https://doi.org/10.3311/PPtr.41673

Periodica Polytechnica Transportation Engineering

Evaluation and Comparison of Classification Models for Predicting Crash Severity

Rossi Passarella[1*], Isbatudinia Isbatudinia[1], Mastura Diana Marieska[2], Dedy Kurniawan[3], Romi Fadillah Rahmat[4], Harumi Veny[5]

[1] Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Jalan Palembang-Prabumulih, KM 32 Inderalaya, 30662 Kabupaten Ogan Ilir, South Sumatera, Indonesia
[2] Department of Informatics, Faculty of Computer Science, Universitas Sriwijaya, Jalan Palembang-Prabumulih, KM 32 Inderalaya, 30662 Kabupaten Ogan Ilir, South Sumatera, Indonesia
[3] Department of Information System, Faculty of Computer Science, Universitas Sriwijaya, Jalan Palembang-Prabumulih, KM 32 Inderalaya, 30662 Kabupaten Ogan Ilir, South Sumatera, Indonesia
[4] Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, 9 Universitas Road, 20155 Medan, North Sumatra, Indonesia
[5] Faculty of Chemical Engineering, MARA University of Technology, Jalan Ilmu 1/1,40450 Shah Alam, Selangor Darul Ehsan, Malaysia
* Corresponding author, e-mail: passarella.rossi@unsri.ac.id

Received: 21 July 2025, Accepted: 11 December 2025, Published online: 26 March 2026

Abstract

Traffic accidents pose a significant challenge to modern transportation systems, impacting both safety and infrastructure. This study aimed to evaluate and compare the effectiveness of seven classification models—Random Forest, XGBoost, Logistic Regression, Decision Tree, Gaussian Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbors' (KNN)—in predicting the severity of traffic accidents. A pre-processed dataset from the Road Safety Data in the UK, involving variable selection, outlier treatment, and data normalization, was utilized. The models underwent parameter tuning using Bayesian optimization, and their performance was assessed based on accuracy, precision, recall, and F1-score. The results indicated that Random Forest achieved the highest accuracy of 99.3% on unseen data, closely followed by XGBoost at 99%. Notably, the Random Forest model in this study outperformed a similar study that used XGBoost. The statistical findings emphasize the benefit of employing a comprehensive set of accident-related variables, advanced pre-processing techniques, and optimized hyper parameter tuning for developing reliable crash severity prediction models. Based on the findings, the Random Forest model is strongly recommended for practical implementation to enhance road safety.

Keywords

traffic accidents, severity prediction, machine learning, classification models, random forest, road safety

1 Introduction

Traffic accidents constitute a serious issue and a great challenge to contemporary transport systems. The consequence of the accidents is profound, not only resulting in thousands of casualties annually but also resulting in massive economic losses (Betele et al., 2025). According to the World Health Organization (WHO), around 1.25 million people die annually as a result of traffic accidents, and the trend persists in several nations (World Health Organization, 2018). Several reasons lead to the vast numbers of accidents, such as human errors, poor road infrastructure conditions, and environmental issues (Chand et al., 2021; Sun et al., 2025). In response to the problems, innovative application of technologies such as artificial intelligence (AI) and machine learning have been employed to enhance security in transport (Chai et al., 2025; Olayode et al., 2023). Utilization of the technologies facilitates the creation of models which can analyze crash risk, forecast the probability of crashes, and estimate the severity of the crash (Ahmed et al., 2021; Berhanu et al., 2024). Despite efforts being taken, prediction and prevention of traffic accidents remain the focal point. Hence, more research will be employed to make transport security models increasingly more precise. Improvement in the accuracy of the models is supposed to effectively counteract the negative effects of traffic accidents (Belete et al., 2025; Berhanu et al., 2024).

Cite this article as: Passarella, R., Isbatudinia, Marieska, M. D., Kurniawan, D., Fadillah Rahmat, R., Veny, H. (2026) "Evaluation and Comparison of Classification Models for Predicting Crash Severity", Periodica Polytechnica Transportation Engineering, https://doi.org/10.3311/PPtr.41673

Recent rapid development in AI technology has opened the door to possibilities for enhanced prediction accuracy in traffic accidents (Akhtar and Moridpour, 2021; Qiao et al., 2025). In their scientific article, Santos et al. (2021) illustrate the potential of integrating various machine learning techniques—including supervised and unsupervised methods—for traffic accident analysis. This approach provides a comprehensive analysis of the factors influencing accident severity. Their study, which used accident data from the Setúbal district in Portugal (the MOPREVIS project), found that the most important factors that affect the severity of accidents are motorcycle and pedestrian accidents (collisions), especially in cities. Meanwhile, factors such as weather conditions and road type are used in the development of hotspot prediction models to map areas prone to future accidents.

Although such advanced data-driven methods have shown great potential, we should remember that there will always be grounds for improvement with respect to the accuracy of the prediction. A potential strategy includes comparing the performance of various machine learning models. We can compare and test the various models in order to find the optimal approach for the application in the prediction of traffic crash severity. Every machine learning model possesses inherent strengths and weaknesses, which depend a great deal on the data's complexity that will be processed as well as on the requirement for interpreting the results of the prediction. A comparative analysis among models is hence a key step in the discovery of the optimal way for predicting traffic crash severity.

In an attempt to make roads safer, the application of machine learning models for predicting the severity of crashes has been the primary focus in the majority of the studies. Some models have been investigated, with particular decision tree-based models including Random Forest, XGBoost, and Decision Tree (Madushani et al., 2023; Shehadeh et al., 2021). The models have been well sought after since they can analyze a wide range of factors contributing to a crash. Aside from decision tree models, SVM is equally used extensively in crash severity classification research, largely owing to the ease with which it can handle data with varied natures (Çeven and Albayrak, 2024; Dimitrijevic et al., 2023). Finally, the use of the Gaussian Naive Bayes also contributes by predicting possible traffic accidents using patterns of existing events (Singh et al., 2025). K-Nearest Neighbor, too, provides a much-needed but different angle by looking at patterns associated with accidents based on similar features among the cases (Khosravi et al., 2024; Özkan and Ulaş, 2024). With the proliferation of methods that have been devised, research is needed to refine classification models in a bid to make them more effective and optimize the prediction accuracy of crash severity is a necessity.

Realizing the significance of the matter, the present study explicitly aims at comparing the performance of seven well-known classification models in predicting the severity of traffic crashes. Random Forest, XGBoost, Logistic Regression, Decision Tree, Gaussian Naive Bayes, SVM, and KNN models are the seven models. In addition, the present study will also carry out parameter optimization via Bayesian optimization in order to enhance the performance of the models. Model performance will be compared with appropriate metrics such as accuracy, precision, recall, and F1-score. It aims at determining the optimal model that would be best suited for accident prediction purposes. Through a comprehensive comprehension of the pros and cons of each model, the present study would be beneficial in providing insight into the optimal accident prediction methods.

Beyond mere model comparison, this research holds great aspirations to make a tangible contribution toward the improvement of a more advanced and data-based traffic safety system. It will be extensively used in denser urban settings as well as more isolated rural areas so as to increase road users' security. This research thereby presents not only a more profound theory but also a practical application that can be used in minimizing accident risks and making roads safer for all.

1.1 Literature review

This section of the literature review is devoted to studies that use machine learning to predict the severity of fatal traffic crashes. Some notable studies, such as the one by Obasi and Benson (2023), have paved the way for the development of more precise and comprehensive prediction models. In addition, the integration of data mining techniques in predicting road accidents has been proven effective by Shakil et al. (2022). In line with these developments, advanced algorithms like LightGBM-TPE, proposed by Li et al. (2022), significantly contribute to visualizing and analyzing accident risk factors in greater depth. Collectively, these studies illustrate how the integration of machine learning methods with other data analysis techniques can provide a richer understanding of significant variables, such as the type of road, the socioeconomic level of the driver, and the location of their residence.

The study titled "Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents" is significant in the sense that it effectively integrated machine learning, econometric methods, and time

series forecasting in examining the severity of accidents in the UK. With it, they illustrated how beneficial the integration could be in achieving a deeper understanding of crash risk factors by labelling significant variables such as the road type, the socioeconomic

Following pioneering work numerous other studies have considered the potential for the application of machine learning in enhancing transport security, with some other but related objectives (Alqahtani and Kumar 2024). For instance, studies on environmental conditions and the nature of accidents have further highlighted the need for the consideration of contextual variables. The study conducted by Gálvez-Pérez et al. (2023) explained that environmental conditions, including the lack of street lighting and urban population, lead to the severity of injury among pedestrians. This finding agrees with another work highlighting the effects of regional factors as well as environmental conditions (Jalilian et al., 2019). In a similar vein, the application of machine learning algorithms in identifying significant risk factors in pedestrian accidents was discussed in (Elalouf et al., 2023). This indicates that the capacity of machine learning in managing intricate data as well as identifying underlying patterns, as highlighted and proves beneficial in all modes of crash scenarios (Azami et al., 2024).

Other research has further ratified the supremacy of some models using machine learning, which is applicable in the case of the model used and tuned. For instance, the researchers found that XGBoost performed optimally when identifying traffic accidents (Xu et al., 2024). From the result, it appears that the model can effectively manage large, complex data and solve issues with data imbalances that are a prevalent issue in accident prediction. Another research study conducted by Liu et al. (2022) ratified these findings by illustrating that the use of XGBoost results in enhanced performance compared to other algorithms in systems for detecting traffic incidents. Another study compared various algorithms, including Gaussian Naive Bayes, Random Forest, Logistic Regression, and Artificial Neural Networks (Obasi and Benson, 2023). Random Forest and Logistic Regression were found by them to make the most accurate prediction. This finding highlights the importance of investigating other classification models.

For instance, Berhanu et al. (2024) analyzed on-route optimization using Random Forest with spatial network analysis. Their work demonstrates how the application of machine learning models can be utilized in the real-world context for the improvement of road safety, which aligns with the mission in (United Nations Economic Commission for Europe, 2021) to make the public safer. Libnao et al. (2023) similarly demonstrated using Gaussian Naive Bayes how the application of machine learning models can be employed in anticipatory actions in traffic management, but they still need to be enhanced with regards to accuracy, which provides avenues for future research such as this one. Even employing the use of machine learning elsewhere in the field of transport, such as in Silagyi and Liu (2023) utilization of SVM for the prediction of plane crashes demonstrates how incredibly flexible and beneficial the use of machine learning can be in the analysis of safety and provides greater context for the analysis carried out by this research. Overall, this literature review confirms that there is a solid foundation for the application of integrated machine learning to predict the severity of traffic accidents based on previous literature (Obasi and Benson, 2023).

Other work goes on to develop and test machine learning methodologies in some other transport aspects of safety. This paper seeks to capitalize on this by comparing the performance of seven different classification models using the machine learning model, among them the optimized XGBoost model and other models that have been determined as significant by this literature review (Random Forest, SVM, Gaussian Naive Bayes, and Logistic Regression). Moreover, this manuscript explicitly adds decision tree models and KNN models and employs parameter optimization using Bayesian optimization to further improve the analysis and potentially further increase the accuracy in prediction over previous studies. This expectedly should provide more complete analysis and practical advice on the selection of the optimal use of machine learning models in the improvement of traffic safety.

2 Data and methods

2.1 Data

Data preparation is one of the key stages in scientific research, particularly when developing prediction models with the help of algorithms for machine learning. For the provision of correct and reliable research findings, data are meticulously collected and processed prior to being utilized for model building. In the present research, crash history data were derived from the official website data.gov.uk, which is maintained by the UK government and administered by the Department for Transport (DfT). Utilized data in this study are Road Safety Data, comprising three primary parts, that is, Road Safety Data-Casualties in 2019 (Data.gov.uk, 2024a), Road Safety Data-Vehicles in 2019 (Data.gov.uk, 2024b), and Road Safety Data-Collisions in 2019 (Data.gov.uk, 2024c). All these three data have varied numbers of rows, as each one represents a variety of entities in the traffic accidents. The Collisions data with general information on the

crashes contain 117,536 rows and 37 variables. Casualties data contain information on the victims in a crash with a total count of 153,158 rows and 21 variables. Vehicles data contain data on vehicles in the crash with a total count of 216,381 rows and 34 variables.

In the second stage, the three data frames have to be merged into one entity in order to get a more complete view of traffic accidents. First, the Collisions dataset must be merged into Casualties using the accident index as the reference. Because one accident might have more than one victim, the links between the collisions and the casualties are one-to-many; each collision in the Collisions dataset can have multiple records in the Casualties dataset. Accordingly, the merged dataset takes a total of 153,158 rows and 57 variables, equal to the victims documented in the Casualties dataset. It implies that each casualty still maintains their accident information, and if an accident includes three casualties, the same accident index will be reflected three times in the merged dataset.

Then, the merged dataset is merged with the Vehicles dataset based on the linking key accident index. It should be mentioned in the merging process that the collision-vehicle relation is one-to-many as well. This is owing to the fact that more than one vehicle may be present in a given accident. In addition, the relation between vehicles and casualties is one-to-many, as each accident might have multiple victims and multiple vehicles present. Thus, during the merging procedure at this level, a casualty has to be matched with a vehicle that features in the same accident, resulting in the duplication of rows in the final dataset. Thusly, the merged dataset contains a total of 233,565 rows and 90 variables representing the matching victims and vehicles per crash.

With more rows than the original dataset as a result of the vehicles and victims' many-to-many relationship, the merged dataset furnishes a more complete data structure in the analysis of the patterns in traffic accidents. This procedure matters as it facilitates the analysis of each victim and vehicle taking into consideration all the respective crash factors. With the merging carried out, the pre-processing procedure comes next, whereby the data will be scrubbed, filtered, and in top shape to be used in future analyses, including the creation of prediction models based on machine learning.

2.2 Pre-processing

During data pre-processing, the variable selection procedure was carried out to identify the most appropriate features for use in developing the prediction model. All the merged dataset initially had a total of 90 variables. Once screened through a selection criterion based on their relation to the influencing factors for traffic accidents using the background knowledge, the selection narrowed down the variables into 14. Some of the variables that have been used include accident severity, speed limit, weather conditions, road surface conditions, light conditions, urban or rural area, road type, junction detail, number of vehicles, vehicle type, age of driver, junction control, number of casualties, and casualty severity. The selection rationale lies in the consideration that they capture major issues that pertain to road safety across the environment, road, road design, vehicle and road user characteristics, and road control. By taking into consideration these variables, the model should be more capable of making more precise forecast estimates concerning crash severity.

Out of the 14 variables that were chosen, accident severity is specified as the target variable. Its selection as a target is well justified since it is more closely linked with the goal of the research, i.e., understanding and predicting the severity of accidents. Furthermore, it holds tremendous importance in road safety and can produce actionable feedback for policy making and accident prevention. Aiming at variable accident severity, this research can contribute meaningfully by addressing road safety and the effects of traffic accidents by expanding the comprehension on the causes of the accidents as well as the efficacy of counter-measures that could be undertaken.

Following that, exploratory data analysis (EDA) was performed in order to gain insight into the patterns, relationships, and anomalies in the data. This analysis is significant in determining the distribution of the data as well as the existence of outliers and missing values. The second pre-processing procedure involved cleaning the data by removing invalid and duplicate records. This reduced the data quantity down to 64,904 rows, as can be visualized in Table 1. Subsequently, outliers were treated on a number of variables with irregular distribution, i.e., speed limit, number of vehicles, number of casualties, age of driver, as well as vehicle type by utilizing the capping technique with Interquartile Range (IQR). Besides, for the sake of ensuring the scale of the variables were the same, all five were normalized using the min-max scaler technique.

Following all the pre-processing phases, the second step involves splitting the data for model testing and training for the baseline model. Data were split between target and predictive variables, and then split with a proportion of 80:20 based on random sampling. Accordingly, the data were utilized as 51,923 rows for training and 12,981 rows for testing. This split ensures that the model learns intricate patterns using sufficient data and still retains some test data for evaluating the performance and generalization.

Table 1 Composition of the initial dataset after pre-processing

Classes label	Amount data	Total data
1	1,298	
2	16,722	64,904
3	46,884	

2.3 Methods

As illustrated in Fig 1, the research involves a systematic series of stages in developing a model predicting the severity of the crash using a machine learning framework. The research starts by collecting data from the Road Safety Data source with information on accidents, vehicles, and victims. Once the data is collected, the second stage involves pre-processing with various critical stages such as variable selection, data cleaning, normalization, and outlier handling. Variable selection involves the selection of the optimal features that will be used in developing the model, and data cleaning involves the elimination of missing values, duplicates, and incorrect data. In an effort to handle non-standardized scales, normalization is conducted on some variables, and outlier handling is carried out using the IQR method so that the model will function in the optimum way.

Upon completion of the pre-processing phase, data handling, if needed, is performed, particularly to deal with class imbalances that could interfere with model performance. Based on the data's distribution, the technique employed during the handling involves oversampling using SMOTE for more instances in the minority class or undersampling for decreasing the majority class's instances. This process ensures that the model developed becomes free from bias against some classes and can make more precise forecasts. On completion of the data-handling procedure, the data is split into two parts, i.e., for training and for testing. This splitting occurs through the random sampling technique so that each class in the dataset will be well maintained in the test and the training data. With this technique, the model can be tested fairly, and the overfitting risk could be minimized so that the results achieved are more credible when used in new data.

We performed the training, testing, and hyper parameter tuning during the model creation phase. We tested seven classification algorithms and checked how accurately they performed: Random Forest, XGBoost, Logistic Regression, Decision Tree, Gaussian Naive Bayes, SVM, and KNN. While training the model, the optimal hyper parameters are searched for the model so that it performs better, with techniques being grid searches, random searches, or Bayesian optimization so that the model will make more accurate predictions. We tested the model, after training it, with an F1 score, accuracy matrix, and recall matrix to check how well the model classified the severity of the accident.

Following the model assessment, the second step is model selection, in which the top model with the highest performance according to the performance metrics assessment is chosen. From the top model, further analysis is conducted in the result analysis. Last but not least, the study concludes based on the findings derived, which will serve as a foundation for policy development and prevention efforts for enhanced traffic safety in the future.

3 Results and discussion

This section describes the results of the research using the previously described methods.

3.1 Scenario

In the present research, the class distribution of the target variable presents itself as imbalanced, which might impact the performance of the prediction model. Thus, a number of data imbalance treatment scenarios were performed so the model could identify patterns more effectively and not be subject to majority-class bias. In this study, the three methods employed were oversampling with SMOTE,

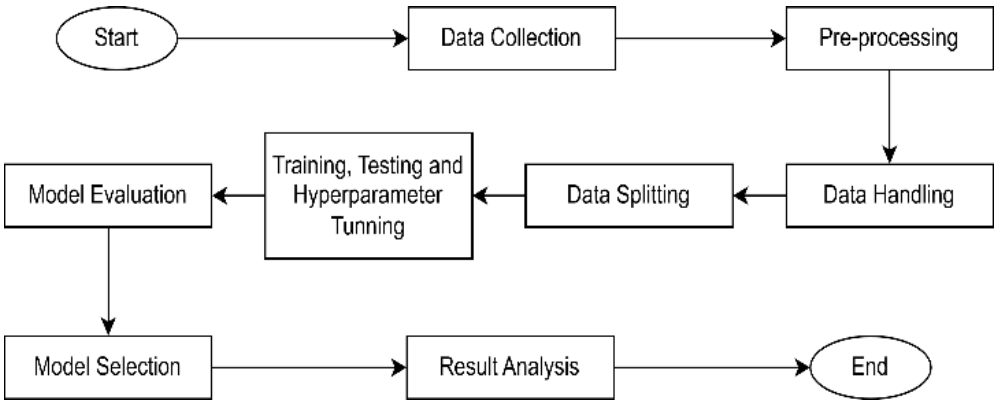

Fig 1 Workflow of the research

undersampling, and the use of a combination of the two as the hybrid sampling. Oversampling with SMOTE seeks to augment the sample count of the minority classes by developing a synthetic sample with the current data distribution so the model will have additional data to train with previously underrepresented classes. On the other hand, the practice of undersampling involves decreasing the sample count for the majority class so the data distribution will be more equal, though the practice leads to possible loss of information. Lastly, the use of the two approaches combines the two by balancing the sample count optimally between the two classes so the chance of bias might be minimized and the model becomes more generalizable.

Following the use of different sampling techniques, the assessment takes place using different methods based on how data imbalance is treated. If sampling were used in treating data imbalance, the primary measure of interest would be accuracy as the technique seeks to balance the data so that the model learns equally well from all classes. By increasing the representation of the minority classes or by decreasing the sample for the majority classes, the model should identify patterns more equally, thereby making more stable outputs. In the case of data imbalance not being treated, the measure changes more toward the use of the F1 score that more accurately addresses the balance between precision and recall, particularly if the minority classes were significantly less than the majority class.

These results for model evaluation are presented in Table 2, comparing the performance metrics both before and after the application of data imbalance handling methods. This provides a better description of the performance of each method. By comparing the methods, one can analyze the degree with which each sampling strategy impacts the improvement in model performance as well as whether some methods compare favorably with others with respect to the dataset at hand. This analysis will help identify the optimal method that can be used in this research so that the resulting model will have optimal accuracy and generalizability when deployed on test data.

Table 2 Comparison of sampling methods using the initial dataset (number of data: 64,904)

Sampling methods	Training (%)		Testing (%)	
	F1 score	Accuracy	F1 score	Accuracy
Original without sampling	97	98	85	93
Oversampling	99	99	90	90
Undersampling	100	100	85	85
Hybrid sampling	100	100	94	94

Depending on the result comparison illustrated in Table 2, the hybrid sampling technique was used as it gives the optimal balance between performance on test and training data. In comparison with other approaches, the F1 score on the test data by the hybrid sampling technique had a high measure with a value of 94% and accuracy with the same measure at a value of 94%, meaning the model can manage class imbalances well with minimal overfitting as in the case of oversampling and the undersampling approaches individually. Following the data treatment using the hybrid sampling, the data were further split into three: 70% for training, 15% for the test, and 15% for the unseen data, with details illustrated in Table 3. Random data division takes place using the random sampling technique with a particular random state for the sake of consistent replication.

3.2 Training model analysis

Training models are a key aid when developing and implementing effective learning improvement programs. In ML, data-driven model development is a significant component in developing precise models. Nevertheless, the performance of ML models varies based on a variety of factors, such as data pre-processing, model selection and assessment, and hyper parameter tuning. In the final analysis, the findings indicate that the use of the hybrid sampling technique gives the optimal performance between the training and test data with greater accuracy and F1 scores compared to other techniques. In comparison with the technique with no imbalance handle or other sampling techniques, the use of the hybrid sampling reduces the chances of overfitting, giving a stable and consistent model.

Then, hyper parameter tuning with Bayesian optimization was carried out as it proves more effective in determining the optimal hyper parameter setting compared to grid search and random search. Bayesian optimization employs a probabilistic model, enabling a more intelligent search with minimal numbers of iteration and, hence, saving computational time and resources. It is more effective

Table 3 Composition of the data set used to develop the prediction model after sieving the data using the hybrid sampling method

Parameters	Classes label			Subtotal
	Class 1	Class 2	Class 3	
Training	30,612	27,971	28,578	87,161
Test	6,606	5,941	6,131	18,678
Unseen	6,643	6,036	5,999	18,678
TOTAL	43,861	39,948	40,708	124,517

Table 4 Comparison of training models for the data of the developed classification models

No	Models	Accuracy (%)
1	Random forest	99.9
2	XGBoost	99.8
3	SVM	99.5
4	Logistic regression	88.8
5	Decision tree	97.8
6	KNN	100
7	Gaussian naive bayes	87.9

Table 5 Comparison of the results of testing the training model with the test data of the developed classification model

No	Models	Accuracy (%)
1	Random forest	99.3
2	XGBoost	99.1
3	SVM	98.9
4	Logistic regression	89
5	Decision tree	97.1
6	KNN	99.5
7	Gaussian naive bayes	87.7

in searching the parameter space compared to computationally expensive grid search and less systematic random search. Seven models were compared based on accuracy, and the models were created as presented in Table 4.

From Table 4, the top pick based solely on the training data results would be the XGBoost model with a 99.8% accuracy. While the Random Forest model beats it by a narrow margin with a 99.9% accuracy, the computational efficiency and overfitting resistance that XGBoost possesses make it a better model. Meanwhile, the KNN model with a 100% accuracy was not used as the model is likely to overfit and as such tends to decrease in performance when tested on new data. It is slower in making predictions as well, particularly for large data. Nevertheless, since this analysis still took place using the training data, the models were tested again using the testing data and new data to ascertain the degree each model can generalize to novel data.

3.3 Testing model analysis

Following the training phase, the second step involves trying the model's performance over the test data in order to test how well the models can generalize learned patterns. This test aims at determining the level at which the model can make correct and reliable predictions when dealing with new data. Furthermore, this test seeks to ascertain the issues such as overfitting or underfitting, which could impact the performance reliability of the model in actual usage. Through a comparison among the performance of different models on the test data, one can ascertain the optimal way in which the dataset is processed and the best resulting prediction outcome. The accuracy scores for the seven models that were tested are listed in Table 5, indicating how each model performed in classifying the test data.

From Table 5, the KNN model had the highest accuracy on the test data, at 99.5%, followed by Random Forest (99.3%) and then XGBoost (99.1%). While KNN performed the best in this test, the accuracy score on the prior training data achieved 100%, which shows the potential for overfitting. In the meantime, Random Forest and XGBoost suffered a marginal decrease in their training accuracy, which indicates that these models are more stable and not so stuck on the patterns in the training data. The SVN model with an accuracy score of 98.9% was still powerful, and Decision Tree (97.1%) suffered more decrease compared with training, which may reflect the inability for generalization. Meanwhile, logistic regression (89%) and Gaussian Naive Bayes (87.7%) are still the least accurate models but still stable for easier interpretation. Random Forest and XGBoost are still the top pick according to these results, as they can achieve a balance between being highly accurate and resistant against overfitting, but still, further proof with the use of unseen data is required to identify the absolute best model.

3.4 Unseen model analysis

After the model is tested against test data, the second step involves testing the performance against unseen data so that the model's generalizability in applying its prediction against entirely novel data can be quantified. Unseen data involves part of the dataset that had never been utilized prior in the testing process or the training process, so the test results at this point can reflect a more objective view of the reliability of the model in real-world contexts. Unseen data testing seeks to ascertain if the model can still sustain accuracy and performance stability or has dropped as a consequence of overfitting of the train data. If the test results and the unseen data have a considerable disparity, then the model might not be generalizable. Thus, the analysis here forms a critical component so as to ensure that the model chosen proves to be reliable prior to implementation in actual prediction. Model accuracy test results on the unseen data are reflected in Table 6 as a comparison between each model's performance in predicting more general accident classes.

From Table 6, the KNN model still had the greatest accuracy on the unseen data, at a level of 99.4%, followed by

Table 6 Comparison of unseen data of the developed classification model

No	Models	Accuracy (%)
1	Random forest	99.3
2	XGBoost	99
3	SVM	98.8
4	Logistic regression	88.8
5	Decision tree	97.2
6	KNN	99.4
7	Gaussian naive bayes	87.5

Random Forest (99.3%) and then XGBoost (99%). While the KNN model still had a remarkably high performance, this model had hit a ceiling previously at 100% on the training data, which still points toward the possibility of overfitting. Random Forest and XGBoost were still stable with a minor decrease in accuracy compared to the test data, meaning they could handle the new data more effectively. Meanwhile, the model with the accuracy level of 98.8% for SVM stays in the race, while Decision Tree (at 97.2%) still shows a minor improvement over the testing data but still lags behind the ensemble models like Random Forest and XGBoost. Still the least accurate models were the logistic regression (88.8%) and the Gaussian Naive Bayes at 87.5%. This indicated that the models using probability and the use of the linear regression were not as effective in grasping the intricate patterns used in this data. Overall, Random Forest and XGBoost are the more stable options since they can strike a balance between being highly accurate and generalizable to the new data and so the top models to use in actual application.

3.5 Model selection

Selection of the optimal model in a system using machine learning relies not only on how well it can be trained with data but how well it generalizes on testing and unseen data. A model that does extremely well on the data it is being trained on but severely declines when tested with novel data runs the risk of overfitting; the model overfits the patterns in the data it's being trained with and does not work optimally in real-world scenarios. On the other hand, a model with consistent accuracy across various phases of testing demonstrates greater resistance to variations in data and hence more suitable to be used in real-world scenarios. Thus, in the search for the optimum model, an extensive analysis by comparing the performance of the models on the data being trained, the test data, and the unseen data must be carried out. With these three phases of testing, a comparison table for all the models' accuracy is presented in Table 7. This identifies the model with the optimal yet balanced accuracy, stability, and generalizability.

Table 7 Comparison of accuracy between classification models built

No	Models	Accuracy (%)		
		Train	Test	Unseen
1	Random forest	199.9	99.3	99.3
2	XGBoost	99.8	99.1	99
3	SVM	99.5	98.9	98.8
4	Logistic regression	88.8	89	88.8
5	Decision tree	97.8	97.1	97.2
6	KNN	100	99.5	99.4
7	Gaussian naive bayes	87.9	87.7	87.5

From Table 7, the performance of differing machine learning algorithms indicates that Random Forest performs the very best in accuracy and generalization. The model had extremely high accuracy for the training data (99.9%) and had very little degradation on test data (99.3%) and previously unseen data (99.3%). That stability indicates that Random Forest can identify patterns well without suffering from too much overfitting, so it can be trusted with dealing with new data with stable performance. Stability in performance is one reason why the optimal model should be chosen, as algorithms that overfit the training data will have a serious plunge in accuracy when tested with previously unseen data.

In comparison, while the highest accuracy on unseen data belongs to KNN with 99.4%, the model achieves 100% accuracy when the data is on the training set, which strongly indicates overfitting. Overfitting will make the model perform extremely well on known data but less consistently when more variations in new data occur. All the same, XGBoost performed exceedingly well with accuracy approximating Random Forest giving 99.8% on the training set, 99.1% on test data, and 99.0% on unseen data but slightly lower stability compared to Random Forest. Other models including SVM, Decision Tree, Logistic Regression, and Gaussian Naive Bayes performed moderately well but not as effectively as Random Forest in being balanced between accuracy and generalizability. Decision Tree boasts relatively good accuracy, giving 97.8% when the data is from the training set, 97.1% when on test data, and 97.2% when the data is from the unseen dataset but more susceptible to overfitting compared to the use of ensemble methods such as Random Forest and XGBoost. On the other hand, logistic regression and Gaussian Naive Bayes have relatively lower accuracy ranging between 88%-89% with limitations in the case of more complicated data.

It can be observed from the plots in Fig. 2 that Random Forest is highly stable and accurate at the different stages of assessment, making the model the most consistent one for use in practical implementations. It demonstrates higher

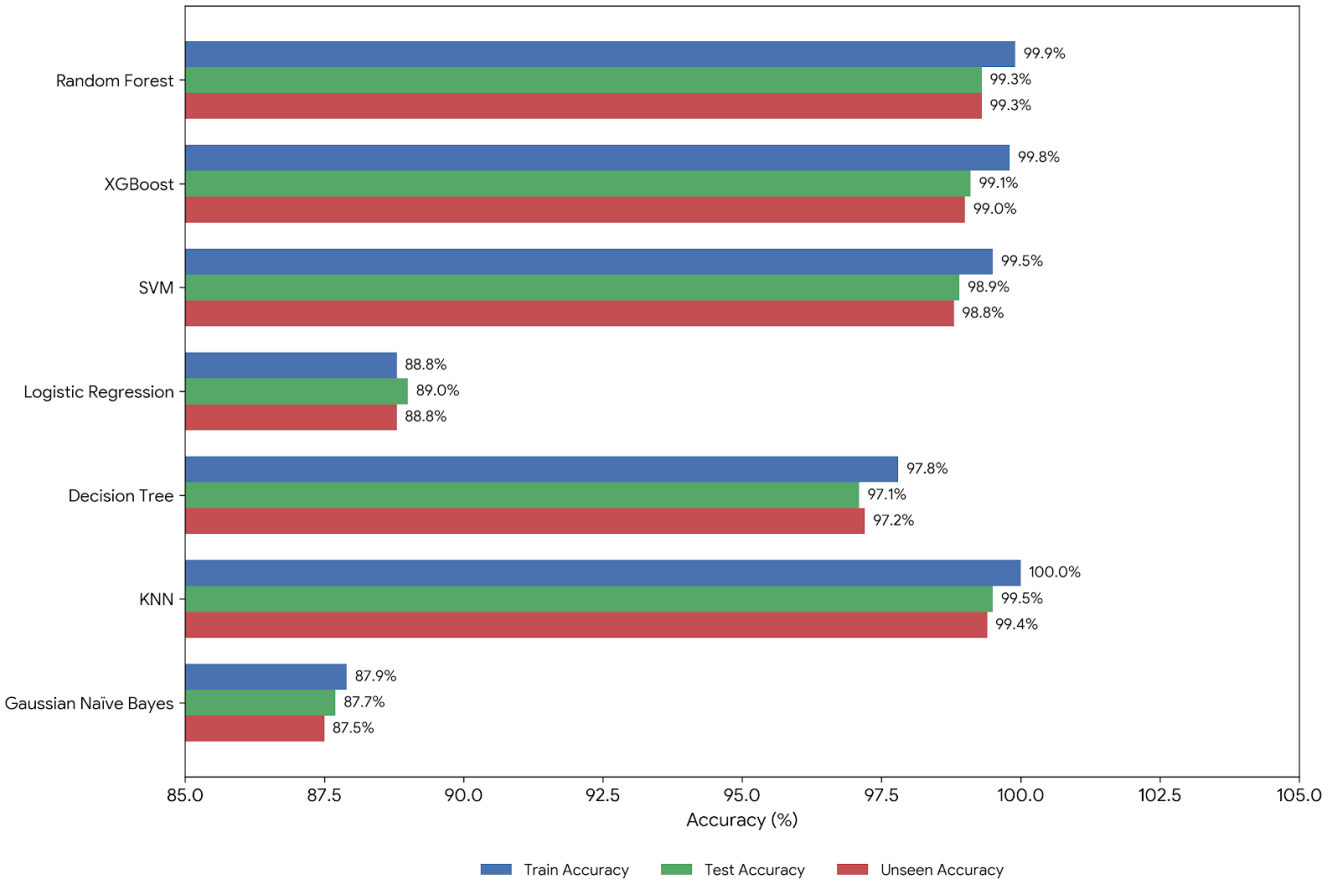

Fig. 2 Comparison of classification model accuracy based on model development

performance on the training data (99.9%) with little deterioration on test data (99.3%) and on unseen data (99.3%) and thus indicates the model's capacity for good generalization of new data patterns. Stability in practical use especially comes in handy since the model will need to be consistent even when it takes data that is not quite the same as the training data used in the model. With a balance between good accuracy, performance stability, and good generalization capacity, Random Forest emerged as the optimal model in this experiment. Aside from being excellent in the case of complex data, the model avoids overfitting as a result of the mechanism of ensemble learning. Its support for numerical as well as categorical variables makes it a model that can be used in a wide variety of model scenarios. Further testing will be done with the model to ensure it continues performing well in real life. This will involve examining how fast it makes the prediction, how well it performs on greater data, and other performance indicators such as precision, recall, and F1 score more critically.

In the previous study (Obasi and Benson, 2023), the variables were primarily location, the count of vehicles and victims, and the characteristics of the vehicle and the driver. This study, by contrast, employs a broader set of variables, including environmental conditions, accident characteristics, and road infrastructure features, which enables the model to have a more comprehensive understanding of accident patterns. This variation in the selection of variables, consequently, impacts the model's accuracy and reliability in the prediction of traffic crash severity. Moreover, the data pre-processing done in this study differs from previous studies. In the previous study, data pre-processing was not detailed, whereas in this study, more advanced methods were employed, including normalization using Min-Max scaling, outlier treatment using the IQR technique, and data balancing using SMOTE and oversampling. All these ensure that the used data is cleaner, structured, and balanced, enhancing the model's performance in identifying patterns.

From the findings of this comparison, it can be concluded that model performance does not solely depend on the selection of the algorithms but in addition largely on other variables including the selection of variables, data pre-processing, and parameter optimization. With a more systematic consideration in the current study, the Random Forest model is capable of making more reliable and more precise predictions than the optimal models in previous studies. This shows that the development

of artificial intelligence models for predicting traffic accidents improves with the improvement in data analysis methods and the optimization of models in machine learning. By taking these into consideration, the current study achieved more precise results and demonstrates that the use of the appropriate procedure at each stage in the models significantly affects the end performance of the model.

3.6 Analysis of the selected model

Following the model selection procedure, the Random Forest model proved optimal based on accuracy on training, test, and unseen data when compared with the other six models. In a further insight into how well the model performed, the classification report and the confusion matrix were examined for the three datasets. Table 8 shows the random forest model performance results on each dataset.

From Table 8, the model performance indicates very high accuracy on all three data tested, i.e., training, test, and unseen. On the training data, the model correctly classified all samples with precision, recall, and F1-score of 100% for all classes, as well as accuracy with a value of 100%, meaning the model could identify patterns in the training data with minimal errors. On the test data, when tested, the model experienced a dip in performance with accuracy being 99% and average precision, recall, and F1-score being equal to 99%. Though still very high, it can be observed that in Classes 2 and 3 the recall and precision fell slightly to a value of 99%, meaning some errors have occurred in classifying some samples. Based on the unseen data, the model still had top performance with accuracy being equally 99% as well as average precision, recall, and F1-score being equal to 99%, meaning the model can identify patterns using the unseen data with almost the same accuracy as with the test data. The fact that the performance between test and unseen data are consistent shows that the model can generalize well and does not suffer a sharp decrease when confronted with the new data.

From Table 9, classification performance indicates excellent classification for the training, test, and unseen data with minimal misclassifications. On the training data, nearly all the samples were correctly classified; for instance, Class 1 had a correct sample count of 30,599 with very few misclassifications to other classes. Classes 2 and 3 were also quite accurate with the majority being correctly classified. On the test data, the model performed well with a moderate rise in misclassification, but predominantly between Classes 2 and 3; for instance, 51 samples of Class 2 were misclassified as Class 3, and 21 samples of Class 3 were misclassified as Class 2. On the unseen data, the model performed well with the same pattern of errors, with Classes 2 and 3 recording higher misclassification errors than Class 1, though the model still had a great classification ratio. It appears as if the model could have been even more accurate if the threshold had been adjusted, or more data had been added in the classes that were more difficult to identify. This is because the majority of the errors occurred between Classes 2 and 3.

The random forest model had excellent performance with a very high accuracy on training, test, and unseen data, evidencing the model's capacity to identify patterns and generalize new data well. The classification report values show that the average precision, recall, and F1 score were 99% to 100%, meaning that the model could make predictions with very little error. While the values for model evaluations are very high, it ought to be observed that the model might have overfit itself on the training data, which could lead to overfitting. Nevertheless, performance on test and unseen data were similar to the training data except with slightly lower accuracy, meaning that the model can still perform well with new data. Analysis using confusion matrix found that the majority were accurately classified on all three data but still with some misclassifications, notably between Classes 2 and 3 that persisted across all test stages. It could be that the model struggled with the differentiation between the two classes since the features used were very similar, meaning the classes were less differentiated. Nevertheless, the misclassifications are still a few against the total data being predicted correctly, so this does not have big implications for the model's performance.

Table 8 Random forest model classification report

	Train (%)			Test (%)			Unseen (%)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
1	100	100	100	100	100	100	100	100	100
2	100	100	100	99	99	99	99	99	99
3	100	100	100	99	100	99	99	99	99
Accuracy			100			99			99
Macro avg	100	100	100	99	100	99	99	99	99
Weighted avg	100	100	100	99	100	99	99	99	99

Table 9 Random forest model confusion matrix

		Predicted								
		Training (number of data)			Test (number of data)			Unseen (number of data)		
		Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Actual	Class 1	30599	8	5	6584	15	7	6617	16	10
	Class 2	48	27884	39	20	5870	51	30	5934	72
	Class 3	9	0	28569	8	21	6102	11	17	5971

Being an ensemble-based model, Random Forest possesses a variety of strengths, including being capable of dealing with complicated data, diminishing prediction variance, and being more overfitting-resistant compared to models based on individual decision trees. Nonetheless, a few enhancements can be added to optimize the outcome, including feature selection so as to identify the significant features in classifying so that the model can more efficiently identify similar classes and minimize the calculations. With consideration being placed on these areas, the model will be more precise in dealing with new data as well as being stable when exposed to changing conditions. In all, Random Forest proved to be the optimal model in this classification since it gave stable yet precise results, whereby there were still some errors in classification that could be minimized via optimization.

4 Conclusion

From the analysis output, we can state that Random Forest is the optimal model for classification in this research. It had very good accuracy for training, test data, and unseen data, with nearly perfect precision, recall, and F1-score. Its performance indicates that the model can identify patterns effectively and possesses good generalization with new data. While some misclassifications exist, especially between classes with similar features, the quantity is low and does not make much contribution to the accuracy. Random Forest's reliability in dealing with intricate data as well as the capacity for overfitting reduction made it suitable for use in this research.

In comparison with the past research, the model presented here demonstrates a considerable performance improvement. While the past research centered more on location and the characteristics of the driver, the present study employs a more comprehensive set of variables encompassing the environment conditions, the features of the accidents, as well as the road infrastructure. Moreover, the present study makes use of more advanced pre-processing and model tuning on Bayesian optimization, enhancing the accuracy and stability in prediction. Evaluation findings demonstrate that the model here has a higher accuracy at 99%, with an MAE at 0.0078 and an RMSE at 0.0982, lower than the past research with an MAE at 0.0874 and an RMSE at 0.1761. This demonstrates that the model developed here is more precise and more reliable for the analysis of traffic accidents. In general, the current research validates that using more useful variables, pre-processing well, and ideal model tuning can help enhance the performance with respect to the prediction of a crash, thereby acting as a more robust reference practice in transport security.

Acknowledgments

We extend our gratitude to the members of the RIS Laboratory who assisted us in the completion of this study.

Data availability statement

Data and code related to this study are available upon request from the corresponding author.

References

Ahmed, S., Hossain, A. 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, London, UK, pp. 390–397. ISBN 978-1-6654-6668-4 https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

Akhtar, M., Moridpour, S. (2021) "A Review of Traffic Congestion Prediction Using Artificial Intelligence", Journal of Advanced Transportation, 2021(1), 8878011. https://doi.org/10.1155/2021/8878011

Alqahtani, H., Kumar, G. (2024) "Machine learning for enhancing transportation security: A comprehensive analysis of electric and flying vehicle systems", Engineering Applications of Artificial Intelligence, 129, 107667. https://doi.org/10.1016/j.engappai.2023.107667

Azami, M. F. A. M., Misro, M. Y., Hamidun, R. (2024) "Road crash dynamics in Malaysia: Analysis of trends and patterns", Heliyon, 10(18), e37457. htpps://doi.org/10.1016/j.heliyon.2024.e37457

Belete, M. D., Alitasb, G. K., Nibretu, S., Dessie, M. E. (2025) "Road traffic accident determinant factor identification in case of East Gojjam, Ethiopia using wrapper feature selection algorithm", African Transport Studies, 3, 100018. https://doi.org/10.1016/j.aftran.2024.100018

Berhanu, Y., Schröder, D., Wodajo, B. T., Alemayehu, E. (2024) "Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks", Results in Engineering, 23, 102737. https://doi.org/10.1016/j.rineng.2024.102737

Çeven, S., Albayrak, A. (2024) "Traffic accident severity prediction with ensemble learning methods", Computers and Electrical Engineering, 114, 109101. https://doi.org/10.1016/j.compeleceng.2024.109101

Chai, H., Dong, K., Liang, Y., Han, Z., He, R. (2025) "Machine learning-based accidents analysis and risk early warning of hazardous materials transportation", Journal of Loss Prevention in the Process Industries, 95, 105594. https://doi.org/10.1016/j.jlp.2025.105594

Chand, A., Jayesh, S., Bhasi, A. B. (2021) "Road traffic accidents: An overview of data sources, analysis techniques and contributing factors", Materials Today: Proceedings, 47, pp. 5135–5141. https://doi.org/10.1016/j.matpr.2021.05.415

Data.gov.uk (2024a) "Road safety data-casualties", [online] Available at: https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-casualty-2024.csv [Accessed: 20 January 2025])

Data.gov.uk (2024b) "Road safety data-vehicles", [online] Available at: https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-vehicle-2024.csv [Accessed: 20 January 2025]

Data.gov.uk (2024c) "Road safety data-collisions", [online] Available at: https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-collision-2024.csv [Accessed: 20 January 2025]

Dimitrijevic, B., Asadi, R., Spasovic, L. (2023) "Application of hybrid support vector Machine models in analysis of work zone crash injury severity", Transportation Research Interdisciplinary Perspectives, 19, 100801. https://doi.org/10.1016/j.trip.2023.100801

Elalouf, A., Birfir, S., Rosenbloom, T. (2023) "Developing machine-learning-based models to diminish the severity of injuries sustained by pedestrians in road traffic incidents", Heliyon, 9(11), e21371. https://doi.org/10.1016/j.heliyon.2023.e21371

Gálvez-Pérez, D., Guirao, B., Ortuño, A. (2023) "Analysis of the Elderly Pedestrian Injury Severity in Urban Traffic Accidents in Spain Using Machine Learning Techniques", Transportation Research Procedia, 71, pp. 6–13. https://doi.org/10.1016/j.trpro.2023.11.051

Jalilian, M. M., Safarpour, H., Bazyar, J., Keykaleh, M. S., Malekyan, L., Khorshidi, A. (2019) "Environmental Related Risk Factors to Road Traffic Accidents in Ilam, Iran", Medical Archive, 73(3), pp.167–172. https://doi.org/10.5455/medarh.2019.73.169-172

Khosravi, Y., Hosseinali, F., Adresi, M. (2024) "Identifying accident prone areas and factors influencing the severity of crashes using machine learning and spatial analyses", Scientific Reports, 14(1), 29836. https://doi.org/10.1038/s41598-024-81121-7

Li, K., Xu, H., Liu, X. (2022) "Analysis and visualization of accidents severity based on LightGBM-TPE", Chaos, Solitons & Fractals, 157, 111987. https://doi.org/10.1016/j.chaos.2022.111987

Libnao, M., Misula, M., Andres, C., Mariñas, J., Fabregas, A. (2023) "Traffic incident prediction and classification system using naïve bayes algorithm", Procedia Computer Science, 227, pp. 316–325. https://doi.org/10.1016/j.procs.2023.10.530

Liu, L., Guevara, A., Sanchez-Galan, J. E. (2022) "Identification and classification of road traffic incidents in Panama City through the analysis of a social media stream and machine learning", Intelligent Systems with Applications, 16, 200158. https://doi.org/10.1016/j.iswa.2022.200158

Madushani, J. P. S. S., Sandamal, R. M. K., Meddage, D. P. P., Pasindu, H. R., Gomes, P. I. (2023) "Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers", Transportation Engineering, 13, 100190. https://doi.org/10.1016/j.treng.2023.100190

Obasi, I. C., Benson, C. (2023) "Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents", Heliyon, 9(8), e18812. https://doi.org/10.1016/j.heliyon.2023.e18812

Olayode, I. O., Du, B., Severino, A., Campisi, T., Alex, F. J. (2023) "Systematic literature review on the applications, impacts, and public perceptions of autonomous vehicles in road transportation system", Journal of Traffic and Transportation Engineering (English Edition), 10(6), pp. 1037–1060. https://doi.org/10.1016/j.jtte.2023.07.006

Özkan, E. K., Ulaş, H. B. (2024) "Comparison of four machine learning methods for occupational accidents based on national data on metal sector in Turkey 3", Safety Science, 174, 106468. https://doi.org/10.1016/j.ssci.2024.106468

Qiao, W., Huang, E., Zhang, M., Ma, X., Liu, D. (2025) "Risk influencing factors on the consequence of waterborne transportation accidents in China (2013–2023) based on data-driven machine learning", Reliability Engineering & System Safety, 257, 110829. https://doi.org/10.1016/j.ress.2025.110829

Santos, D., Saias, J., Quaresma, P., Nogueira, V. B. (2021) "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction", Computers, 10(12), 157. https://doi.org/10.3390/computers10120157

Shakil, F. A., Hossain, S. M., Hossain, R., Momen, S. (2022) "Prediction of Road Accidents Using Data Mining Techniques", In: Proceedings of International Conference on Computational Intelligence and Emerging Power System. Algorithms for Intelligent Systems, Ajmer, India, pp. 25–35. ISBN 978-981-16-4102-2 https://doi.org/10.1007/978-981-16-4103-9_3

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., Hamedat, O. (2021) "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression", Automation in Construction, 129, 103827. https://doi.org/10.1016/j.autcon.2021.103827

Silagyi, D. V., Liu, D. (2023) "Prediction of severity of aviation landing accidents using support vector machine models", Accident Analysis & Prevention, 187, 107043. https://doi.org/10.1016/j.aap.2023.107043

Singh, A., Murzello, Y., Pokhrel, S., Samuel, S. (2025) "An investigation of supervised machine learning models for predicting drivers' ethical decisions in autonomous vehicles", Decision Analytics Journal, 14, 100548. https://doi.org/10.1016/j.dajour.2025.100548

Sun, W., Abdullah, L. N., Khalid, F. B., Sulaiman, P. S. B. (2025) "Classification of traffic accidents' factors using TrafficRiskClassifier" International Journal of Transportation Science and Technology, 17, pp. 328–344. https://doi.org/10.1016/j.ijtst.2024.05.002

United Nations Economic Commission for Europe (2021) "A Foundational Safety System Concept to Make Roads Safer in the Decade 2021-2030", In: The concept of a national road safety system, United Nations, pp 37–44. ISBN 9789210049177 https://doi.org/10.18356/9789210049177c007

World Health Organization (2018) "World health statistics", [pdf] Available at: http://apps.who.int/iris/bitstream/handle/10665/272596/9789241565585-eng.pdf?ua=1, 2018 [Accessed 11 April 2024]

Xu, M., Liu, H., Yang, H. (2024) "Ensemble learning based approach for traffic incident detection and multi-category classification", Engineering Applications of Artificial Intelligence, 132, 107933. https://doi.org/10.1016/j.engappai.2024.107933