Predicting Road Traffic Collisions Using a Two-Layer Ensemble Machine Learning Algorithm

James Oduor Oyoo * , Jael Sanyanda Wekesa , Kennedy Odhiambo Ogada

Posted Date: 8 November 2023

doi: 10.20944/preprints202310.1780.v2

Keywords: Road collision traffic; Data imbalance; Machine Learning; Driving Simulation

Article

Predicting Road Traffic Collisions using a Two-Layer Ensemble Machine Learning Algorithm

James Oduor Oyoo [1,*], Jael Sanyanda Wekesa [2] and Kennedy Ogada Odhiambo [1]

[1] School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology; jimmoyoo@gmail.com
[2] School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology; jael.wekesa@jkuat.ac.ke
[3] School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology; kenogada@gmail.com
* Correspondence: jimmoyoo@gmail.com

Abstract: Road traffic collisions are among the world's critical issues causing many casualties, deaths, and economic losses with a disproportionate burden falling on developing countries. Existing research has been done to analyze this situation using different approaches and techniques on different stretches and intersections. In this paper, we propose a two-layer ensemble machine learning (ML) technique to assess and predict road traffic collisions using data from the driving simulator. The first (base) layer integrates supervised learning techniques namely, k- Nearest Neighbors (k-NN) AdaBoost, Naive Bayes (NB), and Decision Trees (DT). The second layer predicts road collisions by combining the base layer outputs by employing the stacking ensemble method using logistic regression as a meta-classifier. In addition, the synthetic minority oversampling technique (SMOTE) was performed to handle the data imbalance before training the model. To simplify the model, the particle swarm optimization (PSO) algorithm was used to select the most important features in our dataset. The two-layer ensemble model proposed had the best outcomes with an accuracy of 88%, F1 score of 83%, and an AUC of 86% as compared with k-NN, DT, NB, and AdaBoost. The proposed two-layer ensemble model can be used for theoretical as well as practical applications like road safety management for improving existing conditions of the road network and formulating traffic safety policies based on evidence in future.

Keywords: road collision traffic; data imbalance; machine learning; driving simulation

1. Introduction

Globally road traffic crashes take the lives of nearly 1.35 million every year, more than two every minute with more than nine in ten of all deaths occurring in low and middle-income countries. Road traffic collisions have become the leading cause of death for people aged 15- 29 years, world health organization (WHO) estimates that crashes will cause another 13 million deaths and 500 million injuries around the world by 2030 if urgent action is not taken [1]. A World Health Organization (WHO 2018) research report revealed that Kenya was one of the world's worst collisions recorded, accounting for a fatality rate of 27.8 per 100,000, population [2] . Accordingly, the city of Nairobi recorded the highest share of the total road crashes in Kenya. In addition, road traffic collisions in Nairobi Kenya are a cause of significant losses of human life and economic resources. According to the National Transport and Safety Authority (NTSA) Report , 4,690 people lost their lives to road collisions between January 1 and December 13 in 2022 [3]. Additionally, the report notes that pedestrians and riders are dying at much higher rates because of car collisions from time to time in Kenya. WHO recently announced, "Decade of Action for Road Safety 2021-2030", setting the target of preventing at least 50% of road traffic deaths and injuries by 2030 [4]. Significant attention is required to minimize road collisions and as a result, research into building prediction models (PMs)

 doi:10.20944/preprints202310.1780.v2

and traffic collision prevention is critical to improve road safety policies and to reduce fatalities on roads [5].

Since road traffic collisions are random, traditional techniques such as logit and probit models have been widely used to predict these collisions [6]. Although statistical models have good mathematical interpretation and provide a better understanding on the role of individual predictor variables, they have some limitations [7].These traditional approaches are built on assumptions such as requiring a predefined mathematical form, the presence of outliers, and missing values in the dataset, such inferences may be untrue and can negatively affects the outcome of the prediction model [8]. With the advancement in soft computing methods, machine learning techniques have emerged as promising tools in road safety collisions research to overcome the limitations of statistical methods. In contrast to traditional techniques, machine learning (ML) techniques can manage outliers and missing values in the dataset. To predict road collisions, ML techniques have been applied to primary and secondary road collision datasets for different road networks [9,10]. Data unavailability in low and middle-income countries impedes road safety improvements. Access to data is crucial for scientific research on identifying factors causing high road risk and assessing the effectiveness of interventions [11].

Our main objective in this study is to develop and evaluate a crash prediction model that can predict road traffic collisions and their patterns. We perform accident analysis, by applying a two-layer ensemble stacking method using logistic regression as a meta-classifier, and the four most popular supervised machine learning algorithms: NB, k-NN, DT, and AdaBoost, because of their proven accuracy in this field [12–14]. Datasets for this study were acquired from a fixed-base driving simulator [15]. The prediction accuracy, precision, recall, and F1-score of each ML technique were compared and measured to highlight the best fit. Our contribution through this paper is the development of a crash prediction model that can predict the outcome of a collision, which can help the emergency centers to estimate the possible impacts and provide better appropriate medical treatment, enable policymakers to formulate better policies for road safety based on evidence, and to enable better road traffic safety management.

The article is structured as follows, Section 2 focuses on the research methodology and explains data preprocessing, feature selection, and building the ensemble model. Section 3 gives the analysis outcomes. Section 4 discusses the key findings of this research. Lastly in Section 5, we conclude the paper and address future works.

2. Materials and Methods

In this study, we developed an ensemble model with two layers using four base classifiers and a meta-classifier that integrates the base layer models to improve the performance. The four supervised ML algorithms employed to predict road collisions and their patterns are k-NN, DT, AdaBoost, and Naïve Bayes. Subsequently, the logistic regression was integrated as a meta-classifier in the second layer of the model by integrating the outputs of the four first layer models. Figure 1 presents the flowchart adopted in this study. The research methodology has been structured into the following steps: data collection, data preprocessing, building ensemble model, and performance evaluation of the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

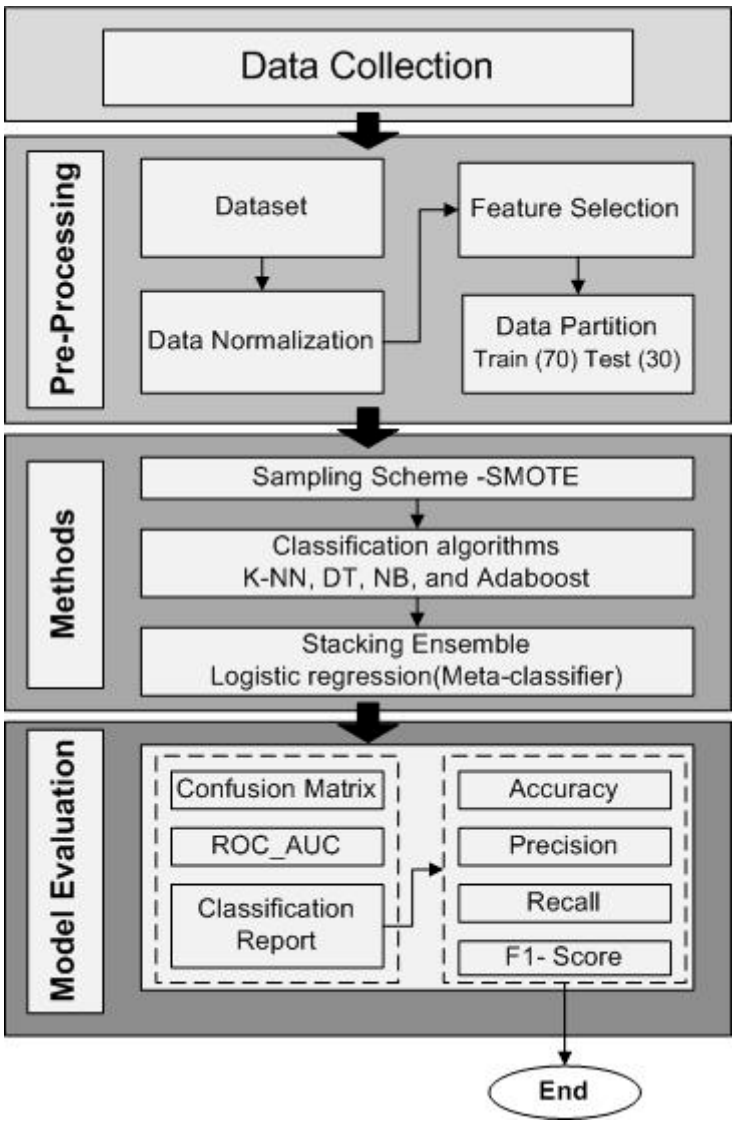

Figure 1. Study workflow diagram.

2.1. Study population and data description

A driving simulator was used to collect data for this study. It is very dangerous to conduct trials in the real world environment, driving simulator provides an excellent tool for collecting data in a safe environment [16,17]. The 3.5km Mbagathi way in Nairobi, Kenya was modelled in the driving simulator at Strathmore University Business School, Institute of healthcare management. The simulations included 80 participants who were selected through the snowball approach. The participants were required to hold a valid driver's license with more than two years of driving experience. An informed consent form was administered to the participant, and they were also briefed on why they were selected and informed of the importance of participating in the study. . Weather, the speed limit, lane width, and road layout served as the primary determinants of the scenarios. The driving simulator has a driving seat, a powerful simulation computer, 3 screens that display the driving scenarios, an observer screen, a 7" Tablet that displays the speedometer, a steering wheel, a clutch, a gear stick, an accelerator, and brakes. Figure 1 shows a participant driving through the simulated road during the experiment. The simulations were based on 2 scenarios that included before treatment and after treatment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

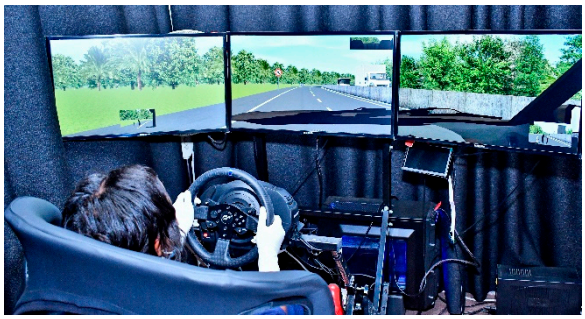

Figure 2. A participant driving on the simulated road scenario at Strathmore University.

2.2. Data preprocessing

The data with 15 features were loaded into the panda data frame object to facilitate various preprocessing procedures. Firstly, the data set was normalized using 15 features, after which missing values were discovered in some of the fields. Since the missing values would affect the performance of the model, we replaced the blank and null feature values by applying the mean value of that feature column [18,19]. The mean values that were used to fill the missing feature records presented no extreme value that could have affected the mean.

Feature selection

Feature selection is a critical factor in obtaining an accurate prediction. Using all the features leads to an inefficient model, as the number of features increases, models struggle for accuracy hence reducing model performance [20]. In this study, we used Sklearn, a Python library to select the features. To obtain the most important features for this study, we employed four algorithms: particle swarm optimization (PSO), univariate feature selection, recursive feature elimination, and feature importance.

1. *Particle swarm optimization (PSO) algorithm:* This technique works by searching for the optimal subset of features. It locates the minimum of a function by creating several 'particles. These particles store their best position as well as the global position. It is this combination of local and global information that gives rise to 'swarm intelligence' [21]. In our study we implemented XGBoost and linear regression algorithms to select the best features.
2. *Recursive feature elimination:* This technique works by selecting the optimal subset of features for estimation by reducing 0 to N features iteratively [22]. The best subset is then chosen based on the model's accuracy, cross-validation score, or Roc-Auc curve.
3. *Univariate feature selection:* This approach works by selecting the optimal features using the univariate statistical tests. It might be considered a stage in the estimator's preprocessing process [23]. In our study, we implemented the chi-squared statistical test using the SelectKBest method.
4. *Feature importance:* It works by classifying and evaluating each attribute to create splits. Decision tree models that are developed on ensembles, for example, extra trees and random forests can be used to rank the relevance of certain features [24]. In our study, we employed the extra trees classifier for feature selection.

After performing the feature selection algorithms, we selected the top six features as shown in Table 1 based on the selected features algorithms.

Table 1. Features having a strong relationship with road collisions.

Univariate feature Selection	Recursive Elimination method	Feature importance	Particle swarm optimization (PSO)
Lane gap	Lane gap	Lane gap	Lane gap
Speed	Speed	Speed	Speed

 doi:10.20944/preprints202310.1780.v2

Brake	Brake	Brake	Brake
Education level	Education level	Education level	Driver Experience
Driver Experience	Driver Experience	Driver Experience	Surface condition
Drivers' Age	Drivers' Age	Drivers' Age	Gender

Three techniques, namely univariate, recursive elimination method and feature importance had the top six common features, while the PSO algorithm had four common features with the other three techniques. For this study, we employed the PSO feature selection method because the performance of the model was not affected when evaluating the model using the features selected by the other three techniques.

2.3. Building the two-layer ensemble model

We evaluated the performance of machine learning approaches, by splitting the dataset in the ration of ratio of 70% training dataset and 30% testing dataset. In our research, we employed four well-known classification algorithms previously used to predict road traffic collisions and the stacking ensemble method to predict road traffic collisions. Stacking is an ensemble method for integrating numerous models with a meta-classifier. Following the development of the base models, the four base models (level-0) – k-NN, AdaBoost, DT, and Naïve Bayes were integrated using a stacking framework for road collision prediction. We selected the four base models because of their proven diversity in predicting road collisions. In the second layer, Logistic regression, was employed as a meta-classifier to classify road collisions from the outputs of the base models. A 10-fold cross-validation technique was used to evaluate how well the models predicted traffic collisions [25]. The proposed two-layer ensemble model has been shown in Figure 3 below. The following section expounds on the four supervised machine-learning techniques and the stacking method employed for our study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

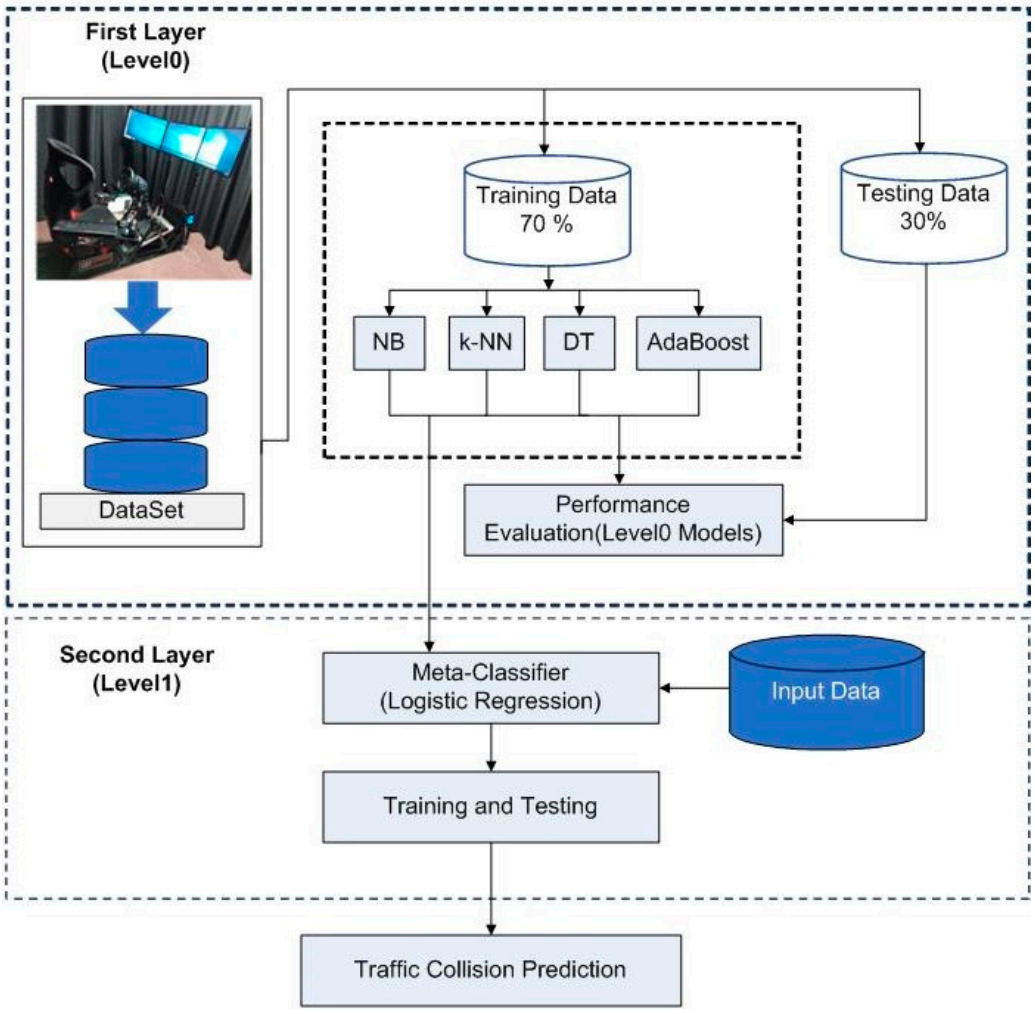

Figure 3. The proposed Two-layer ensemble model.

(i) Naïve Bayesian Classifier (NBC): This algorithm employs the theorem of Bayes. It works by estimating the probability of various classes based on a variety of features and allocates the new class to the class with the highest probability[26]. In our study, Gaussian NB was chosen because the feature set contains continuous variables. The NB is represented by the following formular.

$$P(H|E) = \frac{P(E|H)*P(H)}{P(E)} \qquad (1)$$

Where P(H|E) is the posterior probability of the hypothesis given that the evidence is true, P(E|H) is likelihood of the evidence given that the hypothesis is true, P(H) is prior probability of the hypothesis, and P(E) is prior probability that the evidence is true. The posterior probability is mainly the probability of $'H'$ being true given that $'E'$ is true.

(ii) k-Nearest Neighbors (k-NN): This method can be considered a voting system, where the majority class determines the class label of a new data point among its nearest neighbor [27]. It then analyzes datasets, calculates the distance function and similarities between them, and groups them based on k values. In our study, the k value was obtained by performing several tests with values ranging from 1 to 50, and the prediction performance was compared to the k value. We plotted the accuracies for both training and test datasets as shown in Figure 4. The performance of k-NN shows a drop in both the test and training datasets after adding neighbors, the drop continues for both until a point where they converge. The test dataset improves with an increase in the number of neighbors from iteration 33 until they converge with the training dataset at neighbor 42. In the proposed model, we set the k value at 42 as this yielded the best results and Euclidean distance was selected as distance function [28].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

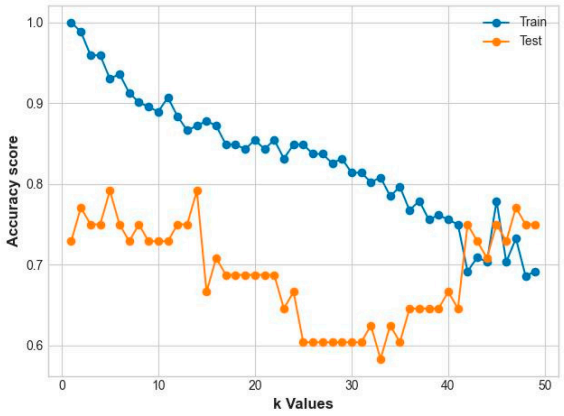

Figure 4. Line plot illustrating k-NN accuracy on train and test datasets at different neighbors.

The distance between the clusters is used to classify new input data, and the closest cluster is allocated. The following formula illustrates the k-NN approach.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)^2} \quad (2)$$

Where x , y , is the two points in n-space, n is the number of input samples, and yi, xi is the distance vectors starting from the original point.

(iii) Decision Trees (DT): This methodology is a nonparametric supervised learning method for classification and regression. The goal is to build a model that predicts the target variable's value by learning simple decision rules based on data attributes [29]. This is shown by the mathematical formula below.

$$Entropy: \sum i = 1 - p * log_2(pi) \quad (3)$$

$$Entropy(S) = -p + log_2 p + -p\text{-}log_2 p - \quad (4)$$

Given S is the sample of training examples, and p+ is the proportion of the positive training examples while p− is the proportion of the negative training examples. DT has an overfitting problem, and to overcome it, we used pruning technique to remove splits with little information gained [DT]. This simplifies the DT by reducing the time cost of training and testing and eliminates the problem of overfitting [30]. In our study, increasing the tree depth in early stages results in a corresponding improved performance for training dataset and reduced performance in test dataset. As the tree depth grows, a corresponding improvement is noted on both training and test dataset up to the depth of 4. Depth 5 reveals that the model overfits the training dataset at the expense of the test dataset. as shown in Figure 5 below. In our study we set the maximum tree depth at 4.

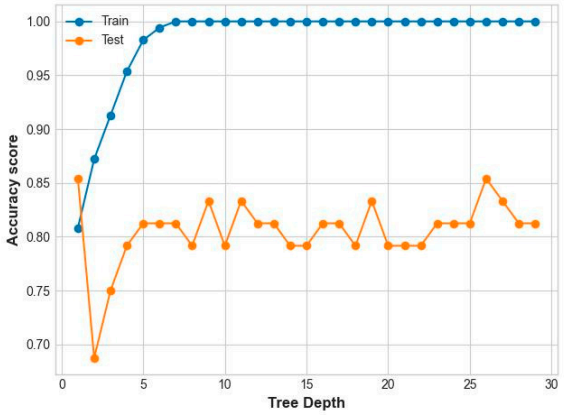

Figure 5. Line plot illustrates DT accuracy on train and test datasets at different tree depths.

(iv) Adaptive Boosting (AdaBoost): AdaBoost is a classification method that calls a given weak learner algorithm repeatedly in a number of rounds. In the training dataset, each instance is weighed, and overall errors are calculated. More weight is given when it is difficult to predict, and less weight

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

is given when it is simple to predict [31,32]. The AdaBoost approach has a weight that is represented as a vector for each weak learner. The input samples are illustrated in the equation 5 as follows:

$$Weight, w_i = \frac{1}{n} \quad (5)$$

Where wi is the i'th training instance weight and n is the number of the training instances.

(v) Stacking ensemble method: Stacking is a method of integrating predictions from various machine learning models on the same dataset, such as bagging and boosting [33]. The stacking technique's architecture consists of two or more models, known as base models or level-0, and meta-models that combine the predictions of the base models, known as level-1 models [34]. For our study, stacking was selected because the employed models are often distinct and fit on the same dataset, then a single model is trained to integrate the outputs of the base as best as possible [35]. In our study, we implemented logistic regression as meta-model to provide a seamless interpretation of the base models' predictions.

2.4. Validation and performance measurement

We performed some steps in our experiment to develop the accident prediction model. The first step was to partition the dataset in the ratio of 70% training and 30% testing data. The accuracy was assessed using a 10-fold cross-validation technique in the second stage. The entire dataset was divided into 10 subsets at random, where each subset was used as testing data along with the other nine subsets.

2.5. Data Oversampling

There are limitations associated with working with a binary classification when dealing with imbalanced dataset [36]. Oversampling have been chosen to mitigate the effect of any underlying samples with under representation. Across most of the datasets considered as imbalanced, sampling strategies have been implemented improve the overall model's accuracy [37,38]. One of the most important aspects to note is that oversampling is not considered to create any new data instances as this can result to overfitting, conversely under sampling may exclude important samples from the learning process meaning that the most useful data instances may be overlooked by the model [39].

In this study, our dataset was imbalanced and therefore we performed a synthetic minority oversampling technique (SMOTE) resampling strategy to handle the data imbalance [40]. The SMOTE algorithm develops synthetic positive cases to enhance the proportion of the minority class [41]. In our scenario, data had 76% instances of no collision and 24% instances of collision as shown in Figure 6.

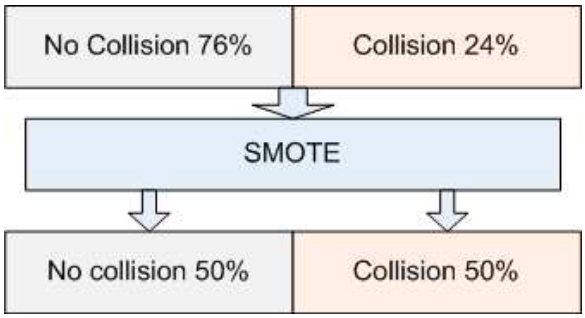

Figure 6. SMOTE methodology diagram.

The dataset before SMOTE is illustrated in Figure 7 as a scatter plot with many points spread for the majority class and a small number of points scattered for the minority class. Majority class, 0 represents No collisions and 1 represents collisions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

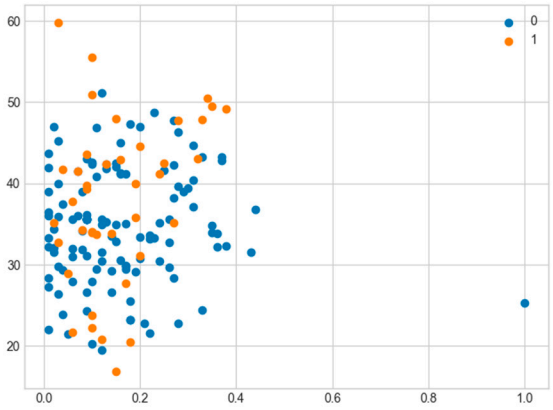

Figure 7. Scatter plot of imbalanced dataset before SMOTE.

The transformed dataset was balanced after SMOTE as shown in the scatter plot Figure 8 below in the ratio of 1:1.

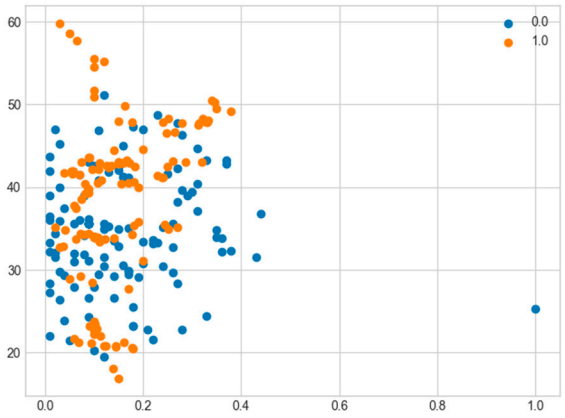

Figure 8. Scatter plot of the balanced dataset after SMOTE.

The crash prediction model's performance was evaluated using the classification report that included computed values of accuracy, precision, recall and F1 score of the algorithms. Our model suffered from underfitting because of using the outputs of the base layer model in the second layer, and to overcome the problem of underfitting in our model, some input features from Table 1 that were used in base layer models were reduced and used together with the output of the base layer models. The reason for this approach was to improve the model. Logistic regression was used to train the level-1 input features as a meta-classifier. The test data set was then used to evaluate the two-layer ensemble model. The model with the highest values of the metrics was considered the best prediction model.

The data generated by the confusion matrix was used to test each model's performance metric. The outcomes of the initial and predicted classifications generated by a classification model comprise the confusion matrix (CM) [42]. Table 2 shows a representation of a confusion matrix.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

Table 2. The architecture of the Confusion Matrix.

Total instances		**Predicted**	
		Negative	**Positive**
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (TN	True Positive (TP)

The confusion matrix layout shown above displays the actual classes in the rows and the predicted class observations in the columns.

The following defines each entity in CM.

In **TN**, the entities that are originally negative are appropriately classified as negative.

In **FN**, the entities that are originally positive are wrongly classified as negative.

In **TP,** the entities that are originally positive are appropriately classified as positive.

In **FP**, the entities that are originally negative are incorrectly classified as positive.

The observations of the confusion matrix for every model were used to calculate the following performance metrics and evaluate model performance based on these metrics:

Accuracy: represents the percentage of the total number of instances that were correctly classified, as shown by the equation below:

$$Accuracy\ (AC) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

Recall: represents the percentage of positive events that were correctly classified as shown by the following equation:

$$Recall\ (R) = \frac{TP}{TP+TN} \quad (7)$$

Precision: the percentage of correctly predicted positive instances, shown by the equation below:

$$Precision\ (P) = \frac{TP}{TP+P} \quad (8)$$

F1- Measure: the performance of the model is measured using the F1 measure that represents the harmonic mean of the Recall and Precision. Its value is in the range of 0 to 1, with 1 denoting the best model and 0 denoting the poorest model. The F1 equation is represented by the equation below:

$$F1 = \frac{2*(R*P)}{R+P} \quad (9)$$

Error rate: represents the frequency of miscalculation of the predictions depicted as represented by the equation below.

$$Error_{rate}(ER) = 1 - Accuracy \quad (10)$$

3. Results

3.1. Results of the classification before SMOTE

Since we are predicting the occurrence of a road traffic collision or not, our problem was a binary classification [43]. In this study the data sample from the driver simulation was split into 30% train data and 70% test data. The model's predictive performance on the test dataset was evaluated by comparing accuracy, precision, and F1 score. The effectiveness of each algorithm has been determined from the simulation driver data by employing AdaBoost, DT, NB, and k-NN as base models using the same selected feature set, then employed the stacking ensemble method using logistic regression as a meta classifier to improve on the model's accuracy. We performed two scenarios where one was without SMOTE, and the second scenario was with SMOTE. Before pruning DT and setting the k value for k-NN, Decision trees achieves the highest accuracy of 87%, followed by Two-layer ensemble with 85%, Naïve Bayes with 83%, AdaBoost and k-NN achieves a similar score of 79% as illustrated in Table 3 before SMOTE technique.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

Table 3. Results before performing SMOTE Analysis.

Model	Accuracy	Precision	Recall	F1 Score
AdaBoost	0.79±0.11	0.76±0.13	0.71±0.12	0.72±0.14
k-NN	0.79±0.08	0.81±0.41	0.66±0.19	0.68±0.25
DT	**0.85±0.12**	0.87±0.27	**0.77±0.22**	**0.80±0.19**
NB	0.83±0.05	0.82±0.20	0.76±0.18	0.78±0.10
Two-Layer Ensemble	0.83±0.06	**0.91±0.25**	0.75±0.19	0.79±0.11

The robustness of the ML model is largely assessed and validated using the area under the receiver operator curve (AUC). When the AUC is higher than 0.7, the developed model is said to have good predictive power. Before SMOTE, the Two-layer ensemble had an area of about 0.87%, followed by NB algorithm at 0.83%, AdaBoost 0.82%, k-NN 0.8%, and DT with 0.77% as shown in Figure 9. The experiment was done before implementing pruning on DT and setting the k value on k-NN.

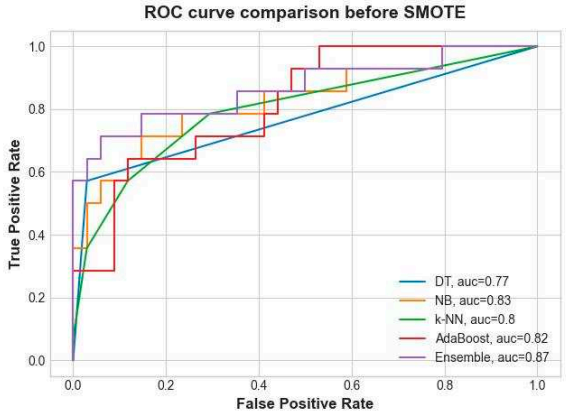

Figure 9. Comparison of the Area Under the Curve (ROC) for the models before SMOTE.

3.2. Results of the classification after SMOTE

The AUC scenarios were also compared, one without any resampling technique and one with a resampling strategy applied. However, after SMOTE resampling strategy, pruning DT due to overfitting, and setting the k value for k-NN, NB had an improved AUC of about 0.86%, AdaBoost remained unchanged while a decrease was noted in the Two-layer ensemble, DT and k-NN as shown in Figure 10.

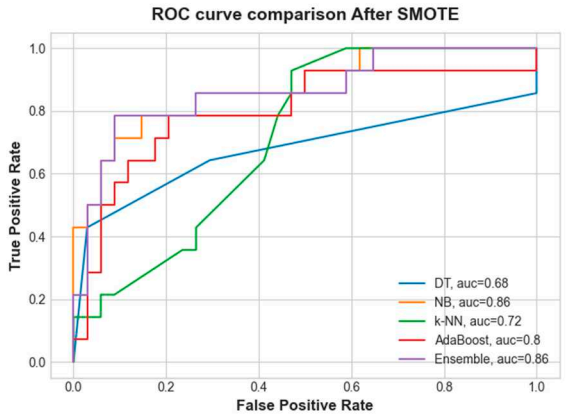

Figure 10. Comparison of the Area Under the Curve (ROC) for the models after SMOTE.

Overall, among all the base models, the recall value was improved on AdaBoost and the proposed two-layer ensemble when applying SMOTE, while a decrease was noted on DT and k-NN, and NB as shown in Table 4. The precision of the model after applying SMOTE was reduced on DT, k-NN, NB, AdaBoost and the two-layer ensemble model. Based on F1 score, a noticeable increase is

 doi:10.20944/preprints202310.1780.v2

noted on the two-layer ensemble model and AdaBoost, while the same was reduced on NB, DT, and k-NN. NB and AdaBoost achieve the highest accuracies of 81% and 79% respectively, followed by DT at 77% while k-NN achieves the lowest accuracy of 72% among the base models as shown in Table 4 after SMOTE. The overall accuracy performance, the two-layer ensemble model achieves of 85% accuracy.

Table 4. Results after SMOTE Analysis for models.

Model	Accuracy	Precision	Recall	F1Score
AdaBoost	0.79±0.09	0.75±0.12	0.73±0.13	0.74±0.08
k-NN	0.72±0.13	0.66±0.12	0.64±0.08	0.65±0.06
DT	0.77±0.08	0.69±0.90	0.68±0.08	0.68±0.10
NB	0.81±0.06	0.72±0.10	0.73±0.12	0.73±0.07
Two-Layer Ensemble	**0.85±0.08**	**0.86±0.09**	**0.82±0.09**	**0.83±0.08**

3.3. Results of the proposed ensemble model

Accuracy is a measure of the effectiveness of a single algorithm, but relying solely on accuracy as a measure of performance index can lead to erroneous conclusions as the model may be biased towards specific collision classes [44]. To solve this limitation in our study, other performance measurement metrics such as recall, F1 score, and precision were evaluated. These performance indicators demonstrate the performance of individual collisions and allow better insights for the model. The outcomes of the "no collisions" and "with collisions" performance measurements are shown in Tables 5 and 6, respectively.

The definition of precision and recall states that the optimum model is one that optimizes both performance measurements. F1 score is also a good performance indicator because it interprets model performance using both precision and recall. In our study, all the models performed well for no collision, while k-NN, DT and AdaBoost performed poorly for collisions. The two-layer ensemble and NB performed well for collisions as shown in Tables 5 and 6.

Table 5. Outcomes of the models for no collisions.

Model	Precision	Recall	F1 Score
k-NN	0.79	0.97	0.87
Decision Trees	0.83	0.86	0.85
AdaBoost	0.83	0.86	0.85
Naïve Bayes	0.85	0.94	0.90
Two-Layer Ensemble	**0.87**	**0.97**	**0.92**

Table 6. Outcomes of the models for collisions.

Model	Precision	Recall	F1 Score
k-NN	0.80	0.31	0.44
Decision Trees	0.58	0.54	0.56
AdaBoost	0.58	0.54	0.56
Naïve Bayes	0.80	0.62	0.70
Two-Layer Ensemble	**0.89**	**0.62**	**0.73**

After evaluating the model using the stacking ensemble method with reduced features, there was a significant improvement in the predictive performance of the models. Table 7 shows the classification accuracy of each model. The two-layer ensemble achieved the highest accuracy of 0.88%, NB had 0.81%, DT 0.81%, AdaBoost 0.79%, while k-NN achieved the lowest score of 0.65%.

 doi:10.20944/preprints202310.1780.v2

Table 7. Outcomes of the models.

Model	Accuracy	Precision	Recall	F1 Score
k-NN	0.65±0.09	0.56±0.12	0.56±0.10	0.56±0.09
Decision Trees	0.81±0.74	0.83±0.12	0.70±0.78	0.73±0.72
AdaBoost	0.79±0.08	0.76±0.10	0.71±0.10	0.72±0.09
Naïve Bayes	0.81±0.10	0.77±0.12	0.76±0.11	0.77±0.09
Two-Layer Ensemble	**0.88±0.08**	**0.86±0.09**	**0.83±0.11**	**0.84±0.79**

Among the base models, NB had the best F1 score performance while k-NN had the lowest. Overall, the best F1 score was achieved by the two-layer ensemble model. Similarly, the proposed two-layer ensemble model had the best recall, while NB had the best recall among the base models, AdaBoost and DT had similar scores, k-NN had the lowest recall score. The two-layer ensemble model had superior precision when compared with the other models, as shown in Table 7. The objective of the ensemble method was to predict road collisions by utilizing a minimal feature set which may be acquired within a short period from the collision scene. Based on this prediction, policy makers, road constructors and health facilities would be able to predict road traffic collisions at any given site hence put all the necessary measures required to avert collisions and save lives. The improved two-layer ensemble model demonstrates that it is the most effective method for predicting road collisions.

4. Discussion

The increase in road traffic collisions necessitates the effective analysis and control of these collisions. The study adopted a unique methodological approach to propose a model that predicts road traffic collisions based on dataset from a driving simulator. The fact that it is very dangerous to conduct trials in a real-world environment and knowing that a driving simulator provides an excellent tool for collecting data in a safe environment devoid of life-threatening risks and damage to property. The dataset from the simulator was downloaded and normalized using 15 features. We then performed feature selection engineering techniques to select the best features, therefore reducing the likelihood of overfitting for our model. The best parameters of each model were determined by a 10-fold cross validation. The training set was partitioned into 10 equal subsets, with one subset serving as testing data and the remaining nine serving as training data. The process was then repeated using the entire ten subsets where the whole data set was used for validation. Our problem was binary classification since our study focused on predicting the occurrence of a collision or no collision [45]. Given the stochastic nature of collisions, they tend to be underrepresented in the dataset, therefore synthetic minority oversampling technique (SMOTE) was used to balance the classes in the training dataset. Crash prediction offers a proactive approach to increase road safety adherence and saving lives. Research into road safety has been of great interest to researchers, industry, and policy makers. Crash prediction remains complex and requires high dimensionality and large datasets to develop models that can effectively predict road traffic collisions [46].

Although depending on accuracy as a measure of models' performance can be misleading, the model might be biased towards one class. In the present study, to overcome these limitations, we determined other performance measures such as precision, recall and F1-scrore. To demonstrate the effectiveness of the proposed model, we compared it to the existing works in literature. Notably, the authors are aware of few works that have focused on crash prediction models based on dataset from the driving simulator [47]. A comparison between the proposed two-layer ensemble approach and other works in literature is presented in Table 8. The strategy was to include similar works that are closely related and deployed the same methodologies. Our study findings align with the existing literature, however if a standard data collection format and a standard feature selection approach is standardized across the globe, the transferability, comparison and usability of these models will be easy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2023 doi:10.20944/preprints202310.1780.v2

Table 8. Comparison of the proposed two-layer ensemble model with works in literature.

Work	Dataset source	Method	Precision	Recall	F1-score	Accuracy
Aldhari et al. [44]	Collected	Ensemble				
		XGBoost	94%	94%	94%	94%
		RF	91%	90%	90%	90%
		LR	65%	65%	65%	65%
yang et al. [45]	Australia road deaths database (ARDD)	Ensemble				
		SVM				88%
		k-NN				87%
		DT				88%
Luo et al. [46]	Driving Simulator	Classification				
		DT				77%
		Gradient boosting decision tree (GBDT)				80%
		Long-short term memory (LSTM)				87%
Mansoor et al. [43]	Canadian Dataset	Ensemble				
		k-NN	62%	70%	66%	67%
		DT	68%	70%	69%	69%
		AdaBoost	72%	72%	72%	71%
		FNN	70%	70%	70%	69%
		SVM	72%	69%	71%	68%
		Two-Layer Ensemble	73%	77%	75%	76%
Proposed	**Driving Simulator**	**Ensemble**				
		k-NN	**56%**	**56%**	**56%**	**65%**
		DT	**83%**	**70%**	**73%**	**81%**
		AdaBoost	**76%**	**71%**	**72%**	**79%**
		NB	**83%**	**76%**	**77%**	**81%**
		Two-Layer Ensemble	**86%**	**83%**	**84%**	**88%**

5. Conclusion

In this paper, we proposed a two-layer ensemble model for predicting road traffic collisions. The employed method: two-layer ensemble was created by combining the outputs of k-NN, DT, AdaBoost, NB, and Logistic regression as a meta-classifier in the two levels. The models were compared with each other in terms of accuracy, precision, recall and F1-score. With the unique combination of the ML classifiers, the two-layer ensemble method achieved a remarkable accuracy of 88% in a 10-fold cross-validation, precision 86%, recall 83% and F1-score of 84%. Since traffic collisions are random, a model that can predict road traffic collisions in a timely manner by using a few input features is required. In practice, crash prediction is an important aspect in emergency services and trauma centers to estimate the potential risks resulting from collisions and accordingly equip the centers and other units with appropriate post-crash care equipment, for policy makers, the findings of this research can be implemented to formulate evidence-based policies as opposed to the cause-and-effect approach that is common in most low- and middle-income countries. The two-layer ensemble model can then be used to predict road collisions and therefore save lives and prevent socio-economic losses. Through validation the proposed two-layer ensemble had the highest

accuracy, the limitation of the proposed approach is the time it takes to run the model, which can be comparatively longer than individual models. Additionally, the dataset in this study was imbalanced, therefore we applied SMOTE resampling strategy, other advanced approaches could have been used to solve the issue of imbalanced dataset. The dataset in this study was based on simulated crash data, we highly advocate a common road collision data collection format to be used by traffic and policy enforcers worldwide. The use of computer science techniques like machine learning for predicting road collisions is found to be more accurate and effective hence if a common data collection methodology is applied, the actual collisions data model can produce more realistic results making the application, transferability, and validation of these models easy. For future work, we propose performing sensitivity analysis to select the best features, then employing the two-layer ensemble model and complementing the traditional approaches for road safety.

6. Patents

No patents resulting from the work reported in this manuscript.

Author Contributions: The authors confirm contributions to the paper as follows: study conception and design: James Oduor Oyoo, Kennedy Ogada Odhiambo, Jael Sanyanda Wekesa; methodology and data collection: James Oduor Oyoo, Kennedy Ogada Odhiambo, Jael Sanyanda Wekesa, findings analysis and interpretation: James Oduor Oyoo, Kennedy Ogada Odhiambo, Jael Sanyanda Wekesa; draft manuscript preparation: James Oduor Oyoo, Jael Sanyanda Wekesa, Kennedy Ogada Odhiambo ; manuscript revision: Kennedy Ogada Odhiambo, Jael Sanyanda Wekesa, James Oduor Oyoo.

Funding: This research received no external funding.

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding author upon a reasonable request.

Acknowledgments: We would also like to thank the NTSA, Kenha, KURA, Annette Murunga, Kevin Otieno, Institute of Healthcare Management- Strathmore University Business School for their support during scenario modelling and development.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. "WHO Death On Roads." [Online]. Available: https://extranet.who.int/roadsafety/death-on-the-roads/#deaths/per_100k
2. "Road traffic injuries." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
3. "NTSA Report on Road Safety 2022." [Online]. Available: https://www.the-star.co.ke/news/2023-01-18-4690-people-died-in-road-accidents-in-2022-report/
4. *Decade of Action for Road Safety*. [Online]. Available: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/decade-of-action-for-road-safety-2021-2030
5. R. E. Al Mamlook, A. Ali, R. A. Hasan, and H. A. Mohamed Kazim, "Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation," in *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, Dayton, OH, USA: IEEE, Jul. 2019, pp. 630–634. doi: 10.1109/NAECON46414.2019.9058268.
6. Z. Li, H. Liao, R. Tang, G. Li, Y. Li, and C. Xu, "Mitigating the impact of outliers in traffic crash analysis: A robust Bayesian regression approach with application to tunnel crash data," *Accid. Anal. Prev.*, vol. 185, p. 107019, Jun. 2023, doi: 10.1016/j.aap.2023.107019.
7. 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, Oct. 2021, doi: 10.1080/17457300.2021.1928233.
8. L. Zheng, T. Sayed, and F. Mannering, "Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions," *Anal. Methods Accid. Res.*, vol. 29, p. 100142, Mar. 2021, doi: 10.1016/j.amar.2020.100142.
9. T. Bokaba, W. Doorsamy, and B. S. Paul, "Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents," *Appl. Sci.*, vol. 12, no. 2, p. 828, Jan. 2022, doi: 10.3390/app12020828.
10. R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity," in *2019 IEEE Jordan International Joint Conference on*

Electrical Engineering and Information Technology (JEEIT), Amman, Jordan: IEEE, Apr. 2019, pp. 272–276. doi: 10.1109/JEEIT.2019.8717393.

11. Y. Berhanu, E. Alemayehu, and D. Schröder, "Examining Car Accident Prediction Techniques and Road Traffic Congestion: A Comparative Analysis of Road Safety and Prevention of World Challenges in Low-Income and High-Income Countries," *J. Adv. Transp.*, vol. 2023, pp. 1–18, Jul. 2023, doi: 10.1155/2023/6643412.
12. M. Al-Nashashibi, W. Hadi, N. El-Khalili, G. Issa, and A. A. AlBanna, "A New Two-step Ensemble Learning Model for Improving Stress Prediction of Automobile Drivers," *Int. Arab J. Inf. Technol.*, 2021, doi: 10.34028/iajit/18/6/9.
13. M. Ameksa, H. Mousannif, H. Al Moatassime, and Z. Elamrani Abou Elassad, "Crash Prediction using Ensemble Methods:," in *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, Kenitra, Morocco: SCITEPRESS - Science and Technology Publications, 2021, pp. 211–215. doi: 10.5220/0010731200003101.
14. P. A. D. Amiri and S. Pierre, "An Ensemble-Based Machine Learning Model for Forecasting Network Traffic in VANET," *IEEE Access*, vol. 11, pp. 22855–22870, 2023, doi: 10.1109/ACCESS.2023.3253625.
15. K. Yang, C. A. Haddad, G. Yannis, and C. Antoniou, "Classification and Evaluation of Driving Behavior Safety Levels: A Driving Simulation Study," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 111–125, 2022, doi: 10.1109/OJITS.2022.3149474.
16. X. Zhang and X. Yan, "Predicting collision cases at unsignalized intersections using EEG metrics and driving simulator platform," *Accid. Anal. Prev.*, vol. 180, p. 106910, Feb. 2023, doi: 10.1016/j.aap.2022.106910.
17. W. Xiao, X. Luo, and S. Xie, "Feature semantic space-based sim2real decision model," *Appl. Intell.*, Jun. 2022, doi: 10.1007/s10489-022-03566-5.
18. M. J. Crowder, A. C. Kimber, R. L. Smith, and T. J. Sweeting, *Statistical Analysis of Reliability Data*, 1st ed. Routledge, 2017. doi: 10.1201/9780203738726.
19. F. A. Shakil, S. M. Hossain, R. Hossain, and S. Momen, "Prediction of Road Accidents Using Data Mining Techniques," in *Proceedings of International Conference on Computational Intelligence and Emerging Power System*, R. C. Bansal, A. Zemmari, K. G. Sharma, and J. Gajrani, Eds., in Algorithms for Intelligent Systems. , Singapore: Springer Singapore, 2022, pp. 25–35. doi: 10.1007/978-981-16-4103-9_3.
20. B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, p. 103375, Sep. 2019, doi: 10.1016/j.compbiomed.2019.103375.
21. Y. Cao, G. Liu, J. Sun, D. P. Bavirisetti, and G. Xiao, "PSO-Stacking improved ensemble model for campus building energy consumption forecasting based on priority feature selection," *J. Build. Eng.*, vol. 72, p. 106589, Aug. 2023, doi: 10.1016/j.jobe.2023.106589.
22. A. Zhang, E. W. Patton, J. M. Swaney, and T. H. Zeng, "A Statistical Analysis of Recent Traffic Crashes in Massachusetts," 2019, doi: 10.48550/ARXIV.1911.02647.
23. A. M Ascensión, O. Ibáñez-Solé, I. Inza, A. Izeta, and M. J. Araúzo-Bravo, "Triku: a feature selection method based on nearest neighbors for single-cell data," *GigaScience*, vol. 11, p. giac017, Mar. 2022, doi: 10.1093/gigascience/giac017.
24. M. Mittal, S. Gupta, S. Chauhan, and L. K. Saraswat, "Analysis on road crash severity of drivers using machine learning techniques," *Int. J. Eng. Syst. Model. Simul.*, vol. 13, no. 2, p. 154, 2022, doi: 10.1504/IJESMS.2022.123344.
25. A. Seraj *et al.*, "Cross-validation," in *Handbook of Hydroinformatics*, Elsevier, 2023, pp. 89–105. doi: 10.1016/B978-0-12-821285-1.00021-X.
26. D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction," *Computers*, vol. 10, no. 12, p. 157, Nov. 2021, doi: 10.3390/computers10120157.
27. J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Phys. Stat. Mech. Its Appl.*, vol. 517, pp. 29–35, Mar. 2019, doi: 10.1016/j.physa.2018.10.060.
28. L. Liu and M. T. Özsu, Eds., "k-Nearest Neighbor Classification," in *Encyclopedia of Database Systems*, Boston, MA: Springer US, 2009, pp. 1590–1590. doi: 10.1007/978-0-387-39940-9_2920.
29. P. Abdullah and T. Sipos, "Drivers' Behavior and Traffic Accident Analysis Using Decision Tree Method," *Sustainability*, vol. 14, no. 18, p. 11339, Sep. 2022, doi: 10.3390/su141811339.
30. Y. Lu, T. Ye, and J. Zheng, "Decision Tree Algorithm in Machine Learning," in *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, Dalian, China: IEEE, Aug. 2022, pp. 1014–1017. doi: 10.1109/AEECA55500.2022.9918857.
31. C. Wang, Y. Wang, and X. Zhang, "A Study of Fatigue Driving Detection System Based on AdaBoost Algorithm," in *2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Hamburg, Germany: IEEE, Oct. 2022, pp. 32–35. doi: 10.1109/AIAM57466.2022.00014.
32. H. Zhao, H. Yu, D. Li, T. Mao, and H. Zhu, "Vehicle Accident Risk Prediction Based on AdaBoost-SO in VANETs," *IEEE Access*, vol. 7, pp. 14549–14557, 2019, doi: 10.1109/ACCESS.2019.2894176.

33. L. Yang and Q. Zhao, "An aggressive driving state recognition model using EEG based on stacking ensemble learning," *J. Transp. Saf. Secur.*, pp. 1–22, May 2023, doi: 10.1080/19439962.2023.2204843.
34. J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, "Crash injury severity analysis using a two-layer Stacking framework," *Accid. Anal. Prev.*, vol. 122, pp. 226–238, Jan. 2019, doi: 10.1016/j.aap.2018.10.016.
35. P. Wu, X. Meng, and L. Song, "A novel ensemble learning method for crash prediction using road geometric alignments and traffic data," *J. Transp. Saf. Secur.*, vol. 12, no. 9, pp. 1128–1146, Oct. 2020, doi: 10.1080/19439962.2019.1579288.
36. A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.
37. Z. Jiang, J. Yang, and Y. Liu, "Imbalanced Learning with Oversampling based on Classification Contribution Degree," *Adv. Theory Simul.*, vol. 4, no. 5, p. 2100031, May 2021, doi: 10.1002/adts.202100031.
38. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
39. D. Lee and K. Kim, "An efficient method to determine sample size in oversampling based on classification complexity for imbalanced data," *Expert Syst. Appl.*, vol. 184, p. 115442, Dec. 2021, doi: 10.1016/j.eswa.2021.115442.
40. Z. Elamrani Abou Elassad, H. Mousannif, and H. Al Moatassime, "Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study," *Traffic Inj. Prev.*, vol. 21, no. 3, pp. 201–208, Apr. 2020, doi: 10.1080/15389588.2020.1723794.
41. F. Sağlam and M. A. Cengiz, "A novel SMOTE-based resampling technique trough noise detection and the boosting procedure," *Expert Syst. Appl.*, vol. 200, p. 117023, Aug. 2022, doi: 10.1016/j.eswa.2022.117023.
42. A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowl.-Based Syst.*, vol. 247, p. 108651, Jul. 2022, doi: 10.1016/j.knosys.2022.108651.
43. M. Mokoatle, Dr. Vukosi Marivate, and Professor. Michael Esiefarienrhe Bukohwo, "Predicting Road Traffic Accident Severity using Accident Report Data in South Africa," in *Proceedings of the 20th Annual International Conference on Digital Government Research*, Dubai United Arab Emirates: ACM, Jun. 2019, pp. 11–17. doi: 10.1145/3325112.3325211.
44. U. Mansoor, N. T. Ratrout, S. M. Rahman, and K. Assi, "Crash Severity Prediction Using Two-Layer Ensemble Machine Learning Model for Proactive Emergency Management," *IEEE Access*, vol. 8, pp. 210750–210762, 2020, doi: 10.1109/ACCESS.2020.3040165.
45. 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, p. 233, Dec. 2022, doi: 10.3390/app13010233.
46. L. Yang *et al.*, "Comparative Analysis of the Optimized KNN, SVM, and Ensemble DT Models Using Bayesian Optimization for Predicting Pedestrian Fatalities: An Advance towards Realizing the Sustainable Safety of Pedestrians," *Sustainability*, vol. 14, no. 17, p. 10467, Aug. 2022, doi: 10.3390/su141710467.
47. T. Luo, J. Wang, T. Fu, Q. Shangguan, and S. Fang, "Risk prediction for cut-ins using multi-driver simulation data and machine learning algorithms: A comparison among decision tree, GBDT and LSTM," *Int. J. Transp. Sci. Technol.*, vol. 12, no. 3, pp. 862–877, Sep. 2023, doi: 10.1016/j.ijtst.2022.12.001.