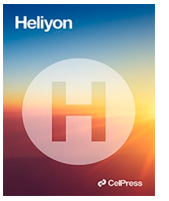
Heliyon 9 (2023) e18812

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents

Izuchukwu Chukwuma Obasi *, Chizubem Benson

European University Cyprus, Center for Risk and Safety in the Environment, Diogenis 6, Engomi, Nicosia, 2404, Cyprus

A B S T R A C T

Traffic accidents pose a significant public safety concern, leading to numerous injuries and fatalities worldwide. Predicting the severity of these accidents is crucial for developing effective road safety measures and reducing casualties. This paper proposes an analytic framework that utilizes machine learning models, including Naive Bayes, Random Forest, Logistic Regression, and Artificial Neural Networks, to predict the severity of traffic accidents based on contributing factors. This study analyzed ten years of UK traffic accident data (2005–2014, N = 2,047,256) to develop and compare different ML models. Results show that the proposed Random Forest and Logistic Regression models achieved an 87% overall prediction accuracy, outperforming Naive Bayes (80%) and Artificial Neural Networks (80%). By employing Random Forest-based feature importance analysis, the study identified Engine Capacity, Age of the vehicle, make of vehicle, Age of the driver, vehicle manoeuvre, daytime, and 1st road class as the most sensitive variables influencing traffic accident severity prediction. Additionally, the suggested RF model outperformed most existing models, attaining a remarkable overall accuracy and superior predictive performance across various injury severity classes. The findings have significant implications for developing efficient road safety measures and enhancing the current traffic safety system. The proposed framework and models can be adapted to various datasets to achieve accurate and effective predictions of traffic accident severity, serving as a valuable reference for implementing traffic accident management and control measures. Future research could extend the proposed framework to datasets containing Casualty Accident information to further improve the accuracy of injury severity prediction.

1. Introduction

Road traffic accidents are a significant cause of fatalities and injuries among individuals. According to the [1], 20–50 million non-fatal injuries and 1.35 million traffic fatalities yearly are attributed to road traffic accidents. As of 2018, road accident injuries had become the leading cause of death for individuals aged 5 to 29. Approximately 93% of global traffic-related fatalities were concentrated in low- and middle-income countries [1]. Among the many groups of affected pedestrians, road users, motorcyclists, and cyclists account for the highest percentages of injuries and fatalities. Individuals in this group of road users are often at a higher risk and are commonly known as vulnerable road users (VRUs) [2].

Given the diverse range of injuries in traffic accidents, injury severity is a vital indicator regularly assessed to gauge road safety performance [3]. Better knowledge of the factors influencing accident injury severity is essential in deploying proactive mitigation methods. Human behaviour, the environment, road conditions, traffic, vehicle characteristics, and crash circumstances influence the severity of injuries resulting from a crash. However, since traffic accidents are random and can vary in severity from one location to another, it is important to conduct a local investigation of the risk factors to predict crashes and implement appropriate preventive measures accurately.

Several statistical-based regression techniques have extensively been used in the literature to model the severity of crash injuries.

* Corresponding author.
E-mail address: io182307@students.euc.ac.cy (I.C. Obasi).

https://doi.org/10.1016/j.heliyon.2023.e18812
Received 22 April 2023; Received in revised form 27 July 2023; Accepted 28 July 2023
Available online 29 July 2023
2405-8440/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Some studies concentrated on using conventional data analysis techniques [4–6]. Even though they can be well-interpreted mathematically and help to understand the function of specific predictor variables better, statistical models can have some limitations, such as poor prediction accuracies, they might produce biased model estimation, etc [7]. ML techniques have recently been introduced as an alternative to traditional data analysis methods to investigate factors associated with injury severity. It enables the extraction of important knowledge from significant amounts of complicated and heterogeneous data. The most widely used machine learning techniques are decision trees (DT) [8], Bayesian networks [9], classification and regression trees (CART) [10], extreme gradient boosting (XGBoost) [9], and random forest (RF) [11].

Machine learning algorithms can discover intricate correlations between the many factors influencing injury severity and make highly accurate predictions. The stacked sparse autoencoder (SSAE) was employed as a deep learning algorithm to forecast injury severity in traffic accidents using the contributing factors [12]. Also [7], compares the eXtreme Gradient Boosting (XGBoost) model to a few conventional ML techniques to determine the severity of crash injuries. Furthermore, a multi-task deep neural network (DNN) framework was proposed by Ref. [9], capable of simultaneously predicting the severity of traffic accident injury, fatality, and property loss. However, the application of machine learning to injury severity prediction also presents several challenges, including limited data availability, which affects the model's performance, data overfitting that can lead to poor performance and reduced accuracy in traffic injury severity predictions, and feature selection challenges.

This research addresses these challenges by examining the current state-of-the-art in predicting road accident injury severity through ML. Furthermore, it seeks to shed light on the primary difficulties and limitations encountered in this field. This study will present an in-depth analysis of the current ML models for predicting injury severity, including a review of the data, features, algorithms, and assessment metrics employed in these models. The paper will also thoroughly review machine learning methods' shortcomings in predicting injury severity. It will emphasize the requirement for ongoing improvements to the precision and dependability of these models.

These are a summary of the paper's contributions.

- The performance of four significant ML algorithms for modelling crash injury severity was assessed, including Naïve Bayes, Artificial Neural Network, Random Forest, and logistic regression.
- We employed the random forest classifier feature importance method to analyse the sensitivity of predictor variables.
- To evaluate our model's predictive abilities across all categories of injury severity and with existing literature in the field, we present several performance statistics.

The remaining sections of this paper are structured as follows. Section 2 presents the proposed methodology and an overview of the dataset containing information on severe traffic injuries. In section 3, the results from the different machine learning models are presented, analyzed, and compared. Section 4 discusses the prediction methods' performance and the classification task's impact on injury severity prediction. Section 5 concludes the paper with the potential applications of the findings in the road safety field.

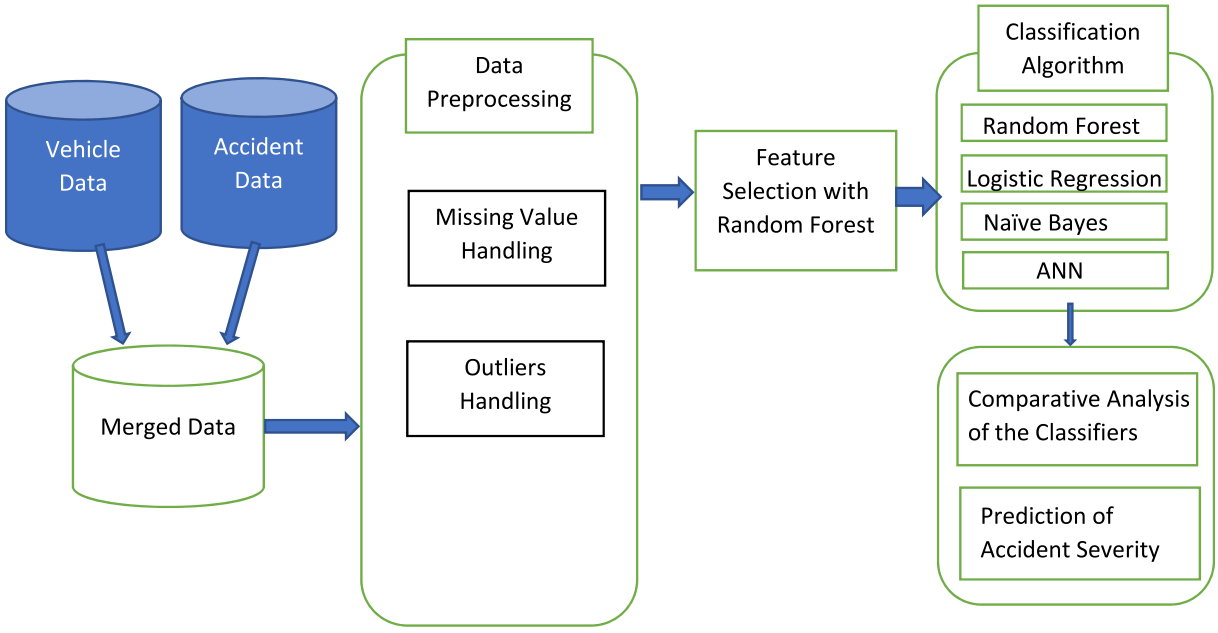


Fig. 1. Framework for the prediction of traffic accident severity.

2. Research methodology

2.1. Overview

In this study, we utilized four distinct classification algorithms to predict the severity of injuries in traffic accidents. These ML classifiers, based on supervised training algorithms, were utilized to classify the datasets. These classifiers have demonstrated promising results owing to their capability to handle multi-dimensional data, adaptability in implementation, versatility, and strong predictive abilities. This research focused on the target variable, which represents the traffic accident severity level and has two possible outcomes: serious and non-serious. We implemented four algorithms to achieve our objectives: Random Forest, Naive Bayes, Logistic Regression, and Artificial Neural Networks.

As illustrated in Fig. 1, the initial step involved collecting a substantial amount of traffic accident records (Accident and Vehicle datasets) from publicly available open-source websites. Second, the datasets are merged. Additionally, the data is cleaned, and descriptive statistical analysis is conducted on the clean data. Then, we employed the random forest classifier feature importance method to analyse the sensitivity of predictor variables. Furthermore, predictive modelling was used to classify and predict the severity of traffic accidents by applying four ML algorithms: Random Forest, Naïve Bayes, Logistic Regression, and Artificial Neural Network.

2.2. Data collection

This research used a dataset from the UK traffic accident records from 2005 to 2014 obtained from a publicly available website. This website disseminates traffic information gathered by several organizations, including the UK Departments of Transportation, traffic cameras, law enforcement organizations, and traffic sensors embedded in road networks. The data collected was from two databases containing the accident and vehicle information. The accident information database provides detailed road safety data regarding the factors contributing to personal injury in road accidents in the United Kingdom between 2005 and 2014. The vehicle information database contains information about the vehicle involved in the road traffic accident from 2005 to 2014, such as the age of the vehicle, sex of the driver, vehicle type, age of the driver, etc.

2.3. Data pre-processing

Data pre-processing plays a vital role in the data mining process. A dataset must be pre-processed or cleaned before any analysis is performed, otherwise it produces subpar, inaccurate, or misleading results because of outliers, redundant values, or missing values [13]. The dataset exhibited missing values in certain attributes from the accident and vehicle datasets. The accident dataset had a total of 0.495% missing values, while the vehicle dataset had a higher proportion at 0.938%. Specifically, the attributes '2nd_Road_Class', 'LSOA_of_Accident_Location', '2nd_Road_Number', and 'Pedestrian_Crossing-Physical_Facilities' contained missing values of 41.23%, 7.08%, 0.85%, and 0.17%, respectively. Additionally, the attributes 'Did_Police_Officer_Attend_Scene_of_Accident', 'Longitude', 'Latitude', and 'Location_Northing_OSGR' had less than 0.01% missing values for each attribute. To address these missing values, the top four variables with the highest percentage of empty columns. Furthermore, the StandardScaler function was employed to standardize the data. The date column also required formatting, as some values were not stored in the correct format. The column was rearranged to represent the date in the Year, Month, and Day format. Additionally, the hours were categorized into specific time groups, such as the 'morning rush' between 05:00–10:00, to enhance data analysis.

2.3.1. Data merging

Data merging is a critical phase in the data preparation process. This process entails combining many datasets into a single coherent dataset for analysis. The merging process may involve identifying common identifiers such as unique identification numbers, dates, or geographic locations to match the datasets. Once the datasets are merged, the data cleaning process can proceed, which may include imputing missing values, standardizing variables, and addressing outliers. In this study, two datasets were merged, and the variable "accident index" was identified as the common identifier for both datasets.

2.3.2. Data cleaning and handling

Data cleaning involves recognizing and addressing parts of incomplete, incorrect, inaccurate, or irrelevant data. This process typically involves modifying, replacing, or deleting problematic data from a database, record set, or table. A significant amount of missing and erroneous data was recorded in the dataset used for this study. First, the problem has been solved by applying the appropriate missing data imputation approach to the two datasets. Then, both datasets were merged to become a single dataset. Furthermore, numerical and categorical data handling is conducted, and new features are selected afterwards.

2.4. Feature selection

An object often includes a variety of attributes, including important, unnecessary, and redundant attributes. Our learning algorithm's performance will only increase with the addition of these linked features. In algorithmic applications, dimensional catastrophes frequently happen because we are unsure which feature is useful for our forecast. To increase the effectiveness of the learning algorithm, particularly for the analysis of complicated data, it is crucial to choose significant characteristics from all features. Different feature selection algorithms have been used in numerous sectors [14]. In this study, the Random Forests (RFs) approach is utilized for

feature selection based on the importance index of each feature. This is done because it can determine the relevance of a single feature variable and performs well on most datasets.

The RFs model is created using decision-making regression trees, which can generate numerous trees. The data for each tree is taken via the bootstrap sampling technique from the bag of set B, and the remaining out-of-bag (OOB) samples are defined as set B, which won't be in the training samples. Let C be a set of B and $\overline{C}$ be a set of $\overline{B}$. Suppose we have a test dataset with n rows and p characteristics represented by an $Xn \times p$ matrix and an n-dimensional label vector y indicating each test's category. The RF algorithm evaluates the significance of features by comparing the errors before and after classification [15]. The algorithm associates each feature X_j with a group of tests that replace the feature's values with rearranged ones. The degree to which the replacement of the original feature influences the outcome is determined by comparing the classification error rates of the original features and the replaced randomly rearranged features in the OOB test set. The discrimination of the randomly rearranged features will decline when the significant features are changed, resulting in a rise in the OOB classification error rate. N OOB sets are used as test sets when N trees are constructed. As a result, the following is the definition of the characteristic importance index Ja:

Where $hk(i)$ signifies a classification label of sample i predicted by dataset Bk, i is a characteristic function, and y_i represents a classification label in the i-th OOB.

2.5. Predictive modelling

This research used predictive modelling to classify and forecast the severity of traffic accidents by applying four ML algorithms: Random Forest, Naïve Bayes, Logistic Regression, and Artificial Neural Network. The Accident Severity feature of the accident served as the response variable. In contrast, the input variables were selected based on their significance as predictors of incident severity using a chi-square feature importance analysis. The analysis used a Jupyter Notebook within the Anaconda Navigator environment (version 6.4.12).

- Logistic Regression is a statistical technique used to forecast a binary outcome, such as yes or no, based on one or more input factors [16]. Although it is made expressly for binary outcomes, it is like linear regression.
- A Random Forest is an ensemble technique that uses many Decision Trees to boost prediction stability and accuracy. To get to the conclusion, it creates an extensive amount of Decision Trees and averages their predictions. This improves the model's overall performance and lessens the overfitting of a single decision tree.
- Naive Bayes is a probabilistic technique for classification problems, especially text categorization. It is based on Bayes' theorem, a mathematical formula that determines the likelihood of an event based on knowledge of the circumstances that might have been connected to it in the past. The Naive Bayes technique assumes that the features (or variables) used for classification are independent, which isn't always the case in real-world situations [17].
- Artificial Neural Network (ANN) is an ML technique modelled after the structure and operation of the human brain. It is made up of several linked nodes or neurons that are arranged in layers. ANN can learn from examples and address complicated issues that are difficult for conventional algorithms to handle. The fundamental idea behind ANN is to train the network using a dataset of labelled instances. The weights of the connections between neurons are changed during training to reduce the error between the predicted and actual outputs [18]. The ANN can forecast the output for fresh inputs once trained.

2.5.1. Data splitting

The stratified sample technique was utilized to divide the data in this research into two groups: a training set and a testing set. The model is built, and its parameters are calculated using the training set containing 80% of the data. It is applied to the testing set to assess the final model's accuracy, consisting of the remaining 20% of the data not included in the training set. Based on the results of the test set, the predictive model's efficiency is assessed.

2.6. Model evaluation metrics

The application of improved tools and techniques in accident research has substantially enhanced our understanding of the elements that affect the frequency and severity of accidents. The continual evolution of these strategies and techniques provides the greatest opportunity for continued development in the discipline [19]. This study compares the performance of various analytics models using various performance measures created from a confusion matrix. The distribution of the observations among the actual classes (rows) and predicted classes are displayed in the confusion matrix, which takes the shape of a contingency table. With the

Table 1
Confusion matrix for binary classification.

Actual class	Predicted class	
	NS (Negative)	S (Positive)
NS (Negative)	TN	FP
S (Positive)	FN	TP

display of the number of true positive, true negative, false positive, and false negative predictions, the matrix offers a means of assessing the precision of a classifier. The metrics used to determine the model's overall performance are accuracy, precision, recall, and F1-score. The confusion matrix is a key tool in classification techniques and is the foundation for assessing a model's predictability. For each machine learning model in this study, the binary confusion matrix is utilized to compute quantitative model performance metrics, as shown in Table 1. As mentioned in the work of [20], numerous metrics for evaluating the performance of the models are presented and defined in the following section.

- Recall or sensitivity $\left(\frac{TP}{TP+FN}\right)$ measures the ability of the classifier to correctly identify positive labels.
- Precision $\left(\frac{TP}{TP+FP}\right)$ evaluates how well the data labels match the positive labels defined by the classifier.
- F-score $\left(\frac{2x(\text{Precision x Recall})}{(\text{Precision}+\text{Recall})}\right)$ measure that combines recall and precision in a weighted average manner.
- Overall accuracy $\left(\frac{TN+TP}{Total}\right)$ indicates the proportion of correct classifications the classifier makes.

3. Result: case study

3.1. Data collection and data description

This study utilized accident data gathered from a UK traffic accident dataset, and it was used to assess the applicability and efficacy of the suggested machine-learning techniques. This dataset is obtained from an open source and is provided by the UK Transport Department for accidents between 2005 and 2014. The data collected was from two databases containing the accident and vehicle information. The accident information database provides comprehensive road safety data concerning the factors and conditions that resulted in personal injuries during road accidents in the United Kingdom. The accident information dataset contains 2,047,256 traffic accident records and 34 attributes, of which "Accident Severity" is considered the response variable. The vehicle information database contains information about the vehicle involved in the road traffic accident. The vehicle information dataset contains 2,177,205 records of the vehicle involved in a traffic accident and 24 attributes of the vehicle, such as the age of the vehicle, sex of the driver, vehicle type, age of the driver, etc. The databases were merged using the accident index as an identifier. After merging the two datasets, we cleaned the gathered traffic accident record data by deleting missing values.

3.2. Feature analysis and selection

In this section, we used the Random Forest feature importance to analyse and select features. As outlined in section 2.2, we have 27 largely independent features after data merging and cleaning. This suggests that the dataset is relatively intricate, and not all features may enhance the forecasting accuracy. Certain features might be irrelevant or duplicated. Therefore, it is essential to conduct feature selection before employing the RF algorithm for predicting the dataset [21].

Before model training, we categorize the features as numerical or categorical based on their types. We apply label encoding to the categorical features to evaluate the relative importance of features, which involves mapping each category to a numerical value. A random selection of 70% of the data was made for the training set, leaving 30% for the testing set. The numerical features in both sets were standardized using the min-max normalization method. After standardization, the numerical features in the training set were

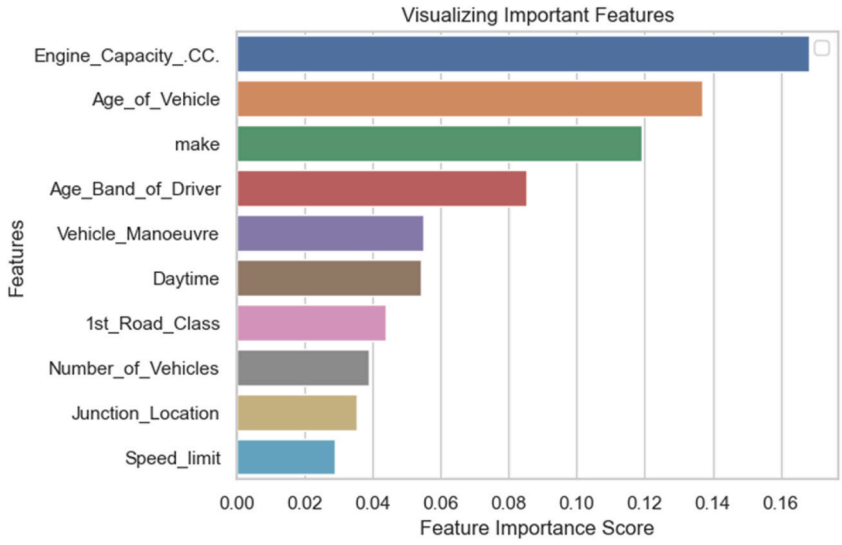


Fig. 2. Variable importance plot using Random Forest Feature Selection.

combined with the categorical features to form the training dataset. Likewise, the standardized numerical features in the test set were merged with the categorical features to create the test dataset. Finally, we employed the RF feature importance. Fig. 2 demonstrates the significance and impact of the input features on the severity of traffic accidents. Using the node purity measure, we arranged the variables in ascending order based on their significance. We relied on the cumulative value curve to determine the importance threshold, specifically the 0.80 value. Observing the importance values of each feature, we found that the 0.80 value corresponded to approximately 0.04. Therefore, we selected the critical value of 0.04 to identify the significant features. The figure displays the outcomes of the feature importance ranking, revealing that the top five features with the most significant impact on accident injury severity are the engine capacity of the vehicle, age of the vehicle, the make of the vehicle, the age of the driver, and the vehicle manoeuvring. This suggests that some vehicle characteristics play a crucial role in determining the severity of accidents. In contrast, accident characteristics such as the time of the day (Daytime), the first road class, the number of vehicles involved, and the junction location have a relatively minor impact on accident injury severity.

3.3. Severity prediction results

3.3.1. Data construction

To construct the data, two main steps were taken. Firstly, the dataset's three levels of injury severity were classified into two classes: serious and non-serious. Secondly, temporally correlated data was constructed.

First, classifying traffic accident injury severity was done by merging fatal and severe injuries into one class, labelled as serious. In contrast, minor injuries were considered another class, labelled as non-serious level. Then, the existing temporal data was used to construct five new features with temporal correlation, including the day of the week, time period, month, hour, and whether the accident occurred on holiday.

3.3.2. Prediction evaluation metrics

In evaluating the model's performance, we employed a matching matrix that comprises false positive, true positive, false negative, and true negative values. A true positive (TP) is recorded when the model accurately detects a serious injury. Conversely, a false positive (FP) happens when the model incorrectly identifies a non-serious injury as serious. A false negative (FN) is observed when the model incorrectly classifies a serious injury as non-serious. Lastly, a true negative (TN) is tallied when the model correctly classifies a non-serious injury. These values are derived from the confusion matrix, as shown in Table 2.

- Recall or sensitivity $\left(\frac{TP}{TP+FN}\right)$ measures the ability to identify positive labels correctly.
- Precision $\left(\frac{TP}{TP+FP}\right)$ evaluates how well the data labels match the positive labels defined by the classifier.
- F-score $\left(\frac{2x(\text{Precision x Recall})}{(\text{Precision}+\text{Recall})}\right)$ measure that combines recall and precision in a weighted average manner.
- Overall accuracy $\left(\frac{TN+TP}{Total}\right)$ shows the proportion of correct classifications the classifier makes.

3.3.3. Predicted result

In this section, the main data features consist of the selected features obtained through feature selection. Subsequently, four algorithms, namely RF, LR, NB, and ANN, are employed to predict the severity of traffic accidents and determine the accuracy of their predictions. The dataset used in this study was split into two parts: a training set consisting of 70% of the incidents and a testing set with 30% of the incidents. The division was carried out through stratified sampling to maintain an equal proportion of serious and non-serious accidents in both sets. After developing the models using the training data, their performance was evaluated on the testing data to gauge their effectiveness. The models' findings regarding the accurate classification of serious and non-serious accidents are presented in Table 3 and Fig. 3(a–d), which compare the predicted and actual outcomes.

Table 4 displays the test dataset's performance for RF, LR, NB, and ANN models. All the models show satisfactory performance, indicating they can generalize when dealing with data. The goal of assessing the performance of these models is to identify the most accurate model among them, as reported by Ref. [22].

In this study, "S" and "NS" represented the positive and negative classes. These labels were used to interpret the recall values. The recall value indicates the model's ability to classify "S" cases correctly. Nevertheless, during the evaluation of the test dataset, LR and RF demonstrated the highest overall classification accuracy of 0.87, outperforming NB and ANN with their respective recall values of 0.80 each.

F-score is another metric also utilized in this study, a commonly used measure of classifier performance as it is the harmonic mean of precision and recall, as reported by Ref. [23]. The findings indicated that the Random Forest classifiers achieved the highest F-score of 0.82, surpassing Artificial Neural Networks, Naïve Bayes, and Logistic Regression, which scored 0.80, 0.80, and 0.81, respectively. Since the F-score balances recall and precision, it can be deduced that the Random Forest classifiers excelled in predicting the severity

Table 2
Confusion matrix for binary classification.

	Predicted: Non-Serious	Predicted: Serious
True: Non-Serious	T: Non-Serious	F: Serious
True: Serious	F: Non-Serious	T: Serious

Table 3
Confusion Matrix for all models.

Model	Actual class	Predicted (test)	
		NS	S
Logistic Regression	NS	28	65,091
	S	56	438,771
Naïve Bayes	NS	11,950	53,169
	S	47,502	391,325
Artificial Neural Network	NS	11,950	53,169
	S	47,502	391,325
Random Forest	NS	3253	61,866
	S	3655	435,172

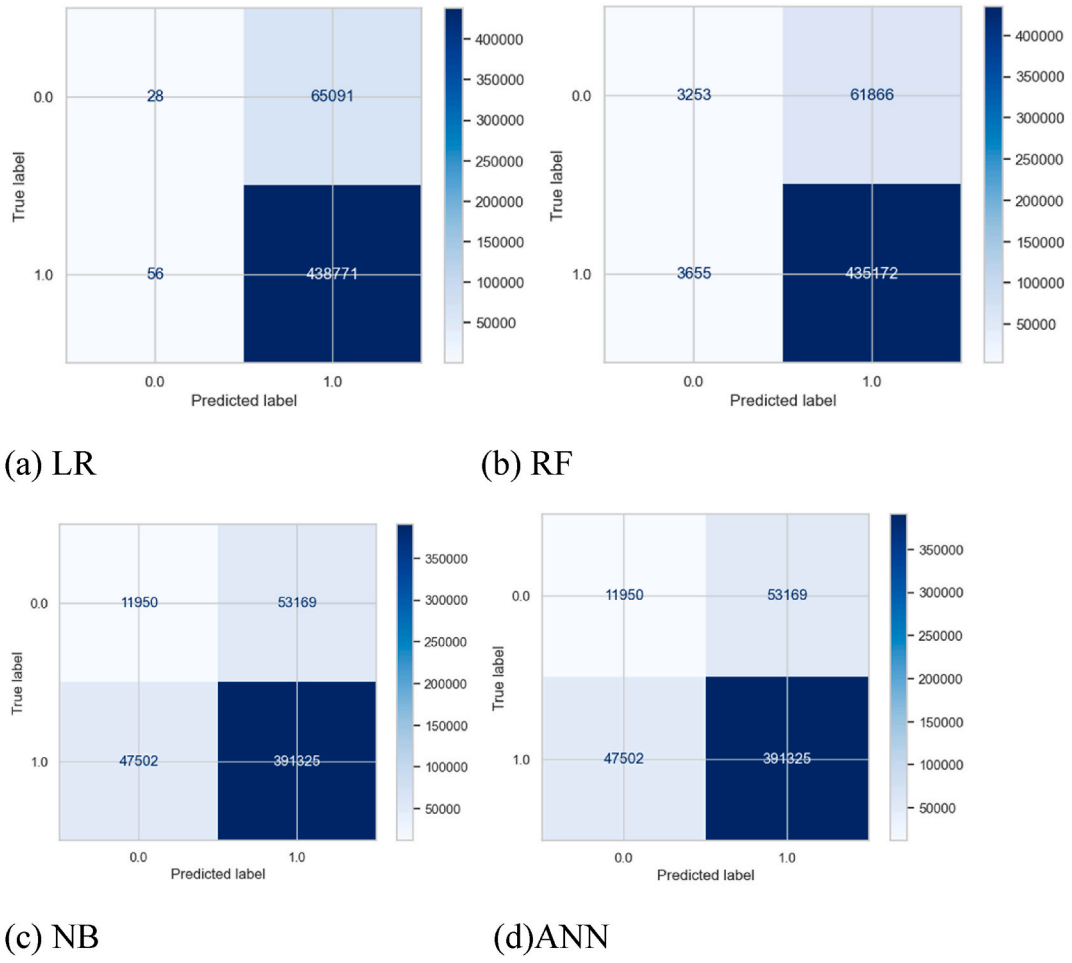


Fig. 3. Confusion matrix generated from the predictions on the test data.

Table 4
Model performance on test data.

Model	Weighted Average Recall	Weighted Average Precision	Weighted Average *F*-score	Weighted Average Overall accuracy
Logistic Regression	0.87	0.80	0.81	0.87
Naïve Bayes	0.80	0.79	0.80	0.80
Artificial Neural Network	0.80	0.79	0.80	0.80
Random Forest	0.87	0.82	0.82	0.87

of traffic accidents.

After evaluating the performance metrics for all models, RF demonstrated superior performance compared to LR, NB, and ANN, making it the best-performing model. Additionally, both LR and RF exhibited F-score values and overall accuracy equal to or higher than NB and ANN.

4. Discussion

Our proposed methods have demonstrated higher prediction accuracy, making them useful tools for accident severity prediction. The advantage of our approach is the utilization of different machine learning techniques in traffic accident injury severity analysis, where each approach is maximized according to its respective strengths. This approach can also solve various traffic-related issues, such as short-term travel time forecasting and traffic flow situation estimation. This is of utmost importance in enhancing the existing traffic safety system within a sustainable transportation framework, such as an intelligent transportation decision system and an intelligent traffic safety management system.

4.1. Importance of the proposed methods

Our study reveals that the LR and RF models exhibited the highest levels of accuracy compared to the other classifiers assessed. This finding aligns with a previous study conducted by Ref. [24], which demonstrated that RF achieved an accuracy rate of 91.98% in identifying employees with an increased risk of fatality at construction sites. Moreover, the study demonstrated that RF surpassed DT, LR, and AdaBoost classifiers. Another investigation discovered that RF was the most effective method for predicting industrial accidents, achieving an accuracy rate of 79% [25]. Furthermore, RF was utilized in a distinct study to forecast the types of occupational accidents that commonly transpire in construction settings [26]. Moreover, by integrating environmental and occupational accident data, the study accomplished a model with a 71.3% accuracy rate. In contrast, for the multi-class classification task, RF emerged as the most effective model for predicting occupational injuries and identifying their causes [27]. The study also highlighted that SVM and RF methods were the most suitable techniques for predicting the severity of work-related injuries due to their notably high accuracy. The RF method combines predictions from multiple classifiers, resulting in a more robust classifier that enhances prediction performance. The research strongly recommends employing these methods. Furthermore, the study indicated that while the ANN model did not perform as well as other models in assessing the severity of traffic accidents, it still achieved a relatively high level of accuracy and successfully identified significant factors in predicting traffic accident severity.

Additionally, this study examined the importance of different features in determining the severity of traffic accidents. The feature importance approach is a contemporary method that aids developers of machine learning models in comprehending and interpret their models. This approach significantly enhances the understanding of classification tasks [28].

This study assessed the importance of features using the RF algorithm. Previous studies, such as the work by Ref. [28], have demonstrated that the RF model surpasses LR in determining feature importance within classification models. The study identified "Engine_Capacity_.CC." as the most significant variable in the dataset. Furthermore, the second most important feature was "Age_of_Vehicle," followed by the feature "make," which ranked third in importance.

Identifying these important features in this research can offer valuable insights for traffic safety management. It enables them to enhance their accident management and control measures by implementing suitable infrastructure improvements, optimizing lighting conditions, setting appropriate speed limits, and providing safety warnings. These actions can contribute to a decrease in both the frequency and severity of traffic accidents.

4.2. Limitations of the proposed methods

The primary constraint of the proposed approach lies in the integrity of the data, particularly concerning the prediction of traffic accident severity. The dataset utilized in this study lacks essential information regarding casualties involved in traffic accidents. It lacks detailed characteristics of the casualties, such as the number of individuals affected by the accident. This absence of crucial data may impede the ability of the proposed methods to attain the intended outcomes. Hence, it is imperative to have datasets that encompass casualty-related information for the proposed methods to effectively contribute to mitigating the adverse consequences of traffic accidents.

Furthermore, various limitations are associated with using a single machine-learning approach. It has been observed that the most favourable outcomes are achieved when multiple analytical techniques are combined, thereby enhancing the analysis of the obtained results [29]. For example, combining a neural network with a rule system [30] or combining a genetic algorithm with a rule system and decision tree [10] has enhanced reliability and accuracy in predicting results compared to a single technique.

4.3. Application in traffic safety management

Sustainable transportation development has traditionally emphasised research into the precise forecast of traffic accident severity, particularly considering the consequences of managing traffic safety. Traffic safety control measures often rely on the limited experience of traffic managers, which can result in deviations from the actual situation. Consequently, strengthening traffic safety in the present system and creating intelligent transportation choice and traffic safety management systems may benefit significantly from improving the accuracy of severity prediction using our suggested methodologies. In contrast to traditional methods that rely on traffic

managers' limited experience, ML algorithms have demonstrated the ability to learn from historical accident data effectively and efficiently.

The proposed ML algorithms in this paper have demonstrated good performance in predicting accident severity, which can serve as an important reference for safety managers in making subjective judgments. For example, safety managers can use the proposed method to identify significant influencing factors of accidents and predict the severity level resulting from these factors. Furthermore, the severity prediction can inform the implementation of accident management and control measures, including improving infrastructure and lighting conditions and implementing speed limits and safety warnings. These predictions can be applied to different datasets, providing a useful tool for safety managers in achieving their goals.

4.4. Future work

In the future, we plan to extend the proposed framework to datasets that include Casualty Accident information. This information has changed how incident severity is recorded and allows for a more consistent classification and reporting of injury severity. However, the change has also led to some previously considered slight injuries now being classified as serious. Adding new features allows for better analysis and prediction of road casualty, which can ultimately help reduce the negative impacts of traffic accidents.

5. Conclusion

The analysis of traffic accident severity is a crucial research topic in road safety. This paper seeks to compare the predictive performance of various machine learning models, such as Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Artificial Neural Networks (ANN), in forecasting injury severity in traffic accidents using various contributing factors. The approach involves using different machine-learning techniques to process and analyse accident data. The study utilized ten years of UK traffic accident data (2005–2014, N = 2,047,256) to develop four ML models for predicting injury severity in individual crashes. The performance of the models was evaluated using various measures such as overall accuracy, precision, recall, and F-1 score.

The Random Forest and Logistic Regression models achieved the highest overall prediction accuracy of 87%, outperforming the Artificial Neural Networks and Naive Bayes models, which had a prediction accuracy of 80%. In addition, a feature importance analysis using Random Forest showed that variables such as Engine Capacity, Age of vehicle, make of vehicle, vehicle manoeuvre, Age of the driver, daytime, and 1st road class were the most influential in the severity of traffic accident prediction. The proposed Random Forest model showed better predictive performance and accuracy than most existing models across different injury severity classes.

Author contribution statement

Izuchukwu Chukwuma Obasi, MSc: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Chizubem Benson, PhD: Contributed reagents, materials, analysis tools or data.

Data availability statement

Data will be made available on request.

Declaration of competing interest

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] World Health Organization, Global Status Report on Road Safety, World Health Organization, 2019.
[2] W. Vanlaar, M. Mainegra Hing, S. Brown, H. McAteer, J. Crain, S. McFaull, Fatal and serious injuries related to vulnerable road users in Canada, J. Saf. Res. 58 (2016) 67–77.
[3] M. Riveiro, M. Lebram, M. Elmer, Anomaly detection for road traffic: a visual analytics framework, IEEE Trans. Intelli. Transport. Syst. 18 (2017) 2260–2270.
[4] Y.K. Arora, S. Kumar, Statistical Approach to Predict Road Accidents in India, 2019. Singapore.
[5] K. Bamel, S. Dass, S. Jaglan, M. Suthar, Statistical Analysis and Development of Accident Prediction Model of Road Safety Conditions in Hisar City, Mohali, India, 2021.
[6] R.K. Jindal, A.K. Agarwal, A.K. Sahoo, Envisaging the road accidents using regression analysis, Int. J. Adv. Sci. Technol 10 (5) (2020) 1708–1716.
[7] A. Jamal, M. Zahid, M.T. Rahman, H.M. Al-Ahmadi, M. Almoshaogeh, D. Farooq, M. Ahmad, Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study, Int. J. Inj. Control Saf. Promot. 28 (4) (2021) 408–427.
[8] C. Gutierrez-Osorio, C. Pedraza, Modern data sources and techniques for analysis and forecast of road accidents: a review, J. Traffic Transport. Eng. 7 (4) (2020) 432–446.
[9] Y. Yang, K. Wang, Z. Yuan, D. Liu, Predicting freeway traffic crash severity using XGBoost-bayesian network model with consideration of features interaction, J. Adv. Transport. 2022 (2022).
[10] S.H.-A. Hashmienejad, S.M.H. Hasheminejad, Traffic accident severity prediction using a novel multi-objective genetic algorithm, Int. J. Crashworthiness 22 (4) (2017) 425–440.
[11] M. Yan, Y. Shen, Traffic accident severity prediction based on random forest, Sustainability 14 (1729) (2022).

[12] Z. Ma, G. Mei, S. Cuomo, An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors, Accid. Anal. Prev. 160 (2021), 106322.
[13] R. Houari, A. Bounceur, M. Kechadi, A. Tari, R. Euler, Dimensionality reduction in data mining : a copula approach, Expert Syst. Appl. 64 (2016) 247–260.
[14] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[15] J. Gan, L. Li, D. Zhang, Z. Yi, Q. Xiang, Emerging technologies in traffic safety risk evaluation, prevention, and control, J. Adv. Transport. 2020 (2020) 13.
[16] G. Mendez, T.D. Buskirk, S. Lohr, S. Haag, Factors Associated with Persistence in Science and Engineering Majors: an Exploratory Study Using Classification Trees and Random Forests, The Research Journal For Engineering Education, 2013.
[17] H. Gao, X. Zeng, C. Yao, Application of improved distributed naive Bayesian algorithms in text classification, J. Supercomput. 75 (2019) 5831–5847.
[18] P. Bhavsar, I. Safro, N. Bouaynaya, R. Polikar, D. Dera, Machine learning in transportation data analytics, Data Anal. Intelli. Transp. Syst. (2017) 283–307.
[19] D. Lord, F. Mannering, The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives, Transport. Res. Pol. Pract. (2010) 291–305.
[20] M. Sokolova, G. Lapalme, A Systematic Analysis of Performance Measures for Classification Tasks, Information Processing & Management, 2009, pp. 427–437.
[21] G. Heinze, C. Wallisch, D. Dunkler, Variable selection - a review and recommendations for the practicing statistician, Biom. J. 60 (3) (2018) 431–449.
[22] A. Oztekin, L. Al-Ebbini, Z. Sevkli, D. Delen, A decision analytic approach to predicting quality of life for lung transplant recipients: a hybrid genetic algorithm-based methodology, Eur. J. Oper. Res. (2018) 639–651.
[23] D. Mathew, Data Mining and Machine Learning Algorithm for Workers' Compensation Early Severity Prediction, 2016.
[24] J. Choi, B. Gu, S. Chin, J.S. Lee, Machine Learning Predictive Model Based on National Data for Fatal Accidents of Construction Workers, Autom. Constr., 2020.
[25] P. Raman, N. Kannan, S. Kumar, K. Raunak, Analysis and prediction of industrial accidents using machine learning, Int. J. Adv. Sci. Technol. (2020) 4990–5000.
[26] K. Kang, H. Ryu, Predicting types of occupational accidents at construction sites in Korea using random forest model, Saf. Sci. (2019) 226–236.
[27] S. Sarkar, V. Pateshwari, J. Maiti, Predictive model for incident occurrences in steel plant in India, in: 8th International Conference on Computing, Communication and Networking Technologies, ICCCNT), Delhi, India, 2017.
[28] M. Saarela, S. Jauhiainen, Comparison of feature importance measures as explanations for classification models, SN Appl. Sci 3 (272) (2021). https://doi.org/10.1007/s42452-021-04148-9.
[29] Z. Yang, W. Zhang, J. Feng, Predicting multiple types of traffic accident severity with explanations: a multi-task deep learning framework, Saf. Sci. 146 (2022), 105522.
[30] F.D. Kakhki, S.A. Freeman, G.A. Mosher, Evaluating machine learning performance in predicting injury severity in agribusiness industries, Saf. Sci. (2019) 257–262.