https://doi.org/127: 10.51153/kjcis.v6i1.127

# Road Accidents Investigation and Forecasting Using Data Mining Techniques

*[1]Usama Fareed, [2]Umair Khadam, [3]Muhammad Munwar Iqbal, [4]Muhammad Javed Iqbal

## Abstract

Investigation of Road accidents is vital because it can uncover the connection between the various properties that lead to a road accident. Factors that influence road accidents can be road elements, climate factors, and traffic attributes. Analysis of road accidents can give data about the involvement of these characteristics, which can be used to beat the accident rate. Data mining is a famous procedure for analyzing the road accident dataset. In this paper, we have used data mining techniques and geometric analysis on a dataset of road accidents to find the impact of attributes like road surface, weather conditions, lighting conditions, and casualty severity on a road accident. The Frequent Pattern (FP) Growth technique was used to discover the association rules. Classification models were made by some decision trees like J48 and Decision Tree (DT), Random Tree, and Hoeffding tree. The results showed that Random Tree Classifier performed well with 90.6% accuracy, followed by Hoeffding Tree with 85.58% accuracy and J48 with 84% accuracy.

**Keywords:** Road Safety, Data Mining, Association, Classification, Frequent Pattern (FP), Decision Tree

## 1. Introduction

Millions of vehicles are running on the road every day. With these vehicles is the possibility that an accident can happen at any time. It has become a challenging task for the Govt. institutions to prevent road accidents. So predicting road accidents using machine learning techniques can help to avoid accidents and to minimize the damage from them [1]. Nowadays, machine learning and deep learning are the main topics for researchers to reduce the rate of accidents. Some researchers use CNN as a deep learning method because it is a speedy and efficient network in different computer applications such as Computer Vision, and image recognition [2]. Another model Extreme Value Theory (EVT) was also used to prevent accidents as EVT has been used on many other grounds (e.g., business, insurance) to foresee the happening of extreme events that are different from the normal. Due to a similar approach, EVT was also brought in for accident prediction on motorway ramps and intersections. However, there are other colliding conditions when the driver is incapable of identifying and react the clash [3]. The EVT predictions are based on Time to Accident (TA) may expose the accident chance of a driver's observation reaction

*[1]Department of Computer Science, University of Kotli AJ&K, Pakistan | usamafarid41@gmail.com
[2]Department of Computer Science, University of Kotli AJ&K, Pakistan | umair_khadim@uokajk.edu.pk
[3]Department of Computer Science, UET, Taxila, Pakistan | munwar.iq@uettaxila.edu.pk
[4]Department of Computer Science, UET, Taxila, Pakistan | javed.iqbal@uettaxila.edu.pk

collapse [4]. Some research has been done on collision prediction using the Internet of Vehicles (IOV) framework. To overcome the accidents problems, they designed a risk calculation proposal that can guess the threat level besides the vehicle by calculating many aspects, including the automobile itself, other vehicles, the surroundings, and the driver. Gathered data was provided for a neural network for assumption, and the outcomes were utilized for extra vehicle management [5]. The main purpose of the research was to spotlight a warning method in IOV-based deep learning models. A neural network was used to get the outcome of numerous features, including surroundings, traffic and road conditions, and the driver himself. In addition, different kinds of warnings can be given to the driver viewing the risk level of the vehicle [6].

Researchers have used a lot of methods to predict accidents. Comparing two geometric techniques, the NB and RENB models have been done to guess the prediction by taking the unobserved data that discover the main reasons for the crash to recover the trafficking shelter [7, 8]. A complete study has been done on predicting road accident difficulty by implementing the (Conv LSTM) Neural Network Model. Datasets that are used in research are vehicle crash data, rainfall data, satellite images, and Traffic cameras data [9]. A comparative study has also been done between the ARIMA model and the ARIMAX model using the Bayesian Information Criteria(BIC) that include the extracted features e.g., humans, vehicles, road conditions, and weather forecasts [10].

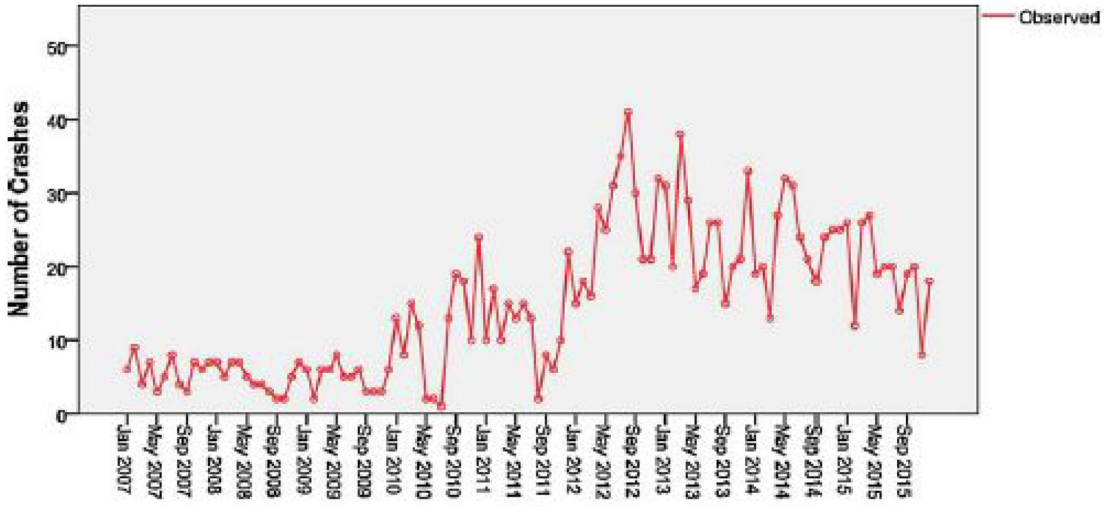


**Figure 1: Number of Crashes in Anambra State, Nigeria [11].**

Some researchers used the Strategic Highway Research Program 2 (SHRP 2) Naturalistic Driving Study (NDS) data to design a security prognostic sculpt. The model can support many applications like the Advanced Driver Assistance System (ADAS) and security hazard profiling of drivers [12]. Figure 1 shows the time series scheme of the comprehensive monthly count data from Jan 2007 to Sep 2015 for the number of crashes in Nigeria.

Expressways have more troubled the Traffic Police Department because accident prediction on

highways or expressways is critical. Different deep learning methods, such as Random Forest and Radiant Boosting Decision Trees, are used to execute accident forecasts [13, 14]. A vital job in traffic crash hindrance is to design an efficient risk forecast structure. If accident risk can be predicted in a specific region, then this information can be broadcast to the drivers close to that area. Still, it is very difficult to predict the risk accurately because many other factors can affect the accident [15].

There are multiple procedures to avert car accidents, such as traffic rules knowledge, a good transportation system, the efficiency of traffic police, and preventative measures in the automobile. Many grounds for accidents are using a cell phone while driving, high speed, listening to radio/ music, or overcrowded roads. However, there are no specific factors on which we can predict accidents correctly. So, this is difficult to use all the factors simultaneously for a prediction. That's why many systems have been designed that use different frameworks to overcome this issue [16]. Classification models are one of the most frequently used techniques in analyzing traffic accidents, where the objective is to design a classifier that can predict the accidents.

This paper will apply decision Tree classification algorithms to understand the most significant accident features. Different kinds of decision trees will apply to model the classifier, such as (Random Tree, J48, Hoeffding Tree, and Decision Stump).

## 2. Literature Review

Several studies have been done to check vehicle-based applications to detect causes of accidents by machine learning. Shrestha et al. [17] used the Apriori Algorithm, Naïve Bayes Classifier, and K means clustering to explore the association, classification, and clustering between the impact factors. They investigate the connection between the features like weather, surface state, light condition, and the drunk driver. A comparative study has been done by Sakhare and Kasbe et al. [18] between Self Organization Map (SOM) and K-Means. SOM is used to find different patterns of accidents that help to predict them in the future and to get better precision. While K means is used for clustering of the data. Results showed that SOP is much better than the K-means because it can easily understand the patterns and predict the accident. Yuan et al. [19] examined the prediction of accidents by using heterogeneous urban data. They collect datasets, such as all vehicle crashes in Iowa, and road and weather conditions, from 2006-13. Dataset was analyzed using four different algorithms, i.e., Deep Neural Network (DNN), Support Vector Machines (SVM), Random Forest, and Decision Tree. It deals with heterogeneous spatial data, and the spatial Graph feature has been used with the help of Eigen Analysis of the road system. Bahiru et al. [20] evaluate the precision of the Naïve Bayes classifiers, CART and J48, to predict the harshness of road accidents. After comparing the above techniques, they conclude that J48 achieved the highest precision. They eradicate the accident year factor at the data processing stage by considering it an irrelevant feature.

Climate and weather conditions are one of the reasons for road accidents. Zou et al. [21] Presented a negative binomial model to investigate the impact of different factors on road accidents. The study includes weather conditions, social development parameters, and the frequency of accidents. The authors claimed that the results showed that the climate increased the accident frequency. In contrast, non-weather conditions such as vehicle conditions and driver behavior also have a considerable effect. The results can also give guidelines to transport authorities to take necessary actions to prevent road accidents. Wenqi et al. [22] created a new prediction model (TAP-CNN) by using the factors that cause traffic accidents, i.e., traffic stream, weather, and light, to erect a state matrix to explain the status of the traffic and CNN model. This model used weather conditions to create a state matrix. The base of the model was Convolution Neural Network. Chen et al. [23] projected a model to expect the threat within a city from a different point of view. They used the Stack De-noising Convolutional Autoencoder model (SDCAE) to forecast the rate of accidents for grid cells by using traffic stream, accidents in the past, and time data. Experiments were performed in real-world traffic big data sets from the major city in China to predict accidents. Comparisons showed that this model performed way better than the baseline models. Rahim et al. [24]Proposed a deep-learning method to predict the severity of an accident using the F1-loss function. The dataset used in the research was collected from Louisiana from 2014 to 2018. The method works on the principle of transforming the variables into images using the Convex Hull algorithm. To get the optimum precision and recall, a CNN model was utilized. The outcome showed improved performance in envisaging the severity of an accident.

In another research, Kaur [25] suggested a new tactic to study road accidents by analyzing the accident data that is collected from traffic systems and data linked with the construction region. Exploratory visualization and correlation analysis were used to examine the regularity of traffic accidents. This study helped them predict the accident, particularly on national highways and normal roads, by estimating the sternness of accidents relying on the type of accident and where the accident happened. Taamneh et al. [26] discuss the major reason for road accidents, that cause deadly severity is gender. Male is the highest possible, age (18-30 years. Generally have more accidents), type of accident ( a car with a pedestrian is common), and place where the accident occurs. Machine learning techniques such as decision trees and MLP were used to conclude the above features. Xiong et al. [27] recommended a new Chain of Road Traffic Incident (CRTI) technique. In which the movement of vehicles before the accidents is observed. CRTI includes the behavior of the driver (speed, lane changing) as well as real-time inputs from road conditions and the environment (weather). The behavior of the vehicle before the accident is observed thoroughly with the help of some other algorithms such as Support Vector Machines (SVM) and the Hidden Markov Model (HMM) so that an accident can be predicted, and an early warning or interference can be made to evade the accident. CRTI can provide a new base for investigating the tactic of timely warning in driver support systems. You et al. [28] presented a different technique to predict the accident rate using data sets that Discrete Loop Detectors gather, and Web Crawl Weather Data's method was applied to spot the risk factor. To select the impact factors Random Forest technique was used. Results found that SVM successfully classifies 76.32% of accidents on

the test dataset and 87.52% on an overall dataset which was much better than earlier research. Li et al. [29] predicted real-time accident threats with the help of LSTM-CNN. SMOTE has been used to overcome the inequality problem. A different data source has been investigated. According to researchers results of this method are far better than other assessment models.

Kurakina et al. [30] presented a geometric technique of forecasting that made it promising to assess the way of assorted causes of the accident rate resulting in the assessment of the effectiveness of the planned measures to perk up road safety. The methods of possible danger allowed for obtaining real and predicted factors of road accident hazards and casualties in a road crash on the road section under study. Changes in accident rates after the execution of road safety measures have been assessed. Asor et al. [31] examined accident data to get the unseen patterns that may be used as a safety measure to reduce accidents in Los Banos Laguna, Philippines. These factors were analyzed by the Decision Tree, Naïve Byes, and Rule induction algorithms. After applying these algorithms, it was seen that the Decision Tree showed the best results with 92.84%. It was found that day and time play an important role in the fatality or severity of the accident. Results showed that the accident place does not have any significant relation to the death of the victim.

One of the major reasons for road accidents is the driver's careless behavior. Uma et al. [32] Presented a prototype design using different cameras, Raspberry Pi, and sensors that can detect the driver's behavior, such as yawning and drowsiness, to prevent an accident. A hybrid IoT and machine learning system are installed in vehicles and connected to the cloud to transmit the driver's behavior to the cloud to take rapid action in an emergency. This system works as a live monitoring system in which every driver's action is monitored continuously. Hashmienejad et al. [33] designed a technique that works with a Novel based method to predict road accidents according to user fondness instead of typical DTs. They modify the multi-objective algorithm (NSGA-II) to get better results. The outcomes showed that its performance is highest (88.2%) than all other classification techniques like ANN, SVM, and Decision Trees. Sinclair and Das used an unverified machine-learning method to examine accidents [34]. Clustering was used to recognize the patterns and links between the factors recorded by the UK Police. Ali et al. [35] described a real-time framework using the Bidirectional long short-term memory(Bi-LSTM) and Ontology and Latent Dirichlet Allocation (OLDA) model to detect traffic accidents. At first, the traffic information was collected by using a query-based search engine. Then sentiment analysis technique was used to classify the traffic events that help in getting the exact details of the accident. In the end, the Bi-LSTM model was trained to detect and analyze traffic accidents. Elyassami et al. [36] proposed a work in which they gathered accident data sets that were given by the Maryland Police. A well define hyper-parameter was used in gradient booting proposed in [37]. They applied three machine learning techniques Random Forest, Gradient Boosted Tree, and Decision Tree. After getting the results it was clear that Gradient Boost Based Model provided the most accurate and influencing features that cause the accidents. The study showed that weather conditions, road conditions, and less visibility are important factors to predict an accident. By using these factors accident risk can be minimized.

Now a day's, traffic accidents have become a major problem in big cities. Every day People lose their lives because of road accidents. It is an important worry for the government and the citizens as well. Considerable work has been done in a paradigm of road safety. Still, there are loopholes to be considered. Some researchers include only a few parameters and exclude other factors that can be the reason for road accidents, e.g., only weather datasets were used, and human behavior was not considered. Some researchers performed the classification on very small and private datasets that can give better accuracy. However, its accuracy decreases when a large dataset is used on the same classification model. Some old techniques were used, such as K-means, which gives results with less accuracy.

**Table 1: Comparison of Different Technologies**

| # | Author | Year | Problem | Solution | Future Work/Drawback |
|---|---|---|---|---|---|
| 1 | Bahiru et al. [20] | 2018 | Predicting Road Traffic Accidents Severity | Applied Classifiers ID3, J48, and Naive Bayes to compare the results | More classifiers such as SVM and random forest should be Included in comparison. |
| 2 | Chen et al. [23] | 2018 | Accident risk prediction | By using Stack Denoise Autoencoder (SDAE) to divide the city into regions | In the future, more datasets such as weather can be added to the research to train the model |
| 3 | Elyassami et al. [36] | 2021 | Road crash analysis and prediction | Decision Trees techniques were used to compare data | Further investigation concerning the driver's Behavior |
| 4 | Hashmienejad [33] | 2017 | Traffic accident severity prediction | Non-Dominated Sorting The genetic Algorithm (NSGA) was used to allow humans to pertain to their inclination. | Feature selection and selection techniques such as PCA and LDA enhance the results. |
| 5 | Kaur and Kaur [25] | 2017 | Forecast of the reason for mishap and a clumsy area on streets and roads | Estimating the severity of accidents on Highways using the R tool. | The severity of accidents can be examined using other factors like limitless speed, shoulder drop-off, etc. |
| 6 | Shrestha et al. [17] | 2017 | Factors that are closely related to fatal accidents. | Applied Apriori algorithm for association and used K means to find clustering. | More data like non-fatal data can be used for more suggestions. |
| 7 | Sakhare and Kasbe [18] | 2017 | Evaluation of road accident data study | A self-organization Map (SOM) was used to locate the patterns. | Required more iteration for the execution of the process. |

| | | | | | |
|---|---|---|---|---|---|
| 8 | Sinclair and Das [34] | 2021 | Examination of a traffic accident in urban areas | K means clustering was used to get blueprints and realize relationships among selected factors. | Further clustering methods will be used for a comprehensive assessment of information. |
| 9 | Taamneh, et al. [26] | 2017 | Accident features contribute to severe injury | Data mining techniques MLP, NB, J48 WEKA applied to predict the brutality of accidents | Studies on pedestrian-vehicle accidents will be performed in the future. |
| 10 | Wenqi et al.[22] | 2017 | Forecasting the traffic accidents | TAP-CNN model was launched with factors like traffic flow, light, and weather. | More features like road alliance and line of traffic will be used to improve the system |
| 11 | You et al. [39] | 2017 | Real-time accident prediction | SVM technique used with web-crawled data to classify the risk status | Only one factor e.g. weather is used. |
| 12 | Yuan et al. [19] | 2017 | Envisage accidents through heterogeneous Urban Data | Classifiers like Random Forest, DNN, and SVM were used to address the heterogeneity of the factors. | In the future optimizing techniques will be explored to predict road accidents in real time. |
| 13 | Asor et al. [31] | 2018 | Security and safety of precious human life along with financial benefits | Naïve Bayes, Decision Tree, and Rule Induction were used. Batch X validation to authenticate model results. | Larger data sets will be used to get improved results, patterns, and analysis. |

In Table 1, a comprehensive comparison of past research papers on road accidents has been made. It shows the problem statement of the papers and then which techniques were used to solve the problem. In the end, it shows the future work or drawbacks of the specific research.

## 3. Proposed Methodology

With a speedy growth in cities, the volume of vehicles is also increased, resulting in severe accidents that led to fatalities and gigantic losses to the respective country's economy. In this entire scenario, predicting traffic accidents and their severity becomes essential to prevent crashes and lessen the hurt from traffic accidents in a positive manner. It is a difficult task to predict the danger of an accident due to the intricate traffic atmosphere, human psychology, and lack of synchronized data.

This research paper covers the major factors such as weather conditions, road surface, lighting conditions, and casualty severity. Data mining techniques will be applied to these factors to get the results. Data mining is a vast field that uses numerous techniques and models to discover the

association in a huge quantity of data. Association rules are the best technique to find important associations between the data of the hefty database. It has a key role in discovering the most repeated items in the database. A standard technique for making association rules is frequent Pattern Growth (FP Growth) which is used to uncover regular datasets. FP Growth is faster and more efficient than Apriori because there is no need for candidate generation. It constructs an FP tree rather than a generation of candidates. It focuses on fragmenting the paths and mining the frequent patterns. Classification has an important role in data mining. Its main purpose is to create a model from a trained dataset for categorizing the records of the unidentified marker. Decision trees are believed to be the latest classification algorithms. It is a support tool that enables to one approach obstacles in a structured and systematic manner. These are also capable of handling multiple outcomes. Figure 2 shows the whole process of data mining the dataset until we get the results.

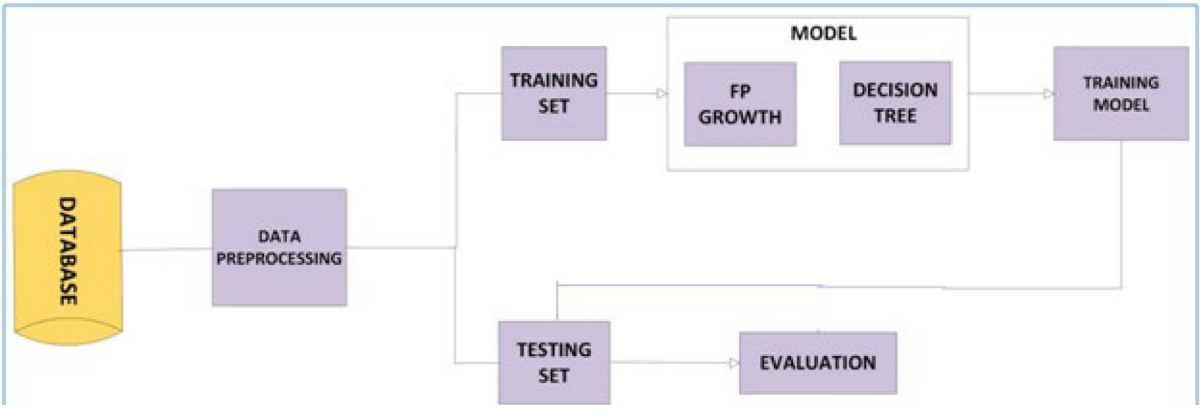


**Figure 2. Proposed Working Model**

### 3.1 Classification Techniques

The classification Techniques include Decision Tree and Classifier Accuracy. A brief discussion of these techniques is given in the next section.

#### 3.1.1 Decision Tree

The decision tree technique is commonly used in data mining. The outcome of this technique is a classification model that calculates the value of an aimed attribute based on the input value. The decision tree builds classification models in the form of trees. Each core node in trees is one of the input variables and has many branches of possible values of that input variable. Each leaf node holds a target attribute value. The decision tree technique was used to understand existing data and forecast new accidents' sternness. The objective is the dynamical assumption of a picking tree until it picks up the balance of flexibility and preciseness. Entropy has been used in this technique, which measures disordered data.

A decision tree can handle non-linear data efficiently. A decision tree can handle both numerical as well as categorical datasets and does not require any significant pre-processing. It can be used for both classification and regression.

### 3.1.2 Comparison between Decision Trees with other Classifiers

#### 3.1.2.1 Decision Tree vs Naive Bayes

- Decision trees are more simple and more flexible than Naive Bayes.
- The structure of Naïve Bayes is generative while Decision Trees is a discriminative model.

#### 3.1.2.2 Decision tree Vs (Support Vector Machine) SVM

- Compared to Decisa ion Tree SVM cannot work efficiently on large datasets.
- SVMs can be very costly with Non-linear kernels.
- The results of SVM are not insightful for a layman as compared to Decthe ision Tree.

### 3.1.3 Classifier Accuracy

A matrix characterizes the performance of a classifier model, called the confusion matrix, which demonstrates the correctly and incorrectly classified instances for each class. The measures used to review a classifier's performance are calculated from the confusion matrix. The most commonly used estimation measure is the accuracy rate, which shows the percentage of correctly classified instances and is calculated using eq (1).

$$Accuracy= \frac{\text{True Positive+True Negative)}}{\text{True Positive+True Negative+False Positive+False Negative}} \quad (1)$$

Specificity indicates the correct negatives values divided by all the negative values

$$Specificty= \frac{TN}{TN+FP}=\frac{TN}{N} \quad (2)$$

The recall is the number of correct classifications divided by the total number of Positives values.

$$Recall= \frac{TP}{TP+FN} \quad (3)$$

Precision is the number of correct positive classification values divided by the total number of positive classification values.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

As decisions were produced, frequent Pattern Growth (FP Growth) was used to make association rules with the prediction rate on the right side. FP Growth was applied in WEKA Tool with minimum support = 0.1 and minimum confidence = 0.9. Table 2 shows the best 10 rules generated by WEKA. These rules are derived based on different conditions or parameters, e.g., road surface, light condition, and weather condition. Generally, it is a prediction that is being made based on some parameters, e.g., as shown in the table, if the road surface is dry and the weather is fine, then there are more chances that an accident will occur, and the severity of the accident will be sight.

**Table 2: 10 Best Association Rules with Highest Confidence Generated by Frequent Pattern (FP) Growth Algorithm**

| **Road Surface = Dry** | **Weather Conditions = Fine** | **Casualty Class = Driver** | **Prediction → High** | **conf: (0.98)** |
|---|---|---|---|---|
| Road Surface = Dry | Weather Conditions = Fine | Light Conditions =Daylight | Prediction → High | conf: (0.97) |
| Road Surface = Dry | Weather Conditions = Fine | Casualty Severity = Slight | Prediction → High | conf: (0.97) |
| Road Surface = Dry | Weather Conditions = Fine | - | Prediction → High | conf: (0.97) |
| Road Surface = Dry | Weather Conditions = Fine | Casualty Severity = Slight | Prediction → High | conf: (0.97) |
| Road Surface = Dry | Weather Conditions = Fine | Type of Vehicle = Car | Prediction → High | conf: (0.97) |
| Streetlights = Present | Light Conditions =Daylight | Type of Vehicle = Car | Prediction → High | conf: (0.91) |
| Type of Vehicle = Car | Casualty Severity = Slight | - | Prediction → High | conf: (0.91) |
| Road Surface = Dry | Type of Vehicle = Car | Casualty Severity = Slight | Prediction → High | conf: (0.91) |
| Type of Vehicle =Car | Weather Conditions = Fine | Casualty Severity = Slight | Prediction → High | conf: (0.91) |

## 4. Results and Discussion

In this study total of 2664 records of road accidents in the 2015 year in Leeds were examined. The dataset contains different attributes like the date of the accident, location of the accident, road surface, type of vehicle, weather conditions, age and sex of casualty, and lighting conditions. Unnecessary attributes like the date and location of the accident and type of vehicle were removed

so we can get more reliable and accurate results. Different features like road surface, lighting conditions, weather conditions, and casualty severity were used to build a model to envisage the occurrence of a road accident. The key function of the research is to classify the most significant factors that lead to road accidents. We used the Weka Tool to build a classifier. Four different algorithms were applied to the dataset. These methods were Decision Tree (DT), J48 and Random Tree, and Hoeffding Tree. To make the prediction more effective, the dataset was estimated in three unusual modes. Firstly, the training set will be the complete dataset. After this precision was estimated by splitting the dataset into 66%.and at the last 10 folds cross-validation mode was used.

Prediction results of the DT J48 model are shown in Table 3 . DT J48 prediction while using the whole dataset as a training set for the road surface, lighting condition, weather condition, and casualty severity are 84.79%, 85.66%, 84.38, and 87.31% respectively. The average prediction accuracy for the J48 model was 85% based on a split 66% accuracy prediction for the road surface, lighting condition, weather condition, and casualty severity were 82.56%, 83.14%, 83.42%, and 87.12%, respectively. The overall prediction for split 66% was 84%. DT J48 prediction using 10-fold cross-validation for the road surface, lighting condition, weather condition, and casualty severity was 84.27%, 84.12%, 84.27%, and 87.31%, respectively. Overall prediction accuracy for 10 folds cross-validation was 85.23%.

**Table 3: Prediction Results of the J48 Model**

| Model | Evaluation Method | Variables | Correctly Classified Instances | Incorrectly Classified Instances | Accuracy | Time |
|---|---|---|---|---|---|---|
| J48 | Using Training Set | Road Surface | 2259 | 405 | 84.79 | 0.01s |
| | | Lighting Condition | 2282 | 382 | 85.66 | 0.02s |
| | | Weather Condition | 2248 | 416 | 84.38 | 0.05s |
| | | Casualty Severity | 2326 | 338 | 87.31 | 0 s |
| | Using Split66% | Road Surface | 748 | 158 | 82.56 | 0s |
| | | Lighting Condition | 752 | 154 | 83 | 0s |
| | | Weather Condition | 754 | 152 | 83.22 | 0s |
| | | Casualty Severity | 782 | 124 | 86.31 | 0s |
| | 10 Folds Validation | Road Surface | 2245 | 419 | 84.27 | 0.02s |
| | | Lighting Condition | 2249 | 415 | 84.42 | 0.04s |
| | | Weather Condition | 2241 | 423 | 84.12 | 0.06s |
| | | Casualty Severity | 2326 | 338 | 87.31 | 0.01s |

The calculation outcomes of the Decision Stump tree model are shown in Table 4. Decision Stump Tree prediction accuracy using whole datasets training set for the road surface, lighting condition,

weather condition, and casualty severity was 83.10%, 80%, 83.48%, and 87.42%, respectively. The average accuracy for the Decision Tree based on the training set was of 83%.usinga a split of 66% data.

**Table 4. Prediction Results for Decision Stump Tree**

| Model | Evaluation Method | Variables | Correctly Classified Instances | Incorrectly Classified Instances | Accuracy | Time |
|---|---|---|---|---|---|---|
| Decision Stump | Using Training Set | Road Surface | 2214 | 450 | 83.1 | 0sec |
| | | Lighting Condition | 2148 | 516 | 80.63 | 0sec |
| | | Weather Condition | 2224 | 440 | 83.48 | 0sec |
| | | Casualty Severity | 2326 | 338 | 87.31 | 0sec |
| | Using Split66% | Road Surface | 738 | 168 | 81.45 | 0sec |
| | | Lighting Condition | 717 | 189 | 79.13 | 0.02sec |
| | | Weather Condition | 749 | 157 | 82.67 | 0sec |
| | | Casualty Severity | 782 | 124 | 86.31 | 0sec |
| | 10 Folds Validation | Road Surface | 2214 | 450 | 83.1 | 0.02sec |
| | | Lighting Condition | 2132 | 532 | 80 | 0sec |
| | | Weather Condition | 2224 | 440 | 83.48 | 0.09sec |
| | | Casualty Severity | 2326 | 338 | 87.31 | 0.02sec |

The accuracy for the road surface, lighting condition, weather condition, and casualty severity was 81.45%, 79.15%, 82.08%, and 86.53%, respectively. Overall accuracy was 82% using split 66% data. Using 10 folds cross-validation, the accurate forecasting for the road surface, lighting condition, weather condition, and casualty severity was 83.10%, 80%, 83.48%, and 87.34%, respectively. The average prediction accuracy for 10 folds cross-validation was 83%.

The accurate prediction of the Random Tree is shown in Table 5. Random Tree accuracy prediction using the complete dataset as the training set for the road surface, lighting condition, weather condition, and casualty severity was 100%, 100%, 100%, and 87.46%, correspondingly. The average correctness for the training set was 96.85%. Based on split 66% data, the accuracy for the road surface, lighting condition, weather condition, and casualty severity was 81.34%, 85.6%, 88%, and 86.53%, respectively. Overall accuracy prediction for split 66% was 85%. Using 10 folds cross-validation, the accuracy for the road surface, lightning condition, weather condition, and casualty severity was 91.32%, 84.75%, 91%, and 87.34%, respectively. Using 10 folds cross-validation shows the overall correctness was 88%.

**Table 5: Estimated Outcomes of Random Tree Classifier**

| Model | Evaluation Method | Variables | Correctly Classified Instances | Incorrectly Classified Instances | Accuracy | Time |
|---|---|---|---|---|---|---|
| Random Tree | Using Training Set | Road Surface | 2664 | 0 | 100 | 0.03s |
| | | Lighting Condition | 2664 | 0 | 100 | 0s |
| | | Weather Condition | 2664 | 0 | 100 | 0.01s |
| | | Causality Severity | 2570 | 94 | 96.85 | 0.02s |
| | Using Split66% | Road Surface | 737 | 169 | 81.34 | 0s |
| | | Lighting Condition | 776 | 130 | 85.65 | 0S |
| | | Weather Condition | 799 | 107 | 88.18 | 0s |
| | | Casualty Severity | 743 | 163 | 82 | 0.03s |
| | 10 Folds Validation | Road Surface | 2433 | 231 | 91.32 | 0.03S |
| | | Lighting Condition | 2258 | 406 | 84.75 | 0.08s |
| | | Weather Condition | 2433 | 231 | 91.32 | 0.04s |
| | | Casualty Severity | 2330 | 334 | 87.34 | 0.14s |

The prediction accuracy of the Hoeffding Tree is shown in Table 6 that holding tree accuracy using the whole dataset as a training set for the road surface, lighting condition, weather condition, and casualty severity is 93%, 97%, 92.98%, and 96.95% respectively. The overall prediction accuracy of training was 95%. Based on the split of 66%, the accuracy for the road surface, lighting condition, weather condition, and casualty severity is 80.13%, 74.50%, 82.56%, and 84.87%, respectively. The average accuracy of prediction using split 66% was 80.5% using 10 folds cross-validation. The prediction accuracy of road surface, lighting condition, weather condition, and casualty severity was 79.65%, 75.78%, 84.30%, and 86.12%, respectively. The average accuracy prediction for 10 folds cross-validation was 81.26%.

**Table 6: Prediction Results of Hoeffding Tree**

| Model | Evaluation Method | Variables | Correctly Classified Instances | Incorrectly Classified Instances | Accuracy | Time |
|---|---|---|---|---|---|---|
| Hoeffding Tree | Using Training Set | Road Surface | 2479 | 185 | 93 | 0.06s |
| | | Lighting Condition | 2585 | 79 | 97 | 0.01s |
| | | Weather Condition | 277 | 187 | 92.98 | 0.01s |
| | | Casualty Severity | 2583 | 81 | 96.95 | 0.02s |
| | Using Split66% | Road Surface | 726 | 180 | 80.13 | 0s |
| | | Lighting Condition | 675 | 231 | 74.5 | 0.01s |
| | | Weather Condition | 748 | 158 | 82.56 | 0s |
| | | Casualty Severity | 769 | 137 | 84.87 | 0.06s |
| | 10 Folds Validation | Road Surface | 2122 | 542 | 79.65 | 0.03S |
| | | Lighting Condition | 2019 | 645 | 75.78 | 0.02s |
| | | Weather Condition | 2246 | 418 | 84.3 | 0.06s |
| | | Casualty Severity | 2296 | 368 | 86.12 | 0.03s |

### 4.1 Comparison of J48, Decision Stump, Random Tree, and Hoeffding Tree

Overall accuracy in predicting the road accidents of J48, Decisions Tree, Random tree, and Hoeffding Tree is shown in Figure 3. Accuracy was measured in three ways, e.g., training dataset, 66% cross-validation, and 10 folds cross-validation. We can see that the results of J48 and Decision Stump Tree for the training set are similar, while the Random Tree and Hoeffding Tree outcomes are 100% and 95%respectively. Moreover, using10-fold cross-validation and split 66% Random Tree gave results far better than J48, Decision Stump Tree, and Hoeffding Tree.


**Figure 3: Overall Prediction Accuracy using all Techniques**

### 4.2 Weighted Average

True positive rate(TP), False positive rate (FP), precision, recall, and F-measure of all algorithms using complete data set as training set is shown in Figure 4. Whereas "DS-Tree" stands for Decision Stump Tree, "R-Tree" stands for Random Tree and "H-Tree" stands for Hoeffding Tree. Precision is a measure of correctness or eminence, while recall assesses entirety or quantity. A high recall value shows that the algorithm returned the most relevant results. When an algorithm returns a more relevant result, it is called high accuracy.

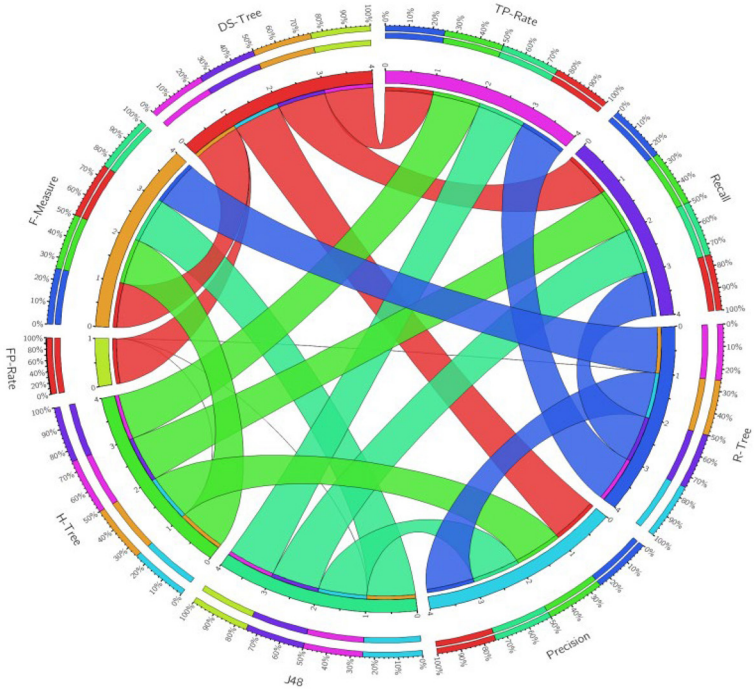


**Figure 4: Weighted Average Using Whole Dataset as Training Set**

Figure 5 shows the weighted average of the True positive rate (TP), False positive rate (FP), precision, recall, and F-measure of all algorithms using 10 folds cross-validation method.

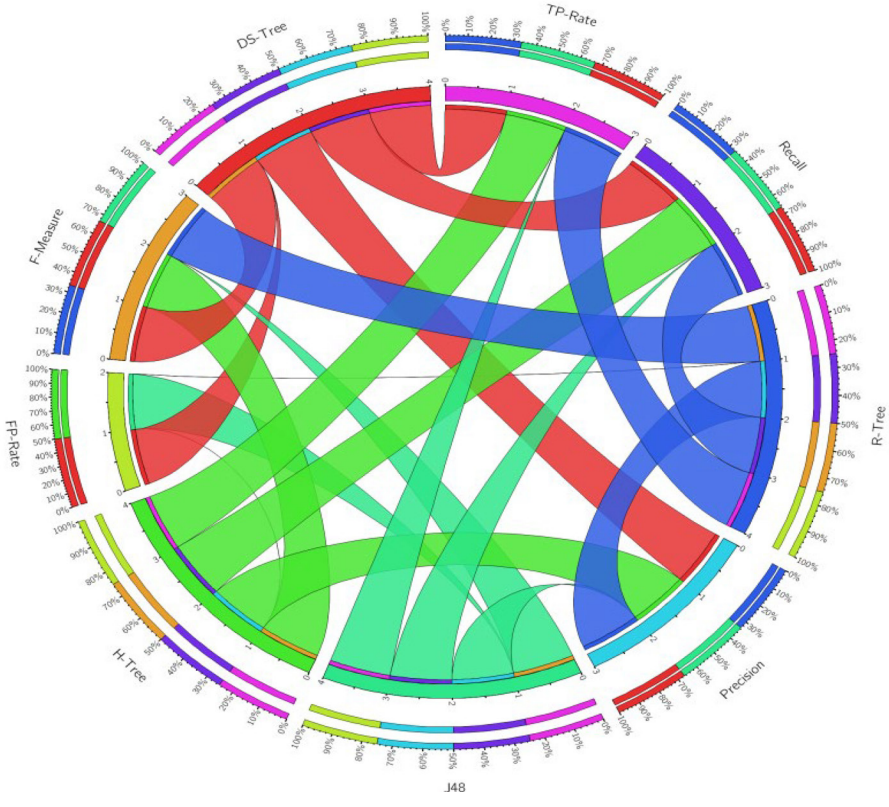


**Figure 5: Weighted Average Using 10 folds cross-validation**

Figure 6 shows the weighted average of the True Positive rate (TP), False Positive rate (FP), precision, recall, and F-measure of all algorithms using the split 66% method.

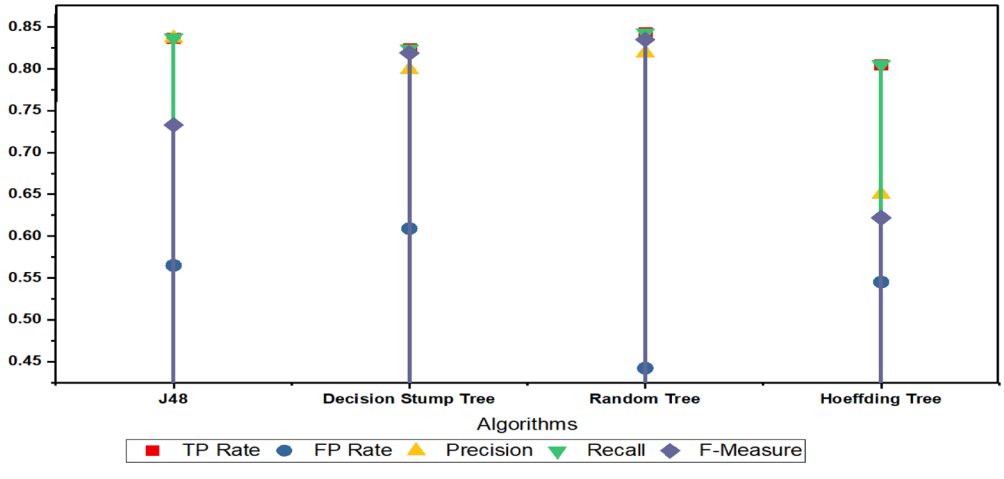


**Figure 6: Weighted Average Using Split 66%**

Many researchers have spent time finding the best-performing classifiers for data mining. Different classifiers have been applied to different datasets and their results have been investigated to choose the best among them. A similar kind of research has been made in which three different classifiers Multilayer Perceptron (MLP), J48, and BayesNet evaluated on a dataset of 150 instances using WEKA Tool [40].

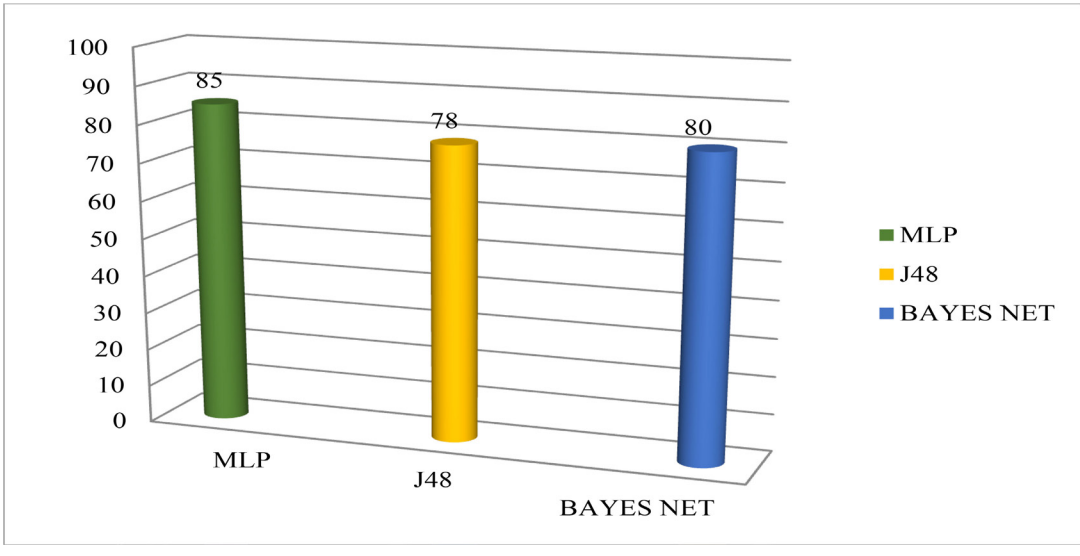


**Figure 7: Accuracy of Classifiers in Previous Research**

The Apriori algorithm was used to generate the association rules. The results found that the accuracy of Multi-Layer Perceptron, J48, and Bayes Net was 85%, 78%, and 80% respectively. While in our study we have applied different Decision Tree algorithms like Decision Stump Random tree J48 and Hoeffding Tree.

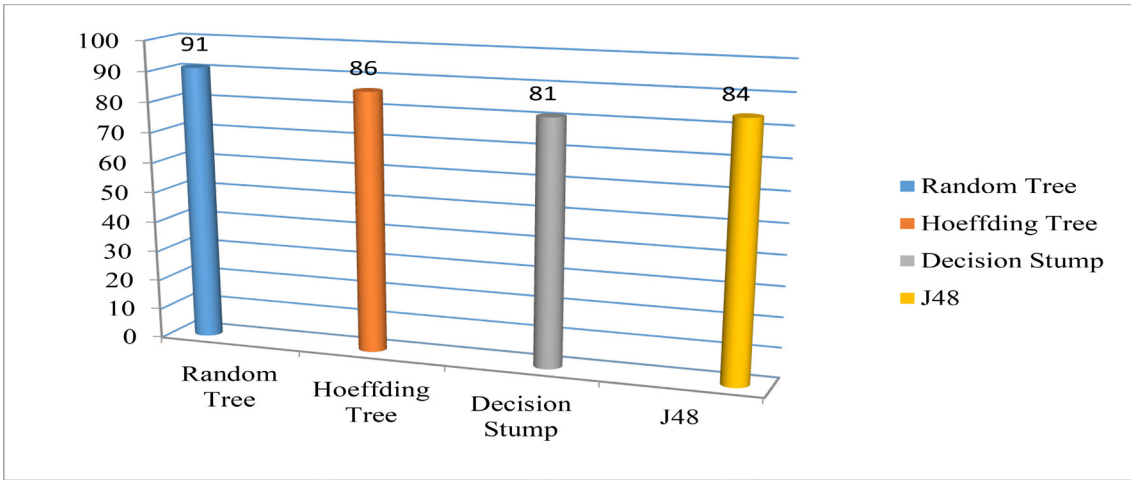


**Figure 8: Accuracy of Decision Tree Classifiers**

The results of all these algorithms were comparatively better than MLP, BayesNet, and J48. And among these Decision Trees, the Random tree outperformed the other algorithms. In another research, the authors have made a comparative analysis of different classifier techniques which include Naïve Bayes, SVM, and Decision Trees. They have used Self Organized Map (SOM) to make homogeneous segments of data. The classification accuracy of SVM, Naïve Bayes, and Decision Tree was 67%, 68%, and 71% respectively [41]. The accuracy of the Decision Tree used in our research is far better than the SVM model. Another thing that made our work unique is the use of Frequent Pattern (FP) Growth for association Mining. It is a far better technique than Apriori. So, it is recommended for researchers and scholars study the Random Tree algorithm for data mining classification. A comparison of the previous and current accuracy of all classifiers is shown in Figure 7 and Figure 8.

## 5. Conclusion

Accident data scrutiny is necessary to discover the features that lead to road accidents and to give some measures to decrease the risks related to that factor. The principle of our paper is to create classifiers that can accurately predict the occurrence of accidents and to make some rules that can be used to recognize the main features that are the reasons for accidents. The rate of accidents can be minimized in the future, and this article used the dataset of 2664 accident records of Leeds during the year 2015. We applied four different Tree Structures e.g., J48, Decision Stump, Random Tree, and Hoeffding Tree, to predict the accurate results. WEKA software was used in this research to establish the classifiers. The correctness of the J48 model for the complete dataset applied as training data, split 66% data and 10 folds cross-validation was 85%, 84%, and 85.23%. For Decision Stump, the prediction accuracy of the complete dataset applied as training data, split 66% data and 10 folds cross-validation was 83%, 82%, and 83%. Random Tree shows the prediction accuracy of the complete dataset applied as training data, split 66% data, and 10 folds cross-validation was 96.85%, 85%, and 88%. Hoeffding Tree accuracy prediction for the complete dataset used as training data, split 66% and 10 folds cross-validation was 95%, 80.5%, and 81.24% correspondingly. The accuracy of the Random Tree is far better than the other models. It outperforms the J48, Decision Stump, and Hoeffding Tree Models. In this study, we learned that road surfaces, weather conditions, and lighting conditions are the major factors that can cause road crashes. There is a tendency that injuries in road accidents often happen to car drivers more than others.

## References

[1] 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, 2021.

[2] M. Zheng et al., "Traffic accident's severity prediction: A deep-learning approach-based CNN network," vol. 7, pp. 39897-39910, 2019.

[3] H. Tariq, M. M. Iqbal, U. Khadam, and M. A. Al Ghamdi, "Analysis of Road Traffic Accidents to Improve Safety and Protection," Technical Journal, vol. 26, no. 04, pp. 32-39, 2021.

[4] C. Wang, C. Xu, Y. J. A. A. Dai, and Prevention, "A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data," vol. 123, pp. 365-373, 2019.

[5] M. John and H. Shaiba, "Analysis of Road Accidents Using Data Mining Paradigm," in Mobile Computing and Sustainable Informatics: Springer, 2022, pp. 215-223.

[6] C. Chen et al., "A rear-end collision prediction scheme based on deep learning in the Internet of Vehicles," vol. 117, pp. 192-204, 2018.

[7] Y. Yan et al., "Crash prediction based on random effect negative binomial model considering data heterogeneity," vol. 547, p. 123858, 2020.

[8] A. Comi, A. Polimeni, and C. Balsamo, "Road accident analysis with data mining approach: evidence from Rome," Transportation research procedia, vol. 62, pp. 798-805, 2022.

[9] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous Spatio-temporal data," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 984-992.

[10] C. C. Ihueze, U. O. J. A. A. Onwurah, and Prevention, "Road traffic accidents prediction modeling: An analysis of Anambra State, Nigeria," vol. 112, pp. 21-29, 2018.

[11] N. Arbabzadeh and M. J. I. T. o. I. T. S. Jafari, "A data-driven approach for driving safety risk prediction using driver behavior and roadway information data," vol. 19, no. 2, pp. 446-460, 2017.

[12] H. Meng, X. Wang, and X. Wang, "Expressway crash prediction based on traffic big data," in Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, 2018, pp. 11-16.

[13] K. Li, H. Xu, and X. Liu, "Analysis and visualization of accident severity based on LightGBM-TPE," Chaos, Solitons & Fractals, vol. 157, p. 111987, 2022.

[14] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 3346-3351: IEEE.

[15] Z. Halim, R. Kalsoom, S. Bashir, and G. J. A. I. R. Abbas, "Artificial intelligence techniques for driving safety and vehicle crash prediction," vol. 46, no. 3, pp. 351-387, 2016.

[16] L. Li, S. Shrestha, and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," in 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 2017, pp. 363-370: IEEE.

[17] A. V. Sakhare and P. S. Kasbe, "A review on road accident data analysis using data mining techniques," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-5: IEEE.

[18] Z. Yuan, X. Zhou, T. Yang, J. Tamerius, and R. Mantilla, "Predicting traffic accidents through heterogeneous urban data: A case study," in Proceedings of the 6th international workshop on urban computing (UrbComp 2017), Halifax, NS, Canada, 2017, vol. 14, p. 10.

[19] T. K. Bahiru, D. K. Singh, and E. A. Tessfaw, "Comparative study on data mining classification algorithms for predicting road traffic accident severity," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1655-1660: IEEE.

[20] Y. Zou, Y. Zhang, and K. Cheng, "Exploring the impact of climate and extreme weather on fatal traffic accidents," Sustainability, vol. 13, no. 1, p. 390, 2021.

[21] L. Wenqi, L. Dongyu, and Y. Menghua, "A model of traffic accident prediction based on convolutional neural network," in 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE), 2017, pp. 198-202: IEEE.

[22] C. Chen, X. Fan, C. Zheng, L. Xiao, M. Cheng, and C. Wang, "Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data," in 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD), 2018, pp. 328-333: IEEE.

[23] M. A. Rahim and H. M. Hassan, "A deep learning based traffic crash severity prediction framework," Accident Analysis & Prevention, vol. 154, p. 106090, 2021.

[24] G. Kaur and H. Kaur, "Prediction of the cause of the accident and accident-prone location on roads using data mining techniques," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-7: IEEE.

[25] M. Taamneh, S. Alkheder, and S. Taamneh, "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates," Journal of Transportation Safety & Security, vol. 9, no. 2, pp. 146-166, 2017.

[26] X. Xiong, L. Chen, and J. Liang, "A new framework of vehicle collision prediction by combining SVM and HMM," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 699-710, 2017.

[27] J. You, J. Wang, and J. J. J. o. m. t. Guo, "Real-time crash prediction on freeways using data mining and emerging techniques," vol. 25, no. 2, pp. 116-123, 2017.

[28] P. Li, M. Abdel-Aty, and J. Yuan, "Real-time crash risk prediction on arterials based on LSTM-CNN," Accident Analysis & Prevention, vol. 135, p. 105371, 2020.

[29] E. Kurakina, S. Evtiukov, and J. Rajczyk, "Forecasting of road accident in the DVRE system," Transportation research procedia, vol. 36, pp. 380-385, 2018.

[30] J. R. Asor, G. M. B. Catedrilla, and J. E. Estrada, "A study on the road accidents using data investigation and visualization in Los Baños, Laguna, Philippines," in 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 96-101: IEEE.

[31] S. Uma and R. Eswari, "Accident prevention and safety assistance using IoT and machine learning," Journal of Reliable Intelligent Environments, vol. 8, no. 2, pp. 79-103, 2022.

[32] S. H.-A. Hashmienejad and S. M. H. Hasheminejad, "Traffic accident severity prediction using a novel multi-objective genetic algorithm," International Journal of crashworthiness, vol. 22, no. 4, pp. 425-440, 2017.

[33] C. Sinclair and S. Das, "Traffic Accidents Analytics in UK Urban Areas using k-means Clustering for Geospatial Mapping," in 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET), 2021, pp. 1-7: IEEE.

[34] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K.-S. Kwak, "Traffic accident detection and condition analysis based on social networking data," Accident Analysis & Prevention, vol. 151, p. 105973, 2021.

[35] S. Elyassami, Y. Hamid, and T. Habuza, "Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees," in 2020 6th IEEE Congress on Information Science and Technology (CiSt), 2021, pp. 520-525: IEEE.

[36] Alim, Affan, Abdul Rafay, and Imran Naseem. "PoGB-pred: prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach." Current Bioinformatics 16.3 (2021): 446-456.

[37] J. You, J. Wang, and J. Guo, "Real-time crash prediction on freeways using data mining and emerging techniques," Journal of modern transportation, vol. 25, no. 2, pp. 116-123, 2017.

[38] S. Hussain, L. Muhammad, F. Ishaq, A. Yakubu, and I. Mohammed, "Performance evaluation of various data mining algorithms on road traffic accident dataset," in Information and Communication Technology for Intelligent Systems: Springer, 2019, pp. 67-78.