International Journal of Transport Development and Integration
Vol. 7, No. 2, June, 2023, pp. 131-138
Journal homepage: http://iieta.org/journals/ijtdi

Analysis of Road Traffic Accidents Severity Using a Pruned Tree-Based Model

Timothy T. Adeliyi[1*], Deborah Oluwadele[1], Kevin Igwe[2], Oluwasegun J. Aroba[3]

[1] Department of Informatics, University of Pretoria, Pretoria 0083, South Africa
[2] Department of Psychology, University of Johannesburg, Johannesburg 2092, South Africa
[3] ICT and Society Research Group, Information Systems, Durban University of Technology, Durban 4001, South Africa

Corresponding Author Email: timothy.adeliyi@up.ac.za

https://doi.org/10.18280/ijtdi.070208

Received: 12 January 2023
Accepted: 11 May 2023

Keywords:
accident severity, machine learning algorithms, pruned tree-based model, road traffic accident

ABSTRACT

Traffic accidents are becoming a global issue, causing enormous losses in both human and financial resources. According to a World Health Organization assessment, the severity of road accidents affects between 20 and 50 million people each year. This study intends to examine significant factors that contribute to road traffic accident severity. Seven machine learning models namely, Naive Bayes, KNN, Logistic model tree, Decision Tree, Random Tree, and Logistic Regression machine learning models were compared to the J48 pruned tree model to analyze and predict accident severity in the road traffic accident. To compare the effectiveness of the machine learning models, ten well-known performance evaluation metrics were employed. According to the experimental results, the J48 pruned tree model performed more accurately than the other seven machine learning models. According to the analysis, the number of casualties, the number of vehicles involved in the accident, the weather conditions, and the lighting conditions of the road, is the main determinant of road traffic accident severity.

1. INTRODUCTION

People are killed in car accidents every day. As a result, there has been a great deal of research into the factors that determine accident severity in road traffic accidents. Despite the United Nations' goal of reducing the number of road traffic deaths and injuries by 2030, between 20 and 50 million people are injured each year, with 1.3 million dying. Despite having approximately 60% of the world's vehicles, low-and middle-income countries account for 93% of these fatalities. Road traffic accidents have a detrimental impact on most countries, costing them up to 3% of their GDP [1]. Furthermore, more males between the ages of twenty and forty die in traffic accidents each year than females, resulting in financial issues in families and increasing poverty [2].

A cross-disciplinary study is required to discover and analyze the factors affecting and contributing to the severity of road traffic accidents and driver injuries, which are global concerns. Literature has identified driver inattention induced by distraction, overloaded attention, and boring driving as human factors contributing to road traffic accidents [3]. Road traffic accidents are exacerbated by distraction, weather conditions, sleep deprivation, improper lane changes, and nighttime driving [2]. While some studies look at the elements that influence road traffic accidents and their severity, others dig deeper to find the most likely causes of road traffic accidents and what determines their severity.

Numerous research has been undertaken in various countries to offer road accident prediction and analysis methods [4]. Cao et al. [5] present a systematic strategy for identifying severe driving episodes concerning time and place using batch clustering and real-time clustering techniques. Fawcett et al. [6] suggested a Bayesian hierarchical model for projecting accident numbers in future years at places within a pool of probable road safety hotspots. The model assists road safety practitioners in determining the location of potential future hotspots, allowing them to take a proactive rather than a reactive strategy for road safety plan implementation.

Models for forecasting road traffic events were also investigated in the literature. By mining seven months of accident data and 1.6 million users' G.P.S. records, Chen et al. [7] created a deep stack denoise autoencoder model to learn human mobility's hierarchical features to predict traffic accident risks. Lu et al. [8] created a road traffic prediction model called TAP-CNN that takes into account traffic accident-impacting factors such as traffic flow, weather, and light. The samples utilized to test the model's accuracy revealed that the model outperformed the classic neural model in forecasting road traffic. You et al. [9] employed the Matched case-control approach and support vector machines (SVMs) methodologies to assess risk status and estimate the likelihood of an accident using discrete loop detector traffic data and web-crawl weather data. The SVMs classifier with web-crawl weather data improves crash prediction accuracy while reducing false alarm rate when compared to the SVMs classifier without data.

Authors in the studies [10-12] investigated the prediction of the severity of the repercussions incurred by road traffic accident victims. This includes employing latent class clustering, an unsupervised probabilistic clustering approach, to investigate the trends, prevalence, and severity of bike motorist collisions in Denmark [10]. In the study [11], a deep learning-based convolutional neural network was used to extract weights of traffic accident features to improve the

accuracy of prediction. Rezaie Moghaddam et al. [12] used artificial neural networks and a deep learning approach to predict road traffic accident severity using the unique traffic accident severity prediction conventional neural network (TASP-CNN) model. Although machine learning techniques and models have been used to predict traffic accidents and their severity, little emphasis has been made on the precision and accuracy of these models' performance. Furthermore, the implication of unknown data in the dataset utilized by these models warrants further investigation because it is commonly assumed that a model performs well based on the output.

This study builds on earlier research to assess the effectiveness of multiple classical machine learning models, addressing previously neglected areas of forecasting factors that influence accident severity. The J48 pruned tree approach is also investigated in the paper as a machine learning model that efficiently deals with unknown data in the dataset. The J48 pruned tree-based classifier model is then used to compare and analyze a dataset of road traffic accidents with seven other cutting-edge classical machine learning models. Nave Bayes, Bagging, KNN, Logistic Model Tree, Decision Tree, Random Tree, and Logistic Regression are among these models. The findings will add to the body of knowledge in this domain by emphasizing the impact of unknown variables in the dataset on the performance of classical machine learning models and how this affects the precision of predicting road traffic accidents and their severity.

2. RELATED WORKS

Despite frantic efforts to design and build vehicles that are less likely to be involved in traffic accidents, road traffic accidents remain one of the primary causes of death in both developed and developing countries. Data mining and machine learning techniques and tools provide powerful techniques and tools for analyzing and forecasting the severity of road accidents and as a result, reduce the rate of road-traffic-induced casualties. Various studies have been conducted to identify and manipulate the factors that influence road traffic accidents. Factors such as physical disabilities, mental disabilities, insufficient driving ability, alcohol drunkenness, inattention to traffic signage and signposts, drug misuse, and lack of focus, using a phone while driving are a few factors identified in the literature. Further, environmental factors such as bad weather and inadequate lighting, road slope, surface, curve, speed limit, intersection kinds, car models, vehicle make, and technical problems have also been identified as potential factors that could cause and impact the severity of road traffic accidents [13-15].

Bucsuházy et al. [3] researched human behavior as factors that may contribute to accidents or have an impact on their causes. The goal of their research was to look into the most common causes of traffic accidents among certain risk groups, such as young drivers. The statistical analysis of Pearson's chi-squared test on the causes of traffic accidents examined age, gender, road familiarity, annual miles, driving behaviors, and a tendency for risky behavior. Following the analysis, the study concluded that two main factors, namely inexperience and speeding or driving without adapting, can cause traffic accidents involving young drivers. Kasahani et al. [16] examined some of the factors that affect drivers that result in severe injuries and identified weariness as a major factor.

The quest to reduce the likelihood and severity of road traffic accidents necessitates cross-disciplinary research, hence the diverse investigation into the various strategies to prevent road accidents and reduce their severity [17]. For instance, Internet of Things (IoT) technology has been recognized for its potential to reduce the impacts of road accidents provided that the accidents are reported to the right rescue quarters and fast actions are taken. Likewise, machine learning models such as the Bayesian networks [18] have been leveraged to determine the association between car makes and the likelihood of road accidents. The literature has also echoed that the decision tree model is more dependable and has a consistent rapid response rate [13, 19]. The tree-based regression model was used to analyze the geometry of road accidents on two or more lane highway rates of injuries and severity and to choose features that are relevant to the risk of traffic accidents. This model has proven to be beneficial in reducing the likelihood of biases from various control factors, warning signs, and speed limits on highway and railway crossings [20].

A multi-task DNN architecture was proposed by Yang et al. [21] for predicting various accident severity levels of injury, death, and property loss. A complete and accurate analysis of the severity of traffic accidents is made possible by the multi-task and deep learning design. By layer-wise relevance propagation, which creates explanations based on the structure and weights of DNN, their framework was able to pinpoint the important components that contribute to the three types of traffic accident severity. Our suggested model accurately forecasts the severity risks of road accidents based on trials done with Chinese traffic accident data. Furthermore, Zhang et al. [22] used the Boruta Algorithm (BA) feature selection technique in conjunction with a Random Forests (RF) classifier to identify the crucial characteristics that affect injury severity. The four classifiers Naive Bayes (NB), KNearest Neighbor (K-NN), Binary Logistic Regression (BLR), and Extreme Gradient Boosting (XGBoost) were then given the influential features to effectively predict injury severity. The vehicle type, the month of the year, the driver's age, and the alignment of the road segment were found to have the most influence on BA's experimental examination. It was discovered that the gender of the driver, the presence of a median, and the presence of a shoulder were all irrelevant.

Likewise, Assi et al. [23] introduced four machine learning models to predict accident severity: feed-forward neural networks (FNN), support vector machines (SVM), fuzzy C-means clustering-based feed-forward neural network (FNN-FCM), and fuzzy c-means based support vector machines (SVM-FCM). The models' injury severity prediction accuracy, sensitivity, precision, and harmonic mean of sensitivity and precision were all assessed. In terms of accuracy and F1 score, the SVM-FCM model outperformed the other generated models for predicting accident severity. Numerous studies [24-26] have investigated the efficiency of random forest in forecasting the severity of road accidents. To compare the performance of classical and ensemble mode machine learning models, Ahmed et al. [27] used well-known assessment metrics such as prediction accuracy, precision, recall, F1 score, and area under the receiver operator characteristic. Random Forest outperformed other methods such as logistic regression (LR), K-nearest neighbor (KNN), naive Bayes (NB), extreme gradient boosting (XGBoost), and adaptive boosting (AdaBoost). Some studies examined and used machine

learning models to predict the severity of traffic accidents, as shown in Table 1.

Table 1. Summary of accident severity using machine learning models

Author	Machine learning algorithm	Data description	Model prediction accuracy
Sameen et al. [28]	RNN	N=1,130 (2009-2015)	73.7%
Zhang et al. [24]	Random Forest	N=5,538	53.9%
AlMamlook et al. [25]	Random Forest	N=271,563 (2010-2016)	75.5%
Labib et al. [29]	AdaBoost	N=43, 089 (2001-2015)	80%
Wahab and Jiang [30]	Random Forest	N=8,516 (2011-2015)	73.91%
Fiorentini and Losa [31]	Logistic Regression	N=6,515 (2005-2018)	85.74%
Assi et al. [23]	SVM-FCM	N=10,000 (2011-2016)	74%
Komol et al. [26]	Random Forest	N=21,158 (2013-2019)	72.3%
Ahmed et al. [27]	Random Forest	N=13775 (2016-2020)	86.8%

This study advances the body of existing knowledge to examine significant features that contribute to accident severity by comparing the Naive Bayes, KNN, Logistic model tree, Decision Tree, Random Tree, and Logistic Regression machine learning models to the J48 pruned tree model.

3. MATERIAL AND METHODS

This section includes an explanation of the machine learning algorithms, the adapted framework for the study as shown in Figure 1, and evaluation metrics. The experiment was carried out on a machine running Windows 10 with an Intel(R) Core (TM) i7-8650U CPU running at 1.90GHz (8 CPUs), 2.1GHz, 8GB of RAM, and a 500GB hard drive.

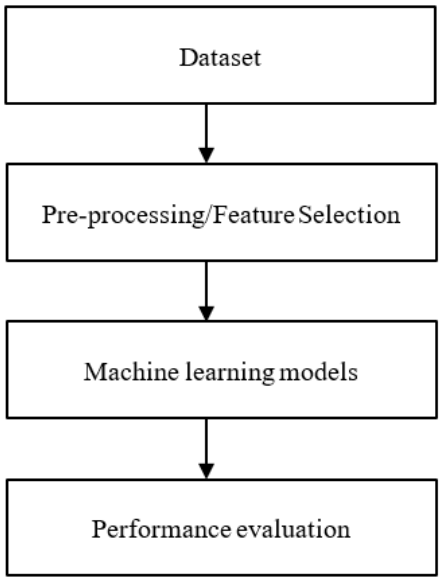

Figure 1. The framework for the road traffic accident using machine learning models

3.1 Dataset

The data set was prepared from manual records of road traffic accidents for the year 2017-2020 https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents [32]. Each record contains explanatory variables called features that provide information about the accident. Among other information, the manual records contain the driving experience of the driver, the age of the driver, the service year of the vehicle, road surface conditions, and weather conditions. In the current study, these variables were analyzed to provide information about accident severity (fatal injury, serious injury, and slight injury), which is the target variable. All the sensitive information, such as the driver's name, has been excluded during data encoding and finally, the data set contains 32 features and 12316 instances of road accidents in an excel format, where each column contains a feature, and each row contains an instance of the road accidents.

3.2 Feature selection and data pre-processing

Table 2. Description of the selected features

	Features	Description
1	Day of week	The day of the week the accident occurred
2	Age band of driver	The age range of the accident-causing driver
3	Sex of driver	This identifies the driver's gender
4	Educational level	The degree to which the accident-causing driver was educated
5	Vehicle driver relation	Indicates if the driver is an employee, an owner, or someone else
6	Driving experience	The driving expertise of the accident-causing driver
7	Type of vehicle	The type of vehicle the accident-causing driver was driving
8	Owner of the vehicle	This describes the title holder of the vehicle
9	Service year of the vehicle	This reveals when the accident-related vehicle was last serviced
10	Defect of vehicle	This reveals the defects of the vehicle that existed before the collision
11	Area accident occurred	This identifies the place where the accident occurred
12	Lanes or medians	The kind of lane the car was traveling in when the accident occurred
13	Road alignment	This defines the terrain of the road where the accident occurred
14	Type of Junction	The kind of road intersection where the accident occurred
15	Road surface type	Identifies the type of road where the accident occurred such as asphalt, dirt, or another.
16	Road surface conditions	This describes the state of the road's surface where the accident happened
17	Light conditions	The light condition at the time the accident occurred
18	Weather conditions	The weather conditions at the time of the accident
19	Type of collision	This displays the nature of the collision of the vehicles
20	Number of vehicles involved	This indicates how many vehicles were involved in the collision
21	Number of casualties	The number of accident-related fatalities
22	Vehicle movement	The driver's behavior before the vehicle collision
23	Cause of accident	The identifies the reason why the car accident occurred

A dataset may contain features that do not provide relevant information about the target variable hence, analyzing such features can produce misleading results Features selection is the process of deleting unimportant features from the data set and choosing the important ones. Only 23 of the 32 features in the road accident data set, which are shown in Table 2, were shown to be useful for predicting accident severity. The choice was supported by preparatory research and subject-matter expertise.

Further to feature selection, data pre-processing was applied to increase the quality of the selected features. Data pre-processing includes i) handling missing values. For example, replacing missing numeric values in a column with the mean of the values in the same column. ii) Feature scaling: For example, the mean of each numerical column is subtracted from each value in the column, and the result is divided by the standard deviation of that column, and iii) Removing the outliers: that is, removing values outside of a specific range when compared to other values in the same column. In the current study, 23 of the chosen features have missing values that have been replaced with unknown ones. The missing data were either not accessible at the time of data collection or there was no information available to fill it in.

3.3 Machine learning algorithms

This section provides a brief description of the selected well-known machine learning algorithms which will be compared based on their performance in classifying accident severity. These algorithms are Naïve Bayes, Bagging, K-Nearest Neighbours, Logistic Model Tree, Decision Tree, Random Tree, Logistic Regression, and J48 pruned tree.

Naïve Bayes

Naïve Bayes classifier is one of the well-known machine learning algorithms that perform well in vast areas of application [33]. Naïve Bayes classifier assumes that k classes $C_1, C_2, \dots\ C_k$ to be predicted are conditionally independent and each class is represented by an n-dimensional vector $X = x_1, x_2, \dots, x_n$ of features. This assumption implies that there are no dependencies among features. Thus, the assumption simplifies the Naïve Bayes computations as shown in Eq. (1)

$$P(X) = \frac{P(X)\ P(C_i)}{P(X)} \tag{1}$$

Furthermore, Nave Bayes is known to demand fewer processing resources and a little amount of training data to estimate the required parameter in the classification process.

Bagging

Bagging is one of the ensembles approaches that combines the performance of several models, trained on randomly sampled data from a given data set. The sampling is done with replacement. The idea is to improve the accuracy of a final model by combining the vote (that is, counting the output) of different models trained on the subset of the data set. In the basic bagging approach, the models are trained in parallel [34]. The process of training each model on a subset of the data set and combining the voting result is called bootstrap aggregating popularly referred to as bagging [35].

K-Nearest Neighbor

K-Nearest neighbor (KNN) [36] relies on the similarity or distance measure to compute the class of a given instance of a data set. K-nearest neighbor makes use of decision rules that provide a nonparametric means of assigning data instances to a class label, based on the k-classes closely related or close to the given instance.

Logistic Model Tree

The Logistic Model Tree (LMT) is a popular classification model. LMT is a machine learning algorithm that combines the classification algorithms Decision Tree (DT) and Logistic Regression (LR). One such system that learns decision-tree classifiers is C4.5. In practice, LMT produces more accurate results than related algorithms like C4.5 and CART. Unlike decision trees, leaf nodes on LMT feature a logistic regression function of linked attributes in addition to just being class labels. The regression function considers a subset of all data attributes [37].

Decision Tree

The decision tree [38] is a commonly used machine learning algorithm that uses a rule-based tree structure to split data into predefined classes. The decision rules for splitting the data are based on the characteristics and classification of the data set. A decision tree learns these rules, deduced from the data set, and predicts the value or class of the target variable by applying the rules to a given instance of the data set belonging to the class. There are several algorithms for decision tree generation. These include Iterative Dichotomiser 3 (ID3) [39], C4.5 [35], and Classification and Regression Tree (CART) [40] which has been recently applied for classifying accident fatality.

Random Forest

Random forest uses the bagging approach by combining the performance of several decision trees. The idea is to improve the accuracy of the random forest model by combining the accuracy of several decision trees. The mean or median of the outputs of the decision trees is taken as the output of a random forest where the value of the target variable is continuous, while the mode is used where the target variable is a discrete class label [31].

Given that a random forest is made of M decision trees, the output $f_{rf}^{M}(X^I)$ of the random forest that takes the ith instance of a data set $X = x_1, x_2, \dots, x_n$ of n-dimensional features is given by:

$$\begin{aligned} &F_{rf}^{M}(X^i) = y = \{\frac{1}{M}\sum_{m=1}^{M} T_{ree(X^i)}\,, \\ &y = continuous\ value\ max\ (count\left(T_{ree(X^i)}\right)), \\ &y = a\ discrete\ class\ label \end{aligned} \tag{2}$$

where, $T_{ree(X^i)}$ is the output of a decision tree when evaluated on the ith instance of X. The function $count$ returns the number of times each class label was predicted by the decision trees while max returns the maximum of a given set of values.

Logistic Regression

Logistic regression is one of the simplest classification approaches which is often used as a baseline model for comparing the performance of classification algorithms. It makes use of the logit function to map the features to the independent variables. Logistic regression is defined by the equation [41]:

$$F(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(b_1x_1+\ b_2x_2+\cdots+\ b_nx_n)}} \tag{3}$$

where, $z = \hat{y} = b_1x_1 + b_2x_2 + \cdots + b_nx_n$ is the linear regression of the output variable on the features.

J48 Pruned Tree

The J48 pruned tree is a variant of the C4.5 algorithm, it is based on tree pruning and supports two methods of pruning, namely, subtree replacement and subtree raising. In subtree replacements, the nodes in the decision tree are replaced with the leaf nodes [42], while the subtree raising involves selecting and moving a subtree upwards in the tree such that the subtree is closer to the root. J48 uses a greedy search strategy, like the C4.5 algorithm, to build decision trees, and it permits tweaking various parameters to improve classification accuracy [43]. A significant benefit of the chosen model is that missing values in the dataset have little to no effect on how a J48 pruned tree is built.

3.4 Performance evaluation

In the experimental comparison of classifiers, the standard performance measures Accuracy, Precision, Recall, F1-Measure, MCC, Roc Area, PRC Area, processing time, Kappa statistic, and mean absolute error were computed [44-46]. To determine how well a classification method performed, a confusion matrix was also plotted.

4. RESULTS AND DISCUSSIONS

This study benchmarked the performance of the J48 pruned tree with seven other machine learning models, including Naive Bayes, Bagging, KNN, Logistic model tree, Decision Tree, Random Tree, and Logistic Regression, in terms of 10 key metrics, including Accuracy, Precision, Recall, F1-Measure, MCC, Roc Area, PRC Area, processing time, Kappa statistic, and mean absolute error, to determine the model that produces predictions with the highest accuracy. The study used 10-fold cross-validation to confirm the findings and prevent overfitting throughout the training phase [31]. In each iteration, a separate fold of the data is kept back for validation while the other 9 folds are utilized for learning. The trained models are then applied to the data in the validation fold to generate predictions. Table 3 shows seven well-known performance evaluation metrics, such as accuracy, recall, F1-measure, Matthews Correlation Coefficient (MCC), Roc Area, and PRC Area. The J48 pruned tree model is the base model with the best performance on the road traffic accidents dataset.

Table 3. The experimental results for the severity of the road traffic accident

	Accuracy	Precision	Recall	F1-Measure	MCC	Roc Area	PRC Area
J48 pruned tree	85.47	0.848	0.855	0.799	0.225	0.612	0.787
Naive Bayes	83.53	0.762	0.835	0.774	0.016	0.610	0.778
Bagging	84.29	0.77	0.843	0.781	0.069	0.577	0.765
KNN	77.58	0.749	0.776	0.761	0.049	0.531	0.747
Logistic model tree	84.74	0.810	0.847	0.788	0.132	0.601	0.782
Decision Tree	84.52	0.793	0.845	0.787	0.114	0.599	0.775
Random Tree	84.51	0.789	0.845	0.784	0.094	0.592	0.773
Logistic Regression	84.51	0.772	0.845	0.775	0.013	0.593	0.771

Table 4 demonstrates that the J48 pruned tree model outperforms the other seven classical models in terms of prediction time. Furthermore, according to the value of Kappa statistics and mean absolute error, the results clearly show that the J48 pruned tree model outperformed the other seven classical models

Table 4. Prediction time, Kappa statistics and mean absolute error calculations

	Time	Kappa statistic	Mean absolute error
J48 pruned tree	0.10	0.1192	0.1615
Naive Bayes	0.03	0.0092	0.1724
Bagging	1.31	0.032	0.1692
KNN	0.10	0.0487	0.1655
Logistic model tree	63.24	0.0636	0.1929
Decision Tree	0.12	0.0608	0.1670
Random Tree	0.33	0.0431	0.1697
Logistic Regression	31.44	0.0023	0.1719

A machine learning model is analysed using a confusion matrix. The statistics related to true positives, false negatives, false positives, and true negatives are reflected [41]. The confusion matrix for the J48 pruned dataset is shown in Figure 2.

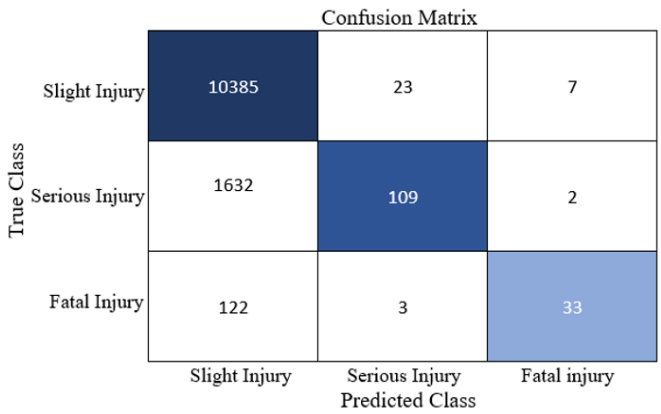

Figure 2. Confusion matrix of the road traffic accident dataset

Based on the confusion matrix of the road traffic accident dataset, the following conclusions are inferred:

i. 10385 are classified as true and have been correctly predicted as a slight injury

ii. 109 are classified as true and have been correctly predicted as a serious injury

iii. 33 are classified as true and have been correctly predicted as a fatal injury

Furthermore, out of 23 selected criteria from the RTA dataset, the top four factors that determine the severity of a road traffic accident are the number of casualties, the number of vehicles involved, the weather conditions, and the lighting conditions. J48 pruned tree classifier was used to determine significant factors that led to different severity of road traffic accidents. According to the studies [47, 48], weather condition is a key factor in determining the severity of accidents, furthermore, the study [49] concur with our study that lightning conditions are a significant factor in accident severity.

5. CONCLUSION

Over time, an increase in the number of individuals purchasing vehicles has increased the frequency and severity of road traffic accidents. This has grown to be a significant problem that requires the attention of traffic law enforcement organizations. This study used the J48 pruned trees to analyze and predict the factors that influence the severity of road traffic accidents. The J48 pruned tree was compared to Naive Bayes, Bagging, KNN, Logistic model tree, Decision tree, Random tree, and Logistic regression. The WEKA data mining tool environment was leveraged for comparison of the various machine learning models as well as for experimentation. The dataset used to train and test the machine learning models contained 23 features. The testing was completed using the trained dataset through the 10-fold cross-validation. The J48 pruned tree model outperformed the other seven traditional machine learning models, based on performance findings utilizing ten popular evaluation metrics. Finally, this study was able to identify significant factors that influence the severity of road traffic accidents. These include the number of casualties, the number of vehicles involved in the accident, the weather conditions, and the lighting conditions of the road. This study will inform road traffic agencies on influencing factors to mitigate for improving traffic safety, as well as raise awareness of the significant factors causing accident severity in communities.

REFERENCES

[1] World Health Organization. (2022). Report on road traffic injuries. https://www.who.int/health-topics/road-safety#tab=tab_1, accessed on March 1, 2023.

[2] Khan, K., Zaidi, S.B., Ali, A. (2020). Evaluating the nature of distractive driving factors towards road traffic accident. Civil Engineering Journal, 6(8): 1555-1580. http://dx.doi.org/10.28991/cej-2020-03091567

[3] Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia, 45: 555-561. https://doi.org/10.1016/j.trpro.2020.03.057

[4] Gutierrez-Osorio, C., Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. Journal of Traffic and Transportation Engineering (English edition), 7(4): 432-446. https://doi.org/10.1016/j.jtte.2020.05.002

[5] Cao, G., Michelini, J., Grigoriadis, K., Ebrahimi, B., Franchek, M.A. (2016). Cluster-based correlation of severe driving events with time and location. Journal of Intelligent Transportation Systems, 20(6): 516-531. https://doi.org/10.1080/15472450.2016.1152891

[6] Fawcett, L., Thorpe, N., Matthews, J., Kremer, K. (2017). A novel bayesian hierarchical model for road safety hotspot prediction. Accident Analysis & Prevention, 99: 262-271. https://doi.org/10.1016/j.aap.2016.11.021

[7] Chen, Q., Song, X., Yamada, H., Shibasaki, R. (2016). Learning deep representation from big and heterogeneous data for traffic accident inference. In Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). https://doi.org/10.1609/aaai.v30i1.10011

[8] Lu, W.Q., Luo, D.Y., Yan, M.H. (2017). A model of traffic accident prediction based on convolutional neural network. In 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE), pp. 198-202. https://doi.org/10.1109/ICITE.2017.8056908

[9] You, J., Wang, J., Guo, J. (2017). Real-time crash prediction on freeways using data mining and emerging techniques. Journal of Modern Transportation, 25: 116-123. https://doi.org/10.1007/s40534-017-0129-7

[10] Kaplan, S., Prato, C.G. (2013). Cyclist-motorist crash patterns in Denmark: A latent class clustering approach. Traffic Injury Prevention, 14(7): 725-733. https://doi.org/10.1080/15389588.2012.759654

[11] Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., Wang, Z. (2019). Traffic accident's severity prediction: A deep-learning approach-based CNN network. IEEE Access, 7: 39897-39910. https://doi.org/10.1109/ACCESS.2019.2903319

[12] Rezaie Moghaddam, F., Afandizadeh, S., Ziyadi, M. (2011). Prediction of accident severity using artificial neural networks. International Journal of Civil Engineering, 9(1): 41-48.

[13] Momeni Kho, S., Pahlavani, P., Bigdeli, B. (2022). Analyzing and predicting fatal road traffic crash severity using tree-based classification algorithms. International Journal of Transportation Engineering, 9(3): 635-652.

[14] Wen, H., Du, Y., Chen, Z., Zhao, S. (2022). Analysis of factors contributing to the injury severity of overloaded-truck-related crashes on mountainous highways in China. International Journal of Environmental Research and Public Health, 19(7): 4244. https://doi.org/10.3390/ijerph19074244

[15] Alkhoori, F.A., Maghelal, P.K. (2021). Regulating the overloading of heavy commercial vehicles: Assessment of land transport operators in Abu Dhabi. Transportation Research Part A: Policy and Practice, 154: 287-299. https://doi.org/10.1016/j.tra.2021.10.019

[16] Kashani, A.T., Moghadam, M.R., Amirifar, S. (2022). Factors affecting driver injury severity in fatigue and drowsiness accidents: A data mining framework. Journal of Injury and Violence Research, 14(1): 75. https://doi.org/10.5249%2Fjivr.v14i1.1679

[17] Kumar, N., Lohani, D., Acharya, D. (2022). Vehicle accident sub-classification modeling using stacked generalization: A multisensor fusion approach. Future Generation Computer Systems, 133: 39-52. https://doi.org/10.1016/j.future.2022.03.005

[18] Chen, T., Wong, Y.D., Shi, X., Wang, X. (2022). Optimized structure learning of bayesian network for investigating causation of vehicles' on-road crashes. Reliability Engineering & System Safety, 224: 108527. https://doi.org/10.1016/j.ress.2022.108527

[19] Ahmed, A.M., Rizaner, A., Ulusoy, A.H. (2018). A novel decision tree classification based on post-pruning with bayes minimum risk. Plos one, 13(4): e0194168. https://doi.org/10.1371/journal.pone.0194168

[20] Wang, D., Liu, Q., Ma, L., Zhang, Y., Cong, H. (2019). Road traffic accident severity analysis: A census-based study in China. Journal of Safety Research, 70: 135-147. https://doi.org/10.1016/j.jsr.2019.06.002

[21] Yang, Z., Zhang, W., Feng, J. (2022). Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. Safety science, 146: 105522. https://doi.org/10.1016/j.ssci.2021.105522

[22] Zhang, S., Khattak, A., Matara, C.M., Hussain, A., Farooq, A. (2022). Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents. PLoS one, 17(2): e0262941. https://doi.org/10.1371/journal.pone.0262941

[23] Assi, K., Rahman, S.M., Mansoor, U., Ratrout, N. (2020). Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. International Journal of Environmental Research and Public Health, 17(15): 5497. https://doi.org/10.3390/ijerph17155497

[24] Zhang, J., Li, Z., Pu, Z., Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access, 6: 60079-60087. https://doi.org/10.1109/ACCESS.2018.2874979

[25] AlMamlook, R.E., Kwayu, K.M., Alkasisbeh, M.R., Frefer, A.A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 272-276. https://doi.org/10.1109/JEEIT.2019.8717393

[26] Komol, M.M.R., Hasan, M.M., Elhenawy, M., Yasmin, S., Masoud, M., Rakotonirainy, A. (2021). Crash severity analysis of vulnerable road users using machine learning. PLoS One, 16(8): e0255828. https://doi.org/10.1371/journal.pone.0255828

[27] Ahmed, S., Hossain, M.A., Bhuiyan, M.M.I., Ray, S.K. (2021). A comparative study of machine learning algorithms to predict road accident severity. In 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), London, United Kingdom, pp. 390-397. https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

[28] Sameen, M.I., Pradhan, B., Shafri, H.Z.M., Hamid, H.B. (2019). Applications of deep learning in severity prediction of traffic accidents. In GCEC 2017: Proceedings of the 1st Global Civil Engineering Conference, Kuala Lumpur, Malaysia, pp. 793-808. https://doi.org/10.1007/978-981-10-8016-6_58

[29] Labib, M.F., Rifat, A.S., Hossain, M.M., Das, A.K., Nawrine, F. (2019). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, pp. 1-5. https://doi.org/10.1109/ICSCC.2019.8843640

[30] Wahab, L., Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS One, 14(4): e0214966. https://doi.org/10.1371/journal.pone.0214966

[31] Fiorentini, N., Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7): 61. https://doi.org/10.3390/infrastructures5070061

[32] Bedane, T.T., Assefa, B.G., Mohapatra, S.K. (2021). Preventing traffic accidents through machine learning predictive models. In 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, pp. 36-41. https://doi.org/10.1109/ICT4DA53266.2021.9672249

[33] Budiawan, W., Saptadi, S., Tjioe, C., Phommachak, T. (2019). Traffic accident severity prediction using naive bayes algorithm-a case study of Semarang toll road. In IOP Conference Series: Materials Science and Engineering. IOP Publishing, 598(1): 012089. https://doi.org/10.1088/1757-899X/598/1/012089

[34] Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science, 14: 241-258. https://doi.org/10.1007/s11704-019-8208-z

[35] González, S., García, S., Del Ser, J., Rokach, L., Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion, 64: 205-237. https://doi.org/10.1016/j.inffus.2020.07.007

[36] Cunningham, P., Delany, S.J. (2021). K-nearest neighbour classifiers-a tutorial. ACM Computing Surveys (CSUR), 54(6): 1-25. https://doi.org/10.1145/3459665

[37] Maulana, M.F., Defriani, M. (2020). Logistic model tree and decision tree J48 algorithms for predicting the length of study period. PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic, 8(1): 39-48. https://doi.org/10.33558/piksel.v8i1.2018

[38] Mqadi, N., Naicker, N., Adeliyi, T. (2021). A SMOTe based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection. International Journal of Computing and Digital Systems, 10(1): 277-286. http://dx.doi.org/10.12785/ijcds/100128

[39] Zhang, H., Zhou, R. (2017). The analysis and optimization of decision tree based on ID3 algorithm. In 2017 9th International Conference on Modelling, Identification and Control (ICMIC), Kunming, China, pp. 924-928. https://doi.org/10.1109/ICMIC.2017.8321588

[40] Abbasi, E., Li, Y., Wu, X., Craig, B. (2021). Using classification and regression trees (CART) to identify factors contributing to vehicle crash severity in a port city. International Journal of Transportation Systems, 6.

[41] Hansrajh, A., Adeliyi, T.T., Wing, J. (2021). Detection of online fake news using blending ensemble learning. Scientific Programming, 2021: 3434458. https://doi.org/10.1155/2021/3434458

[42] Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H. (2012). A comparative study of reduced error pruning method in decision tree algorithms. In 2012 IEEE International Conference on Control System, Computing and Engineering, Penang, Malaysia, pp. 392-397. https://doi.org/10.1109/ICCSCE.2012.6487177

[43] Jankovic, R. (2019). Classifying cultural heritage images by using decision tree classifiers in WEKA. In Proceedings of the 1st International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding Co-Located with 15th Italian Research Conference on Digital Libraries (IRCDL 2019), Pisa, Italy, pp. 119-127.

[44] Naicker, N., Adeliyi, T., Wing, J. (2020). Linear support vector machines for prediction of student performance in school-based education. Mathematical Problems in Engineering, 2020: 4761468. https://doi.org/10.1155/2020/4761468

[45] Mqadi, N.M., Naicker, N., Adeliyi, T. (2021). Solving misclassification of the credit card imbalance problem using near miss. Mathematical Problems in Engineering, 2021: 7194728. https://doi.org/10.1155/2021/7194728

[46] Bharati, S., Rahman, M.A., Podder, P. (2018). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, pp. 581-584. https://doi.org/10.1109/CEEICT.2018.8628084

[47] Al-Harbi, M., Yassin, M.F., Bin Shams, M. (2012). Stochastic modeling of the impact of meteorological conditions on road traffic accidents. Stochastic Environmental Research and Risk Assessment, 26: 739-750. https://doi.org/10.1007/s00477-012-0584-y

[48] Drosu, A., Cofaru, C., Popescu, M.V. (2020). Influence of weather conditions on fatal road accidents on highways and urban and rural roads in romania. International Journal of Automotive Technology, 21: 309-317. https://doi.org/10.1007/s12239-020-0029-4

[49] Ijaz, M., Zahid, M., Jamal, A. (2021). A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accident Analysis & Prevention, 154: 106094. https://doi.org/10.1016/j.aap.2021.106094