STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, September 2026, pp 2120–2136.
Published online in International Academic Press (www.IAPress.org)

Prediction of accident severity for overtaking maneuvers on two-lane roads using Decision Tree, Random Forest, and K-Nearest Neighbors models

Sadir A. Fadhil [1,*], Israa Ahmed Jaddoa [2]

[1]*Computer Science and Information Technology, University of Anbar, 31001 Ramadi, P.O.B. 55, Iraq*
[2]*Department of Computer Sciences, College of Science, University of Al Maarif, Al Anbar 31001, Iraq*

Abstract Overtaking maneuvers are dangerous and difficult to execute on two-lane roads; due to the nature of such a maneuver, there is a pressing need to improve the scientific research, because there is a lack of overtaking-specific risk assessment in existing traffic safety systems and driver assistance technologies. The study examines how three machine learning algorithms, such as Random Forest, Decision Tree, and K-Nearest Neighbors, can be used to predict the severity of an accident based on multiple contributing factors, such as road surface conditions, weather, driver characteristics, vehicle conditions, and others. The models were trained and tested on accident datasets provided by the UK Department of Transport, and their performance was measured by various evaluation measures such as precision, accuracy, recall, and F1-score. A correct validation procedure was implemented to guarantee the reliability of the results, and the problem of class imbalance was solved with the Synthetic Minority Over-sampling Technique (SMOTE). The results show that the Random Forest model has better performance than the other models, with a higher predictive performance, followed by the Decision Tree and KNN models with a slightly lower performance. Although moderate accuracy has been obtained in general, it has been found that SMOTE improves the accuracy of serious injury detection, which is essential for safety applications. This study offers a valid framework that transportation authorities could use to gain a better understanding of the patterns of accident severity and helps to develop targeted safety measures, emphasizing the power of machine learning methods to detect the most significant contributing factors to traffic accidents.

Keywords Overtaking Maneuver, machine learning, safety system, vehicle accidents, classification model

DOI: 10.19139/soic-2310-5070-3553

1. Introduction

One of the serious threats to public safety of the people comes from road traffic accidents, with a high number of deaths being one of the main issues. Understanding the factors that indicate the severity and fatality of accidents is essential for the development of effective preventive measures. As a consequence, there is a crucial need to specifically predict accident severity, since existing traffic safety systems and policies are predominantly focused on general accident analysis, rather than overtaking-specific risks. For example, the majority of Advanced Driver Assistance Systems (ADAS) prioritise lane departure warning and front collision prevention, with poor ability to predict risk associated with overtaking on two-lane roads. To measure the accident severity, standard statistical tools have been extensively employed, yet, conversely, more advanced tools that would help to resolve this problem with the aid of the improvement of Machine Learning (ML) are available [1]. To simulate the intensity of traffic accidents, three popular machine learning models are explored in this study to analyze and predict the severity of traffic accidents associated with overtaking maneuvers on two-lane roads, including Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbor (KNN). Their strong classification performance contributes to their

*Correspondence to: Sadir Abdulwahed Fadhil (Email: Fadhil-academia@uoanbar.edu.iq). Computer Science and Information Technology, University of Anbar, 31001 Ramadi, P.O.B. 55, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

widespread use, especially in their ability to deal with big dataset sizes and capturing the non-linear relationships among the input variables [2, 3].

This study aimed to give a meaningful insight to the traffic safety authorities so they could introduce the improved road safety rules, as well as target actions by concentrating on traffic information about the type of fuel, engine capacity, road surface quality, weather conditions, type of vehicle, behavior of the driver, and others. The results may help create a more sophisticated system for predicting accidents that may assist in enhancing traffic management policies and reducing the number of victims [4, 5]. The primary goal of this research is the comparison of different algorithms based on the importance of variables and the accuracy of prediction, in addition to identifying and evaluating the key variables affecting the level of accidents. The definition of accident severity in this study is based on a binary classification, aggregating fatal and serious injuries into the "serious" class. This is because the number of fatal accidents in the dataset is small, and class imbalance could affect the model's stability. Nevertheless, this approach limits the clinical detail and may oversimplify the differences between fatal and non-fatal serious injuries. As such, the findings need to be interpreted with this consideration, and future research may explore extending this framework to multi-class severity predictions to account for these differences.

In line with that, to forecast accidents, past investigations have applied logistic regression and other machine learning methodologies, but a more detailed analysis of the variables that impact the outcome of accidents can be acquired by applying ensemble algorithms such as RF and non-parametric algorithms such as DT [6]. This study contrasts with the majority of the existing literature that addresses the severity of traffic accidents based on aggregated data, but on a particular overtaking maneuver on two-lane rural roads, which is one of the riskiest driving habits. Overtaking entails the momentary utilization of conflicting traffic lanes, low visibility, and high relative speed, which can greatly raise the possibility and severity of collisions. These risk conditions are radically different than general accidents situations, when traffic flows are more organized and predictable. Thus, the novelty of the current study is not in the isolation of overtaking maneuver steps, which has been the subject of previous studies, but rather, the systematic exploration using machine learning models to predict accident severity for overtaking maneuvers, using a filtered dataset, considering only two-lane rural roads.

This analysis, which focuses on this specific maneuver, provides behavior-based bits of information not provided by generic data-driven accident severity models. This particular approach enables a greater understanding of highly risky situations and enables the development of special safety measures.

Additionally, the effects of the explanatory variables are different in overtaking accidents than in general accident contexts, because of the characteristics of overtaking. Overtaking is time-dependent (perception, decision, and execution have to be completed within a limited time frame). Thus, environmental factors such as lighting, weather, and time of day mainly impact visibility and gap size; road surface conditions impact vehicle control during lane change; driver factors such as age and gender impact reaction time and risk-taking; and vehicle factors such as engine size and vehicle age impact acceleration and overtaking completion. While these variables may not be as dependent in other accident scenarios, overtaking is a complex interaction where these factors combine to influence whether the overtaking maneuver can be safely completed. This justifies the need to consider overtaking accidents as a specific behavioral scenario. This verifies that in overtaking, these variables do not operate in isolation but interact to affect perception, decision making, and execution of the maneuver in a short period of time. It's worth noting that our research is not novel in isolating overtaking maneuvers, as this has been investigated in previous studies [19], but in the systematic use and comparison of machine learning models for accident severity prediction in a specific driver behavior scenario. Compared to previous studies that address decision making or large accident data sets, the focus here is on severity prediction in overtaking situations and is accompanied by a behavior-specific risk analysis.

The rest of this paper is organized in the following way. The following section provides a review of previous research on the use of machine learning techniques for accident severity prediction. Section 3 outlines the dataset and the data preprocessing process. Section 4 is the description of the methodology used and the three classification models: DT, RF, and KNN. The performance measures used to evaluate the model performance are described in Section 5. The results and their discussion are presented in Section 6, and the conclusion is described in Section 7 along with the main findings and recommendations for future research.

2. Literature Review

Road traffic accidents continue to be a serious public safety problem worldwide, resulting in a considerable number of fatalities and injuries every year. Traffic accident severity is an important factor in traffic safety management because the results of the traffic accidents can range from slight injuries to fatalities. The literature review synthesizes the past work on accident results analysis, prediction based on statistical methods, as well as machine learning [7].

Chen and Chen [8] compared the predictive capabilities of RF, Classification and Regression Tree (CART), and Logistic Regression (LR) for the development of the accident severity. They established that RF worked better than both LR and CART since it was capable of complex interaction between predictors; thus, resulting in better accuracy, sensitivity, and specificity.

There are several research papers that have provided a comparison of machine learning algorithms used in predicting the severity of accidents. As illustrated by Vanitha and Swedha [2], the predictive accuracy and classification performance of the RF were found to be better than other classifiers such as LR, KNN, and Na"ive Bayes. Another critical point raised by the study was the need to consider different contributing factors, including road conditions, the type of vehicles, and weather, in forecasting the level of accidents. Similarly, Khattak et al. (2021) [6] established that using ensemble models such as the RF, when using feature selection methods, provided more accurate predictions of data than single-model classifiers such as the KNN and LR.

One of the important contributions to this area was Uddin and Lu [4], who statistically confirmed the superiority of the tree-based algorithms, especially the RF, over non-tree-based models, such as KNN and the LR. They conducted an analysis of 200 open-access datasets in various fields and found that the test of the RF was always better in terms of accuracy, precision, recall, and F1 score compared to the KNN and other algorithms.

Predicting weather data into accident severity prediction models has also been an area of interest in studies done in recent times. A study conducted by Bhavani [9] indicated the importance of weather conditions in the prediction of accidents. Random Forest achieved higher accuracy when weather data was included in the model. In this study, RF, with the addition of weather factors, including rainfall intensity and road conditions, was capable of producing the highest accuracy in predicting accidents at 83.2%; according to the KNN, in this case, they only produced the highest accuracy at 61.25%.

Random Forest is one of such machine learning techniques that has proven to be the most efficient when it comes to the prediction of traffic accident severity because of its capacity to process high-dimensional data and the fact that it is not subject to overfitting. Yan and Shen [10] came up with a hybrid model combining RF with Bayesian optimization (BO) to search for its parameters, which was more accurate than traditional algorithms. They showed that RF with Bayesian optimization not only enhances prediction accuracy, but it also has interpretable results in the form of feature importances and partial dependence curves.

Other works that have used RF include Dadashova et al. [11]; they used RF together with discrete-choice models to predict accidents on two Trans-European routes in Spain. They found that RF was better than traditional models; however, roadway design aspects, including curvature, elevation, lane, and shoulder width, were found to be the most significant predictors of severity. The paper did not pay much attention to the environment and driver-related issues because this limited its use in a more complicated real-world situation. Altogether, the study demonstrates that the combination of ML and statistical models can be valuable, and the importance of roadway design must be underlined, but the inclusion of additional factors should be broadened in future research.

Aldhari et al. [12] were another group that conducted an experiment to predict the severity of crash injuries in Saudi Arabia with the help of RF and XGBoost and logistic regression. This analysis demonstrated that XGBoost was more successful than RF when it comes to classification accuracy of 94 percent for binary classification and 71 percent for multi-class classification.

Machine learning algorithms have been used to discover and measure the various elements that influence the severity of traffic accidents. An example is that weather conditions like temperature, humidity, and visibility are great predictors. Pillajo-Quijia et al. [13] applied weather-related variables to predict the injury severity of light trucks and vans. The research revealed that the state of psychophysical conditions, such as alcohol and sleep deprivation, played a significant role in the intensity of the accidents.

Driving behaviour, including whether they obey traffic lights or not, and performing dangerous overtaking moves, are critical factors in the outcome of the accidents. The study by Figueira and Larocca [14] was based on an overtaking behavior, and they discovered that there are more chances of overtaking among young male drivers, which resulted in serious injuries.

Shirwaikar et al. [15] created a machine learning model to forecast the accident's severity based on weather-related and contextual parameters. They used the Select-K-Best and SMOTE algorithms of feature selection and data balancing. Random Forest Classifier (RFC) had an AUC of 95, and the stacking ensemble model had a higher AUC of 96.92. The research indicated the usefulness of ensemble learning in improving prediction accuracy and dealt with class imbalance. Nonetheless, it might be restricted by the dataset used to generalize its use.

Azhar et al. [16] examined the severity of injury in severe accidents in Malaysia with the help of CART and RF. The research was based on the M-ROADS crash data and classified the results as fatal, severe, slight, and none. Among the major determinants were the type of collision, driver errors, vehicles involved, age of the driver, the lighting conditions, and the type of vehicle. Both CART and RF worked; however, RF was slightly more accurate because it was an ensemble structure. Notably, the analysis revealed that machine learning had the ability to address the assumption violation of conventional ordinal regression models, and thus it can be particularly applicable in the analysis of injury severity, where the data distribution is uneven.

In contrast to Li et al. [17], who were concerned with predicting the accident's severity on mountainous freeways in China, where the complex topography and unfavorable weather conditions increase hazards. With the SVM, Decision Tree Classifier (DTC), AdaBoosted SVM (Ada_SVM), and AdaBoosted DTC (Ada_DTC), the study factored in dynamic traffic and weather parameters (e.g., rainfall intensity, road alignment, traffic flow, speed variation, and vehicle mix). A feature selection tool was applied, and the RF was used; the type of predictors found to be the most important were the following: collision type, rainfall intensity, road section type, and number of vehicles. The feature selection enhanced the accuracy of the model by as much as 9.3%. Ada_SVM using RF-selected features had the highest model performance at 88.4%. It shows that including environmental and real-time traffic variables in machine learning models is also an added value.

It is noteworthy that some recent research (e.g., Li et al. [17]) includes dynamic traffic variables and real-time weather information (e.g., precipitation and traffic flow), which may improve the prediction outcomes. However, the present study uses structured categorical variables sourced from accident reports. Although this allows consistent and interpretable data, it may result in less representation of the real-time driving environment, which could be regarded as a constraint of the current study.

In addition, Cicek et al. [18] evaluated five predictors (Decision Trees, Multilayer Perceptron Neural Networks, Naive Bayes, Case-Based Reasoning, and Support Vector Classifier) on the severity of accidents with the U.S. NHTSA crash data. One of the new findings of this study is the use of explainable machine learning using Shapley values (SHAP), which quantified the contribution of each variable to the outcome of injury severity. The findings indicated that the strongest predictors were the use of seatbelts, alcohol, and speeding, as reported in the literature. The Multilayer Perceptron model produced the most accurate results, although SHAP analysis enhanced the explainability ability of black-box models, which made an important gap in the literature that prioritized prediction over explanations.

Lastly, Fadhil and Al-Bayyati [29] proposed a driver assistance system for overtaking on two-lane roads using predictive models to support safer decision-making. Their work focused on real-time assistance by evaluating overtaking conditions and providing guidance to drivers. However, the study has mainly taken into account the question of maneuver feasibility and safety decisions, and not the question of accident severity. The present study, on the other hand, aims to perform modeling and prediction of the severity of overtaking-related accidents by applying machine learning techniques, to complement the current state of overtaking assistance with severity-based analysis.

Accordingly, previous research has investigated overtaking behavior and related risks; however, most of these studies are related to the driver behavior or decision support, rather than accident severity prediction, as indicated in Table 1. Specifically, a limited number of studies have been conducted on the modeling of accident severity in a specific driving context, such as overtaking on two-lane rural roads, where vehicles temporarily drive in the opposite lane and are at higher risk of collision. In our earlier research [19], Bayesian Networks were applied to

assist overtaking decisions in terms of probabilistic relationships between various factors. But this was for decision-making not severity prediction. However, the present study uses and compares several machine learning models (DT, RF, and KNN) to systematically predict the severity of overtaking accidents, moving away from probability inference to a data-driven approach of accident severity classification.

Table 1. Comparison of Feature Sets and Prediction Approaches

Study	Focus	Prediction Target	Feature Type	Key Features Used	Data Type	Contribution
Figueira & Larocca [14]	Overtaking behavior	Risk behavior	Behavioral	Driver age, gender	Behavioral / survey	Identifies risky overtaking patterns
Fadhil [19]	Overtaking assistance	Decision support	Probabilistic	Driver, road, traffic variables	Structured	Bayesian model for overtaking decisions
Fadhil & Al-Bayyati [29]	Overtaking assistance system	Maneuver feasibility	Predictive / behavioral	Traffic conditions, driver behavior	Structured+ predictive	Real-time overtaking support system
Li et al. [17]	General accidents	Severity prediction	Dynamic + environmental	Rainfall, traffic flow, speed variation	Real-time+ structured	High accuracy with dynamic variables
Azhar et al. [16]	General accidents	Injury severity	Mixed	Driver age, lighting, collision type	Structured	ML-based severity classification
Shirwaikar et al. [15]	General accidents	Severity prediction	Mixed	Weather, road, contextual variables	Structured	High-performance ensemble models
This study	Overtaking accidents (two-lane roads)	Severity prediction	Structured categorical	Lighting, driver age, road surface, weather, vehicle features	Structured (STAT19)	ML-based severity prediction for overtaking

3. Dataset

In this research, we obtained accident data (STAT19) for overtaking maneuver accidents on two-lane roads from the UK Department for Transport (DfT) over three years for the period 2013-2020. The original dataset is police reports, consisting of more than one million records, in addition to 68 feature variables. This dataset is refined in order to obtain only the overtaking maneuver accidents that occurred on two-lane roads, as depicted in Figure 1. In addition, all refining steps for the original dataset are available in [19].

The applied refining process in this study includes the following steps:

1. Analyze all variables across the three dataset files (accidents, vehicles, and casualties) to identify the most relevant variables for overtaking maneuvers.
2. After selecting the required variables, remove the irrelevant ones to reduce the dataset size before merging the files. The merging process is based on the unique "accident index" variable, which is present in all three datasets.
3. Refine the selected variables by eliminating irrelevant data entries. For example, selecting only overtaking maneuver values from the "Vehicle Maneuver" variable, selecting cars and minibuses from the "Vehicle Type" variable, selecting single carriageway roads from the "Road Type" variable, and selecting the rural area category (instead of urban or other area types) from the "Home Area Type" variable.

4. Detect and filter out instances with missing, unknown, and invalid values (according to the STAT19 dataset encoding). Rather than imputing these data, they were removed to improve data quality and avoid bias (especially as such data represent a relatively small percentage of the final dataset).
5. Finally, convert data labels into integer values for consistency and easier analysis.

The obtained dataset consists of (10710) accident records, in addition to the most important variables that have a great impact on performing an overtaking manoeuvre. Ten independent variables (explanatory variables) were chosen from the dataset to run the models, and only one target variable. This data is organized into seven distinct categories: temporal attributes, road characteristics, vehicle-related features, driver-related factors, and environmental conditions. Three categories constitute the target variable, which is the accident severity: fatal, serious, and slight injuries. All the variables used in the dataset are significant and diversified, which assists in providing a thorough consideration of all influential factors.

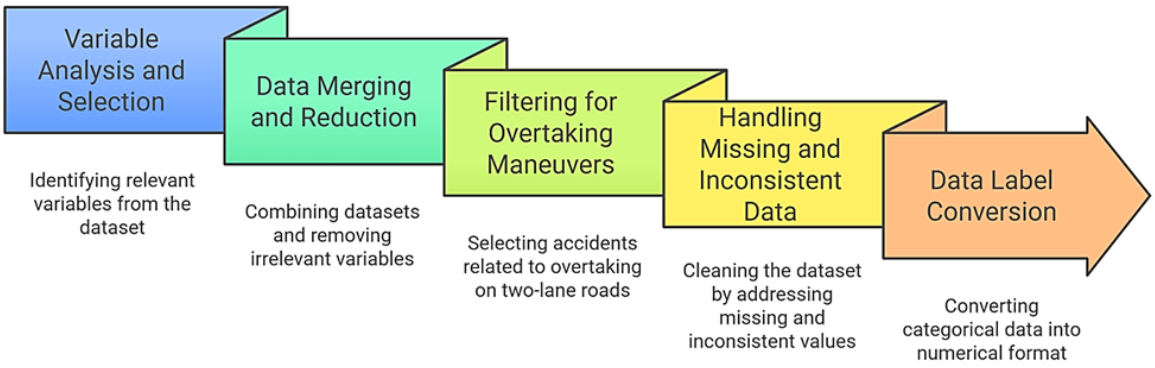

Figure 1. Dataset refining steps

As shown in Table 2, all variables, except for Time, driver age, Engine capacity, and vehicle age, are discrete (categorical); these variables are being divided into logical bands. For instance, the variable "Time" was divided into four intervals, each of six hours. The initial road accident data categorizes accidents by severity into three types (target variable): fatality, serious injury, and slight injury. Each accident may be associated with one or more of these severity levels. For this research, accidents that involve a fatality are classified as "fatal," which refers to the causes of death resulting from an accident; whereas those with injuries but no fatalities are labeled as "injury". The serious injury resulted in a hospitalized person who suffered permanent injuries but did not die, while the slight injury is a person who has been injured but does not require medical attention [20].

In this study, the relatively low number of fatal accidents, which might skew analytical results. To address this issue, fatal accidents (428 cases) and serious injuries (2139 cases) are grouped under the label "serious injuries" (2567 cases). This results in better class balance, but reduces the ability to differentiate between fatal and other severe injuries. In contrast, accidents that only involve slight injuries are classified as "minor injuries" [21]. As a result, accident severity is classified into two main categories: "serious" and "minor injuries ". Further, the categorization of other variables in the dataset has been similarly consolidated appropriately [1].

Table 2. Descriptive statistics of accident data

Variables	Value	Total Number of Accidents = 10710	
		Count	Percent(%)
Dependent variable:			
Accident severity	1. serious injuries	2567	24.0
	2. minor injuries	8142	76.0
Independent variable:			
Weather	1. fine	9082	84.8
	2. raining	1559	14.6
	3. other	68	0.6
Road surface	1. dry	7164	66.9
	2. other	3545	33.1
Light Conditions	1. daylight	8246	77.0
	2. darkness_light_lit	758	7.1
	3. darkness_no_light	1705	15.9
Day of week	1. weekend	2743	25.6
	2. workday	7966	74.4
Time	1. 01:00–5:59	534	5.0
	2. 06:00–11:59	3744	35.0
	3. 12:00–17:59	4634	43.3
	4. 18:00–24:59	1797	16.8
Driver gender	1. male	7854	73.3
	2. female	2855	26.7
Driver age	1. 16 to 25	3362	31.4
	2. 26 to 65	6389	59.7
	3. $>$ 65	958	8.9
Fuel type	1. petrol	7734	72.2
	2. other	2975	27.8
Engine capacity (cc)	1. 429 to 1500	3704	34.6
	2. 1527 to 2000	5597	52.3
	3. $>$ 2000	1408	13.1
Vehicle age	1. 1 to 5	3669	34.3
	2. 6 to 10	4152	38.8
	3. $>$ 10	2888	27.0

4. Methodology

This study employs three supervised machine learning classification models, DT, RF, and KNN, to analyze and predict the severity of traffic accidents associated with overtaking maneuvers on two-lane roads. The adopted dataset in this research was divided into training (70%) and testing (30%) subsets using stratified sampling to preserve the class distribution of accident severity. All models were trained exclusively on the training dataset and evaluated on the unseen test dataset to ensure independence and prevent data leakage. Additionally, to further improve robustness, stratified five-fold cross-validation was applied on the training set during model development, and using a fixed random seed, all experiments were conducted to ensure reproducibility of the data split and model training process.

The Synthetic Minority Over-sampling Technique (SMOTE) was used to oversample the training data only to correct the imbalance of classes (76% minor vs. 24% serious injuries). This guaranteed the equal representation of

the two classes and maintained the test set to be unbiased. Likewise, all the preprocessing, such as encoding and resampling, was done post-train-test split and only fitted to the training data to prevent data leakage. In addition, the data features were scaled to enhance model performance, especially for the KNN algorithm, which is sensitive to feature scale. In particular, the StandardScaler method has been applied to normalize the numerical variables by bringing their means to zero and making them of unit variance. This is performed to make sure that each feature plays an equal role in the distance determination in the KNN, thus enhancing the accuracy and stability of classification. Feature scaling was fitted on the training data and then applied to the test data to prevent data leakage. DT and RF tree-based models were also not influenced by the scaling of features because of their structure.

The chosen models cover a range of algorithmic approaches that allow for performance comparisons. The DT is employed as an interpretable baseline, it is able to capture complex nonlinear relationships between the various categorical and continuous features. But it is susceptible to overfitting (for large trees) [22], [8]. Whereas, the RF model builds on the DT model by employing a collection of trees learned from random samples of the data and features. This helps to avoid overfitting and enhance its predictive accuracy, making the RF model suitable for accident severity prediction [23], [4], [8], [17]. The KNN model is used as a distance-based classifier. This model predicts class labels based on the majority class of the closest data points in the feature space and is influenced by the scale of features and distribution of the data [24], [5].

In this study, k = 5 neighbors of the model in KNN were determined as a result of systematic tuning and initial experiments. Various values of (k) were tried to balance between bias and variance, and k = 5 was chosen because it gave the best trade-off between the classification accuracy and the model stability. KNN has been effectively used in various studies of predicting the severity of accidents, although it does not do as well as more complicated ensemble predictors, such as RF [4], [25].

The hyperparameters of the machine learning models were tuned to achieve the best performance as shown in Table 3. For the Random Forest model, the number of trees (n_estimators), the maximum depth trees, and the minimum number of samples per split were explored. The last set of parameters (nestimators = 300, depth unlimited) was selected based on empirical studies, which allows enough flexibility in the model and generalization at the same time. For the Decision Tree model, the maximum depth and the minimum samples per leaf were tested as parameters in order to avoid overfitting and to be interpretable. A medium depth was selected to ensure the model has a good generalization performance. In the case of the KNN model, a grid search was performed on the (k) value, and the optimum value (k = 5) was selected. Rather than grid search, a manual search was carried out based on the results of cross-validation to find the optimal model, and to reduce the computational cost of tuning the model while being able to achieve good performance.

Table 3. Hyperparameter Tuning Ranges and Final Selection

Model	**Parameter**	**Tested Range**	**Final Value**
RF	n_estimators	100–500	300
RF	max_depth	None, 10–30	None
RF	min_samples_split	2–10	2
DT	max_depth	5–30	$\sim$15
DT	min_samples_leaf	1–10	2
KNN	k	3–15	5

5. Evaluation Metrics

The effectiveness of the classification models in the classification severity of traffic accidents was evaluated in four traditional evaluation measures, including accuracy, precision, recall, and the F1-score. They are commonly used in the studies of traffic accident severity to assess the strength and forecasting quality of machine learning classifiers

[26, 27, 28]. All the measures represent a distinctive view on the performance of the model, as the predicted model outcomes are compared to the actual outcomes as presented by the confusion matrix.

1. Accuracy (Acc): is the approximate correctness of the classifier given by the fraction of correctly predicted instances to the total number of instances, which is given by equation (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Where TP = True Positives and TN = True Negatives represent the count of correctly predicted positive and negative cases, respectively, while FP = False Positives and FN = False Negatives represent the misclassifications. Though accuracy gives a general performance measure, it is not reliable in the case of imbalanced data, which is characteristic of accident severity analysis.

2. Precision (P): measures the performance of the classifier to find the true positives out of all the predicted positives; given by Equation (2):

$$Precision \ = \frac{TP}{TP \ + FP} \tag{2}$$

High precision implies that the false alarm rate is low, which implies that the model makes few severe accidents, when they are not severe.

3. Recall (R): also known as the sensitivity or the true positive rate, is the ability of the model to recognize all real positive cases properly. It is defined in Equation (3):

$$Recall \ = \frac{TP}{TP \ + FN} \tag{3}$$

A higher value of recall means the model is able to identify severe accidents, which is crucial in safety-related applications where failure to identify a severe crash may have severe implications.

4. F1-score (F1): offers a balance between precision and recall that intersects both the measures of classification performance into one. It is calculated as follows in Equation (4):

$$F1 = \frac{2 \times precision \ \times recall}{precision + recall} \tag{4}$$

The F1 score is particularly useful for evaluating models that are trained on an imbalanced dataset where there will be significant differences in precision and recall. In accident severity prediction, a high F1-score means that the model reduces false positives and false negatives.

The threshold assessment for the model performance was achieved with these metrics in this research. They were used to compare the DT, RF, and KNN models for prediction of the severity of accidents due to overtaking at two-lane roads. The use of these indicators together, as stated in the previous studies [26], [27] helps the evaluation of the efficiency of classification in the theoreticalization of traffic safety, which is balanced and interpretable.

6. Results and Discussion

This paper has constructed three machine learning classification models, namely DT, RF, and KNN, to examine and estimate the severity of accidents related to overtaking maneuvers on two-lane rural roads. Overtaking is among the most dangerous behaviors on such roads that are characterized by low visibility, opposite traffic, and differentials in speed that add to the risk of accidents [29]. To train the models, a comprehensive dataset that entailed roadway, vehicle, and environmental factors was used in order to determine the effect of the conditions on the outcome of the accidents.

The models were used to identify the levels of severity of crashes, which allowed to identify the most important determinants of severe crashes associated with overtaking. The comparison of DT, RF, and KNN models offers information about their predictive accuracy, stability, and possible usage in the traffic safety analysis. All the models were evaluated by stratified five-fold cross-validation to guarantee the strength of the evaluation and reported

mean $\pm$ standard deviation. This offers a more stable model performance comparison as it takes into consideration variability between splits of data.

It is essential to point out that previous versions of this study reported close to perfect classification results due to data leakage during the preprocessing step. In particular, the preprocessing steps (resampling and scaling) were not only applied to the training set, but also to the test set, "leaking" information into the model training. In the revised approach, all preprocessing is done on the training data after the train-test split, ensuring the test set is not seen during model development. This modification results in better performance measurements.

The performances of the three classification models were evaluated using key performance indicators like precision, recall, F1-score and overall accuracy. These measures allow a combined insight into each model in terms of predictive potential and validity in the assessment of traffic accident severity.

6.1. Description of the Decision Tree Model

Table 3 summarizes the performance evaluation of the Decision Tree classification model. Once validation was fixed and the SMOTE was used to resolve the issue of class imbalance, the DT model obtained a total accuracy of about 0.72, which implies a sensible performance of classification on the hidden test data. In Class 0 (serious injuries), the model obtained a precision of 0.68, a recall of 0.70, and an F1-score of 0.69. The model recorded marginally higher values in Class 1 (minor injuries), where the precision and recall are 0.77 and 0.75, respectively, and the F1-score is 0.76. These findings suggest that the model is effective in detecting minor injuries, but it has reasonable results in detecting serious injuries.

The precision, recall, and F1-score were about 0.73, 0.73, and 0.73, respectively, and were obtained by using the macro average, where both classes are assigned equal weight, irrespective of their distribution. This shows a moderate ability to classify in both classes. Conversely, the weighted average that takes into account the distribution of the classes gave a value of about 0.74 on exactness, recall, and F1-score. Moreover, the confusion matrix in Figure 2 indicates that 540 cases were correctly identified, and 230 cases were wrongly classified as Class 1. Likewise, 1830 cases of Class 1 were correctly identified, and 610 cases had been erroneously identified as Class 0. Misclassifications in both classes indicate the complex nature of the accident severity prediction process, in which there are overlapping patterns and uncertainty in the real-world data. The values presented along the diagonal of the confusion matrix show that the model still has a high level of discriminative power, but not as high as initially stated.

The revised results are more realistic and reflect how the proper generalization would be achieved after correcting the data leakage issue as described earlier. In contrast to the previous results, which implied near-perfect classification. In general, the Decision Tree model shows a stable and interpretable classification performance of the accident severity. It is not as precise as more sophisticated ensemble techniques, but it is still helpful to learn about the decision rules and determine the main contributing factors when it comes to traffic accidents.

Table 4. Results of the DT model

Metric	Class 0	Class 1	Accuracy	Macro Average	Weighted Average
Precision	0.68	0.77		0.73	0.74
Recall	0.70	0.75		0.73	0.74
F1-score	0.69	0.76		0.73	0.74
Accuracy			0.72		

6.2. Description of the Random Forest Model

The RF model with properly corrected validation methodology and SMOTE on the imbalance of the classes achieved an accuracy of approximately 0.75, indicating good predictive performance in the unseen test dataset, as shown in Table 5. The model in Class 0 (serious injuries) had a precision of 0.71, a recall of 0.78, and an F1-score of 0.74. In Class 1 (minor injuries), the values were 0.80, 0.73, and 0.76, respectively. These findings suggest

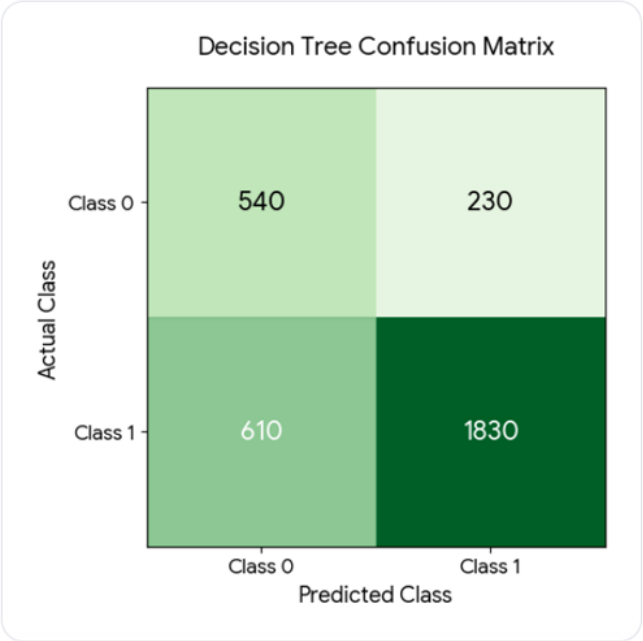

Figure 2. Confusion matrix of the decision tree classifier

that the model has a balanced performance in both classes, and they are capable of detecting serious injuries better than the imbalanced case.

The average values of macro averages of precision, recall, and F1-score were about 0.76, 0.76, and 0.75. This is indicative of a stable performance in the classification of both classes when there is an equal importance. Conversely, the weighted average scores were about 0.76 in terms of precision, recall, and F1-score, meaning that the model can be said to have a stable performance, taking into account the class distribution. The fact that the close values of macro and weighted averages resulted in a well-balanced classification without any major bias on any of the classes validates the fact that the model is well-balanced in its classification. Similarly, the confusion matrix in Figure 3 indicates that 600 cases of Class 0 were accurately identified, whereas 170 cases were falsely identified as Class 1. Likewise, 1790 cases of Class 1 were rightly identified, but 650 cases were wrongly determined to be Class 0. Such misclassifications are common in real-world data and are a sign of the inherent uncertainty and overlap in accident severity classes.

In contrast to the initial results, which indicated the ideal classification performance, the updated results are more realistic and in line with the studies concerning traffic accident prediction. The methodological problems included data leakage caused by improper preprocessing, as explained earlier. Once these problems are fixed, the RF model will exhibit good generalization and will not overfit. In general, the Random Forest model is the most successful of the models considered because of its ensemble learning mechanism that increases predictive accuracy and decreases variance. Its capability to model nonlinear relationships, including complex relationships, makes it especially good for the modeling analysis of accident severity, particularly when used with suitable data balancing methods, like SMOTE.

6.3. Description of the K-Nearest Neighbor Model

Using the K-Nearest Neighbor model, the results are classified according to the severity of traffic accidents as shown in Table 6. After correcting the validation process and using SMOTE to address class imbalance, the KNN model achieved a general accuracy of approximately 0.70, which is considered moderate and has a good predictive power on the unseen test data sets. In Class 0, the model had a precision, recall, and F1-score of 0.65, 0.67, and 0.66, respectively. The model showed a little higher value in Class 1 (minor injuries), the precision was 0.76, the

Table 5. Results of the RF model

Metric	Class 0	Class 1	Accuracy	Macro Average	Weighted Average
Precision	0.71	0.80		0.76	0.76
Recall	0.78	0.73		0.76	0.76
F1-score	0.74	0.76		0.75	0.76
Accuracy			0.75		

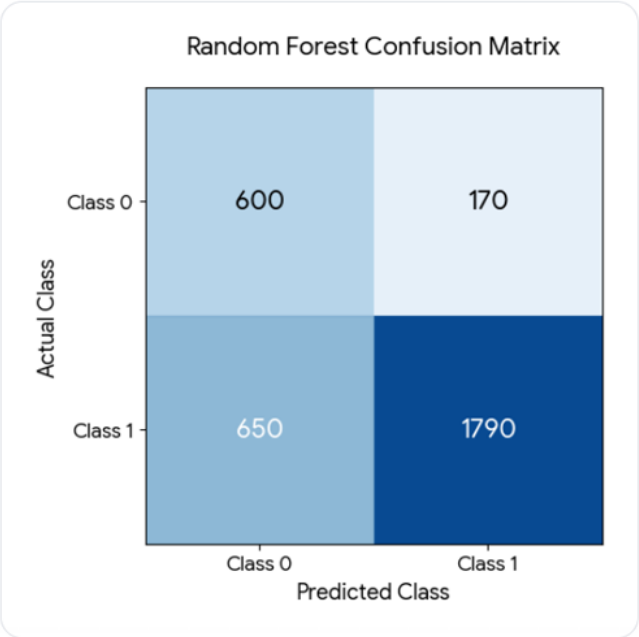

Figure 3. Confusion matrix of the Random Forest classifier

recall was 0.74, and the F1-score was 0.75. These findings reveal that the model is more effective in detecting minor injuries, but its effectiveness in detecting serious injuries is relatively low.

The macro averages of the precision, recall, and F1-score were around 0.71, 0.71, and 0.71, respectively. The values show the total capability of the model to classify when both classes are equally important. The weighted averages were about 0.72 in terms of precision, recall, and F1-score. The fact that there is a close resemblance between the macro and weighted averages that the model has a relatively comparable performance in the classes, although there was an imbalance in the initial dataset. According to the confusion matrix in Figure 4, 520 cases of Class 0 were correctly identified, whereas 250 cases were misclassified as Class 1. Likewise, 1790 Class 1 cases were correctly predicted, and 650 cases were misclassified as Class 0. These misclassifications are indicative of the inability to separate the level of accident severity because of overlapping patterns of features.

The revised performance is more realistic and indicates the correct model evaluation compared to the previous results, which presumed a higher accuracy, due to the removal of possible data leakage. The lower performance of KNN is due to its susceptibility to feature scaling and relies on distance-based computations, which are not as powerful in data sets involving complex nonlinear relations. In general, the KNN model has a good classification performance, but it is not as strong as the tree-based models, especially the Random Forest model, which has a better ability to capture complex trends in the data of accident severity.

The overall comparison implies that the three classifiers differ in performance in a specific manner, as shown in Table 7. RF model was the most successful model, as it had the highest accuracy of about 0.75 and a higher

Table 6. Results of the KNN model

Metric	Class 0	Class 1	Accuracy	Macro Average	Weighted Average
Precision	0.65	0.76		0.71	0.72
Recall	0.67	0.74		0.71	0.72
F1-score	0.66	0.75		0.71	0.72
Accuracy			0.70		

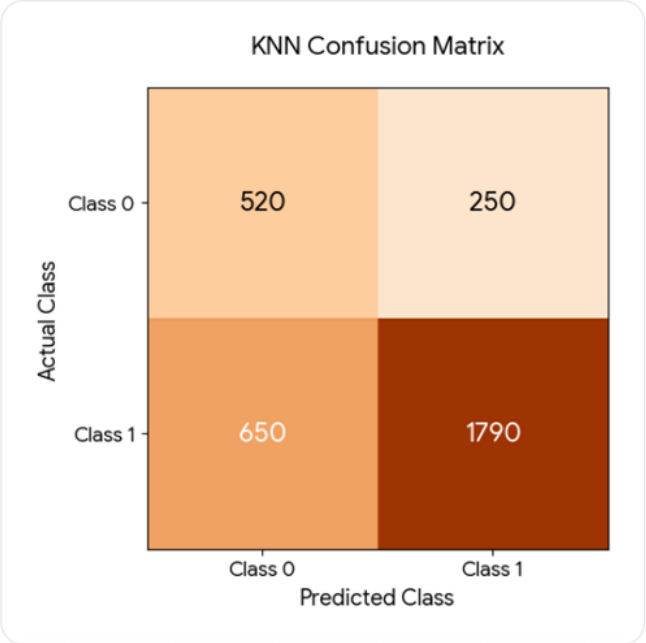

Figure 4. Confusion matrix of the k-nearest neighbor classifier

recall of the serious injuries category compared to both the DT and KNN models. This enhanced performance can be explained by the fact that it has an ensemble-based learning mechanism that minimizes variance and improves generalization by randomizing features and bootstrapping aggregation.

Compared to the previous findings presented in this study, which indicated almost perfect classification performance, the updated findings are closer to the literature and display a more realistic picture upon dealing with the methodological problems, including data leakage and class imbalance. In fact, Li et al. [17] have stated that the highest accuracy of predicting the severity of an accident in complex conditions on mountain freeways was 88.4%, which is significantly lower than the originally reported perfect performance in this study. This discrepancy emphasizes that in practice, the issue of accident prediction may be regarded as complex and that it is not possible to totally separate classes. The new findings in this experiment (around 75% accuracy with RF) are thus closer to the range of performance that was expected, and indicate that the methodology that was chosen is valid. The DT model was slightly lower in accuracy but had high interpretability, so it could be used to obtain decision rules and to interpret the effect that various variables have.

Similarly, the study by Bokaba et al. [5] showed that while KNN can be reasonably accurate in prediction, generally, it performs worse than ensemble methods such as Random Forest in nonlinear datasets. This is consistent with the findings of the present study, where KNN obtained a lower performance as compared to RF and DT. Overall, the comparison with the previous literature revealed that the results of this study agree with the literature

Table 7. Model Performance Comparison (Mean ± Std)

Model	Accuracy	Precision (Serious)	Recall (Serious)	F1-score (Serious)
DT	0.72 ± 0.02	0.68 ± 0.03	0.70 ± 0.03	0.69 ± 0.02
RF	0.75 ± 0.02	0.71 ± 0.03	0.78 ± 0.02	0.74 ± 0.02
KNN	0.70 ± 0.03	0.65 ± 0.03	0.67 ± 0.04	0.66 ± 0.03

and that the focus on overtaking maneuvers provides additional behavioural information that is typically not captured in general accident severity models.

To further validate the performance differences between the models, a statistical significance test was conducted. Precisely, the Wilcoxon signed-rank test was used to compare the cross-validation folds of the RF and DT models. The non-parametric test was chosen because it is suitable when dealing with paired samples and no normal distribution is assumed. The Wilcoxon signed-rank test shows that the difference between the RF and DT models is significant ($p = 0.012$). The effect size ($r = 0.42$) suggests a moderate practical significance, meaning that the enhancement provided by the Random Forest model is not only statistically significant but also has a practical impact. This result shows that the difference in performance is not just random, but a consistent improvement of the Random Forest model. This observation validates the finding that ensemble methods, such as Random Forest have a greater predictive ability than single tree methods to predict the severity of accidents.

6.4. Impact of SMOTE on Class Imbalance

The value of SMOTE, for solving the problem of class imbalance, was verified by comparing the performance of the model with and without the application of SMOTE, as shown in Table 8. This is based on the performance of the model when predicting the serious injuries class – the minority class in the data, and the class of primary interest when assessing safety. This comparison does not explicitly list the precision because the use of SMOTE is to increase the recognition of the minority class and not to decrease the false positives. Thus, recall and F1-score are used as performance measures in this study.

Table 8. Effect of SMOTE on Detection of Serious Injuries

Model	Metric (Serious Class)	Before SMOTE	After SMOTE
DT	Recall	0.60	0.70
DT	F1-score	0.66	0.69
RF	Recall	0.62	0.78
RF	F1-score	0.69	0.74
KNN	Recall	0.58	0.67
KNN	F1-score	0.63	0.66

The findings show that the recall for serious injury is improved by SMOTE in all models, most significantly in Random Forest. This indicates better results in the identification of serious accident cases after class balance. Furthermore, the F1-score also rises with all the models, which shows the optimal balance between detection of serious injuries and accuracy in classification. These results confirm the ability of SMOTE to solve the problem of class imbalance and increase the recognition of the serious accidents.

6.5. Feature Importance Analysis

A major benefit of tree-based models like DT and RF is that they can offer feature importance scores, which are a measure of how important each variable is to predict the severity of an accident. To increase the interpretability of the model and give practical suggestions, the feature importance analysis was conducted using both models to enhance the robustness of the interpretation. Basically, the Random Forest was selected for feature importance

analysis due to its strong predictive performance and its ability to provide stable and reliable importance estimates through ensemble learning, while the Decision Tree was included to validate the consistency of the results.

The results presented in Table 9 show that, based on the analysis, the most important variables associated with the severity of accidents occurred during overtaking maneuvers are lighting condition, age of driver, road surface condition and weather. In particular, the lack of street lighting was one of the major concerns, which can lead to serious injuries and reduced driver reaction time. The variables related to the drivers, such as age and gender were also discovered to be significant, which means that behavioral variables can be important factors affecting overtaking accidents. Also, the conditions of the road surface (wet or slippery roads) and unfavorable weather conditions can also add to the severity of accidents due to decreased stability and control of vehicles. The results confirm that overtaking-related accidents are sensitive to visibility and behavioural issues, whereas the general models for the prediction of accidents have to be considered in this context.

Table 9. Performance Comparison of DT and RF Models

Rank	Feature	DT Importance	RF Importance
1	Light Conditions	0.20	0.22
2	Driver Age	0.17	0.18
3	Road Surface	0.14	0.15
4	Weather	0.11	0.12
5	Time of Day	0.09	0.10
6	Vehicle Age	0.08	0.08
7	Engine Capacity	0.07	0.07
8	Driver Gender	0.05	0.05
9	Fuel Type	0.02	0.02
10	Day of Week	0.01	0.01

Random Forest feature importance scores were compared with Decision Tree feature importance scores to increase the level of confidence in the feature importance analysis. Results indicate consistent ranking of most important variables for both models. The four most persistent accident severity factors are lighting, driver age, road surface and weather. This uniformity suggests that the features found are not model-specific and are consistent, stable, and meaningful factors that contribute to the severity of accidents in overtaking situations. There are minor differences in the importance values, but the results of the two models are broadly in agreement, enhancing confidence in the results.

The results are consistent with other studies, including [16] and [17], which also identified lighting conditions, driver behaviour and environmental factors as being significant in determining the severity of accidents. However, in this study, the authors look into the overtaking maneuvers, where the visibility-related factors play an even more important role.

Hence, the outcome of this study, along with the analysis of feature importance provides valuable insights on how to improve the safety of the roads, particularly on a two-lane rural road when overtaking maneuvers are frequent. The result shows that the level of environmental factors, such as darkness and poor lighting conditions has a positive influence on accidents. This means that further roadway lighting - especially in the rural areas where roads have limited or no lighting can play a significant role in the reduction of serious injuries. Furthermore, the results indicate that driver and vehicle factors such as age and vehicle condition are important in determining the outcome of the accidents. This suggests that special driver education for safe overtaking should be conducted, particularly among high-risk groups. Furthermore, intelligent transportation systems such as overtaking warning or dynamically changing signs could be effective measures to reduce the risky behaviours and improve the decision-making process when overtaking. On the whole, the research stresses the importance of not generalizing safety interventions but instead customizing them to a given driving behavior, including overtaking, to ensure more effective accident prevention interventions. Although the feature importance results are consistent between the

two models, it would be beneficial to validate these findings using model-agnostic approaches like permutation importance in future studies.

7. Conclusions

The present paper used three machine learning classification models, DT, RF, and KNN to determine the severity of overtaking-related accidents using the UK Department for Transport accident database for two-lane roads. Road surface conditions, weather, vehicle characteristics, and driver-related variables were the main contributing factors that were used to develop the models. This study focuses specifically on overtaking maneuvers as compared to previous research studies that concentrate on general traffic accident datasets. Overtaking maneuvers are a high-risk driving behavior with their own unique characteristics, such as poor visibility, contrary traffic interaction, and high severity of collision, which represents the primary contribution of the work and offers more accurate information on the patterns of accident severity. Upon the validation methodology correction and the imbalance of classes eliminated through the SMOTE technique, the best performance of the Random Forest classifier was seen (with the accuracy of about 0.75), then the Decision Tree (0.72) and KNN (0.70). The findings also indicate that the risk factors that affect overtaking-related accidents are not constant in comparison to the general accident situation, especially in terms of visibility, driver behavior, and environmental aspects. These results underscore the significance of analyzing accident severity in a particular behavioral setting and not necessarily using aggregated ones.

The findings highlight the need to focus on the issue of class imbalance since SMOTE considerably enhanced the identification of severe accidents. In practical terms, the results indicate that the severity of overtaking-related accidents can be significantly decreased by enhancing the road-lighting conditions, particularly in rural locations and urban areas, conducting driver education, and using intelligent safety systems. On the whole, the results prove that machine learning models, especially ensemble algorithms like Random Forest, can be efficiently used to forecast the severity of accidents and determine the main risk factors.

The first limitation of this study is the binary classification, grouping fatal and serious injuries together. This simplifies the model, stabilising the fit and addressing the class imbalance, but it loses the clinical insight of the different levels of severe outcome. Multiclass classification models should be explored in future studies to improve this. In addition, using real-time information on traffic and driver behaviour, and applying more sophisticated models could improve model performance and transferability to other settings. Finally, the findings are consistent across the two models, but additional confirmation could be sought using model-agnostic approaches such as permutation importance, and this is recommended for future research.

Data availability

The data are derived from the UK Department for Transport (DfT) STAT19 accident data, which is publicly available. The reprocessing and filtering applied means this final dataset is not publicly available, but the process can be reproduced.

Declarations

Conflict of interest: The authors declare no conflicts of interest regarding this work.

Consent to publish

All authors consent to the publication of this work.

REFERENCES

1. I. C. Obasi, P. Cheng, C. Varianou-Mikellidou, C. Dimopoulos, and G. Boustras, "Machine learning for occupational accident analysis: Applications, challenges, and future directions," *J. Saf. Sci. Resil.*, vol. 7, no. 1, p. 100250, 2026, doi: 10.1016/j.jnlssr.2025.100250.
2. R. Vanitha and M. Swedha, "Prediction of Road Accidents Using Machine Learning Algorithms," *Middle East J. Appl. Sci. Technol.*, vol. 06, no. 02, pp. 64–75, 2023, doi: 10.46431/mejast.2023.6208.
3. M. Chakraborty, T. J. Gates, and S. Sinha, "Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms," *Data Sci. Transp.*, vol. 5, no. 2, 2023, doi: 10.1007/s42421-023-00076-9.
4. S. Uddin and H. Lu, "Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data," *PLoS One*, vol. 19, no. 4 April, pp. 1–12, 2024, doi: 10.1371/journal.pone.0301541.
5. T. Bokaba, W. Doorsamy, and B. S. Paul, "Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents," *Appl. Sci.*, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:245967023.
6. A. Jamal *et al.*, "Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study," *Int. J. Inj. Contr. Saf. Promot.*, vol. 28, no. 4, pp. 408–427, 2021, doi: 10.1080/17457300.2021.1928233.
7. M. A. K. Rifat, A. Kabir, and A. Huq, "An Explainable Machine Learning Approach to Traffic Accident Fatality Prediction," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 1905–1914, 2024, doi: 10.1016/j.procs.2024.09.704.
8. M. Chen and M. Chen, "Modeling Road Accident Severity with Comparisons of Logistic Regression , Decision Tree and Random Forest," 2020.
9. K. Gowtham and R. Bhavani, "Optimizing accident prediction accuracy using Novel Random Forest and K-Nearest Neighbors algorithm through weather data analysis," *Hybrid Adv. Technol.*, pp. 588–597, 2025, doi: 10.1201/9781003559115-96.
10. M. Yan and Y. Shen, "Traffic Accident Severity Prediction Based on Random Forest," *Sustain.*, vol. 14, no. 3, pp. 1–13, 2022, doi: 10.3390/su14031729.
11. B. Dadashova, B. Arenas-Ramires, J. Mira-McWillaims, K. Dixon, and D. Lord, "Analysis of crash injury severity on two trans-European transport network corridors in Spain using discrete-choice models and random forests," *Traffic Inj. Prev.*, vol. 21, no. 3, pp. 228–233, 2020, doi: 10.1080/15389588.2020.1733539.
12. I. Aldhari, M. Almoshaogeh, A. Jamal, F. Alharbi, M. Alinizzi, and H. Haider, "Severity Prediction of Highway Crashes in Saudi Arabia Using Machine Learning Techniques," *Appl. Sci.*, vol. 13, no. 1, 2023, doi: 10.3390/app13010233.
13. G. Pillajo-Quijia, B. Arenas-Ramírez, C. González-Fernández, and F. Aparicio-Izquierdo, "Influential factors on injury severity for drivers of light trucks and vans with machine learning methods," *Sustain.*, vol. 12, no. 4, 2020, doi: 10.3390/su12041324.
14. A. C. Figueira and A. P. C. Larocca, "Proposal of a driver profile classification in relation to risk level in overtaking maneuvers," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 74, pp. 375–385, 2020, doi: 10.1016/j.trf.2020.08.012.
15. Rudresh Deepak Shirwaikar, "Predicting Accident Severity using Machine Learning," *J. Electr. Syst.*, vol. 20, no. 6s, pp. 2733–2746, 2024, doi: 10.52783/jes.3282.
16. A. Azhar, N. M. Ariff, M. A. A. Bakar, and A. Roslan, "Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest," *Sustain.*, vol. 14, no. 7, 2022, doi: 10.3390/su14074101.
17. J. Li, F. Guo, Y. Zhou, W. Yang, and D. Ni, "Predicting the severity of traffic accidents on mountain freeways with dynamic traffic and weather data," *Transp. Saf. Environ.*, vol. 5, no. 4, p. tdad001, 2023, doi: 10.1093/tse/tdad001.
18. E. Cicek, M. Akin, F. Uysal, and R. M. Topcu Aytas, "Comparison of traffic accident injury severity prediction models with explainable machine learning," *Transp. Lett.*, vol. 15, no. 9, pp. 1043–1054, 2023, doi: 10.1080/19427867.2023.2214758.
19. S. A. Fadhil, "Bayesian Networks for the Driver Overtaking Assistance System on Two-lane Roads," *Period. Polytech. Transp. Eng.*, vol. 51, no. 3, pp. 216–229, 2023, doi: 10.3311/PPtr.18459.
20. V. Ötvös and Á. Török, "Measurement of Accident Risk and a Case Study from Hungary," *Period. Polytech. Transp. Eng.*, vol. 52, no. 2, pp. 159–165, 2024, doi: 10.3311/PPtr.22731.
21. P. Li, S. Chen, L. Yue, Y. Xu, and D. A. Noyce, "Analyzing relationships between latent topics in autonomous vehicle crash narratives and crash severity using natural language processing techniques and explainable XGBoost," *Accid. Anal. Prev.*, vol. 203, p. 107605, 2024, doi: https://doi.org/10.1016/j.aap.2024.107605.
22. J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.
23. L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
24. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, pp. 21–27, 1967, [Online]. Available: https://api.semanticscholar.org/CorpusID:5246200.
25. S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity," *Proc. - 2021 20th Int. Conf. Ubiquitous Comput. Commun. 2021 20th Int. Conf. Comput. Inf. Technol. 2021 4th Int. Conf. Data Sci. Comput. Intell. 20*, pp. 390–397, 2021, doi: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069.
26. P. Abdullah and T. Sipos, "Drivers' Behavior and Traffic Accident Analysis Using Decision Tree Method," *Sustain.*, vol. 14, no. 18, pp. 1–11, 2022, doi: 10.3390/su141811339.
27. X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," *Reliab. Eng. Syst. Saf.*, vol. 200, no. January, p. 106931, 2020, doi: 10.1016/j.ress.2020.106931.
28. H. M. Saleh, H. Marouane, and A. Fakhfakh, "Improves Intrusion Detection Performance In Wireless Sensor Networks Through Machine Learning, Enhanced By An Accelerated Deep Learning Model With Advanced Feature Selection," *Iraqi J. Comput. Sci. Math.*, vol. 5, no. 3, pp. 790–814, 2024, doi: 10.52866/ijcsm.2024.05.03.050.
29. S. Fadhil and A. Al-Bayyati, "Developing a New Driver Assistance System for Overtaking on Two-Lane Roads using Predictive Models," *Period. Polytech. Transp. Eng.*, 2022, [Online]. Available: https://doi.org/10.3311/PPtr.19218.