00, 0.00	TF: 0.31, 0.69, 0.00, 0.00 TF-IDF: 0.28, 0.10, 0.00, 0.00
Recall (Not security relevant)	TF: 0.58, 0.29, 0.19, 0.19 TF-IDF: 0.55, 0.77, 0.97, 1.00	TF: 0.58, 0.29, 0.19, 0.19 TF-IDF: 0.52, 0.77, 0.97, 1.00	TF: 0.45, 0.26, 0.97, 1.00 TF-IDF: 0.58, 0.77, 0.97, 1.00
F1-Score (Security relevant)	TF: 0.42, 0.52, 0.58, 0.60 TF-IDF: 0.31, 0.15, 0.00, 0.00	TF: 0.42, 0.52, 0.58, 0.60 TF-IDF: 0.34, 0.15, 0.00, 0.00	TF: 0.33, 0.56, 0.00, 0.00 TF-IDF: 0.32, 0.15, 0.00, 0.00
F1-Score (Not security relevant)	TF: 0.54, 0.35, 0.27, 0.28 TF-IDF: 0.49, 0.59, 0.67, 0.68	TF: 0.54, 0.35, 0.27, 0.28 TF-IDF: 0.48, 0.59, 0.67, 0.68	TF: 0.43, 0.33, 0.67, 0.68 TF-IDF: 0.51, 0.59, 0.67, 0.68

It is inferred that with increasing n-gram size, the security relevant specifications and assets identification improves. The results can be improved further with greater sample size as well as contextual refining of the specifications. Also, with image analysis of architecture, the correlation between specifications and architecture can be mapped accurately to determine assets efficiently.

Table 2. Identified assets with BERT based significance score

Customer security specifications	Identified assets (2–5-g)	BERT based significance score
This system use communication resources which includes but not limited to, HTTP protocol for communication with the web browser and web server and TCP/IP network protocol with HTTP protocol. This application will communicate with the database that holds all the booking information. Users can contact with server side through HTTP protocol by means of a function that is called HTTP Service. This function allows the application to use the data retrieved by server to fulfill the request fired by the user	Application	0.8027
	Database	0.7975
	Booking information	0.7974
	HTTP service	0.7944

6 Conclusion and Future Work

Cybersecurity awareness is a gradual process and there is a demand for strategically engaging methods to help users realize the criticality. However, conventional cybersecurity procedures being majorly manual, limit their reach from widely available resources and information that could improve awareness. Change in architecture and other requirements by OEM requires revisiting entire cybersecurity process which can be exhausting. With relevant case studies and predictive assessments provided by ICAAS, OEMs can have preliminary yet holistic vision of security relevance for their product and design better architecture with standardized security requirements. Lack of timely availability of information on supplies has negative impact on products [14]. But with supplies recommendation of ICAAS, OEMs can pre-plan on secure and suitable supplies for their products.

At present, the data acquisition process of ICAAS has been experimented at a preliminary level and has scope of further improvement as discussed in previous section. Given the application of ICAAS, enormous data may be required for training the model. In such a case, Generative Adversarial Networks (GAN) can be used for careful data augmentation in future work.

Given the infrastructure and computational complexities, ICAAS can also be rendered as a cloud-based service to be subscribed by OEMs. Certain modules of ICAAS such as supplies recommender will require vital dataset on supplies from diverse vendors and supply chains. Although, synthetic data is a solution to limited availability of dataset, certain vital yet sensitive information may be needed from real sources for which privacy may be a major concern. It is therefore believed that, in future, Federated Learning based decentralized approach can be integrated in ICAAS to resolve privacy concerns

related to vital supply chain data sharing or OEM data sharing on Cloud based service. Use of such decentralized approach in ICAAS can help in training its modules by using multiple local datasets without exchanging data. Therefore, the aforementioned plans of integrating GAN and Federated Learning in ICAAS will be undertaken in future to develop and investigate ICAAS's ability as a holistic, efficient, and usable cybersecurity awareness and assessment system.

References

1. Kamongi, P., Gomathisankaran, M., Kavi, K.: Nemesis: Automated architecture for threat modeling and risk assessment for cloud computing. In: The Sixth ASE International Conference on Privacy, Security, Risk and Trust (PASSAT) (2014)
2. Pandey, P., Collen, A., Nijdam, N., Anagnostopoulos, M., Katsikas, S., Konstantas, D.: Towards automated threat-based risk assessment for cyber security in smart homes. In: 18th European Conference on Cyber Warfare and Security (ECCWS) (2019)
3. Jøsang, A., Stray, V., Rygge, H.: Threat poker: Gamification of secure agile. In: Drevin, L., Von Solms, S., Theocharidou, M. (eds.) Information Security Education. Information Security in Action. WISE 2020. IFIP Advances in Information and Communication Technology, Vol. 579. Springer, Cham (2020)
4. Gasiba, T., Lechner, U., Pinto-Albuquerque, M., Porwal, A.: Cybersecurity awareness platform with virtual coach and automated challenge assessment. In: Computer Security. Cyber-ICPS SECPRE ADIoT 2020. Lecture Notes in Computer Science, Vol. 12501. Springer, Cham (2020)
5. Puys, M., Thevenon, P.H., Mocanu, S.: Hardware-in-the-loop labs for SCADA cybersecurity awareness and training. In: The 16th International Conference on Availability, Reliability and Security (ARES). Association for Computing Machinery, New York, NY, USA, Article 147, 1–10 (2021)
6. Khan, W.Z., Khurram Khan, M., Arshad, Q.-u.-A, Malik, H., Almuhtadi, J.: Digital labels: Influencing consumers trust and raising cybersecurity awareness for adopting autonomous vehicles. In: IEEE International Conference on Consumer Electronics (ICCE), pp. 1–4 (2021)
7. Melnyk, S.A., Schoenherr, T., Speier-Pero, C., Peters, C., Chang, J. F., Friday, D.: New challenges in supply chain management: cybersecurity across the supply chain. Int. J. Prod. Res. **60**(1), 162–183 (2022)
8. NIST best practices in supply chain risk management (Conference Materials). Cyber supply chain best practices. https://csrc.nist.gov/CSRC/media/Projects/Supply-Chain-Risk-Management/documents/briefings/Workshop-Brief-on-Cyber-Supply-Chain-Best-Practices.pdf. Last Accessed 09 July 2022
9. Boyens, J., Paulsen, C., Bartol, N., Winkler, K., Gimbi, J.: Key practices in cyber supply chain risk management: observations from industry. https://csrc.nist.gov/publications/detail/nistir/8276/final. Last Accessed 09 July 2022
10. Patil, S.: The supply chain cybersecurity saga: challenges and solutions. https://niiconsulting.com/checkmate/2022/02/the-supply-chain-cybersecurity-saga-challenges-and-solutions/. Last Accessed 09 July 2022
11. Nather, W.: How much security do you really need?, https://blogs.cisco.com/security/how-much-security-do-you-really-need. Last Accessed 09 July 2022
12. Adeptia. AI-Based Data Mapping. https://adeptia.com/products/innovation/artificial-intelligence-mapping#:~:text=AI%20mapping%20makes%20data%20mapping,to%20create%20intelligent%20data%20mappings. last accessed 2022/07/09

13. Bozdal, M., Samie, M., Aslam, S., Jennions. I.: Evaluation of CAN bus security challenges. Sensors **20**(8), 2364 (2020)
14. Leyden, J.: Toyota shuts down production after 'cyber attack' on supplier. https://portswigger.net/daily-swig/toyota-shuts-down-production-after-cyber-attack-on-supplier. Last Accessed 09 July 2022
15. MITRE ATT&CK. https://attack.mitre.org/. Last Accessed 10 July 2022
16. Testing Guide—OWASP Foundation. https://owasp.org/www-project-web-security-testing-guide/assets/archive/OWASP_Testing_Guide_v4.pdf. Last Accessed 10 July 2022
17. National Checklist Program. https://ncp.nist.gov/repository. Last Accessed 10 July 2022



A Study on Written Communication About Client-Side Web Security

Sampsa Rauti(✉), Samuli Laato, and Ali Farooq

University of Turku, Turku, Finland
{tdhein,sadala,alifar}@utu.fi

Abstract. Today, web services services are widely used by ordinary people with little technical know-how. End user cybersecurity in web applications has become an essential aspect to consider in web development. One important part of online cybersecurity is the HTTPS protocol that encrypts the web traffic between endpoints. This paper explores how the relevant end user cybersecurity instructions are communicated to users. Using text-focused analysis, we study and assess the cybersecurity instructions online banks and browser vendors provide with regards to HTTPS. We find that security benefits of HTTPS are often exaggerated and can give users a false sense of security.

Keywords: HTTPS · web application security · cybersecurity education · security guidance

1 Introduction

As online services are often created for and widely used by laypeople with little technical knowledge, end user cybersecurity has become a crucial and relevant aspect to consider in the overall security of information systems (IS) [1,6,17,18]. One of the most popular tools for accessing online services is the web browser. Here, HTTP (Hypertext Transfer Protocol) is the means browsers use to connect to websites. HTTPS (Hypertext Transfer Protocol Secure) is a HTTP connection using modern encryption (currently TLS)[1], securing the connection and preventing man-in-the-middle attacks between communication endpoints. In most browsers, a HTTPS connection to a website has conventionally been indicated with an URL address beginning with HTTPS rather than HTTP, and a small padlock symbol in the address bar [12]. Already introduced in 1994, HTTPS has been steadily growing in popularity. In September 2022, Google reported that HTTPS is used as a default protocol by almost 80% of all web sites[2].

While using HTTPS is indeed important and users should be aware of it, it does not guarantee full protection. For example, malicious websites may simply

[1] https://tools.ietf.org/html/rfc2818.
[2] https://w3techs.com/technologies/details/ce-httpsdefault.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1154–1166, 2023.
https://doi.org/10.1007/978-3-031-27409-1_106

purchase a cheap HTTPS certificate which makes popular browsers display them as secure despite the content of the website being dangerous. Furthermore, there are many layers of communication between HTTPS and the end user, which may be targeted by adversaries. Recent work has discussed attacks such as man-in-the-browser which are able to completely circumvent the protection offered by HTTPS [23]. As a consequence, there also exists a danger of overemphasizing the security provided by HTTPS in end user cybersecurity communication.

The aim of this work is to investigate how essential end user cybersecurity knowledge is communicated in security critical web applications, in particular bank websites. We analyze and evaluate the cybersecurity guidance they provide with regards to HTTPS using text-focused analysis. Consequently, we formulate the following research questions:

RQ1: *How do bank websites and popular browser vendors communicate to users about HTTPS?*
RQ2: *Do the online banks and browser vendors over- or under-emphasize the security benefits provided by HTTPS?*

2 Background

Accelerated by technology trends such as the utilization of cloud services, a multitude of services are offered online [19]. These consist of old services being transformed online (e.g. banking [16]) and new services emerging such as Internet of things (IoT) management systems and social media [20]. Furthermore, many desktop applications are being replaced with web applications, which are accessible everywhere and updated automatically. At the same time, web security relies heavily on users' knowledge about the environment, including their ability to detect potentially malicious websites and avoid them. One of the key visual cues in browsers indicating that to users that a website is secure is the padlock symbol on the address bar. However, while the users may easily assume that this symbol indicates a completely secure web browsing experience, the padlock merely means that the connection to the server uses the HTTPS protocol. Thus, a detailed analysis on how the meaning of HTTPS and encryption is communicated to users is needed.

2.1 Advantages and Misconceptions of the HTTPS Protocol

HTTPS has become a significant element in ensuring secure web browsing. Google has campaigned in favor of secure web[3], advocating adoption of HTTPS encryption for websites. Amidst all the hype surrounding the secure web, however, it has often been forgotten that HTTPS and TLS only secure the end-to-end connection, not the security of the client (browser) or the security and integrity of web pages at endpoints.

[3] https://security.googleblog.com/2018/02/a-secure-web-is-here-to-stay.html.

HTTPS encrypts the communication in transit, but does not provide any protection when the unencrypted data is handled on the client or server side or when it is stored in databases. Therefore, HTTPS does not fully guarantee security, safety or privacy, although users may think so based on many cybersecurity instructions. For example, attacks with malicious browser extensions can effortlessly be implemented on the client side when HTTPS in being used [21].

Moreover, the certificate and necessary infrastructure for HTTPS are easy to obtain for any service provider, also for scammers, and they only guarantees the authenticity of the domain name or the party (e.g. company) maintaining the website. However, users are in no way protected from a website that is malicious to begin with, before it is sent to the client over an encrypted connection.

Motivating and governing HTTPS usage has been incorporated into browsers and web concepts in many ways. These include limitations and guidelines given to developers, such as disallowing mixing HTTP and HTTPS (mixed content) on websites and requiring it as a part of progressive web apps. However, HTTPS has also been acknowledged in cybersecurity communication aimed at end users. Examples of this include directing users to look for a padlock symbol in the address bar to make sure the connection is secure, labeling websites not using HTTPS insecure, and introducing additions like *HTTPS-Only* mode[4] in Firefox and the HTTPS Everywhere[5] extension.

2.2 End User Cybersecurity Behavior

A major research direction in cybersecurity research concerns the end users and their behavior. This research has focused on aspects such as security policy violations [25], personal data exposure and collection [22], and the impact of personality on cybersecurity behavior [24] among others. It is important to understand the security awareness level of end users, as it is a paramount component in the overall security of IT systems [6]. Therefore, ensuring end users are up-to-date on relevant cybersecurity issues and respective behavior and culture is essential.

There are several factors that impact end-users security behaviour (e.g. see [7,9]). These include formal and non-formal education [3,15], offering end users privacy policies that explain potential issues [22,25], information dissemination [2] and security indicators [12]. Textual information on recommended cybersecurity behaviors are offered by almost all internet browsers and online banking websites, which are in focus in this study.

Researchers have suggested that knowledge of security threats [4,14] is a crucial part of cybersecurity awareness. However, a recent work (e.g. [5]) suggests that merely knowledge of security threats does not guarantee secure behavior. In addition to threat knowledge, users need to have necessary skills to act in a secure way. Thus, behavioral guidance is needed. This can be achieved through cues and nudges implemented as part of information systems that guide user

[4] https://blog.mozilla.org/security/2020/11/17/firefox-83-introduces-https-only-mode.

[5] https://www.eff.org/https-everywhere.

behavior to a more secure direction [9]. These cues and nudges can be icons, sounds, popups and other sensory cues that inform end users about the state of cybersecurity. The information they convey can indicate either that things are secure or that they are not.

Several studies (e.g. [11,13,26]) show that people have a flawed understanding about the internet. This in itself is a cybersecurity concern. Browsers are a way to browse the internet and researchers have suggested a number of ways to improve end user security. Krombholz et al., [13] summarized a myriad of literature on security indicators in internet browsers and banking apps, and demonstrated that these indicators have advanced on multiple fronts to provide understandable knowledge to end users. The aim of these indicators is not per se to reflect the technical reality, but rather to direct end users towards desired secure behavior. In a work published in 2015, security experts suggest checking for HTTPS as one of the top six measures users should take for their security [10]. To nudge the users towards paying attention to the HTTPS connection, browsers such as Google Chrome and Mozilla Firefox display a padlock symbol in the address bar as an indication of the HTTPS connection. Furthermore, browsers may issue warnings to users if they are about or enter passwords of credit card information on a HTTP site [8].

In summary, end user cybersecurity behavior is influenced by several parties (e.g. browsers, legislators, news outlets) and in many ways (e.g. nudging, informing). It is paramount to ensure that the actions to increase secure end user behavior work as intended and do not, in fact, have adverse effects. In particular, the communication of HTTPS and the padlock symbol are worthy to investigate in this regard.

3 Materials and Methods

In order to respond to the presented research questions, we focus on cybersecurity communication aimed at the users in (1) web browsers; and (2) banks. We analyze these from the perspective of how well they match the technical implementation of HTTPS and the real security aspects it provides. Thus, looking at Fig. 1, our focus is on the middle box and its relationship with the technical implementations. Accordingly, our study differs from other cybersecurity user studies which focus on end users via interviews or surveys [6].

3.1 Data Sources

We investigate the communication to the users via semantic analysis of two sources. First, how six of the most popular internet browsers (Google Chrome, Firefox, Opera, Safari, Microsoft Edge and Internet Explorer[6]) communicate to their users about HTTPS. These browsers were selected based on popularity as

[6] Popularity of browsers fetched from Kinsta at https://kinsta.com/browser-market-share/ on 5th of March, 2021.

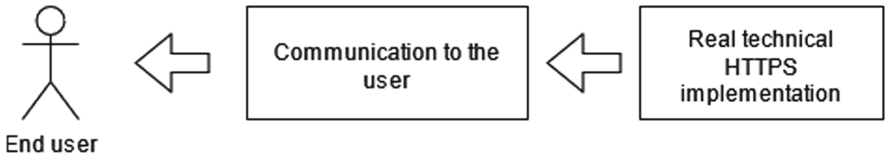


Fig. 1. A visualization how HTTPS technology and implementation are explained and communicated to end users. Instead of end users most often directly being aware of what is going on, they obtain their information through second-hand sources such as the cybersecurity guidance that internet browsers provide.

measured by the number of active users globally. We fetched the instructions that the browsers deliver to their users from official sources, which varied between the browser providers. In case varied instructions were given to the PC and mobile version of the selected browser, we preferred the PC version for continuity's sake. The cybersecurity instructions were glanced through and all information relating to HTTPS or the lock symbol on the address bar were stored for more detailed analysis.

Second, we studied how critical high-security web sites, in this case online banks, communicate about HTTPS to end users. Similarly to the web browsers, the banks were also selected for analysis based on their popularity in the target country. We searched a list of the world's 100 largest banks and via random sampling selected 20 banks for analysis.

In order to abide by the standards of ethical research, we have redacted the names of the banks in this work. This is done to avoid targeting specific companies with potentially damaging results.

3.2 Analysis

With these two sets of data were are able to provide an overview of how HTTPS systems are communicated to end users and identify potentially problematic terminology and user guidance. In order to extract potential problems from the selected set of HTTPS related communication, we approached the texts from the perspective of the technical implementation of HTTPS which is depicted on the right hand side in Fig. 1. Following the semantic analysis approach, we focused on all communication that was not aligned with the technical implementation. We wrote down these identified issues and classified them into clusters. We present these clusters including direct quotes from the browsers' communication in the following section.

4 Results

Guided by our research method, we identified two separate categories of how the security provided by HTTPS is communicated to end users. We identified issues with (1) terminology, and (2) user guidance. In the following, we discuss these two separately.

4.1 Issues with Terminology

Table 1 shows the terminology online banks use to describe the security provided by the HTTPS protocol on their pages. We can see that the most common terms to describe HTTPS are "secure website" and "secure connection". In what follows, we will look at the potential problems with this terminology.

Table 1. How is security or privacy provided by HTTPS described? Terms used on 20 studied online bank cybersecurity guidance pages.

Term	N
Secure website/webpage/site	10
Secure connection	2
Authentic certificate	1
Encrypted connection	1
Legitimate site	1
Secure session	1
Secure transaction	1
Secure data transmission	1

Is the page secure? When cybersecurity guides talk about *secure web pages*, they usually imply that HTTPS and the TLS connection are used. However, it may not immediately be clear to the user that a web page or web application delivered using a secure connection can still be insecure in many ways. For example, a web application can be poorly implemented and contain injection vulnerabilities that leak the user's private data to other users, web pages can be laced with malware, or the owner of the website may simply be a scammer who has acquired a certificate. In all of these cases, the connection may be secure but the web page itself is not.

Accordingly, when cybersecurity guidance calls a web page secure, they merely refer to that the browser connects the remote site using a secure protocol and therefore, attackers cannot tamper with the data between the communication endpoints. However, for the user, security of a web page arguably also means that the web page (the HTML document) they have downloaded for viewing and interact with would be safe to use without compromising their private data and online transactions. Unfortunately, this is not the case. The conception of secure web page can easily become too broad in the user's mind, which makes it problematic to divide web pages into secure and insecure ones just based on their HTTPS usage. Likewise, calling the web where every website would use HTTPS *secure web* can create a false sense of security.

Is the connection secure? Based on the above, calling web pages secure can be confusing and even harmful for users. There is more to the story, however, because HTTPS does not even guarantee a secure connection in the sense

users may understand it. If implemented and utilized correctly, TLS guarantees security on the transport layer, preventing man-in-the-middle-attacks that aim to spy on or tamper with the data sent over the network. However, there is also an alternative interpretation as to what *end-to-end encryption* and secure connections mean.

Whether the connection is secure depends where the end-points of the connection are considered to be and where the "middle" of man-in-the-middle attacks is located. For example, a user might expect every point between the user interface and web server to be secure. Alternatively, the secure connection could be expected to begin when the web application forms a HTTP connection to the server. In both of these scenarios the "connection" is potentially compromised, because the data in the user interface and the data sent from the web application can easily be read and modified for example by a malicious browser extension or an independent piece of malware that has hooked into the browser. These attacks happen on the layers where there is no TLS protection and HTTPS is therefore useless. It is important to understand that TLS is only meant to encrypt the data *during delivery*, not when it is stored or used. The attacker can strike before the application layer data is encrypted or again after the encryption has been removed. From this perspective, Microsoft Edge promises a little too much in its in-browser description of the secure connection, stating that "[...] information (such as passwords or credit cards) will be securely sent to this site and cannot be intercepted".

In our sample of online banking websites and browsers, the studied browser vendors used more accurate terminology than the online banks. The browser vendors did not talk about secure websites, but only call the connection secure. However, there was one exception among the browsers. Google Chrome's help page seems to talk about secure connection and private connection interchangeably, which may further confuse readers. Browser vendors also did not go into detail about what parts of data transmission are guaranteed to be secure, which leaves the term "secure connection" vague and open for misunderstandings.

To summarize, the security terminology revolving around the use of HTTPS in online banks' websites and browsers' instructions largely use overoptimistic and exaggerated language when it comes to cybersecurity. While scaring users with threat scenarios may not be wise either, the used terminology makes unwarranted promises about security. This can have negative impact on end users' cybersecurity awareness and give rise to a false sense of security.

4.2 Problems with Guidance

Table 2 shows cybersecurity guidance given on the studied bank websites on how end users can make sure the website and the connection are secure and legitimate. As can be seen from the Table, almost all the banks list "HTTPS" in the web address as a sign of a secure website and connection. Not only is this problematic because HTTPS does not guarantee the security and integrity of a website itself, but it is outright misleading, because at least the Google Chrome browser has discontinued the practice of displaying the "HTTPS" prefix in the

address. Unfortunately, not many security guidance pages have been updated to reflect this change.

Table 2. How to make sure a website or connection is secure? Cybersecurity guidance given on the studied bank websites.

Bank ID	HTTPS in the address bar	Lock symbol	Check the address is correct	Check the certificate is legitimate
1	X	X		
2	X	X		
3	X	X		
4	X	X		
5	X	X		
6	X	X		
7	X	X		
8	X	X	X	
9	X	X		X
10	X	X		
11	X	X		X
12				X
13	X	X		
14		X		
15	X	X		X
16	X	X	X	
17	X	X		X
18	X	X		X
19	X	X		X
20	X	X		X

Another popular alleged sign of a secure website and connection is the padlock symbol. However, even together with HTTPS, this is not an indication of secure or authentic webpage as fraudsters can easily obtain certificates that makes the site appear secure. Almost half of the cybersecurity guide pages only mention the combination of HTTPS and padlock as a sign of security, which is utterly insufficient.

Checking the address in the address bar was only mentioned 2 times, and users were instructed to click the padlock icon to confirm the certificate of the webpage or the bank only in 8 cases. In majority of fraud and phishing scenarios, the displayed URL is something which cannot and has not been fabricated. Therefore, it is concerning that users are not instructed to check and verify the

address. Clicking on the padlock and checking that the certificate is legitimate is good advice as well, although it is questionable whether the user wants to go through the trouble of checking this. The user may also not be able to differentiate between a genuine certificate and a fake that the scammer has procured for their fraudulent site. Consequently, users should be made more aware of what the correct URL for their bank's website is and what the correct certificate looks like. Unsafe practices such as searching for the bank's name in the search engine and possibly clicking a link leading to a fake banking site should be strongly discouraged by the cybersecurity instructions, but this was not the case.

Not surprisingly, the guidance provided by the browser vendors is more accurate than cyber security instructions of online banks. For example, they contain information on secure certificates and explain how to check their authenticity. However, at times they still contain claims that can be seen as exaggerated, such as padlock symbol indicating that entering sensitive information is fully protected[7,8].

5 Discussion

5.1 Theoretical and Practical Implications

We summarize the key contributions of this work in Table 3. These relate primarily to three areas: (1) cybersecurity communication; (2) security indicator design; and (3) end user cybersecurity. Below we discuss these implications in further detail and how elucidate how they connect to extant literature.

Table 3. Key contributions

Contribution area	Key contributions
Security communication	The security instructions for end users on the world's most popular Bank's websites are outdated Education on how systems work should not be replaced by blind trust on security indicators It is problematic if end users learn to trust that every time something is wrong with their system they see an indicator
Security indicator design	Security indicators may provide a false sense of security In addition to guiding behavior security indicators could be designed to guide learning about potent security measures
End user cybersecurity	There is a shared responsibility between banks, the government and other related agencies to educate the crowds about the current trends in cybercrime and provide knowledge on how to stay protected. Banks should not fall behind in inadequate security communication that leads to a false sense of security

[7] https://support.google.com/chrome/answer/95617.
[8] https://help.opera.com/en/latest/security-and-privacy/.

With regards to cybersecurity communication, we contribute in to the literature on security indicators in web browsers [13]. Through the performed analysis of cybersecurity communication in the world's largest bank's webpages we offer a unique viewpoint to the literature that largely focuses on empirical user studies [7,9].

With regards to security indicators and their design, our work offers a fresh perspective reminding of the potential dangers of simplified communication. For example, Krombholz et al., [13] found that end users oftentimes underestimate the security benefits of using HTTPS. Based on our findings, blindly trusting the padlock symbol to make web browsing secure at times when it is quick and cheap to get a HTTPS certificate for any website is unwise. Furthermore, it is problematic if end users learn to trust that every time something is wrong with their system they see an indicator of sorts.

Finally, with regards to end user cybersecurity, our findings align with previous work in that knowledge about cybersecurity threats and education on how the systems work on a general level is needed [4,14]. Our findings further contradict the argument that security indicators would be better than nothing. In fact, we argue that they may even have a negative impact on cybersecurity for the following reasons:

- They can lure individuals into a false sense of security.
- They may make end users lazy and to not bother to learn how systems actually work.

5.2 Limitations and Future Work

Our empirical work has the following limitations. First, we reviewed the cybersecurity communication of the most popular online banks and browsers, but it may very well be that this is not the primary source of information for many end users. Other sources including alternative websites, social media, news sites, formal education and mouth-to-mouth sources need to also be considered. To account all these, interview studies with end users could be conducted, an approach adopted by related work (e.g., [13]). Second, we analysed the online banks' and browsers' end user cybersecurity communication specifically with regards to HTTPS. Of course, other important aspects regarding end user cybersecurity behavior and communication exist, and future work could explore these.

6 Conclusion

When used and implemented correctly, HTTPS and TLS are essential technologies to safeguard data when it is transmitted between the user's browser and the server. While saying that HTTPS is secure is not wrong, it is a misconception that using the protocol would keep the user data safe inside the browser or even at every point of the data transmission. HTTPS is only one important piece of cybersecurity, and users and web service providers need to be educated on the threats HTTPS does not protect against and the necessary countermeasures.

HTTPS will no doubt become even more prevalent in the future when a new version of HTTP, HTTP/2 is adopted more widely. Although the protocol does not require mandatory encryption, in practice it is required by most client implementations. Hopefully, we will soon to be able to move to a web where every site uses HTTPS and trustworthy certificates by default, and developers as well as users can concentrate more on other security issues.

As the world becomes increasingly digital and complex, the pitfall of simplifying things too much for end users via security indicators and visual cues becomes more prominent. Based on our findings here, we stress the paramount importance of end user cybersecurity education as opposed to luring users to potential false sense of security through teaching them to rely on oversimplified security indicators.

In conclusion, we are not arguing that cybersecurity communication to end users should disclose everything about the technical implementation. However, end user communication should make sure to provide a realistic view of the used security measures so that users are not lead into a false sense of security.

References

1. Carlton, M., Levy, Y.: Expert assessment of the top platform independent cybersecurity skills for non-it professionals. In: SoutheastCon 2015, pp. 1–6. IEEE (2015)
2. Dandurand, L., Serrano, O.S.: Towards improved cyber security information sharing. In: 2013 5th International Conference on Cyber Conflict (CYCON 2013), pp. 1–16. IEEE (2013)
3. Farooq, A., Hakkala, A., Virtanen, S., Isoaho, J.: Cybersecurity education and skills: exploring students' perceptions, preferences and performance in a blended learning initiative. In: 2020 IEEE Global Engineering Education Conference (EDUCON), pp. 1361–1369. IEEE (2020). https://doi.org/10.1109/EDUCON45650.2020.9125213
4. Farooq, A., Isoaho, J., Virtanen, S., Isoaho, J.: Information security awareness in educational institution: an analysis of students' individual factors. In: 2015 IEEE Trustcom/BigDataSE/ISPA, vol. 1, pp. 352–359. IEEE (2015)
5. Farooq, A., Jeske, D., Isoaho, J.: Predicting students' security behavior using information-motivation-behavioral skills model. In: IFIP International Conference on ICT Systems Security and Privacy Protection, pp. 238–252. Springer (2019)
6. Farooq, A., Kakakhel, S.R.U.: Information security awareness: comparing perceptions and training preferences. In: 2013 2nd National Conference on Information Assurance (NCIA), pp. 53–57. IEEE (2013)
7. Farooq, A., Ndiege, J.R.A., Isoaho, J.: Factors affecting security behavior of Kenyan students: an integration of protection motivation theory and theory of planned behavior. In: 2019 IEEE AFRICON, pp. 1–8. IEEE (2019)
8. Felt, A.P., Barnes, R., King, A., Palmer, C., Bentzel, C., Tabriz, P.: Measuring {HTTPS} adoption on the web. In: 26th USENIX Security Symposium (USENIX Security 17), pp. 1323–1338 (2017)

9. Howe, A.E., Ray, I., Roberts, M., Urbanska, M., Byrne, Z.: The psychology of security for the home computer user. In: 2012 IEEE Symposium on Security and Privacy, pp. 209–223. IEEE (2012)
10. Ion, I., Reeder, R., Consolvo, S.: "... no one can hack my mind": Comparing expert and non-expert security practices. In: Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 327–346 (2015)
11. Kang, R., Dabbish, L., Fruchter, N., Kiesler, S.: "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In: Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 39–52 (2015)
12. Kraus, L., Ukrop, M., Matyas, V., Fiebig, T.: Evolution of SSL/TLS indicators and warnings in web browsers. In: Cambridge International Workshop on Security Protocols, pp. 267–280. Springer (2019)
13. Krombholz, K., Busse, K., Pfeffer, K., Smith, M., von Zezschwitz, E.: "if https were secure, i wouldn't need 2fa"-end user and administrator mental models of https. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 246–263. IEEE (2019)
14. Kruger, H.A., Kearney, W.D.: A prototype for assessing information security awareness. Comput. Secur. **25**(4), 289–296 (2006)
15. Laato, S., Farooq, A., Tenhunen, H., Pitkamaki, T., Hakkala, A., Airola, A.: Ai in cybersecurity education-a systematic literature review of studies on cybersecurity moocs. In: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), pp. 6–10. IEEE (2020). https://doi.org/10.1109/ICALT49669.2020.00009
16. Li, F., Lu, H., Hou, M., Cui, K., Darbandi, M.: Customer satisfaction with bank services: the role of cloud services, security, e-learning and service quality. Technol. Soc. **64**, 101487 (2021)
17. Li, L., He, W., Xu, L., Ash, I., Anwar, M., Yuan, X.: Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. Int. J. Inf. Manag. **45**, 13–24 (2019)
18. Lombardi, V., Ortiz, S., Phifer, J., Cerny, T., Shin, D.: Behavior control-based approach to influencing user's cybersecurity actions using mobile news app. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp. 912–915 (2021)
19. Malar, D.A., Arvidsson, V., Holmstrom, J.: Digital transformation in banking: exploring value co-creation in online banking services in India. J. Glob. Inf. Technol. Manag. **22**(1), 7–24 (2019)
20. Newman, N.: The rise of social media and its impact on mainstream journalism (2009)
21. Rauti, S.: A survey on countermeasures against man-in-the-browser attacks. In: International Conference on Hybrid Intelligent Systems, pp. 409–418. Springer (2019)
22. Rauti, S., Laato, S.: Location-based games as interfaces for collecting user data. In: World Conference on Information Systems and Technologies, pp. 631–642. Springer (2020)
23. Rauti, S., Laato, S., Pitkämäki, T.: Man-in-the-browser attacks against IoT devices: a study of smart homes. In: Abraham, A., Ohsawa, Y., Gandhi, N., Jabbar, M., Haqiq, A., McLoone, S., Issac, B. (eds.) Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020), pp. 727–737. Springer International Publishing, Cham (2021)
24. Shappie, A.T., Dawson, C.A., Debb, S.M.: Personality as a predictor of cybersecurity behavior. Psychol. Popul. Med. Cult. (2019)

25. Siponen, M., Vance, A.: Neutralization: new insights into the problem of employee information systems security policy violations. In: MIS Quarterly, pp. 487–502 (2010)
26. Wu, J., Zappala, D.: When is a tree really a truck? Exploring mental models of encryption. In: Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018), pp. 395–409 (2018)

It's All Connected: Detecting Phishing Transaction Records on Ethereum Using Link Prediction

Chidimma Opara[1(✉)], Yingke Chen[2], and Bo Wei[3]

[1] Teesside University, Middlesbrough, UK
c.opara@tees.ac.uk
[2] Northumbria University, Newcastle Upon Tyne, UK
[3] Lancaster University, Lancaster, UK

Abstract. Digital currencies are increasingly being used on platforms for virtual transactions, such as Ethereum, owing to new financial innovations. As these platforms are anonymous and easy to use, they are perfect places for phishing scams to grow. Unlike traditional phishing detection approaches that aim to distinguish phishing websites and emails using their HTML content and URLs, phishing attacks on Ethereum focus on detecting phishing addresses by analyzing the transaction relationships on the virtual transaction platform. This study proposes a link prediction framework for detecting phishing transactions on the Ethereum platform using 12 local network-based features extracted from the Ether receiving and initiating addresses. The framework was trained and tested on over 280,000 verified phishing and legitimate transaction records. Experimental results indicate that the proposed framework with a LightGBM classifier provides a high recall of 89% and an AUC score of 93%.

Keywords: Phishing detection · Ethereum Network · Link prediction · Graph representation

1 Introduction

Blockchain, a distributed ledger, has captured the attention of industry and academia since its introduction in 2008. The most well-known use of blockchain technology is on cryptocurrency platforms, such as Bitcoin and Ethereum. In blockchain systems, transactions are messages sent from the initiator (source address) to the receiver (target address) [1]. By preserving a secure and decentralized transaction record, its use on these cryptocurrency platforms ensures record authenticity, security, and confidence without needing a third party. Buterin, credited as the creator of Ethereum, was among the first to recognize the full potential of blockchain technology, which extended beyond enabling secure virtual payment methods. After Bitcoin, the Ethereum network's Ether cryptocurrency is the second most popular digital currency [11].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1167–1178, 2023.
https://doi.org/10.1007/978-3-031-27409-1_107

Phishing is a well-known social engineering technique that tricks Internet users into disclosing private information that can be fraudulently used. Researchers have been working on detecting and preventing phishing on the Internet for the last two decades. Nevertheless, the primary environments have been emails [2] and websites [7,8]. With the advancement of blockchain technology, phishing scams on cryptocurrency transactions have increased exponentially, necessitating a focus on detecting phishing in the virtual transaction environment.

Phishing detection methods in virtual transaction environments differ from traditional websites in target objects and data sources. Specifically, unlike on traditional websites, phishing detection focuses on distinguishing malicious web content, while on virtual transaction platforms, the focus is on detecting phishing addresses. In other words, while detecting phishing on traditional websites relies on the analysis of the content of the web page (URL, HTML, and network attributes), the detection framework in virtual transaction environments utilizes the transaction records between Ethereum addresses to distinguish between phishing and non-phishing addresses. Therefore, using phishing detection approaches for traditional phishing attacks on web pages and emails will be unsuitable for mitigating attacks on the Ethereum platform.

Existing phishing detection techniques on the Ethereum platform have focused on two approaches: 1. extracting statistical features from the amount and time stamp attributes, and 2. applying network embedding techniques to the above attributes. These approaches are based on the assumption that the amount of Ether sent between addresses and the record of time spent are the most important factors to consider when detecting phishing addresses. However, these approaches are limited as they depend on detecting large amounts of value because they imply a legitimate transaction. Using transaction amount as a criterion gives rise to a high misclassification of legitimate transactions with a low transaction amount. Also, phishing transactions in which significant amounts have been transacted are wrongly classified.

This paper uses a different approach to detecting phishing addresses on a virtual transaction platform. Intuitively, detecting phishing in the virtual transaction environment aims to alienate the bad actors. Therefore, instead of modelling the relationship between the transaction amount and the transaction time, we focused on the relationship between the addresses of the transactors to establish a pattern between them using statistical features. Our proposed approach is not dependent on the specific amount transacted but on the presence of any transaction to and from a suspicious node. Furthermore, the method proposed in this paper removes the extra complexity of using network embedding techniques while providing a high AUC score.

Specifically, we propose a link prediction model that predicts whether a relationship exists between two transacting addresses on the Ethereum platform based on their node data. The node data in this paper comprises labelled node pairs (Ether transferring node address - Ether receiving node address) corresponding to possible transaction links and outputs 12 tailored features based on the node pairings. These features represent graph edges and are divided into pos-

itive and negative samples based on their node labels. Subsequently, the graph edges with corresponding labels are fed into a LightGBM classifier to obtain link predictions.

The main contributions of this work are as follows:

- This paper proposes a link prediction model that uses only the addresses of the receiving and sending addresses and extracts features from the local network on the Ethereum platform. The proposed approach is not dependent on the specific amount transacted but on the presence of any transaction to and from a suspicious node. Furthermore, the method proposed in this paper removes the extra complexity of using network embedding techniques while providing high recall and AUC scores.
- The proposed framework's efficiency in identifying phishing nodes was validated by extensive experiments using real-world datasets from the Ethereum transaction network. Additionally, the results of the experiments demonstrate that the proposed link prediction method outperformed state-of-the-art feature-based node classification techniques.

The remainder of the paper is divided into the following sections: The next Section summarises related papers on proposed solutions for identifying phishing using traditional methods and elaborates on phishing identification on Ethereum. Section 3 discusses the proposed model in detail. Section 4 presents the research questions and evaluation criteria used to examine the proposed phishing detection framework. Section 5 contains the complete results of the proposed model's evaluations. Finally, Section 6 concludes the paper and discusses future work.

2 Related Works

Most state-of-the-art approaches to detect phishing transactions on Ethereum use graph embedding techniques. Graph modelling techniques have been applied in many domains; the blockchain ecosystem is not left behind. Zhuang et al. [14] designed a graph classification algorithm to model semantic structures within smart contracts and detect inherent vulnerabilities. Liu et al. [5] proposed a GCN-based blockchain address classifier using graph analytics and an identity inference approach.

On the Ethereum platform, Wu et al. [12] proposed a technique for detecting fraud on the Ethereum platform using a concatenation of the statistical features extracted from the transaction amounts and timestamps and automated features from a novel network-embedding approach called trans2vec for downstream phishing classification.

Wang et al. [10] proposed a transaction subgraph network to identify Ethereum phishing accounts (TSGN). The TSGN, inspired by random walk, uses a weight-mapping mechanism to retain transaction amount information in the original transaction network for downstream network analysis tasks. 1621 transaction networks centred on phishing nodes and 1641 transaction networks centred on normal nodes were expanded into subgraphs using the proposed TSGN

and applied to various graph classification algorithms, such as manual attributes, Graph2Vec, and Diffpool. Based on the deep learning method Diffpool, TSGN achieved the best classification performances of 94.35% and 93.64%, respectively.

Yuan et al. [13] approached phishing identification as a graph classification challenge, enhancing the Graph2Vec approach using line graphs and achieving high performance. The technique proposed by Yuan et al. focuses on structural elements extracted from line graphs, thereby omitting information from the graph direction, which is critical for identifying phishing schemes.

The study by Lin et al. [4] modelled the Ethereum network transaction data as a temporally weighted multi-digraph. These graphs were then applied to a random walk-based model to obtain explicable results regarding the interaction between network transactions for phishing account detection.

3 Methodology

This section elaborates on the architecture of the proposed phishing link prediction framework for Ethereum.

3.1 Problem Definition

Given a directed Multigraph, $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{Y})$ of a transaction network, where $\mathcal{V}$ represents a set of nodes that correlate to the target and source addresses on Ethereum. In this study, the source address is analogous to the address initiating the transaction, while the target address is the recipient. The variable $\mathcal{E}$ corresponds to the transaction relationship between the target and source addresses, where $\mathcal{E} = \{e_{u,w}, u, w \in V, u \neq w\}$. Edge attributes e_u, w include local network-based features a_u, w such as node PageRanking, degree centrality, and betweenness centrality. The label of each transaction in the Ethereum network is $\mathcal{Y}_{A_x} \in \mathbb{R}^{|G| \times |\gamma|}$. On the Ethereum platform, $gamma = 1$ represents a phishing transaction, while $\mathcal{Y} = 0$ represents a legitimate transaction.

3.2 Proposed Phishing Detection Framework

Figure 1 provides an overview of the proposed phishing detection framework, which consists of three core parts: transaction graph construction, extraction of network features, and the link phishing detection classifier.

Transaction Network Construction/Feature Extraction As shown in Figure 1, we first construct a large-scale Ethereum transaction network. The nodes are the network addresses, and the edges are the addresses' intrinsic local network-based characteristics. A transaction has two directions: out and in. The out-transactions of an account transfer Ether from the account to other accounts, and the in-transactions of an account receive Ether from other accounts. Specifically, the proposed model considers the relationship between the

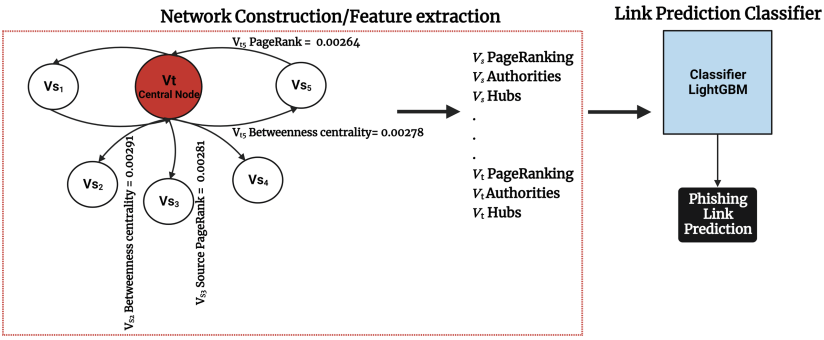


Fig. 1. The Phishing detection framework for transactions on the Ethereum platform.

From transaction address (initiating node) and the *To transaction address* (receiving node) to determine the maliciousness of a transaction.

Subsequently, we extract intrinsic features that link a target address V_t to all its corresponding source addresses V_s. Table 1 details the 12 features extracted. **The PageRank feature**, obtained from both the source and target nodes, ranks the given nodes according to the number of incoming relationships and the importance of the corresponding source nodes. PageRank is essential because it rates each node based on its in-degree (the number of transactions transferred to the node) and out-degree (the number of transactions transferred by the specified node).

The HITS algorithm is one of the essential link analysis algorithms. It produces two primary outcomes: authorities and hubs. In this study, the HITS algorithm calculates the worth of a node by comparing the number of transactions it receives (authorities) and the number of transactions it originates (hubs). As the primary objective of phishing addresses is to obtain as much Ether as possible, and they may not transmit any ether, the value of the Authorities and Hubs will play a crucial part in distinguishing phishing addresses from legitimate ones.

Degree centrality offers a relevance score to each node based on the number of direct, 'one hop' connections it has to other nodes. In an Ethereum network, we assume that legitimate nodes are more likely to have faster connections with nearby nodes and a higher degree of centrality. This assumption is based on the observation that there are likely more legitimate nodes in a given network than phishing nodes.

$$d_v = \frac{deg(v)}{n}, \text{for } v \in V \tag{1}$$

where 'deg(v)' is the degree of node 'v' and 'n' is the number of nodes in set V.

In an Ethereum network, **betweenness centrality** quantifies the frequency with which a node is located on the shortest path between other nodes. Specifically, this metric identifies which nodes are "bridges" connecting other nodes in a network. This is achieved by placing all the shortest paths and then counting

Table 1. Description of features based on the local network

Features	Description
V_s PageRank	Ranking of the source nodes in the graph based on the number of transactions and the importance of the nodes making those transfers.
V_s Authorities	Estimates the source node value based on the incoming transactions.
V_s Hubs	Measures the source node value based on outgoing transactions.
V_s Betweenness centrality	Measures how often a source node V_s appears on the shortest paths between nodes in the network.
V_s Closeness centrality	Measures the average distance from a given source node to all other nodes in the network.
V_s Degree centrality	Measures the fraction of nodes V_s is connected to.
V_t PageRank	Ranking of the target nodes in the graph based on the number of transactions and the importance of the nodes making those transfers.
V_t Authorities	Estimates the value of V_t based on the incoming transactions.
V_t Hubs	Measures the value of V_t based on outgoing transactions.
V_t Betweenness centrality	Measures how often a target node V_t appears on the shortest paths between neighbouring nodes in the network.
V_t Closeness centrality	Measures the average distance from a given target node to all other nodes in the network.
V_t Degree centrality	Measures the fraction of nodes V_t is connected to.

how often each node falls on one. Phishing nodes are more likely to have a low betweenness centrality rating because they may impact the network less.

$$c_B(v) = \sum_{p,q \in V} \frac{\sigma(p, q|v)}{\sigma(p, q)} \tag{2}$$

where V is the set of nodes, $\sigma(p, q)$ is the number of shortest (p, q)-paths, and $\sigma(p, q|v)$ is the number of those paths passing through some node v other than p, q.

Essentially, **closeness centrality** assigns a score to each node based on its "closeness" to every other node in the network. This metric computes the shortest pathways connecting all nodes and provides each node with a score based on the sum of its shortest paths. The analysis of proximity centrality values revealed that network nodes are more likely to impact other nodes rapidly.

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{3}$$

where 'd(v, s)' is the shortest-path distance between 'v' and 's,' and 'n' is the number of nodes in the graph.

The Link Prediction Classifier As stated earlier, the objective of link prediction is to determine the presence of phishing transactions using the intrinsic features of the local network. Subsequently, we employed the LightGBM classifier for the downstream task to detect phishing transactions. Please note that other shallow machine learning classifiers can be used at this stage. However, we chose to use LightGBM because research has shown that it provides a faster training speed and more efficiency compared to other shallow machine learning algorithms [3]. In addition, it utilizes less memory and has a higher degree of precision than all other boosting techniques. They have also been proven to be compatible with larger datasets [6].

4 Research questions and Experimental Setup

This section discusses the research questions, dataset, hyperparameters and metrics used to set up and evaluate the proposed model and its baselines.

4.1 Research questions

- **RQ1:** How accurate is the proposed link prediction model for detecting phishing transactions compared with other time and amount feature-based state-of-the-art approaches?
- **RQ2:** What are the technical alternatives to the proposed link prediction model, and how effective are they?
- **RQ3:** How important are the features used in the proposed link prediction model for detecting phishing transactions between Ethereum addresses?

Data Source/Preprocessing The dataset used in this paper was obtained from the xblock.pro website.[1] It contains 1,262 addresses labelled phishing nodes and 1,262 non-phishing nodes crawled from Etherscan. Each address contains the transaction information between the target node and its corresponding source nodes. Note that transactions exist between a specific target node and multiple source nodes. This observation is not surprising because a single phishing address can receive multiple Ethers from different non-phishing addresses.

Existing studies use only the first node address and the Ether received for graph construction. This research aims to look beyond the first-node address and examine all transaction records carried from and to the addresses. This

[1] http://xblock.pro/#/dataset/6.

approach removes the challenges of a few datasets and demonstrates the importance of studying the connectivity between outgoing and incoming transactions from phishing and non-phishing nodes.

Consequently, 13,146 transactions were extracted from 1,262 phishing addresses and 286,598 from 1,262 legitimate addresses. As it is clear that the number of legitimate transactions is considerably higher than the number of phishing transactions, the synthetic minority oversampling technique (SMOTE) was adopted to address the imbalance in the training set. A random number of minority classes were added until both classes were equally represented. To prevent bias in the results, the instances were normalized in the dataset to appear similar across all records, leading to cohesion and higher data quality.

After oversampling the minority class, our final corpus contained a balanced dataset of **286,598** phishing and benign instances.

Hyperparameter Setting A combination of hyperparameters is required to classify the link prediction model using LightGBM. A grid search was used to determine the optimal hyperparameters of the models by setting the number of estimators to 10,000 and the learning rate to 0.02. In addition, the default value for the number of leaves was set at 31, and the application type was set to binary.

Evaluation Metrics The performance of the link prediction model was evaluated using $Recall = \frac{(TP)}{(TP+FN)}$ and $F1_{score} = \frac{(Precision \times Recall)}{(Precision+Recall)}$ where TP, FP and FN represent the numbers of True Positives, False Positives and False Negatives, respectively. Also, the Area Under the Curve (AUC) score was calculated, representing the degree or measure of separability. A model with a higher AUC is better at predicting True Positives and True Negatives. Finally, to assess the performance of the proposed model and its baseline on the corpus, the dataset was divided into 80% for training and 20% for testing.

5 Results

This section discusses the experiments conducted to evaluate the proposed phishing link prediction method and the results of answering each research question.

5.1 Comparing The Proposed Model with State-Of-The-Art Baselines (RQ1)

To demonstrate a thorough evaluation of our methods, a comparison of the performance of the link prediction model with the existing state-of-the-art feature-based approaches was conducted. These methods include those utilized by Wu et al. [12], who used non-embedding techniques to extract local information from addresses to detect phishing. The time features, amount features, and time plus amount features are among the retrieved features.

Table 2. Result of the proposed model and other state-of-the-art non-embedding models

Models	Recall	F-1 Score	AUC Score
Proposed Model	**0.890**	**0.697**	**0.930**
[12] (Time Features Only)	0.302	0.326	0.835
[12] (Amount Features Only)	0.321	0.358	0.848
[12] (Time + Amount Features)	0.478	0.494	0.865

Result: Table 2 presents the outcomes of the approaches (balancing recall, F1-Score and AUC score). The proposed model demonstrated the best recall performance for this dataset. The results indicate that the proposed method can detect phishing transactions with a satisfactory level of recall and AUC score by utilizing only locally based information collected from analysis of the relationship between the transaction addresses.

The proposed model also performed the best in the F1-score, demonstrating that the phishing class's overall precision and recall performance is robust. In other words, the proposed model not only detects phishing cases accurately but also avoids incorrectly labelling too many legitimate addresses as phishing. This shows that the proposed strategy for phishing strikes a balance between precision and recall. Compared to the other models, the time-features-only model performed the worst, indicating that it could not correctly identify most phishing classes.

Investigating False Positives and False Negatives From the results in Section 5.1, we found that the proposed model inaccurately classified 287 legitimate links as phishing links and 2702 phishing instances were incorrectly classified as legitimate. To investigate false positive links (i.e., legitimate transactions that were wrongly classified as phishing) and false negatives (i.e., phishing transactions that were incorrectly identified as legitimate), we performed a manual analysis on a subset of 100 addresses and their corresponding edges from the false positives and false negatives obtained from the result discussed above.

Our analysis shows that most false-positive and false-negative transactions involve phishing addresses transferring Ether to a legitimate address. This type of transaction is uncommon and only occurs when the phishing address attempts to establish credibility with the legitimate target address. Although this type of transaction is genuine, as the legitimate address duly receives the ether, the model is bound to misclassify it because it originates from a phishing address. Exploring the maliciousness of specific addresses in the Ethereum network and determining their validity will be a top priority for future work.

5.2 Alternative Technical Options for The Proposed Link Prediction Model (RQ2)

The selected shallow machine learning classifier of the detection framework also influences the detection performance. Consequently, this study considers logistic regression, naive Bayes, and decision trees as the baseline classifiers. Using the extracted features as input, Table 3 details the detection outcomes of the three classifiers using the extracted features as input.

Table 3. Result of the proposed model and its alternative options

Models	Recall	F-1 Score	AUC Score
Proposed Model	**0.890**	**0.697**	**0.930**
Logistics Regression	0.838	0.146	0.694
Naive Bayes	0.982	0.106	0.605
Decision Tree	0.865	0.467	0.890

Result: From the results, it is clear that the performance of the predicted model using the LightGBM classifier is superior to that of other classifiers owing to its suitability for the link prediction task. The proposed model produced an average F1 score, a recall rate, and an AUC score of 83%. Across all the evaluative parameters, logistic regression was the alternative option with the lowest performance. This low performance is because logistic regression requires modest or no multicollinearity among independent variables.

5.3 Feature Importance (RQ3)

In addition to our analysis, an investigation of the features that were informative for the classification outcomes of the proposed model was conducted. We employed a sensitivity analysis technique to determine the impact of each feature on categorization output. In sensitivity analysis, the variability of changes in results is determined by the input variability [9]. In this study, the effect of each feature was determined using the one-at-a-time method. This strategy measures the model output statistics for each change in the entry category. The efficiency of each feature is then estimated based on the sensitivity of the classification model.

Result: In Table 4, it is evident that the absence of the target node's closeness centrality and target degree had the most significant impact on the model's declining recall. Eliminating the source node's betweenness centrality and target PageRank had the opposite effect on the link prediction model's recall. With the removal of the source hub, the model's F1-Score and recall are unaffected. Not analyzing the target node's PageRank and betweenness centrality reduced the F-1 score by approximately 0.008. Therefore, removing these features reduced the effectiveness of the model.

Table 4. Result of The Sensitivity Analysis

Features	Recall	F-1 Score
V_s PageRank	0.885	0.696
V_s Authorities	0.884	0.698
V_s Hubs	0.885	0.693
V_s Betweenness centrality	0.887	0.696
V_s Closeness centrality	0.884	0.695
V_s Degree centrality	0.884	0.694
V_t PageRank	0.886	0.688
V_t Authorities	0.883	0.698
V_t Hubs	0.883	0.695
V_t Betweenness centrality	0.883	0.688
V_t Closeness centrality	0.881	0.694
V_t Degree centrality	0.883	0.697

In summary, the most significant characteristics of the proposed model are the target node's (recipients) PageRank, betweenness, and closeness centralities. Eliminating these three features reduces the F1-score by approximately 0.008. This result is not surprising, given that the primary objective of attackers on the Ethereum platform is to coerce victims into sending them ETH.

5.4 Limitations

This study has some limitations. First, the proposed feature sets depend entirely on a specific dataset and may not be easily adapted to another dataset without minor adjustments. Second, network embedding techniques, such as Node2Vec and trans2Vec, might automate the feature extraction process from large-scale network data.

Nonetheless, network-embedded models consume more resources. In addition, unlike our proposed model, which uses features extracted from the local network, network embedding techniques are challenging to explain. Also, timestamps can easily be added to evolve the phishing detection technique into a time-series classification.

6 Conclusion and Future Work

This paper proposes a systematic study for detecting phishing transactions in an Ethereum network using link prediction. Specifically, a three-step approach for identifying the connections between network nodes using extracted local network features were demonstrated. We extracted 12 features based on the influence and relationships between the addresses in the network and used them as inputs for a

LightGBM classifier. Experiments on real-world datasets demonstrated the effectiveness of the proposed link prediction model over existing feature-based state-of-the-art models in detecting phishing transactions. In the future, we intend to conduct further studies on the impact of the proposed link prediction model on other downstream tasks such as gambling, money laundering, and pyramid schemes.

References

1. Chen, W., Guo, X., Chen, Z., Zheng, Z., Lu, Y.: Phishing scam detection on ethereum: Towards financial security for blockchain ecosystem. In: IJCAI, pp. 4506–4512. ACM (2020)
2. Gutierrez, C.N., Kim, T., Della Corte, R., Avery, J., Goldwasser, D., Cinque, M., Bagchi, S.: Learning from the ones that got away: detecting new forms of phishing attacks. IEEE Trans. Dependable Secur. Comput. **15**(6), 988–1001 (2018)
3. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30 (2017)
4. Lin, D., Wu, J., Xuan, Q., Chi, K.T.: Ethereum transaction tracking: inferring evolution of transaction networks via link prediction. Phys. A: Stat. Mech. Its Appl. **600**, 127504 (2022)
5. Liu, X., Tang, Z., Li, P., Guo, S., Fan, X., Zhang, J.: A graph learning based approach for identity inference in dapp platform blockchain. IEEE Trans. Emerg. Top. Comput. (2020)
6. Minastireanu, E.A., Mesnita, G.: Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur (2019)
7. Opara, C., Chen, Y., et al.: Look before you leap: detecting phishing web pages by exploiting raw url and html characteristics. arXiv:2011.04412 (2020)
8. Opara, C., Wei, B., Chen, Y.: Htmlphish: enabling phishing web page detection by applying deep learning techniques on html analysis. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
9. Pannell, D.J.: Sensitivity analysis of normative economic models: theoretical framework and practical strategies. Agric. Econ. **16**(2), 139–152 (1997)
10. Wang, J., Chen, P., Yu, S., Xuan, Q.: Tsgn: Transaction subgraph networks for identifying ethereum phishing accounts. In: International Conference on Blockchain and Trustworthy Systems, pp. 187–200. Springer (2021)
11. Wood, G., et al.: Ethereum: a secure decentralised generalised transaction ledger. Ethereum Proj. Yellow Pap. **151**(2014), 1–32 (2014)
12. Wu, J., Yuan, Q., Lin, D., You, W., Chen, W., Chen, C., Zheng, Z.: Who are the phishers? Phishing scam detection on ethereum via network embedding. IEEE Trans. Syst. Man Cybern.: Syst. (2020)
13. Yuan, Z., Yuan, Q., Wu, J.: Phishing detection on ethereum via learning representation of transaction subgraphs. In: International Conference on Blockchain and Trustworthy Systems, pp. 178–191. Springer (2020)
14. Zhuang, Y., Liu, Z., Qian, P., Liu, Q., Wang, X., He, Q.: Smart contract vulnerability detection using graph neural network. In: IJCAI, pp. 3283–3290 (2020)

An Efficient Deep Learning Framework FPR Detecting and Classifying Depression Using Electroencephalogram Signals

S. U. Aswathy[1(✉)], Bibin Vincent[2], Pramod Mathew Jacob[2], Nisha Aniyan[2], Doney Daniel[2], and Jyothi Thomas[3]

[1] Department of Computer Science and Engineering, Marian Engineering College, Thiruvananthapuram, Kerala, India
aswathy.su@gmail.com

[2] Department of Computer Science and Engineering, Providence College of Engineering, Alappuzha, Kerala, India
{nisha.a,doney.d}@providence.edu.in

[3] Department of Computer Science and Engineering, Christ University, Bangalore, India
j.thomas@christuniversity.in

Abstract. Depression is a common and real clinical disease that has a negative impact on how you feel, how you think, and how you behave. It is a significant burdensome problem. Fortunately, it can also be treated. Feelings of self-pity and a lack of interest in activities you once enjoyed are symptoms of depression. It can cause a variety of serious problems that are real, and it can make it harder for you to work both at home and at work. The main causes include family history, illness, medications, and personality, all of which are linked to electroencephalogram (EEG) signals, which are thought of as the most reliable tools for diagnosing depression because they reflect the state of the human cerebrum's functioning. Deep learning (DL), which has been extensively used in this field, is one of the new emerging technologies that is revolutionizing it. In order to classify depression using EEG signals, this paper presents an efficient deep learning model that allows for the following steps: (a) acquisition of data from the psychiatry department at the Government Medical College in Kozhikode, Kerala, India, totaling 4200 files; (b) preprocessing of these raw EEG signals to avoid line noise without committing filtering; (c) feature extraction using Stacked Denoising Autoevolution; and (d) reference of the signal to estimate true and all. According to experimental findings, The proposed model outperforms other cutting-edge models in a number of ways (accuracy: 0.96, sensitivity: 0.97, specificity: 0.97, detection rate: 0.94).

Keywords: Electroencephalogram · Autoencoder · Classification · Convolutional Neural Network · Depression

1 Introduction

The World Health Organization estimates that more than 322 million people worldwide experience depression, making it the mental disorder that is most responsible for causing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1179–1188, 2023.
https://doi.org/10.1007/978-3-031-27409-1_108

disability worldwide. Patients with depression are frequently identified by symptoms like a sense of sadness, helplessness, or guilt; a lack of interest or energy; changes to one's appetite, sleeping habits, or daily routines.

Numerous factors, including poverty, unemployment, traumatic life events, physical illnesses, and problems with alcohol or drug use, are thought to be the root causes of depression. Additional primary causes of depression are believed to include recent occurrences like the Covid 19 pandemic and its effects, including lockdowns, quarantine, and social seclusion. Given that depression has never before threatened public health, has some detrimental effects on depressed people, including suicide, and that prompt and more effective treatment can be obtained with early diagnosis, It is imperative to create an efficient and trustworthy method of identifying or even anticipating depression [6, 7].

EEG signals, which by nature are nonstationary, extremely complex, non-invasive, and nonlinear, reflect the state and function of the human brain. Due to this complexity, it would be challenging to see any abnormality with the naked eye. These traits have caused physiological signals to be seen as practical tools for the early detection of depression [8]. Deep learning is defined as a hierarchy of algorithms that includes a subset of hidden neurons. These models enable computers to create complex concepts out of simple statements. The following layers are built using the learned concepts. Furthermore, pattern and data structure recognition in these methods is carried out by multiple processing layers.

Recent applications of this multi-layer approach span a variety of industries, from the automotive industry, IoT, and agriculture to diverse applications in medicine. Deep learning solutions are increasingly being used in related contexts as a result of the challenges associated with manually analysing EEG signals, the limitations of machine learning techniques, and deep learning architecture's capacity to automate learning and feature extraction from input raw data [9–11]. These methods enable the quickest extraction of implicit nonlinear features from EEG signals. This study presents a useful DL model for detecting depression from EEG signals.

1.1 Key Highlights

The following are the objectives of this paper's effective deep learning for classifying depression.

Create a deep learning model that is effective at classifying depression from EEG signals by using data from the real-time Kozhikode, Kerala repository of hospital. This training produces results that are satisfactory.

- Using auto encoder, a CNN-based network variant, to extract features.
- Using T-RFE, feature vectors have been created using those three EEG signals.
- Using 3D CNN classification, effective depression and non-depression detection is achieved.

2 Literature Review

A novel computer model for EEG-based depression screening is presented by Acharya et al. [1] using convolutional neural networks (CNN), a type of deep learning technique. The suggested classification method does not call for feeding a classifier with a set of semi-manually chosen features. It automatically and adaptively distinguishes between the EEGs obtained from depressed and healthy subjects using the input EEG signals. 15 patients with depression and 15 patients with normal EEGs were used to test the model. Using EEG signals from the left and right hemispheres, the algorithm had accuracy of 93.5% and 96.0%, respectively. The deep model put forth by Thoduparambil et al. [2] incorporates Convolution Neural Network and Long Short Term Memory (LSTM). And employed to find depression. To learn the local characteristics and the EEG signal sequence, CNN and LSTM are used, respectively. Filters and the input signal are convolved to produce feature maps in the convolution layer of the deep learning model. After the LSTM has learned the various patterns in the signal using all the extracted features, fully connected layers are then used to perform the classification. The memory cells in the LSTM allow it to remember the crucial details over time. Additionally, it has a variety of features for changing the weights while working out. Han et al. [3] built a psychophysiological database with 213 subjects (91 depressed patients and 121 healthy controls). A pervasive prefrontal lobe three-electrode EEG system was used to record the electroencephalogram (EEG) signals of all participants while they were at rest using the Fp1, Fp2, and Fpz electrode sites. 270 linear and nonlinear features were extracted after denoising with the Finite Impulse Response filter, which incorporates the Kalman derivation formula, Discrete Wavelet Transformation, and an Adaptive Predictor Filter. The feature space was then made less dimensional using the minimal-redundancy-maximum-relevance feature selection method. The depressed participants were separated from the healthy controls using four classification techniques (Support Vector Machine, K-Nearest Neighbor, Classification Trees, and Artificial Neural Network). A computer-aided detection (CAD) system based on convolutional neural networks was proposed by Li et al. [4]. (ConvNet). However, the local database should serve as the cornerstone for the CAD system used in clinical practise, so transfer learning was used to build the ConvNet architecture, which was created through trial and error. They also looked at the function of different EEG features, In order to identify mild depression, spectral, spatial, and temporal information is used. They found that the EEG's temporal information significantly improved accuracy and that its spectral information played a significant role. DepHNN (Depression Hybrid Neural Network), Sharma et al. present a novel EEG-based CAD hybrid neural network for depression screening in 2021 [5]. The suggested approach makes use of windowing, LSTM architectures, and CNN for temporal learning and sequence learning, respectively. Neuroscan was used to collect EEG signals from 21 drug-free, symptomatic depressed patients and 24 healthy people for this model. The windowing technique is used by the model to accelerate computations and reduce their complexity.

3 Methodology

Figure 1 depicts the overall architecture of the suggested framework. The Department of Psychiatry at the Government Medical College in Kozhikode and Kerala, India, collected EEG signals from participants (aged 20–50) and stored them in a real-time repository for data collection. This is sufficient for training and testing a better deep learning model. 15 of the participants were healthy, and 15 had depression. The dataset's use in this study received approval from a senior medical ethics panel. Additionally, the same written consent was given by every subject. The EEG signals were produced by the brain's bipolar channels FP2-T4 (in the right half) and FP1-T3 (in the left half). While at rest for five minutes with their eyes open and closed, each subject provided data. 256 Hz was used as the sampling rate for the EEG signals. With the aid of a notch filter, 50 Hz power line interference was eliminated. The dataset contained 4200 files from each of 15 depressed and 15 healthy individuals. There were 2000 sampling points per file. Following the collection of the data, the raw signals may need to have noise and other artefacts removed before moving on to the next stage. b) Bad channels were identified and eliminated during the preprocessing stage because many algorithms will fail in the presence of appallingly bad signals egregiously bad signals [12–14]. There is a complex relationship between bad channels and referencing, as will be discussed below. The overall goals of this stage are to I eliminate line noise without committing to a filtering strategy, (ii) robustly reference the signal in relation to an estimate of the "true" average reference, (iii) identify and interpolate bad channels in relation to this reference, and (iv) retain enough data to allow users to re-reference using a different method or to undo the interpolation of a specific channel.

Once these signals have undergone preprocessing, they are then passed on to c) feature extraction, where sufficient features are extracted with the aid of Stacking Denoising Autoencoder (SDAE) [15]. An artificial neural network architecture called a stacked autoencoder consists of several autoencoders and is trained using greedy layer-wise training. The middle layer, the output layer, and the input layer are all included in each autoencoder. In the stacked autoencoder, the middle layer's output serves as the next autoencoder's input. The stacked autoencoder is extended by the SDAE. SDAE's input signals are tainted by noise. In this study, a quick model of SDAE with two autoencoders is used to decode and recover the blurred original input EEG X= [X1, X2, Xk] from noise [16, 17]. The frequency bands alpha (8–12 Hz), beta (12–30 Hz), theta (4–8 Hz), and delta were used to separate the signals (0.5–4 Hz). Higuchi Fractal Dimension (HFD), correlation dimension (CD), approximate entropy (EN), Lyapunov expo-nent (LE), and detrended fluctuation analysis were some of the features that were extracted (DFA). These characteristics were extracted from each frequency band in order to obtain a total of 24 parameters for each subject. Based on topographical brain regions, the features were compared and averaged over designated channels. After features have been extracted, step d) feature selection uses transform-recursive feature elimination to reduce dimensionality (T-RFE). LSSVM, a fast-training variant of SVM known as the least square

The T-RFE algorithm is implemented using a support vector machine in order to lower the high computational cost. Additionally, due to the low risk of overfitting, the linear LSSVM based EEG feature selection and classification approach in our prior work

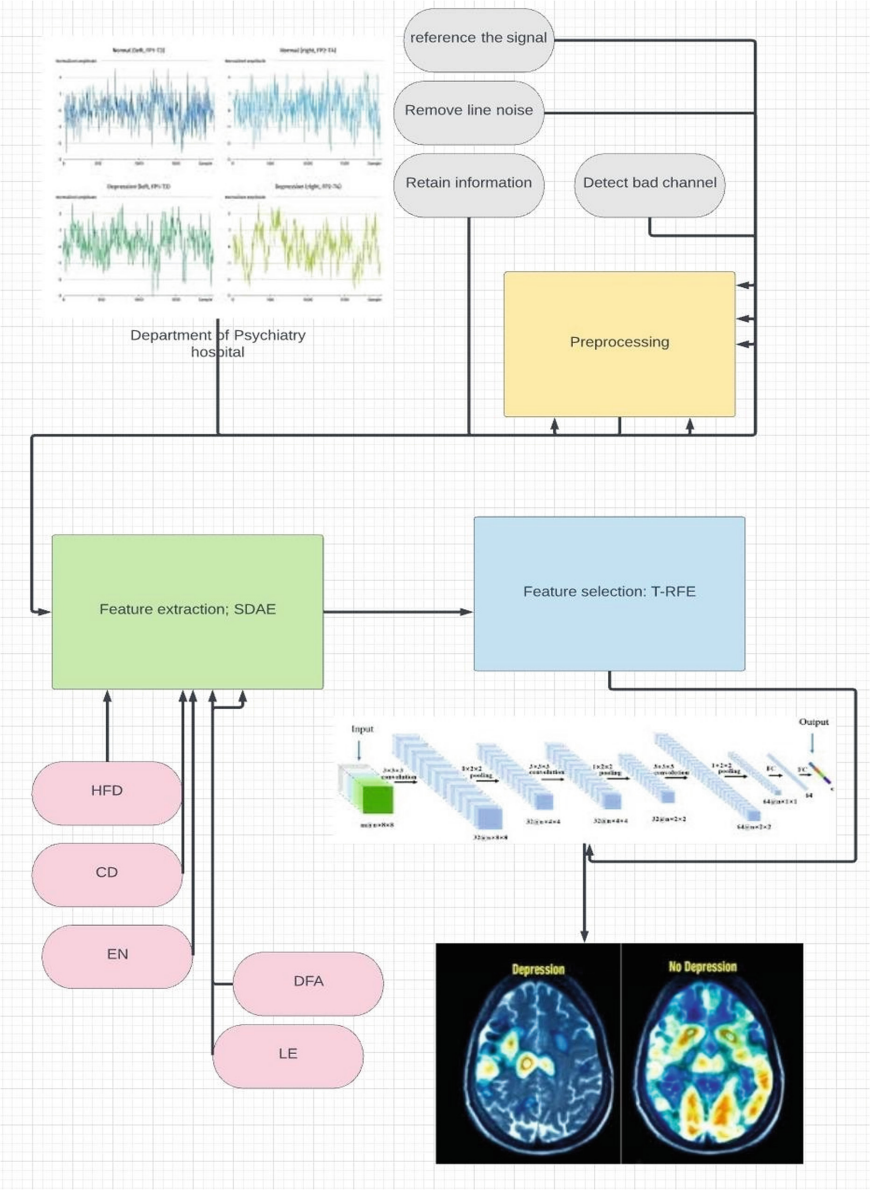


Fig. 1. Overall architecture of proposed framework

has demonstrated better performance than its nonlinear form. These feature vectors are finally provided to the 3D CNN for classification [18, 19]. The 6 6 64 partial direct coherence (PDC) matrices, which are the input of the 3D CNN, represent the EEG signals' connectivity. The PDC matrices are computed using equation (5), i.e., $f = 0.625b$, where $b = 1, 2, \ldots, 64$, over six DMN channels at each (40/64)-Hz frequency bin. A 3D

CNN will be used to classify depression from the EEG signal in comparison to a healthy control (HC) given the 3D PDC input. Our recommended 3D One fully connected layer, a global average pooling layer, three dropout layers, three rectified linear unit (ReLU) activation layers, three batch normalization (BN), and three convolutional layers make up the overall architecture of CNN. A nonlinear activation function follows each layer of convolution (ReLU).

The model is implemented using pytorch, an open source Python library for building deep learning models, and Google Collaboratory, an open source Google environment for developing deep learning models. Hardware requirements include Ryzen 5/6 series processors, 1TB HDDs, NV GPUs, and Windows 10 OS. The proposed model is compared to a number of other models, including VGG16, VGG19, Resnet50, Googlenet, Inception v3, ANN, Alexnet, and standard CNN, on a number of different metrics, including accuracy, sensitivity, specificity, recall, recall rate, precision, F1-score, detection rate, TPR, FPR, AUC, and computation time.

4 Result

To evaluate the effectiveness of a machine learning classification algorithm, confusion matrix is used. Figure 2 gives the confusion matrix. From the result we can see the following details in confusion Matrix: True Positives, True Negatives, False Positives False Negatives. It is displayed as a matrix. This comparison between actual and expected values is provided. We receive a 3×3 confusion matrix for classes of 3. We can assess the model's performance using metrics like recall, precision, accuracy, and AUC-ROC curve.

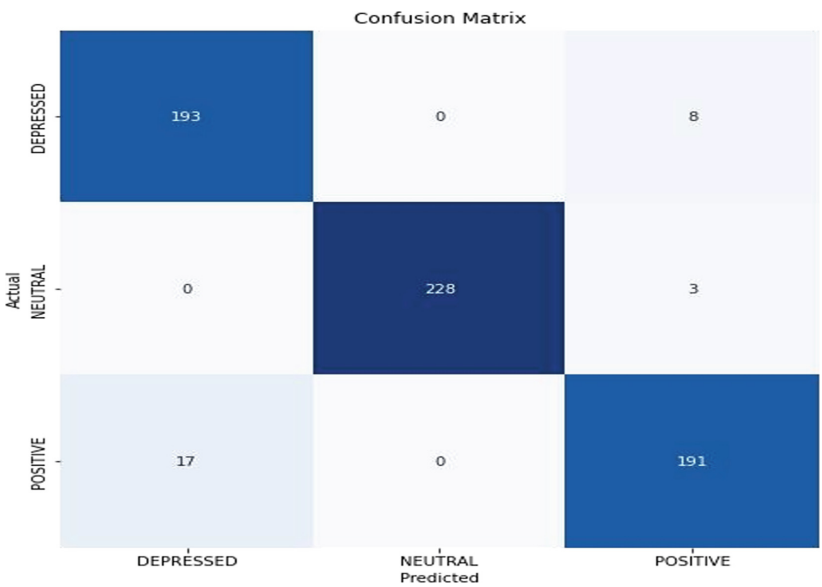


Fig. 2. Confusion Matrix of the proposed Method

Let's figure out the depressed class's TP, TN, FP, and FN values.

TP: It should be the same for both the actual and predicted values. Thus, cell 7's value is the TP value for the depressed class. 17 is the Value.

FN: the total value of the corresponding row, minus the TP value. (Cell 8 + Cell 9) = FN. It is equal to 0 + 191 = 191.

FP: the total of the values in the relevant column, excluding the TP value. (Cell 1 + Cell 4) = FP. 193+0 = 193 is the value.

TN = (cell 2 + cell 3 + cell 5 + cell 6); FN: the sum of values of all columns and rows excluding those for which we are calculating the values. The sum is 239, which is 0+8+228+3.

Similar calculations are made for the neutral class and the results are as follows:

TP: 228(cell 5) (cell 5)

FN: (cells 4 and 6) 0+1=1

FP: (cell2 + cell 8) 0+0 = 0.

TN: 193+8+17+191 = 409 (cells 1–3, 7–9 total).

Similarly, the value/matrix for the positive class is calculated as follows:

TP: 191 (cell 9) (cell 9)

FN: (cells 7 and 8) = 17+0 = 1

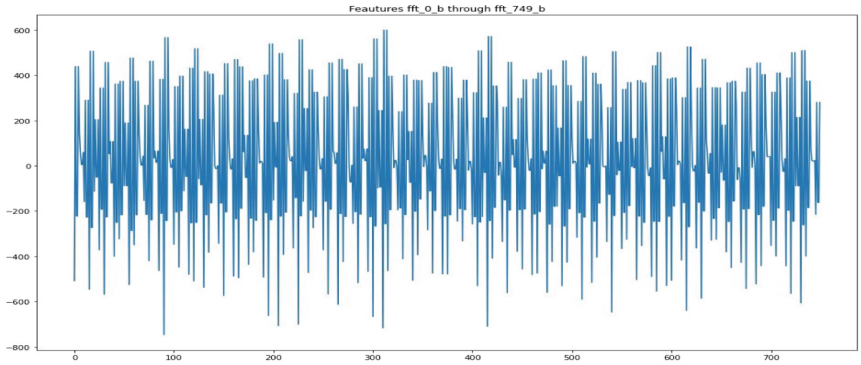
FP: (cells 3 and 6) 8+3 = 11

TN: 193+0+0+228 = 409 (cells 1–2, 3–4, and 5).

These are the data gathered from the confusion matrix mentioned above (Fig. 3).

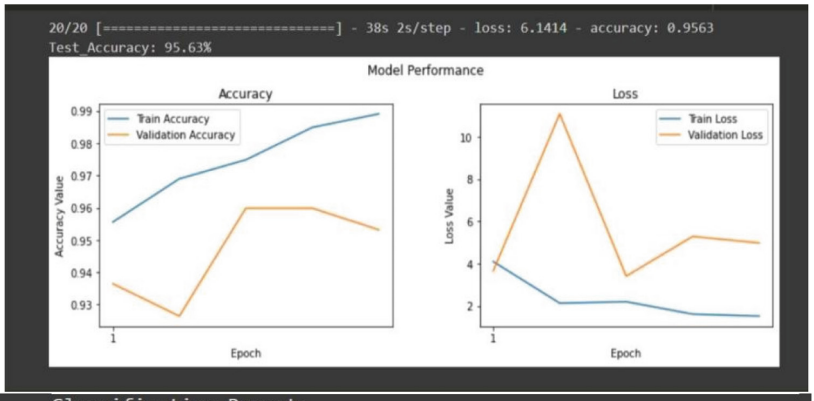
5 Conclusion

This study concentrated on identifying and predicting depression using EEG signals and deep learning algorithms. According to the SLR method, which was employed in this study, a thorough review was carried out, in which some studies that were specifically focused on the subject were evaluated and had their key elements examined. Discussion also includes open questions and potential future research directions. Given our goals and the fact that most articles compared the outcomes of two or more deep learning algorithms on the same prepared dataset, the taxonomy was created by combining all deep learning techniques used in all studies. It was discovered after analysing 22 articles that were the result of a thorough, elaborate, SLR-based refinement that CNN-based deep learning methods, specifically CNN, 1DCNN, 2DCNN, and 3DCNN, are by far the more preferable group among the various adopted algorithms, accounting for almost 50% of the total in sum. With approximately one-third of the total, CNN won this classification. Only the CNN- based category outperformed the combined models of the two LSTM blocks and CNN-based algorithms mentioned earlier in this sentence. Additionally, it was found that different researchers used different feature extraction methods to create models for AQ3 that were more appropriate. The majority of papers utilising these techniques aimed to extract local features end-to-end using convolutional layers. The analysis shows that all studies gather EEG signals, clean them of noise and artefacts, extract the necessary features from the pre-processed signals, and then use one or more deep learning techniques to categorise depressive and healthy subjects. In conclusion,




```
[34] data['label'].value_counts()

NEUTRAL      716
DEPRESSED    708
POSITIVE     708
Name: label, dtype: int64
```




```
Classification Report:
----------------------
              precision    recall  f1-score   support

   DEPRESSED       0.92      0.96      0.94       201
     NEUTRAL       1.00      0.99      0.99       231
    POSITIVE       0.95      0.92      0.93       208

    accuracy                           0.96       640
   macro avg       0.95      0.96      0.95       640
weighted avg       0.96      0.96      0.96       640
```

Fig. 3. Gives the experimental results of the proposed method

In accordance with our objectives, we aimed to present a thorough analysis of the SLR method and in relation to the SLR method in order to provide future research with a strong foundation.

References

1. Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adeli, H., Subha, D.P: Automated EEG-based screening of depression using deep convolutional neural network. Comput. Methods Prog. Biomed. **161**, 103–113 (2018)
2. Thoduparambil, P.P., Dominic, A., Varghese, S.: M:EEG-based deep learning model for the automatic detection of clinical depression. Phys. Eng. Sci. Med. **43**(4), 1349–1360 (2020)
3. Dhas, G.G.D., Kumar, S.S.: A survey on detection of brain tumor from MRI brain images. In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), July, pp. 871–877. IEEE (2014)
4. Cai, H., Han, J., Chen, Y., Sha, X., Wang, Z., Hu, B., Gutknecht, J.: A pervasive approach to EEG-based depression detection. Complexity (2018)
5. Li, X., La, R., Wang, Y., Niu, J., Zeng, S., Sun, S., Zhu, J.: EEG-based mild depression recognition using convolutional neural network. Med. Biol. Eng. Comput. **57**(6), 1341–1352 (2019)
6. Sharma, G., Parashar, A., Joshi, A.M.: DepHNN: a novel hybrid neural network for electroencephalogram (EEG)-based screening of depression. Biomed. Signal Process. Contr. **66**, 102393 (2021)
7. Ahmadlou, M., Adeli, H., Adeli, A.: Fractality analysis of frontal brain in major depressive disorder. Int. J. Psychophysiol. **5**(2), 206–211 (2012)
8. Aswathy, S.U., Dhas, G.G.D. and Kumar, S.S., 2015. Quick detection of brain tumor using a combination of EM and level set method. Indian J. Sci. Technol. **8**(34)
9. Geng, H., Chen, J., Chuan-Peng, H., Jin, J., Chan, R.C.K., Li, Y.: Promoting computational psychiatry in China. Nat. Hum. Behav. **6**(5), 615–617 (2022)
10. Puthankattil, S.D., Joseph, P.K.: Classification of EEG signals in normal and depression conditions by ANN using RWE and signal entropy. J. Mech. Med. Biol. **12**(4), 1240019 (2012)
11. Stephen, D., Vincent, B., Prajoon, P.: A hybrid feature extraction method using sealion optimization for meningioma detection from MRI brain image. In: International Conference on Innovations in Bio-Inspired Computing and Applications, December, pp. 32–41. Springer, Cham (2021)
12. Hosseinifard, B., Moradi, M.H., Rostami, R.: Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. Comput. Methods Progr. Biomed. **109**(3), 39–45 (2013)
13. Bairy, G.M., Bhat, S., Eugene, L.W., Niranjan, U.C., Puthankatti, S.D., Joseph, P.K.: Automated classification of depression electroencephalographic signals using discrete cosine transform and nonlinear dynamics. J. Med. Imag. Hlth Inf. **5**(3), 635–640 (2015)
14. Acharya, U.R., Sudarshan, V.K., Adeli, H., Santhosh, J., Koh, J.E., Puthankatti, S.D.: A novel depression diagnosis index using nonlinear features in EEG signals. Eur. Neurol. **74**(1–2), 79–83 (2015)
15. Aswathy, S.U., Abraham, A.: A Review on state-of-the-art techniques for image segmentation and classification for brain MR images. Curr. Med. Imag. (2022)
16. Mumtaz, W., Qayyum, A.: A deep learning framework for automatic diagnosis of unipolar depression. Int. J. Med. Inf. **132**, 103983 (2019)

17. Liao, S.C., Wu, C.T., Huang, H.C., Cheng, W.T., Liu, Y.H.: Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns. Sensors (Basel). 14,17(6), 1385 (2017)
18. Wan, Z.J., Zhang, H., Huang, J.J., Zhou, H.Y., Yang, J., Zhong, N.: Single-channel EEG-based machine learning method for prescreening major depressive disorder. Int. J. Inf. Tech. Decis. 18(5), 1579–603 (2019)
19. Duan, L., Duan, H., Qiao, Y., Sha, S., Qi, S., Zhang, X.: Machine learning approaches for MDD detection and emotion decoding using EEG signals. Front. Hum. Neurosci. 14–284 (2020)



Comparative Study of Compact Descriptors for Vector Map Protection

A. S. Asanov[1], Y. D. Vybornova[1], and V. A. Fedoseev[1,2(✉)]

[1] Samara National Research University, Moskovskoe Shosse, 34, 443086 Samara, Russia
asanov.as@ssau.ru, vicanfed@gmail.com
[2] IPSI RAS—Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardeyskaya 151, Samara 443001, Russia

Abstract. The paper is devoted to the study of compact vector map descriptors to be used as a zero watermark for cartographic data protection, namely, copyright protection and protection against unauthorized tampering. The efficiency of the investigated descriptors in relation to these problems is determined by the resistance of their values to map transformations (adding, deleting vertices and objects, map rotation, etc.). All the descriptors are based on the use of the Ramer-Douglas-Peucker algorithm that extracts the significant part of the polygonal object determining its shape. The conducted study has revealed the preferred descriptor for solving the copyright protection problem, as well as several combinations of other descriptors identifying certain types of tampering. In addition, a modification of the Ramer-Douglas-Peucker algorithm, which is more efficient than the basic algorithm, is proposed.

Keywords: Zero watermarking · Vector map protection · GIS · Ramer-Douglas-Peucker · Compact descriptor

1 Introduction

Today's digital economy widely applies cartographic data, which are mainly stored and processed in geographic information systems (GIS) [1], as well as published through specialized web services (public cadastral map of Rosreestr, 2GIS, Yandex.Maps, etc.). Creating and updating thematic digital maps of a certain area is a time-consuming task. The most frequently used data sources for its solution are paper maps, satellite images, outdated digital maps, adjacent digital maps of another thematic category, and open vector data of unconfirmed reliability (for example, OpenStreetMap). Despite the development of technologies for automating routine operations, in particular artificial intelligence methods, the creation of a digital map is still largely done manually by experienced cartographic engineers. This is due to the complexity of the task of combining heterogeneous, underdetermined, and also often contradictory data. Therefore, the value of some types of digital cartographic data is very high, which makes the problems of protection of these data urgent [2–6]. The increased volume of vector data, to which access (open or by subscription) is provided by web-mapping services to the broad masses of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1189–1198, 2023.
https://doi.org/10.1007/978-3-031-27409-1_109

users, adds to the urgency of the protection issues. 10–15 years ago such services were few, and public access was available only for viewing rasterized vector data using the Tile Map Service (TMS) protocol [7]. Now many services provide access to vector data using Mapbox Vector Tiles technology [8] or allow users to download data in KML or geoJSON.

The main problems of vector map protection, as well as other data in general, are copyright protection and protection against unauthorized modification [9, 10]. The first one is aimed to confirm the rights of the author or a legal owner of the data in case of theft. The second problem is aimed to detect a forged map or map fragment. Cryptographic means (i.e., digital signatures) [9, 11], as well as digital watermarks [2, 3] are mainly used to solve these problems.

In this paper, we focused on the use of a special approach to vector map protection within the framework of digital watermarking technology—the so-called *zero watermarking* [12–14]. This approach resembles a hashing procedure: some identifying information (called a *descriptor*) is computed for the protected map and then stored separately in a publicly accessible database. At the "extraction" stage, the descriptor is recomputed for a potentially forged map. After that, the obtained descriptor is compared to the original one queried from the database. For example, in [6] the feature vertex distance ratio (FVDR) is calculated and then combined with digital watermark bits for greater secrecy. In [15] the digital watermark is constructed based on triangulation and the calculation of local features within each triangle.

In practice, depending on the problem of data protection being solved, the zero watermark must either be resistant to all transformations of the vector map (for copyright protection) or must be destroyed under certain transformations and thus signal unauthorized changes.

The goal of this study is to determine which characteristics of a vector map are useful for zero watermark formation, depending on the problem and types of map tampering to be detected with this digital watermark.

For example, a digital watermark based on the ratio of distances between feature vertices of polygons will theoretically make it possible to detect only the addition and removal of vertices not included in the set of feature points. Also, a digital watermark based on the average distance from the center of the bounding box to each vertex lets us detect also the addition and removal of feature points. Based on the results of this study, we can make recommendations for map descriptor selection, depending on the specifics and conditions of the use of map data.

2 The Ramer-Douglas-Peucker Algorithm and Its Modification

2.1 Description of Algorithms

In all the descriptors studied in this paper, the Ramer-Douglas-Peucker algorithm [16, 17], is used as an integral part. This algorithm is aimed to reduce the complexity of vector polygonal objects by reducing their point number. The use of this algorithm in descriptors used for zero watermarking theoretically should provide robustness to small changes in map objects. Below we consider this algorithm in more detail, as well as

its modification developed to eliminate the disadvantages that appear when using this algorithm in descriptors.

The input of the Ramer-Douglas-Peucker algorithm is the most distant vertices in the object (points A and B in Fig. 1). Next, the algorithm finds the vertex farthest from the segment that connects the vertices selected in the first step (point C in Fig. 1). Then the ratio of the distance from a point to this segment to the length of the segment itself is calculated (the ratio of CD to AB in Fig. 1). If the obtained value is less than a predefined threshold: $CD/AB < \alpha$, then all previously unmarked vertices are considered as non-feature and can be discarded from the point set of the optimized object. If $CD/AB \geq \alpha$, then the algorithm recursively calls itself for two new segments, CA and CB. The result of the algorithm is an object consisting only of feature vertices.

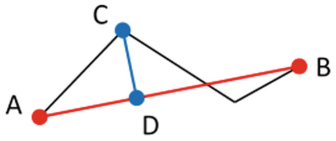


Fig. 1. Illustration of the Ramer-Douglas-Peucker algorithm

The disadvantage of this algorithm in relation to the considered problem of constructing informative descriptors is the complexity of selecting α so that the algorithm results in a sufficient number of feature points to preserve the shape of the object. This problem appeared in practical tests and is investigated in detail in the comparative study described in the next subsection.

To eliminate this drawback, we decided to introduce a small modification to the algorithm. At the first iteration, C is always recognized as a feature point. Then at further iterations, we calculate the threshold value as $\alpha CD/AB$. In this relation, AB and CD are segments of the next object, at the first iteration of the algorithm. The such modification increases the robustness of the feature point set to insignificant map modifications of the map. This fact was confirmed in our experimental study (see Sect. 2.2).

2.2 Comparative Study of the Original and Modified Algorithm

In order to compare the original and modified Ramer-Douglas-Peucker algorithm, we implemented a study with the following scenario:

1. The redundancy reduction algorithm is applied to the original map, and the set of feature vertices is stored.
2. From 10 to 90% of non-feature vertices are added to each object of the original map.
3. The redundancy reduction algorithm is applied to the modified map. Ideally, its result should be equivalent to the one obtained in Step 1.
4. The error is found as the sum of erroneously deleted and erroneously retained vertices divided by the total number of vertices in the map.

The experiment was repeated for different α, different fractions of added vertices (from 10 to 90%) and different versions of the algorithm. In our experiments, we used

an urban building map with introduced correction of absolute coordinates and cleared from semantic data (see Fig. 2). This map contains 4996 polygonal objects.



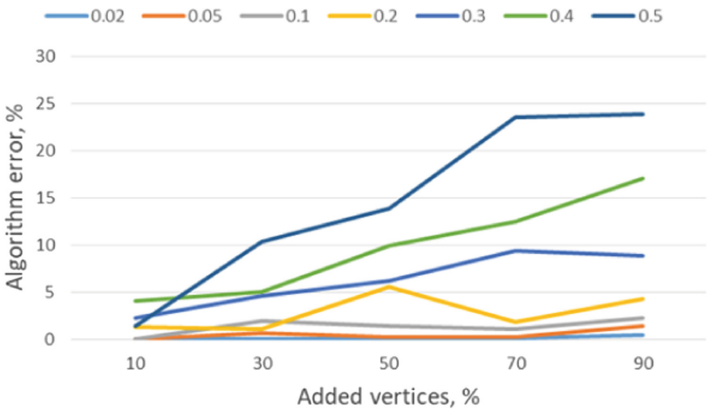
Fig. 2. Fragment of a test map used in the experiments

The results of the experiment are shown in Fig. 3. As can be seen from the graphs, the modified algorithm has higher accuracy than the original one. It should also be noted that the error is less than 1% for small α, so in further experiments we used values $\alpha = 0.02$ (the primary option) and $\alpha = 0.1$ (to increase the speed of calculation).

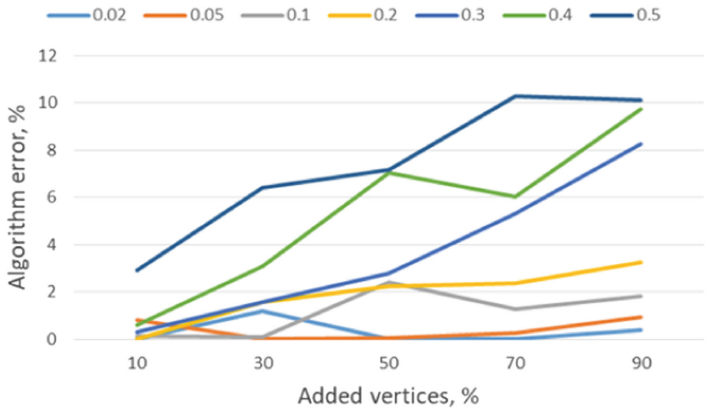
3 Description of the Compact Descriptors to be Analyzed

In this paper, we call *a compact descriptor* some numerical value (a real number or a vector of real numbers) that characterizes an area of the map containing an indefinite number of polygonal objects. We do not focus on any of the two data protection problems described in Sect. 1 when selecting a set of descriptors and analyzing them. Obviously, descriptors suitable for problem 1 will be ineffective for problem 2. Descriptors representing the first group must be robust to various vector map modifications (their range is not infinite in practice, but is determined by the specifics of data use), while those representing the second group must be fragile to the distortions that need to be detected. Therefore, we investigated a wide range of descriptors in order to make recommendations for both problems:

1. Average ratio of distances between feature vertices.
   In each object, the distances between all adjacent points are calculated. Then the ratio of these distances between pairs of adjacent segments is found, for normalization the smaller segment is always divided by the larger one, regardless of the order.



a)



b)

Fig. 3. Dependence of the algorithm error on the percentage of vertices added with different α in the Ramer-Douglas-Peucker algorithm (**a**) and its modification (**b**)

The ratios are summed and divided by the number of segments:

$$d = \frac{2}{N} \sum_{i=0}^{N/2-1} \min\left(\frac{r_{2i}}{r_{2i+1}}, \frac{r_{2i+1}}{r_{2i}}\right).$$

where N is the number of segments of one object, $r_i, i = 0 \ldots N-1$—segment lengths.

This equation specifies the way to calculate certain measures in each object. The descriptor of the fragment is the average value d among all objects.

2. Average ratio of the bounding boxes areas.

   For each object in a map fragment, the area of the bounding box is calculated, then the pairwise ratio of these areas is found, always smaller to larger, irrespective of order. The ratios are summed and divided by the number of objects in the fragment.

3. Average distance between the centers of masses of objects within a group.

Initially, the distances between the centers of masses of the objects are calculated. Then they are divided by each other in the ratio smaller to larger, regardless of the order, summed up and divided by the number of objects in the fragment.

4. Average ratio of the number of feature vertices within a group.

   When calculating this descriptor, the ratio of the number of vertices in the objects to each other is found, then all the ratios are summed up and divided by the number of objects.

5. The average ratio of the distances from the center of mass to the upper right corner of the bounding box.

   The distance from the center of mass to the upper right corner of the bounding box is calculated for each object in the map area. Then the pairwise ratios of distances are found, summed and divided by the number of objects. Similarly to the previous ones, the ratio of values is smaller to larger.

6. The average ratio of the distances from the center of the bounding box to each vertex.

   Each object has an average distance from the center of the bounding rectangle to each vertex, then the distances on the map section are divided into pairs in the ratio of lesser to greater, and the average of these ratios is calculated.

   As one can see, all these descriptors do not depend on the map scale, and their values are in the range from 0 to 1. Before calculating these descriptors, each map object should be optimized by the modified Ramer-Douglas-Peucker algorithm.

   It should also be noted that in practice when detecting distortions of a digital map, it is of considerable interest to know which part of the map has undergone changes. To be able to localize changes using descriptors, the following approach was used. The original map was divided into equal square areas. At each of them, a compact descriptor was calculated, taking into account the characteristics of all polygonal objects in the given area. When comparing the descriptors, the areas were considered separately, which allows to carry out the localization of changes.

4 Experimental Investigation

4.1 Map Transformations

As part of our work, a series of experiments were conducted to investigate the robustness of the selected compact descriptors. In these experiments, the map was divided into equal sections, and then the above descriptors were calculated for each of them. Next, the map distortion procedure was performed. Both the type of distortion and its level, determined by a scalar parameter, were changed. Next, the descriptors were also calculated on the distorted map, and the relative change of the descriptor was estimated. The descriptor was considered robust to a certain distortion if the relative error was less than 1% for all values of the parameter.

We used the distortions listed below (their parameters are specified in parentheses):

1. Map rotation (angle from 0 to 360°).
2. Adding vertices without changing object shape (fraction from 10 to 100%).
3. Adding non-feature vertices that change object shape (fraction from 10 to 100%).
4. Removal of arbitrary vertices (fraction from 5 to 40%).

5. Removal of non-feature vertices (fraction from 5 to 40%).
6. Changing the order of vertices—cyclic shift (number of points).
7. Adding copies of existing map objects (fraction from 10 to 100%).
8. Adding new four-point map objects (fraction from 10 to 100%).
9. Random object deletion (fraction from 10 to 90%).

4.2 Summary of the Obtained Results

The results of the series of experiments are shown in Table 1. It uses the following notations: "+−" means that the descriptor is robust to the given distortion on the whole set of parameter values, "−" means fragility, "+−" means robustness at a subset of parameter values. Finally "!" means that the robustness changes very chaotically and one cannot reliably predict either its robustness or fragility for different maps.

Table 1. Summary table on the robustness of the studied descriptors.

Distortion/descriptor	1	2	3	4	5	6	7	8	9
1	+	+	!	−	+	+	−	−	+−
2	+−	+	+	+−	+	+	+−	+−	+−
3	+	+	!	−	+	+	−	−	−
4	+	+	!	−	+	+	−	−	−
5	−	+	!	+−	+	+	−	−	+−
6	+−	+	!	−	+	+	−	−	−

As one can see from the table, the most robust among the studied descriptors is descriptor 2. This fact means that it is the most effective descriptor for copyright protection. The other descriptors are robust to only certain distortions, so they can be used to detect those types of distortions to which they are fragile. One way to detect a particular kind of distortion is to combine several kinds of descriptors that differ in just one distortion. Here are a few examples:

- We can use descriptors 4–5 to detect map rotation (distortion 1). If the descriptor 4 value compared to the previously stored value is not changed, unlike the descriptor 5 value, then only rotation could happen to the map.
- We can use descriptors 1 and 4 to detect the removal of a small number of objects (distortion 9), because this is the only distortion for which these descriptors give different results.
- Descriptor 5 can be used in combination with descriptor 1 to detect distortion 4 (removing object vertices).

It should be noted that when adding vertices that do not change the shape of the object and removing non-feature vertices, all descriptors were stable only due to the use of the Ramer-Douglas-Peucker algorithm. Otherwise, only descriptor 2 would be stable to these types of distortions.

4.3 Detailed Results for Descriptor 2

Let us focus in more detail on the results shown by descriptor 2 and summarized in Table 1. This descriptor turned out to be robust to adding vertices with and without changing the object shape, shifting and removing non-feature vertices, and rotating the map by 90, 180, and 270 degrees. The graph in Fig. 4 shows that there is a dependence of feature deviation on rotation angle: the further the angle value is from 90° and its multiples, the greater the difference between the feature and the original one.

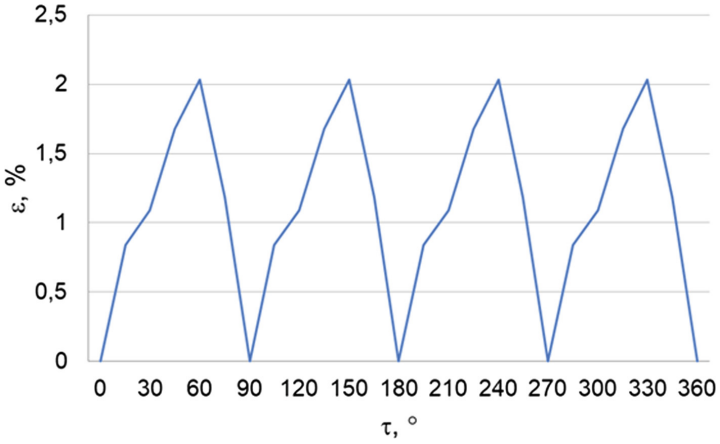


Fig. 4. Effect of map rotation angle (distortion 1) on the relative change in the value of descriptor 2

When we remove arbitrary vertices, add new objects or copies of existing objects, and remove objects, the value of the descriptor changes, but there is a clear dependence on the percentage of distortion, which is reflected in Figs. 5 and 6. Therefore, firstly, for small deviations, the descriptor can be correlated with the original value, and secondly, this nature of the graphs in the presence of a priori information about the nature of distortions allows you to estimate the level of distortion introduced.

5 Conclusion

A series of experiments on the practical applicability of various compact vector map descriptors for solving vector map data protection problems have been conducted in this paper. All investigated descriptors have shown that they (by themselves or in combination with some others) can be informative for the detection of certain distortions of a vector map. For example, descriptor 2 (average ratio of areas of bounding boxes) showed quite high robustness to almost all types of distortions, which allows us to highly evaluate the prospects of its use for copyright protection.

It should also be noted that the Ramer-Douglas-Peucker algorithm used at the preliminary stage plays an important role in the information content of each of the studied descriptors. In our paper, we have presented its modification to increase the stability of its work in conditions of distortions.

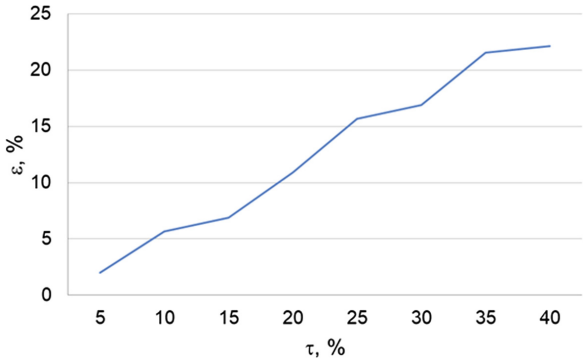


Fig. 5. Effect of the fraction of deleted nodes (distortion 4) on the relative change in the value of descriptor 2

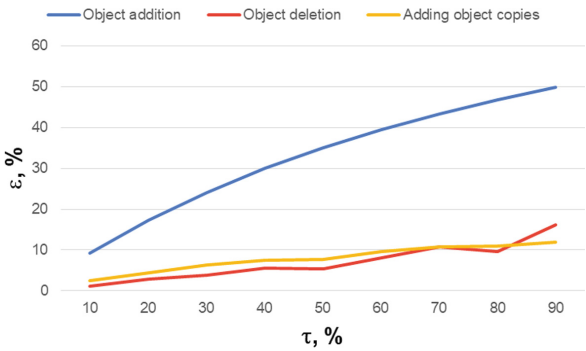


Fig. 6. Effect of the proportion of objects added or removed (distortions 7–9) on the relative change in the value of the descriptor 2

Acknowledgments. This work was supported by the Russian Foundation for Basic Research (project 19-29-09045).

References

1. Bolstad, P.: GIS fundamentals : a first text on geographic information system, 5th edn. Eider Press, Minnesota (2016)
2. Vybornova, Y.D., Sergeev, V.V.: A new watermarking method for vector map data. In: Eleventh International Conference on Machine Vision (ICMV 2018), pp. 259–266 SPIE (2019)
3. Peng, Y., Lan, H., Yue, M., Xue, Y.: Multipurpose watermarking for vector map protection and authentication. Multim. Tools Appl. **77**(6), 7239–7259 (2017). https://doi.org/10.1007/s11042-017-4631-z
4. Abubahia, A.M., Cocea, M.: Exploiting vector map properties for gis data copyright protection. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 575–582 IEEE, Vietri sul Mare, Italy (2015)

5. Abubahia, A., Cocea, M.: A clustering approach for protecting GIS vector data. In: Zdravkovic, J. et al. (eds.) Advanced Information Systems Engineering, pp. 133–147 Springer International Publishing, Cham (2015)
6. Peng, Y., Yue, M.: A zero-watermarking scheme for vector map based on feature vertex distance ratio. JECE **35**, 35 (2015)
7. Tile Map Service Specification – OSGeo. https://wiki.osgeo.org/wiki/Tile_Map_Service_Specification. Last accessed 01 July 2022
8. Vector Tiles | API. https://docs.mapbox.com/api/maps/vector-tiles/. Last accessed 01 July 2022
9. Dakroury, D.Y., et al.: Protecting GIS data using cryptography and digital watermarking. IJCSNS Int. J. Comput. Sci. Netw. Secur. **10**(1), 75–84 (2010)
10. Cox, I.J. et al.: Digital Watermarking and Steganography. Morgan Kaufmann (2008)
11. Giao, P.N., et al.: Selective encryption algorithm based on DCT for GIS vector map. J. Korea Multim. Soc. **17**(7), 769–777 (2014)
12. Ren, N., et al.: Copyright protection based on zero watermarking and Blockchain for vector maps. ISPRS Int. J. Geo-Inf. **10**(5), 294 (2021)
13. Zhou, Q., et al.: Zero watermarking algorithm for vector geographic data based on the number of neighboring features. Symmetry **13**(2), 208 (2021)
14. Xi, X., et al.: Dual zero-watermarking scheme for two-dimensional vector map based on delaunay triangle mesh and singular value decomposition. Appl. Sci. **9**(4), 642 (2019)
15. Li, A., et al.: Study on copyright authentication of GIS vector data based on Zero-watermarking. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 1783–1786 (2008)
16. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. Comput. Graph. Image Process. **1**(3), 244–256 (1972)
17. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica **10**(2), 112–122 (1973)



DDoS Detection Approach Based on Continual Learning in the SDN Environment

Ameni Chetouane[1(✉)] and Kamel Karoui[1,2]

[1] RIADI Laboratory, ENSI, University of Manouba, Manouba, Tunisia
chetouaneameni@gmail.com, kamel.karoui@insat.rnu.tn
[2] National Institute of Applied Sciences and Technology, University of Carthage, Carthage, Tunisia

Abstract. Software Defined Networking (SDN) is a technology that has the capacity to revolutionize the way we develop and operate network infrastructure. It separates control and data functions and can be programmed directly using a high-level programming language. However, given the existing and growing security risks, this technology introduces a new security burden into the network architecture. Intruders have more access to the network and can develop various attacks in the SDN environment. In addition, modern cyber threats are developing faster than ever. Distributed Denial of Service (DDoS) attacks are the major security risk in the SDN architecture. They attempt to interfere with network services by consuming all available bandwidth and other network resources. In order to provide a network with countermeasures against attacks, an Intrusion Detection System (IDS) must be continually evolved and integrated into the SDN architecture. In this paper, we focus on Continual Learning (CL) for DDoS detection in the context of SDN. We propose a method of continually enriching datasets in order to have a better prediction model. This is done without interrupting the normal operation of the DDoS detection system.

Keywords: Software Defined Networking (SDN) · Network security · Security threats · DDoS · Machine Learning (ML) · Continual Learning (CL)

1 Introduction

Over the past several decades, traditional network architecture has largely remained unchanged and has proven to have some limitations. Software Defined Networking (SDN) is an open network design that has been proposed to address some of traditional networks' key flaws [1]. Network control logic and network operations, according to SDN proponents, are two separate concepts that should be split into layers. Therefore, SDN introduced the control plane and data plane

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1199–1208, 2023.
https://doi.org/10.1007/978-3-031-27409-1_110

concepts: the centralized control plane manages network logic and traffic engineering operations, whereas the data plane only controls packet transfer among networks [2]. Although the characteristics of SDN, such as logical centralized control, global network awareness, and dynamic updating of forwarding rules, make it easy to identify and respond to attacks on the network. However, because the control and data layers are separated, new attack opportunities arise, and the SDN can become the target of various attacks such as Distributed Denial of Service (DDoS) [3]. These attacks are designed to cripple networks by flooding cables, network devices, and servers with unauthorized traffic. Several DDoS attacks have occurred, resulting in downtime and financial losses [4]. Therefore, an Intrusion Detection System (IDS) must be integrated into the SDN environment. It examines network data, analyzes it, and looks for anomalies or unwanted access [5]. For the past few years, IDS based on Machine Learning (ML) has been on the rise. However, the results of the different ML methods depend highly on the dataset. A number of public datasets have been used, including NSL-KDD [6]. However, before using these datasets to traina ML intrusion detection model, the authors do not consider the quality of the datasets. These datasets are also outdated and are not specific to the SDN environment. In addition, one of the most challenging aspects of cybersecurity is the changing nature of security dangers [7]. New attack vectors grow as a result of the development of new technologies and their exploitation in novel or unconventional ways. This involves making certain that all cybersecurity components are continually updated to guard against potential vulnerabilities. In this paper, we propose a method for detecting DDoS in the SDN environment based on Continual Learning (CL). The majority of CL research is focused on the computer vision and natural language processing areas, with the network anomaly detection domain receiving less attention [8]. The contributions in this paper include:

- The proposition of CL system to detect DDoS in the SDN environment based on dataset enrichment. This is accomplished without interfering with the detecting system's normal operation.
- The proposition of three metrics to verify the usefulness of the new dataset in terms of quality, quantity, and representativity.

The remainder of the paper is organised as follows. The related works are presented in Sect. 2. The proposed system is described in Sect. 3. In Sect. 4, we present the case study. Section 5 concludes this paper presenting future work.

2 Related Works

DDoS attacks are one of the most serious risks in SDN [9]. Several ML approaches to detect DDoS in SDN have been tried and tested. In [10], the authors proposed a method to detect DDoS in SDN based on ML. They evaluated different important feature selection methods. The best features are selected based on the performance of the SDN controller and the classification accuracy of the machine

learning approaches. To identify SDN attacks, a comparison of feature selection and ML methods has also been developed. The experimental results show that the Recursive Feature Elimination (RFE) approach is used by the Random Forest (RF) method to train the most accurate model, which has an accuracy rate of 99.97%. Ashodia et al. [11] suggested a ML technique to detect DDoS in SDN that combines Naive Bayes (NB), Decision Trees (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF). The experiment results demonstrate that Decision Tree and Random Forest algorithms offer superior accuracy and decision rates in comparison with other algorithms. The authors in [12] used various machine learning techniques such as DT, NB, and LR for DDoS detection in SDN. The proposed method includes different steps such as data preprocessing and data classification using ML classifiers. Compared to other algorithms, the machine learning algorithm with the greatest results was DT, which had an accuracy rate of 99.90%. The authors in [6] employed Decision tree (DT) and Support Vector Machine (SVM) techniques for DDoS detection in SDN. The authors identified and selected crucial features for additional detection. The SVM classifier and DT module are then used to forward the dataset to the next step. The classifiers classify the traffic dataset into two categories: attack and normal, according to the flag value (0 or 1). Otherwise, the controller will choose the route for the regular traffic packets. Employing the SVM and DT classifiers, the controller will broadcast the forwarding table to handle the payload when a DDoS problem is detected. According to the experiments, SVM performs better in a simulated environment than the decision tree.

3 Proposed System

CL brings together research and methods that deal with the issue of learning when the distribution of the data changes over time and knowledge fusion over limitless data streams must be must be considered [13]. In a previous work, we evaluated the performance of various ML approaches for DDoS detection in the SDN environment. We compared various methods, such as DT, RF, NB, SVM, and KNN. These methods are commonly used for DDoS detection in SDNs and perform well with high accuracy [14]. We found that the RF method performed better than the other methods. Therefore, we try to enhance the learning process of this method for DDoS detection in SDN. Our goal is to provide our model with new predictive capabilities without forgetting what has been learned previously. We propose a method for continual dataset enrichment and deployment of new models whenever we have a better predictor model. This is done without interrupting the detection system's operation. The flowchart of the process of CL is presented in Fig. 1.

Before explaining the different steps of the proposed system, we present the notation that will be used.

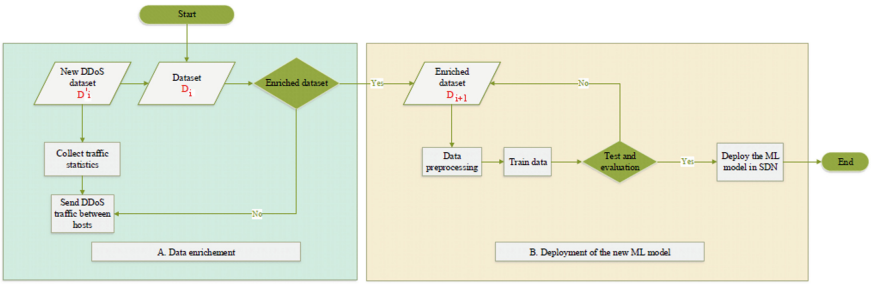


Fig. 1. The Continual Learning process.

3.1 Notation

- $P = \{p_k\}$: This set represents the security policy of the institution. It gathers the types of attacks that the institution would like to protect itself against. This set is chosen by the security administrators of the institution.
- D_i: the initial dataset.
- $D_i.type$: the set of types of attacks presented in D_i.
- $D_i.dat$: the data presented in D_i.
- D'_i: the newly generated dataset.
- $D'_i + .type$: the set of types of intrusions presented in D'_i.
- $D'_i.dat$: the data presented in D'_i.
- D_{i+1}: the new dataset which is obtained by combining D_i and D'_i.
- $D_{i+1}.type$: the set of types of intrusions of the new dataset which is obtained by combining $D_i.type$ and $D'_i.type$.
- $D_{i+1}.dat$: the data presented in D_{i+1}.
- $D_{i+1Diff}.type = |D_i.type - D'_i.type|$: is the difference between $D_i.type$ and $D'_i.type$. It includes attack types that belong to $D'_i.type$ and do not belong to $D_i.type$. The set $D_{i+1Diff}.type$ is used to display the new attack types generated in D'_i.
- $D_{i+1Union}.type = |D_i.type \cup D'_i + .type|$: is the set of union of $D_i.type$ and $D'_i.type$. It includes the types of attacks that belong to $D_i.type$ and $D'_i.type$.
- $D_{i+1Inter}.type = |P \cap D_{i+1}.type|$: is the set of intersection of P and $D_{i+1}.type$. It includes the types of attacks that belong to both P and $D_{i+1}.type$.
- $D_{i+1Diff}.dat = |D'_i.dat - D_i.dat|$: is the difference between $D_i.dat$ and $D'_i.dat$. It includes the data that belong to $D'_i.dat$ and do not belong to $D_i.dat$. The set $D_{i+1Diff}.dat$ is used to display the new data generated in D'_i.
- $D_{i+1Union}.dat = |D_i.dat \cup D'_i.dat|$: is the set of union of $D_i.dat$ and $D'_i.dat$. It includes the data that belong to $D_i.dat$ and $D'_i.dat$.

3.2 Dataset Creation

In order to achieve CL, we propose to enrich a selected dataset "D_i". We create a new dataset "D'_i" by generating new DDoS traffic based on the attack types

presented in the security policy P. This is done without interrupting the detection system in operation. We propose to generate DDoS traffic between hosts and collect the traffic statistics from the switches. The generated DDoS traffic is new and is not included in the selected dataset "D_i". Then, we place the obtained traffic statistics into a "D'_i" dataset. We combine the two datasets to obtain the new dataset "D_{i+1}". We propose a method to check whether this dataset is efficient or not. After checking the usefulness of "D_{i+1}" we train the ML model with this new dataset. Once our ML model is selected and trained, it is placed in the SDN architecture. In addition, we can use external SDN-based public datasets available online to enrich the initial dataset.

3.3 Dataset Effectiveness

After combining the two datasets, we propose a method based on the use of metrics to determine the effectiveness of the new dataset D_{i+1} in terms of quality, quantity, and representativity. In the first step, we focus on the effectiveness in terms of quality of the new dataset, which is presented in our case by the types of attacks. We present a metric called quality "qual($D_{i+1}.type$)" to verify the effectiveness of $D_{i+1}.type$. The proposed metric determines whether the dataset "D_{i+1}" obtained by combining the two datasets is enriched or not with respect to "D_i" based on the types of attacks. In other words, the combination is able to handle new types of attacks. The proposed metric is calculated as follows:

$$qual(D_{i+1}.type) = \frac{|D_{i+1Diff}.type|}{|D_{i+1Union}.type|} \quad 0 \leq qual(D_{i+1}.type) \leq 1 \tag{1}$$

– Where $|D_{i+1Diff}.type|$ represents the number of elements of $D_{i+1Diff}.type$ and $|D_{i+1Union}.type|$ is the number of elements of $D_{i+1Union}.type$.

For the effectiveness of D_{i+1} in terms of quantity, we propose a metric called quantity "quan($D_{i+1}.dat$)" that defines the number of occurrences of the new attack types in the new dataset D_{i+1}.

$$quan(D_{i+1}.dat) = \frac{|D_{i+1Diff}.dat|}{|D_{i+1Union}.dat|} \quad 0 \leq quan(D_{i+1}.dat) \leq 1 \tag{2}$$

We also provide another metric called representativity "rep($D_{i+1}.type$)", to assess how representative the new dataset D_{i+1} with respect to all searched attack types P. The proposed metric is calculated as follows:

$$rep(D_{i+1}.type) = \frac{|D_{i+1Inter}.type|}{|P|} \quad 0 \leq rep(D_{i+1}.type) \leq 1 \tag{3}$$

– Where $|D_{i+1Inter}.type|$ represents the number of elements in $D_{i+1Inter}.type$ and $|P|$ is the number of elements in P.

After the calculation of the different metrics, we move on to the next step, which is the evaluation of the obtained values, which are considered to be decision values. We used the method presented in [15] for evaluating the values of

decision-making attributes. The author proposed two approaches for aggregating attribute values based on two levels of classification: individual attribute classification and global classification. The author aggregated measures into a single measure that is a good indicator for making a decision. The obtained measurement is reversible. We use two types of classification. We start with the classification of each value related to each metric. We associate a metric value ($qual(D_{i+1}.type)$, $quan(D_{i+1}.dat)$, $rep(D_{i+1}.type)$ a binary value based on the different intervals presented in Table 1.

Table 1. Individual classification of metric values

Class	Conditions	Associated binary value
Low	$0 \leq$ metric value<0.25	00
Medium	$0.25 \leq$ metric value<0.5	01
High	$0.5 \leq$ metric value<0.75	10
Excellent	$0.75 \leq$ metric value< 1	11

Then we used the bit alternation method in the global classification that allows constructing a metric for decision making [15]. Before alternating the individual classes of each metric, we order the metrics. If we consider three factors $qual(D_{i+1}.type =$ '10', $quan(D_{i+1}.dat) =$ '11' and $rep(D_{i+1}.type) =$ '11'. We assume that the data quality is more important than the data quantity and the representativity. The sequence M = '111011' with an integer value of 59 is obtained by applying the bit alternation of the three factors ($qual(D_{i+1}.type)$, $quan(D_{i+1}.dat)$, $rep(D_{i+1}.type)$). The procedure is carried out by alternating the bit sequences (Fig. 2).

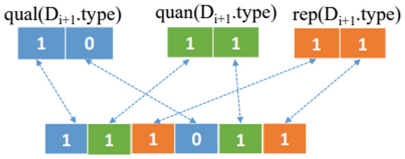


Fig. 2. The bit alternation method.

Finally, we define the threshold of acceptability as S_{use}. The choice of the threshold S_{use} is not part of the overall objectives of this research.

If M $> S_{use}$, we can conclude that the new dataset D_{i+1} resulting from the combination of D_i and D'_i is useful and effective. Once we verify that the new dataset is useful, we train the ML model with the new data D_{i+1}. Then, we evaluate the performance of this new model using the standard metrics, namely

accuracy, precision, and recall. These metrics are generally employed to assess the performance of ML methods [2]. If the new ML model performs well, we will deploy it in the SDN controller.

4 Case Study

In this section, we apply the proposed system to a case study.

4.1 Dataset Creation

In this section, we try to create a new dataset D_1' which includes new DDoS traffic. We first select an initial dataset called "DDoS attack SDN dataset" D_1 [16] and try to enrich it. This dataset contains both benign TCP, UDP and ICMP traffic and harmful traffic, which consists of TCP Syn attacks, UDP Flood attacks and ICMP attacks. There are 23 features in all in the data collection. The class name in the last column determines whether the traffic is legitimate or malicious. Besides, we used the mininet emulator to create the SDN traffic dataset, namely the "new DDoS dataset". This dataset was produced by augmenting the Ryu controller with a Python program made using the Ryu API [17] and the Mininet emulator [18]. It regularly gathers various flow and port statistics and keeps track of all the switches in the topology. The statistics it gathers are also saved in a file. We generate a DDoS attack using hping3 [19]. For generating attacks, four types of floods are generated, an ICMP flood, a TCP SYN flood, a UDP flood, and a LAND attack. ICMP flood, TCP SYN flood, and UDP flood are presented in the first dataset D_1. We try to generate new DDoS attacks based on these types to learn more about these attacks and get new results from another SDN domain. Therefore, the ML model can learn these types of DDoS attacks from the samples provided by D_1 and D_1'. We also generate LAND DDoS attacks that are not presented in the selected dataset D_1. We can use other available datasets and try to enrich them to get other combinations of DDoS types. The characteristics of the used datasets are presented in Table 2.

Table 2. Characteristics of the used datasets

Dataset D_i	Number of samples $\lvert D_i.dat \rvert$	Number of features	Types of attacks $D_i.type$
DDOS attack SDN	104345	23	TCP syn UDP flood ICMP flood
new DDoS dataset	969691	21	TCP syn UDP flood ICMP flood LAND attack

4.2 Dataset Effectiveness

In this section, We try to determine the effectiveness of the new dataset obtained by combining the initial dataset and the new generated dataset. First of all, we present the values of the different notation fields presented earlier in Sect. 3.1:

- $P = \{ICMP, UDP, TCP, LAND\}$: It presents the intrusions types that we aim to detect. These intrusions are considered the most dangerous for the SDN environment [20].
- D_1: "DDoS attack SDN dataset".
- $D_1.type$: {TCP, UDP, ICMP}.
- $D_1.dat$: the data presented in D_1.
- D'_1: "new DDoS dataset".
- $D'_1.type$: {TCP, ICMP, UDP, LAND}.
- $D'_1.dat$: the data presented in D_2.
- D_2: the enriched dataset which is obtained by combining D_1 and D'_1.
- $D_2.type$: {TCP, ICMP, UDP, LAND}.
- $D_2.dat$: the data presented in D_2.
- $D_{2Diff}.type$ = {LAND}.
- $D_{2Union}.type$ = {TCP, UDP, ICMP, LAND}.
- $D_{2Inter}.type$ = {TCP, UDP, ICMP}
- $D_{2Diff}.dat$ = {LAND.dat}
- $D_{2Union}.dat$ = {TCP.dat, UDP.dat, ICMP.dat, LAND.dat}.

Then, we calculate the different metrics using equations (4), (5) and (6). We obtain the following results:

$$\text{qual}(D_2.type) = 0.25,\ \text{quan}(D_2.dat) = 0.4\ ,\ \text{rep}(D_2.type) = 0.75.$$

The next step consists in associating binary values to the values of the three metrics. We obtain the following factors (see Table 1): $qual(D_2.type)$ = '01', $quan(D_2.dat)$ = '01', $rep(D_2.type)$= '11' . There are several ways to define the order of importance of different factors depending on the institution. In our case study, we assume that $qual(D_2.type)$ is more important than $quan(D_2.dat)$ and $quan(D_2.dat)$ [15]. We use the bit alternation approach to determine the decision value (see Fig. 2). For example, if we suppose that the threshold of acceptability $S_{use} = 10$. We obtain the following sequence: 001111, which corresponds to the integer value M = 15. We can see that M > S_{use}.

As a result, the new dataset D_2 formed by combining the initial dataset D_1 and the newly generated dataset D'_1 is effective. Therefore, we can say that the initial dataset D_1 is enriched. We train the Random Forest (RF) model with new training data D_2. In the next step, we assess the performance of the new trained RF method using the standard metrics, namely accuracy, precision, and recall. The RF model gives good results, with a value of 99% for the three metrics. We can note that this model performs well. Therefore, we deploy it in the SDN architecture without interrupting the system operation or the DDoS detection process.

5 Conclusion

In this paper, we propose a Continual Learning (CL) system for DDoS detection in the SDN environment. We apply CL to the DDoS detection system in the SDN environment to make it self-adapting to modern threats and reduce recycling costs. We create a new DDoS dataset and combine it with a selected dataset. Then, we propose a method to verify whether the dataset obtained by combining the selected dataset with the newly generated dataset is useful or not; in other words, whether we managed to enrich the selected dataset. We propose three metrics, called quality, quantity, and representativity, to determine the effectiveness of the new dataset. We use the bit alternation method to integrate the three metrics and make a decision about the usefulness of the new dataset. We train our Machine Learning (ML) model with the enriched dataset. In the next step, we evaluate the performance of the new ML model using the standard metrics, namely accuracy, precision, and recall. The new model performs well, so we deployed it on the SDN controller. In future work, we will use Deep Learning (DL) methods for DDoS detection in SDN.

References

1. Kreutz, D., Ramos, F.M., Verissimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-defined networking: a comprehensive survey. Proc. IEEE **103**(1), 14–76 (2014)
2. Chetouane, A., Karoui, K.: A survey of machine learning methods for DDoS threats detection against SDN. In: International Workshop on Distributed Computing for Emerging Smart Networks, pp. 99–127. Springer (2022)
3. Kreutz, D., Ramos, F.M.V., Verissimo, P.: Towards secure and dependable software-defined networks. In: Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking, pp. 55–60 (2013)
4. Sachdeva, M., Singh, G., Kumar, K., Singh, K.: Measuring impact of DDoS attacks on web services (2010)
5. Liao, H.J., Lin, C.H.R., Lin, Y.C., Tung, K.Y.: Intrusion detection system: a comprehensive review. J. Netw. Comput. Appl. **36**(1), 16–24 (2013)
6. Sudar, K.M., Beulah, M., Deepalakshmi, P., Nagaraj, P., Chinnasamy, P.: Detection of distributed denial of service attacks in SDN using machine learning techniques. In: 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–5. IEEE (2021)
7. What is cybersecurity?
8. Amalapuram, S.K., Tadwai, A., Vinta, R., Channappayya, S.S., Tamma, B.R.: Continual learning for anomaly based network intrusion detection. In: 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp. 497–505. IEEE (2022)
9. Eliyan, L.F., Di Pietro, R.: Dos and DDoS attacks in software defined networks: a survey of existing solutions and research challenges. Futur. Gener. Comput. Syst. **122**, 149–171 (2021)
10. Nadeem, M.W., Goh, H.G., Ponnusamy, V., Aun, Y.: DDoS detection in SDN using machine learning techniques. Comput. Mater. Contin. **71**(1), 771–789 (2022)

11. Ashodia, N., Makadiya, K.: Detection of DDoS attacks in sdn using machine learning. In: 2022 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1322–1327. IEEE (2022)
12. Altamemi, A.J., Abdulhassan, A., Obeis, N.T.: DDoS attack detection in software defined networking controller using machine learning techniques. Bull. Electr. Eng. Inform. **11**(5), 2836–2844 (2022)
13. Ring, M.B. et al.: Continual learning in reinforcement environments (1994)
14. Aslam, M., Ye, D., Tariq, A., Asad, M., Hanif, M., Ndzi, D., Chelloug, S.A., Elaziz, M.A., Al-Qaness, M.A., Jilani, S.F.: Adaptive machine learning based distributed denial-of-services attacks detection and mitigation system for SDN-enabled iot. Sensors **22**(7), 2697 (2022)
15. Karoui, K.: Security novel risk assessment framework based on reversible metrics: a case study of DDoS attacks on an e-commerce web server. Int. J. Netw. Manag. **26**(6), 553–578 (2016)
16. Ahuja, N., Mukhopadhyay, D., Singal, G.: DDoS attack SDN dataset (2020)
17. mgen S. Natarajan. Ryu application api
18. Mininet emulation software (2018)
19. S. Natarajan. hping3
20. Sen, S., Gupta, K.D., Manjurul Ahsan, M.: Leveraging machine learning approach to setup software-defined network (SDN) controller rules during DDoS attack. In: Proceedings of International Joint Conference on Computational Intelligence, pp. 49–60. Springer (2020)



Secure e-Voting System—A Review

Urmila Devi(✉) and Shweta Bansal

Computer Science and Engineering Department, K.R. Mangalam University, Gurugram, Haryana, India
urmilabibyan@gmail.com, shweta.bansal@krmangalam.edu.in

Abstract. Security is a requirement which is needed in all aspects of life. In this computer world when everything is online then online threats are common to all. e-Voting system is now the need of the hour. As in this internet era everyone wants everything just one click away including shopping, online food ordering and even finding life partners online then why not voting online. In traditional voting systems everyone has to go to cast their votes on polling booths but with few reasons only few go for voting. There are many reasons like—people do not want to stand in long queues, they are out of station on the day of voting etc. So to keep all these in mind online voting can solve such problems and make more vote casting to make the election successful. But there are many security measures that come into consideration for the online voting system implementation such as—Ballot stuffing threats, Denial-of-service attack, double spending problem, 51% attacks and many more. This paper will give a comprehensive survey about the possible threats in online voting system and different types of cryptographic techniques like—Asymmetric key cryptography, Elliptic-Curve Cryptography (ECC), Identity Based Encryption (IBE) cryptography etc. which can be used to prevent from such attacks and also provide a comparison summary based on the e-Voting security requirements.

Keywords: e-Voting System · Online Voting · Cryptography techniques · Security threats · Ballot Stuffing · Double spending attack

1 Introduction

Voting is a process to elect the fair and correct person for the nation. In the traditional voting system people used to cast their votes on paper ballots. After that EVM (Electronic voting Machine) came into picture for voting. But there are many reasons only few people come to cast their votes. Due to that election fair decisions may not come. So to make elections successful, an online voting system can be remedy for the same. But whenever we talk about the online transactions then online attacks come into the mind automatically. Like–

- Whether online voting is safe?
- What if my vote doesn't count?
- What if someone has changed my vote in between?

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1209–1224, 2023.
https://doi.org/10.1007/978-3-031-27409-1_111

- What if the online voting server goes down or out of service while casting the vote?
- What if the entire voting system gets hacked by illegitimate persons where they can temper the entire voting process?

So there are many more questions that come to mind whenever we try to implement any application online and want to share through the internet. In this paper we will discuss the security requirements of e-Voting systems, possible online threats in e-Voting systems and different types of cryptography techniques.

1.1 Cryptography

Cryptography is the study of secure data communication techniques that allows sender to send and intended recipient can use the message. While transferring the information it will be encrypted and at time of receiving it will be decrypted by the legitimate recipient (Fig. 1).

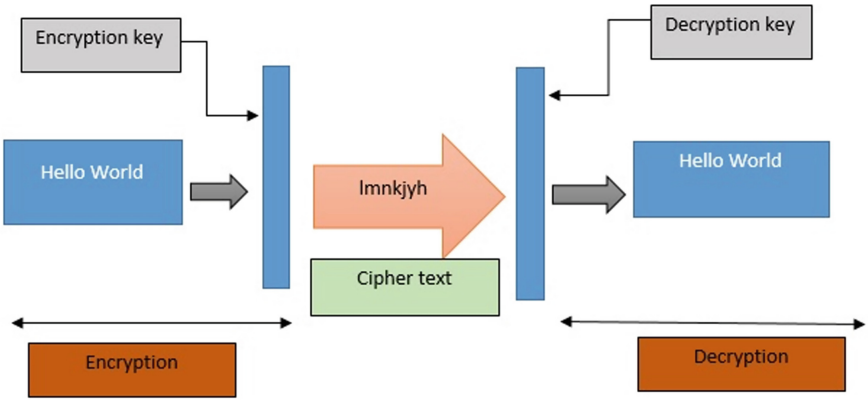


Fig. 1. Cryptography process

In cryptography below terminology is being used–

1. **Encryption**—It is a process to convert the actual information into different forms by using the encryption key.
2. **Encryption Key**—This is a procedure or logic which is being used to transform the actual message into a different one.
3. **Cipher text**—When Plain text gets converted in different forms then that changed form will be called as cipher text.
4. **Decryption**—It is a process to decrypt the information to get the actual meaning by using the decryption key.
5. **Decryption Key**—This is a procedure or logic to convert the cipher text into plain text to get the actual meaning.

Now, when data is being encrypted then hackers can hack that logic which is being used to transform the message. Then there are mainly two types of key exchange concept—Public and private key.

- **Public Key cryptography**—Public-key cryptography, sometimes known as Asymmetric cryptography, is the study of cryptographic systems that employ pairs of linked keys. A public key and its accompanying private key make up each key pair. Cryptographic algorithms that are based on one-way functions are used to create key pairs. The private key must be kept hidden in order for public-key cryptography to be secure; nevertheless, security is not compromised if the public key is freely distributed. There are many known and used asymmetric key cryptographic algorithms like—Diffie–Hellman key exchange protocol, ElGamal, Elliptic-Curve Cryptography, RSA etc. In Fig. 2 Public key cryptography is demonstrated where at time of encryption public keys will be used but at time of decryption only one private key will be used which will be specific to the receiver.
- **Private Key cryptography**—Symmetric cryptography is one type of encryption (also known as secret key cryptography or private key cryptography). Since symmetric cryptography is substantially quicker than asymmetric cryptography, it is ideally suited for bulk encryption. Symmetric cryptography uses a shared key between both parties (which is kept secret). The shared secret key must be exchanged by both parties before any communications can start. A distinct shared key is necessary for each communication between two entities. Other communication partners are kept in the dark regarding the key. Advanced Encryption Standard (AES), Triple Data Encryption Standard (3DES), and Rivest Cipher 4 (RC4) are popular secret-key cryptographic algorithms. As shown in Fig. 3 only one private is being used to encrypt and decrypt the message.

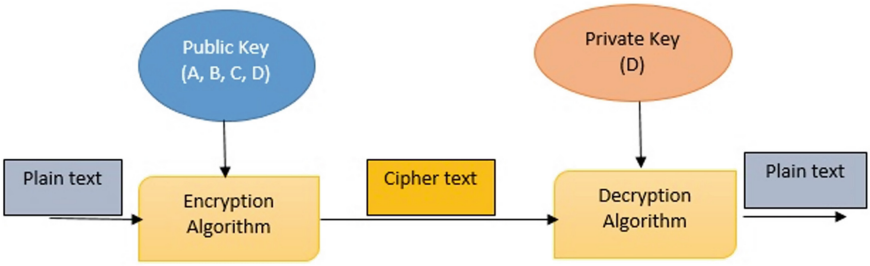


Fig. 2. Public key/Asymmetric Key Cryptography

So in cryptography generation of keys and their security is more important to make the message free from online threats.

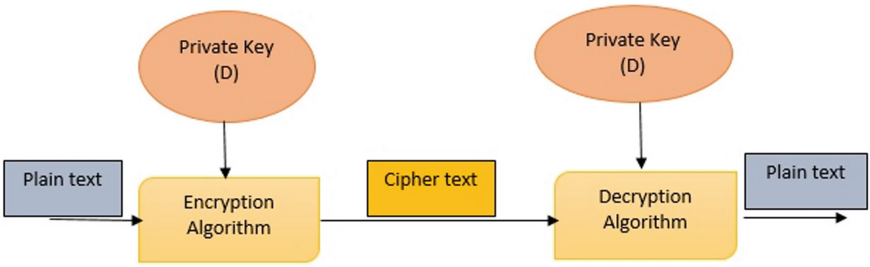


Fig. 3. Private Key/Symmetric Key Cryptography

1.2 e-Voting System

Online voting system has mainly three phases (Fig. 4).

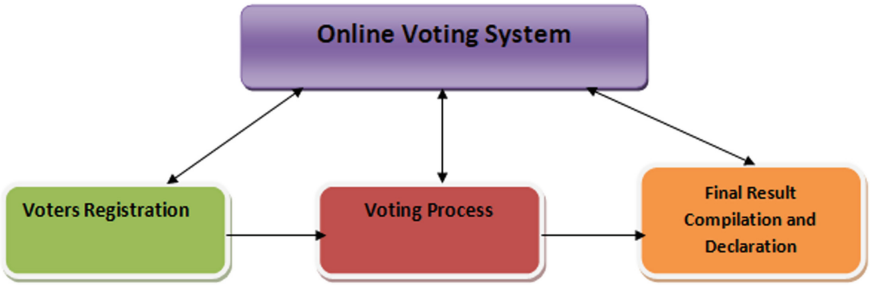


Fig. 4. Online voting process

Security required at every stage of the online voting process. At time of voter registration the correct person has to be registered where valid ID proofs will be used for the same. In the second phase, vote casting is a very important process as we need to take care about the attacks related to ballots, and online services. System should not go down at the time of the process. Almost all the attacks happen at this stage only. At last in the last phase voting results will be announced where bulletin board security is most important along with the prevention from the result manipulation.

1.3 e-Voting System Security Requirements

e-Voting system has requires below mainly security checks (Table 1).

Table 1. e-Voting system security requirements

S. No	Security Requirements	Remarks
1	Voter Authenticity	To ensure that the voter is authentic to cast the vote by checking the valid ID proofs.
2	Registration	The voter registration shall be done in person only. However, the computerized registration database shall be made available to polling-booths all around the nation.
3	Voter Anonymity	Ensure that votes must not be associated with voter identity.
4	System Integrity	Ensure that system cannot be changed during the election process.
5	Data Integrity	Ensure that each vote is recorded as intended and cannot be tempered once it get stored.
6	Privacy	Make sure that no one should be able to determine how many individuals voted.
7	No Vote-Selling	Ensure not to change the votes once it gets stored.
8	Reliability	Election systems should work robustly, without loss of any votes, even in the face of numerous failures, including failures of voting machines and total loss of network communication. The system shall be developed in a manner that ensures there is no malicious code or bugs.
9	Availability	System should be available at any cost at time of the process.
10	System Accountability	Ensure that system operations are logged and audited.
11	System Simplicity	System should be simple so that bug identification can be found in an easy way.
12	Ballot secrecy	The voting ballot should be secure. The voting system should not reveal the vote which an encrypted ballot corresponds to.
13	Receipt-Freeness	The voting system should not provide the Voter any evidence to prove to a third party how she voted.

2 Related Work

Authors [1] studied e-Voting systems based on blockchain technology. They used asymmetric key cryptography to encrypt the votes. They proposed a framework where they used the Aadhar Id as Virtual Id for the voters for their identification and authentication. They found the choices of key-length impact the file size of the encoded record in the encryption process. They ensure the security in the e-Voting process by using the Blockchain technology with cryptography techniques. But in their study transparency, auditability and coercion resistance is missing which can be a topic for future researchers.

Blockchain has been adopted to address significant challenges, such as trust in diverse domains, including voting, logistics and finance. However, transaction malleability has been identified as a threat for blockchain, which can potentially lead to an inconsistent state that can result in further attacks such as double-spending. In this context, this paper

[2] is focused on investigating the feasibility of transaction malleability within a typical blockchain application aiming to identify scenarios that may lead to a successful transaction malleability attack. The author's objective in doing so is to highlight conditions which cause such attacks to facilitate the development of protection mechanisms for them. Specifically, this paper presents a successful simulation of transaction malleability attack within the context of blockchain-based electronic voting. The evaluation has identified the impact of parameters, such as network delay and block generation rate in achieving a successful transaction malleability attack, which highlights future directions of research.

Authors [3] studied the e-Voting system and proposed a secure and verifiable voter registration and authentication scheme using the Fuzzy Vault algorithm. They found their proposed mechanism prevents ballot stuffing attack. The investigator has taken the idea from Direct Recording Electronic with Integrity and Enforced Privacy (DRE-ip) and tried to provide the improvement in DRE-ip algorithm. They also introduced an improved non-interactive zero-knowledge (NIZK) proof that boosts the efficiency of the system and proposed two methods to store the ballots using blockchain and cloud server. They used the exponential ElGamal cryptosystem to encrypt a vote to provide the coercion-resistance system.

Authors [4] studies remote e-Voting systems and proposed a new model for remote e-Voting systems for large scale elections. The investigators found that their scheme can be implemented in IoT devices. They ensures the Double-Voting Prevention, Anonymity, Coercion-Resistance and Receipt Freeness property of the system. They found by using their d-BAME new model voters can cast their vote through mobile devices in less than a minute where encryption key size is 4096 bits.

Authors [5] studied Remote e-Voting systems and proposed a new idea which can provide authentication, anonymity and transparency by using the Smart Contract with Elliptic Curve Cryptography by using the Ethereum Blockchain Network. They found this idea can provide a solution to those who are living in remote areas but cannot vote. By using Blockchain based Remote e-Voting system it can be possible.

Authors [6] studies the e-Voting system thoroughly and provides a summary about the requirements of e-Voting system and explains the literature where others have done work based on blockchain technology. Authors also discussed the security part of the e-Voting system. As per their study they mentioned, not a single proposed scheme suited for all the requirements for e-Voting which can be a future research area for the new generation.

Authors [7] studies e-Voting systems and provides an Android based mobile application using blockchain technology with Deep Learning to provide the Coercion-Resistance in the system. The investigators have used face detection techniques. They suggested face detection can be used with facial emotions for better Coercion Detection. They also suggested this application be developed for another mobile-based operating system.

Authors [8] studies the Blockchain based e-Voting system and addresses the few issues which can occur while voting. Issues like manipulations of the election results may occur. They proposed a model with two layer encryption techniques to secure the results. The investigators make sure in their model that results can be counted after

the participation of all stakeholders at the end. Due to blockchain technique use they ensure the voter's privacy also.

The Elliptic-Curve Cryptographic (ECC), pairings, and Identity Based Encryption (IBE) cryptography algorithms were used by the authors [9] to study the electronic voting system and suggest a new electronic voting protocol, named VYV for Verify-Your-Vote. In their protocol, they have made sure that the electronic voting mechanism is secure and private. Using the ProVerif tool, they also attempted to demonstrate the security of the VYV protocol. The investigator concluded that their suggested protocol was resistant to coercion. They made recommendations for future work to ensure coercion resistance.

Using biometric validation and authentication, authors [11] proposed a novel approach to prevent bogus voting by studying blockchain-based electronic voting systems. The scientists also suggested a system that would recognise voting trends using deep learning techniques in order to comprehend human psychology.

Authors [12] have proposed a new technique to avoid the phishing attacks in online voting systems by sharing the share 1 and share 2 of the images using Visual Cryptography.

Authors [14] studied the electronic voting system and proposed an e-Voting model which is implemented by using the CryptDB. They proposed OTP based authentication and to store the votes used CryptDB which is a database where information will get stored in encryption form. They ensured the security in authentication using OTP technique and authorization by using the voter's location.

Authors [15] studied the e-Voting system and proposed a system named SecureBallot which is an open source system too. They used symmetric encryption along with public key cryptography. They combined Advanced Encryption Standard (AES) and Rivest, Shamir, Adleman (RSA) cryptographic techniques to achieve security in the electronic voting process. They ensure voter's authentication, registration and system and data integrity and coercion resistance in the system.

Authors [16] studied the e-Voting system and proposed a secured voting system by using the RSA (Rivest, Shamir, Adleman) Key Encapsulation Mechanism with two layers-Symmetric and Public key Layers. They also compared the RSA with the ECC (Elliptic curve cryptography) and talked about the limitations of the ECC due to larger encrypted message size which resulted in high complexity of the system. They ensured the Voters authenticity, integrity, privacy, system auditability and receipt freeness in the e-Voting process. They also suggested future researchers use the blockchain technology to make the system robust as RSA is limited to give protection against quantum computer attacks.

Authors [17] studied the voting system and proposed a model which is an android app with the blockchain technology. They proposed a system based on two blockchains—one for VOTERS and VOTES blockchain. They ensured the authentication of the voters by giving the PIN at time of registration based on the NID (National Identification Number).

Authors [18] studied the e-Voting system and focused on few challenges like transparency and auditability in the voting process. They have achieved these two goals by using the Multi-Agent technique of Artificial Intelligence with the integration of Blockchain technology by introducing the new model named as ABVS (Auditable Blockchain Voting System). They found e-Voting provides numerous benefits other

than their targeted goals like—Prevention of fraud and acceleration Result Processing. They suggested we can achieve these goals by using the Smart Contract concept also for future researchers.

Authors [19] studied blockchain based e-Voting and propped a new model "iSAY" with new proposed framework "iZigma". In their model they basically proposed a polling system where a general public can be a part of the legislative decision making process. The investigators used the Natural language Processing (NLP) technique along with the Machine learning algorithm to identify the public opinion and that polling system saved on blockchain. So they used three main techniques in their model—Blockchain, NLP and Machine learning algorithms. They suggested if we use machine learning with blockchain for e-Voting process then the system can be more robust.

Authors [20] studied blockchain based e-Voting system and did a systematic analysis of existing systems with implementation of blockchain with the different parameters like—blockchain type, the consensus approach used, and the scale of participants. They found a number of potential research opportunities in this field for future prospects.

Authors [21] studied blockchain based e-Voting systems and did a review on existing e-Voting systems. The authors suggested the Hyperledger Sawtooth framework can be used to implement the realistic, robust and practical e-Voting system.

2.1 Possible Attacks in e-Voting Systems

See Table 2.

Table 2. Attacks in e-Voting systems

S. No	Attacks	Remarks
1	Malleability Attack [2]	To change the unique Id of a transaction in Blockchain before confirming from the Network. In This attack miners do not realize that the transaction hash has been changed or not and they can allow addition of the wrong transaction in the network. Authors have stimulated the Malleability Attack to give a better understanding and also gave a summary on this attack's consequences.

(*continued*)

Table 2. (*continued*)

S. No	Attacks	Remarks
2	Ballot stuffing attack [3]	Ballot stuffing is an illegal practice where one person can submit multiple votes during the voting process where only one vote is allowed to be submitted. Authors have discussed the ballot stuffing issue and proposed a new model using the secure multi-party computation and non-interactive zero-knowledge (NIZK) proof concepts.
3	Denial of Service Attack [10]	Denial of Service Attack is a type of attack where an attacker targets the network and makes it inaccessible from the users of the network. Basically attackers use flooding to make the service down for the network and can use the same for their use.
4	Non-Repudiation Attack [2]	Non-Repudiation Attack refers to a scenario where a statement of an author is an obligation or provides dispute to its authorship. This is questionable practice on authorship.
5	51% Attack Discussion/Majority Attack [13]	Whenever any transaction is being added to the blockchain network then there should be a 51% majority for the same. But when an attacker is successful to take the 51% nodes in the network by their side then the network can be attacked easily. This type of problem is called 51% attack.
6	Double Voting Problem [4]	In the e-Voting system the ballot should prevent the Double voting problem.
7	Trash Attack [3]	Bull Bulletin boards should be compromised in the e-Voting system, this can be called a trash attack.

2.2 Cryptographic Techniques Used to Secure the e-Voting System

See Table 3.

Table 3. Cryptographic techniques

S. No	Cryptographic Techniques	Remarks
1	Elliptic-Curve Cryptography (ECC) [9]	It is a key based technique for encryption of data and uses the algebraic structure for the same.
2	Identity Based Encryption (IBE) cryptography [9]	Identity-based encryption (IBE) is a type of public-key cryptography in which a third-party server creates a public key that may be used for both encrypting and decrypting electronic messages using a straightforward identification, like an email address.
3	Asymmetric key cryptography [1]	Asymmetric cryptography, sometimes known as public-key cryptography, is the study of cryptographic systems that employ pairs of linked keys. A public key and its accompanying private key make up each key pair.
4	Homomorphic Encryption [8]	Instead of having to decrypt each cipher text individually at a high computational expense, the Homomorphic encryption method can be utilized to construct numerous cipher texts and then decrypt them.
5	Visual Cryptography [12]	Visual information (images, text, etc.) can be encrypted using the cryptographic technique known as visual cryptography such that when the information is decrypted, a visual image is produced.

2.3 Comparison of Related Literature Based on the Security Requirements

See Table 4.

Table 4. Literature comparison based on security requirements

Year	Technology used in Literature	Voter Authenticity	Voter Anonymity	Data Integrity	Privacy	No Vote-Selling/ Coercion Resistance	Reliability	Availability	Ballot secrecy	Receipt-Freeness
2020	Blockchain Technology with Virtual ID of Aadhar [1]	Yes	No	Yes	Yes	No	No	Yes	No	No
2021	Direct-Recording Electronic (DRE) machines using secure multi-party computation and non-interactive zero-knowledge (NIZK) proof. [3]	Yes	No	No	Yes	No	No	Yes	No	No
2021	Blockchain Technology using ElGamal cryptography [4]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
2021	Blockchain technology using Elliptic Curve Cryptography [5]	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No

(*continued*)

Table 4. (*continued*)

Year	Technology used in Literature	Voter Authenticity	Voter Anonymity	Data Integrity	Privacy	No Vote-Selling/ Coercion Resistance	Reliability	Availability	Ballot secrecy	Receipt-Freeness
2021	Blockchain technology and Face Detection Technique using Deep Learning [7]	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No
2021	Blockchain Technology [8]	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No
2018	Elliptic-Curve Cryptographic (ECC), pairings, and Identity Based Encryption (IBE) cryptography algorithms [9]	Yes	No	Yes	Yes	No	No	No	No	No
2021	Blockchain technology with IOT based devices [10]	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No

(*continued*)

Table 4. (*continued*)

Year	Technology used in Literature	Voter Authenticity	Voter Anonymity	Data Integrity	Privacy	No Vote-Selling/ Coercion Resistance	Reliability	Availability	Ballot secrecy	Receipt-Freeness
2021	Biometric Validation and Deep learning technique to know the voting patterns [11]	Yes	Yes	Yes	Yes	No	No	No	No	No
2021	CryptDB and OTP based Authentication [14]	Yes	No	Yes	Yes	No	No	No	Yes	No
2021	Advanced Encryption Standard (AES) and Rivest, Shamir, Adleman(RSA) cryptographic technique used [15]	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No

(*continued*)

Table 4. (*continued*)

Year	Technology used in Literature	Voter Authenticity	Voter Anonymity	Data Integrity	Privacy	No Vote-Selling/ Coercion Resistance	Reliability	Availability	Ballot secrecy	Receipt-Freeness
2022	Rivest, Shamir, Adleman (RSA) Key Encapsulation Mechanism with two layers-Symmetric and Public key Layers [16]	Yes	No	Yes	Yes	No	No	No	Yes	Yes
2020	Blockchain Technology with Unique PIN number based on NID(National Identification Number) [17]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No

Yes means that requirement is fulfilled by the proposed model in the literature
No means that requirement is not fulfilled by the proposed model in the literature

Conclusion

Security is always an alarming point in any kind of online application implementation. Since we are talking about an online voting system then there are many more challenges to meet all security requirements. As all the transaction will be online to secure the vote ballots, vote storage, authorities' logins to avoid the unauthorized access and to secure the final results from the illegitimate access. As per the discussed requirement there are challenges in maintaining and securing the vote ballot by doing stuffing or any other illegal activity. As per the review of related work there are main key security feature which needs attention in all the implementation like–

1. Voters Authentication
2. Voters Anonymity
3. No vote selling/Coercion Resistance
4. Ballot Stuffing
5. Receipt-Freeness
6. Reliability
7. Denial-of-Service Attack

As per the latest literature reviews there are still improvement needed to make the secure e-Voting system. Many has used different-different cryptographic techniques used to achieve the security in system. But there are many loopholes where any intruder can access the system. Few paper has discussed about the blockchain based e-Voting system but they did not focus on the all the security requirements.

For future work security can be achieved by using the Blockchain technology with Artificial Intelligence techniques. A secure model can be designed by using the Intrusion detection techniques using AI technique at time of all phases and to store the votes and results blockchain can be used so that no one cannot change the stored votes. As Blockchain technology comes with the peer-to-peer distributed ledger nature where transactions cannot be hacked as once any information gets stored in the blockchain it is next to impossible to change the transaction.

This paper finally summarize all the possible threats, security requirements along with the different comparison from existing literature and at the end suggested future work direction to achieve/make the secure and robust e-Voting system.

References

1. Roopak, T. M., Sumathi, R.: Electronic Voting based on virtual ID of Aadhar using blockchain technology. In: 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020—Conference Proceedings, ICIMIA, pp. 71–75 (2020)
2. Khan, K.M., Arshad, J., Khan, M.M.: Simulation of transaction malleability attack for blockchain-based e-Voting. Comput. Electr. Eng. **83**, 106583 (2020)
3. Panja, S., Roy, B.: A secure end-to-end verifiable e-Voting system using blockchain and cloud server. J. Inf. Secur. Appl. **59** (2021)
4. Zaghloul, E., Li, T., Ren, J.: D-BAME: distributed blockchain-based anonymous mobile electronic voting. IEEE Internet Things J. **8**(22), 16585–16597 (2021)

5. Rathore, D., Ranga, V.: Secure remote e-Voting using blockchain. In: Proceedings—5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021, ICICCS, pp. 282–287 (2021)
6. Buyukbaskin, L.A., Sertkaya, I.: Requirement analysis of some blockchain-based e-Voting schemes. Int. J. Inf. Secur. Sci. **9**(4), 188–212 (2020)
7. Pooja, S., Raju, L.K., Chhapekar, U., Chandrakala, C.B.: Face detection using deep learning to ensure a coercion resistant blockchain-based electronic voting. Eng. Sci. **16**, 341–353 (2021)
8. Taş, R., Tanriöver, Ö.Ö.: A manipulation prevention model for blockchain-based e-Voting systems. Secur. Commun. Netw. (2021)
9. Chaieb, M., Yousfi, S., Lafourcade, P., Robbana, R., Chaieb, M., Yousfi, S., Lafourcade, P., Verifiable, R. R. V. A.: Verify-your-vote : a verifiable blockchain-based online voting protocol to cite this version:HAL Id : hal-01874855 Verify-Your-Vote : A Verifiable Blockchain-based Online Voting Protocol (2018)
10. Rathee, G., Iqbal, R., Waqar, O., Bashir, A.K.: On the design and implementation of a blockchain enabled e-Voting application within IoT-oriented smart cities. IEEE Access **9**, 34165–34176 (2021)
11. Prabhakar, E., Kumar, K.N., Karthikeyan, S., Kumar, A.N., Kavin, P.: Smart online voting and enhanced deep learning to identify voting patterns. Int. Res. J. Modern. Eng. Technol. Sci. **4**, 162–165 (2021)
12. Nisha, S., Madheswari, A.N.: Prevention of phishing attacks in voting system using visual cryptography. In: 1st International Conference on Emerging Trends in Engineering, Technology and Science, ICETETS 2016—Proceedings (2016)
13. Singh, S., Wable, S., Kharose, P.: A review of e-Voting system based on blockchain technology. Int. J. New Pract. Manag. Eng. **10**(4), 9–13 (2022)
14. Vemula, S., Kovvur, R.M.R., Marneni, D.: Secure e-Voting system implementation using CryptDB. SN Comput. Sci. **2**(3), 1–6 (2021). https://doi.org/10.1007/s42979-021-00613-9
15. Agate, V., De Paola, A., Ferraro, P., Lo Re, G., Morana, M.: SecureBallot: a secure open source e-Voting system. J. Netw. Comput. Appl. **191**(May), 103165 (2021)
16. Ahubele, B.O., Oghenekaro, L.U.: Secured electronic voting system using RSA Key encapsulation mechanism. Eur. J. Electr. Eng. Comput. Sci. **6**(2), 81–87 (2022)
17. Kumar, D.D., Chandini, D.V., Reddy, D.: Secure electronic voting system using blockchain technology. Int. J. Smart Home **14**(2), 31–38 (2020)
18. Pawlak, M., Poniszewska-Marańda, A.: Blockchain e-Voting system with the use of intelligent agent approach. In: ACM International Conference Proceeding Series, pp. 145–154 (2019)
19. Wattegama, D., Silva, P.S., Jayathilake, C.R., Elapatha, K., Abeywardena, K., Kuruwitaarachchi, N.: "iSAY": Blockchain-based intelligent polling system for legislative assistance. Int. J. Adv. Comput. Sci. Appl. **12**(1), 233–239 (2021)
20. Huang, J., He, D., Obaidat, M.S., Vijayakumar, P., Luo, M., Choo, K.K.R.: The application of the blockchain technology in voting systems. ACM Comput. Surv. **54**(3) (2021)
21. Vivek, S.K., Yashank, R.S., Prashanth, Y., Yashas, N., Namratha, M.: e-Voting systems using blockchain: an exploratory literature survey. In: Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020, pp. 890–895 (2020)

Securing East-West Communication in a Distributed SDN

Hamdi Eltaief(✉), Kawther Thabet, and El Kamel Ali

PRINCE Research Lab, University of Sousse, ISITCom, ISITCom Route principale n°1, Hammam, Sousse 4011, Tunisia
hamdi.eltaief@issatso.rnu.tn , ali.kamel@isitc.u-sousse.tn

Abstract. The multi-controller software-defined networks are vulnerable to a false data injection attack. It creates topology inconsistency among controllers. We present a security architecture using multi-controller SDN and blockchain technology. The architecture aims to ensure secure and trustworthy inter-controller communication based on a reputation mechanism. We offer a reputation mechanism to compute the trust value of each node in the topology and to determine the problem's source, whether it is the switch or the master controller. The master controller creates block, the redundant controllers validate, and mining decides to transfer the block to the blockchain or reject it according to the reputation values of nodes. The comparative study between the existing mining decision process and the proposed one clearly specifies the trust value of each node, and the detection of the source of the fraudulent flow rules injection is more trustworthy.

Keywords: Securing communication · Multicontroller SDN · Blockchain · Reputation mechanism · Trustworthy

1 Introduction

The SDN programmable network is a paradigm that consists of separating the control plane and the data plane of a network. It can be distributed when it has several controllers each of them is responsible for a one domain. There are two types of communication in this network (vertical, horizontal). OpenFlow protocol manages the the vertical communication. Controllers exchange network topology information between them via their horizontal communication (east-west communication). The type of communication allows controllers to share any information about network topology (link state, devices, flow table, network hosts, ...). However, multicontroller SDN could be targeted by false data injection attack. This could cause routing malfunctions and routing loops. So, we present a security architecture using multi-controller SDN and blockchain technology. The presented architecture is aimed at ensuring secure and trustworthy intercontroller communication based on a reputation mechanism. The master

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1225–1234, 2023.
https://doi.org/10.1007/978-3-031-27409-1_112

controller creates block then the redundant controllers validate the blocks, and the mining decides to transfer the block to the blockchain or reject it according to the reputation values of the nodes. The reputation mechanism employs fading reputation strategies to manage and customize malicious nodes (controllers or switches) to detect the source of the false data injection. The comparative study between the existing mechanism and the proposed one clearly specifies the trust value of each node and demonstrates the detection of the source of the false data injection.

2 Related Work

Li et al. [1] propose a novel secure multi-controller rule enforcement verification mechanism in SDN. it adopt blockchain technology to provide privacy protection for forwarding behaviors. They present a signature scheme with appropriate cryptographic primitives.

Derhab et al. [2] present a security architecture using multi-controller SDN and blockchain technology. they present a reputation mechanism which rates controllers according strategies.

Tong et al. [3] present a novel an architecture using SDN and blockchain technology. They construct the control layer combining horizontal and vertical communication.Therefore, Tong et al. [3] propose an SDN architecture using blockchain technology. They modified the control layer, and add blockchain to recording the network information to ensure data integrity and reliability and prevent malicious administrator Threats.

Latah et al. [4] they focus on data plane authenticatin. It can be exploited to attack SDN's control plane. They propose a protocol for authenticating SDN's data plane including SDN switches and hosts. We also provide a proof-of-concept that demonstrates the applicability and feasibility of our protocol in SDNs.

Boukria et al. [5] propose an architecture using SDN and blockchain technology to avoid false flow rules injection in SDN data layer devices. The blockchain technology provides the controller integrity and authentication of flows circulated between the controller and switches.

Bose et al. [6] propose a mechanism to prevent the threats of DDoS at the switch level by embedding an security using blockchain onto the interaction channels of data and control planes.

3 Proposed Solution

In the literature approach, we suppose that the problem(inject false data attack) comes only from the controllers. But in reality, we can not be sure that this attack maybe comes from other nodes in the architecture (switches). To solve the problem encountered by the literature approaches, we propose a new approach to determine the source of the problem. So, we add the switch in the process of the reputation mechanism. For every test, we increase or decrease the reputation of the different devices (Controller or Switch). Figure 1 presents a multicontroller SDN network.

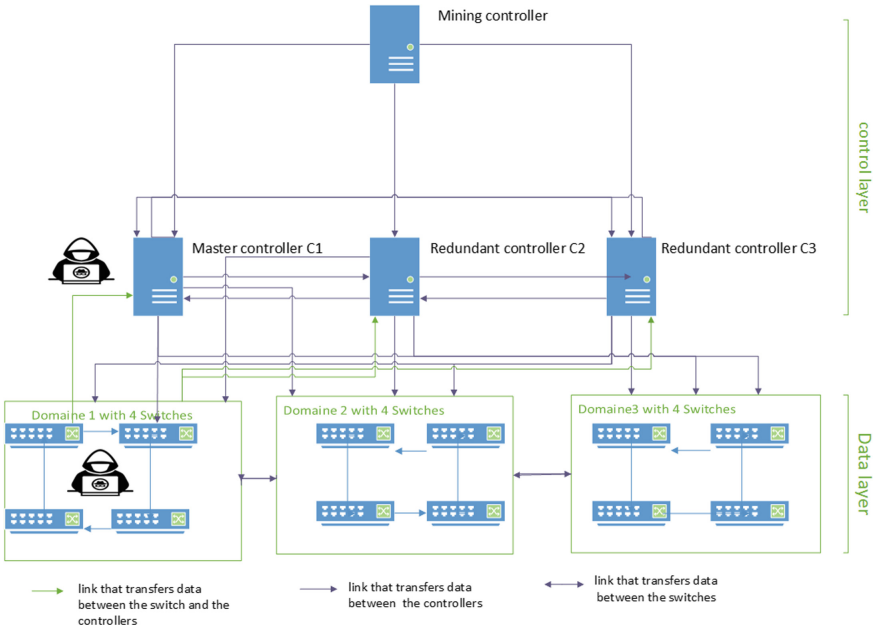


Fig. 1. The used architecture of multicontroller SDN.

3.1 The Principal Idea

The principal idea is to consider the reputation value of the switches before deciding to decrease or increase the reputation value of the master. So, we explore the Flow of the switch in a specific interval (Redundant or not Redundant). To understand from where the fails arrived (did the fails arrive through the master controller or the switch), to take the appropriate decision.

3.2 Mining Decision Process

The diagram sequence in Fig. 2 presents the mining validation process of the master controller. We describe this process as follow:

(1) Switch sends a request to the master, to the two redundant controllers, and the mining.
(2) Master creates a block and then sends it to the two redundant controllers.
(3) The two redundant controllers create new blocks.
(4) The two redundant controllers validate the two blocks with the block sent by the master.
(5) Mining collects the information from the miners (redundant controllers) and according to the result of the miners the mining takes a decision: If the consensus is reached the block will be validated and shared with the other controller from the Blockchain, Else the mining determines the source of

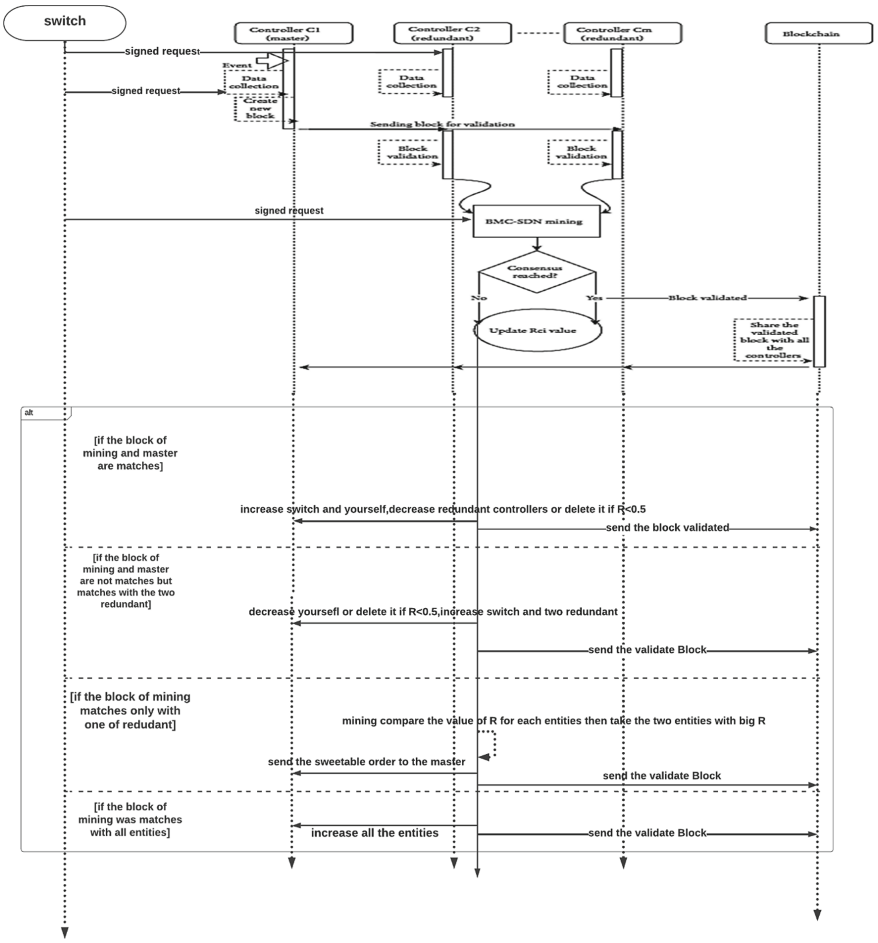


Fig. 2. Diagram sequence of the mining decision process

the problem. If the switch is the source of the problem, the block will be validated and shared with the other controller from the Blockchain. If the master is the source of the problem the block is rejected.

(6) The mining updates the trust values (R) of all nodes. If the new trust values (R) of the source of the problem (master or switch) is <0.5 the mining decides to replace the source of the problem.

3.3 The Reputation Process

Mining collects the information from the miners (redundant controllers) and according to the results of the miners, the mining takes a decision. If the consensus is not reached, the block is considered invalid and the mining does not

append this block to the blockchain. In this case, we have a different scenario as follows:

- When the block computed by each redundant controllers (c2,c3) match with the block computed by the master (c1).
- When the block computed by each redundant controllers (c2,c3) not match with block computed by the master (c1).
- when the block computed by the c2 match with the block computed by the master and the block computed by c3 not match with the block computed by the master.
- when the block computed by the c3 match with the block computed by the master and the block computed by c2 not match with the block computed by the master.

These different scenarios are considered in the Reputation Function 3.

4 Evaluation

We present in this section an evaluation of the proposed idea. The purpose of these evaluation is to highlight the gain brought by using the proposed idea

```
Function Reputation (var TMining: Table, TBlock: Table, Dk, SB: integer): Table
Variables o, p: integer;
        R: reel;
Begin
        o=Dk-1
          p=10*SB-1
            If TBlock [o, p] == valid and TBlock [o+1, p] ==valid
            and TBlock [o+2, p] ==valid and TBlock [o+3, p] ==valid or invalid
            R. Calcul (var TMining, TFlow)
            Else If TBlock [o, p] ==invalid and TBlock [o+1, p] ==valid
            and TBlock [o+2, p] ==valid and TBlock [o+3, p] ==valid or invalid
            R. Calcul (var TMining, TFlow)
            Else If TBlock [o, p] ==valid and TBlock [o+1, p] ==valid
            and TBlock [o+2, p] ==invalid and TBlock [o+3, p] ==valid or invalid
            R. Calcul (var TMining, TFlow)
            Else If TBlock [o, p] ==valid && TBlock [o+1, p] ==invalid
            and TBlock [o+2, p] ==valid and TBlock [o+3, p] ==valid or invalid
            R. Calcul (var TMining, TFlow)
            Else If TBlock [o, p] ==invalid && TBlock [o+1, p] ==invalid
            and TBlock [o+2, p] ==invalid and TBlock [o+3, p] ==valid or invalid
            R. Calcul (var TMining, TFlow)
End
```

Fig. 3. Reputation function

compared to the existing one. We consider for experimental evaluation only one domain (master controller c1, two redundant controllers c2, c3 and one switch).

4.1 Algorithm Mining Process

Figures 4 and 5 present the algorithm mining process of our approach.

```
Type cellule=STRUCTRE
  R, RP, RPi, Ri: real r, r1, r2, r3: Boolean
  P, Pi: Integer
Fin STRUCTURE
Type Mining=STRUCTURE
  C1, C2, C3, S1, D1, Ci, So, Dk: String
End STRUCTURE
Type Flow=STRUCTURE
  C1, C2, C3, S1: String
End STRUCTURE
Type Block=STRUCTURE
  R, r1, r2, r3: String
End STRUCTURE
```

Fig. 4. Structure declaration

4.2 Evaluation Results

Figure 6 shows the different values of trust R of the three controllers (mining does not consider the reputation of the switch) according to the reputation mechanism proposed in [2]. Every 10 blocks (block0: initialization, block10: valid, block20: valid, block30: valid, block40: valid, block50: invalid, block60: invalid, block70: valid, block80: valid), we change the trust value of all the nodes. We notice that when the block of the master controller is valid, we increment the value of R of the other controllers. We also notice that there is a causal relationship between the three controllers: we increment all the controllers or decrement them according to the mining validation process.

The results presented in Fig. 7 shows the reputation values of the switch after each 10 blocks (block0: initialization, block10: valid, block20: valid, block30: valid, block40: invalid, block50: invalid, block60: invalid, block70: valid, block80: valid). We note that when block 40, 50 and 60 are invalid in mining($R < 0{,}5$ see Fig. 6), the R of the switch is <0,5 for the same block 40, 50 and 60.

```
Algorithm MiningProcess

Variables

    i, j, M, N: integer, W3 ←0.8: reel, W1 ←0.2: reel, RT1 ←0: integer, RT2 ←1: integer,
Topo, Add, Rep: Table
    Tableau TMining [M, N]: Mining
    Tableau TFlow [M, N]: Flow
    Tableau TBlock [M, N]: Block
    bi: Table

Begin \*the beginning of algorithme
    write ("the proposed topology", topo);
    Add←  Addition (TMining)
   Write ("the continues of the table TMining:", Add);
    For each (domain Dk) do
      For each (TBlock bi) do
        If (the master is validated) then
          Add to the blockchain the block;
        End if
       For each (10 Blocks) do
         Reputation (TMining, TBlock)
        End for
      End for
    End for
     Rep←  Reputation (TMining, TBlock) \\* the different validation process of reputation
mechanism is presented in section 4.4
     Write ("the up date of all the devices is:", Rep);
End \* end of the algorithme
```

Fig. 5. Algorithm mining process

So, according to the proposed reputation mechanism, we can conclude that the source of the problem is the switch, not the master controller. The mining validation process of our reputation mechanism does not reduce the reputation value of the controllers. So, our mining decision process is more trustworthy than the process presented by Abdelouahid Derhab and al. in [2].

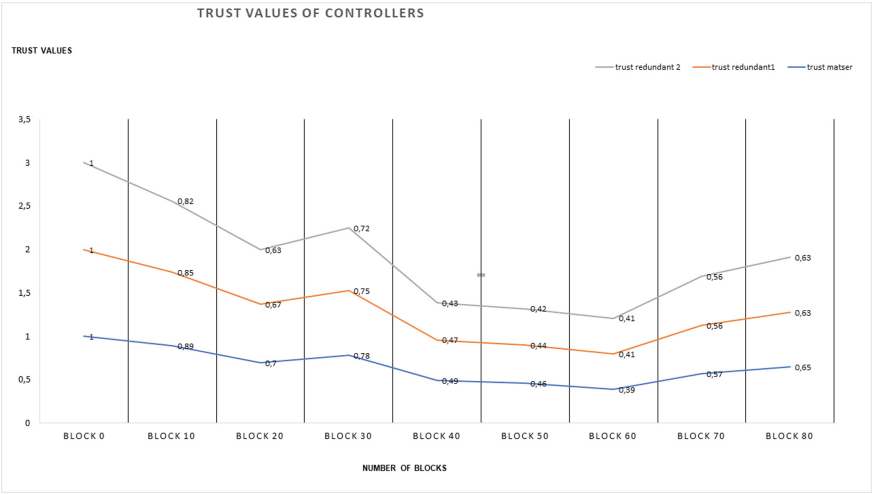


Fig. 6. Trusts values of the controllers

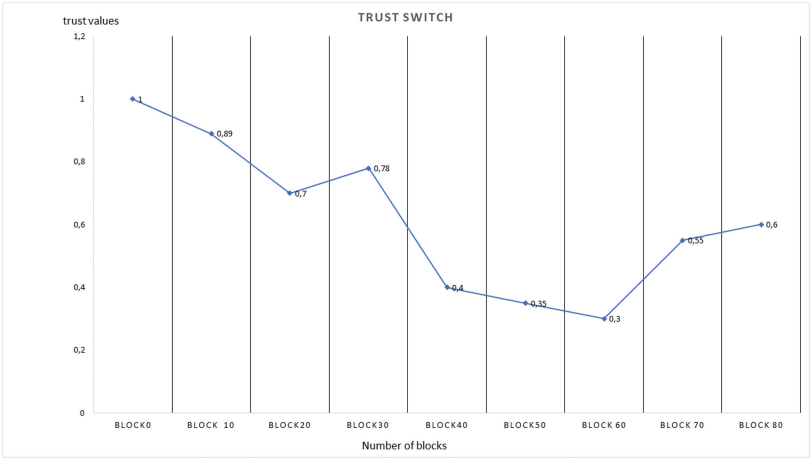


Fig. 7. Trusts values of the switch

Figure 8 shows the difference between the trust c1, according to the reputation mechanism proposed by [2] and the proposed one. When blocks 40, 50, and 60 are invalid in mining, the reputation value of the master (c1) is >0,5. Indeed, the problem comes from the switch, not from the master controller. Figure 7 shows that for the blocks 40, 50, and 60 the reputation values of the switch is <0,5. In this case the mining decides to eliminate the source of the problem (the

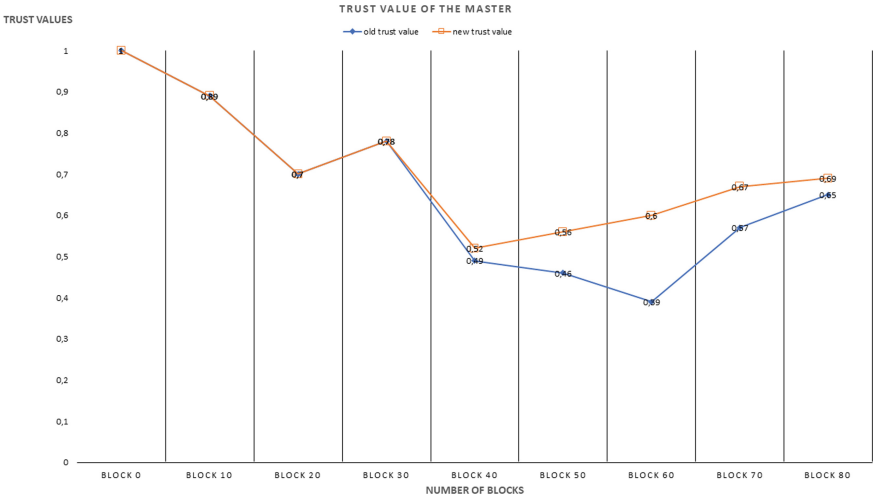


Fig. 8. Trusts values of the master

switch) and increases the reputation value of the master. However, in the case of the mechanism presented in [2] the mining do not detects that the source of the problem is the switch and decreases the reputation value of the master (see Fig. 6 old trust value). So, our mining decision process is more trustworthy than the process presented by Abdelouahid Derhab and al. in [2].

5 Conclusions

In this paper, we have presented an architecture to secure software defined networks. In this architecture, The master controller creates block, the redundant controllers validate, and the mining decides to transfer the block to the blockchain or reject it according to the reputation values of nodes. The reputation mechanism employs fading reputation strategies to manage and customize malicious nodes (controllers or switches) to detect the source of the false data injection. The comparative study between the existing mining decision process and the proposed one clearly specifies the trust value of each node, and the detection of the source of the fraudulent flow rules injection is more trustworthy. In future work, we will propose a simulation results of the proposed solution.

References

1. Li, P., Guo, S., Wu, J., Zhao, Q., Valenza, F.: BlockREV: blockchain-Enabled Multi-Controller Rule Enforcement Verification in SDN. Wiley, Security and Communication Networks (2022)
2. Derhab, A., Guerroumi, M., Belaoued, M., Cheikhrouhou, O.: BMC-SDN: blockchain-based multicontroller architecture for secure software-defined networks. Wirel. Commun. Mob. Comput. (Wiley) (2021)

3. Tong, W., Tian, W., Dong, X., Yang, L., Ma, S., Lu, H.: B-SDN: A novel blockchain-based software defined network architecture. In: International Conference on Networking and Network Applications (NaNA), pp. 206–212 (2020)
4. Latah, M., Kalkan, K.: DPSec: a blockchain-based data plane authentication protocol for SDNs. In: Second International Conference on Blockchain Computing and Applications (BCCA), pp. 22–29 (2020)
5. Boukria, S., Guerroumi, M., Romdhani, I.: BCFR: blockchain-based controller against false flow rule injection in SDN. In: IEEE Symposium on Computers and Communications (ISCC), pp. 1034–1039 (2019)
6. Bose, A., Aujla, G.S., Singh, M., Kumar, N., Cao, H.: Blockchain as a service for software defined networks: a denial of service attack perspective. In: IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 901–906 (2019)

Implementing Autoencoder Compression to Intrusion Detection System

I Gede Agung Krisna Pamungkas, Tohari Ahmad(✉), Royyana Muslim Ijtihadie, and Ary Mazharuddin Shiddiqi

Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
tohari@if.its.ac.id

Abstract. Intrusion Detection System helps catch dangerous incoming packets to a computer network. Despite its substantial role, achieving a high-performance system is still challenging, which may cause insecure whole computer networks. Therefore, a reliable system, which can recognize such threatening attacks, is required. Inspired by previous studies, in this research, we take the Autoencoder to extract and reduce the packet's features and then analyze the results using the Softmax classifier. Furthermore, we remove dropouts and increase the number of batches. The experimental results show that this approach can improve the system's performance. Nevertheless, some works should still be carried out in the future to increase the system's reliability.

Keywords: IDS · Autoencoder · Intrusion · Feature Extraction · Network Security · Network Infrastructure

1 Introduction

For many years, almost all applications have utilized computer network technology, connecting a device to others. This configuration has made data transfer more accessible and faster, which can help many applications to run, such as online meetings and data sharing. Nevertheless, it is vulnerable to attacks, which can destroy the whole system, causing data leakage and costing much to recover. Therefore, a scheme that can prevent computer networks from attacks is needed, one of which is the Intrusion Detection System (IDS). It must be able to recognize unsafe packets, which can be done by analyzing the activity within the network. An IDS works by detecting incoming packets to the network, analyzing them, and deciding whether they are harmful to the system.

IDS generally can be grouped into network-based IDS (NIDS) and host-based IDS (HIDS). A NIDS observes the computer network infrastructure and analyzes its activity represented by the packet flow in the network. Specifically, the packet's header and body are evaluated for possible attacks. Differently, a HIDS focuses on monitoring the system activity of a machine where it has been installed. Furthermore, an HIDS checks the files' integrity and finds suspicious activity based on the log files [1].

Both IDS types evaluate the network packets using one of two approaches: signature and anomaly. In the first approach, the IDS relies on the signatures of existing attacks

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1235–1243, 2023.
https://doi.org/10.1007/978-3-031-27409-1_113

and rules specified by the security administrator. It only compares the incoming data with those stored patterns; thus, it cannot detect unknown attacks. Differently, in the second approach, IDS finds how far the data deviate from typical behaviours. This IDS type usually employs machine learning to recognize intrusion [1].

Many machine learning algorithms applied to IDS focus on feature extraction or classification, such as that done by Megantara and Ahmad [2], which implemented ANOVA for extracting features. In the previous research, Pamungkas et al. took Autoencoder for the feature extraction process, and Softmax, *k*-NN, Naïve Bayes for classifying the features [3]. Based on that research [3], in this paper, we design a scheme to improve its performance by still using Autoencoder with the Softmax classifier, whose results are also compared to others [4].

The remaining parts of this paper are organized as follows. Section 2 studies the implementation of Autoencoder applied to IDS for selecting features. Section 3 describes the proposed method, while Sect. 4 provides the experimental results and compares them with other research. Finally, the conclusion is given in Sect. 5.

2 Previous Works

Some IDS has implemented Autoencoder to extract features, such as that was done by Farahnakian and Heikkonen [4], and Potluri and Diedrich [5]. In their research, Autoencoder compresses data and recovers them using an encoder and a decoder, respectively. The encoder generates codes whose number of features is fewer than the original ones, while the decoder extracts them as closely as possible from the initial ones, as illustrated in Fig. 1.

Autoencoder comprises code, encoder, and decoder. The encoder compresses the given input to construct code, which is lower dimensional output. On the contrary, the decoder extracts this input [4], translating the data, so there must be a loss, which should be kept minimal. Consequently, the output generated by Autoencoder may be slightly different from the input. Furthermore, it transforms categorical data taken from the KDDCup99 dataset to numerical using one-hot encoding. This approach increases the feature numbers. For example, FTP, HTTP, HTTPS and SMTP can be mapped to (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1), respectively. They use Autoencoder architecture to extract features with four layers (32, 32, 32, 32), and Softmax for classifying both the multi and binary classes.

In other research, Potluri and Diedrich [5] also transform categorical to numeric data. Differently, they take the NSL dataset, employing label encoding, resulting in the same number of features. For example, FTP, HTTP, HTTPS and SMTP are converted to 1, 2, 3, and 4, respectively. Additionally, the Autoencoder architecture is implemented for 2-layers (20, 10) feature extraction, and the Softmax is for classification, which is divided into four types, one binary and three multiclass. The first multiclass classification comprises 1 normal, 1 attack type (DoS) and 1 remaining attacks; the second also consists of 3 classes, 1 normal, 2 attack types (DoS, Probe) and 1 remaining attack; the third multiclass is 1 normal and 4 attack types.

Still using the NSL, Al-Qatf et al. [6] also implemented one-hot encoding to convert categorical data, similar to [4]. Here, they only implemented 1 layer of Autoencoder

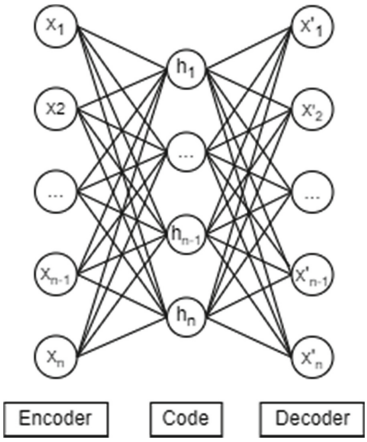


Fig. 1. General Autoencoder Model

architecture for the feature extraction. The SVM classifier is applied for binary and multiclass classification. They evaluated various compression scales, between 10 and 120. In the following research, Larsen et al. [7] developed multipath neural network-based IDS. They find the method can detect spoofing attacks. Specific IDS for mobile environments and cloud computing has been investigated [8]. It is claimed that the system has a high performance. In addition, they show that derived features improve the system's capability to detect attacks. A protocol-based IDS is proposed [9] and implemented in the application layer. It does not require collecting the network traffic and can recognize attacks in unknown networks.

Based on those studies, it is found that Autoencoder has been implemented in various schemes. This includes the layer number and classification type implemented in different datasets. This Autoencoder-based IDS research can be improved further.

3 Autoencoder for IDS

Here, the Autoencoder is implemented to decrease the feature number to 25%, 50%, and 75% of the initial number, whose primary process flow, as in [3], is described as follows (see Fig. 2).

3.1 Preprocessing

The one-hot encoder converts categorical into numerical data **2**with binary coding, causing an increase in the number of features; each category transforms into a new column and only has 0 and 1 values for that out and in the category, respectively. Then the MinMaxScaler is applied to normalize the data.

3.2 Autoencoder Process

This is the focus of this research, in which the Autoencoder is trained to produce the model for training and testing. This method compresses the features to 75%, 50%, and

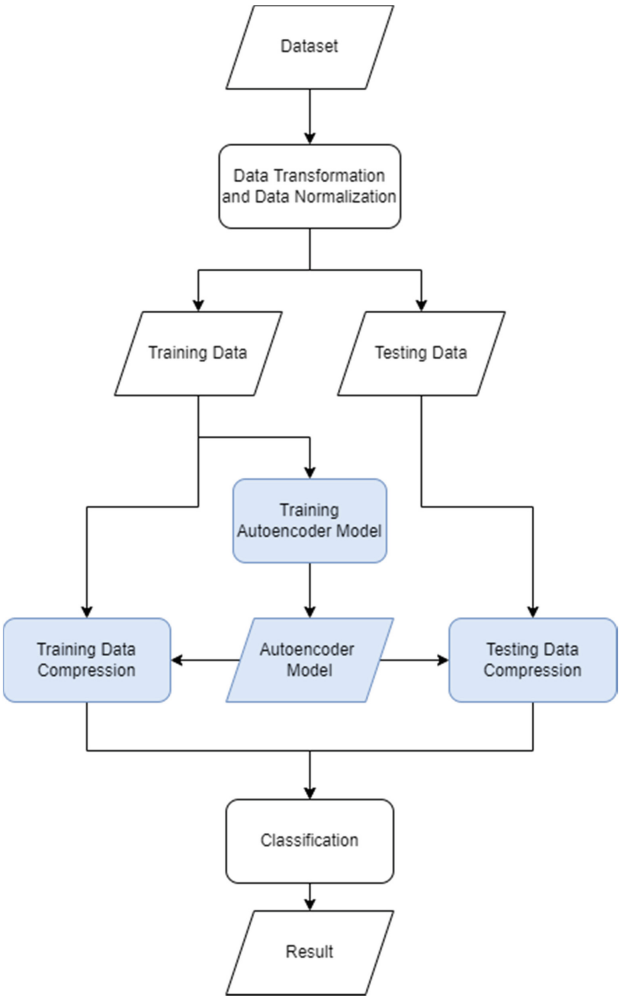


Fig. 2. Primary process flow, inspired by [3]

25%, each comprising three different layers, which are one-layer, two-layer, and three-layer. In the beginning, the first Autoencoder is trained using the training data, which results in compressed data. This process is repeated to the following Autoencoder using the data compressed in the previous Autoencoder, resulting in the second compressed data. Next, the third Autoencoder is trained using these second compressed data.

After Autoencoder has been trained, there are three Autoencoders, consisting of one, two, and three layers. Next, they are trained with previously specified compressed ratios. Compressing training and testing data with trained Autoencoder produces fewer features than the initial data used for the classification.

3.3 Evaluation Process

The performance is measured using the Softmax classifier, having two layers. The neuron number of the first layer is the same as that of compressed data, while the second layer has the same as that of classes. Unlike [3], which uses dropout with 100 batches, this method is without dropout and is performed in 250 batches. This improved classification model is shown in Fig. 3.

4 Experimental Result

Similar to other research, the experiment uses NSL-KDD, UNSW-NB15, Kyoto, and KDDCup99 datasets, evaluating multiclass and binary classification, except the Kyoto dataset, which only has binary data. Furthermore, this study defines True Positive (TP) as the correct detection of an attack, True Negative (TN) as the correct detection of normal, False Positive (FP) as the incorrect detection of normal, and False Negative (FN) as the incorrect detection of attack packets.

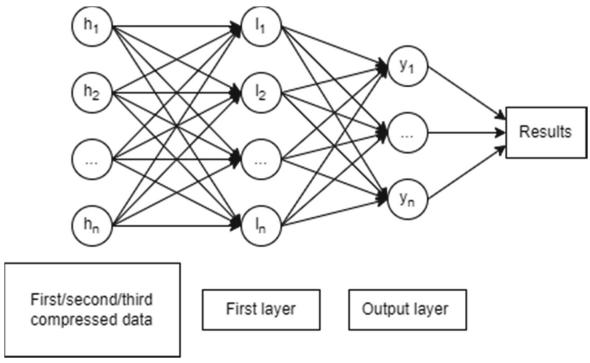


Fig. 3. Classification model without using dropout, improving that of [3]

Same as [3], KDDTrain+.txt, which has 125973 data and KDDTest+.txt, which has 22544 data, are taken for training and testing, respectively. Initially, NSL-KDD comprises 23 classes, categorized into 5, consisting of one normal and four attacks. These are Use to Root (U2R), Denial of Service (DoS), Remote to Local (R2L), and Probe. Meanwhile, the Kyoto dataset taken from 20151231.txt contains 309068 data. As a binary, this dataset has two classes: normal and attack, denoted as 1 and -1, respectively.

The third dataset is UNSW-NB15, taken from UNSW_NB15_training-set.csv for training with 82332 data, and UNSW_NB15_testing-set.csv for testing with 175341 data. This dataset generally comprises ten classes, consisting of one normal and nine attack types.

The last dataset, KDDCup99, is from kddcup.data_10_percent_corrected with 494021 data and corrected with 311.029 data, for training and testing processes, respectively, similar to [3]. The KDDCup99 has redundant data, and after being cleaned, it is

Table 1. The Experimental Result Implementing 25% Compression

Dataset	Layer	Binary			Multiclass		
		Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
NSL-KDD	1	77.20	77.20	65.50	79.09	75.55	92.21
NSL-KDD	2	78.60	78.60	68.38	82.34	79.32	95.03
NSL-KDD	3	80.78	80.78	69.72	83.10	81.08	90.58
UNSW-NB15	1	87.90	82.84	97.38	66.61	98.82	80.61
UNSW-NB15	2	87.78	85.50	97.08	67.29	98.89	80.65
UNSW-NB15	3	87.48	85.38	96.86	67.06	99.61	79.50
Kyoto	1	99.44	99.82	94.64	–	–	–
Kyoto	2	99.48	99.92	94.05	–	–	–
Kyoto	3	99.18	99.59	94.12	–	–	–
KDDCup99	1	94.79	89.60	97.97	92.92	92.44	97.15
KDDCup99	2	94.54	89.03	97.92	92.85	92.66	96.73
KDDCup99	3	95.44	89.95	98.80	94.29	93.66	97.97

becoming 145586 for training and 77291 for testing. Furthermore, the available classes of KDDCup99 are the same as the NSL-KDD dataset, as both have the same configuration.

4.1 Compression

As previously described, this experiment uses 25%, 50%, and 75% compression of Autoencoder and some layers, similar to [3], as in the following experimental scenarios.

Scenario 1. The Autoencoder compresses each dataset by 25%, reducing the number of their features. As in [3], in NSL-KDD, this number decreases the amount of data from 122 to 30; in UNSW-NB15, it is from 196 to 49; in Kyoto, it is from 43 to 11; and in KDDCup99, it is from 119 to 30. These experimental results are given in Table 1 with binary and multiclass classification.

In this scenario, it is shown that the best accuracy on binary classification is 99.48%, reached by the two-layer Autoencoder using the Kyoto dataset. Furthermore, that of multiclass is 94.29% from 3 layers autoencoder using the KDDCup99 dataset.

Scenario 2. In this scenario, the Autoencoder compresses data by 50%. By using that parameter value, the number of features of NSL-KDD, UNSW-NB15, Kyoto, and KDDCup99 drops respectively from 122 to 60, from 196 to 98, from 43 to 21, and from 119 to 60, as shown in Table 2.

Different from the previous scenario, here, the highest accuracy on binary classification is obtained by two layers of Autoencoder using the Kyoto dataset, which is 99.83%, and that of multiclass is in 3 layers of Autoencoder using the KDDCup99 dataset, which is 93.56%.

Table 2. The Experimental Result Implementing 50% Compression

Dataset	Layer	Binary			Multiclass		
		Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
NSL-KDD	1	79.41	66.27	96.77	80.24	77.91	94.69
NSL-KDD	2	78.67	65.19	96.49	82.48	83.11	94.89
NSL-KDD	3	78.90	66.22	95.66	82.34	80.21	94.96
UNSW-NB15	1	89.20	85.50	97.08	68.28	98.90	80.99
UNSW-NB15	2	89.05	85.38	96.86	68.15	99.09	80.69
UNSW-NB15	3	88.55	84.54	97.10	68.58	99.08	80.40
Kyoto	1	99.83	99.95	98.42	–	–	–
Kyoto	2	99.83	99.97	98.18	–	–	–
Kyoto	3	99.80	99.93	98.16	–	–	–
KDDCup99	1	93.31	85.44	98.13	93.15	91.09	97.32
KDDCup99	2	93.21	85.47	97.96	93.22	89.66	97.91
KDDCup99	3	93.60	85.29	98.70	93.56	91.37	97.24

Scenario 3. In scenario 3, the compression is 75%. Here, the number of features of NSL-KDD, UNSW-NB15, Kyoto, and KDDCup99 decreases respectively from 122 to 90, from 196 to 147, from 43 to 32, and from 119 to 90, as depicted in Table 3.

Here, the binary classification reaches 99.87% of accuracy, obtained by 3 layers Autoencoder using the Kyoto dataset, and that of multiclass is 93.47% in 2 layers autoencoder using the KDDCup99 dataset.

4.2 Overall Performance

In NSL-KDD, the highest accuracy in binary and multiclass is when the data are compressed by 25%, with three layers of Autoencoder reaching 80.78% and 83.10%, respectively; for UNSW-NB15, they are obtained from 3 layers of Autoencoder compressed by 75%, which are 89.34% and 69.63%, respectively. Almost similar to those results, in the Kyoto dataset, the best accuracy is compression by 75%, with a three-layer Autoencoder of 99.87%. Furthermore, in KDDCup99, the highest accuracy in binary and multiclass is compression by 25%, with three layers of Autoencoder reaching 95.44% and 94.29%, respectively.

From those three scenarios, the highest accuracy in binary is of 3-layer Autoencoder and the features reduced to 75% with the Kyoto dataset. In multiclass, it is an autoencoder in 3 layers and features reduced to 25% using the KDDCup99 dataset. For the subsequent evaluation, we compare the experimental results with [3–5], and [6], as shown in Table 4 for binary, and Table 5 for multiclass classification.

Tables 4 and 5 show that the method is more applicable to binary classification with an accuracy, sensitivity, and specificity of 99.87%, 99.98%, and 98.44%, respectively, using

Table 3. The Experimental Result Implementing 75% Compression

Dataset	Layer	Binary			Multiclass		
		Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
NSL-KDD	1	79.18	66.47	95.98	79.59	75.18	95.49
NSL-KDD	2	80.12	67.63	96.63	78.43	76.89	95.46
NSL-KDD	3	80.11	67.86	96.31	80.43	77.76	95.68
UNSW-NB15	1	87.60	82.78	97.86	68.86	98.62	81.39
UNSW-NB15	2	88.77	84.73	97.39	69.48	98.96	81.00
UNSW-NB15	3	89.34	85.84	96.79	69.63	98.94	81.02
Kyoto	1	99.87	99.97	98.63	–	–	–
Kyoto	2	99.84	99.97	98.31	–	–	–
Kyoto	3	99.87	99.98	98.44	–	–	–
KDDCup99	1	93.21	85.18	98.13	93.38	89.76	98.29
KDDCup99	2	92.72	83.79	98.20	93.47	88.29	98.46
KDDCup99	3	93.30	84.94	98.43	93.28	89.78	97.70

Table 4. Performance of binary classification obtained by some research

Research	Accuracy (%)	Sensitivity (%)	Specificity (%)	Dataset
Pamungkas et al. [3]	99.42	99.95	89.86	Kyoto
Farahnakian and Heikkonen [4]	96.53	95.65	–	KDDCup99
Potluri and Diedrich [5]	–	97.50	96.50	NSL-KDD
Al-Qatf et al. [6]	84.96	76.57	–	NSL-KDD
Proposed method	99.87	99.98	98.44	Kyoto

the Kyoto dataset. In multiclass classification, [4] is the best, whose accuracy is 94.71%, and sensitivity is 94.42% in KDDCup99. It is slightly different from this research, with accuracy and sensitivity of 94.29% and 93.66%, respectively. It is because the number of features we use is 119, different from [4], which is 117. Nevertheless, their method is more complex because using four-layer Autoencoder, compressing 117 features to 32 (32, 32, 32, 32) differs from this research with three-layer Autoencoder compressing 119 features to 30 (30, 30, 30).

5 Conclusion

In this research, it can be inferred that this proposed method generates better results for binary classification using the Kyoto dataset. On the other hand, KDDCup99 has

Table 5. Performance of multiclass classification obtained by some research

Research	Accuracy (%)	Sensitivity (%)	Specificity (%)	Dataset
Pamungkas et al. [3]	93.55	90.63	97.82	KDDCup99
Farahnakian and Heikkonen [4]	94.71	94.42	–	KDDCup99
Potluri and Diedrich [5]	–	97.50	95.00	NSL-KDD
Al-Qatf et al. [6]	80.48	68.29	–	NSL-KDD
Proposed method	94.29	93.66	97.97	KDDCup99

better results on multiclass classification, although it is still lacking compared to [4]. This pattern is similar to [3]. However, this method is the best in terms of the number of layers.

This research also shows that removing dropouts and increasing the number of batches can improve performance. We may use other classification methods in future work to get the best results, with more variant classification and more evaluation. Furthermore, execution time may be considered in the actual implementation since it is crucial in deciding what action must be done once an attack is detected.

References

1. Warzynski, A., Kolaczek, G.: Intrusion detection systems vulnerability on adversarial examples. In IEEE Int. Conf. Innov. Intell. Syst. Appl. INISTA (2018)
2. Megantara, A.A., Ahmad, T.: ANOVA-SVM for selecting subset features in encrypted internet traffic classification. Int. J. Intell. Eng. Syst. **14**(2), 536–546 (2021)
3. Pamungkas, I.G.A.K., Ahmad, T., Ijtihadie, R.M.: Analysis of autoencoder compression performance in intrusion detection system. Int. J. Saf. Secur. Eng. **12**(3), 395–401 (2022)
4. Farahnakian, F., Heikkonen, J.: A deep auto-encoder based approach for intrusion detection system. In: Int. Conf. Adv. Commun. Technol. ICACT, pp. 178–183 (2018)
5. Potluri, S., Diedrich, C.: Accelerated deep neural networks for enhanced Intrusion Detection System. In: IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA (2016)
6. Al-Qatf, M., Lasheng, Y., Al-Habib, M., Al-Sabahi, K.: Deep learning approach combining sparse autoencoder with SVM for network intrusion detection. IEEE Access **6**, 52843–52856 (2018)
7. Larsen, R. M. J. I., Pahl, M.-O., Coatrieux, G.: Authenticating IDS autoencoders using multipath neural networks. In: 5th Cyber Security in Networking Conference (CSNet) (2021)
8. Faber, K., Faber, L., Sniezynski, B.: Autoencoder-based IDS for cloud and mobile devices. In: IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid) (2021)
9. Huang, Y.-L., Hung, C.-Y., Hu, H.-T.: A Protocol-based Intrusion Detection System using Dual Autoencoders. In: IEEE 21st International Conference on Software Quality, Reliability and Security (QRS) (2021)



Secure East-West Communication to Authenticate Mobile Devices in a Distributed and Hierarchical SDN

Maroua Moatemri(✉), Hamdi Eltaief, Ali El Kamel, and Habib Youssef

University of Sousse, ISITCom, PRINCE Research Lab, ISITCom Route principale n 1, Hammam, Sousse 4011, Tunisia
marouamoatemri@gmail.com, hamdi.eltaief@issatso.rnu.tn, ali.kamel@isitc.u-sousse.tn, Habib.youssef@fsm.rnu.tn

Abstract. With the evolution of networks, the number of mobile devices in the networks is continuously increasing. While moving from one SDN domain to another, any mobile device must keep having access to the network services without interruption. To make that possible, the users' authentication information must be known by the controllers. To do so, the controllers must communicate the users' authentication information in a secure environment to keep them protected against any possible threat. we present in this paper a secure east-west communication to authenticate mobile devices in a distributed and hierarchical SDN. we present a comparison study between the flat architecture and the hierarchical one according to the authentication delay. The results shows that the hierarchical SDN architecture provides a secure East-West communication and minimize the used authentication delay of the mobile devices.

Keywords: East-west communication · Authentication delay · SDN · Multi-controller

1 Introduction

The field of computer networks is an innovating field. Due to their rigid architectures and complex management, traditional networks are no longer able to handle the new needs introduced with the permanent evolution of networks. Thus, a new architecture of networks has been introduced, Software Defined networks (SDN), which will allow a dynamic and flexible management of the networks [5]. A software defined network separates the control plane and the data plane [3]. An SDMN has been defined [2] to define the SDN to wireless mobile networks. Network services are time sensitive and every second has a huge importance to guarantee a good QoS. Thus, a mobile user needs to have continuous access while moving freely within the network. To be able to do so, the mobile authentication information must be shared, in a minimal time, between the controllers. In this paper, we propose a secure east-west communication to authenticate the mobile

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1244–1253, 2023.
https://doi.org/10.1007/978-3-031-27409-1_114

devices in a distributed and hierarchical SDN. The proposed solution ensure less authentication delay compared to a flat architecture. This paper is organized as follow. Section 2 presents a brief overview of the related work about the different architectures used for a distributed SDN. Section 3 defines the proposed solution to authenticate mobile devices in a distributed and hierarchical SDN. The performance evaluation is in Sect. 4. We conclude in Sect. 5.

2 Related Work

If Only one controller is responsible of all the network. This will cause problems in terms of scalability. With the continuous evolution of the networks, they are becoming larger and larger. Thus, due to its limited capacity, it will be impossible for one controller to manage all the network. Another problem that occurs with the single controller SDN is the single point of failure. So, distributed SDN are proposed [5]. When implementing a distributed SDN, several architectures could be adopted. The first architecture is the flat design. This design consists of several controllers and each of them is responsible of a specific domain. To maintain a general view, all the controllers communicate through their east-west bound. This design solves the scalability problem of the single controller's SDN since it adds more controllers. In another hand, it introduces new problems, such as complicated controller management and the addition of the control overhead due to the frequent communication between controllers. Aiming to solve these problems, the hierarchical design has been proposed [10]. Most solutions propose a two layer controller design: one root controller and several domain controllers. Each domain controller manages its own domain and manages the intra-domain routing operations. Then we have the root controller that handles all the domain controllers and keep a global view of the network. This will allow it to communicate with the domain controllers when needed. He is also responsible of managing the inter-domain rooting operations [8].

3 The Proposed Approach

3.1 Overview

The idea consists of applying secure east-west communication to authenticate the mobile devices in a distributed SDN, which we proposed in [6] on a different network architecture design. We will adapt the idea to a hierarchical design. The proposed architecture is presented in Fig. 1.

We have a hierarchical architecture with three-layer controllers. The first layer consists of controller "C". These controllers will only be responsible for managing their domains and will not save the users' authentication information. The second layer of controllers is composed with domain controllers "DC". Every domain controller is responsible for managing group controller "C" called virtual domain controller set "VDCset". Each DC contains the authentication information of the users belonging to its VDCset. The third layer contains the root controller "RC". The root controller has a global view of the network [1]. Compared

with flat architecture, hierarchical architecture presents several advantages. The first advantage concerns the load of the root controller. Adding the domain controllers will reduce the root controller's load. This is explained by the fact that it will not be the only network entity taking all the decisions. Secondly, the users' authentication information will be distributed to the domain controllers. This will reduce the risk that an intruder steals all the user's information at once.

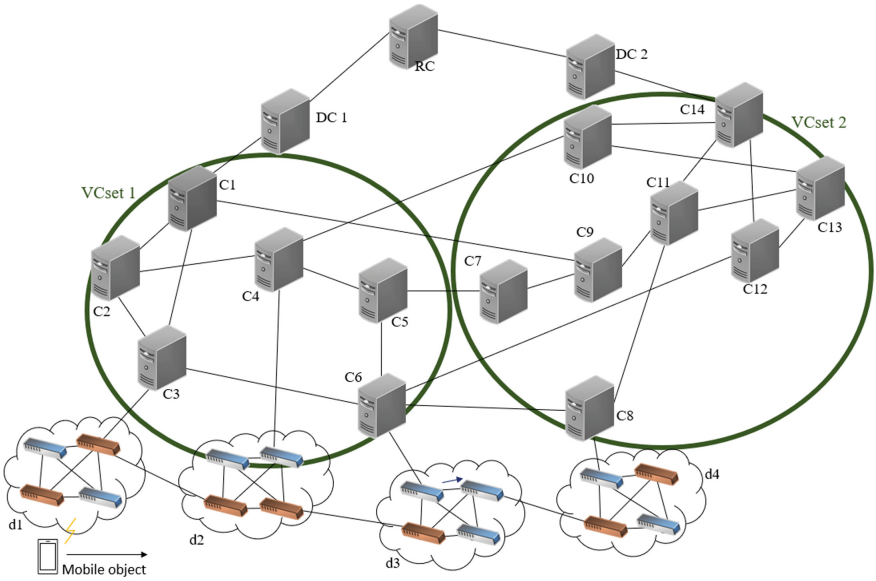


Fig. 1. The proposed architecture

3.2 The Proposed Approach

When a new user enters the network, he will need authentication. This authentication is realized through the Extensible Authentication Protocol protocol [12] and the IEEE 802.1x authentication [9]. The authentication process is done by an authentication server (RADIUS, for example) [7] through a challenge-response authentication. Once the user enters the infrastructure, another authentication will take place to allow him to benefit from the services he needs. The same type of authentication will take place with an authentication server. The first authentication process is described in the Fig. 2. Since this authentication could take a long time, when a user needs to re-authenticate, he will need to use a signed token, MSK (Master Session Key), containing his authentication information [1]. In fact, when a mobile needs to initiate a re-authentication request after moving to a foreign domain, he will need to provide his signed token. The foreign domain will decrypt it using the authentication server's public key. This will

allow him to verify the authenticity of the token. Then, he will extract the MSK and authenticate the mobile. This process is presented in the Fig. 3. When the foreign domain does not belong to the same VDCset, the same process will take place, except that the root controller will be the one decrypting and extracting the MSK. This process called the inter-VDCset re-authentication is presented in Fig. 3b.

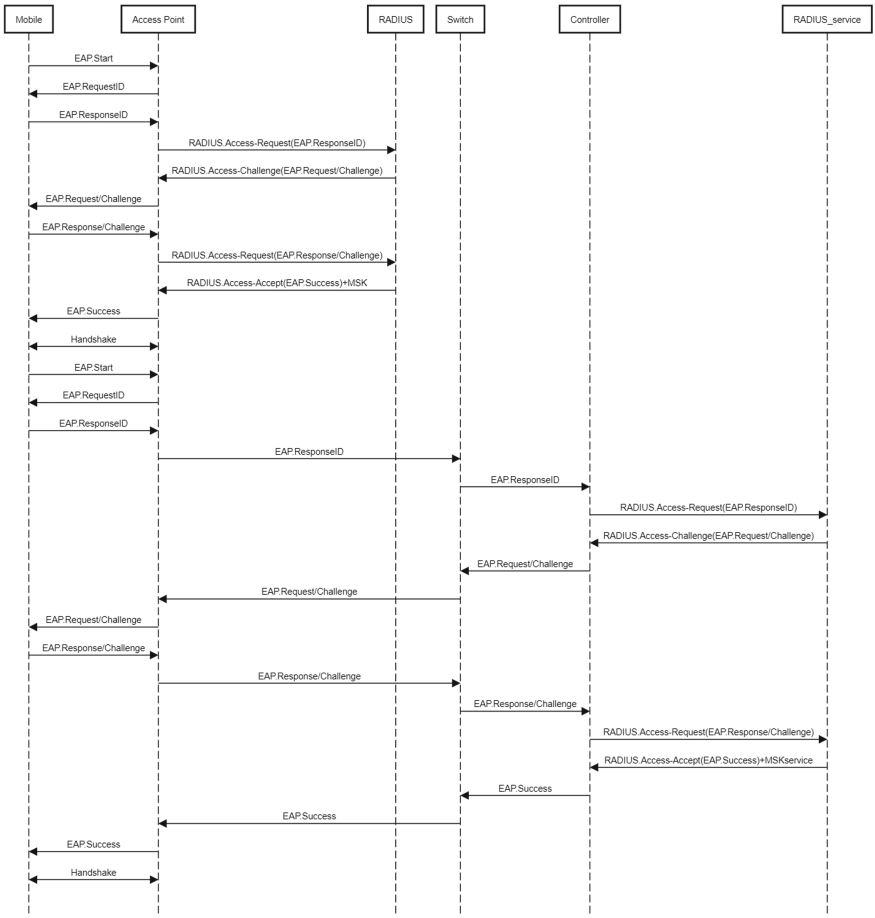


Fig. 2. The first authentication process

4 Performance Evaluation

We present an analytical performance evaluation of The proposed approach according to the authentication delay and it is compared with the flat architecture.

Table 1. Analytical evaluation notation.

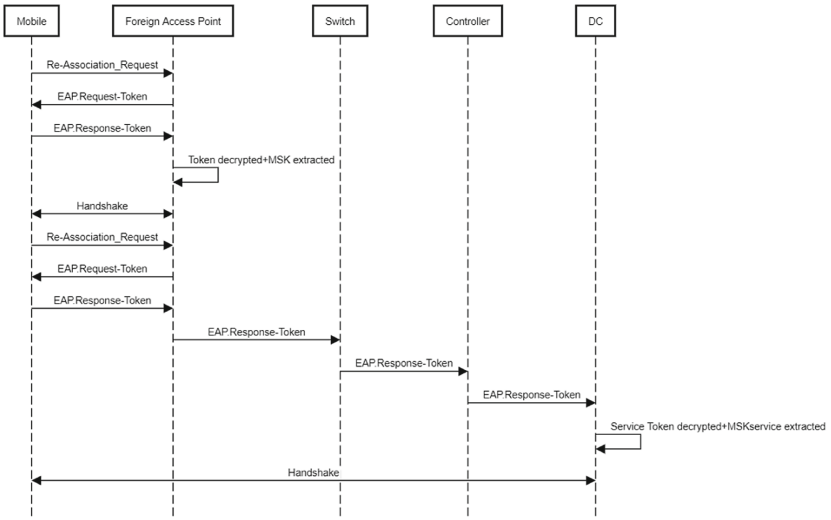
Symbol	Representation
n	The VDCset
Ck	The border controller (C1 or C14)
DC_n	Domain Controller of VDCset n
RC	Root controller
FC	Foreign controller
T_{Access_Auth}	Access authentication time
$T^{Infrastructure}_{Access_Auth}$	Infrastructure access authentication time
$T^{Service}_{Access_Auth}$	Service access authentication time
$T^{Service}_{intraVDCset_n}$	Service access authentication time when moving in the same VDCset
$T^{Service}_{interVDCset}$	Service access authentication time when moving in a different VDCset
Distance (Ci,Cj)	The distance separating two neighbor controllers
Distance (Ck,DC_n)	The distance separating the boarder controller and its DC
distance (DC_n,RC))	The distance separating the DC and the RC
P_n	Shortest path between the foreign controller and the DC
T_{cal}	The total time for encryption, decryption, hash,...
V_{prop}	Propagation velocity
$size_{req}$	Size of the message of the request
$size_{reply}$	Size of the message of the acknowledgement
$datarate$	Transmission speed

4.1 Analytical Evaluation

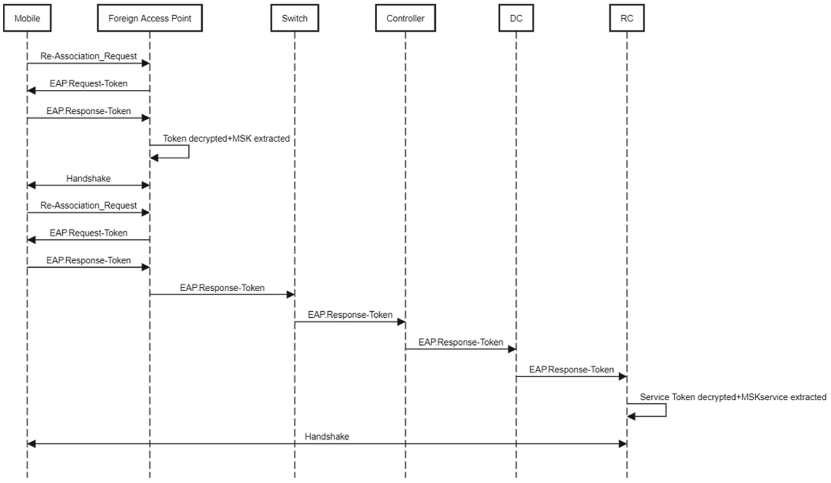
The notation used is presented in the Table 1. Through the diagrams presented in 3.2, we conclude that to calculate the authentication delay, we need to consider two different times, namely, the infrastructure access authentication time , and the service access authentication time. The access authentication time is expressed in Eq. 1.

$$T_{Access_Auth} = T^{Infrastructure}_{Access_Auth} + T^{Service}_{Access_Auth} \quad (1)$$

The service access authentication time is calculated differently depending on the type of movement done by the mobile device. If the mobile device moves to the domain of a foreign controller within the same VDCset, the time to authenticate and access the service by the foreign controller is calculated using the Eq. 2. It is the sum all the communication cost needed to attend to the border controller



(a) Re-Authentication: IntraVDCset



(b) Re-Authentication: InterVDCset

Fig. 3. The re-authentication process

and the cost value of the communication with the DC_n.

$$T^{Service}_{intraVDCset_n} = (\sum_{(C_i,C_j)\in P_n} 2*\frac{(distance(Ci,Cj))}{V_{prop}} + \frac{(size_{req} + size_{reply)}}{datarate} + T_{Cal}) + 2 * \frac{(distance(C_k, DC_n))}{V_{prop}} + \frac{(size_{req} + size_{reply)}}{datarate} + T_{Cal} \quad (2)$$

The time to authenticate the mobile devices by the foreign controller (when a mobile device moves to the domain of a foreign controller in another VDCset) is the sum of the time of all communications cost needed to attend the border controller, the time cost of the communication with the DC_n and, the time cost of the communications with the RC. It is expressed using the Eq. 3.

$$T^{Service}_{interVDCset} = (\sum_{(C_i,C_j)\in P_n} 2*\frac{(distance(C_i,C_j))}{V_{prop}} + \frac{(size_{req} + size_{reply)}}{datarate} + T_{Cal}) + 2 * \frac{(distance(C_k, DC_n))}{V_{prop}} + \frac{(size_{req} + size_{reply)}}{datarate} + T_{Cal} + 2 * \frac{(distance(DC_n, RC))}{V_{prop}} + \frac{(size_{req} + size_{reply)}}{datarate} + T_{Cal} \quad (3)$$

4.2 Experimentation

To test the proposed approach using the hierarchical architecture design and compare it with the flat architecture design used in [6], we will use the NSFNET [11] composed with 14 nodes and 21 links. The proposed architecture with the approximate distances in kilometers [4] is presented in Fig. 4. We suppose that the two domain controllers are equidistant from the border controllers and that the root controller is also equidistant from the two domain controllers. Given that the infrastructure access authentication time is done at the data plane through the access point and $T^{Infrastructure}_{Access_Auth}$ computed for both approaches used for the comparison, we will only consider the service authentication time which takes place within the control plane. The comparison is realized analytically based on the average access authentication time computed with the Eq. 4. It represents the access authentication time in each foreign controller (FC) that the mobile visited during his route by the number of times the mobile changed the domains. We used the equation in [6] to calculate the average time of the flat architecture. Table 2 presents the different scenarios used to compare the two architectures. We refer to the controller where the mobile is registered with source node.

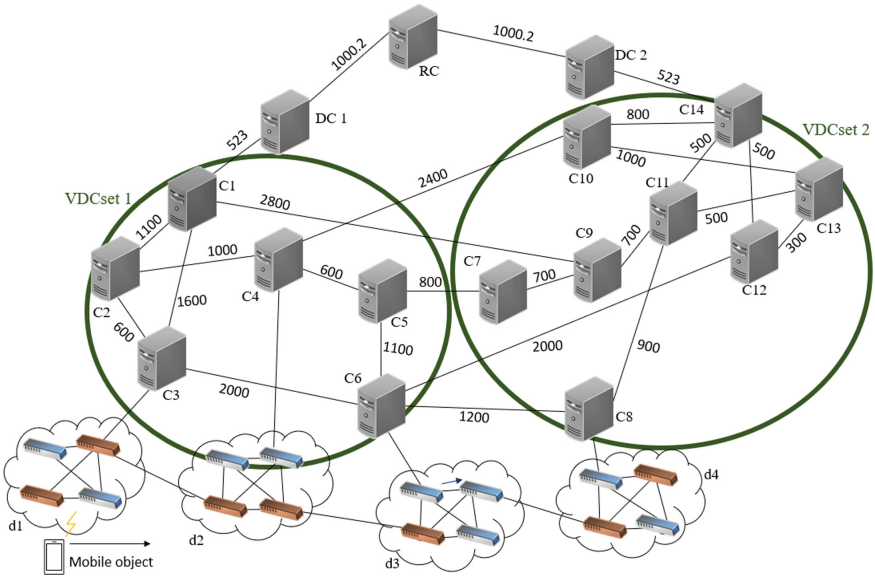


Fig. 4. The proposed architecture with approximate distances

$$Average\ T^{Service}_{Access_Auth} = \frac{\sum_{(FC)\in Route} T^{Service}_{Access_Auth}}{(number\ of\ FC \in Route) - 1} \tag{4}$$

The analytical results are presented in Fig. 5. Our proposed approach presents an improvement in most scenarios compared to the flat architecture approach. The average access authentication time for the proposed hierarchical approach

Table 2. Test scenarios.

Scenario	Source node	Route
1	8	8-6-12-13-11
2	9	10-4-5-7
3	8	8-6-3-2-1
4	9	8-6-3-2-4-5
5	12	2-1-10-14-13-12
6	6	3-5-7-9-1
7	1	3-6-12-14-16-11
8	3	7-5-6-8-11
9	13	6-3-1-10-13
10	5	1-9-7-5-6-3

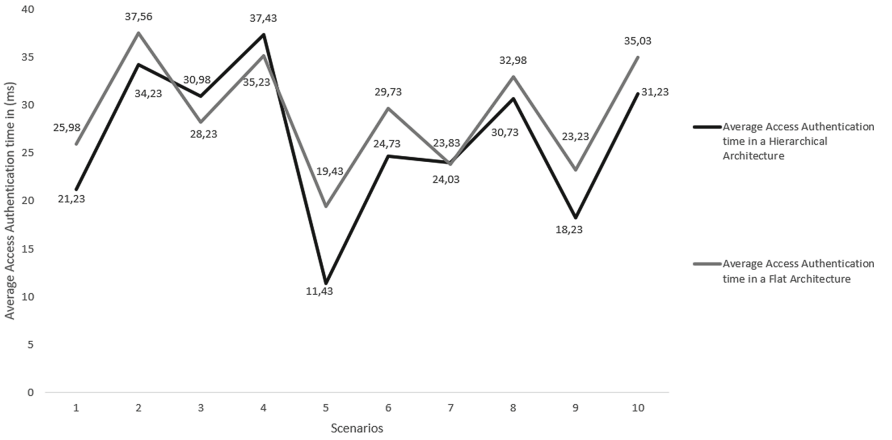


Fig. 5. Analytical comparison of the average access authentication time

is equal to 26.43 ms, compared with 29.12 ms obtained for the flat architecture approach. It presents an improvement of 2.69 ms.

5 Conclusion

The main idea of this paper is to compare our proposed hierarchical approach and the existent flat approach according to the access authentication time of mobile devices in distributed SDN architecture. This time needs to be minimal to ensure a good Quality of Service (QoS) for users without limiting their movements. The experimentation results show that the proposed hierarchical SDN architecture provides a secure inter-controller solution and minimizes the used authentication time of the mobile devices in distributed SDN.

References

1. Aissaoui, H., Urien, P., Pujolle, G.: Low latency of re-authentication during handover: re-authentication using a signed token in heterogeneous wireless access networks. In: 2013 International Conference on Wireless Information Networks and Systems (WINSYS), pp. 1–7. IEEE (2013)
2. Chen, M., Qian, Y., Mao, S., Tang, W., Yang, X.: Software-defined mobile networks security. Mob. Netw. Appl. **21**(5), 729–743 (2016)
3. Foundation, O.N.: Software-defined networking (SDN) definition,
4. Ghose, S., Kumar, R., Banerjee, N., Datta, R.: Multihop virtual topology design in WDM optical networks for self-similar traffic. Photonic Netw. Commun. **10**(2), 199–214 (2005)
5. Hu, T., Guo, Z., Yi, P., Baker, T., Lan, J.: Multi-controller based software-defined networking: a survey. IEEE Access **6**, 15980–15996 (2018)

6. Moatemri, M., Eltaief, H., Kamel, A.E., Youssef, H.: Secure east-west communication to approve service access continuity for mobile devices in a distributed SDN. In: International Conference on Computational Science and its Applications, pp. 283–297. Springer (2022)
7. Rubens, A., Rigney, C., Willens, S., Simpson, W.A.: Remote authentication dial in user service (RADIUS). RFC 2865 (2000)
8. Sarmiento, D.E., Lebre, A., Nussbaum, L., Chari, A.: Decentralized SDN control plane for a distributed cloud-edge infrastructure: a survey. IEEE Commun. Surv. Tutor. **23**(1), 256–281 (2021)
9. Smith, A.H., Zorn, G., Roese, J., Aboba, D.B.D., Congdon, P.: IEEE 802.1X remote authentication dial in user service (RADIUS) usage guidelines. RFC 3580 (2003)
10. Togou, M.A., Chekired, D.A., Khoukhi, L., Muntean, G.M.: A hierarchical distributed control plane for path computation scalability in large scale software-defined networks. IEEE Trans. Netw. Serv. Manag. **16**(3), 1019–1031 (2019)
11. Tomovic, S., Radonjic, M., Radusinovic, I.: Bandwidth-delay constrained routing algorithms for backbone SDN networks. In: 2015 12th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), pp. 227–230 (2015)
12. Vollbrecht, J., Carlson, J.D., Blunk, L., Aboba, D.B.D., Levkowetz, H.: Extensible authentication protocol (EAP). RFC 3748 (2004)



Cyber Security Issues: Web Attack Investigation

Sabrina Tarannum, Syed Md. Minhaz Hossain(✉), and Taufique Sayeed

Department of Computer Science and Engineering, Premier University, Chattogram 4000, Bangladesh
saaabrin@gmail.com, minhazpuccse@gmail.com, taufique.sayeed@puc.ac.bd

Abstract. In these recent days, remote work will likely be the biggest cyber security challenge. Remote employment will still be common because of the numerous COVID-19 requirements. After businesses encouraged remote work owing to pandemic concerns, malicious actors now have an easier time finding insecure or incorrectly configured systems that connect to the internet. Web assaults are actions taken against websites and web-based applications with the intent to steal sensitive data, interrupt web service systems, or seize the targeted web systems. Web attacks are a growingly significant subject in digital forensics and information security. It has been noted that invaders are gaining the capacity to get around safety measures and initiate several complex attacks. One of the biggest obstacles is effectively responding to new and unidentified threats, despite several attempts to handle these attacks using a variety of technologies. The objective of this paper is to review the research on web attacks focusing primarily on attack detection methods on various areas of study. The goal is to do explore on web attack investigation, identification methods focusing on different areas such as vulnerabilities, prevention, detection technologies and protection. This paper also examines several related research problems and potential future paths for online assault detection that could aid in more precise research development.

Keywords: Web attacks · Security · Vulnerabilities · Prevention · Intrusion detection · Machine learning techniques

1 Introduction

Different security issues and cyber assaults have risen at an exponential rate in recent years due to the ever-increasing demand for digitization. Today, a lot of online sites and apps rely on web services to easily communicate information with one another. Web services give companies and people a method to reuse functionality across services by giving them a mechanism to send different types of data across the network. Since many scenarios are controlled by user input, there is a dynamic interaction between the user and the online service. This dynamic nature frequently raises concerns. Millions of Internet users utilize a variety of applications on a regular basis, including e-governments, e-commerce, social networking sites, blogs, content management systems, and web emails, among many others, and for all of these applications, web applications are crucial. According to reports, 92% of web-based applications are weak points, 75% of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1254–1269, 2023.
https://doi.org/10.1007/978-3-031-27409-1_115

information security attacks target web apps, 70% of web-based assaults are successful, and web apps can encounter up to 27 attacks every minute [2]. Hacking private and personal information is the primary motivation behind these assaults. Attacks take place as a result of vulnerabilities in the database server, security measures, and web server. Unauthorized users can also gain administrative access rights to the web application or the server due to poor development and configuration of the web application. Additionally, the design of the Hypertext Transfer Protocol (HTTP) is important if it cannot keep up with the intricate structure of online applications. The contributions of this paper are summarized as follows.

- We build a review and concise discussion on a variety of machine learning system for finding various intrusion detection systems and summarize web attack identification based on different machine learning and deep learning techniques.
- Then we summarize various types of attacks according to our study depending on different focus area such as web security, vulnerability, prevention attacks detection and protection.
- Finally, we go through and briefly review some related research issues and future directions for web attacks detection that might benefit both academics and business professionals to carry out more research and development in pertinent application areas.

2 Background

A web attack uses a website's weaknesses to gain access without authorization, grab private data, upload malicious content, or change the website's content. There is various indication that a web is affected such as, If the end users are unable to access the victim's website, properly entered URLs directing to incorrect websites (spoofing), unusually slow network performance, frequent server reboots, abnormalities in the log files. In this section we give an overview of attack indication, types and methods use for detecting those attacks.

2.1 Web Attacks and Security Risks

The major types of web attacks are Injection Attacks, DNS Spoofing, Session Hijacking, Phishing, Brute force, Denial of service, Dictionary Attack, URL interpretation, File Inclusion Attacks, Man in the Middle Attack. Other well-known security incidents in the realm of cyber security include privilege escalation, password attacks, insider threats, advanced persistent threats, crypto jacking attacks, various web application attacks, etc. A data breach, often known as a data leak, is a type of security incident that involves unauthorized data access by a person, application, or service [31].

2.2 Web Security Resistance Strategies

There are various methods to find the detection of web attacks. Among them host-based intrusion (HIDS) and network based intrusion (NIDS) system are quite familiar. While

an NIDS examines and keeps track of network connections for suspicious activity, a HIDS monitors critical files on a single system. Based on the detection approach, Application Intrusion Detection Systems (AIDS) can be broadly divided into two categories: Signature and Anomaly based Detection System as shown in Table 1.

Table 1. Categories of Application Intrusion Detection Systems (AIDs)

	Characterization	Advantages	Disadvantages
Signature-based detection	Can detect known model of attacks. Attacks are identified via feature matching	A low rate of false alarms and rapid detection	Cannot locate the attack if the signature database does not already contain the predetermined attack
Anomaly-based detection	Based on unusual behavior of the system	Can detect unknown attacks	High number of fake alarm and low accuracy

Beside these, a hybrid intrusion detection system considers both anomaly-based and signature-based detection methods. In a hybrid detection system, signature-based detection is used to identify assaults that follow a known pattern, and anomalous intrusion detection is used to identify brand-new attacks. The basic terms related to web attacks are described in Table 2.

2.3 Machine Learning Task in Web Attack

Data science, deep learning, and computational statistics are all subfields of machine learning (ML), which is frequently referred to as a branch of "Artificial Intelligence." The main goal is to educate computers how to learn from data. Machine learning models frequently consist of a collection of rules, procedures, or sophisticated "transfer functions" in order to find intriguing data patterns or to detect or forecast behavior. These capabilities could be crucial in the field of web security. The following describes the use of machine learning for identifying various types of web attacks.

A WAF does not offer enough protection from zero-day attacks. The effectiveness of machine-learning-based WAFs, which can be used in addition to or replacement of signature-based approaches, was researched and demonstrated by Applebaum et al. The tested system had a 98.8% accuracy rate [11]. To increase attack detection and lower the false alarm rate, Hussein et al. applied machine learning algorithms such Nave Bayes, K-means, and Bayes Net [15]. Betarte et al. demonstrated how machine learning methods might enhance MODSECURITY's recognition capacity by decreasing fake positives and rising true positives [19]. The authors in [17] provide research on how machine learning approaches can be utilized to handle the difficulty and analyze the aberrant behavior connected to phishing online attacks. For more effective and precise web assault detection, Ren et al. used a hidden Markov model and a BOW technique. They demonstrated that Bag of Words has a higher recognition rate, a lesser number of fake alarms, and is less expensive [24].

Table 2. A summary of key terms and areas related to web attacks

Key Terms	Description
Web attacks	An action that undermines the security, integrity, confidentiality, or availability of information is referred to as a web attack. It may also cause harm to the networks that provide the information
Intrusion	An activity that is used for compromising data security of a system
Web Anomaly	Web anomalies are outliers, noise, deviations, and exceptions
Data Breach	The deliberate or accidental disclosure of secure data to a hostile environment is referred to as a data leak or spill
Machine Learning	A key part of artificial intelligence (AI), which is concerned with the study of methods to complete a given task without using explicit instructions
Deep Learning	An important component of AI's machine learning creates security models often using artificial neural networks with multiple data processing layers
Detection Models	Models use features as inputs and apply machine learning algorithms to get a predetermined result for wise decision-making
WAF	Web Application Firewalls. This firewall solution frequently monitors and filters data packets for the presence of malware or viruses
Mod Security	Embeddable Firewalls. With no modifications to the current infrastructure, it shields web applications from a range of assaults and permits HTTP traffic monitoring and real-time analysis
Snort	A system for detecting and preventing intrusions. To detect potentially malicious activity, SNORT employs a rule-based language that blends anomaly, protocol, and signature inspection approaches

2.4 Methods for Identifying Web Attacks

This sector presents some research on approaches for monitoring common assaults on websites and web applications. There are two categories for the studies analyzed: (i) The detecting group using rules, patterns, or signatures (ii) the detection based on anomaly and (iii) a hybrid intrusion detection system.

2.4.1 Detection Based on Signatures

Díaz-Verdejo et al. proposed a method to reduce the false positive in various Signature based intrusion detection system. It is also showed that for a joint decision using various SIDS the accuracy or the detection rate can be improved [14].

2.4.2 Detection Based on Anomaly

Riera et al. projected a systematic review to determine the present state of web anomaly detection technology. Their work shows how anomaly detection methods can be used to stop and identify internet attacks [13].

2.4.3 Hybrid Detection System

The majority of intrusion detection techniques for web attacks are signature-based. However, these approaches may fail to detect unknown threats due to missing attributes or poor profiling. Hussein et al. presented a system that blends both signature- and anomaly-based intrusion detection techniques in order to decrease warnings received and find novel attacks [15]. At first, author used Snort to analyze the dataset. Then for the next step Naïve Bayes, K-means and Bayes Net various algorithms are applied for recognizing anomaly based attacks.

2.4.4 Detecting Web Attacks with Deep Learning

A branch of machine learning is called deep learning, a computational model that takes cues from the biological neural networks in the human brain. In the reference paper, Yao Pan et al. proposed a system that evaluate the viability of both a semi-supervised and an unsupervised strategy for web assault detection. They also used deep learning method for the proposed system. They conclude that with limited domain expertise and labeled training data, the proposed method may effectively and precisely identify threats such as SQL injection, denial of services and cross-site scripting [27].

The summary of the state-of-the-art for detecting web attacks are as shown in Table 3.

3 Research Issues and Future Directions

This study releases several research issues and challenges in the area of web attacks. In the following, we summarize these issues and challenges.

3.1 Hybrid Learning Scheme

The greater part of intrusion detection techniques for web attacks detection is signature-based. However, these approaches may fail to detect unknown threats due to missing attributes or poor profiling. Beside this, in anomaly based detection system the accuracy is not high enough but can detect unknown attacks. So, a hybrid technology combining of signature and anomaly based invasion recognition system or a combination of machine learning and deep learning algorithm can be useful to extract the abnormality from the problem sphere, which can solve the limitations of a specific detection system.

3.2 Generalized Strategy for Web Attacks Detection

How to manage a high volume of incoming traffic when each packet needs to be verified with every signature in the database is another issue with detection techniques. As a result, processing all of the traffic takes a long time and reduces system throughput. Sometimes techniques are so specific that its knowledge depends on particular operating system, version, and application. So generalized technique is needed so that not to tie in specific environments.

Table 3. Summary of review findings

References	Types of Attack	Summary of the Paper	Area of Focus
[1]	Cyber Attacks	Specific web browser forensics methods and suggested workable investigative tools	Vulnerability detection
[2]	Web attacks	Guideline for identifying web attacks	Attack Prevention
[3]	Denial-Of-Service, various injection attacks, spoofing	Dynamic analysis was mostly used to provide solutions, with static analysis coming in second	Attack Prevention
[4]	Web Application Attacks	Web application security, with the goal of systematizing the available solutions into a broad picture that encourages further study	Web security, Protection
[5]	Web Attacks	Technologies, procedures, and tactics for intrusion detection look into new attack kinds, defenses, and contemporary academic research	Attack Detection
[6]	Web Attacks	With a particular emphasis on data accessibility, cyber security and cyber risk management	Attack Detection and Prevention
[7]	Injection Attacks	Contribute to the community by creating a strategy for preventing common injection attacks on web apps	Attack Protection

(*continued*)

Table 3. (*continued*)

References	Types of Attack	Summary of the Paper	Area of Focus
[8]	Zero-day Attacks	Machine learning-based hybrid approaches to successfully learn and identify intrusions have been presented	Attack Detection
[9]	Zero-day Attacks	Extends previous work by including individual request outlier explanations into an end-to-end pipeline	Attack Detection
[10]	Web Attack	This paper examines several common online assault monitoring and detection tools and methods that have been created and used in practice	Attack Detection
[11]	Web Application Attacks, Zero-day attacks	The research focuses on determining whether machine-learning-based WAFs are effective in thwarting the existing attack patterns that target web application frameworks	Attack Protection
[12]	SQL injection attacks	Efficiency of Web application firewall Mod security is evaluated in the proposed literature	Attack Prevention
[13]	Web Attacks	An analysis of the effectiveness of anomaly detection methods for preventing and identifying web attacks	Attack Detection and Prevention
[14]	Web Attacks	Signature based intrusion detection technique	Attack Detection

(*continued*)

Table 3. (*continued*)

References	Types of Attack	Summary of the Paper	Area of Focus
[15]	Web Attacks	To decrease acquired alerts and find new attacks, a methodology that combines both signature-based and anomaly-based techniques has been developed	Attack Detection
[16]	Phishing attacks	Different tools and methods for identifying phishing attacks	Attack Detection
[17]	Phishing attacks	This study examines how machine learning techniques can be used to search for unusual activity associated with phishing online attacks as a possible solution to the issue	Attack Detection
[18]	DoS Attacks	Developed a new method for filtering and detecting huge numbers of attack packets. This method includes novel data structures and algorithms. In order to be effective in real-time attack response, the suggested method focuses a strong emphasis on minimizing storage space and processing time	Attack Detection
[19]	Web Attacks	Examine the advantages of machine learning techniques for evaluating WAF	Attack Detection
[20]	Web Attacks	In this study, big data was used to detect web threats using machine learning techniques	Attack Detection and Analysis

(*continued*)

Table 3. (*continued*)

References	Types of Attack	Summary of the Paper	Area of Focus
[22]	Network Attacks	Focuses on how the online and offline performance of Snort IDS is impacted by multithreading, standard rule set setups, and real-time data shipping	Attack Detection System and Analysis
[23]	Web Attacks	In this research, a Web Gene Tree (WGT)-based MTD technique is proposed	Vulnerability detection
[24]	Web Attacks	With the help of hidden Markov algorithms, this article effectively detects web attacks using a BOW model to extract features	Web Attack Detection method
[25]	Web Forensic	To develop a new model, issues with the online forensics procedure's method, technique, application, and software that handles web activities have been looked into and analyzed	Vulnerability detection
[26]	Zero-Wall	This study suggests Zero-Wall, an unsupervised method for effectively identifying zero-day Web threats that integrates with an on-the-go WAF	Attack Detection

(*continued*)

Table 3. (*continued*)

References	Types of Attack	Summary of the Paper	Area of Focus
[27]	Web Attack	Three new findings are presented in this work that is related to the study of autonomous intrusion detection systems	Vulnerability detection
[28]	Web Attack	The goal of this work is to implement the web attack recognition model utilizing the Core Rule Sets of Mod Security in order to provide the capabilities of Snort web attack detection	Attack Detection
[29]	Brute Force Attack	The work investigates the deficiency of web attacks detection using ensemble learning and big data	Vulnerability detection
[30]	Web Application Attacks	The aim of this paper is to look into the methods and tools used to stop attacks. In order to solve the flaws of current technology and provide more useful solutions, data mining and machine learning approaches are also researched	Attack Detection and Prevention
[31]	Cyber Attack	In order to provide intelligent services in the field of cyber-security, specifically for intrusion detection, authors used a variety of well-known machine learning classification techniques, such as Bayesian Network (BN), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Random Tree (RT), Decision Table (DTb), and Artificial Neural Network (ANN)	Intrusion Detection

3.3 Analysis in Intrusion Detection Solutions

In order to provide data-driven judgments, security models based on machine learning frequently require a lot of static data. Systems for detecting anomalies rely on building such a model while taking into account both regular behavior and anomalies according to their patterns. A vast and dynamic security system's usual behavior, however, is not well understood and may alter over time, which may be seen as a gradual increase in the dataset. In numerous situations, the patterns in incremental datasets may shift. This frequently leads to a significant number of false positive alarms. In order to forecast unknown assaults, a recent regressive behavioral trend is more likely to be interesting and pertinent than one from the past. Therefore, effectively using the model in intrusion detection solutions could be another issue.

3.4 Proposed Technique to Solve Attack

There are various techniques used to solve the attacks as shown Fig. 1.

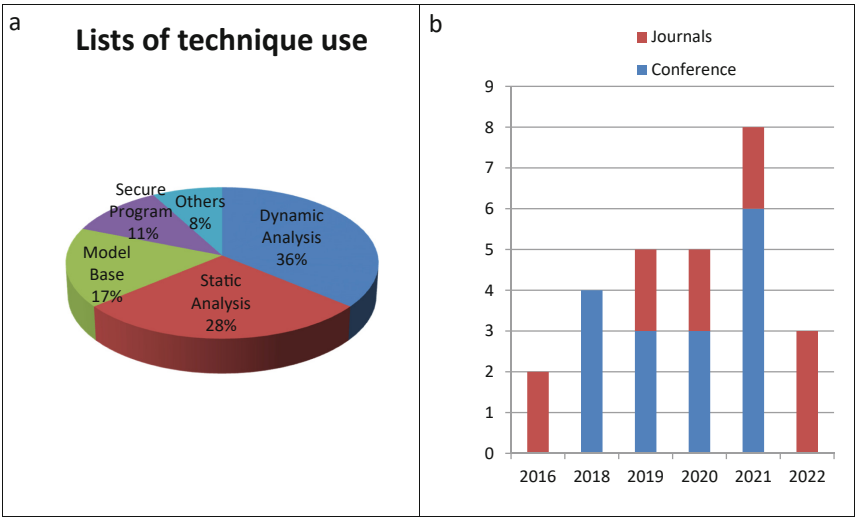


Fig. 1. **a** Lists of techniques used to solve attacks, **b** Publication year vs. publication count.

Dynamic Investigation: Dynamic analysis refers to recognize output with respect to a predefined input in runtime.

Static investigation: Static analysis deal with programming. In order to check for vulnerabilities, differences in the program are found.

Model Based: Different types of model use such as hybrid model to [33, 34] detect attacks, or model for feature selection so that unknown or new attacks are easily detected.

Secure Programming: Different coding like data mining, machine learning etc. are used to test the proposed model for detecting attacks.

Others: Different types of tools like SNORT, MODSECURITY, firewalls etc. used for detection, prevention of web service attacks.

One of the key ideas in cyber security is the classification or prediction of attacks. Important modules that are in charge of developing a prediction model to categorize threats or assaults and predict the future for a specific security risk. The development of a data-driven security model for a specific security challenge based on the idea of online assaults, as well as proper empirical evaluation to assess the model's efficacy and efficiency and evaluate its usability in the real-world application area, may be future work.

3.5 Experiment Benchmarks

The following machine learning algorithms are used as workbench for measuring accuracy in detecting intrusions:

- **Logistic Regression**: This statistical model is frequently used for categorization and predictive analytics based on a collection of independent variables, logistic regression calculates the likelihood of an event occurring, such as voting or not voting.
- **K-Nearest Neighbors**: The k-nearest neighbor algorithm, often known as KNN. It is a non-parametric and supervised learning classifier that employs proximity to classify or predict the grouping of a single data point. While it may be used for either regression or classification issues, it is most commonly utilized as a classification technique, based on the idea that comparable points can be discovered nearby.
- **Naïve Bayes**: The Bayes theorem calculates the likelihood of an event occurring given the chance of another event occurring. The following equation expresses Bayes' theorem mathematically:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- **Support Vector Machines**: SVM locates a hyperplane that defines a boundary between data kinds. This hyper-plane is nothing more than a line in two dimensions. Each data item in the dataset is plotted in an N-dimensional space in SVM, where N is the number of features/attributes in the data. Find the best hyperplane to split the data. To apply SVM to multi-class situations, we can develop a binary classifier for each data class.
- **Decision Trees**: The most powerful and widely used tool for categorization and prediction is the Decision Tree. A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label.
- **Random Forest**: The "forest" it creates is an ensemble of decision trees, which are often trained using the "bagging" approach. The bagging approach is based on the premise that combining learning models improves the final output. Random forest has the significant benefit of being applicable to both classification and regression problems, which comprise the majority of contemporary machine learning systems. It also resists over fitting, which is common in decision trees.

3.6 Experiment Results

NSL-KDD dataset [32] is used for testing intrusion detection systems. Dataset produces normal and anomalous request including various types of attacks. The features in this data set are used to characterize right and erroneous system operation. Machine learning methods employ these properties to create models that categorize the accuracy of the system's execution state. Table 4 and Fig. 2, shows the machine-learned model with different algorithms to forecast undetected trace which reflects a genuine system execution.

Table 4. Comparison of different machine learning models for intrusion detection

	Accuracy (%)	Precision (%)	Recall (%)
Logistic Regression	87.624	83.568	91.608
K Neighbors Classifier	98.936	99.056	98.670
Gaussian NB	91.605	92.532	89.296
Support Vector Machines	97.289	97.547	96.646
Decision Tree	99.869	99.847	99.872
Random Forest	99.876	99.932	99.805
PCA	99.821	99.898	99.721

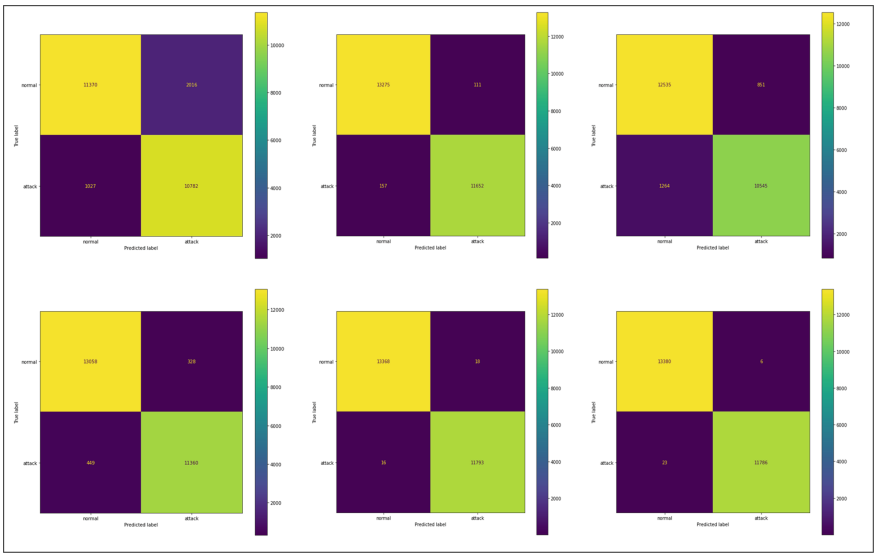


Fig. 2. Performance evaluation of different machine learning algorithms for intrusion detection system.

4 Discussion

Here are the web threat areas that were the focus of the research articles. Web service vulnerabilities and web service attacks are the main topics of this study. Figure 2 shows that 5 research (16.67%) and 13 (43.33%) respectively, focus on web service vulnerabilities and web service assaults detection. 12 articles (40%) concentrate on the development of various combination tactics or attacks to test the robustness of web services and assess the web service contingency mechanisms. The strategies are then divided into Attack Detection or Prevention and Vulnerability detection or Prevention to provide more detail. Additionally, crucial are attack detection and prevention; 5 studies (16.67%) concentrate on the vulnerabilities. A few papers concentrate on various algorithms, models, and tools for the detection and prevention of various web threats. Finally, as an extension, we utilize some machine learning algorithms on NSL-KDD dataset for intrusion detection and achieve the highest accuracy of 99.876%, precision of 99.932% and recall of 99.805% for random forest classifier.

5 Conclusion

The biggest challenges with using online services to send data are those with privacy and data protection. The security of online services must be maintained by taking into account three components of information security, including confidentiality, integrity, and availability. Attacks on the web are very aggressive and more likely to affect business. Application intrusion detection systems and web application firewalls are two detection methods that are effective in catching known threats with high accuracy. This is because the majority of commercial devices uses signature-based technology and predefined regulations, which are the main causes of their reliance on these technologies. However, the vast majority of strategies were created to progressively fend against fresh and undiscovered threats. In order to achieve the requisite effectiveness, the methodologies employed for anomaly-based assaults are currently being developed. The number of attacks will be greatly reduced with the incorporation of anomaly-and signature-based detection systems. In this paper, we focused on the analysis of numerous web attacks and various methods, instruments, and machine learning and deep learning algorithms for their detection and prevention. Although the real-time detection capabilities of those technologies are relatively constrained, they provide invaluable insights into attack detection through the study of successful assaults and the identification of previously undiscovered ones. To provide a future research agenda for the study of web threats, we have further highlighted and discussed a number of significant security analysis challenges.

References

1. Rasool, A., Jalil, Z.: A review of web browser forensic analysis tools and techniques. Res. J. Comput. **1**(1), 15–21 (2020)
2. Calzavara, S., Focardi, R., Squarcina, M., Tempesta, M.: Surviving the web: a journey into web session security. In: The Web Conference 2018—Companion of the World Wide Web Conference, WWW 2018. Association for Computing Machinery, Inc., pp. 451–455 (2018). https://doi.org/10.1145/3184558.3186232

3. Mouli, V.R., Jevitha, K.P.: Web services attacks and security—A systematic literature review. In: Procedia Computer Science. Vol. 93. Elsevier B.V., pp. 870–877 (2016). https://doi.org/10.1016/j.procs.2016.07.265
4. i, X., Xue, Y.: A survey on web application security. Tech. rep., Vanderbilt University (2011). http://www.truststc.org/pubs/814.html
5. Ozkan-Okay, M., Samet, R., Aslan, O., Gupta, D.: A comprehensive systematic literature review on intrusion detection systems. IEEE Access **9**, 157727–157760 (2021)
6. Cremer, F., Sheehan, B., Fortmann, M., et al.: Cyber risk and cybersecurity: a systematic review of data availability. Geneva Pap Risk Insur Issues Pract. Published online 2022
7. Ibarra-Fiallos, S., Higuera, J.B., Intriago-Pazmino, M., Higuera, J.R.B., Montalvo, J.A.S., Cubo, J.: Effective filter for common injection attacks in online web applications. IEEE Access **9**, 10378–10391 (2021)
8. Maseno, E.M., Wang, Z., Xing, H.: A systematic review on hybrid intrusion detection system. In: Maglaras, L. (ed.) Security Communication Networks, pp. 1–23 (2022)
9. Sejr, J.H., Zimek, A., Schneider-Kamp, P.: Explainable detection of zero day web attacks. In: Proceedings - 2020 3rd International Conference on Data Intelligence and Security, ICDIS 2020. Institute of Electrical and Electronics Engineers Inc., pp. 71–78 (2020)
10. Dau, H. X., Trang, N. T. T., Hung, N.T.: A survey of tools and techniques for web attack detection. J. Sci. Technol. Inf. Secur. **1**(15), 109–118 (2022). https://doi.org/10.54654/isj.v1i15.85211
11. Applebaum, S., Gaber, T., Ahmed, A.: Signature-based and machine-learning-based web application firewalls: a short survey. In: Procedia CIRP. Vol 189. Elsevier B.V., pp. 359–367 (2021)
12. Mukhtar, B.I., Azer, M.A.: Evaluating the modsecurity web application firewall against SQL injection attacks. In: Proceedings of ICCES 2020 - 2020 15th International Conference on Computer Engineering and Systems. Institute of Electrical and Electronics Engineers Inc. (2020)
13. Riera, T.S., Higuera, J.R.B., Higuera, J.B., Herraiz, J.J.M.: Montalvo JAS. Prevention and fighting against web attacks through anomaly detection technology. A systematic review. Sustainability **12**(12) (2020)
14. Díaz-Verdejo, J., Muñoz-Calle, J., Alonso, A.E., Alonso, R.E., Madinabeitia, G.: On the detection capabilities of signature-based intrusion detection systems in the context of web attacks. Appl. Sci. **12**(2) (2022)
15. Hussein, S.M.: Performance evaluation of intrusion detection system using anomaly and signature based algorithms to reduction false alarm rate and detect unknown attacks. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI) pp. 1064–1069 (2016)
16. Lyashenko, V., Kobylin, O., Minenko, M.: 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S & T). IEEE (2018)
17. Ortiz Garces, I., Cazares, M.F., Andrade, R.O.: Detection of phishing attacks with machine learning techniques in cognitive security architecture. In: Proceedings—6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019. Institute of Electrical and Electronics Engineers Inc., pp. 366–370 (2019)
18. Quỹ phát triển khoa học công nghệ quốc gia (Vietnam), Institute of Electrical and Electronics Engineers. RIVF 2019 Conference Proceedings : The 2019 IEEE-RIVF International Conference on Computing and Communication Technologies : Danang, Vietnam, March 20–22 (2019)
19. Betarte, G., Pardo, A., Martinez, R.: Web application attacks detection using machine learning techniques. In: Proceedings—17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018. Institute of Electrical and Electronics Engineers Inc., pp. 1065–1072 (2019)

20. Zuech, R.: Machine Learning Algorithms for the Detection and Analysis of Web Attacks (2021)
21. Sarker, I.H.K., Badsha, A.S.M., Alqahtani, S., Watters, H., Ng, P., Alex: Cyber security data science: an overview from machine learning perspective. J. Big Data (2020)
22. Thorarensen, C.: A Performance Analysis of Intrusion Detection with Snort and Security Information Management. Master's thesis, Link¨oping University, Database and Information Techniques (2021)
23. Zhang, Y., Ma, D., Sun, X., Chen, K., Liu, F.: WGT: Thwarting web attacks through web gene tree-based moving target defense. In: Proceedings—2020 IEEE 13th International Conference on Web Services, ICWS 2020. Institute of Electrical and Electronics Engineers Inc., pp. 364–371 (2020)
24. Ren, X., Hu, Y., Kuang, W., Souleymanou, M.B.: A web attack detection technology based on bag of words and hidden markov model. In: Proceedings—15th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2018. Institute of Electrical and Electronics Engineers Inc., pp. 526–531 (2018). https://doi.org/10.1109/MASS.2018.00081
25. Varol, A.: Institute of Electrical and Electronics Engineers. Portugal Section., Institute of Electrical and Electronics Engineers. In: 7th International Symposium on Digital Forensics and Security, 10–12 June 2019, Barcelos, Portugal
26. Tang, R., Yang, Z., Li, Z., Meng, W., Wang, H., Li, Q., Sun, Y., Pei, D., Wei, T., Xu, Y., Liu, Y.D.: Zerowall: Detecting zero-day web attacks through encoder-decoder recurrent neural networks. In: IEEE INFOCOM 2020—IEEE Conference on Computer Communications, pp. 2479–2488 (2020)
27. Pan, Y., et al.: Detecting web attacks with end-to-end deep learning. J. Internet Serv. Appl. **10**(1), 1–22 (2019). https://doi.org/10.1186/s13174-019-0115-x
28. Yang, C., Shen, C.H.: Implement web attack detection engine with snort by using modsecurity core rules (2009)
29. Zuech, R., Hancock, J., Khoshgoftaar, T.M.: Investigating rarity in web attacks with ensemble learners. J. Big Data **8**(1), 1–27 (2021). https://doi.org/10.1186/s40537-021-00462-6
30. Varol, A., Karabatak, M., Varol, C.: Fırat Üniversitesi, Institute of Electrical and Electronics Engineers. Turkey Section, Institute of Electrical and Electronics Engineers. In: 6th International Symposium on Digital Forensic and Security: Proceeding Book , 22–25 March 2018, Antalya, Turkey
31. Alqahtani, H., Sarker, I.H., Kalim, A., Minhaz Hossain, S.M., Ikhlaq, S., Hossain, S.: Cyber intrusion detection using machine learning classification techniques. In: Chaubey, N., Parikh, S., Amin, K. (eds.) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore (2020)
32. NSL-KDD dataset. https://www.kaggle.com/datasets/hassan06/nslkdd. Accessed 20 April 2022
33. Hossain, S.M.M., Sen, A., Deb, K.: Detecting spam SMS using self attention mechanism. In: Vasant, P., Weber, G.W., Marmolejo-Saucedo, J.A., Munapo, E., Thomas, J.J. (eds.) Intelligent Computing & Optimization. ICO 2022. Lecture Notes in Networks and Systems, vol. 569. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-19958-5_17
34. Hossain, S.M.M., et al.: Spam filtering of mobile SMS using CNN–LSTM based deep learning model. In: Hybrid Intelligent Systems. HIS 2021. Lecture Notes in Networks and Systems, vol. 420. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96305-7_10



Encrypting the Colored Image by Diagonalizing 3D Non-linear Chaotic Map

Rahul(✉), Tanya Singhal, Saloni Sharma, and Smarth Chand

Delhi, India
agg.tanya00@gmail.com, salonisharma0820@gmail.com

Abstract. Nowadays data transmission takes up several other forms as compared to simple text data; images are one of the prominent modes of data transmission. The transmission of digital data in the form of images results in image encryption and decryption is of prime interest for recent research work. Protecting the image during transmission from several attacks and unauthorized access is very important. Images are of the most eye-catching class of information in the security field where efficient encryption and decryption mechanisms need to be devised. To enhance the security of the system and to decrease the probability of attacks on image encryption a large key space-based approach is designed. This proposed technique uses a 3D nonlinear chaotic map by using the diagonalization of the cubic map and logistic map. The proposed system enhances overall security by utilizing the randomness of the trilinear interpolation chaotic system.

Keywords: 3D Nonlinear Chaotic Map · Cubic Map · Image Encryption · Key Space · Logistic Map

1 Introduction

With the contemporary growth of tools, methods, and processes in every field of information and processing technology security becomes an important factor to consider. The transmission and transformation of data into digital content requires an efficient mechanism that provides a quality focus along with security [1]. Most of the recent applications use the image as a mode of transmission such as bio-medical imaging, remote-sensing, multimedia systems, holographic storage, surveillance systems (military and police identification systems), etc. [2]. Securing images for transmission is significant as the transmitting images can simply be retrieved by any unauthorized or unlawful person/system [3]. Efficient cryptographic mechanisms need to be deployed which may use a varied key space. Factors like amplitude values, phase change, frequency domains, wavelength deviation, and degree of polarization may be altered in one or the other way to increase the vitality of key space. These mechanisms may be clustered into two different categories: information hiding and encryption. A few common techniques of data hiding are steganography and watermarking [4, 5]. Image encryption is a wider phenomenon that is classified as optical, spatial, transform domain, and compressive sensing image encryption. In the case of spatial encryption techniques such as

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1270–1283, 2023.
https://doi.org/10.1007/978-3-031-27409-1_116

meta-heuristics, chaotic maps, elliptic curves, and automata are frequently used. The wavelength, gyrator, fractional Fourier, and Fresnel transform are commonly used in the transform domain [6, 7]. Image encryption comes up as a promising solution to overcome this challenge. The security protocols applied to image encryptions are secured in comparison to text encryption [8, 9]. Image encryption comes with the limitation of big size, spatial redundancy, and high correlation. Some motivational features behind image encryption include fast speed and high reliability. Thus, a traditional encryption algorithm may not work very well on image encryption. To overcome these limitations and lossless transmission chaotic maps may be used which provides sensitivity. There exist several methods which utilize the inherent features of chaotic maps such as high sensitivity and simple structure to provide better results.

Image encryption uses a color image which is represented using a color matrix containing multiple rows and columns. The discrete values of this matrix hold a pixel value which is the outcome of the quantization process demonstrating the degree of the pixel position corresponding to its color [10, 11]. The matrix values are a triplet of three image components listed as RGB components to define each pixel area value. Image security is critical because of techniques that are required to convert original images into cipher images. The cipher image is difficult to understand and hence provides additional security. One of the most prominent theories which developed a basis for image encryption is the "Chaos theory" [12]. The ability to enhance security by applying mechanism which includes mixing capabilities, high sensitivity, heterogeneous control parameters, and randomness [13]. The first encryption mechanism using chaotic encryption is given by Matthews in 1989. This algorithm results in low complexity by employing "a multiple-definite function for a linear chaotic map" [14]. After this algorithm, several researchers described this problem in chaotic systems using their concepts and implementations. Each has varied results and improved performance. This proposed work describes a multi-chaotic trilinear interpolation system that may simultaneously encrypt various RGB components of color images. The chaos system consists of six maps $T_1(x)$, $T_2(x)$, $T_2(y)$, $T_3(x)$, $T_3(y)$, *and* $T_3(z)$ to increase the conjugation of the items $x_i^3, x_i^2, y_i^2, x_i y_i, y_i^3, z_i^2 x_i, \ldots z_i^3$ and expand the security of the system. The image's blocks are shuffled using the first map $T_1(x)$, while the second and third maps $T_2(x)$, $T_2(y)$ scrambled the position in rows and columns, respectively. Finally, $T_3(x)$, $T_3(y)$ and $T_3(z)$ will change the pixels values twice to make the system more complex, and hence more secure.

A detailed discussion of all the methods using chaotic maps is explained in the literature review which is presented in Sect. 2. The proposed chaos-based encryption system is presented in Sects. 3 and 4. The encryption and decryption algorithms are illustrated in Sect. 5. The results of implementing the proposed work are highlighted in Sect. 6. The analysis of key space along with a discussion of performance parameters used in the study is given in Sect. 7. Finally, the work is concluded and imminent work is exposed to researchers in Sect. 8.

2 Literature Review

Image-based encryption becomes an important technique for information security. Several approaches have been devised in literature each with its pros and cons. This section highlights some novel and efficient approaches used in past. Image encryption in medical diagnosis is carried out by Lima et al. [15] where a multi-parameter cosine transformation is used. The transformation is carried out using similarity vectors that establish the relation between the 3D image cosine transform and the Laplace function of the lattice graph. The given work is unique in the sense that it uses 3D transformation whereas the majority of work used in past focuses on 2D images and transformation. The results thus obtained prove to be a promising method with a given key size and greater encryption security.

The attractive features of parallel computing to speed up the image encryption process are applied by Wang et al. in [16]. To overcome the limitations of high time complexity and poor permutation index the authors suggested a fusion-based approach. A parallel diffusion algorithm is designed that guarantees low time-space complexity. The diffusion algorithm is divided into cycles where each cycle adopts permutation and combination over computational models to ensure improvement in efficiency using parallel image encryption. The authors of [17] proposed a color image encryption method that uses the traditional Arnold transformation along with three-channel interference. The three channels used in interference are the basic RGB components decomposed with three different phase shifts. Encryption is applied using random phase masks. For image decryption corresponding light wavelengths are used to ensure the feasibility, security, and effectiveness of the proposed work. An exhaustive literature survey has been carried out by the authors of [18]. A comprehensive study is presented for more than 50 states of art techniques. Apart from description classification is carried out which results in 10 classes of color image encryption. A Hopfield chaotic method (HPNN) [19] was designed by the authors which uses neural network-based color image encryption. In the proposed work a composite map consisting of logistic and tent map is used for key generation and distribution. The conventional Arnold function transformation is used for image scramble. The HPNN is primarily deployed in the diffusion phase to generate a self-diffusion chaotic matrix.

Zhang et al. [20] in their research work uses spatiotemporal chaos for image encryption. It uses a large dataset to provide validation for the presented work and the results show that this approach provides a larger key space to enhance security and better chaotic behaviors. A stream cipher-based image encryption [21] method is proposed which uses one-time pads for key generation. To generate a secure random key piecewise linear ordering is used. The two performance measures NPCR and UACI are used to validate the results. Equal and singular modulus decomposition is used in [22]. The proposed method utilizes the triple masking mechanism using several transformations to enhance security. The key selection is purely random that is implemented using MATLAB software. A block scrambling method for digital image encryption is given in [23]. The three basic color components (R, G, and B) are scrambled for encryption. The image is divided into blocks where each block is independently encrypted and diffusion is performed later. A magic cube color image encryption [24] is applied over medical images to ensure safety and prevention of attacks of medical data. A fused magic cube

technique is used for compounding where the replacement of each block is provided to ensure flexibility and randomness.

3 Proposed Chaos System

The cubic map and logistic map are used to generate the new 1D, 2D, and 3D equations that make up the trilinear interpolation chaotic system. The proposed chaos system strengthens the components' cubic and quadratic connection $y_i^2, x_i^2, x_i y_i, x_i^3,$ and z_i^3 and provided more security to the system. The form of the proposed system is simply as follows:

$$f(x, y, z) = \sum_{i=0}^{3} \sum_{j=0}^{3} \sum_{k=0}^{3} \gamma_{ijk} x^i y^j z^k \tag{1}$$

The geometric representation of trilinear interpolation is shown in Fig. 1. The sum of the products of the values at each corner and the portion of the volume diagonally opposite the corner equals the product of the value at the desired point and the entire volume.

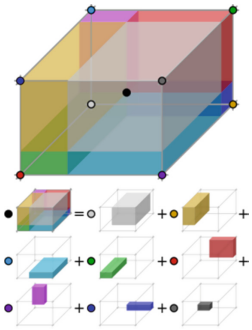


Fig. 1. Visualization of trilinear interpolation.

Within a constrained range, expression (2) will provide six chaotic maps (0,1). Beyond the range, the control parameters γ_{ijk} exhibit no chaotic behavior.

$$\left\{ \begin{array}{c} 1D \\ x_{n+1} = \mu x_n (1 - x_n)(2 + x_n) \\ 2D \\ x_{n+1} = \mu_1 x_n (x_n + 2)(1 - x_n) + \gamma_1 y_n^2 \\ y_{n+1} = \mu_2 y_n (y_n + 2)(1 - y_n) + \gamma_2 \left(x_n y_n + x_n^2\right) \\ 3D \\ x_{n+1} = \lambda x_n (2 + x_n)(1 - x_n) + \beta y_n^2 x_n + \alpha z_n^3 \\ y_{n+1} = \lambda y_n (2 + y_n)(1 - y_n) + \beta z_n^2 y_n + \alpha x_n^3 \\ z_{n+1} = \lambda z_n (2 + z_n)(1 - z_n) + \beta x_n^2 z_n + \alpha y_n^3 \end{array} \right\} \tag{2}$$

The system goes into chaos and generates a chaotic series in the range from 0 to 1 subject to: $1.41 < \mu < 1.59$, $2.71 < \mu_1 < 3.6$, $2.68 < \mu_2 < 3.51$, $0.13 < \gamma_1 < 0.25$, $0.11 < \gamma_2 < 0.17$, $3.49 < \lambda < 3.83$, $0 < \beta < 0.026$ and $0 < \alpha < 0.017$. A bifurcation diagram is presented in Fig. 2. Bifurcation parameters are displayed on the horizontal axis and the vertical axis represents the values (x_n, y_n, z_n) visited asymptotically.

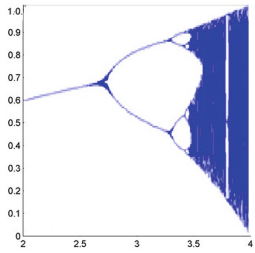


Fig. 2. Bifurcation diagram of sequences in expression (2)

4 Diagonalizing of the Proposed Chaotic Map

The system (1) can be represented as the vector

$$T = (T_1(x), T_2(x), T_3(x), T_2(y), T_3(y), T_3(z)) \tag{3}$$

$$T\left(\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}\right) = \begin{pmatrix} T_1 & 0 & 0 \\ T_2 & T_2 & 0 \\ T_3 & T_3 & T_3 \end{pmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{4}$$

$$T_i \begin{bmatrix} x_{n+1} \\ y_{n+1} \\ Z_{n+1} \end{bmatrix} = \left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right) \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} mod\ n \tag{5}$$

$$T_i \begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = A \times \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} mod\ n \tag{6}$$

where n is a prime number. The matrix A is invertible if the value of the determinant |A| is unequal to zero and the condition of $gcd\ (|A|, n) = 1$ is fulfilled. System (1) has the inverse as $x_{n+1} = (A^{-1} \times x_n) mod\ n$. Finding a new x', y', *and* z' system without cross terms is known as diagonalizing Eq. (3). Every quadratic form, according to the Principal Axes Theorem, can be diagonalized. Thus, the quadratic form of 2D can be diagonalized as $\left(x'_{n+1}\right)^T D x'_{n+1}$ and $\left(y'_{n+1}\right)^T Q y'_{n+1}$, where D *and* Q are each 2×2 matrices with 4 parameters. The number of parameters rises to 9 when the 3 variables x, y, *and* z are diagonalized in their 3D form.

5 Encryption and Decryption Algorithms

5.1 The Image Encryption Algorithm (IEA)

Red, blue, and green are the 3 parts of a color image I that are subdivided into $M \times N$ matrices R, B, and G, each of which has M rows and N columns comprising the pixel values. Figure 3, illustrates the proposed technique which employs a strategy known as the Pixel Transform Table (PTT) by creating a chaotic sequence for block shuffles, pixel substitutions, and permutations. The PPT strategy is illustrated as follows:

Pixel Transform Table (ρ, jD − map)

Input: Trilinear interpolation chaotic
1. Set the chaotic parameters to zero
2. For $j = 1$ *to n*
3. For $i = 1 \rightarrow \rho$
4. Generate the sequence $S_j(i)$ using $jD - map$
5. $S_j = \{(1), S(2),..., S_j(\rho)\}$
6. $S^j = \text{Sort}(S_j, \rho)$
7. Find the values' positions of [] in S[], then create the transfer $T_j = \{t_1, t_2,.., \rho\}$

Output: Random map T_j

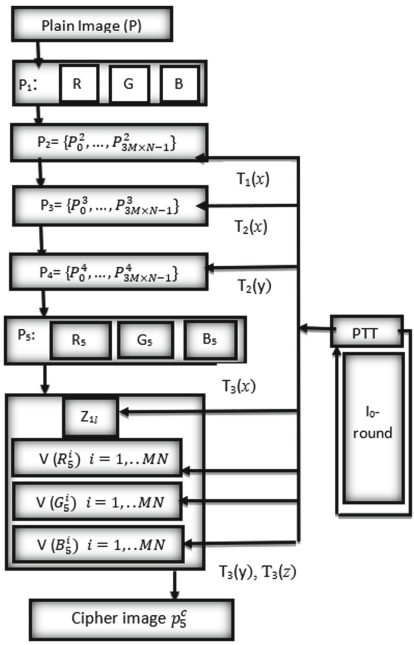


Fig. 3. A diagram of the proposed trilinear chaotic encryption algorithm

The algorithm involves three phases, as follows:

Phase-1 Each matrix (R, B, G) in a color image P is separated as (α, β) blocks with size $\frac{M}{\alpha} \times \frac{N}{\beta}$, as shown in Figure 4. These pixel blocks' positions are randomly shifted and three matrices $R_1, B_1,$ and G_1 are combined to form a new matrix $P_1 = [R_1, B_1, G_1]$. P_1 is rewritten in a one-dimensional array $P_2 = \{p_0, p_1, \ldots, p_{3MN-1}\}$. With initial parameters, μ and x_0, the chaotic map (1D) shuffles the P_2 entries.

1. Find the initial parameters μ and x_0 in 1D.
2. Let $T_1(x) \longleftarrow \text{PTT}\left(\frac{M}{\alpha} \times \frac{N}{\beta},\ \mathbf{1D}\right)$.
3. Based on $T_1(x)$, arrange the one-dimensional vector $P_2 = \{p_0, p_1, \ldots, p_{3MN-1}\}$.
4. Reconstruct the three M× N matrices $R_2,\ B_2$, and G_2 from the vector P_2.
5. Reconstruct the blocks in each matrix R_2, G_2 and B_2 to create $R_3, G_3,$ and B_3.

Phase-2: In this phase, the logistic map was used to shuffle the pixel coordinates in rows and columns.

The 2D map enters a chaotic state for given initial values of $\mu_1, \mu_2, \gamma_1, \gamma_2, x_0,$ and y_0.

1. For some initial values $\mu_1, \mu_2, \gamma_1, \gamma_2, x_0, y_0$
2. Let $T_2(x) \longleftarrow \text{PTT}(M \times N,\ 2D), T_2(y) \longleftarrow \text{PTT}(M \times N,\ 2D)$.
3. Combine the matrices $R_3,\ G_3$ and B_3 vertically to obtain the matrix 3M × N denoted P_3.
4. Reordering the rows in matrix P_3 based on $T_2(x)$, its rows will be listed in the new order, $t_1, t_2 \ldots, t_{3M-1}, t_{3M}$ We earn a new 3M × N denoted P_4.
5. Decompose P_4 into three matrices $R_4, G_4,$ and B_4.
6. Combine, $R_4, G_4,$ and B_4 horizontally to obtain a new matrix.
7. Reordering the columns of P_4 using $T_2(y)$, its columns will be listed in the new order, $t_1, t_1 \ldots, t_{N-1}, t_N$. We earn a new 3M × N denoted P_5.

Phase-3: Pixel values having been changed by pixel permutation using the 3D logistic map, the formula 3D enters a chaotic state for given initial values $\lambda, \beta, \alpha, x_0, y_0, z_0$. Let D_{now} be the newly decrypted value following the initial diffusion, P_{now} is the current plain value, D_{pre} is the preceding cyphered value, and P_{pre} is the preceding plain value.

The initial values $D(V(R_5))_0, P(V(R_5))_0, D(V(B_5))_0, P(V(B_5))_0, D(V(G_5))_0$ and $P(V(G_5))_0$ are equal to zero.

1. Disjoint the induced matrix P_5 into $R_5, G_5,$ and B_5 M × N matrices from left to right. Then, construct three vectors $V(R_5), V(G_5),$ and $V(B_5)$, each with length MN.
2. For the initial values $\lambda, \beta, \alpha, x_0, y_0, z_0$ do
3. $T_3(x) \longleftarrow \text{PTT}(3MN,\ 3D)$,
4. $T_3(y) \longleftarrow \text{PTT}(3MN,\ 3D)$
5. $T_3(z) \longleftarrow \text{PTT}(3MN,\ 3D)$
6. $For\ I = 1 \longrightarrow 3MN$

$$z_{1l} = (T_3(x)\ x\ 10^{14,3})mod\ 3$$
$$z_{2l} = (T_3(y)\ x\ 10^{14,256})mod\ 256$$
$$z_{3l} = (T_3(z)\ x\ 10^{14,256})mod\ 256$$

7. if z_{1l}=0 then

8. Repeat

$$D_jV(R_5) = (P_j(\,V(R_5) \times z_{2l} + D_{j-1}V(R_5) \times P_{j-1}V(R_5) \times z_{3l}) mod 256$$

Until all the values in V(R_5) changed.

9. Else if $z_{1l} = 1$
10. Repeat

$$D_jV(B_5) = (P_j(\,V(B_5) \times z_{2l} + D_{j-1}V(B_5) \times P_{j-1}V(B_5) \times z_{3l}) mod 256$$

Until all the values in V(B_5) changed.

11. if $z_{1l} = 2$

Repeat

$$D_jV(G_5) = (P_j(\,V(G_5) \times z_{2l} + D_{j-1}V(G_5) \times P_{j-1}V(G_5) \times z_{3l}) mod 256$$

Until all the values in V(G_5)have been changed.

12. Construct three ciphered $vectors\ V(R_5^c), V(G_5^c), V(B_5^c)\ (i = 1,2,\ldots,MN)$. They are converted into three matricesR_5^c, G_5^c, and B_5^c with the components of ciphered image P_5^c of size $M \times N$.

IEA: Plain Image (P) $\longrightarrow$ *Cipher Image* (P_5^c)

IEA: $T(P)$ $\longrightarrow$*Cipher Image* (P_5^c)



(a) *(b)* *(c)*

Fig. 4. **a** The color image P is divided into (5, 4) blocks, **b** shuffling the color image P (5, 4) blocks, **c** shuffling the color image P ($\alpha = 25$, $\beta = 20$) blocks.

5.2 The Image Decryption Algorithm (IDA)

IDA works in a reverse manner from the image encryption algorithm (IEA), is comparable to the former. The proposed chaotic system is a bijection, and its matrix has its inverse for determinant 1.

$$IDA : Cipher\ Image\left(P_5^c\right) \rightarrow Plain\ Image\ P$$
$$IDA : T^{-1}\left(P_5^C\right) \rightarrow P$$

6 Results and Analysis

To encrypt the 256 × 256 Baboon image and its RGB components, the following initial parameters were chosen as follows. Figure 5a–d illustrates the "Baboon" color images along with the RGB parts of each color image before encryption.

1*D* ($\mu = 1.57$, $x_0 = 0.12634456278912345$)

$2D$ ($\mu_1 = 3.33$, $\mu_2 = 3.34$, $\gamma_1 = 0.17$, $\gamma_2 = 0.14$, $x_0 = 0.23451686789876543$, $y_0 = 0.12343285678987654$)

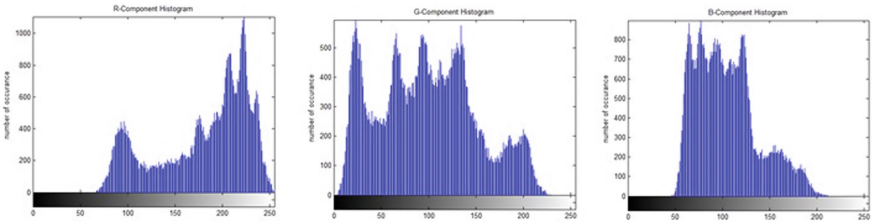
$3D$ ($\alpha = 0.01$, $\beta = 0.020$, $\lambda = 3.66$, $x_0 = 0.34569607898765432$, $y_0 = 0.4560147898765321$, $z_0 = 0.56789876543210634$)



(a) *(b)* *(c)* *(d)*

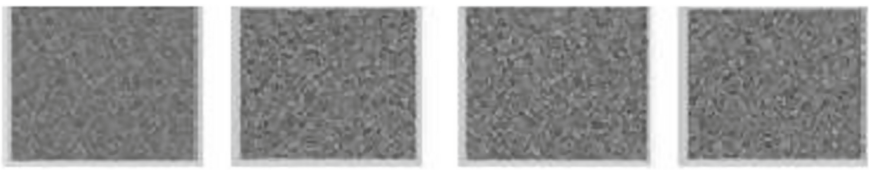
Fig. 5. a–d Displays "Baboon" color images along with the RGB parts of each color image before encryption.

Figure 6a–c show the "Baboon" histograms and their RGB components before encryption; the histograms indicate how the number of pixels in the "Baboon's" plain images correlates to the number of pixels at each color density level.



(a) *(b)* *(c)*

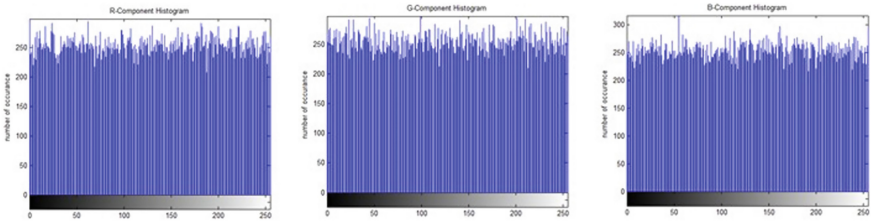
Fig. 6. Before encryption, the RGB histograms of the "Baboon" image.



(a) *(b)* *(c)* *(d)*

Fig. 7. Displays "Baboon" color images along with the RGB parts of each color image after encryption.

Figure 7a–d illustrates the "Baboon" color images along with the RGB parts of each color image after encryption. Figure 8 below display the histograms of "Baboon's" RGB components following encryption. The histograms show how the number of pixels at each color density level is rather evenly associated with the number of pixels in the ciphered RGB.



(a) *(b)* *(c)*

Fig. 8. After-encryption RGB component histograms for the "Baboon" image.

7 Security Analysis

7.1 The Key Size

To prevent brute-force attempts, the encryption must use a sufficient number of unique keys altogether. The proposed algorithm uses six values as x^{1D}, x^{2D}, y^{2D}, x^{3D}, y^{3D}, z^{3D}, and eight parameters μ, μ_1, μ_2, γ_1, γ_2, λ, β, α, as secret keys.

The work proposed in [25] demonstrated that for the precision 10^{-17}, the keys $K_{x_0^{1D}} = K_{x_0^{2D}} = K_{y_0^{2D}} = K_{x_0^{3D}} = K_{y_0^{3D}} = K_{z_0^{3D}} = 10^{17}$ and $K_\mu = K_{\mu_1} = K_{\mu_2}, K_{\gamma_1} = K_{\gamma_2} = K_\lambda = K_\beta = K_\alpha = 0.5 \times 10^{17}$. Assume that the plain color images have a size of 128×128. The number of iterations over six maps I_0 is $6(3 \times M \times N) = 6(3 \times 128 \times 128) \approx 2^{18} \approx 10^6$. The total key space will reach $\approx 1.953 \times 10^6 \times 10^{235} = 1.953 \times 10^{121}$. The key space of the proposed work is larger than 2^{138}, 2^{58}, 10^{140}, 2^{256}, 10^{79}, 4.2×10^{122} [3, 9, 11, 19, 21, 25] and it is greater than $2^{448} = 7.8 \times 10^{134}$ as discussed in [26]. The diagonalization form (2) has total 24 parameters and initial values as keys. The key in the suggested work raises it to 10^{415}. The proposed techniques describe a key space that is robust enough against brute-force attacks.

7.2 The Sensitivity Analysis of the Secret Keys

Different cipher pictures result from minor key variations. Using the incorrect key to decrypt the image results in the creation of another image. With the correct key of $\lambda = 3.66$, the encrypted image of the baboon is shown in Fig. 9. Figure 10 on the other hand shows how to decrypt an image of a baboon using the erroneous encryption key, which is $\lambda = 3.660000000000000001$. Making the algorithm responsive to the key was successful. A slight change to the key will produce a completely different decryption result, making it impossible for the attacker to access the correct original image.

7.3 Analyzing the Correlation of Adjacent Pixels

The level of pixel association is used to evaluate the correlation between pixels. In general, the performance of the encryption algorithm will be worse and vice versa depending on how strongly close pixels in the ciphered image are correlated. The following formula



Fig. 9. Result of using the right parameters to decrypt the R, G, and B components of the Baboon image.

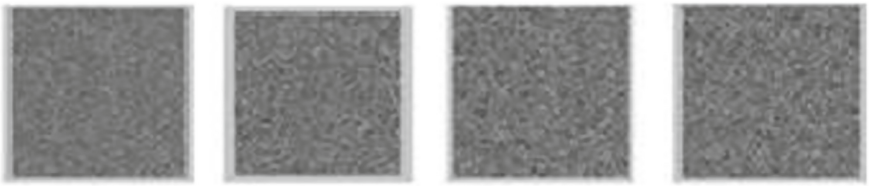


Fig. 10. Result of using the wrong parameters to decrypt the R, G, and B components of the Baboon image.

is used to determine the correlation between 3000 randomly chosen neighboring pixels in the vertical, horizontal, and diagonal directions.

$$E(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i) \tag{7}$$

$$E(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i) \tag{8}$$

$$D(x) = \sum_{i=1}^{N}(x_i - E(x))^2 \tag{9}$$

$$D(y) = \sum_{i=1}^{N}(y_i - E(y))^2 \tag{10}$$

$$cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(x))(y_i - E(y)) \tag{11}$$

$$r_{xy} = \frac{cov(x, y)}{\sqrt{D(x)}\sqrt{D(y)}} \tag{12}$$

Table 1 shows the adjacent pixels values, x and y, in three directions: horizontal direction (H-D), vertical direction (V-D), and diagonal direction (D-D). As a result, the ciphered image's adjacent pixels are random and the encryption's effects are resistant to statistical assault. The adjacent pixel in case of the original image, exhibits a high level of concentration and are close to "1" on the other hand in the encrypted image it shows the value towards "0".

Table 1 Neighbouring pixel correlation values obtained for the original and encrypted versions of the Baboon's image with components (H-D, V-D, and D-D)

Components	Color	Plain image	Cipher image
H-D	*R*	0.9413	−0.0019
	G	0.8796	−0.0056
	B	0.9164	0.0018
V-D	*R*	0.9527	−0.0021
	G	0.9283	−0.0047
	B	0.9563	0.00201
D-D	*R*	0.6471	−0.0021
	G	0.9567	−0.0036
	B	0.9355	0.0015

7.4 Peak Signal-to-Noise Ratio (PSNR) Analysis

The primary purpose of PSNR in picture reconstruction is as a quality indicator which can be measured in the equation shown below:

$$PSNR_{R,G,B} = 20 \times \log_{10}\left(\frac{255}{sqrt\left(MSE_{R,G,B}\right)}\right) \tag{13}$$

$$MSE_{R,G,B} = \sum_{i}\sum_{j}\frac{C_{R,G,B}(i,j) - C'_{R,G,B}(i,j)}{MxN} \tag{14}$$

The difference between the plain image and the ciphered image is expressed as the mean square deviation (MSD), which ranges from 0 to 255. Table 2 also displays the differences in PSNR values between the plain and the encrypted image. The proposed work exhibits greater resistance to statistical attacks.

Table 2 Results of PSNR for Baboon's image

Baboon's image	Proposed algorithm		
	R	G	B
PSNR	8.3985	9.4578	8.9701

8 Conclusion

In this work, we proposed a method for image security with a key space of six values as $x_0^{1D}, x_0^{2D}, y_0^{2D}, x_0^{3D}, y_0^{3D}, z_0^{3D}$, and eight parameters μ, μ_1, μ_2, $\gamma_{1,2}$, λ, β, α, as secret keys to counterattack various brute-force attacks. The proposed cryptosystem for image encryption relies on a trilinear chaotic system, which combines many multidimensional chaos systems. The trilinear system produces six entirely random bijections, $T_1(x)$, $T_2(x)$, $T_2(y)$, $T_3(x)$, $T_3(y)$, and $T_3(z)$. As a small variation to the encryption key will result in an entirely different decryption result which makes it impossible for an attacker to gain access to the right original image. In the ciphered image, the adjacent pixels are randomly chosen, which make it resistant to statistical attacks. The two neighbouring adjacent pixels of the original image is high level tend toward "1", while in the encrypted image, it is close to "1". The computational comparison of the proposed algorithm with existing cryptosystems revealed that the proposed method has a very high key space and a high level of security.

In the future the proposed technique will be tested with multiple datasets, and we will also perform comparison of the proposed work with state of art techniques.

References

1. Singh, A., Jain, K.: An automated lightweight key establishment method for secure communication in WSN. Wirel. Pers. Commun. 1–21 (2022)
2. Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., ... Nee, A.Y.C.: Enabling technologies and tools for digital twin. J. Manuf. Syst. **58**, 3–21 (2021)
3. Basu, A., Talukdar, S.: On the implementation of a digital image watermarking framework using saliency and phase congruency. In: Computer Vision: Concepts, Methodologies, Tools, and Applications, pp. 1391–1430. IGI Global (2018)
4. Smith, M., Miller, S.: Facial recognition and privacy rights. In: Biometric Identification, Law and Ethics, pp. 21–38. Springer, Cham (2021)
5. Singh, A., Jain, K.: An efficient secure key establishment method in cluster-based sensor network. Telecommun. Syst. **79**(1), 3–16 (2021). https://doi.org/10.1007/s11235-021-00844-4
6. Pan, J.S., Sun, X.X., Yang, H., Snášel, V., Chu, S.C.: Information hiding based on two-level mechanism and look-up table approach. Symmetry **14**(2), 315 (2022)
7. Sajedi, H.: Applications of data hiding techniques in medical and healthcare systems: a survey. Netw. Model. Anal. Health Inform. Bioinform. **7**(1), 1–28 (2018).
8. Matin, A., Wang, X.: Video encryption/compression using compressive coded rotating mirror camera. Sci. Rep. **11**(1), 1–11 (2021)
9. Wang, X., Liu, C., Jiang, D.: A novel visually meaningful image encryption algorithm based on parallel compressive sensing and adaptive embedding. Expert Syst. Appl. **209**, 118426 (2022)
10. Kakkar, A.: A survey on secure communication techniques for 5G wireless heterogeneous networks. Inf. Fusion **62**, 89–109 (2020)
11. Rathore, M.S., Poongodi, M., Saurabh, P., Lilhore, U.K., Bourouis, S., Alhakami, W., ... Hamdi, M.: A novel trust-based security and privacy model for Internet of Vehicles using encryption and steganography. Comput. Electr. Eng. **102**, 108205 (2022)

12. Tan, Y., Qin, J., Tan, L., Tang, H., Xiang, X.: A survey on the new development of medical image security algorithms. In: International Conference on Cloud Computing and Security, June, pp. 458–467. Springer, Cham (2018)
13. Roy, M., Chakraborty, S., Mali, K.: A chaotic framework and its application in image encryption. Multimed. Tools Appl. **80**(16), 24069–24110 (2021). https://doi.org/10.1007/s11042-021-10839-7
14. Khan, J.S., Ahmad, J.: Chaos based efficient selective image encryption. Multidimension. Syst. Signal Process. **30**(2), 943–961 (2018). https://doi.org/10.1007/s11045-018-0589-x
15. Lima, V.S., Madeiro, F., Lima, J.B.: Encryption of 3D medical images based on a novel multiparameter cosine number transform. Comput. Biol. Med. **121**, 103772 (2020)
16. Wang, X., Feng, L., Zhao, H.: Fast image encryption algorithm based on parallel computing system. Inf. Sci. **486**, 340–358 (2019)
17. Chen, W., Quan, C., Tay, C.J.: Optical color image encryption based on Arnold transform and interference method. Opt. Commun. **282**(18), 3680–3685 (2009)
18. Ghadirli, H.M., Nodehi, A., Enayatifar, R.: An overview of encryption algorithms in color images. Signal Process. **164**, 163–185 (2019)
19. Wang, X.Y., Li, Z.M.: A color image encryption algorithm based on Hopfield chaotic neural network. Opt. Lasers Eng. **115**, 107–118 (2019)
20. Zhang, Y.Q., He, Y., Li, P., Wang, X.Y.: A new color image encryption scheme based on 2DNLCML system and genetic operations. Opt. Lasers Eng. **128**, 106040 (2020)
21. Liu, H., Wang, X.: Color image encryption based on one-time keys and robust chaotic maps. Comput. Math. Appl. **59**(10), 3320–3327 (2010)
22. Khurana, M., Singh, H.: Asymmetric optical image triple masking encryption based on gyrator and Fresnel transforms to remove silhouette problem. 3D Res. **9**(3), 1–17 (2018)
23. Hosny, K.M., Kamal, S.T., Darwish, M.M.: A color image encryption technique using block scrambling and chaos. Multimed. Tools Appl. **81**(1), 505–525 (2021). https://doi.org/10.1007/s11042-021-11384-z
24. Rani, N., Sharma, S.R., Mishra, V.: Grayscale and colored image encryption model using a novel fused magic cube. Nonlinear Dyn. **108**(2), 1773–1796 (2022). https://doi.org/10.1007/s11071-022-07276-y
25. Teng, L., Wang, X.: A bit-level image encryption algorithm based on spatiotemporal chaotic system and self-adaptive. Opt. Commun. **285**(20), 4048–4054 (2012)
26. Kumari, M., Gupta, S., Sardana, P.: A survey of image encryption algorithms. 3D Res. **8**(4), 1–35 (2017)



Study of Third-Party Analytics Services on University Websites

Timi Heino, Sampsa Rauti(✉), Robin Carlsson, and Ville Leppänen

University of Turku, Turku, Finland
{tdhein,sjprau,crcarl,ville.leppanen}@utu.fi

Abstract. The accelerated digitalization and the increased use of online services for everyday tasks, online privacy issues are more important than ever before. This also goes for universities which are increasingly moving information and services online. Our study provides a technical overview of prevalence of third-party analytics on university websites. Websites of 40 universities from eight different countries around the world are analyzed to reveal third-party analytics services they use. The study shows that most universities, especially in many technologically advanced western countries, have alarmingly high number of analytics services on their websites. The results emphasize the need for web developers and data protection officers to better assess what kind of data their websites deliver to third parties. This is especially important for universities, as kind of example institutions tasked with advancing the common good.

Keywords: Online privacy · Third-party analytics services · Tracking · University Websites

1 Introduction

The digitalization of society continues and online services are increasingly being used to take care of ordinary everyday tasks. Universities, too, have benefited from this development and moved many services and lots of information online for potential applicants, students, researchers and other interested parties alike. In order to serve their visitors better, website maintainers use third-party analytics services. Universities are no exception. While the analytics services reveal who the visitors are and how they behave when browsing the website, they also cause privacy concerns [6,9,13].

Although universities are not always in state ownership or publicly funded, they have the role of generating and sharing information. They are institutions issuing degrees, conducting research and having societal impact. While universities usually acknowledge the importance of social responsibility in their curricula and guidelines for researchers, they should also act socially responsibly as organizations. This is why, one could argue, it is not part of their role to deliver

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1284–1292, 2023.
https://doi.org/10.1007/978-3-031-27409-1_117

personal data about website visitors to third-party analytics services. Instead, they should aim to be exemplary in online privacy. Because of the accelerated digitalization caused by the COVID-19 pandemic, such privacy issues are more important than ever before [11].

While the use of analytics services in higher education institutional websites has not been studied that much (see e.g. [7]), many university libraries seem to have taken the privacy of their websites more seriously, advising caution in the use of third-party analytics services [2,10]. These critical voices seem to be overpowered by a multitude of articles recommending Google Analytics [3,5,14], often without sufficient consideration of privacy consequences when using a third-party analytics service.

Striving to fill the obvious gap in this research area, we analyze websites of 40 universities to find out what kind of third-party analytics services they use. The five top universities from eight countries around the world were included in the sample. The current study provides a technical analysis of prevalence of third-party analytics on university websites and also allows us to compare university websites around the world in terms of visitors' privacy.

The rest of the paper is structured as follows. Section 2 outlines the setting of the study, explaining how the websites were chosen and analyzed. Section 3 presents the results of the study, comparing the use of third-party analytics services in different countries. Section 4 discusses the impact of the findings in terms of user privacy and web development. Finally, Sect. 5 concludes the paper.

2 Setting of the Study

In order to choose the universities for the current study, we first picked 8 countries around the world: Australia, Canada, Chile, China, Finland, Germany, India, and South Africa. For each selected country, we then chose the top five universities[1] The websites of the selected universities were then analyzed to find third-party analytics services.

When browsing a website, we went through each link in the main menu. We also used the search option if one was available and opened one search result link. All cookies were always accepted at the beginning of the visit if a cookie banner popped up. This approach is of course not an exhaustive analysis, but it does give a clear picture of how many analytics services are invoked even during a brief visit to a given university website.

Chrome developer tools (DevTools), which can be used to analyze web requests made by a website, were used when browsing university websites, and all third-party requests (requests that go outside the university's own server/domain name) recorded during the visit were manually analyzed. Domain names of requests, payloads, and request chains were inspected to distinguish requests related to tracking and analytics from other requests. Such benign requests were,

[1] The top five universities for each country were chosen according to the listing at https://www.topuniversities.com/university-rankings/world-university-rankings/2022.

among other things, HTML and CSS element downloads and requests associated with the normal functionality of the website. A sample view of Chrome DevTools is shown in Fig. 1.

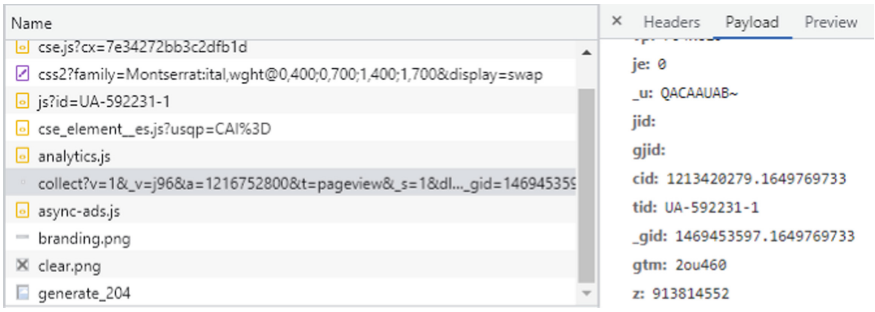


Fig. 1. A sample view of Chrome DevTools with a list of web requests and a payload of a request delivered to Google Analytics.

It is worth noting that our definition of "analytics service" is quite broad here. Requests that can be used for tracking and profiling, such as third-party requests caused by embedded videos, advertisement and social media share buttons were included in our data, although their primary purpose may not always be collecting analytics.

3 Results

Table 1 shows the average number of analytics services per university website and the number of unique analytics services per country. Of the studied countries, universities in Australia had the most analytics on their websites, with a staggering average of 12.6 services per analyzed website. On the website of one Australian university, our brief test browsing revealed 16 different services potentially collecting personal data. Even the website with least analytics included 8 different analytics services. While the universities vaguely state in their privacy policy documents that analytics are used for "business and learning analytics purposes" and "improving our services", this is hardly a valid justification for using 16 different third-party services. Australian university websites also had the widest array of analytics services in the study, 27 in total. Google Analytics, Facebook and Twitter were present on all 5 studied Australian websites. There were also a couple of analytics services that seemed to be exclusive to Australia, such as Tealium.

Finland takes the second place with an average of 7.0 analytics services per website. The university with the largest number of analytics services on their website included 10 unique services, still a large number difficult to justify. Finnish universities used 13 unique analytics services in total. Not surprisingly,

Table 1. The average number of used analytics services per university website and the number of different services per country.

Country	Services per website (average)	Number of different analytics services
Australia	12.6	27
Finland	7.0	13
Canada	5.6	14
South Africa	5.4	12
Chile	3.0	7
India	2.2	4
China	1.4	4
Germany	1.4	4

Google Analytics was present on all websites (5), followed by Facebook (4). Canada comes third, with 5.6 analytics services per university website on average. The highest number of analytics services on a single website was 8. The Canadian websites included 14 unique analytics services, overtaking Finland in this respect. The scenery of different services resembles Finland, with Google Analytics rampant everywhere (5).

Thus far, it seems technologically advanced western democracies all have quite a high number of analytics services on their university websites. Although South Africa may not exactly fit in this category, it still had 12 unique services in total on the university websites. Close to Canada, South Africa had 5.4 third-party analytics services per university website on average. The assortment of analytics services also remains similar to the western countries with services such as Google Analytics (5), Facebook (3), Hotjar (2) and LinkedIn Insight (2).

Chile had 7 different analytics services on the websites, with an average of 3.0 services per website. In Chile, Google's services constituted a major part of the different analytics services on the analyzed websites, but there were also two other third party services, Hotjar and Livechatinc. Compared to Chile, India had less unique analytics services (4) planted into university websites, averaging 2.2 services per website. Furthermore, all the used services (Google Analytics, AdSense, YouTube, DoubleClick) belonged exclusively to Google. One of the Indian universities, Indian Institute of Technology Delhi, did not appear to have any analytics services on its website.

Finally, Germany and China are tied, having 4 unique analytics services on university websites, and both averaging only 1.4 services per website. It is interesting to note how well Germany does compared to other western countries. Germany is also unique in the sense that the websites do not include services primarily meant for collecting analytics (such as Google Analytics). Instead, the found services were Google AdSense, Google Ad Services and Adobe Dynamic Tag Management Assets. Germany is also exemplary in the sense that it has

replaced third-party analytics services with Matomo, an open source solution that collects analytics locally [4]. Matomo was found on the websites of 4 German universities.

Considering China's usual eagerness to harvest data, it may at first be surprising there are not many analytics services present on the analyzed websites. The reason for this is probably China's censorship and the fact that the Great Firewall blocks many western services (for instance, Google Analytics) [1]. Instead, China uses its own centralized information collection. Tracking functionality of Baidu, a large Chinese technology company, was found on 4 of the 5 studied websites. Interestingly, Baidu also seems to be one of the few private companies licensed to gather and deliver data to China's governmental analytics [8]. Only one university, Shanghai Jia Tong University, appeared to have Google's services, such as Google Analytics and DoubleClick, on its website. On the whole, China does extremely well in not leaking information outside its own borders.

Table 2 shows the numbers of university websites using different analytics services. For example, 28 of 40 studied university websites were found to use Google Analytics. It is easy to see that Google enjoys a position of overwhelming dominance when it comes to university website analytics. Out of the 9 most used services, 6 are provided by Google. With the exception of one German and three Chinese universities, every time a website contained any third-party analytics, Google's services were present. This gives Google a broad front-row view to general interests and behavior of academic website visitors around the globe, but also the possibility to track individuals and collect their personal data. Other major analytics providers are Facebook, LinkedIn Insight, Hotjar and Twitter.

Figure 2 gives some insight into what and how many different analytics services are used in the studied countries. The figure shows, for example, that Australia has a wide collection of different analytics services, many of which are not used in other countries. It is also easy to see how Indian websites only make use of Google's services, and how few services Germany and China employ.

4 Discussion

The results of the current study do not flatter the technologically advanced western countries like Australia and Finland, where university websites are replete with analytics. Looking at western countries, Germany is a clear winner when it comes to university website analytics and privacy. Perhaps past issues in Germany – such as the extensive surveillance by the East German secret police, Stasi, as well as the much more recent incident of alleged tapping of Chancellor Angela Merkel's phone – have had some influence on the situation. Germany has a long tradition of discussing privacy issues and healthy skepticism towards data collection. Privacy is considered a civil right, not just an option. It would be desirable that Germany's strong privacy practices would also spread to other countries, especially to those where the use of analytics services has gotten out of hand. Something can also be learned from China's protectionist approach in preventing data leaks outside of its borders.

Table 2. The numbers of websites using different analytics services.

Analytics service	Universities using the service
Google Analytics	28
Doubleclick	20
YouTube	17
Google Tag Manager	16
Facebook	15
Google Adsense	11
LinkedIn Insight	11
Hotjar	11
Google Ad Services	9
Twitter	7
Siteimprove analytics	5
New Relic	4
Microsoft	4
Baidu	4
Adobe	3
AppNexus	3
The Trade Desk	3
Tealium	3
Quantcast	2
Snapchat	2
LivechatInc	2
Mpulse	1
Coveo Analytics	1
Static.srcspot	1
Qualaroo	1
Crazyegg	1
Tiktok	1
Addthis	1
Yahoo	1
ClickDimensions	1
Sharethis	1

The websites of universities and public sector bodies in general are not a bad place to start the change for better. Public sector bodies and publicly funded institutions in particular should be exemplary by improving privacy of their websites and online services [12]. It is also important to ask the question of whether it is ethical for universities to use analytics and track users [7]. A university website

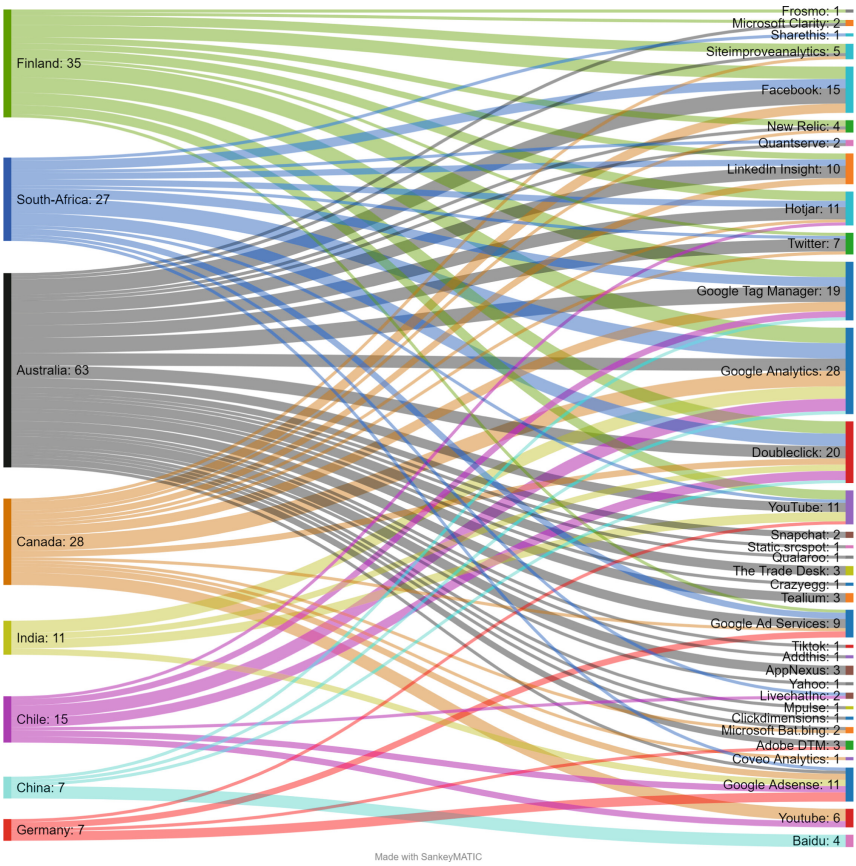


Fig. 2. An alluvial diagram showing the flow of data from the studied countries to different analytics services. The numbers indicate the found analytics service instances per country (on the the left side) and detected instances per analytics service (on the right side). The diagram clearly illustrates how Australia, for example, branches to many different third-party services, while Germany and China only use a few.

should not be a profit-making commodity benefiting third parties but rather a platform for advancing the common good. Institutions with societal impact and often public funding should not be giving away their users' browsing behavior and personal data to analytics providers which use it to gain profit and power. The findings of our study show that too often, users browsing university websites and looking for information have to surrender their data to third-parties, probably without fully realizing this fact. This is also especially problematic because one party, Google, seems to be receiving this information from almost all the websites studied.

It is important to note that it is not some insignificant technical details but identifying information (such as IP addresses and user or device identifiers) that

are often delivered to third parties. What is more, analytics services can also use the context information about the page the user is visiting. In some cases, this can lead to sensitive personal data leaking out. Consider, for example, a student who is searching information on how to reach an accessibility planning officer to get help with special arrangements for his or her studies. Similarly, a student could be looking to talk with psychiatrist or trying to report a harassment case. Even the fact that the student visits pages related to these themes is highly sensitive in itself. Web developers may not always realize this dangerous connection between analytics and delicate web content. Also, as this content with the help on content management systems, university staff creating the potentially sensitive content page may not even have deeper technical skills or knowledge of the use of analytics services on the university website or how web beacons and analytics even work in general. Moreover, it may not be easy to content creators to comprehend the fact that e.g. embedding an instructional YouTube video on a web page leads to all kinds of information being leaked. Our study did not delve into what kind of personal information exactly may be leaked to third parties but this topic, along with a closer look at what kinds of potentially sensitive pages university websites include, is an interesting subject for further research.

In modern web development, analytics and social media buttons are routinely added to websites, and many web platforms and content management systems make this very easy. At the same time, developers often forget privacy – the fact that embedding e.g. social media buttons or YouTube videos on a websites costs visitors their personal data and privacy is not sufficiently taken into account. If analytics really are needed, web developers and data protection officers should consider using analytics tools that store the data locally, such as Matomo [4,10]. Inspecting third-party requests with Chrome DevTools (like we did in the current study) or tools such as Web Evidence Collector[2] should also be an integral part of the testing phase in the web development process. Moreover, to avoid websites teeming with different analytics services, the purpose of each used service should be clearly documented and justified.

5 Conclusions

In the current study, we have provided an overview of the analytics use on university websites around the globe. The high numbers of used analytics services, especially in technically advanced western countries, raise questions about user privacy and how universities portray themselves online. While there were also some positive signs, such as adoption of local analytics in a couple of universities, the findings clearly indicate that web developers and data protection officers should pay closer attention to what data their websites send out and where. Analyzing data flows to third-party services and building websites with user privacy in mind need to become essential parts of web development.

[2] https://edps.europa.eu/edps-inspection-software_en.

Acknowledgements. This research has been funded by Academy of Finland project 327397, IDA – Intimacy in Data-Driven Culture.

References

1. Chandel, S., Jingji, Z., Yunnan, Y., Jingyao, S., Zhipeng, Z.: The golden shield project of China: A decade later-an in-depth study of the great firewall. In: 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 111–119. IEEE (2019)
2. Chandler, A., Wallace, M.: Using Piwik instead of Google analytics at the Cornell university library. Serials Libr. **71**(3–4), 173–179 (2016)
3. Farney, T., McHale, N.: Introducing google analytics for libraries. Libr. Technol. Rep. **49**(4), 5–8 (2013)
4. Gamalielsson, J., Lundell, B., Butler, S., Brax, C., Persson, T., Mattsson, A., Gustavsson, T., Feist, J., Lönroth, E.: Towards open government through open source software for web analytics: the case of Matomo. JeDEM-eJournal eDemocracy Open Government **13**(2), 133–153 (2021)
5. Griffin, M., Taylor, T.I.: Employing analytics to guide a data-driven review of LibGuides. J. Web Libr. **12**(3), 147–159 (2018)
6. Heino, T., Carlsson, R., Rauti, S., Leppänen, V.: Assessing discrepancies between network traffic and privacy policies of public sector web services. In: Proceedings of the 17th International Conference on Availability, Reliability and Security, pp. 1–6 (2022)
7. Jordan, K.: Degrees of intrusion? A survey of cookies used by UK Higher Education institutional websites and their implications (2018). https://ssrn.com/abstract=3142312
8. Liang, F., Das, V., Kostyuk, N., Hussain, M.M.: Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. Policy Internet **10**(4), 415–453 (2018)
9. Mayer, J.R., Mitchell, J.C.: Third-party web tracking: policy and technology. In: 2012 IEEE Symposium on Security and Privacy, pp. 413–427. IEEE (2012)
10. Quintel, D., Wilson, R.: Analytics and privacy. Inf. Technol. Libr. **39**(3) (2020)
11. Sharma, T., Bashir, M.: Use of apps in the COVID-19 response and the loss of privacy protection. Nat. Med. **26**(8), 1165–1167 (2020)
12. Thompson, N., Ravindran, R., Nicosia, S.: Government data does not mean data governance: lessons learned from a public sector application audit. Government Inf. Q. **32**(3), 316–322 (2015)
13. Wambach, T., Bräunlich, K.: The evolution of third-party web tracking. In: International Conference on Information Systems Security and Privacy, pp. 130–147. Springer (2016)
14. Yang, L., Perrin, J.M.: Tutorials on google analytics: how to craft a web analytics report for a library web site. J. Web Libr. **8**(4), 404–417 (2014)

A Systematic Literature Review on Security Aspects of Virtualization

Jehan Hasneen[1], Vishnupriya Narayanan[2], and Kazi Masum Sadique[3](✉)

[1] Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka 1000, Bangladesh
[2] Department of Computer Science, Blekinge Institute of Technology (BTH), Karlskrona, Sweden
[3] Department of Computer and Systems Sciences, Stockholm University, Borgarfjordsgatan 8, 164 07 Kista, Sweden
sadique@dsv.su.se

Abstract. Cloud computing is an emerging technology where organizations can have the flexibility of their infrastructure and store and retrieve data from cloud storage. The main advantages of cloud computing are speed, cost reduction, and scalability. Cloud computing is mainly developed on the virtualization of resources with reusability capabilities. In virtualization, computer systems are deployed virtually as software and hardware resources are shared between virtual machines. It allows multiple operating systems and software applications to run on a single server. Virtualization plays an essential role in the cloud environment. Though cloud computing improves the performance and quality of work for organizations, it has security, privacy, and trust-related threats. In this paper, we have surveyed the security aspects of virtualization. We have discussed cloud computing and virtualization, followed by a detailed analysis of virtualization security. We have also proposed a multi-agent-based model to improve virtualization security.

Keywords: Virtualization · Security · Vulnerability · Cloud Computing · Cloud Security · VMM

1 Introduction

Cloud computing is a popular paradigm for reliable data storage, easy application accessibility, and reduced cost for service. It is famous for easy deployment, management, and less onside resource requirement [1, 2]. Cost efficiency, easy monitoring, and support made cloud computing a huge success. Most IT organizations are stepping into implementing cloud architecture. The main technology behind Cloud computing is virtualization [3, 4]. Virtualization offers the abstraction layer between multiple instances of single physical hardware. It also offers monitoring services to N number of virtual machines. It also provides useful features like performance isolation, service consolidation, and live migration [5, 6]. Virtualization requires high initial investments, but it offers efficient hardware uses. Despite the numerous benefits of Virtualization technology, it has its own security vulnerabilities and issues [7, 8]. In this paper, we will discuss

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1293–1302, 2023.
https://doi.org/10.1007/978-3-031-27409-1_118

virtualization security threats and vulnerabilities along with their viable countermeasures. Below we have discussed different terms related to virtualization security before discussing the security aspects of virtualization in detail.

1.1 Cloud Computing

Cloud computing is the paradigm for outsourcing IT services which reduces complexity and cost and increases reliability, scalability, and availability [1]. There are different types of cloud service and deployment models. We have presented cloud service model based on sizes in Fig. 1 and cloud deployment model based on accessibility in Fig. 2.

Cloud Service Models. Below we have presented different types of cloud service models [9, 10]:

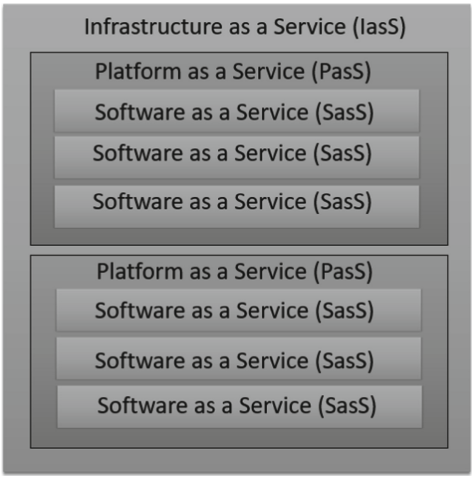


Fig. 1. Cloud service models based on size.

Software as a Service (SaaS). Here the customer can access the application provided by the cloud provider. The customer cannot control the infrastructure, such as servers, OS, or storage. Customers only have the right to access the application running on the Cloud Provider side with limited application configuration permissions that enable the customers to execute necessary changes. These are security concerns regarding this framework data access risk, identity theft, and control over your information.

Platform as a Service (PaaS). Here the customer can create their application and deploy the application in the environment offered by the cloud service provider by using the libraries, programming language tools, and services. But like SaaS, customers will not have access to the servers, OS, or storage rather they have full access to the deployed application configuration.

Infrastructure as a Service (IaaS). In an IaaS model, the customer is provided with full access to processing, networking, and storage capacity. With this, customers can deploy or develop applications with their operating systems. Here customers have more control over the environment.

Cloud Deployment Models. The cloud deployment model is divided into four categories based on organizational needs and strategy [6]. All four categories of cloud deployment models are discussed below.

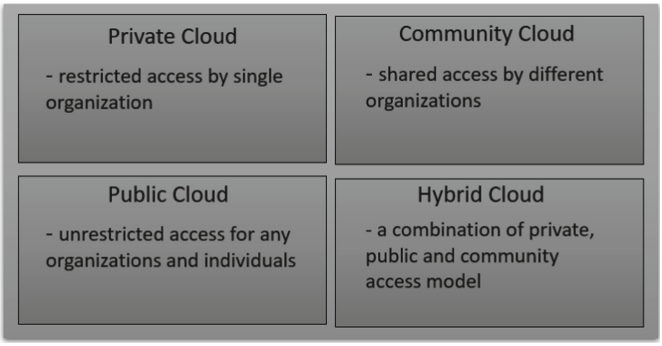


Fig. 2. Cloud deployment models based on access.

Private Cloud. A single organization uses a private cloud. The data owner has better control and security in a private cloud than in other cloud deployment models. Access to data from external can be controlled using firewall devices. The cloud servers can be deployed internally or externally, but access to the data needs to be secure using firewalls and virtual private network (VPN) devices.

Community Cloud. A community cloud deployment model is where several similar organizations share the same database and/or resources. The organizations follow the same security and privacy model. An example could be civic registration numbers.

Public Cloud. It is the cloud model where resources are available for public use. It primarily considers the pay-per-use model and is suitable for small organizations. But security is a significant concern in this type of deployment model.

Hybrid Cloud. A hybrid cloud deployment model is a combination of private, community and public clouds. In this model, organizations use private clouds for their critical data and community cloud for shared data and resources with others. It may also include public access for unprotected data. A miss configuration of this type of model increases security vulnerabilities.

2 Research Goals and Research Questions

We have considered specific research goals and research questions for this systematic literature review work. Our research goal and research questions are described below.

2.1 Research Goal

Our research goal for this study was to identify security issues in virtualization and to discuss countermeasures. To achieve this goal, we set the following questions for our research.

2.2 Research Questions

We have considered the following research questions for this systematic literature review:

- What is the current state of security of virtualization?
- What are the challenges and open issues within the security domain of virtualization?
- What are the possible mitigation measures?

3 Research Methodology

In this paper, we have performed a systematic literature review within the virtualization security domain. A systemic literature review is a process of finding previous research within a certain domain [11]. It also helps researchers to gather any domain knowledge in their early research career. To perform the systematic literature review, we have considered a few selection criteria during our search. We have used the following inclusion and exclusion criteria.

3.1 Inclusion Criteria

The following inclusion criteria are considered:

- Only journal and conference papers are considered.
- Papers that are written in English are considered.
- Papers which somehow related to virtualization and security aspects are included.
- Papers published in digital format are considered.

3.2 Exclusion Criteria

The following exclusion criteria are considered:

- Books and other sources are excluded.
- Papers that are not written in English are excluded.
- Papers unrelated to security, virtualization, cloud computing, and containers are excluded.

Three electronic databases are used to search the related articles and papers. The name of the databases and the search keywords are presented in Table 1.

4 Results and Discussion

In this study, we have identified 29 related papers for analysis. After analyzing the papers, we identified two types of virtualization security vulnerability aspects: administrative vulnerability and technical vulnerability. We have also discussed mitigation measures followed by our proposed model.

Table 1. Selected databases table.

Database	Searched in
IEEE Xplore	Metadata only
Scopus	Article title, abstract, keywords
Web of science	Topic

4.1 Administrative Aspects

As virtualized cloud deployment models vary among private, public, and shared forms, databases and resources could be located anywhere under any kind of administration. It is not a very easy task to have proper administration/management on virtually deployed systems. Lack of adequate administration makes virtual machines vulnerable to security threats.

4.2 Technical Aspects

The technical vulnerabilities are connected to both hardware and software. Below we pointed out several security vulnerabilities domains of virtualization, which are also partially listed in [12]:

- Virtual Machine (VM) Sprawl - the uncontrolled proliferation of VMs: The creation of duplicate virtual machines, bringing them up and leaving them abandoned in the network, creates VM sprawling. VM sprawling opens up weak points for hacking the virtualized system [12, 14].
- User's/Organization's Sensitive Data Within a VM: Virtual Machines (VMs) from different clients operating with sensitive data need proper separation. Co-locating sensitive VMs clients in the same physical server increases vulnerabilities.
- Control and visibility on VMs: The control and visibility of VMs within the unauthorized community is a security threat because a dishonest user can access sensitive personal or organizational data and change the system's behavior [14].
- Use of Pre-Configured images (Golden Image): The concept of golden image is used for quick installing systems. But it could introduce security vulnerabilities because a golden image is not continuously updated to protect the system against new threats [14].
- Resource Exhaustion (DDoS attacks): Attackers can exhaust system resources. For example, a DDoS attack can stop the service by sending massive unwanted traffic to the system [14].
- Hypervisor (VMM) Security (VM Escape): Hypervisor works as a platform and provides the structure to operate and run Virtual Machines (VMs). Attack on hypervisor security can be fatal in terms of a single-point failure of the entire infrastructure [13].

4.3 Mitigation Measures

The following mitigation measures [14] can reduce vulnerabilities of virtualization:

- Secure Identity Management and Authentication: The identity management model of the cloud service users can be improved further to remove access of unauthenticated users.
- Access control: The access control for users can be improved with minimal access for users and more border access for the administrators.
- Secure Programming: While developing cloud services, it is essential to consider security in the first place and not at the end. It reduces backdoor access.
- Ensure Network Security: As we saw, there are different types of cloud service models, and access to data can be further improved with network security. For example, having Virtual Private Network (VPN) will remove malicious third-party access to sensitive data.
- System Hardening (Hypervisor): In this process, access to the system is improved by considering different best practices. For example, replacing the default password, regular updates of the system, encrypted data transmission, etc.
- Others (Physical access, Logging, Monitoring, and VM separation): Restricted physical access to the data center is crucial. Logging and monitoring are always important, and separating VMs reduces the risk of data leakages.

Virtualization security vulnerabilities can be mitigated if we can ensure all the above-mentioned criteria.

4.4 Proposed Model

As discussed above, virtualization vulnerabilities can be introduced by human mistakes and malicious user and system activities. Automated security management is worthwhile here. One of the ways to automate monitoring and defense is a multiagent system. Multiagent systems work collaboratively; they are also capable of self-decision-making [15]. Based on our analysis of the results, we have proposed a multiagent-based virtualization security enchantment model in Fig. 3.

Our proposed multiagent-based model has nine components: duplication monitoring agent, resource exhaustion monitoring agent, update monitoring agent, authentication management agent, authorization management agent, identity management agent, collaboration management agent, network monitoring agent, and integrity monitoring agent. Below we have described each of the blocks.

Duplication Monitoring Agent: The duplication monitoring agent should monitor the system for duplication of virtual machines within the same network. A duplicate entry can be a redundant system. In case a duplication is detected without an entry as redundant, the duplication monitoring agent should share the details of those systems with the system administrator with a warning that duplicate entries without a redundant flag are not allowed.

Resource Exhaustion Monitoring Agent: The resource exhaustion monitoring agent is responsible for continuous monitoring of access and use of resources. If any unexpected steps are found in the use of system resources, the resource exhaustion monitoring agent should block the request from the source, which is identified as a malicious user or system.

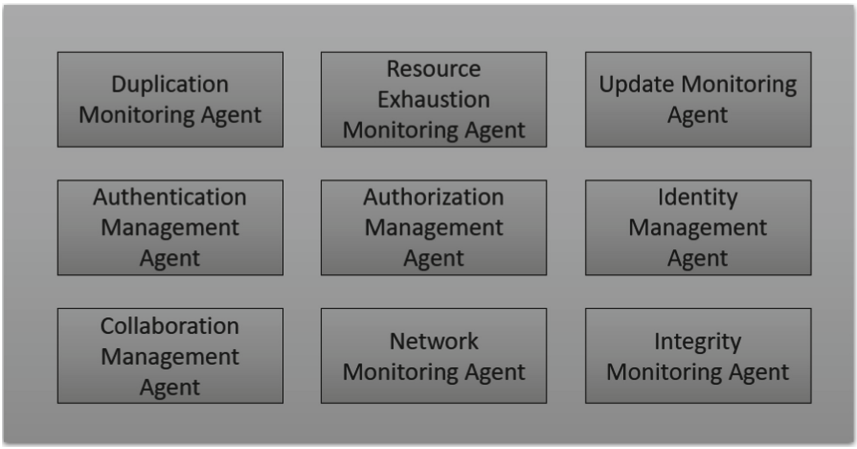


Fig. 3. Multiagent-based model for virtualization security enhancement

Update Monitoring Agent: The update monitoring agent is responsible for monitoring system update and compatibility of updates with the currently running service in the system. If the update monitoring system finds that the update is a security related update and is compatible with the current services, it will possibly automatically update the system after sending a notification to the system administrator. If the update monitoring agent became unable to verify the compatibility, it shouldn't perform the update but only notify the system administrator that new security update is available, and a manual compatibility test is expected by the system administrator.

Authentication Management Agent: The authentication management agent is responsible for ensuring authenticated access to the system. The authentication management agent should block any system configuration without authentication.

Authorization Management Agent: The authorization management agent is responsible for verifying and maintaining authorization entries. If any unauthorized access is detected, the source should be immediately blocked by the authorization management agent.

Identity Management Agent: The identity management agent is responsible for managing the identity information of users and systems accessing different virtual machines. It consciously updates its entry, and requests from compromised systems should be flagged in the database.

Collaboration Management Agent: The collaboration management agent is responsible for monitoring the collaboration between different virtual machines. It is also responsible for internal collaboration between different agents for making group decisions.

Network Monitoring Agent: The network monitor agent monitors the network to ensure malicious traffic toward the host machine, which may create resource exhaustion. The agent notifies the network administrator and the system administrator about malicious activity within the network and specifically towards the guest or host machine is concerned.

Integrity Monitoring Agent: The integrity monitoring agent ensures data integrity in the virtual environment. It would notify the system administrator if any integrity error occurred.

5 Related Works

As we have proposed a multi-agent-based virtualization security model in this paper, we have also reviewed related works. So far, we have yet to find an article that discusses explicitly any similar solution to what we have proposed above. But some papers discussed multi-agent-based cloud security solutions. In [16], the authors proposed a multiagent-based cloud monitoring system, where a master agent collaborates with the collector agent to collect information about the cloud environment and communicates with the worker agents to perform reactive tasks. In [17], the authors performed a systematic literature review on cloud storage security frameworks where multiagent solutions are proposed. They have identified several papers that discussed cloud data security challenges and possible solutions for the improvement of cloud data security using multiagent. In the end, they have proposed their model for cloud data storage security using multiagent. Another article [18] discussed cloud data storage security based on the multiagent system. The authors suggested client-side data encryption in the proposed solution, and on the cloud side, they presented the multiagent solution to check the data integrity. In [19], the authors proposed a new multiagent based distributed intrusion detection system (MAS-DIDS) for cloud computing environments. In their model, the agents negotiate for interaction and communication to create alarms for any unwanted behavior within the cloud platform. Another article also proposes a multiagent-based intrusion detection system for cloud security [20]. The authors presented a model with several agents at different layers: the interface layer, the mediation layer, and the control layer. The agent at the interface layer checks the packets from the network and sends them to the following agent, called the monitoring agent, who sends it further to the analysis agent, where fundamental analysis is performed based on specific rule sets. A supervisor agent at the control layer is proposed for collaboration between the intrusion detection agent and the intrusion prevention agent. The intrusion detection agent creates alert and advisory reports, while the intrusion prevention agent is responsible for blocking the attackers. As we mentioned earlier, from our literature review on multiagent based virtualization security solutions, we couldn't find any article that proposed a similar solution to ours.

6 Conclusions and Future Works

We have performed a systematic literature review on virtualization security in this paper. We have identified 29 papers not listed in the reference because of limitations for the number of references. Considering our research questions, we have pointed out several security vulnerabilities and related security measures. We have extended our findings with mitigation measures based on our analysis of the identified papers. In the end, we have proposed a multiagent-based model for the enhancement of virtualization security. In our future work, we will extend our proposed multiagent-based model with testbed

implementations. Virtualization and cloud computing services are provided to different individuals and organizations, and those can include private data. The feasibility of a multiagent-based solution can also be explored in future research.

References

1. Abdelrahem, O., Bahaa-eldin, A.M., Taha, A.: Virtualization security: a survey. pp. 32–40 (2016)
2. The NIST definition of cloud computing. https://csrc.nist.gov/publications/detail/sp/800-145/final. Last accessed 31 Dec 2020
3. Riddle, A.R., Chung, S.M.: A survey on the security of hypervisors in cloud computing. In: Proceedings of the 2015 IEEE 35th International Conference on Distributed Computing Systems Workshops (ICDCSW 2015), pp. 100–104 (2015). https://doi.org/10.1109/ICDCSW.2015.28
4. Dimitrov, M., Osman, I.: The impact of cloud computing on organizations in regard to cost and security. pp. 29–30 (2012)
5. Li, Y., Li, W., Jiang, C.: A survey of virtual machine system: current technology and future trends. In: 3rd International Symposium on Electronic Commerce and Security (ISECS 2010), pp. 332–336 (2010). https://doi.org/10.1109/ISECS.2010.80
6. Tangirala, S.: Efficient big data analytics and management through the usage of cloud architecture. J. Adv. Inf. Technol. 302–307 (2016). https://doi.org/10.12720/jait.7.4.302-307
7. Di Pietro, R., Lombardi, F.: Virtualization technologies and cloud security: advantages, issues, and perspectives. arXiv (2018)
8. Tank, D., Aggarwal, A., Chaubey, N.: Virtualization vulnerabilities, security issues, and solutions: a critical study and comparison. Int. J. Inf. Technol. **14**(2), 847–862 (2019). https://doi.org/10.1007/s41870-019-00294-x
9. IaaS, PaaS and SaaS—IBM Cloud service models. https://www.ibm.com/cloud/learn/iaas-paas-saas. Last accessed 31 Oct 2022
10. SaaS vs PaaS vs IaaS: what's the difference & how to choose. https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/. Last accessed 31 Oct 2022
11. Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M.: Lessons from applying the systematic literature review process within the software engineering domain. J. Syst. Softw. **80**, 571–583 (2007). https://doi.org/10.1016/j.jss.2006.07.009
12. Top 11 virtualization risks identified. https://www.networkcomputing.com/data-centers/top-11-virtualization-risks-identified. Last accessed 10 Nov 2022
13. Asvija, B., Eswari, R., Bijoy, M.B.: Security in hardware assisted virtualization for cloud computing—state of the art issues and challenges. Comput. Netw. **151**, 68–92 (2019). https://doi.org/10.1016/j.comnet.2019.01.013
14. Cloud security alliance: best practices for mitigating risks in virtualized environments. pp. 1–35 (2015)
15. Dorri, A., Kanhere, S.S., Jurdak, R.: Multi-agent systems: a survey. IEEE Access **6**, 28573–28593 (2018)
16. Grzonka, D., Jakóbik, A., Kołodziej, J., Pllana, S.: Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security. Futur. Gener. Comput. Syst. **86**, 1106–1117 (2018)
17. Talib, A.M., Atan, R., Abdullah, R., Murad, M.A.A.: Security framework of cloud data storage based on multi agent system architecture: semantic literature review. Comput. Inf. Sci. **3**(4), 175 (2010)

18. Arki, O., Zitouni, A., Eddine Dib, A.T.: A multi-agent security framework for cloud data storage. Multiagent Grid Syst. **14**(4), 357–382 (2018)
19. Achbarou, O., El Kiram, M.A., Bourkoukou, O., Elbouanani, S.: A new distributed intrusion detection system based on multi-agent system for cloud environment. Int. J. Commun. Netw. Inf. Secur. **10**(3), 526 (2018)
20. Achbarou, O., El Kiram, M.A., Elbouanani, S.: Cloud security: a multi agent approach based intrusion detection system. Indian J. Sci. Technol. **10**(18), 1–6 (2017)

Detection of Presentation Attacks on Facial Authentication Systems Using Intel RealSense Depth Cameras

A. A. Tarasov[1], A. Y. Denisova[1,2(✉)], and V. A. Fedoseev[1,2]

[1] Samara National Research University, Moskovskoe shosse, 34, Samara 443086, Russia
anna3138@yandex.ru
[2] IPSI RAS—Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardeyskaya 151, Samara 443001, Russia

Abstract. In this paper, we explore the prospects for using Intel RealSense depth cameras to solve the problem of presentation attack detection in facial authentication systems. Studies known to date use quantized depth data. Such an approach makes it impossible to connect these data with anthropometric facial features and scene geometry. In addition, in recent papers, some researchers declared the limitation of using depth cameras in this problem only at small (less than 1 m) distances to the face. In this regard, we have made our collection of 480 samples containing bonafide images and two types of presentation attacks. The samples were taken at distances up to 2 m, as well as in different face positions relative to the camera axis. We also designed a set of features based on texture analysis, anthropometric properties, and face oval localization. The conducted experiments show the advantage of using the original, not quantized depth data. Our study also refutes the thesis that depth cameras can be used for presentation attack detection only at short distances. Additionally, we demonstrate the applicability of the proposed features on one of the publicly available datasets.

Keywords: Presentation attack · Facial dataset · Liveness detection · Facial authentication · Depth data · Intel RealSense · Attack detection · Machine learning · Data collection

1 Introduction

Facial authentication systems are the most popular biometrics-based identification systems used in different spheres of life from banking to home security [1]. The main reason for their wide distribution is the easy non-contact way of authentication. However, face recognition systems are vulnerable to presentation attacks in which illegal user demonstrates a fake photo of the target person to the system. To present presentation attacks, intruders use different presentation attack devices such as printed photos, mannequins, 3D masks, etc. The availability of a person's photos in social networks and other Internet sources as well as a variety of presentation attack methods increase the intruder's

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1303–1314, 2023.
https://doi.org/10.1007/978-3-031-27409-1_119

opportunities. Thus, the security of face recognition systems is under threat without presentation attack detection.

The presentation attack detection (PAD) problem is also known as liveness detection because the goal is to determine whether the input image is a living person's photo or not. There are different kinds of liveness detection mechanisms. The earliest one is image texture analysis [2–4]. It takes into account the difference in the texture of images captured one time (real images) and images captured two or more times (fake images). This approach is suitable only for printed and screen photo detection.

Another PAD mechanism is based on human-computer interaction [5–9]. In this case, some actions inherent to a living person are detected, for example, eye blinking, lip movement, or facial expression changes. For better security, the recognition system generates random queries to the user to perform some appropriate actions, and then the recorded actions are recognized. Randomness prevents facial movement falsification. Nevertheless, the human-computer interaction approach suffers from a long time of data acquisition and sophisticated implementation.

The next PAD approach is life information estimation [10–12]. The physical and biological characteristics of a living person may be measured using video sequences and additional devices registering data about the person simultaneously with the image acquisition process. For example, the heartbeat, micro motions of muscles, and blood flow can be used for liveness detection. The problem with such methods is their high complexity. Also, most of them require contact with the human body across additional devices.

One more approach to liveness detection is image quality analysis [13–15]. Its key idea is the analysis of image quality degradation during deceiving face reproduction. The additional blur, different reflectance properties, and other features distinguish fake images from real ones. However, this approach requires an expensive registration camera and it is still vulnerable to high-quality presentation attack methods.

The last but not least approach is multimodal PAD when additional devices such as depth, thermal or infrared cameras are used to detect features of a living person [16–18]. Depth cameras provide information about the 3D profile of the analyzed object, thermal cameras give information about the temperature of the object and infrared cameras may be effective in distinguishing real skin patterns from artificial materials. In our opinion, the use of additional imaging devices is the most promising way of liveness detection because it keeps the non-contact image acquisition process and it is suitable for a wider range of presentation attack devices than the other approaches.

In our paper, we strive to show that even low-cost depth cameras such as Intel RealSence may provide sufficient PAD effects in collaboration with traditional RGB cameras. However, the conditions of shooting and data preparation have to be investigated. Because of the lack of open access Depth and RGB datasets with different shooting conditions and raw Depth data, we collected our dataset composed from RGB and Depth images acquired by the Intel RealSence camera. We evaluated PAD using images captured with different distances to the object and its position in the frame to discover the potential of Depth data. The other aspect, that we focus on, is data normalization. As most of the Depth data is normalized to the range 0 to 255, the capabilities of raw Depth data remain unexplored.

2 Related Works

In the history of depth-based PAD, the paper of Erdogmus and Marcel [18] is one of the pioneer works. The authors analyzed RGB and Depth images coregistered by the eye position in terms of texture properties. For both modalities, they computed local binary pattern features (LBP) and tested several classifiers such as χ^2-distance, linear discriminant analysis (LDA), and support vector machines (SVM). Sun et al. [19] investigated the other fundamental feature of color and depth images. They applied canonical correlation analysis to define RGB and Depth image pairs with the best correlation. For fake faces, the face contours significantly differ from face contours in RGB images, and the correlation between the images is low. As for classifiers, Sun et al. exploited SVM with radial basis functions and χ^2 kernels. Naveen et al. [20] proposed a combination of local and global features. Local features included discrete cosine transform (DCT) coefficients in eyes and nose areas. Global features were LBP and Binarized Statistical Image Features (BSIF). Both global and local features were computed independently for color and depth images and concatenated into a single feature vector. Naveen et al. used a simple Euclidean distance classifier. The mentioned methods used normalized Depth data and ignored the real distances in the depth signal.

Raghavendra and Busch [21] and Albakri and Alghowinem [22] attempted to provide depth analysis without normalization. Raghavendra and Busch used BSIF features and some local depth features in eye and nose regions. These local features include the absolute depth change value in the eye region and the increase of the nose region in the depth image in comparison to the color image. Raghavendra and Busch applied two independent SVM classifiers and aggregated the classification results. Albakri and Alghowinem tested the simplest depth features such as absolute depth values in five face points. If the depth is equal in two or more points then they classified the image as a presentation attack. This method is suitable for print and screen attacks when the depth value is almost the same for each face point. Zhou et al. [9] proposed a more sophisticated method for constructing depth features but applied an additional near-infrared (NIR) modality to detect presentation attacks. They provided face volume analysis using a Depth map and improved the RANSAC algorithm. Finally, they detected presentation attacks by eye blinking in NIR images and face volume analysis in Depth images.

Lately, researchers concentrated their efforts on neural network based methods used for both feature extraction and classification. For example, Wang et al. [23] used a convolutional neural network (CNN) for feature extraction in RGB images and LBP features for depth images. The entire feature vector was classified by the SVM method. Ewald et al. [24] used three separate CNNs for RGB, NIR, and Depth modalities to detect presentation attacks. George et al. [25] proposed the MCCNN method of liveness detection on RGB, NIR, Depth, and thermal (TIR) data. However, the performance of MCCNN on RGB and Depth data only was worse than in the case of using all modalities. It is worth mentioning that all these CNN-based methods were trained and investigated using Depth data normalized in the range [0,255].

Due to the provided analysis, we can see that the potentials of RGB and Depth data for PAD are underestimated. Our research aims to analyze depth-based features calculated without preliminary depth normalization. We try to answer the question of whether depth data without normalization are beneficial for PAD.

The other goal of our paper is the investigation of object position influence on PAD results. The review [22] provided by Albakri and Alghowinem showed that the problem of existing depth-based PAD methods is the little registration distance to the object. Most of the methods require the registration distance to be less than 1 m. However, it is more convenient to use registration distances of more than 1 m in access control systems. Thus, we decided to study PAD ability when the registration distance is over a meter.

The list of existing publicly available RGB and depth datasets is shown in Table 1. Existing datasets with depth and RGB data usually provide normalized in the range [0,255] depth values and all of these datasets are obtained with the distance to depth camera less than 1 m. To meet our research goals we collected our dataset which included depth data without normalization captured for different face positions in the frame and different registration distances.

Table 1. Publicly available depth and RGB datasets.

Dataset	# subjects	# videos	Depth camera type	Attack types
3DMAD [26]	17	255	Microsoft Kinect	3D Mask
WMCA [25]	72	6716	Intel RealSense SR300	Print, Replay, 2D Mask, 3D Mask, etc.
CASIA-SURF [27]	1000	21000	Intel RealSense SR300	Print, Cut
CSMAD [28]	14	246	Intel RealSense SR300	3D Mask
HQ-WMCA [29]	51	2904	Intel RealSense D415	Print, Replay

3 Dataset Preparation

In our work, we used the Intel RealSense D435 camera. This camera captures depth data with a resolution of up to 1280 × 720. The data is stored in a 16-bit format in millimeters. The resolution of the RGB module is 1920 × 1080. According to camera characteristics, the optimal range for measuring depth data is from 30 cm to 3 m. Intel also distributes the *librealsense* library to work with cameras of this family. This library has some utils for frame synchronization between depth and RGB data streams both in time and in geometry. Figure 1 shows a matched pair of RGB and depth images (the latter one is depicted in pseudo-colors for clarity).

When collecting the dataset, the following conditions varied:

- distance to a face: 1, 1.5, and 2 m;
- the authenticity of a person: a bonafide, a printed photo, and a printed photo cropped along the contour with holes for the eyes;
- position of a face relative to the frame center: 9 rectangular sectors dividing the frame area into equal parts;
- different people.



Fig. 1. An example of a geometrically matched pair of images

In total, the collected dataset contains 480 pairs of synchronized RGB and depth images. 178 of them are bonafide examples, and 302 of them represent presentation attacks. Figure 2 shows some examples of RGB images. The assembled dataset is available for free download at Github[1].

4 Features Used for Liveness Detection

In this work, we did not aim to construct an ideal feature family that would provide the best quality of presentation attack detection. Instead, we studied the feasibility of high detection quality using the Intel RealSense D435 data taken under various conditions. To do this, we considered a set of features justified by reasonable considerations that have fast calculation algorithms. On this set, as well as on its subsets, the quality of classification of presentation attacks was tested under various shooting conditions.

To calculate features, we were seeking a human face in the RGB image using Multitask Cascaded Convolutional Neural Network (MTCNN) [30]. For each found face, a bounding box was allocated, of such a size that it included both the face itself and the background behind the head. Further, the obtained bounding box coordinates were transformed into depth image space. Then we calculated the features in the depth image inside this bounding box. The total feature set consisting of 12 feature groups (both scalar and vector) is shown in Table 2.

As one can see, there are 3 feature subsets in this table. The first one (A1–A4) includes features traditionally used in texture analysis problems, including the PAD problem [25]: variance, gradient histogram, and correlation coefficients. B1–B3 are designed to take into account the anthropometric features of a person and count on the use of source (non-normalized) depth data. Finally, C1–C5 are based on dividing the analyzed bounding box into two regions and comparing their characteristics. The first region includes only pixels that are guaranteed to lie inside the face oval. The second one, as a consequence, should include all pixels that lie outside the face oval. In addition, it also includes a certain (preferably small) number of pixels close to the border of the face oval, for which there is no reliable confidence that they are inside the contour.

To find the face oval contour, we used Face Alignment Network (FAN) [31]. The landmarks obtained at the output of this network were combined into a single closed polygon. The disadvantage of the C1–C5 features is the high computational complexity

[1] https://github.com/vicanfed/depth-pad-dataset.



Fig. 2. RGB images from the collected dataset

Table 2. Groups of features used in the research.

Code	Feature description	Length
A1	Global variance	1
A2	Variance over K non-overlapping regions	K
A3	M samples of depth gradient histogram	M
A4	Correlation coefficients in 4 directions	4
B1	Global difference *max* – *min*	1
B2	Differences *max* – *min* in K non-overlapping regions	K
B3	Mean values in K non-overlapping regions	K
C1	Average depth inside the face oval minus average depth outside	1
C2	Median depth inside the face oval minus median depth outside	1
C3	*max* – *min* difference inside the face oval	1
C4	Variance outside the face oval	1
C5	Variance within the face oval	1

of finding these contours. It may turn out that the use of this procedure is not reasonable in practice, especially given the small total length of this group compared with others (only 5 features). Therefore, in our research, we compared two options: the calculation of features C1–C5 by contours found by FAN, and their calculation by a rough approximation of the area inside the face oval. Specifically, in a second way, we used a rectangle located in the center of the bounding box and occupied 1/4 of the area inside the bounding

box (see Fig. 3). The results of the experiments comparing these two options are given in Sect. 5. However, looking ahead, we may note that the use of a rough approximation based on their results is recognized as rea.

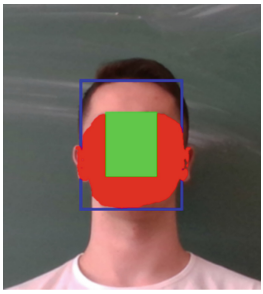


Fig. 3. Areas used to calculate the features: *blue* – original bounding box, *red* – area inside the face oval, obtained by FAN, *green* – rough approximation of the red area by a rectangle

5 Experiments and Discussion

5.1 Experimental Conditions

In the preliminary studies, several classical classifier models were tested: linear SVM, SVM-RBF, Random Forest, kNN. The best results on the selected set of features were shown by Random Forest, so next, we used only this classifier model. The quality of training models was assessed using cross-validation by splitting the dataset into 10 subsets. Accuracy and F1 were used as the main indicators.

Even before the experiments, it was clear that the set of features specified in Table 2 is redundant since many of them have a similar meaning, and also since for large K and M, the total length of the feature vector can become comparable to the sample size, which will inevitably lead to overfitting.

Therefore, in our studies, we formed feature vectors dynamically using a quasi-optimal sequential feature addition procedure. In this procedure, at each step, we add one feature group providing the best quality to the current feature vector. This procedure stops when the classification quality stops growing. It does not guarantee optimality, and it can also lead to different results when any parameters change. However, it helps to evaluate the potential of our set of features in various conditions and to evaluate the usefulness of individual groups of features in solving the PAD problem.

The total length of the full feature set, according to Table 2, is $3K + M + 11$, with K being the square of an integer. In our studies, we selected optimal K values for every composition of input data and study conditions. For that, we analyzed classification quality on three corresponding groups of features (A2, B2, B3). Also, to avoid the excessive complication of the experimental scenario, when calculating feature A3, we simply used $M = K$.

5.2 Source vs. Normalized Depth Data

The purpose of the first experiment was to test the thesis about the advantage of source (non-normalized) depth data over normalized data, used by other scholars (see Table 1).

For normalized data, we defined $K = 25$ experimentally. The chain (B2, A3, C5, A1, B3, A4) with a length of 66 features showed the best result in the sequential procedure. For non-normalized features, the best result is provided by $K = 64$, and the best chain is obtained after the 5th iteration: (A2, B1, A3, B2, C3) with a length of 140 features. Comparative measures of classification quality are shown in Table 3. For both types of data, Table 3 shows both the final result and the result after the first iteration. The impressive difference in quality is not due to the difference in the lengths of the feature vectors.

Table 3. Comparative results of PAD quality using source and normalized depth data.

Type of depth data	Iter. No.	Total length	Accuracy	Precision	Recall	F1
Normalized	1	25	0.917	0.931	0.937	0.934
Normalized	6	66	0.935	0.959	0.937	0.948
Source	1	64	0.944	0.957	0.954	0.955
Source	5	140	0.967	0.977	0.970	0.973

5.3 Fast Calculation of Features Inside of the Face Oval

As noted in Sect. 4, we conducted a study aimed at comparing two methods for calculating features C1–C5 (exact and approximate): in terms of processing time and classification quality. This study was performed by a PC with the following specifications: AMD Ryzen 5 2600 3.9 GHz, RAM: 16 Gb, GPU: Nvidia GeForce GTX 1070.

The average time for accurate feature calculation was 4.465 s per image, and the time for simplified calculation was 1.642 s. The analysis of the classification quality for the two methods showed controversial results. On the source data, the accuracy of classification for C1–C5 was 0.873 with an accurate calculation and 0.842 with a simplified calculation. At the same time, on the normalized data, the result turned out to be the opposite: 0.844 versus 0.87. Taking into account the fact that both quasi-optimal chains obtained in the previous subsection use only one group from C1–C5, and the difference in the accuracy of classification by C1–C5 is 2–3% with almost a threefold time saving, we considered it acceptable in practice to use a simplified calculation.

5.4 Dependence of the Distance to the Object on Detection Quality

One of the most important studies was to test the possibility of using the Intel RealSense camera to detect presentation attacks at distances of more than 1 m. As mentioned in Sect. 2, this case was not considered in previous papers. This will potentially let us use

the proposed method in access control systems, where it is often not possible for a person to approach the camera.

The study was carried out on the best chain of features obtained for source data. The classifier was trained with data taken from all distances, and the quality assessment was carried out by cross-validation with subsequent grouping by distances to the object. The results presented in Table 4 show a decrease in the quality of the classification as the distance increases. However, both F1 and Accuracy exceed 0.9 even at a distance of 2 m from the object, which confirms the efficiency of the algorithm at such a distance.

Table 4. Comparative results of PAD quality at specific distances.

Distance to face	Accuracy	Precision	Recall	F1
1 m	0.982	0.972	1.0	0.986
1.5 m	0.970	0.981	0.971	0.976
2 m	0.923	0.987	0.889	0.935

Table 5. PAD quality on WMCA data.

Feature subset	Total length	Accuracy	Precision	Recall	F1
Selected for WMCA	141	0.944	0.943	0.989	0.965
Selected for our dataset	66	0.924	0.924	0.986	0.954

5.5 Feature Check on the WMCA Dataset

In addition to the applicability of the Intel Realsense camera for solving the PAD problem, including at different distances, we were also interested in the applicability of the features used in this work to other datasets. For this, an additional study was carried out on the WMCA dataset [25]. This dataset, in addition to RGB and Depth, also includes images captured in the infrared and thermal ranges. These channels were not used in our study. Also, WMCA includes new types of presentation attacks that are not represented in the dataset we have collected. At the same time, in the WMCA database, the depth data are normalized, and all faces were captured with the same position on the center of the frame at a very short distance (40 cm).

When using WMCA data, we found that the best classification quality is provided at $K = 64$. The quasi-optimal chain of features was obtained at the 6th step and had the following composition: (B2, C4, C5, C3, B3, A3). For comparison, Table 5 also shows the classification results for the chain that performed best on our features. As follows from the table, both options provide high classification accuracy, which indicates the operability of our set of features for solving the PAD problem on data obtained under various conditions.

6 Conclusion

In this paper, we considered the prospects for using cameras of the Intel RealSense D400 family to solve the problem of presentation attack detection. Since the available datasets contain depth data quantized on the interval [0, 255], it became necessary to collect our collection. The collected dataset contains 480 pairs of synchronized RGB and depth images and is available for free download. We proposed 12 groups of features using textural analysis methods, anthropometric facial features, and methods for highlighting face contours. The conducted experiments showed the unconditional advantage of the original depth data over the quantized ones used in other studies. In addition, the thesis that depth cameras can be used to solve the PAD problem only at distances of less than 1 m is refuted. In addition, the applicability of the proposed features on the WMCA dataset is shown.

Acknowledgments. This work was supported by the Russian Foundation for Basic Research (project 19-29-09045).

References

1. Zhang, M., Zeng, K., Wang, J.: A survey on face anti-spoofing algorithms. J. Inf. Hiding Priv. Prot. **2**(1), 21 (2020)
2. Daniel, N., Anitha, A.: Texture and quality analysis for face spoofing detection. Comput. Electr. Eng. **94**, 107293 (2021)
3. Shu, X., Tang, H., Huang, S.: Face spoofing detection based on chromatic ED-LBP texture feature. Multimed. Syst. **27**(2), 161–176 (2020). https://doi.org/10.1007/s00530-020-00719-9
4. Sthevanie, F., Ramadhani, K.N.: Spoofing detection on facial images recognition using LBP and GLCM combination. J. Phys.: Conf. Ser. **971**(1), 012014 (2018)
5. Hadiprakoso, R.B.: Face anti-spoofing method with blinking eye and HSV texture analysis. IOP Conf. Ser.: Mater. Sci. Eng. **1007**(1), 012034 (2020)
6. Li, Y., Wang, Y., Zhao, Z.: Face anti-spoofing methods based on physical technology and deep learning. In: International Conference on Computer Vision, Application, and Design (CVAD 2021), vol. 12155, pp. 173–184 (2021)
7. Singh, A.K., Joshi, P., Nandi, G.C.: Face recognition with liveness detection using eye and mouth movement. In: Proceedings of International Conference on Signal Propagation and Computer Technology, pp. 592–597 (2014)
8. Ng, E.S., Chia, Y.S.: Face verification using temporal affective cues. In: Proceedings of the 21st International Conference on Pattern Recognition, pp. 1249–1252 (2012)
9. Zhou, J., Ge, C., Yang, J., Yao, H., Qiao, X., Deng, P.: Research and application of face anti-spoofing based on depth camera. In: 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI), pp. 225–229 (2019)
10. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: Proceedings of International Conference on Image Analysis and Signal Processing, pp. 233–236 (2009)
11. Smiatacz, M.: Liveness measurements using optical flow for biometric person authentication. Metrol. Meas. Syst. **19**(2), 257–268 (2012)
12. Li, X., Komulainen, J., Zhao, G.: Generalized face anti-spoofing by detecting pulse from face videos. In: Proceedings of IEEE 23rd International Conference on Pattern Recognition, pp. 4239–4244 (2016)

13. Chang, H.H., Yeh, C.H.: Face anti-spoofing detection based on multi-scale image quality assessment. Image Vis. Comput. **121**, 104428 (2022)
14. Galbally, J., Marcel, S.: Face anti-spoofing based on general image quality assessment. In: Proceedings of 22nd International Conference on Pattern Recognition, pp. 1173–1178 (2014)
15. Galbally, J., Marcel, S., Fierrez, J.: Image quality assessment for fake biometric detection: application to iris, fingerprint, and face recognition. IEEE Trans. Image Process. **23**(2), 710–724 (2014)
16. Mohamed, S., Ghoneim, A., Youssif, A.: Visible/infrared face spoofing detection using texture descriptors. MATEC Web Conf. **292**, 04006 (2019)
17. Sun, L., Huang, W.B., Wu, M.H.: TIR/VIS correlation for liveness detection in face recognition. In: International Conference on Computer Analysis of Images and Patterns, pp. 114–121 (2011)
18. Erdogmus, N., Marcel, S.: Spoofing 2D face recognition systems with 3D masks and anti-spoofing with Kinect. In: 2013 International Conference of the BIOSIG Special Interest Group (BIOSIG), pp. 1–8 (2013)
19. Sun, X., Huang, L., Liu, C.: Multimodal face spoofing detection via RGB-D images. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2221–2226 (2018)
20. Naveen, S., Fathima, R.S., Moni, R.S.: Face recognition and authentication using LBP and BSIF mask detection and elimination. In: 2016 International Conference on Communication Systems and Networks (ComNet), pp. 99–102 (2016)
21. Raghavendra, R., Busch, C.: Novel presentation attack detection algorithm for face recognition system: application to 3d face mask attack. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 323–327 (2014)
22. Albakri, G., Alghowinem, S.: The effectiveness of depth data in liveness face authentication using 3D sensor cameras. Sensors **19**(8), 1928 (2019)
23. Wang, Y., Nian, F., Li, T., Meng, Z., Wang, K.: Robust face anti-spoofing with depth information. J. Vis. Commun. Image Represent. **49**, 332–337 (2017)
24. Ewald, K.E., Zeng, L., Mawuli, C.B., Abubakar, H.S., Victor, A.: Applying CNN with extracted facial patches using 3 modalities to detect 3D face spoof. In: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 216–221 (2020)
25. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans. Inf. Forensics Secur. **15**, 42–55 (2019)
26. Erdogmus, N., Marcel, S.: Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6 (2013)
27. Zhang, S., et al.: Casia-surf: a large-scale multi-modal benchmark for face anti-spoofing. IEEE Trans. Biom., Behav., Identity Sci. **2**(2), 182–193 (2020)
28. Bhattacharjee, S., Mohammadi, A., Marcel, S.: Spoofing deep face recognition with custom silicone masks. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–7 (2018)
29. Heusch, G., George, A., Geissbühler, D., Mostaani, Z., Marcel, S.: Deep models and shortwave infrared information to detect face presentation attacks. IEEE Trans. Biom., Behav., Identity Sci. **2**(4), 399–409 (2020)

30. Jiang, B., Ren, Q., Dai, F., Xiong, J., Yang, J., Gui, G.: Multi-task cascaded convolutional neural networks for real-time dynamic face recognition method. In: Liang, Q., Liu, X., Na, Z., Wang, W., Mu, J., Zhang, B. (eds.) Communications, Signal Processing, and Systems, pp. 59–66. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-6508-9_8
31. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1021–1030. IEEE, Venice (2017)

Big Data Between Quality and Security

Hiba El Balbali(✉), Anas Abou El Kalam, and Mohamed Talha

Cadi Ayyad University, National School of Applied Sciences, Marrakech, Morocco
elbalbalih@gmail.com, a.abouelkalam@uca.ac.ma, mohamed.talha@icloud.com

Abstract. Data is one of the most precious assets an organization can have; it may have a huge impact on its longterm performance, or even, existence. Hence, data quality and data security remain a real challenge. Ensuring the quality and security of data should never be considered as an expense, but as a wise investment. On the one hand, data should be protected to prevent attacks or violations of their confidentiality, integrity, and availability. On the other hand, it must be of high quality to ensure the efficiency of the decision-making process. Unfortunately, the majority of existing works deal with quality and security separately whereas these two areas are closely related and may be jointly addressed;this can be handled in three ways: quality and security could be mutually blocked, security can be used to improve quality, and inversely, quality can serve to enhance security. In this paper, we present a study on big data quality and security, the conflict between them as well as our proposed approach that uses Artificial Intelligence to enforce quality in the service of security by improving the quality of log data before applying Machine Learning algorithms to detect threats.

Keywords: Big Data · Big Data Security · Big Data Quality · Machine Learning

1 Introduction

Big Data refers to the set of tools and technologies that facilitate the processing of this large volume of data generated at high speed and with great variety. It is commonly described through five characteristics.

- Volume: it refers to the large amount of data generated.
- Variety: it refers to the diversity of the data.
- Velocity: it refers to the speed at which data is generated and processed.
- Value: it refers to the profits derived from the data.
- Veracity: it refers to the credibility of the data.

According to several studies [1–4], Big Data faces many challenges, including quality and security. These challenges are mainly due to huge data amounts and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1315–1326, 2023.
https://doi.org/10.1007/978-3-031-27409-1_120

heterogeneity, reliability of data and their sources, and so on. All these characteristics can lead to security breaches or quality deterioration. Furthermore, assessing and ensuring quality can be a security barrier and backward.

Unfortunately, the majority of existing works treat quality and security separately, even though the two are linked and can be addressed through three approaches (see Fig. 1): The conflict between quality and security in Big Data, which states that ensuring quality requires certain flexibility which may compromise security. Quality at the service of Security that can be ensured by improving the quality of data used in security approaches; and, finally, Security at the service of Quality: strengthening certain security properties may enhance quality.

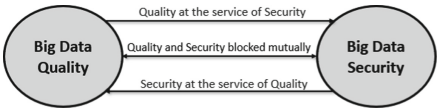


Fig. 1. Big Data between Quality and Security.

Subsequently, this paper is organized as follows: The next section will be devoted to the quality of Big Data. The third section will then address the security of Big Data. The fourth section will be dedicated to the discussion of the conflict that may arise between quality and security in the context of Big Data and highlights quality at the service of Security and vice versa. The fifth section will introduce our AI-based proposed approach to enforce quality in the service of security. Finally, the last section will conclude the paper and present some perspective works.

2 Big Data Quality

2.1 Quality Concept in the Big Data Context

Data quality is a multidimensional concept that represents the evaluation of information and is characterized by a set of measurable dimensions. Effective data quality management is critical to any consistent data analysis, as data quality is crucial to gaining actionable and accurate insights from available data.

Several studies were interested in data quality. The authors of [5] defined data quality as "Fitness for use", i.e. data of good quality is data suitable for use by data consumers. According to [6], other definitions have been introduced; such as compliance with requirements. The General Administration of Quality Control explained that data quality represents the degree to which a set of inherent characteristics meets specifications [7].

Besides that, let us precise that data quality is generally assessed using several dimensions. A dimension is a set of attributes that represent a single aspect of data quality [5]. The authors of [2] defined a dimension as a measurable

property of data quality that represents some aspect of data, such as accuracy, precision, consistency, etc. Dimensions are evaluated using metrics; a metric is a quantifiable instrument that defines how a dimension is measured [8].

Regarding the existing dimensions, no agreement has been made on their exact number. The authors of [5] conducted a study that allowed them to identify 179 dimensions of data quality. They did a second study on the importance of these dimensions and they reduced it to 20 dimensions including Completeness, Accuracy, Consistency etc. We summarize the most important Data Quality dimensions as follows [9]:

- Completeness: it verifies that the data is sufficient to reach findings. This can be determined by ensuring that no data is missing in a data set. Completeness can be measured using the Eq. (1)

$$Comp = Nmv/N \tag{1}$$

Nmv : Number of missing values
N: Total number of values

- Accuracy: it describes the similarity between a value v and a value v', considered as the correct representation of the actual phenomenon that v aims at representing [8]. Accuracy can be measured using the equation (2).

$$Acc = Ncv/N \tag{2}$$

Ncv : Number of correct values
N: Total number of values

- Consistency: it verifies that data from all systems in an institution is synchronized and reflects the same information. Consistency can be measured using the equation (3).

$$Cons = Nvrc/N \tag{3}$$

Nvrc: Number of values that respects the constraints
N: Total number of values

Following this same principle, we can consider the following functions to quantify the following dimensions:

- Uniqueness: it ensures that there are no duplicates present in the data. Uniqueness can be measured using the equation (4).

$$Uni = Ndv/N \tag{4}$$

Ndv: Number of duplicated values
N: Total number of values

– Validity: it checks if the rules have been fulfilled, and the data is correct (format, type, and range). Validity can be measured using the equation (5).

$$Val = Nvv/N \tag{5}$$

Nvv: Number of valid values
N: Total number of values

– Timeliness : it verifies if the information is accessible when it is needed. Timeliness can be measured using the equation (6).

$$Tim = Navn/N \tag{6}$$

Navn: Number of accessible values when needed
N: Total number of values

2.2 Big Data Quality Challenges

Data quality in Big Data faces several challenges [2,6] :

– Heterogeneity: data is collected from different sources (IoT, industries, scientific results, etc.). This diversity brings several types of data and complex data structures:
- Structured data: data is predefined and formatted according to a well-defined and precise model like relational databases.
- Semi-structured data: data is not stored in a relational database but has some organizational characteristics that make it easier to evaluate. It constitutes an intermediate form between structured and unstructured data, such as delimited files. They contain tags or other markers to separate items.
– The large volume of data: the volume of data is enormous and it is difficult to assess the quality of the data in a reasonable time, especially the unstructured ones.
– Data Change: data changes very quickly and can become obsolete, this requires higher demands on processing technology. False information can result if data is not captured in real-time.
– Security : according to [11], privacy and confidentiality are barriers to data accessibility. To ensure the quality of the data, it is necessary to have access to it and to have the authorization to carry out transformations, which can be blocked by the principles of security.

The authors of [12] proposed a data quality assessment process that consists of eight components, starting with determining the goals of data collection, selecting necessary data quality dimensions, and determining an evaluation reference. Once the preparation of the quality assessment is complete, the data acquisition phase begins, followed by the data cleaning step.

3 Big Data Security

Big data security refers to all the policies, models, and mechanisms (e.g., measures, tools) used to protect both the data and the analytics processes from information theft, DDoS attacks, ransomware, or other malicious activity that can come from offline or online sources and can crash a system. It covers both internal and external, as well as accidental and malicious threats.

Big Data projects usually have three main stages. Each of these stages should be guarded by solid security standards. Failure to implement security precautions when storing and processing Big Data might lead to data breaches. The first step is "Data Acquisition". While in transit, data might be corrupted or intercepted. The next phase is "Data Storage". At this point, data can become a victim of fraud or be held hostage (both in the Cloud or On-Premise servers) [13]. "Data Consumption" is the final stage. The information could be used by vicious intruders to gain access to computers.

Big Data features, namely 5Vs, may cause several security issues and threaten security principles. For example, data integrity concerns may come as a result of Hardware or Software errors, intrusions, etc. which can lead to many fatal consequences such as Data Poisoning, losses, thefts, ...

Generally, the availability of data is fulfilled by applying multiple data replications and this can affect data integrity; e.g., since any legitimate modification or update must be reported correctly on all sources. In addition, the replication process itself must ensure consistency and integrity (no loss or modification during the process deployment). Moreover, availability can harm privacy by simplifying the combination and the analysis of information and the deduction of sensitive information on individuals [14].

Regarding confidentiality, it is generally ensured with security mechanisms such as access control and encryption. However, the underlying attacks can bypass access control and direct access to data, and diversified sources of big data contain more confidential data than can be violated by unauthorized users [15].

4 Big Data Between Quality and Security

4.1 The Conflict Between Quality and Security

In the big data context, data quality and data security are two critical fields that are generally not addressed jointly and are sometimes even opposed to each other. Inherent quality features like data heterogeneity, the huge volume of data produced, etc. make it hard to assess and maintain security. Inversely, implementing security mechanisms such as anonymization or encryption may impact its quality (impoverish the data, make it inaccessible, etc).

More precisely, ensuring data quality requires access to data and this can be blocked by the principles of confidentiality and integrity. On the other hand, preserving security using approaches like encryption or data mirroring - which refers to the process of copying data in real-time as an exact copy to test integrity

- can complicate or blur the assessment of data quality, especially for consistency and accuracy. As a result, quality is an obstacle to security and vice versa.

Hence, the strengthening of data security mechanisms at the expense of data quality processes and the adoption of certain security tolerances to improve data quality are two strategies that require vigilant arbitration [2].

In our recent work [16], we introduced an approach using the PolyOrBAC framework [17] that allows each organization to set its security policy independently. We extended the framework by including the Web Service Agreement Specification [18] to automate the process of negotiating and generating access agreements. The resultant framework allows enabling better data storage, eliminates duplicates, and avoids several complex data quality processes.

4.2 Quality at the Service of Security in Big Data

Several security mechanisms and tools are based on data to make a decision. For example: blocking or authorizing access by an authentication system, generating an alarm by an IDS/IPS, correlating logs by a SIEM, learning phase of an application firewall or a behavior-based IDS, etc. Hence, the more the data used by these mechanisms is reliable and of good quality, the more appropriate the security decisions will be.

Let us take a concreate example. Quality at the service of security can be implemented through approaches like SIEM. It collects log and event data - produced by an organization's applications, security devices, and host systems - such as antivirus events, firewall logs, etc. SIEM is only as effective as the information put in it-the adage "Garbage In, Garbage Out" pertains in this situation [19]; When high-quality log data are fed into a SIEM system, we get high-quality security insights about the network. These insights can help strengthen network security protocols. Moreover, regarding intrusion detection, systems mainly aim to classify observed traffic as either legitimate or malicious. Machine learning approaches are appropriate to handle such problems, they learn from data samples to categorize or uncover patterns in the data. The systems rely highly on the quality of the knowledge base which means that it is necessary to learn from high-quality data.

4.3 Security at the Service of Quality in Big Data

As mentioned above, security is defined by three characteristics: confidentiality, integrity, and availability. However, these features are included in the ISO/IEC 25012 standard [20] as quality dimensions. As a result, strengthening the security characteristics leads to better quality.

Regarding integrity, it is a common property between quality and security. In data security, integrity refers to the protection of data against unauthorized changes, while in data quality, it is attached to consistency, accuracy and, completeness [5]. Thus, improving integrity in security help enhance quality. Moreover, when multiple data sources provide the same types of information, the most reliable ones are used [21]. Reputation and trust of data sources are two

properties managed by security. These two properties are widely used to assess the quality of data [16].

Implementing a secure environment can help improve and maintain Big Data quality. In fact, access control management can help control authorized users' access to data and, as a result, optimize data quality. On this point, in [16], we proposed an approach that consists of controlling access in a collaborative environment; such a system is used as a distribution platform that enhances quality by identifying and removing duplicates.

5 An AI-Based Approach to Manage Quality and Security of Data in the Big Data Context

In this section we propose using the Artificial Intelligence (AI) to enforce quality in the service of security and vice versa. AI is the imitation of human intelligence, its goal is to teach computers to think and behave like humans. AI underpins all computer learning and is the future of complicated decision-making. Machine Learning (ML) goes further; this branch of AI allows machines to learn on their own without relying on commands. To understand a data set in a ML system, we usually group instances initially. This process is called Clustering. A clustering algorithm consists in separating data according to their properties or functionalities and gathering them in different clusters according to their similarities. Moreover, In ML, a typical task is the study and construction of algorithms that can learn from and predict data [22]. Three datasets, in particular, are often used at different phases of model development: training, validation, and test sets. A training dataset is the data sample used to fit a model, it is the initial data fed into ML algorithms to teach them how to predict. The validation dataset represents the data used to provide an unbiased evaluation of a model adjusted on the training dataset while tuning model hyperparameters. The test dataset is the sample of data used to evaluate the final model fit on the training dataset.

High-quality training data is mandatory to build a high-performing machine learning model. The quality of the data provided in any machine learning project will definitely have a huge effect on its success; a model that learns from high-quality training data will produce correct and reliable outputs, unlike one that is based on low-quality data.

Besides, ML can be used in security in a variety of ways, including log analysis, prediction, and clustering security events. For example, for analyzing logs (generated from computers, networks, firewalls, applications servers, and other IT systems), a useful prediction might be to classify whether a specific log event or set of events is causing a serious incident that needs attention. Another helpful prediction would be to identify an event that helps to explain the fundamental cause of a problem. However, there are some challenges such as log volumes that are instantly growing and logs that tend to be noisy and unstructured. All of these challenges can lead to uncertain and inaccurate results. In terms of security, since it is a very sensitive field that cannot allow errors, it is necessary to employ high-quality data that will be fed into AI algorithms. The proposed approach includes two major steps (see Fig. 2):

– The first step in the data quality phase are defining the thresholds and necessary dimensions that best characterize our needs (Completeness, Accuracy, Precision, Consistency, and so on), as well as the metrics to quantify them, that will later be employed in the quality evaluation process. Based on the established threshold, we will be able to decide whether the quality level attained is acceptable or not. Next, the data collection process will start; Structured, semi-structured, and unstructured data are collected from audit, and traceability processes of an information system, and stored in an appropriate database. The next step concerns the evaluation of the collected data, a data profiling will be carried out, and based on its results, we will start improving the quality of our data. Our Big Data environment is therefore made up of audit and traceability data produced by a given information system; generally, these data are too large, heterogeneous, and generated at high speed. This process's output consists of prepared data that will serve as an input to the next phase.
– The second part concerns Data Security, we want to implement a set of ML algorithms to classify any new event entering the system according to its security threat level. This phase starts with Feature Selection; it is the process of selecting a subset of pertinent variables for use in model construction. Removing irrelevant data increases learning accuracy, decreases computation time, and enhances understanding of the learning model or data. Once pertinent features are determined, we will create and evaluate our model.

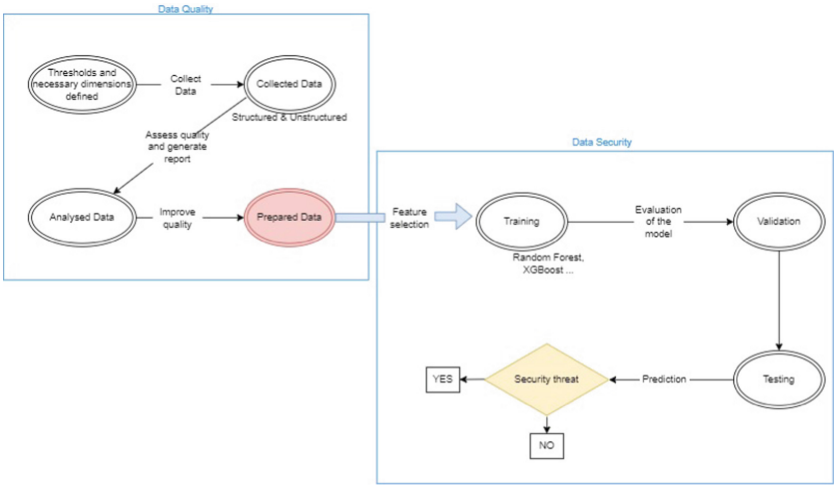


Fig. 2. AI-based proposed approach

5.1 Data Preprocessing

To provide appropriate input data to our models, it is necessary to conduct a serie of preprocessing steps in order to guarantee a good accuracy. We chose to test ML models on the UNSW-NB15 dataset. This dataset was created by the cyber-security research group at the Australian Center of Cyber Security (ACCS). It is a a combination of real modern regular activities and synthetic contemporary attack behaviors. They extracted 100 Gigabytes of data representing 2,540,044 records with different types of attacks, namely, Fuzzers, Analysis, Backdoors, DOS, Exploits, Generic, Reconnaissance, Shellcode, and worms. Each observation is described by 49 features, including the label (normal or attack). The size of the normal traffic represents 87,34 % of the dataset whereas the malicious traffic represents 12,64 %. The preprocessing phase is summarized below:

- Merging data files: the data came in 4 separated files.
- Adding the header.
- Verification and correction of data types: some variables were of the wrong type.
- Handling missing values.
- Removal of duplicates: the dataset contained 480.591 duplicated rows.
- Encoding categorical features: most of ML algorithms require variables to be numeric.
- Feature Scaling: Normalization of the data is necessary to avoid algorithmic inefficiencies.
- Feature Selection: to avoid overfitting, improve accuracy and reduce computing time. we used 'Featurewiz' python library. It uses Searching for Uncorrelated List Of Variables (SULOV) and eXtreme Gradient BOOSTing (XGBoost). This method selected 17 relevant features (see Table 1).

5.2 Machine Learning Algorithms Used

Once the preprocessing phase is done, we created and evaluated our models:

- eXtreme Gradient Boosting (XGBoost) which is based on sequential ensemble learning and decision trees, it combines the results of a set of simpler and weaker models to provide a better prediction. It includes a large number of hyperparameters that can be modified and tuned for better results. The method seeks to optimize a weak classifier (f_0) from which it builds a new classifier (f_1) by introducing a residual:

$$f_1(x) = f_0(x) + h(x) \tag{7}$$

  By repeating the operation several times, we obtain a complex classifier F:

$$F(x) = \sum_{i=1}^{n} \alpha_i f_i \tag{8}$$

Table 1. Selected features using Featurewiz.

Feature	Description
sttl	Source to destination time to live value
service	http, ftp, smtp, ssh, dns, ftp-data, irc, and -
sport	Source port number.
res_bdy_len	The content size of the data transferred from the server's HTTP service.
ct_srv_dst	No. of connections that contain the same service and source address in 100 connections.
sbytes	Source to destination transaction bytes.
state	The state and its dependent protocol (ACC, CLO, CON, ...)
dbytes	Destination to source transaction bytes.
ackdat	The time between the SYN_ACK and the ACK packets of the TCP.
smeansz	Mean of the flow packet size transmitted by the src.
ct_dst_sport_ltm	No of connections of the same destination address and the source port in 100 connections.
dstip	Destination IP address.
ct_flw_http_mthd	No. of flows that has methods such as Get and Post in http service.
dttl	Destination to source time to live value.
ct_dst_ltm	No. of connections of the same destination address in 100 connections.
stime	Record start time.
attack_cat	The name of each attack category.

– Random Forest is made up of a large number of individual decision trees that work together as an ensemble. Each individual tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model. When building each individual tree, it employs bagging and feature sampling in an attempt to create an uncorrelated forest of trees whose prediction is more precise than that of any individual tree.

5.3 Results and Discussion

These algorithms were developed using Jupyter Notebook on a Windows 10 operating system with 16 GB RAM and an i7 CPU. We applied the algorithms on full dataset and on selected features to see the impact of Feature Selection. To evaluate our models, we used different metrics:

– Accuracy: it describes the model's performance over all classes.
– Precision: it measures the model's accuracy in classifying a sample as positive.

– Recall: it measures the ability to detect positive samples.

Table 2 summarizes the results obtained and presents a comparison between the models when used with full datased and with selected features. The best accuracy is obtained with XGBoost classifier when used with selected features (17 relevant features) with 99,93% and 28,35s for training time, and 0,18s for prediction time. Feature selection not only improves accuracy, but also reduces computing time.

Table 2. Results with Full Dataset and Selected Features

	Full Dataset			Selected features		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
XGBoost	99,89%	99,90%	98,91%	99,93%	99,89%	99,89%
Random Forest	98,75%	98,20%	97,64%	99,80%	98,43%	98,52%

6 Conclusion

With the emergence of big data, data in various industries continues to make tremendous development. How to ensure and maintain big data quality and security is considered as an important issue. The majority of existing works deal with quality and security separately whereas the two subjects are linked. In this paper, we presented different concerns of quality and security in the context of Big Data. We highlighted the challenges of each of them, the existing solutions in literature and we discussed the conflict between quality and security. We emphasized quality in the service of security and vice versa. And Finally we presented our AI-based proposed approach, we conducted a comparison of models when used with full dataset and with selected features. For future work, we will improve the performance of our models, and we will use other algorithms to build a powerful AI-based solution that handle, both, quality and security.

References

1. Lee, I.: Big data: dimensions, evolution, impacts, and challenges. Bus. Horiz. **60**(3), 293–303 (2017)
2. Talha, M., El Kalam, A.A., Elmarzouqi, N.: Big data: trade-off between data quality and data security. Procedia Comput. Sci. **151**, 916–922 (2019)
3. Talha, M., Elmarzouqi, N., Abou El Kalam, A.: Quality and security in big data: challenges as opportunities to build a powerful wrap-up solution. J. Ubiquitous Syst. Pervasive Netw. **12**(1), 09–15 (2019)
4. Bastien, L.: Défis Big Data - Quels sont les principaux challenges - relever dans le cadre d'un project Big Data? LeBigData.fr, 06-Jun-2017

5. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996)
6. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Sci. J. **14**, 2 (2015)
7. General Administration of Quality Supervision, Quality management systems - Fundamentals and vocabulary (GB/T19000-2008/ISO9000:2005). ISO9000 (2008)
8. Arolfo, F., Vaisman, A.: Data quality in a big data context. In: Advances in Databases and Information Systems, pp. 159–172. Springer International Publishing, Cham (2018)
9. Taleb, I., Kassabi, H.T.E., Serhani, M.A., Dssouli, R., Bouhaddioui, C.: Big data quality: a quality dimensions evaluation. In: 2016 International IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld) (2016)
10. Taleb, I., Serhani, M.A., Dssouli, R.: Big data quality assessment model for unstructured data. In: 2018 International Conference on Innovations in Information Technology (IIT) (2018)
11. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM **40**(5), 103–110 (1997)
12. Ghosh, P.(guha): Impact of Data Quality on Big Data management. Dataversity. Accessed 16 Dec 2021
13. PGS Software — Application development, outsourcing offshore software development company, outsourcing. NET, Java, nearshoring. Accessed 14 July 2021
14. Bertino, E., Ferrari, E.: Big data security and privacy. In: Studies in big data, pp. 425–439. Springer International Publishing, Cham (2018)
15. Fang, W., Wen, X.Z., Zheng, Y., Zhou, M.: A survey of big data security and privacy preserving. IETE Tech. Rev. **34**(5), 544–560 (2017)
16. Talha, M., Abou El Kalam, A.: Big data: towards a collaborative security system at the service of data quality. In: Hybrid Intelligent Systems, pp. 595–606. Springer International Publishing, Cham (2022)
17. Abou El Kalam, A., Deswarte, Y., Baina, A., Kaaniche, M.: Access control for collaborative systems: a web services based approach. In: IEEE International Conference on Web Services (ICWS 2007) (2007)
18. Andrieux, A. et al.: Web services agreement specification (WS-Agreement). Global Grid Forum 2, 2004
19. cybersecurity.att.com (2020)
20. ISO/IEC 25012:2008: Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model (2009)
21. Talha, M., Elmarzouqi, N., Abou El Kalam, A.: Towards a powerful solution for data accuracy assessment in the big data context. Int. J. Adv. Comput. Sci. Appl. **11**(2) (2020)
22. Ron, D.: Mach. Learn. **30**(1), 5–6 (1998)

Learning Discriminative Representations for Malware Family Classification

Ayman El Aassal(✉) and Shou-Hsuan Stephen Huang

Computer Science Department, University of Houston, Houston, TX 77204, USA
aelaassal@uh.edu, shuang@cs.uh.edu

Abstract. The increasing number of data breaches and cyberattacks in recent years highlights the importance of malware detection research. It is common for malware to have many mutations to evade signature-based detection software. The proliferation of malware variants makes malware detection more challenging. Thus, an essential aspect of the detection process is accurately classifying malware into families of similar variants. Classifying malware into families will result in faster detection and more efficient handling of malware. This research uses a dynamic approach for family classification by analysing malware behaviour. We use a graph to model the run-time behaviour of malware where the edges represent the transition of system calls. With the graph, we then use representation learning methods to extract different families' latent characteristics automatically. This approach is evaluated for malware family classification using an existing malware dataset.

Keywords: Malware Classification · Graph Modeling · Representation Learning · Machine Learning

1 Introduction

An increased interest in malware detection has emerged in recent years due to the rising number of malware attacks and the constant surge of new malware variants. Many types of malware, such as Viruses, Adware, Trojans, Ransomware, Bots, and Worms, are spreading worldwide. It is challenging to defend against cyber-attacks when millions of new malware attacks are discovered and reported monthly [1]. This problem has motivated researchers to leverage advances in machine learning and neural networks to develop large-scale automated frameworks for malware classification and detection [2]. This paper addresses the first step in the detection process: classifying malware into families to simplify the detection.

Malware analysis and classification play a significant role in any detection framework, and its goal is to provide insights into malware behaviour in targeted systems [2]. Malware can be categorized into families based on their malicious payload, propagation method, and run-time behaviour. This grouping allows for a more efficient detection process since there is no need to run an in-depth analysis on malware variants belonging to a known family, costing time and effort.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1327–1336, 2023.
https://doi.org/10.1007/978-3-031-27409-1_121

Malware may be analysed statically or dynamically, and a significant amount of research has been done on both types of analysis over the years. In static analysis, security researchers and analysts inspect malware samples without executing them in a controlled environment. They may use analysis tools and disassemblers to investigate and extract features from different sections of the Portable Executable (PE) header and its assembly code [3, 4]. However, hackers may hide or obfuscate the malicious payload of their malware by using methods like packing, polymorphism, and metamorphism. These obfuscation methods may disrupt the static analysis process and generate a signature different from the original malware [5–7].

In dynamic analysis, researchers inspect malware by executing them in a controlled and isolated environment called sandboxes, in which they use monitoring tools to analyse the behaviour of malware [8]. This behaviour includes the native function calls executed by the malware, the files created or edited, the communications with remote servers over the network, etc. The advantage of dynamic analysis is its robustness against the previously mentioned obfuscation methods because it requires running the malware samples in a sandbox. This approach allows an investigated sample to unpack its code and execute any payload as if running in a target system.

This research focuses on malware dynamic behavioural analysis using a graph model. We analyse the run-time execution logs of malware samples collected from the Malrec dataset [9] and generate their behaviour graphs. We then apply graph representation learning methods to extract discriminative feature vector representations from these graphs. We then feed these feature vectors to a machine learning model to classify malware into families of similar variants. We demonstrate that this approach has promising results in this field.

To summarize the main objectives of this research:

- We propose a new approach for malware family classification using graph-based dynamic behaviour analysis, which does not require manual feature engineering.
- We use a novel approach to evaluate the performance of graph representation learning for malware family classification.
- The application of representation learning on graphs for dynamic malware analysis is novel in the research literature. This technique may also be used to evaluate other software behaviours.

This paper is organized as follows: Sect. 2 reviews existing malware classification and detection research. Section 3 defines malware behaviour graphs and describes our approach to extracting feature vectors and classifying malware into different families. Section 4 shows the results of the methods introduced in this project and compares them with recent endeavours in the literature on dynamic malware analysis. Finally, Sect. 5 summarizes this research and discusses future work.

2 Related Work

2.1 Static Analysis

There is a growing body of research on static malware analysis mixed with machine learning for malware family classification. Several studies analysed malware binaries

to extract features that can be used for differentiating between malware families. For instance, Yuan [10] proposed a byte-level malware classification method based on deep learning in which they transform malware binaries into Markov images and then classify them into families. They tested their model using 10-fold cross-validation on the Microsoft Kaggle [11] and Drebin [12] datasets and achieved 99.26% and 97.36% accuracy, respectively. Furthermore, Verma [13] introduced a cost and time-effective binary texture analysis. Their method achieved fast feature extraction and classification with high performance. It is also robust towards code obfuscation and class imbalance. However, it is not effective in detecting unknown malware.

According to Aslan [14], traditional machine learning may not be enough to detect new malware variants due to the concealing techniques used by hackers. To solve this problem, they proposed a hybrid deep learning architecture based on transfer learning of two wide-ranging pre-trained networks in an optimized way. The hybrid model was evaluated on multiple datasets and achieved 97.78% accuracy on the MalImg dataset.

2.2 Dynamic Analysis

Several studies have explored dynamic analysis to detect new malware samples. The primary approach consists in analysing the execution of malware samples to extract discriminative features that may be used to train a machine learning or deep learning model. For instance, Ding [15] generated system call dependency graphs from malware samples to build Maximum Common Subgraphs (MCSG) representing different malware families. Then they used a matching algorithm that compares the MCSGs of malware families and a sample's dependency graph. Their model was evaluated on ~1200 malware samples with six families and benign samples that run on Windows XP.

Existing literature on dynamic malware detection also focuses on cyber threats to Android phones. Specifically, Zhou [16] collected the run-time system calls of Android applications and extracted feature vectors from this data through statistical methods. They propose a new machine learning model that uses the Monte Carlo algorithm to adjust its weights and reach convergence. Furthermore, Alzaylaee [17] analysed the execution of Android malware samples on a set of real phones with different versions of the Android OS. The authors used stateful input generation to enhance code coverage and extracted 420 static and dynamic features to train their deep learning model. Their framework achieved up to 97.8% accuracy using dynamic features.

Our research introduces a new method for dynamic malware analysis based on graph classification using representation learning. This method does not require manual feature engineering and achieves results comparable to state-of-the-art. We train and test our models using the Malrec dataset [9], collected while taking measures to prevent anti-monitoring methods. We describe our approach in more detail in the following section.

3 Methodology

3.1 Program Behaviour

We define the behaviour of a program as the list of operations it performs when running on a machine. These operations include but are not limited to: allocating memory

and creating files, mapping open ports or programs, pinging IP addresses, sending and receiving files, editing Windows registries, etc. The operating system provides native API functions that allow programs to execute these operations. These low-level functions, or system calls, enable a program to access hardware resources using kernel-level privileges, Fig. 1. In this research, we concentrate on the behaviour of system calls.

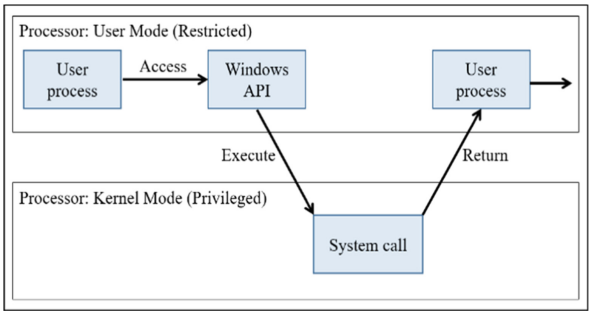


Fig. 1. Illustration of a system call execution

3.2 Graph Modelling of Program Behaviour

Several malware datasets include a log of system call executions for each malware in the set. Each system function includes some parameters, typically in hex format, and the parent process of the instance. In this study, we choose to use the system calls logs from the Malrec dataset [9] as the behaviour of the malware. We try to capture the pair of calls that share one or more parameters within a short "distance." The dependency among the system calls can be viewed as a dynamic signature of a program. A formal definition of a graph capturing the behaviour of the system calls is provided below.

Definition 1 Let $V = \{v_1, v_2,\ldots, v_S\}$ be all the system functions of a given operating system where S is an integer. A **log** of system calls generated by a program is a sequence $(c_1, c_2,\ldots, c_M)$ where M is an integer and C_i is a call of function v_j with appropriate arguments for some $j = 1, 2, \ldots, S$.

Note that we use a pair of curly brackets to represent a set and parentheses for a sequence. The sequence is of length M, typically larger than S, the total number of the system functions. To simplify the notation, we use v_i for both the function and a call to the function.

Definition 2 Let $C = (c_1, c_2,\ldots, c_M)$ be a log of system calls of length M; two function calls c_i and c_j, are said to be **dependent** on each other if they share at least one non-null function argument.

Note that the elements in sequence C may not be unique, i.e., a function call may appear in a log more than once. For example, given a set of four API functions $\{v_1, v_2, v_3, v_4\}$, we can have a log of six system calls $C = \{c_1, c_2, c_3, c_4, c_5, c_6\}$ such as: $(c_1, c_2, c_3, c_4, c_5, c_6) = (v_1, v_2, v_3, v_4, v_1, v_2)$. If function calls c_4 and c_5 share one non-null argument, they are considered dependent.

Definition 3 Let $C = (c_1, c_2, \ldots, c_M)$ be a log of system calls of length M, and window W be a positive integer number, a **behaviour graph** G of window W of C is defined as (V, E) where vertices V is defined as the set of all unique system calls in log C and the set of edges E is defined as:

$$\left\{(c_k, c_l) = \left(v_i, v_j\right) \middle| \, v_i \text{ and } v_j \text{ are dependent}, 1 < |k - l| \leq W, \textit{and } v_i, \; v_j \in V\right\}$$

In the definition of the behaviour graph, we include only system call pairs within a distance of W, a small window size. Two dependent calls that are far apart are less likely to be related. The size of the window W was chosen empirically to 10 in our experiments. The small window size also makes the algorithm run more efficiently.

3.3 Graph Representation Learning Methods

While graphs can capture the behaviour of the system calls, it is difficult to analyse the behaviour in graphs. A typical solution is to embed a graph into a multidimensional space. Representation learning approaches learn a mapping function that embeds nodes or (sub)graph structures as points into a low-dimensional space. The goal is to optimize the mapping function so that the geometric relationship between the embeddings reflects the graph structure [18]. These embeddings can be used as feature vectors for downstream classification tasks. This representation allows for efficient, automatic, and task-independent feature learning, and we will demonstrate that it achieves high accuracy in classifying malware behaviour graphs into families of similar variants. To that end, we implement the following graph embedding methods proposed in the literature, namely Graph2vec [19] and GL2vec [20].

Graph embedding with Graph2vec algorithm approach: This method aims to generate a vector representing a graph based on its rooted subgraphs. It is inspired by the popular Word2vec [21] and Doc2vec [22] algorithms based on the skip-gram model. They are used in Natural Language Processing to embed documents and words into feature vectors. In Graph2vec, the documents are graphs, and words are its rooted subgraphs. Given a node n of a graph G and an integer D, a **rooted subgraph** of degree D with node n as root is a subgraph that contains all the nodes that are reachable with a path of length D or less from the root N [19].

The Graph2vec approach starts by extracting the rooted subgraphs from all the graphs in the dataset using the Weisfeiler-Lehman (WL) method, also called the colour refinement method [23]. This method recursively generates a label encoding for each node in the graph based on the encoding of its neighbours. Then the Skip-Gram model is applied to these subgraphs to automatically generate the feature vectors (also called embeddings) of each behaviour graph in the dataset.

In this research, we implement Graph2vec and apply it to our dataset of malware behaviour graphs to automatically generate embeddings (feature vectors) that capture the dependencies between system calls and represent them as feature vectors. These vectors are then used to train a machine learning classifier, namely Random Forest and SVM, to classify malware into families of similar variants.

Chen [20] mention the limitations of Graph2vec. They argue that this approach only considers node labels and ignores edge information since the WL method generates

rooted subgraphs based only on the node labels. They also add that it fails to capture similarities in the local structure of the nodes in a graph. We describe in what follows the GL2vec approach and how it answers these limitations. We also encourage the readers to check the provided references for more details about these algorithms.

Graph embedding with GL2vec algorithm approach: Chen [20] used the concept of a line graph in conjunction with the original graph to solve the limitations of the Graph2vec method.

Given a graph $G = (V, E)$, its line graph $L(G) = (LV, LE)$ represents the adjacency of the nodes in graph G. In other words, the edges are represented as nodes and are connected by an edge if they share a common endpoint. The L(G) node labels carry the corresponding edge information from the original graph G, which could be edge labels, features, or weights.

The authors proposed to apply Graph2vec on both graph G and its line graph L(G) and then concatenate their embeddings to create one final feature vector. The intuition behind this approach is to capture the edge information and structural properties that were not represented in the embedding of graph G alone. Formally GL2vec of graph G is defined as follows:

$$GL2vec(G) = Graph2vec(\mathrm{L}(G)) + Graph2vec(G)$$

In the context of malware family classification, we apply this approach to generate feature vectors representing the structural properties of malware behaviour graphs. We then use these vectors to train two machine learning classifiers and compare their performance against the same classifiers trained with embeddings generated using Graph2vec.

4 Performance Evaluation

4.1 The Dataset

The lack of new and available malware datasets hinders the malware detection literature, especially for dynamic analysis. Researchers often find it challenging to collect many malware samples, so they often download a repository of samples from sources like VXheaven and Virusshare. However, these malware collections may be outdated and target older operating systems such as Windows XP [24]. Furthermore, previous studies may select an arbitrary security provider like Avast, Norton, or Kaspersky to get the ground truth for their dataset. However, the detection rate and assigned labels usually vary between antivirus software. To illustrate this point, we use VirusTotal.com, which returns the analysis results for over 80 antivirus software for any given file. We analyse the VirusTotal reports of more than 66k malware samples from the Malrec dataset [9], and the results show that 67% of samples were detected by less than 30 antivirus software. In addition, there is no consensus on the assigned family for 35% of malware samples. These issues may complicate the comparison of different approaches proposed in the malware detection literature. Thus, we suggest using majority voting to collect the ground truth labels of malware samples.

From the Malrec dataset, we collected 7000 malware samples from 14 different families (500 malware per family). We confirmed their assigned families by majority voting between the antivirus software correctly identifying the samples as malicious. We split this data into 80% training and 20% testing to evaluate the methods described in Sect. 3.

4.2 The Results

We summarize the results achieved using our approach in Table 1. We show the performance of two popular machine learning classifiers, Random Forest (RF) and SVM, for malware family classification when trained with feature vectors generated by the methods described in Sect. 3.

Table 1. Results achieved by RF and SVM when trained with different feature vectors

Method	Classifier	Accuracy (%)	F1 (%)	AUC (%)
Graph2Vec	SVM	96.64	96.65	99.75
	RF	94.43	94.54	99.82
GL2Vec	SVM	98.86	98.88	99.93
	RF	99.50	99.52	99.99

We can see from these tables that Random Forest achieves better results than SVM in most cases when trained on the same type of features. They show that Random Forest performs the best when trained with GL2vec embeddings and reaches 99.50% accuracy, followed by Graph2vec embeddings with 94.43% accuracy. The SVM classifier follows the same pattern with 98.86% and 96.64% accuracy when trained with GL2vec and Graph2vec embeddings, respectively. The difference in results between these two approaches confirms that GL2vec is better at capturing the structural properties of malware behaviour graphs. Figure 2 further illustrates this point by showing the confusion matrix generated by this approach. The classifier achieved a high detection rate of at least 98% for each of the 14 malware families in our dataset.

4.3 Comparison of Malware Dynamic Analysis Results

At the beginning of this section, we describe how the lack of available datasets hinders the malware detection literature. The absence of benchmark datasets is especially true in the case of dynamic analysis, whether for Windows, Linux, or Android malware. Researchers usually compare their approach using the MalImg [25] and the Microsoft malware classification challenge [11] datasets for static analysis. However, they do not contain information about the run-time execution of malware, which is needed for dynamic analysis. This issue complicates the comparison of dynamic analysis methods proposed in the literature. Table 2 illustrates this point, showing a few recent research

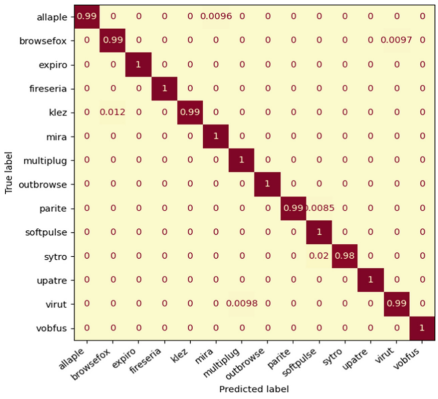


Fig. 2. Multiclass Confusion Matrix of Random Forest trained with GL2vec embeddings

Table 2. Recent results in malware dynamic analysis research

Method	Dataset	Platform	Features	Acc. (%)	F1 (%)
Di [26]	VxHeaven	Windows	Dynamic	97.73	97.19
Ding [15]	Malware.lu	Windows	Dynamic	96.40	–
Darabian [27]	VirusTotal	Windows	Static & Dynamic	99.00	97.00
Zhou [16]	Virusshare, Baidu	Android	Dynamic	97.70	98.48
Current Approach	Malrec	Windows	Dynamic	99.50	99.52

endeavours conducted in this field, the datasets used, and their results for malware classification and detection.

It is challenging to compare the research referenced in Table 4. However, we can conclude that the approach introduced in this paper meets the standards of the current state of research in dynamic malware analysis in terms of performance.

5 Conclusion

This research focused on the classification of malware into families of similar variants. We modelled the run time execution log of malware into behaviour graphs and introduced different approaches to extract feature vectors for the downstream classification task. The results of this work demonstrated that graph representation learning methods could generate expressive feature vectors that lead to high performance by the machine learning classifier.

We evaluated two graph embedding approaches and used two popular machine learning classifiers to compare the classification results. We were pleasantly surprised by the outcome of the two embedding approaches, especially GL2vec, since the SVM and RF models achieved an accuracy of 98.86% and 99.50%, respectively. We also compared our

approach with other works in the malware dynamic analysis literature. We showed that the methods proposed in this paper achieve results comparable with the state-of-the-art.

The application of graph representation learning methods to malware classification and detection is in the early stage of the research. However, the current results are promising, and we hope to explore this field further. Once we can classify malware into families, we can then test if a program belongs to a malware family.

Acknowledgment. This work was supported by the National Science Foundation (1950297, 1433817); the U. S. Department of Education (P200A210119); and the National Security Agency (H98230-22-1-0323).

References

1. AV-Test: The Independent IT-Security Institute (2022) Malware Statistics & Trends Report. https://www.av-test.org/en/statistics/malware/. Accessed 29 Oct 2022
2. Gibert, D., Mateu, C., Planes, J.: The rise of machine learning for detection and classification of malware: research developments, trends and challenges. J. Netw. Comput. Appl. **153**, 102526 (2020)
3. Solis, D., Vicens, R.: Convolutional neural networks for classification of malware assembly code. In: Recent Advances in Artificial Intelligence Research and Development: Proceedings of the 20th International Conference of the Catalan Association for Artificial Intelligence, Deltebre, Terres de L'Ebre, Spain. p. 221 (2017)
4. Kinable, J., Kostakis, O.: Malware classification based on call graph clustering. J. Comput. Virol. **7**, 233–245 (2011)
5. Hai, N.M., Ogawa, M., Tho, Q.T.: Packer identification based on metadata signature. In: Proceedings of the 7th Software Security, Protection, and Reverse Engineering/Software Security and Protection Workshop, pp. 1–11 (2017)
6. Ucci, D., Aniello, L., Baldoni, R.: Survey of machine learning techniques for malware analysis. Comput, Secur. **81**, 123–147 (2019)
7. Euh, S., Lee, H., Kim, D., Hwang, D.: Comparative analysis of low-dimensional features and tree-based ensembles for malware detection systems. IEEE Access **8**, 76796–76808 (2020)
8. Singh, J., Singh, J.: A survey on machine learning-based malware detection in executable files. J. Syst. Architect. **112**, 101861 (2021)
9. Severi, G., Leek, T., Dolan-Gavitt, B.: Malrec: compact full-trace malware recording for retrospective deep analysis. In: International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 3–23 (2018)
10. Yuan, B., Wang, J., Liu, D., et al.: Byte-level malware classification based on markov images and deep learning. Comput. Secur. **92**, 101740 (2020)
11. Ronen, R., Radu, M., Feuerstein, C., et al.: Microsoft malware classification challenge (2018). arXiv Prepr. arXiv180210135
12. Arp, D., Spreitzenbarth, M., Hubner, M., et al.: Drebin: Effective and explainable detection of android malware in your pocket. In: NDSS. pp 23–26 (2014)
13. Verma, V., Muttoo, S.K., Singh, V.B.: Multiclass malware classification via first-and second-order texture statistics. Comput. Secur. **97**, 101895 (2020)
14. Aslan, Ö., Yilmaz, A.A.: A new malware classification framework based on deep learning algorithms. IEEE Access **9**, 87936–87951 (2021)
15. Ding, Y., Xia, X., Chen, S., Li, Y.: A malware detection method based on family behavior graph. Comput. Secur. **73**, 73–86 (2018)

16. Zhou, Q., Feng, F., Shen, Z., Zhou, R., Hsieh, M.-Y., Li, K.-C.: A novel approach for mobile malware classification and detection in Android systems. Multim. Tools Appl. **78**(3), 3529–3552 (2018). https://doi.org/10.1007/s11042-018-6498-z
17. Alzaylaee, M.K., Yerima, S.Y., Sezer, S.: DL-Droid: deep learning based android malware detection using real devices. Comput. Secur. **89**, 101663 (2020)
18. Hamilton, W.L., Ying, R., Leskovec, J.: Representation learning on graphs: methods and applications (2017). arXiv Prepr. arXiv170905584
19. Narayanan, A., Chandramohan, M., Venkatesan, R., et al.: Graph2vec: Learning distributed representations of graphs (2017). arXiv Prepr. arXiv170705005
20. Chen, H., Koga, H.: Gl2vec: Graph embedding enriched by line graphs with edge features. In: International Conference on Neural Information Processing, pp. 3–14 (2019)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). arXiv Prepr. arXiv13013781
22. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp 1188–1196 (2014)
23. Rieck, B., Bock, C., Borgwardt, K.: A persistent weisfeiler-lehman procedure for graph classification. In: International Conference on Machine Learning. pp 5448–5458 (2019)
24. Salehi, Z., Sami, A., Ghiasi, M.: MAAR: Robust features to detect malicious activity based on API calls, their arguments and return values. Eng. Appl. Artif. Intell. **59**, 93–102 (2017)
25. Karthikeyan, L., Jacob, G., Manjunath, B.: Malware images: visualization and automatic classification. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security, p. 4 (2011)
26. Xue, D., Li, J., Lv, T., et al.: Malware classification using probability scoring and machine learning. IEEE Access **7**, 91641–91656 (2019)
27. Darabian, H., Homayounoot, S., Dehghantanha, A., et al.: Detecting cryptomining malware: a deep learning approach for static and dynamic analysis. J. Grid Comput. **18**, 293–303 (2020)



Host-Based Intrusion Detection: A Behavioral Approach Using Graph Model

Zechun Cao[1(✉)] and Shou-Hsuan Stephen Huang[2]

[1] Texas A&M University-San Antonio, San Antonio, TX 78224, USA
zcao@tamusa.edu
[2] University of Houston, Houston, TX 77204, USA
shuang@cs.uh.edu

Abstract. Data breach incidents are becoming a global threat, costing companies millions with each attack. Most host-based intrusion approaches analyze system logs, such as the system call log, to monitor host network traffic. However, the system call log does not capture some critical behavior of the user. This research explores host-intrusion detection based on the user file-accessing log, which provides an additional dimension of user behavior for data breaching. This paper hypothesizes that an intruder behaves inside a host computer differently from "normal" users. We propose to use a graph model to model file-accessing behavior at a high level of abstraction. We then derive a set of behavioral features from the graph model for machine learning algorithms to identify the intruders. We validated our hypothesis with an existing user activity dataset by adopting an anomaly detection approach. The results based on our approach report an Area Under the Curve (AUC) of the receiver operating characteristic curve value of 0.98, with the one-class Support Vector Machine model trained on only normal users' data.

Keywords: Intrusion detection · Graph model · Machine learning · Behavioral model.

1 Introduction

As our modern society is more than ever dependent on digital technology, critical industries, such as banking, healthcare, energy, and transportation, rely on computers and networks to store crucial data files and facilitate their operations. Meanwhile, the stakes are high for malicious users to evade or penetrate the network intrusion detection and attack the victims' hosts. Once the adversaries reach the victims' hosts, they often search extensively to locate sensitive data for

This work was partly supported by the National Science Foundation grants (1950297, 1433817); the U.S. Department of Education grant (P200A210119); and the National Security Agency grant (H98230-22-1-0323).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1337–1346, 2023.
https://doi.org/10.1007/978-3-031-27409-1_122

financial gain or political and military reasons. The theft of critical and sensitive information on victims' hosts causes a tremendous loss for companies and poses a significant threat to society. Recently, Equifax disclosed that a massive data breach in 2017 might have impacted 143 million consumers, nearly 44 percent of the U.S. population. In this incident, attackers got their hands on names, social security numbers, birth dates, addresses, driver's license numbers, and about 209,000 credit card numbers, causing the impact of this breach to last for years [15]. In such catastrophic data breaches, the attackers conducted extensive searches on the hosts and successfully located files containing credentials to elevate their rights or permissions [4]. To defend against such sophisticated and determined attackers, we need to develop an effective host-based intrusion detection system (HIDS) to protect the hosts with critical and sensitive data.

Despite the uptick in data breach incidents, most existing HIDSs either can be evaded by variants of known attacks or show an alarming tendency to generate huge volumes of false positives [20]. This paper presents a host-based intrusion detection method based on the anomaly detection approach. We propose using a graph to model a user's behavior to overcome the aforestated challenges in the anomaly detection approach. We hypothesize that intruders exhibit unusual behaviors compared to normal users, which our graph model for intrusion detection can capture. We summarize the main contributions of our work as follows:

- This paper proposes and formally defines a graph model based on activity traces to describe and model a user's file-accessing behavior in a system.
- We propose a method to derive behavioral features from the activity trace and the intruder's behavior deviations.
- We build machine learning models with anomaly detection algorithms to validate our hypothesis on a real-world user activity dataset.

We organize the remainder of this paper as follows. Section 2 summarizes the related work to our research. Section 3 defines the proposed activity trace and graph model with formal definitions and describes the derived behavioral features. Section 4 presents our method's evaluation results with a real-world user activity dataset. Section 5 concludes this paper and discusses how our work can be extended.

2 Related Work

There were two main approaches employed by the intrusion detection techniques in the HIDS: misuse detection and anomaly detection. The misuse-detecting approach defines abnormal behavior and relies on signature-matching algorithms to detect anomalies [20]. It compares audit data on the host, such as operating system calls and user commands, with a database of signatures associated with known attacks. Earlier detection methods on the misuse-based approaches [14, 17] were circumvented by sophisticated attackers targeting a system as they only detect known attacks.

Recent work about HIDS demonstrated that anomaly detection techniques effectively detect host intruders by leveraging various data sources in the system. Verma and Bridges [19] evaluated the distance between entries in the host logs to detect anomalous events in the system. However, their approach has shortcomings in describing normal users with diverse behavioral patterns. In [7], the authors developed statistical features on system call traces to detect attacks with machine learning algorithms. Based on their evaluation, using statistical features to capture dynamic and complex user behaviors did not achieve competitive performance. Similarly, [8] used file access logs to derive statistical features and evaluated the approach using a balanced class dataset as a binary classification task. Nevertheless, their approach relied on using attackers' data in the training process, which is very challenging in practical scenarios as attackers' data is scarce. Researchers also presented various methodologies to model host user behaviors by leveraging various data sources [10,13,16,18]. However, these proposed approaches depended on external resources, such as the host's operating system [13,18], keyboard device [10], and a specific set of system commands [16].

It is worth noting that numerous network intruder detection methods [2,3,5,6,9,11] were proposed in the past with decent performance. They studied features extracted from network connections and adopted machine-learning models to identify anomalous users. However, there was no discussion about an abstract model representing user behavior, which limited their work to only specific applications. Another issue in [3,5] is that applying the final model to time series datasets is challenging, which is problematic for user activity logs with timestamps.

3 Methodology

This section provides formal definitions of our graph model, and the behavioral features derived from the model. Due to space limitations, the algorithms for constructing the trace and graph are not presented here.

3.1 Model Definitions

Our model assumes that each record in the user activity log is associated with a system timestamp. To explain our model's definitions, we first define and use the file access log as an example of the user's activity log. Then, we define the file transition trace built by the file access log. Lastly, we define the directed graph associated with the trace.

Definition 1. (File Access Log) Given an integer $n \in \mathbb{N}$, a file access log $L = (f_1, f_2, \ldots, f_n)$ is a list ordered by the files' timestamps, where $f_i (1 \leq i \leq n)$ is a file identifier in the file system.

Modern file system auditing software can log additional information for each file identifier, so we assume each file identifier f in the file access log can have various attributes. We use $f.attr$ to denote the attribute $attr$ of the file f. For

example, *f.timestamp* is the timestamp of the file f being visited, and *f.duration* indicates how long a user stays working on a file f before moving to the following file in the log. Note that the file access log may contain a consecutive sequence of identical file identifiers if it accesses the file repeatedly. To focus on the user's transitions among files and eliminate repeating file identifiers, we use one entry to represent each consecutive same-file-identifier entries group to compose a trace. Next, we formally define the trace below.

Definition 2. (Trace) Given a file access log $L = (f_1, f_2, \ldots, f_n)$, we define an equivalence relation $f_i \equiv f_{i+1}$ $(1 \leq i < n)$ if f_i's file identifier is identical to f_{i+1}'s file identifier. Based on the equivalence relation, L can be partitioned into equivalence classes C_j sorted according to the timestamps, where $1 \leq j \leq m$. We define a trace $T = (r_1, r_2, \ldots, r_m)$, where $m \leq n$, and r_i $(1 \leq i \leq m)$ is the first element of each equivalence class C_i.

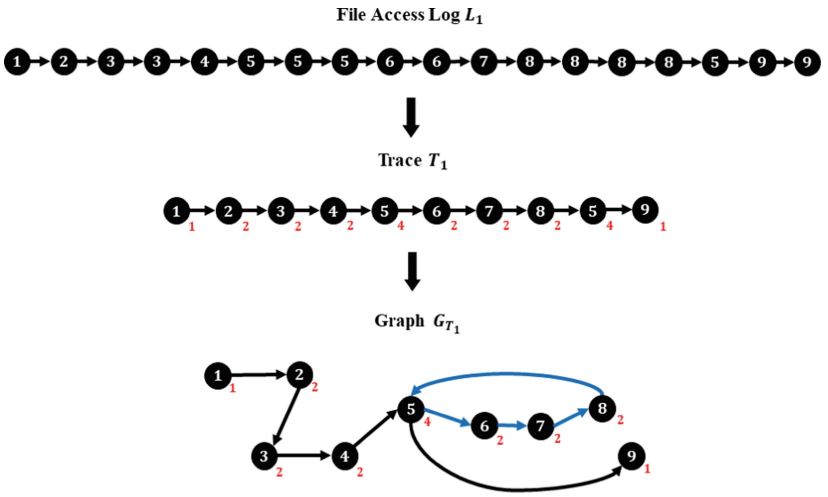


Fig. 1. Most trace elements have a degree (in red) of 2, except the element labeled "5" has a degree of 4 because the user revisits it often. There is only one cycle (colored in blue) in graph G_{T_1}.

As trace T_1 is shown in Fig. 1, each element in T_1 has a previous and next element, except for the first and last elements. Also, two consecutive elements are distinct. Arrows connecting two elements indicate the user's transitions. Except for the first and last elements, each element in the trace connects with two other elements. Note that a trace is a file access log with constraint, so each element r in a trace inherits the attributes from corresponding files in the log. For example, *r.duration* denotes the time duration the user stays on the element r. Determining the degree of an element r in a trace is challenging, as a trace is a sequence of elements with a linear topological structure. However, if we view

a trace as a walk over its elements, we can map its elements and transitions as vertices and edges in a directed graph. Next, we define the graph as a directed graph associated with the trace.

Definition 3. (Graph) Given a trace $T = (r_1, r_2, \ldots, r_m)$, we define its associated graph as a directed graph $G_T = (V, E)$, where $V = \{r_i \mid r_i \in T\}$ is a set of vertices, and $E = \{(u, v) \mid (u, v) = (r_i, r_{i+1}), \forall i, 1 \leq i \leq m-1, r_i \in T\}$ is a set of edges.

We use $\|V\|$ and $\|E\|$ to denote the number of the graph's vertices and edges. Note that a surjective-only mapping function exists that maps all the elements in T to the vertices V in G_T, such that the nodes with identical file identifiers should map to the vertex with the same identifier in the graph. Therefore, there may be fewer vertices in the graph G_T than the nodes in the associated trace T. Given an element r in a trace, and its corresponding vertex v in the associated graph, we use $v.indegree$ to denote the indegree of vertex v, and the v's outdegree as $v.outdegree$. Then, we define the degree of vertex v as $v.degree = v.indegree + v.outdegree$. Since vertex v is mapped from the element r in the trace, we have $r.degree = v.degree$.

Figure 1 shows an example trace, T_1, with its constructed graphs, G_{T_1}. For simplicity, we label the elements and their mapped vertices with integers. We also label the value of the degree attribute next to each element and its corresponding vertex. In trace T_1, the transitions are sequential among mostly unique elements. Therefore, it is evident that most elements have a degree of 2, except for the elements labeled "1", "5" and "9". Also, because of the transition in the trace from element "8" back to "5", there is one cycle formed in G_{T_1}, whose edges are marked in blue.

3.2 Behavioral Features Derived from Graph Model

This section provides details of the five behavioral features extracted from our graph model. We remark on the rationale for designing these features, but space limitation prohibits us from giving the implementation and complexity analysis here.

Number of Vertices (NV) Based on our model's definitions, the number of vertices, $\|V\|$, of a given graph $G_T = (V, E)$, indicates the number of unique files a user traverses in the trace. With the graph model, we now conveniently use the graph property $\|V\|$ to describe the number of unique files users explore in a file system. Despite its simplicity, feature NV is meaningful because a normal user visits fewer unique files than an intruder in a given duration. After all, the intruder has the motivation to explore a large portion of the file system.

Graph Connectivity (GC) The motivation for finding a graph's maximum vertex degree is that a normal user tends to move frequently among related files, whereas an intruder mainly performs linear searches. In a graph built from a trace, the directed edges represent files' transitions in the user's activity log. Thus, we may see vertices with a high vertex degree value in a normal user's graph due to the user's frequent transitions. In contrast, most of the vertices

in an intruder's graph may have a value of vertex degree as 2, as they only connect to their two immediate neighbors. We define a graph's connectivity as the normalized maximum vertex's degree, with the vertex degree as the sum of its indegree and outdegree values.

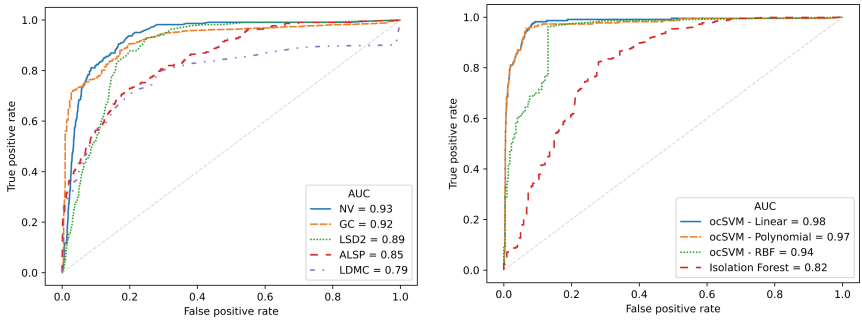
Longest Segment with Degree-2 Nodes (LSD2) In a file access trace, revisiting an element increases the element's degree. An element has a degree of 2 if it is only visited once, except the first and the last. An intruder in a file system often searches broadly without revisiting a file unless it is considered valuable. In contrast, a normal user may work on only a few files for an ongoing task and repeatedly moves among them. The intuition that leads us to study the trace by finding its LSD2 is distinguishing these two file access patterns. The LSD2 denotes the longest segment's length in a trace with only degree-2 elements. As we defined previously, an element r in a trace is a degree-2 element if its corresponding vertex in the graph has indegree and out-degree values summed as 2. For example, in Fig. 1, we have $T1.LSD2 = 4$ because every element in the segment from "6" to "9" has a degree of 2.

Average Length of Shortest Path (ALSP) To understand the intuition of finding the ALSP for a graph, assume that we have a fully-connected graph in which an edge connects each pair of distinct vertices. For each pair of vertices, we can always find the shortest path with a length of 1. Similarly, when a normal user works on a set of related files, the user will likely find the "shortcut" between a pair of files. In contrast, if we find the shortest path between a pair of vertices in an intruder's graph, it is likely to be close to the number of transitions along the associated trace, as most vertices are linearly connected. Thus, the shortest path length for an intruder's graph tends to be higher than for a normal user. Given a graph, we can find the lengths of the shortest paths between all pairs of vertices, and we define ALSP as the average length by the number of pairs of vertices.

Longest Duration Maximal Clique (LDMC) As a normal user, it is common to work on several tasks, and each task may involve multiple related files. While staying on one task, a normal user typically moves among related files and spends time on them before shifting to the next task. Therefore, if a normal user spends time on tasks, we expect to find several maximal cliques in the graph with large duration values. However, as an intruder usually hungrily searches for valuable information, it is unlikely to find a maximal clique with long-duration vertices. Consequently, we believe that a normal user's longest duration stays on a maximal clique and is likely to be larger than an intruder.

4 Experiments and Results

In this section, we validate our hypothesis that the proposed graph model can capture the user's behavior and reveal discernable differences between normal users and attackers. We set up our experiments in a host-based intrusion detection scenario using a real-world file system access dataset, Windows-Users and - Intruder simulations Logs (WUIL) [1]. The experiments are conducted on the



(a) ROC curves of the individual graph-derived feature.

(b) ROC curves of combined graph-derived features with isolation forest, one-class SVM with linear, polynomial and RBF kernels.

Fig. 2. Performance evaluation of individual feature and anomaly detection models with combined features

preprocessed dataset containing 23,050 normal users' and 222 attackers' samples, with five features derived from the graph model built by each sample, namely NV, GC, LSD2, ALSP, and LDMC. It is not uncommon that even if a malicious activity is detected, system administrators may not be willing to share information about the problem [12]. To cope with the scarcity issue of the attackers' data, we evaluate the performance of our method by adopting anomaly detection algorithms and training the detection model with only normal users' data. We use the machine learning library *scikit-learn* in Python for the experiment implementation. All experiments are performed on a Windows machine with a 6-core CPU running at 3.4 GHz, and 32 GB of RAM.

4.1 Behavioral Features and Intrusion Detection Evaluation

We first examine the feature discrimination capacity by adopting each feature as a single-attribute classifier to classify attackers from normal users. We measure TPRs and FPRs at various threshold settings for each classifier and plot the ROC curve in Fig. 2a. The ROC curves are shown and sorted by their AUC values for analysis and comparison. In Fig. 2a, all five features show decent classification performance with an average AUC value of 0.88. This indicates that our proposed graph model can effectively capture different behaviors exhibited by normal users and attackers. Note that the ROC curves of the Number of Vertices (NV) and Graph Connectivity (GC) have the highest AUC values among the five features. Compared to the other features, the NV and GC are graph features that are not dependent on the knowledge and assumptions for a specific task. This result shows us that the proposed graph model may be able to effectively model behaviors in various domains.

Since the distribution of the target class is highly imbalanced in our dataset, we adopt anomaly detection algorithms to evaluate the combined performance of

all five features. Unlike classification algorithms, anomaly detection algorithms are unsupervised and identify attackers as anomalies by training the model with only the normal users' samples. We use two anomaly detection algorithms based on their implementations in the *scikit-learn* library: One-Class Support Vector Machine and Isolation Forest.

We select the models with median AUC values from the group 10-fold cross-validation results and plot the ROC curves in Fig. 2b. We notice that the one-class SVM algorithm performs better than the Isolation Forest with all three kernel functions. The best model, one-class SVM with linear kernel function, shows excellent performance with an AUC value of 0.98, and the TPR reaches 98% with an FPR of 9%. Moreover, our proposed approach achieves higher performance than the current work using statistical features, in which the reported best model has an AUC value of 0.94. It is worth mentioning that, however, our approach leverages all normal users' logs to train a single detection model, which contrasts with the existing masquerader detection research. Additionally, we adopt unsupervised anomaly detection algorithms for the model evaluation, which is different from the supervised classification algorithms in existing work [8].

5 Contribution and Future Work

This paper proposed and formally defined a behavioral graph model to describe user behaviors in a host computer system. We hypothesized that intruders have different cyber behavior patterns than normal users in a system, which our behavioral graph model can recognize to detect adversaries in our computer system. We derived five features from the graph model and validated the model using an anomaly detection algorithm on a file access log dataset to validate our hypothesis. The experiment results support our hypothesis by showing that all five features derived from our graph model are discriminative in distinguishing attackers from normal users. By adopting the anomaly detection algorithm one-class SVM with linear kernel function, we train a model with all the features and achieve a high AUC value of 0.98, TPR of 98% with the 9% FPR from its ROC curve.

The proposed approach can achieve higher performance with a lower dimensional feature space than the existing work relying on statistical features. As future work, we are interested in exploring more task-invariant graph-derived features, such as the Graph Connectivity and the Number of Vertices, to validate our approach in other tasks. Furthermore, the user file access approach presented in this paper may be combined with other host-based intrusion systems based on system call logs to be even more effective in detecting intruders.

References

1. Camiña, J.B., Hernández-Gracidas, C., Monroy, R., Trejo, L.: The Windows-Users and -Intruder simulations Logs dataset(WUIL): an experimental framework for masquerade detection mechanisms. Expert. Syst. Appl. **41**(3), 919–930 (2014). https://doi.org/10.1016/j.eswa.2013.08.022, https://linkinghub.elsevier.com/retrieve/pii/S0957417413006349
2. Cao, Z., Huang, S.H.S.: Detecting intruders and preventing hackers from evasion by tor circuit selection. In: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference On Big Data Science and Engineering (TrustCom/BigDataSE), pp. 475–480. IEEE, New York, NY (2018). https://doi.org/10.1109/TrustCom/BigDataSE.2018.00074, https://ieeexplore.ieee.org/document/8455942/
3. Chitrakar, R., Huang, C.: Selection of candidate support vectors in incremental SVM for network intrusion detection. Comput. Secur. **45**, 231–241 (2014). https://doi.org/10.1016/j.cose.2014.06.006, https://linkinghub.elsevier.com/retrieve/pii/S0167404814000996
4. Government Accountability Office: DATA PROTECTION Actions Taken by Equifax and Federal Agencies in Response to the 2017 Breach. Tech. Rep. GAO Publication No. 18–559, Washington, D.C.: U.S. Government Printing Office (2018)
5. Gu, J., Lu, S.: An effective intrusion detection approach using SVM with naïve Bayes feature embedding. Comput. Secur. **103**, 102158 (2021). https://doi.org/10.1016/j.cose.2020.102158, https://linkinghub.elsevier.com/retrieve/pii/S0167404820304314
6. Gu, J., Wang, L., Wang, H., Wang, S.: A novel approach to intrusion detection using SVM ensemble with feature augmentation. Comput. Secur. **86**, 53–62 (2019). https://doi.org/10.1016/j.cose.2019.05.022, https://linkinghub.elsevier.com/retrieve/pii/S0167404819301154
7. Haider, W., Hu, J., Xie, M.: Towards reliable data feature retrieval and decision engine in host-based anomaly detection systems. In: 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), pp. 513–517. IEEE, Auckland, New Zealand (2015). https://doi.org/10.1109/ICIEA.2015.7334166, http://ieeexplore.ieee.org/document/7334166/
8. Huang, S.H.S., Cao, Z., Raines, C.E., Yang, M.N., Simon, C.: Detecting intruders by user file access patterns. In: Liu, J.K., Huang, X. (eds.) Network and System Security. Lecture Notes in Computer Science, vol. 11928, pp. 320–335. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-36938-5_19, http://link.springer.com/10.1007/978-3-030-36938-5_19
9. Huang, S.H.S., Cao, Z.: Detecting malicious users behind circuit-based anonymity networks. IEEE Access **8**, 208610–208622 (2020). https://doi.org/10.1109/ACCESS.2020.3038141, https://ieeexplore.ieee.org/document/9258912/
10. Killourhy, K., Maxion, R.: Why did my detector do that? In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Jha, S., Sommer, R., Kreibich, C. (eds.) Recent Advances in Intrusion Detection. Lecture Notes in Computer Science, vol. 6307, pp. 256–276. Springer Berlin Heidelberg, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15512-3_14, http://link.springer.com/10.1007/978-3-642-15512-3_14

11. Kim, G., Lee, S., Kim, S.: A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert. Syst. Appl. **41**(4), 1690–1700 (2014). https://doi.org/10.1016/j.eswa.2013.08.066, https://linkinghub.elsevier.com/retrieve/pii/S0957417413006878
12. Liu, M., Xue, Z., Xu, X., Zhong, C., Chen, J.: Host-based intrusion detection system with system calls: review and future trends. ACM Comput. Surv. **51**(5), 1–36 (2019). https://doi.org/10.1145/3214304, https://dl.acm.org/doi/10.1145/3214304
13. Maxion, R., Townsend, T.: Masquerade detection using truncated command lines. In: Proceedings International Conference on Dependable Systems and Networks. pp. 219–228. IEEE Computer Society, Washington, DC, USA (2002). https://doi.org/10.1109/DSN.2002.1028903, http://ieeexplore.ieee.org/document/1028903/
14. Mishra, A., Nadkarni, K., Patcha, A.: Intrusion detection in wireless ad hoc networks. IEEE Wireless Commun. **11**(1), 48–60 (2004). https://doi.org/10.1109/MWC.2004.1269717, http://ieeexplore.ieee.org/document/1269717/
15. Newman, L.: How to protect yourself from that massive Equifax breach. Wired (2017). https://www.wired.com/story/how-to-protect-yourself-from-that-massive-equifax-breach/
16. Salem, M.B., Stolfo, S.J.: Modeling User Search Behavior for Masquerade Detection. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) Recent Advances in Intrusion Detection. Lecture Notes in Computer Science, vol. 6961, pp. 181–200. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-23644-0_10, http://link.springer.com/10.1007/978-3-642-23644-0_10
17. Schonlau, M., DuMouchel, W., Ju, W.H., Karr, A.F., Theus, M., Vardi, Y.: Computer intrusion: detecting masquerades. Stat. Sci. **16**(1), 58–74 (2001)http://www.jstor.org/stable/2676780, publisher: Institute of Mathematical Statistics
18. Schonlau, M., Theus, M.: Detecting masquerades in intrusion detection based on unpopular commands. Inf. Process. Lett. **76**(1–2), 33–38 (2000). https://doi.org/10.1016/S0020-0190(00)00122-8, https://linkinghub.elsevier.com/retrieve/pii/S0020019000001228
19. Verma, M.E., Bridges, R.A.: Defining a metric space of host logs and operational use cases. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 5068–5077. IEEE, Seattle, WA, USA (2018). https://doi.org/10.1109/BigData.2018.8622083, https://ieeexplore.ieee.org/document/8622083/
20. Zanero, S.: Behavioral Intrusion Detection. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Aykanat, C., Dayar, T., Körpeoğlu, İ.(eds.) Computer and Information Sciences - ISCIS 2004. Lecture Notes in Computer Science, vol. 3280, pp. 657–666. Springer Berlin Heidelberg, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30182-0_66, http://link.springer.com/10.1007/978-3-540-30182-0_66

Isolation Forest Based Anomaly Detection Approach for Wireless Body Area Networks

Murad A. Rassam[1,2(✉)]

[1] Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
m.qasem@qu.edu.sa

[2] Faculty of Engineering and Information Technology, Taiz University, 6803 Taiz, Yemen

Abstract. Anomalous data detection is an important task for ensuring the quality of data in many real-world applications. Medical healthcare services are one such application where Wireless Body Area Networks (WBAN) is used to track human health situations. Such tracking is achieved by collecting and monitoring the basic physiological vital signs and making them available to the healthcare givers to assess the criticality status of patients, especially in Intensive care units (ICU). Various anomaly detection approaches have been proposed for detecting anomalies collected in WBAN such as statistical, machine learning and deep learning techniques. However, the lack of ground truth data made the job of training such models a difficulty in supervised settings. In this paper, an Isolation Forest-based anomaly detection approach for WBAN (iForestBAN-AD) model is proposed. The iForest technique is fully unsupervised and does not employ any distance measure or density function like most existing techniques and rather detects anomalies based on the concept of isolation. To evaluate the proposed approach, experiments on data samples from real world physiological network records (Physionet) were conducted. The results show the viability of the proposed approach as it achieves around 95% AUC and outperforms many of the existing baseline unsupervised techniques on multivariate dataset samples.

Keywords: Anomaly Detection · Wireless Body Area Network · Isolation Forest · Unsupervised Learning · Internet of Medical Things

1 Introduction

Remote and pervasive vital signs monitoring become a necessity in societies where the average lifetime increases and the number of elderly people who need continuous monitoring are exponentially increasing especially in Europe. Such an increase creates an overload in the healthcare sectors and urges the need for pervasive systems that can monitor large numbers of patients easily. Furthermore, the increase of patients who require ICU admission and monitoring requires automated systems to handle the continuous monitoring of patients in such units and facilitates the decision-making process by doctors and healthcare givers.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1347–1357, 2023.
https://doi.org/10.1007/978-3-031-27409-1_123

Internet of Medical Things (IoMT) is the concept of collecting, analyzing and storing health-related data by tiny sensors that constitute the body area sensor networks. Such data includes many vital signs observations such as blood pressure (BP), oxygen saturation (SPO2), and pulse rate among others [1]. Figure 1 shows different sensors implanted over the human body to measure the vital signs used to monitor the health condition of patients at home or in the ICUs.

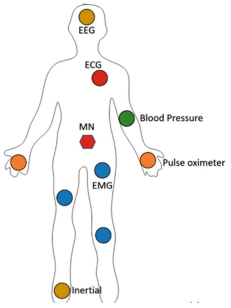


Fig. 1. Wireless Body Area Networks [2]

Ensuring the quality of collected data in WBANs for healthcare monitoring applications is a prominent research area in which the anomaly detection concept is employed to detect anomalous observations that arise due to various reasons. Several anomaly detection approaches for WBANs have been introduced in the literature based on statistical, machine learning and other techniques such as [3–6]. However, such approaches employed techniques that are computationally heavy and therefore require a considerable amount of time that can be critical for the healthcare monitoring situation. In addition, some of the existing approaches have not considered the situation of more than one parameter and monitor individual signs separately.

To this end, this paper considers the situation of detecting anomalous observations in multivariate healthcare data by utilizing the concept of isolation. To achieve this goal, the Isolation Forest (iForest) algorithm is employed where 6 vital signs of healthcare data recorded at the ICU are considered altogether to build a model for efficient detection. The isolation concept as explained and employed in [7] can achieve a low linear time complexity and a small memory requirement because it does not depend on any distance measure calculations.

The contribution of this paper is as follows:

- Proposing a new anomaly detection model for WBAN based on the iForest technique.
- A comparative analysis of the proposed model with baseline existing models in the literature.

The rest of this paper is organized as follows: Sect. 2 reviews and analyzes the literature on anomaly detection for WBAN. Section 3 introduces the proposed model and presents a background on the isolation concept. Section 4 presents the experimental

evaluation results and compares the proposed model with existing literature. Section 5 concludes this paper.

2 Related Works

Various anomaly detection schemes have been proposed to detect the abnormal readings collected by WBANs to facilitate accurate decisions by healthcare givers. Such schemes were designed based on different approaches such as machine learning approaches [8, 9], statistical approaches [3, 10–13] and game-based approaches [14] among others. Statistical approaches can be used in two modes: parametric-based and non-parametric-based. Similarly, Machine learning approaches are used in supervised and unsupervised settings. However, the lack of ground truth data to train ML models in supervised modes and the nature of the anomaly detection problems made the unsupervised approaches a choice.

Unsupervised ML models such as [15, 16] have been used to detect anomalous data observations in WBANs. In such models clustering algorithms such as K-Means, hierarchal clustering and fuzzy C-Means clustering techniques are used. However, in such models, authors assumed that the clusters are well distinguishable and therefore a distinct line between normal and anomalous readings is clear. This assumption is not realistic in the case of physiological readings where the abnormal readings for a patient can be considered normal for another.

A study in [3] introduced an approach for detecting continuous changes in readings such as modifications, forgery, and insertions in electrocardiogram (ECG) data. A Markov model with different window sizes (5% and 10%). Using only univariate data, the study reported 99.8% and 98.7% of true negatives with 5% and 10% windows size, respectively. The Markov-based models usually have small time execution but the space complexity is high.

Two types of correlation exist in healthcare data and any other time series data which are temporal and spatial correlation. Temporal correlation refers to the strong relationships between data observations of the same variable according to the time stamp. Spatial correlation refers to the relationship between more than one variable at the same time stamp. Authors in [4, 17] considered the correlation that exists among data observations temporally and spatially. The results of the proposed approaches are found to be enhanced when both types of correlation are utilized. However, those approaches have high computational complexities and cannot be utilized in real-time.

In [13], authors proposed a model for anomaly detection in WBAN by adopting the data sampling approach with the Modified Cumulative Sum (MCUSUM) technique. The proposed approach aimed to enhance the speed of detection by the sampling method, while the use of the MCUSUM algorithm aimed to enhance the security of detection. Although the proposed approach in this work enhances the detection efficiency, the stationary process incurred by MCUSUM made it difficult to detect random and emergent anomalies. In addition, the linear statistical-based approaches are always parametric which makes them unsuitable for real-world applications.

One-class machine learning approaches such as [18, 19] are the most suitable unsupervised learning approaches for anomaly detection in sensor systems. It depends on

the availability of normal observations to build a model that can detect any abnormalities in future sensor readings. An example of such approaches. However, most such approaches are implemented to consider only univariate variables of the physiological measurements separately.

Isolation forest-based approaches have been utilized in the literature as good candidates to develop efficient anomaly detection models. In [7] authors employed the concept of isolation to detect anomaly detection efficiently without the need of using distance functions. The experimental evaluation of the iForest proposed in this study using different datasets shows that the iForest approach outperforms the one-class support vector machine (OCSVM) and the local outlier function (LOF) approaches. Furthermore, authors in [20] designed a model for distributed anomaly detection in wireless sensor networks based on the isolation principle. The authors claimed that the proposed isolation principle approach helps to reduce the computational complexity and therefore reduces the energy consumption besides it achieves better detection accuracy results as it utilizes the spatial and temporal correlation in a distributed fashion. Another study [21] utilized the isolation principle in combination with the concept of drifting to detect anomalies in data streams. Both studies [20, 21] evaluated the isolation principle using several real-world datasets to report its viability.

To the best of our knowledge, the isolation principle has not been used for detecting anomalies in the context of the data streams collected by WBAN. Therefore, in this paper, we aim to prove that the isolation forest (iForest) algorithm is a good candidate for detecting anomalies in the context of WBAN more effectively compared to existing models in the literature.

3 Proposed Approach

To discuss the design of the proposed iForestBAN-AD model, we adopt the scenario that several m sensors nodes are placed on multiple different positions of the patient body to collect various physiological observations of the vital signs as in Fig. 2.

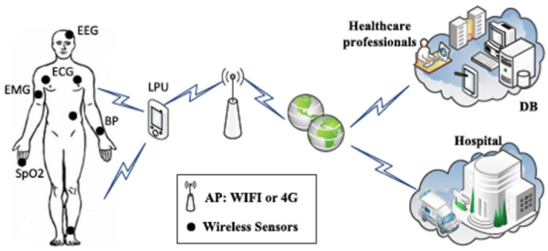


Fig. 2. A scenario for WBAN deployment [12]

As shown in Fig. 2, the collected observations are sent to the Local Processing Unit (LPU) which has enough resources for processing data received from sensors and detects anomalous observations before sending the data to the healthcare professionals or hospital management. If the observations are found to be anomalous, an alarm or

any kind of notifications is sent to the healthcare givers to check if it is a sign of health degradation or faulty measurements.

3.1 Principle of Isolation Forest

According to [7], the term, isolation, refers to "separating an instance from the rest of others". The principle of isolation-based anomaly detection models is to measure each data observation's susceptibility to being isolated whereas anomalies are those that have this degree of susceptibility. To model the idea of isolation, a tree structure of the observations that naturally isolates data is modelled. This tree is composed of random binary trees of recursively partitioned observations. To detect anomalous instances, we rely on the fact that anomalies tend to form shorter path trees because anomalies are less in the count and then it results in a smaller number of partitions represented by shorter paths in a tree structure. Moreover, anomalies are those instances with distinguishable feature values which are to be separated early in the partitioning process. As a result, when a shorter path length of the forest of random trees is produced for some points, these points are most likely to be anomalies.

3.2 IForestBAN-AD Model

Figure 3 presents the different phases of the proposed *iForestBAN-AD* model. The details of the proposed model and its different phases are given in the following subsections.

Data collection/loading: in this stage, the real-world physiological dataset samples are collected and loaded to train and evaluate the proposed model. More details on the dataset will be given in Sect. 4.

Data preprocessing: some preprocessing steps are applied to the dataset samples to make them suitable for machine learning operations. Such steps include removing null values of the data observations, scaling the data in the range (0,1) using the min and max functions, and labelling the data observations as 0/1 classes. Such labelling will be used only for the evaluation processes to test the efficacy of the proposed model in terms of accuracy of detection. It is worth mentioning that the iForest algorithm is trained unsupervised, as we will see in the following subsection, to determine the anomaly score that will be used later to decide the abnormality of the data instances.

iForestBAN-AD Engine: the engine of the proposed anomaly detection model based on iForest algorithm consists of two main stages: the training stage and the evaluation stage. The description of the iForest algorithm is adopted from [7].

Let $X = \{x1,\ldots,xn\}$ be the dataset observations. A sample of instances $X' \subset X$ is used to build an isolation tree (iTree). The sample X' is divided recursively by the random selection of an attribute q and a split value p, the process continues until one of the stop criteria is fulfilled which are: (1) one instance only remains in the node, or (2) the node had data items of the same values.

An iTree by definition is a binary tree, in which every node in that tree is composed of zero or two daughter nodes.

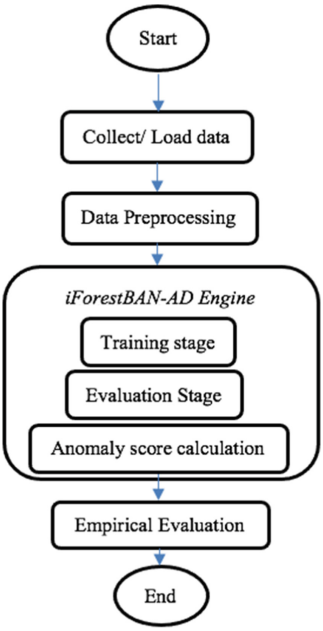


Fig. 3. The *proposed iForestBAN*-AD Model

1. **Training stage**

In this stage, iTrees are formed by dividing a subsample X′ recursively until all instances are isolated. A simple abstraction of the training process is given in the pseudocode in Table 1.

The details of iTree(X′) procedure of constructing the iTrees can be found in [7]. As the result of the training process, a number of trees (forest) are returned and is ready to be tested.

2. **Evaluation stage**

In the testing stage, as detailed in [7], a single path length h(x) is derived by counting the number of edges e from the root node to an external node as instance x traverses through an iTree.

3. **Anomaly Score Calculation**

The anomaly score s of an instance x is defined as:

$$s(x, \psi) = 2^{\frac{-E(h(x))}{c(\psi)}} \tag{1}$$

Table 1. Pseudocode of Training iForest Algorithm

Algorithm 1 : *iForest(X,t,ψ)*

Inputs: *X* - input data, t - number of trees, ψ - subsampling size

Output: a set of t iTrees

1: Initialize Forest

2: for i=1 to t do

3: X′ ←sample(X,ψ)

4: Forest← Forest ∪ iTree(X′)

5: end for

6: return Forest

where E(h(x)) is the average of h(x) from a set of iTrees, h(x)single path length, c is the average path length of unsuccessful searches.

4 Experimental Evaluation

After the anomaly score is obtained, a set of experiments using different dataset samples is implemented to verify the efficacy of the proposed model. Before describing experiments and reporting the results, a short description of the dataset is given.

Figure 4 shows the variation of data observations for subject 330 in the MIMIC II. In Fig. 4a, all features are depicted, whereas in Fig. 4b–g the variations in Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Mean Blood Pressure, Pulse Rate, and Oxygen Saturation, respectively. Some features are highly correlated such as HR and Pulse which are identical. Furthermore, a high positive correlation is noticed between ABPDias in Fig. 4d and ABPMean in Fig. 4e.

To evaluate the proposed iForestBAN-AD model, Samples from subjects 330 and 441 are used. The proposed model was developed using various python libraries such as pandas, Numpy, and Sklearn. The settings of the parameters of the iForest algorithm used in this research is reported in Table 2.

The Area under Curve (AUC) scores and the MAE metrics are used to evaluate the proposed model, and to compare its performance with existing unsupervised models in the literature namely One-Class Support Vector Machines (OSVM), KMeans, and the Local Outlier Factor (LOF). For the evaluation of the proposed model as an unsupervised

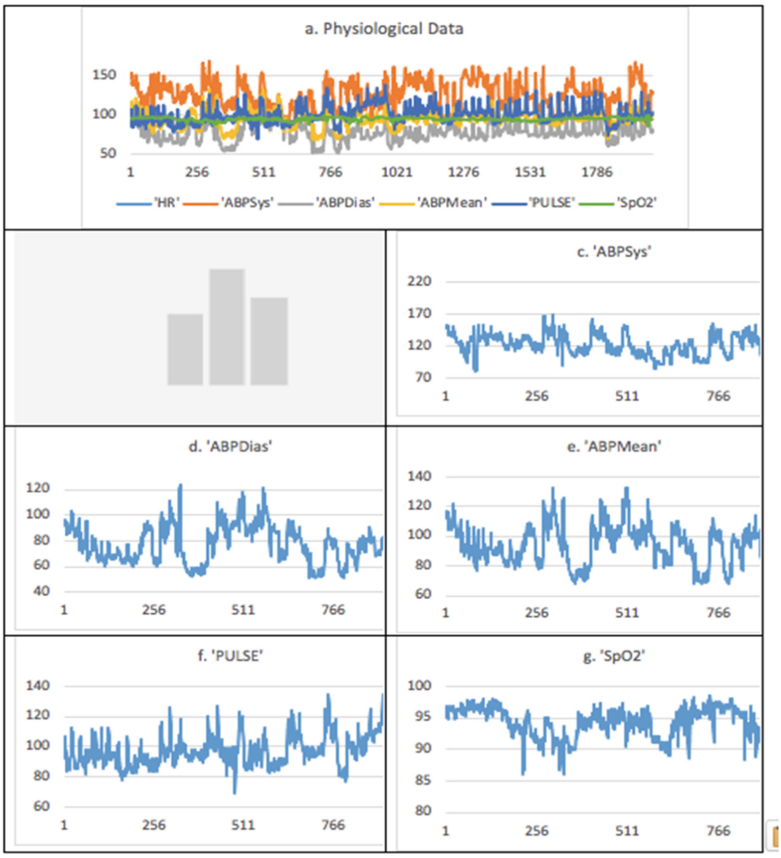


Fig. 4. Sample of physiological data for Subject 330

Table 2. Parameter settings of the iForest algorithm used in this rpaper.

Parameter	Value
n_estimators	100
max_samples	'auto'
contamination	variable
max_features	1.0
random_state	42
verbos	0

model, we assume that anomaly labels are unavailable in the training stage and are only available in the evaluation stage to calculate the evaluation measure, AUC.

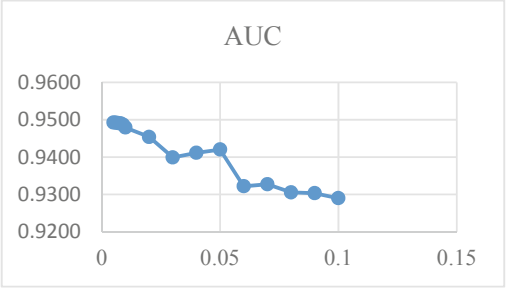


Fig. 5. Accuracy versus contamination ratio for Subject 330 records.

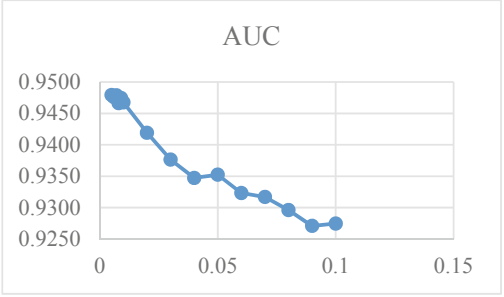


Fig. 6. Accuracy versus contamination ratio for Subject 441 records.

Figures 5 and 6 show the AUC scores of the proposed model on subject 330 and subject 441, respectively. The AUC scores are depicted versus the contamination parameter which is found to be the only parameter that slightly affects the performance. The contamination parameter controls the threshold for deciding when a scored data instance should be considered an anomaly. The experimental evaluation clearly shows that the best achieved AUC score is around 95% for both data subjects which indicates that the performance is stable with different data samples. The results further show that the AUC increases by a decrease in the contamination parameter which controls the ratio of anomalous measurements in the subsample. The number of trees t is made constant and equals 100. The average Mean Absolute Error (MAE) reported is 0.228 and 0.227 for subject 330 and subject 441 records, respectively. It is also noticed that the t value of the MAE decreases with the decrease of contamination parameter for this dataset records.

The performance of the proposed model was empirically compared with 3 existing unsupervised anomaly detection models proposed in the literature named KMeans, LOF and the OCSVM using the subject 330 records as shown in Table 2. The comparison shows that the proposed model outperforms those models as it achieves 95% AUC compared to 72% for OCSVM, 51% for LOF and 41% for KMeans. The OCSVM scored second and the best of the other candidates whereas KMeans is the worst (Table 3).

Table 3. Comparison of AUC on Subject 330 records

	OCSVM	KMeans	LOF	iForest (Proposed)
AUC	0.72	0.41	0.51	0.95

5 Conclusion

Ensuring the quality of vital signs observations collected by WBANs is crucial to facilitate taking timely and accurate decisions by healthcare givers in the IoMT applications. In this paper, the unsupervised iForest algorithm was used to design a model for detecting anomalous data observations in WBANs and therefore ensuring data quality. The experimental evaluation on a real-world physiological data records has proved that the proposed approach outperformed existing baseline unsupervised approaches. Furthermore, the concept of isolation reduces the computational burden that is usually resulted from the employment of machine and deep learning models as it does not require distance measures calculations. In future, the concept drifting of data needs to be investigated together with the isolation concept in order to consider the context of patient in near real time.

References

1. Santos, M.A., Munoz, R., Olivares, R., Rebouças Filho, P.P., Del Ser, J., de Albuquerque, V.H.C.: Online heart monitoring systems on the internet of health things environments: A survey, a reference model and an outlook. Inf. Fusion **53**, 222–239 (2020)
2. Al-Mishmish, H., Alkhayyat, A., Rahim, H.A., Hammood, D.A., Ahmad, R.B., Abbasi, Q.H.: Critical data-based incremental cooperative communication for wireless body area network. Sensors 18, 3661 (2018)
3. Khan, F.A., Haldar, N.A.H., Ali, A., Iftikhar, M., Zia, T.A., Zomaya, A.Y.: A continuous change detection mechanism to identify anomalies in ECG signals for WBAN-based healthcare environments. IEEE Access **5**, 13531–13544 (2017)
4. Mohamed, M.B., Makhlouf, A.M., Fakhfakh, A.: Correlation for efficient anomaly detection in medical environment. In: 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 548–553. IEEE, (Year)
5. Salem, O., Serhrouchni, A., Mehaoua, A., Boutaba, R.: Event detection in wireless body area networks using Kalman filter and power divergence. IEEE Trans. Netw. Serv. Manag. **15**, 1018–1034 (2018)
6. Saneja, B., Rani, R.: An integrated framework for anomaly detection in big data of medical wireless sensors. Mod. Phys. Lett. B **32**, 1850283 (2018)
7. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation-based anomaly detection. ACM Trans. Knowl. Disc. Data (TKDD) **6**, 1–39 (2012)
8. Lau, B.C., Ma, E.W., Chow, T.W.: Probabilistic fault detector for wireless sensor network. Expert Syst. Appl. **41**, 3703–3711 (2014)
9. Saraswathi, S., Suresh, G.R., Katiravan, J.: False alarm detection using dynamic threshold in medical wireless sensor networks. Wirel. Netw. **27**(2), 925–937 (2019). https://doi.org/10.1007/s11276-019-02197-y

10. Zhang, H., Liu, J., Pang, A.-C.: A Bayesian network model for data losses and faults in medical body sensor networks. Comput. Netw. **143**, 166–175 (2018)
11. GS, S., Balakrishnan, R.: A Statistical-based light-weight anomaly detection framework for wireless body area networks. Comput. J. **65**, 1752–1759 (2022)
12. Salem, O., Alsubhi, K., Mehaoua, A., Boutaba, R.: Markov models for anomaly detection in wireless body area networks for secure health monitoring. IEEE J. Selec. Areas Commun. **39**, 526–540 (2020)
13. Boudargham, N., El Sibai, R., Bou Abdo, J., Demerjian, J., Guyeux, C., Makhoul, A.: Toward fast and accurate emergency cases detection in BSNs. IET Wirel. Sensor Syst. **10**, 47–60 (2020)
14. Arfaoui, A., Kribeche, A., Senouci, S.M., Hamdi, M.: Game-based adaptive anomaly detection in wireless body area networks. Comput. Netw. **163**, 106870 (2019)
15. Ahmad, B., Jian, W., Ali, Z.A., Tanvir, S., Khan, M.: Hybrid anomaly detection by using clustering for wireless sensor network. Wirel. Pers. Commun. **106**, 1841–1853 (2019)
16. Qu, H., Lei, L., Tang, X., Wang, P.: A lightweight intrusion detection method based on fuzzy clustering algorithm for wireless sensor networks. Advances in Fuzzy Systems 2018, (2018)
17. Albattah, A., Rassam, M.A.: A correlation-based anomaly detection model for wireless body area networks using convolutional long short-term memory neural network. Sensors **22**, 1951 (2022)
18. Shahid, N., Naqvi, I.H., Qaisar, S.B.: One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh environments. Artif. Intell. Rev. **43**(4), 515–563 (2013). https://doi.org/10.1007/s10462-013-9395-x
19. Zhang, Y., Meratnia, N., Havinga, P.: Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks. In: 2009 international conference on advanced information networking and applications workshops, pp. 990–995. IEEE, (Year)
20. Ding, Z.-G., Du, D.-J., Fei, M.-R.: An isolation principle based distributed anomaly detection method in wireless sensor networks. Int. J. Autom. Comput. **12**(4), 402–412 (2015). https://doi.org/10.1007/s11633-014-0847-9
21. Togbe, M.U., Chabchoub, Y., Boly, A., Barry, M., Chiky, R., Bahri, M.: Anomalies detection using isolation in concept-drifting data streams. Computers **10**, 13 (2021)

Author Index

A
Abdellaoui, Zaineb 786
Abou El Kalam, Anas 1315
Afzal, Rafi 441
Agarwal, Shubham 1022
Agnihotri, Shikha 814
Ahmad, Tohari 1235
ahmed, Imen Mohamed ben 636
Akter, Bonna 127
Al Ghawi, Sana 453
Alam, Mohammad Jahangir 237
Alaya, Bechir 337
Al-Barmani, Zahraa 63
Albuquerque, I. M. C. 412
Alcântara, J. P. M. 412
Al-Janabi, Samaher 63, 79, 152, 222
Alqatawneh, Ibrahim 198
Al-Rajab, Murad 198
Amador-Angulo, Leticia 746
Aniyan, Nisha 1179
Arfaoui, Nouha 401
Aroba, Oluwasegun Julius 948
Arshath Raja, R. 213
Arunkumar, S. 96
Asanov, A. S. 1189
Asif, Asifuzzaman 237
Aslam, Sultan Md. 981
Assafra, Khadija 337
Aswathy, S. U. 1179
Avanija, J. 872
Avasthi, Madhav 595
Ayed, Yassine Ben 501

B
Bacanin, Nebojsa 302
Bafna, Prafulla 1013, 1022
Baker, Mohammed Rashad 625
Bal Raju, M. 940
Bandyopadhyay, Sanghamitra 521
Bansal, Shweta 1209
Barman, Mala Rani 127
Barros, Mateus F. 644
Basheer, Aysha 521
Ben Aouicha, Mohamed 249
Benmohamed, Emna 762
Benslimane, Djamal 281
Bernábe Loranca, M. Beatriz 379
Berriche, Wassim 1122
Bhanu, B. Balaji 852, 862
Bhuvaneswari, S. 106
Biswal, Sumitra 1143
Biswas, Al Amin 166
Boaullegue, Ridha 177
Bouazizi, Samar 762
Boukhris, Imen 543
Boulaares, Soura 281
Bozhenyuk, Alexander 391, 668

C
Campos, Alexis 13
Canán, Alberto Carrillo 379
Cao, Zechun 1337
Carlsson, Robin 1284
Carneiro, Davide 328, 703
Castillo, Oscar 746
Cebrián-Hernández, Ángeles 260
Chaibi, Nesrine 291
Challa, Nagendra Panini 852
Chand, Smarth 1270
Chandra Das, Badhan 166
Chang, Wui-Lee 510
Chari, Soham 655
Chen, Yi-Li 566
Chen, Yingke 1167
Chetouane, Ameni 1199
Chidambaram, S. 213, 831, 841

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1359–1363, 2023.
https://doi.org/10.1007/978-3-031-27409-1

Chinsamy, Kameshni K. 948
Chitteti, Chengamma 872
Chowdhury, Anupam 981
Correia, Ricardo 703
Costa, Helder Gomes 432, 804, 907

D
da Silva, Welesson Flávio 432
Dahal, Roshan 831
Dang, Quang-Vinh 1114
Daniel, Doney 1179
Danilchenko, Eugenia V. 618
Danilchenko, Vladislav I. 575, 618
Das, Badhan Chandra 714
Das, Monidipa 521
Das, Samar 981
Dasho, Anni 302
de Mello, João Carlos Correia Baptista Soares 432
de P. Canuto, Anne Magály 644
de Souza, Hudson Hübner 432
de Souza, Luciano Azevedo 432
Deepak, Gerard 96, 116, 270
Denisova, A. Y. 1303
Derbel, Mouna 786
Devi Priya, R. 991
Devi, Urmila 1209
Devipriya, R. 929
Dhali, Aditi 127
Dhanvardini, R. 116
Dhattarwal, Aayush 313
do Canto Souza, Wesley 432
Dorji, Sonam 831, 841

E
El Aassal, Ayman 1327
El Balbali, Hiba 1315
El Kamel, Ali 1244
Eladel, Asma 291
ElBehy, Ichrak 357
Eltaief, Hamdi 1225, 1244
Erromh, Mohamed Ali 543

F
Faiz, Sami 281
Farhat, Mariem 625
Farook, S. 862
Farooq, Ali 1154
Faruqui, Nuruzzaman 188, 366
Fedoseev, V. A. 1189, 1303
Ferdib-Al-Islam, 441
Fkih, Fethi 423, 493
Frikha, Mondher 625

G
Gargouri, Faiez 585
Garnica, Carmen Cerón 379
Gaskó, Noémi 756
Gayathri, V. P. 680
Gerasimenko, Evgeniya 391, 668
Ghorbel, Ahmed 786
Ghozzi, Faiza 585
Gorgônio, Arthur C. 644
Goyal, Saliya 1013
Gupta, Daya Sagar 1132
Gupta, Deeya 1067
Gupta, Pranav 1067
Gupta, Ritu 824
Gyeltshen, Pema 841

H
Hadj Taieb, Mohamed Ali 249
Hadriche, Abir 357
Haider, Thekra 152
Hajdarevic, Zlatko 302
Hamza, Sihem 501
Haranath, A. Prem Sai 862
Hasan, Khan Mehedi 441
Hasan, Omlan 981
Hasib, Khan Md. 714
Hasneen, Jehan 1293
Heino, Timi 1284
Hemanth, C. 53, 726
Henrietta, H. Mary 21
Hindhuja, V. 1041
Hong, Tzung-Pei 566
Hossain, Syed Md. Minhaz 1254
Huang, Wei-Ming 566
Hussain, Shaik Asif 453

I
Ijtihadie, Royyana Muslim 1235
Islam, Md. Jahidul 237
Islam, Md. Rahatul 441
Ito, Keisho 1093
Iyer, Nalini C. 1102

J
Jacob, Pramod Mathew 1179
Jani, Rafsun 714
Janicijevic, Stefana 302
Jarray, Ridha 357
Jasmy, Ahmad Jasim 198
Jayabharathi, C. 1041, 1053
Jihad, Kamal H. 625
Jiménez-Rodríguez, Enrique 260
Jmail, Nawel 357
Jovanovic, Luka 302
Júnior, João C. Xavier 644
Jurme, Tshewang 841

K
Kadhuim, Zena A. 222
Kalamani, M. 1041, 1053
Kamalam, Gobichettipalayam Krishnaswamy 680, 690
Kamel Ali, El 1225
Kanakachalam, Sruthi 690
Kanber, Shashikantha G. 736
Kannan, Beulah Divya 270
Kar, Jayaprakash 1067
Karotia, Akanksha 532
Karoui, Hend 501
Karoui, Kamel 1199
Karthik, B. Venkata Phani 872
Kavitha, K. 971
Keshavamurthy, Bettahally N. 350
Kherallah, Monji 636
Khoi, Bui Huy 142, 1082
Kirupa, P. 991
Kishibuchi, Ryohei 1093
Kotecha, Ketan 106
Krisna Pamungkas, I Gede Agung 1235
Kulkarni, Sushil 480
Kumain, Hitesh Mohan 655
Kureychik, Viktor M. 575, 618

L
Laato, Samuli 1154
Lacerda, M. G. P. 412
Lahmar, Ines 177
Lakshminarayana Reddy, D. 881
Lavanya, K. R. 893
Leandro, Andrés 607
Leppänen, Ville 1284
Lima-Neto, F. B. 412
Ling, Jill 510
Ltifi, Hela 762
Lung, Rodica Ioana 556, 776
Luque, Gabriel 607

M
Maalej, Rania 636
Madhavi, K. Reddy 852
Mahapatra, Aishwarya 1067
Mahbubur Rahman, Md. 166
Mahmood, Saif 79
Majed, Hadeer 79
Makka, Shanthi 940
Makwakwa, Tsepo G. 948
Malhan, Shivani 814
Mallek, Hana 585
Malwad, Dnyaneshwar S. 464
Manjith, R. 1053
Manoj Kumar, S. 991
Marjanovic, Marina 302
Marleni Reyes, M. 379
Melin, Patricia 13
Minhaz Hossain, Syed Md. 981
Mishra, Pooja Manghirmalani 480
Moatemri, Maroua 1244
Mohammed, Ghada S. 152
Monteiro, José 328
Monteiro-Filho, J. B. 412
Mouthami, K. 959, 991
Mridul, Aunik Hasan 237
Musfique Anwar, Md. 166
Mythili, S. 1041, 1053

N
Naik, Sachin 595
Nair, Akhil R. 726
Nair, Karthika S. 726
Nakouri, Haïfa 543
Narayanan, R. Lakshmi 971
Narayanan, Vishnupriya 1293
Naresh, Chandragiri 852
Naseeba, B. 852, 862, 872
Ngan, Nguyen Thi 142, 1082
Nimisha, C. 1041
Noman, Syed Muhammad 198

O
Oliveira, Óscar 328
Omote, Kazumasa 1093

Ong, Sing-Ling 510
Opara, Chidimma 1167

P
Pamarthi, Sasi Preetham 862
Panwar, Surya Nandan 1013
Patel, Nikhil 188, 366
Patil, Nagamma 736
Pawar, Samruddhi 1022
Peter, Geno 1031
Phelgay, Thinley 841
Pidadi, Parth 464
Pooja, R. I. 959
Pousia, S. 1041, 1053
Praghash, K. 213, 831, 841, 1031
Pramod, Dhanya 3
Priya, R. Devi 1031
Priyadarshini, J. Sheeba 96

R
Ragamathana, R. 53
Rahman, Mushfiqur 237
Rahman, Shahid 441
Rahul, 1270
Rai, Rajesh 831
Raj, A. Stanley 21
Raja, Sandhiya 690
Rajbongshi, Aditya 127
Ramasahayam, Shravya 940
Rao, B. Narendra Kumar 862
Rassam, Murad A. 1347
Rathore, Hemant 655
Ratnoo, Saroj 313
Rauti, Sampsa 1154, 1284
Reddy, Sujan 736
Rhouma, Delel 423, 493
Ritu, Ritu 917
Roboredo, Marcos Costa 804
Rodzin, Sergey 391

S
S. Barreto, Cephas A. da 644
Sadique, Kazi Masum 1293
Sai, Katari 872
Sai, Pothuri Hemanth Raga 872
Sailhan, Francoise 1122
Sakthivel, K. 929, 1002
Saminathan, A. 929, 1002
Sánchez, Daniela 13
Sandeep, V. R. 736
Sangeetha, R. G. 53
Sangeetha, R.G. 726
Sanim, Mostofa Shariar 441
Santhanavijayan, A. 96, 116
Sara, Umme 127
Saravanan, M. 929, 1002
Sarvavnan, M. 971
Sassi, Salma 281
Sayeed, Taufique 1254
Sequeira, Evander Darius 350
Shakil, Rashiduzzaman 127
Shanto, Md. Shariful Islam 714
Sharma, Saloni 1270
Shekhawat, Hema 1132
Shiddiqi, Ary Mazharuddin 1235
Shine, K. Nithin 726
Shoba Bindu, C. 881, 893
Shraddha, B. H. 1102
Shreecharan, D. 213
Singh, Gaurav 464
Singhal, Tanya 1270
Sousa, Cristovão 703
Souza, Luciano Azevedo de 804
Stephen Huang, Shou-Hsuan 1327, 1337
Stonier, Albert Alexander 1031
Subramaniyaswamy, V. 106
Suciu, Mihai 556
Sujith, J. G. 929, 1002
Sunil, C. K. 736
Sunitha, Lingam 940
Susan, Seba 532
Swarnkar, Latika 1067

T
T, Rashmi 655
Tahir, Sheikh Badar ud din 366
Talha, Mohamed 1315
Tallón-Ballesteros, Antonio J. 260
Tarannum, Sabrina 1254
Tarasov, A. A. 1303
Targino, Victor V. 644
Thabet, Kawther 1225
Tharanyaa, J. P. Shri 1053
Thomas, Jyothi 1179
Tobgay, Thinley 841
Trivedi, Sandeep 188, 366
Trung, Ha Duyen 32, 43
Tsai, Yu-Chuan 566
Tsuzuki, Yuto 1093

Turki, Houcemeddine 249

V
Vanitha, Pazhanisamy 680
Venugopal, Gayatri 3, 595
Vignesh, N. 106
Vijesh, S. 929, 1002
Villar-Dias, J. L. 412
Vincent, Bibin 1179
Vybornova, Y. D. 1189

W
Wali, Heera 1102
Wangchuk, Kinley 831
Wazarkar, Seema 350
Wei, Bo 1167

Y
Yada, Shohei 1093
Yahia, Mohamed 177
Yan, Cao 1093
Youssef, Habib 1244
Yuvaraj, N. 213, 959, 1031

Z
Zaied, Mourad 291
Zaier, Aida 177
Zidi, Salah 337
Zivkovic, Miodrag 302
Zrigui, Mounir 337