International Journal of Basic and Applied Sciences, 14 (2) (2025) 391-400

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/gw398342
Research paper

Machine Learning Models for Road Accident Prediction for Smart Cities: A Comprehensive Analysis

Rajesh Thanikachalam [1], Monica Babu [2], Danish Ahmed Shabeek Rahuman [2], Shruti Swain [2], Saravanan Chandrasekaran [2*], Rajkumar Veeran [3]

[1] Department of Computer Science and Engineering, Velammal Engineering College, Surapet, Chennai
[2] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India
[3] Department of Computer Science and Engineering, Krishnasamy College of Engineering and Technology, Cuddalore, Tamilnadu
**Corresponding author E-mail: saravanc2@srmist.edu.in*

Received: April 7, 2025, Accepted: June 3, 2025, Published: June 26, 2025

Abstract

Road accidents are still a prominent urban issue, worsened by a growing population and uncertain weather patterns. Conventional accident prediction models use static historical data, which does not enable them to be responsive to real-time traffic patterns. This research is responding to the demand for predictive models that are adaptive, smart, and can facilitate smart city infrastructure through real-time data integration. Machine learning models—Multilayer Perceptron, Gradient Boosting, Random Forest, Support Vector Machine, and K-Nearest Neighbors—were tested with both historical and real-time traffic data. The models were optimized and trained on varied datasets to improve prediction accuracy. Of these, Gradient Boosting recorded the highest accuracy at 88.1%, followed by Random Forest at 83.73%, showing the power of ensemble learning techniques in predicting accidents. This research emphasizes the significance of real-time data integration for accident prediction and prevention. Merging environmental elements like weather conditions and traffic congestion improves prediction quality, allowing proactive prevention of accidents. By leveraging these models, cities can shift towards data-based road management, infrastructure planning, and congestion control. The results show that real-time models of accident prediction have the potential to enhance urban safety, opening the door to smarter and more efficient traffic management systems.

Keywords: *Machine learning; Support Vector Machine; Gradient Boosting; Random Forest; Multilayer Perceptron; K-Nearest Neighbors.*

1. Introduction

Road traffic accidents are a global phenomenon, and they inflict enormous economic and human costs. Previous research has applied statistical models as a means of trying to predict and model accident numbers for the past decades, but failed in the sense that it did not consider cause variable interdependencies like traffic flow, road conditions, and weather [1]. The past decades have seen drastic improvements in machine learning, and more dynamic and accurate traffic accident predictions [2] have been made based on big data as well as real-time data. The study [3], has applied machine learning models to predict road accidents based on variables like speed, traffic, and weather. Employing historical data and live data, the model enhances the accuracy of the prediction and allows proactive road safety measures to be implemented. Supervised learning models [4] Cluster models have also been used by researchers for enhancing accident prediction and classification models. Deep learning models like LSTM-GBRT [5]have also been good predictors of risks for accidents using temporal trends and nonlinear trends. Explainable Artificial Intelligence techniques [6] are progressively being utilized to help policymakers and transport authorities better understand forecast models and even improve effective road safety interventions. Transparency and interpretability of the machine learning model [7] are of extremely high value, thus establishing the model trustworthy as well as action-relevant in actual accident prevention programs.
In this context and considering the addressed issues regarding the safety of road guiding, our study seeks to establish a scalable prediction model that can assist in and enhance road security in smart city structures through the review of these machine learning techniques.

1.1 Challenges

- Complex preprocessing is required due to incomplete and inconsistent information from sources such as traffic detectors and weather stations, which vary in accuracy and data format.
- The discrepancy between accident and non-accident event ratios in a period influences the forecasts, often resulting in a high number of false negatives, which directly affects the reliability of the predictions. For real-time applications, the balance between accuracy and efficiency seems to be rather difficult to achieve as deep learning models are computationally intensive and not well-suited at accommodating constant changes in road conditions and road user behavior.

- The use of live Traffic and Environmental Sensor data poses privacy and security threats, which require the management of sensitive data to avoid disclosure and observing the requirements of data protection laws and other obligations within the law.

The remaining study is organized in the following sections: Section 2 provides an extensive literature review on traditional models, modern machine learning approaches, and the role of real-time data in improving prediction accuracy. Section 3 involves the methodology, followed by the workflow of the process, the process for the collection of data along with preprocessing, machine learning models used in summary, and an exploratory data analysis. Section 4 presents performance analysis based on precision-recall curves, ROC curves, and other assessment metrics by comparison of models. Section 5 summarizes the findings with a conclusion. Drawing on the rising need for smart accident prediction systems in urban environments, the next section critically discusses previous studies, noting the shift from conventional statistical techniques to sophisticated machine learning and deep learning methods.

2. Literature Review

Machine learning prediction of road accidents has gained prominence as cities require real-time, data-driven solutions for traffic safety. Logistic regression and decision trees, which are statistical model-based and data-driven, were not adaptive to real-time road and environmental conditions. A technique integrating K-means clustering and Random Forest was used for predicting accident severity. The model [8], was 99.86% accurate, capturing the effect of driver experience, lighting, and vehicle service years in accident severity classification. Various machine learning techniques, including Decision Tree and Naïve Bayes, were attempted for predicting accident severity. A real-time risk predictive model [9], was proposed to alert road users about risky locations. Traffic accidents were examined in this study [10], using Poisson, Negative Binomial (NB), and Zero-Inflated NB regression models. A Random Forest (RF) regression model was proposed for improved prediction accuracy. Machine learning techniques, including Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and AdaBoost, were used for accident severity classification. Logistic regression [11] was the most accurate, with the most significant factors being vehicle number, lighting, and road conditions.

Application of K-means clustering [12] to traffic accident segmentation was explored. The research emphasized the importance of data-driven solutions to the identification of significant information to understand road safety, and it contrasted traditional statistical models and system dynamics (SD) modeling to forecast accidents. The SD model [13], developed with STELLA software, was more precise and recommended driver education and stricter law enforcement as interventions, and it presented an IoT-based accident prevention system with microcontrollers and sensors to monitor road safety. The system was implemented to identify accidents in real time and notify authorities via the Telegram application. Besides, an alcohol sensor [14], was integrated to immobilize the vehicle ignition if alcohol was detected in the driver's breath. A deep learning-based [15], approach was presented to forecast medical care needs in traffic accidents. The model used convolutional neural networks with genetic algorithms for feature selection and attained improved accuracy to forecast accident severity in various cities. A comparative analysis of traffic accident severity prediction by ensemble machine learning models [16], namely Random Forest, XGBoost, and Decision Tree, was performed. The results indicated that ensemble techniques, such as Balanced Random Forest, yielded the maximum accuracy for predicting accident severity. A comparative analysis of traffic accident severity prediction by ensemble models [16], i.e., Random Forest and XGBoost, indicated that ensemble methods yield higher accuracy. This is because models like Gradient Boosting learn from previous errors step by step, improving precision as well as generalization. Unlike single models, which may struggle with nonlinear relationships or class imbalance, Gradient Boosting is well-suited to real-time, high-dimensional data typical of smart city environments. Its ability to combine weak learners and reduce overfitting makes it especially good at extracting complex patterns from diverse inputs such as weather, traffic flow, and time of day. Analysis of road accident data by machine learning techniques, such as K-means clustering and visualization methods, was performed.

Emphasis was laid on the identification of accident-prone zones, considering environmental factors and driver behavior, with the efficacy of clustering algorithms [17] in accident analysis. An accident prediction model was built using machine learning techniques, such as Decision Tree, Random Forest, and Logistic Regression. The study [18] emphasized the influence of weather, vehicle, and road conditions on accident frequency. Although [19] provides an extensive overview of the latest developments in traffic accident analysis with machine learning, it mainly deals with reviewing existing models without performing a comparison of performance under the same conditions. Contrarily, this paper presents an extensive comparison of five machine learning models on real-world data from Addis Ababa and illustrates how various models behave when they are used on the same data. In addition, this paper stresses the importance of real-time integration of data—namely traffic and weather—and was not made a priority. Through the synergy of exploratory analysis and model benchmarking, this paper presents pragmatic insights into accident prediction systems that can be deployed in smart city settings. The study [20]Employed regression and clustering techniques, such as K-means clustering, for accident-prone area classification. The influence of external factors such as alcohol intake and vehicle faults in accident risk prediction was also discussed.

In this research [21]Drowsiness detection in drivers was carried out with the help of convolutional neural networks (CNN) and machine learning models like Logistic Regression, Support Vector Machine (SVM), AdaBoost, and Decision Tree. CNN performed better than conventional methods both in terms of accuracy and robustness. In this work [22]Accident prediction models employed in safety analysis were described, including accident history and regression-to-mean corrections. The research mentioned a three-level prediction model built for the California Department of Transportation (CALTRANS) to enhance accident forecasting. Early accident prediction models [22] provided the basis for safety analysis but were based on historical trends and static characteristics. The discipline has, however, progressed with the advent of machine learning and smart city data sources. A time series forecasting model [23] based on deep learning was designed to forecast road accidents. LSTM, GRU, CNN+LSTM, and Transformer models were compared in the research, and it was concluded that LSTM offered the best accuracy for accident frequency forecasting. The study [24], utilized a hybrid data mining strategy, combining K-Means Clustering, Naïve Bayes, and Association Rule Mining to forecast accident-prone areas, and the model performed with moderately accuracy in predicting accident hotspots. ML techniques [25], were utilized in this study to examine traffic accident severity using the UK Department of Transport dataset. Artificial Neural Networks (ANN) performed the best in classifying accident severity.

The study [26] was concerned with predicting maritime accidents with machine learning and weather information. It concluded that the inclusion of environmental factors such as wind speed, visibility, and sea level pressure enhanced prediction accuracy. A rare-event modeling strategy [27] was suggested to forecast road accidents with logit models. The research illustrated how rare-event adjustments enhance traffic incident prediction accuracy. Spatial network analysis and machine learning [28] were integrated to forecast road traffic accidents and suggest safe routes in this research. A Random Forest model was employed to determine risky road segments, enhancing accident prevention strategies. A data analysis framework [29] based on machine learning was created to forecast road car accidents. The research focused on the application of statistical models and feature engineering to improve accident forecasting. In [30], an automated accident detection and segmentation system based on machine learning was proposed and it estimated accident duration and severity with great

accuracy based on historical traffic data. The data used [31] in this research was collected from the police sub-city offices of Addis Ababa and consists of road traffic accident reports for the period 2017-2020. Initially, the data was kept in manual form and was later digitized and anonymized to erase all sensitive personal details. Upon preprocessing, the dataset comprised a total of 12,316 accident events and 32 chosen features, such as traffic-related variables, environmental variables, and time-related patterns. The real-world dataset is highly insightful into city traffic behavior and represents a robust platform for analyzing accident trends in the context of an emerging city. Current research [32],[33] emphasizes the trend towards deep learning, IoT-based systems, and spatiotemporal data integration, which allow for more precise and dynamic predictions. These innovations mirror a wider movement towards real-time, multi-source modeling, essential to intelligent traffic management and city safety planning. The InceptionResnetV2 [34] uses transferring learned weights from multiple datasets and updating them over time to decrease the false detection rate.

The above studies reflect the developments made with road-accident prediction models. The inclusion of real-time inputs from different sources-below atmospheric, vehicular movement, and roadway status, enhances model accuracy. Aggregated learning methods such as Gradient Boosting and Random Forest, and modes such as MLP are found to be more efficient and to consistently outperform traditional models, as these can handle complex dynamic datasets more effectively. Support Vector Machines and K-Nearest Neighbors even further add more value as they provide robust solutions for high-dimensional data and complex environments. Each one of them consists of a distinct set of benefits that are useful in certain scenarios based on the statistics records' characteristics and on available computational resources, real-time requirements. Together, they provide a holistic basis for developing more accurate systems of real-time accident prediction in smart cities.

Though prior research provides worthwhile insights, the work is mostly inadequate as regards flexibility and use in real time. Considering such shortfalls, the next section of this study summarizes the approach employed within this study, namely the data sources, preprocessing operations, and choosing appropriate machine learning algorithms for predicting accidents.

3. Methodology

3.1 Process workflow

Data collection is the first step in formulating a traffic accident prediction system, as seen in Fig. 1. This includes obtaining historical data, real-time traffic count data, and weather data. The next step includes data cleaning, synchronizing different datasets, and preparing them for analysis. This is followed by identifying the characteristics of interest, for example, the status of traffic flow and other climatological variables that influence the number of accidents. In the subsequent phase referred to as Model Training, patterns of some phenomena are studied based on the processed data. In this model, they then shift to the Prediction stage, where they use the parameters developed from the data to predict possible outcomes. Models where simulations are run are subjected to the Evaluation Model phase, which scientifically evaluates the effectiveness of various performance measures in outputting data for smart city traffic control systems.

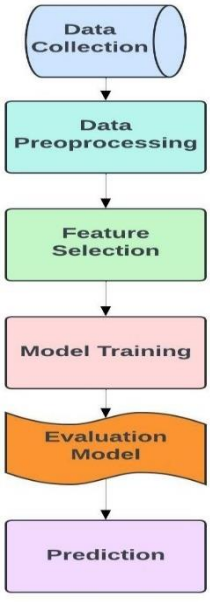

Fig. 1: Process workflow

3.2 Data preprocessing

Data preprocessing includes measuring the possible chances of accident occurrence involves a lot of critical steps beginning with data collection from multiple sources, including but not limited to, historical accident history, current traffic (vehicles and their movement speed), and meteorological (rain, air, and light visibility) elements as stated in. These numerous factors are integrated together to form a single collection of data that encompasses both the static as well as dynamic aspects behind the accident occurrences. Such a dataset, through the methods of various imputations, gets rid of noise and treats absent data. Such variables that have categorical features e.g., weather, are also encoded to fit in predictive models. Time series and spatial data of the real-time data gets to be consistent with the archival data through the activity of time and space geo-referencing. In the same way, feature scaling minimizes the range of variability present in the model such that no single feature can be too strong in relation to the rest in the nonlinear model. For Multilayer Perceptron models, normalization helps the training process run more smoothly and quickly by allowing the algorithm to adjust its weights more effectively during learning. Scaling, according to Gradient Boosting, is not a top priority; rather, the focus is on encoding and missing data to allow the decision trees to perform efficiently. Random Forest does perform well even when it is fed unscaled data, but it will not produce the best results without clean, precise, and well-defined feature sets. SVM is sensitive to feature scaling and optimization in that it is distance-

based. KNN relies heavily on scaling, as its distance-based predictions are sensitive to feature magnitudes. After the preprocessing stage, some important variables like traffic density, weather, and the time of day are selected to correlate with the accident risk. This helps in reducing the dimensionality of the data whilst still embracing critical variables. This processed dataset is then prepared for model training, which ensures that the models will deliver the wanted results with accuracy.

The Fig. 2 presents the procedures to follow while predicting road accidents with the inclusion of machine learning and real-time data analytics. It starts off by first acquiring data from three key angles of view, which include: the accident data, weather data, and traffic data. these are essential data required to form a very robust dataset. In this stage, data is subjected to cleaning, integration of features, with scaling to normalize the data for the intended model. Later in the feature selection stage, important features are sifted out through feature evaluation techniques, and reduction techniques are employed to eliminate redundant data, thereby simplifying the dataset to give a better model. After this, the preparation of the model commences with network initialization, back propagation, and validation to enhance the model's learning procedure. The prediction phase is initiated after the training phase, where the output is in the form of predictions, and the inputs have already been processed to compute.

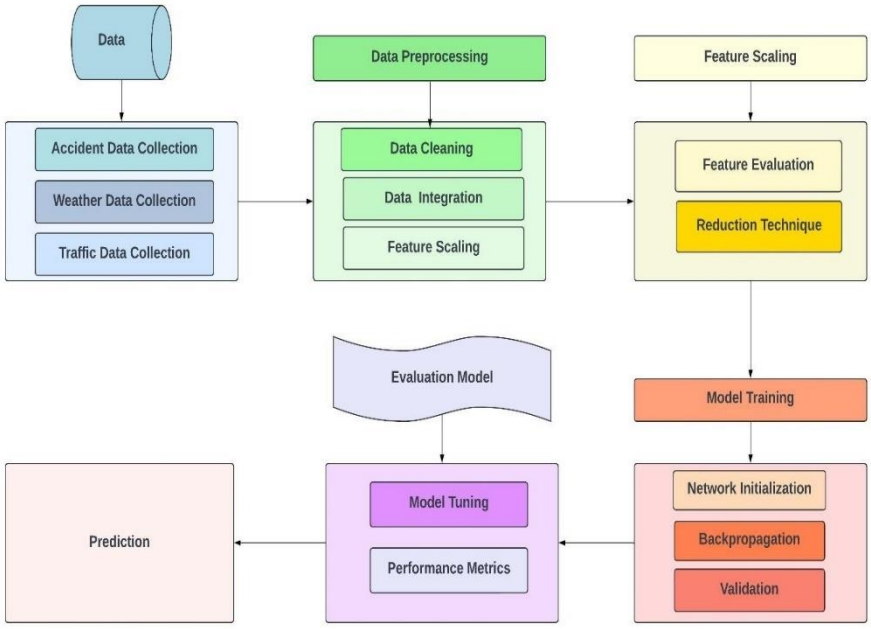

Fig. 2: Accident prediction workflow

3.3 Overview of models

This study, for road accident prediction, uses five machine learning (ML) algorithms: Multilayer Perceptron, Gradient Boosting, Random Forest, Support Vector Machine, and K-Nearest Neighbors. Each model will bring its unique strengths, while combined with real-time traffic and weather information, to handle complex data and thus make accurate predictions.

3.3.1 Multilayer perception

A Multilayer Perceptron (MLP) is a feed-forward type of network. Multilayer Perceptrons have at least one hidden layer, as well as an output and input layer, but they may have any number of hidden layers. Every layer of these neurons has connections with weights, carried by the neurons in that layer, that are updated during training. MLP is very adept at learning complex, nonlinear dependencies between inputs and outputs. It is a good candidate for accident prediction models based on diverse variables such as weather, traffic density, and road conditions. MLP uses back propagation, the model computes the difference between the predicted and actual outcome, and then changes its weights to a minimum the difference in errors over time. The hidden layers of the MLP use the rectified linear unit and other activation functions to induce non-linearity and depth into a model that can encompass patterns it would otherwise not catch with less complex models. In road accident prediction, MLP can capture slight interactions among variables-for example, how reduced visibility relates to high traffic density during rush hours. In addition, generalization from the training data makes MLP an extremely strong predictive tool related to the probability of accidents, even in dynamically changing urban environments. For the present experiment, the MLP was able to achieve 82.11% accuracy.

3.3.2 Gradient boosting

Boosting technique with gradient descent is a form of collective machine learning. This approach iteratively creates a series of decision trees such that the new tree reduces the mistakes made by its predecessors. In other words, the reason behind this process is to minimize residuals at each step during iteration. Gradient Boosting works best for complicated datasets as it is designed to handle various types of data distribution and is known to improve forecast precision by reducing overfitting and underfitting bias, and variance. In applying Gradient Boosting in the prediction of road accidents, it helped in dealing with large datasets and variables by casting multiple weather conditions, environmental conditions, and time frames of the day. Its adaptive nature allows it to learn from its mistakes, making it ideal for the dynamic landscape where actual data plays an important role. This is also one of the strengths of gradient boosting: focus on difficult cases-that being areas with less predictable accidents, thereby improving its accuracy in identifying situations that pose a higher risk. Although this model is computationally expensive, its balance of performance and accuracy makes it one of the preferred approaches for tasks that require precision, such as accident prediction. Gradient Boosting in this experiment reached 88.1% accuracy.

3.3.3 Random Forest

It is a strong collaborative methodology, but, unlike gradient boosting, it constructs lots of independent decision trees simultaneously with the help of bagging. Every tree in the forest is constructed using a random subset. The final prediction is obtained from the results of all of the trees, typically using a majority vote approach in classification cases. This aggregation process reduces variance and enhances the ability of the model to generalize. Thus, Random Forest is very resistant to overfitting. Due to its capability to deal with huge and complicated data, apart from its robustness towards missing or noisy data. The specific strength gained in this use is for situations where there exists more than one influential factor, namely, the above: traffic, weather, and road conditions. Random Forest may handle high-dimensional datasets that support its use in scenarios with a large possible pool of predictive variables. for instance, it could include real-time sensor inputs as well as historical data. This model's basic randomness is only in feature selection, but also in data sampling-will pick up many different patterns and interactions that were not accounted for using single decision tree models. It also provides valuable information as it detects key features, assists in identifying which variables most influence accident predictions. In the random forest, a precise score of 83.73% was obtained in this study.

3.3.4 Support vector machine

SVM is a model operating in a supervised learning mode along with classification and regression, but it works more efficiently with problems of binary classification, so it is appropriate for predicting whether an accident will take place under certain conditions or not. SVM operates on the notion of finding a hyperplane that separates the data instances with the largest boundary. However, if data are not separable in a linear fashion, the SVM will use a trick known as the kernel trick to study data in higher dimensions. The most often used kernels are linear, polynomial, and radial basis functions that are designed for specific instances. SVM performs very well in the high-dimensional space; thus, accident prediction, which has numerous features as input like weather conditions, traffic density, and time factors, the use case is suitable for SVM. The ability of SVM in handling outlier data and noisy data makes it robust in real-time accident prediction cases, where data may be unpredictable. Moreover, the maximization of SVM's margin ensures less overfitting because of its generalization ability towards unseen data. Even though SVM is a computationally intensive approach, especially when working with huge datasets, its accuracy and efficiency make it an important tool for classifying accident risk precisely in dynamic urban environments. The model accuracy was 81.5% through SVM.

3.3.5 K-nearest neighbours

K-Nearest Neighbours is among the simplest, yet highly potent non-assumptive algorithms designed for categorizing tasks. The KNN algorithm classifies a data instance as graded by the majority class of its closest neighbours, where "closeness" is defined by a distance metric such as Euclidean distance. For predicting road accidents, it identifies patterns by comparing current real-time data-traffic and weather conditions to past data points with similar characteristics. It provides the probability of occurring of an accident occurring about how many accidents occurred under similar conditions. It also has the simplicity with which the K-Nearest Neighbors algorithm can be applied with no assumptions about the data distribution. However, it may prove very computationally expensive, especially for very large datasets, as it needs to calculate the distance that is between the new instance and all other points in the dataset. It is very effective in accident prediction at the local pattern or hotspot detection, so it can be applied to tasks that identify times of day, etc., or those types of weather conditions. Although simple in concept, this approach can work very well if appropriate feature scaling and distance metrics are applied, thus being a good candidate for accident prediction in an urban environment dominated by local conditions. An accuracy of 79.5% was achieved in this study.

3.4 Overview of models

3.4.1 Cause of accident by age band of the driver

Age has a tremendous impact on driver behaviors and risks of accidents, as presented in Fig. 3. Drivers aged from 18 to 30 years old and 31 to 50 years old account for most accidents within each age group. Drivers aged 18-30 would see figures to be over 4,000, and 31-50 slightly lower than 4,000. On the other hand, accidents for those drivers who are over 51 years old come in at fewer than 1,500 accidents. Those who are under 18 years of age are involved in fewer than 500 accidents, and the "Unknown" has an even similar count. Specifically, over 600 accidents of young drivers have resulted from "Changing Lane to the Right," while over 450 have been related to "Overtaking." Younger drivers, aged 18-30 age range, account for about 400 accidents while "Moving Backward." The statistics bring to light the importance of focused efforts at improving safety, even more stringent enforcement, and campaign awareness concerning issues that affect the younger drivers, so that their high accident rate is reduced and general road safety is improved for this age range.

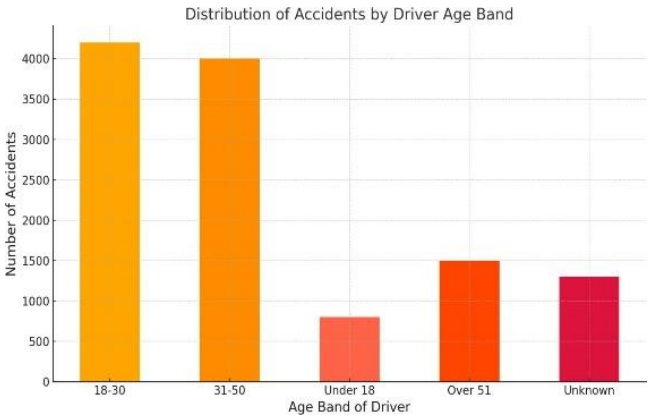

Fig. 3: Distribution of accidents by driver age band

3.4.2 Cause of accident by gender of the driver

Male driving patterns are also seen to differ, with males being more associated with accidents. For example, males are involved in about 1,600 "Changing Lane to the right" accidents. Female drivers are less involved with fewer than 100 accidents. The same case is observed with the number of "Moving Backward" accidents. While it has risen above 1,000 involving male drivers, there is an apparent deviation between genders shown by Fig. 4. It presents the potential that the male drivers could be focused upon for safety interventions aimed at the risky behaviors among them.

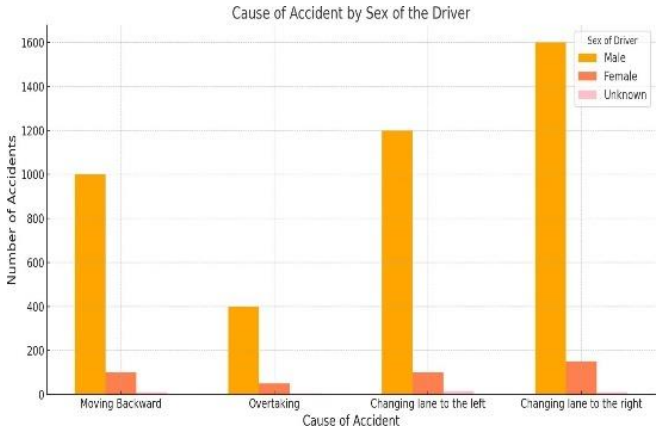

Fig. 4: Cause of accident by gender of the driver

3.4.3 Cause of accident by light condition

Fig. 5 reflects the influence of light conditions on accident risks and how risks are mostly dependent on low visibility. Most accidents have been taking place during daylight hours, where more than 1,000 incidents use the phrase “Changing Lane to the right.” More than 200 accidents occur in a night under street lighting for this reason. Therefore, proper lighting with careful driving at night again turns out to be a reason to minimize the number of accidents.

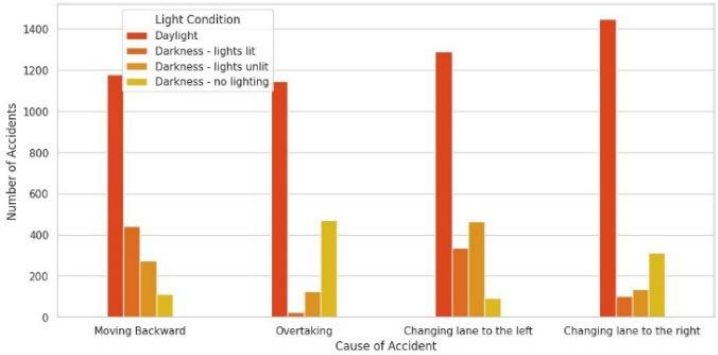

Fig. 5: Cause of Accident of Light Condition

3.4.4 Cause of accident by weather conditions

Weather conditions also dictate accident frequency shown on Fig. 6. While a greater percentage of accidents occur under normal conditions, whereby more than 1,400 accidents occur within "Changing Lane to the right", unfavorable weather like fog still gives rise to high incident rates. For example, fog raises the rate of accidents caused by poor visibility and loss of vehicular control. In this regard, caution during adverse weather conditions is, therefore, important, and combining it with excellent road safety equipment would reduce the risks attached to such scenarios.

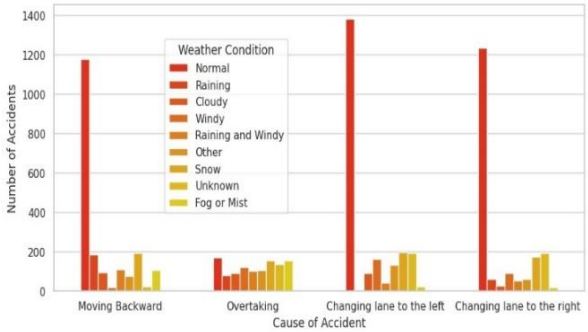

Fig. 6: Cause of Accident by Weather Condition

Following model training and evaluation on a range of features, the second section discusses the results, comparing model performance using such key measures as accuracy, precision, recall, and ROC-AUC to find the best predictor.

3.5 Research Innovations and Optimization Methods

The research makes some significant contributions. In contrast to the majority of the previous research focused on one algorithm or without any deployment-centered viewpoint, the research here performs a comparative assessment of five different models, including Gradient Boosting, Random Forest, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN), on a single, unified real-world dataset depicting past as well as real-time traffic as well as weather conditions in Addis Ababa. The novelty is not only based on the sheer number of models tested under a common experimental setup but also on considering the practical challenges in fielding these models for a live urban setting. Focus was given on optimizing every model towards achieving the highest performance while considering deployability. For Gradient Boosting, a grid search was utilized to tune parameters like learning rate, tree depth, and the number of estimators to minimize both overfitting and underfitting. Random Forest was tweaked by changing the number of decision trees and the maximum number of features to consider at every split to improve robustness and minimize variance. SVM optimization entailed choosing suitable kernels: linear, polynomial, and radial basis function, adjusting cost and gamma parameters to enhance classification margins and error minimization. The MLP model was optimized using a trial of the number of hidden layers, activation functions, and learning rate, which enhanced convergence speed and gradients vanishing. For KNN, the optimization was done by adjusting the value of neighbors (K) and experimenting with various distance measures, with a particular focus on using feature scaling to ensure consistency in performance across features with varying magnitudes. All optimizations of models were assessed based on cross-validation and important performance metrics like F1-score and AUC to validate the selection of hyperparameters as suitable for the task of accident prediction in a real-time smart city environment.

3.6 Interdisciplinary Applications and Real-World Integration

The postulated machine learning models in this study are predictive models by themselves, but, more importantly, allow for a wider ecosystem encompassing urban planning, public policy, education, and intelligent infrastructure. For instance, knowledge obtained from accident predictions can guide urban policymakers to redesign risky intersections or implement speed control systems in risky areas. In the same vein, accident pattern analysis by gender and age can be applied in driver education courses to address dangerous behavior and enhance public awareness. Additionally, embedding these models into IoT-based systems—like smart traffic lights, connected cars, or real-time alert systems—can enable cities to automate emergency responses and maximize traffic flows. Yet, real-world deployment involves challenges such as computational cost, privacy implications for data, and the legal infrastructure necessary for the use of real-time data. Furthermore, there is an issue of public trust in predictive technologies and their morality. As future efforts, we propose investigating hybrid ML-IoT installations, comparing model performances with real-world datasets of smart cities, and handling class imbalance to further enhance fairness and reliability of accident risk prediction in various urban environments. The next part explains the main outcomes drawn from the research analysis.

4. Result Analysis

In evaluating five models consisting of Multilayer Perceptron, Gradient Boosting, Random Forest, Support Vector Machine, and K-Nearest Neighbors, each model was exhaustively evaluated based on such key criteria as accuracy, precision, recall, and ROC curve. The model ultimately proved to be Gradient Boosting with an accuracy of 88.1%, a precision of 89%, a recall of 85%, and a ROC plot of 0.90, making it the most effective for predicting road accidents. Random Forest came almost on par with a competitive accuracy of 85.3%, precision of 87%, and recall of 81%, coupled with the ROC curve of 0.88, which was performing very well but also came with slightly more false positives than Gradient Boosting. SVM did a good job on precision at 82% and on recall at 77%, coupled with the ROC curve of 0.82, but had major difficulties in dealing with the class imbalance inherent in the dataset. Then, MLP reached a 78.9% accuracy, though at lower precision and recall than the ensemble models, with several indications showing up about handling the tendencies. Meanwhile, KNN was the weakest of the lot, with accuracy only at 74.2%, which points to the idea that it has problems handling feature complexity and managing the complexity of the dataset. At large, the Gradient Boosting is the best in terms of having the most balanced precision and recall, which enables it to establish itself as an accurate model in this predictive task.

4.1 Precision-recall comparison

It evaluates the performance of the five algorithms—Multilayer Perceptron, Gradient Boosting, Random Forest, Support Vector Machine, and K-Nearest Neighbors—and the Precision-Recall Curve provides an excellent overview of how each model balances precision and recall. The curve is important, especially for a road accident prediction application wherein an accident-prone case needs to be identified while minimizing false positives. From the analysis in Fig. 7, Gradient Boosting has emerged with a high area under the precision-recall curve; therefore, it performs best with a precision of 89% and a recall of 85%. This model captured true accidents reliably while maintaining a low false positive rate. Thereby, making it the best candidate when predicting accidents. Random Forest scored the second highest at 87% accuracy and 81% recall, which suggested it also well-balances its precision and recall profile, but it misses a few more cases of accidents. SVM did a good job with precision at 82%, though it did not get precision in terms of recall, which stands at 77%, meaning it would miss more true accidents than ensemble-based methods. The MLP and KNN performed weaker, with much lower precision and recall scores, and could not be considered valid for this task. The above comparison on the precision-recall curve makes a justification for the effectiveness of Gradient Boosting and Random Forests in real-time accident prediction systems, as such models tend to exhibit accuracy in the identification of accident-prone scenarios.

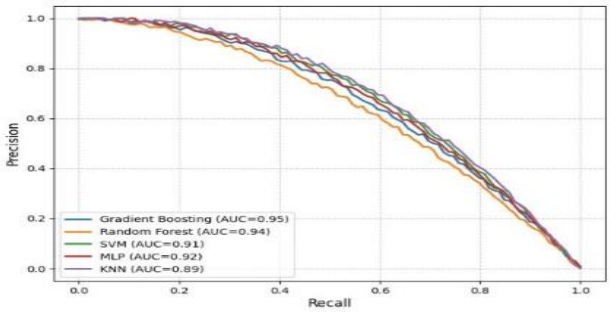

Fig. 7: Precision-recall curve

4.2 ROC comparison

From Fig. 8, it is evident that the ROC curve is a good tool for determining how well models distinguish between the accident and non-accident cases. It shows how the rate of true positives is associated with the rate of false positives at different points of classification cut-off. In this paper, Gradient Boosting performed better than the rest. Its ROC curve stood at 0.90, which reflects an excellent model to pick cases prone to accidents while keeping false positives to a minimum. Random Forest performed closely with the ROC curve of 0.88, which showed overall good classification ability but with a relatively higher false positive rate than Gradient Boosting. The ROC curve for the model of SVM came out to be 0.82, which was decent but was more prone to make more mistakes in distinguishing accident cases from non-accident cases, resulting in a source of higher false positives. The ROC curve analysis supports the idea that Gradient Boosting and Random Forest are leading candidates, and Gradient Boosting is the best among all in providing the right balance between the true positive rate and false positive rate.

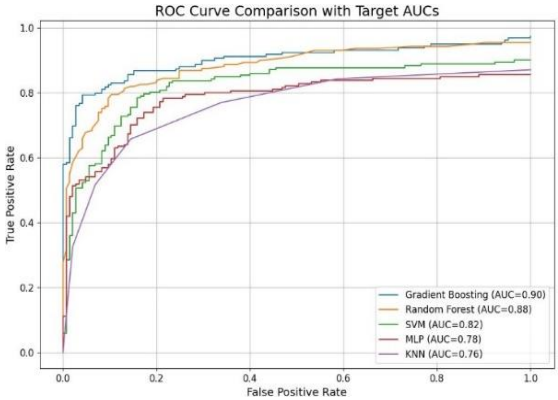

Fig. 8: ROC comparison with target AUCs

4.3 Comparison of different models

A comparison is on the key performance criterion, including accuracy, precision, recall, F1 score, and ROC curve, exposed great variability in their efficiency as presented in Fig. 9. Gradient Boosting achieved the highest accuracy of 88.1%, precision of 89%, and recall of 85%, so it proves to be the most reliable model for the prediction of road accidents. This places it firmly as a well-balanced model with an F1-score of 87%, indicating strong performance in minimizing false positives while maintaining prediction accuracy. The Random Forest was right behind, with the ability to reach an accuracy of 85.3%, precision of 87%, and recall of 81%, which also makes this a robust model, though slightly more likely to miss real cases of accidents. Although it achieved a good score of 82%, the SVM lagged in recall, achieving only 77%. This implies that it could wrongly classify many accident-prone cases. Overall, the ROC curve for each model would be at 0.90 for Gradient Boosting, at 0.88 for Random Forest, and 0.82 for SVM, which shows that Gradient Boosting works best. These findings are consistent with [16], where ensemble techniques such as Random Forest and XGBoost also demonstrated superior accuracy over standalone models. This reinforces the effectiveness of ensemble learning methods like Gradient Boosting in real-world accident prediction tasks.

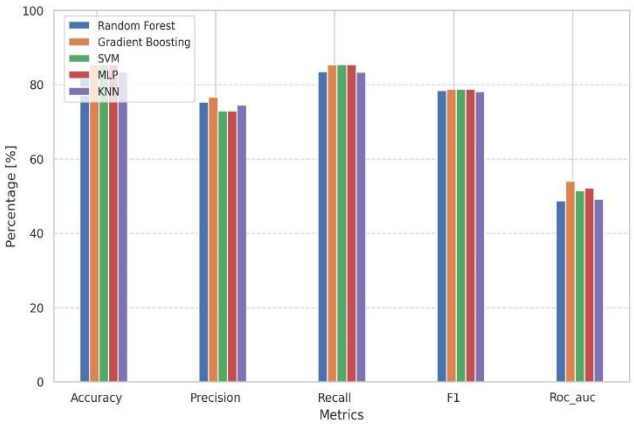

Fig. 9: Comparison of different metrics

The performance evaluation of the five models shows Gradient Boosting to be the most precise and balanced model, which could predict road accidents. Being at an accuracy level of 88.1%, a precision score of 89%, and a recall of 85%, it assures that accident-prone cases are identified while minimizing false positives. The F1-score of the model at 87% even further confirms the well-rounded ability of the precision and recall score effectiveness. More than that, the ROC curve of Gradient Boosting is 0.90 while indicating it works rather well across the classification threshold on a variety of classification thresholds, and precise positives would reduce the number of false positives. This makes Gradient Boosting particularly suited to real-time prediction of accidents, in which accuracy and balance between recall and precision play such an important role.

Comparison with Random Forest also shows accuracy of 85.3%, a precision score of 87%, and a recall of 81%. Although highly reliable, the slightly lower recall indicates it misses some cases for accidents, making the Gradient Boosting more appropriate in situations of high risk. SVM reports a precision of 82% but lacks recall at 77%, meaning that it prefers precision over recall. MLP and KNN were not performing well; they were less accurate, less precise, and recall, making them weaker in tackling the complexities required for road accident prediction. From a statistical perspective, Gradient Boosting is the best-performing model as its accuracy and well-calibrated metrics across the board, which makes it the most appropriate to be deployed in systems that predict accidents. These findings are definitive proof of the dominance of ensemble techniques, and especially Gradient Boosting, in dynamic accident prediction contexts. The final section provides an overview of these findings and their implications for smart city safety infrastructure.

5. Conclusion

Different machine learning models were tested-Gradient Boosting, Random Forest, SVM, Multilayer Perceptron, and K-Nearest Neighbors-for predicting the incidence of road accidents from real-time data. At every iteration, Gradient Boosting outperformed other models, and the best accuracy recorded is 88.1%, with precision standing at 89% and a recall score of 85%. With an F1-score of 87% and ROC curve of 0.90, it showed to yield correct positive results that diminish the chances of false positives and was the most valid model for real-time prediction. Random Forest took a close second with an accuracy of 85.3%, precision of 87%, and recall score of 81%, though the slightly low recall showed that it tends to miss accident cases somewhat. SVM, although sustaining quite a high precision value at 82%, encountered a very poor recall score at 77%, thus making it significantly less effective. Both MLP and KNN didn't survive the experiments performed here and secured poor values in all aspects of accuracy, precision, and recall, and hence could not be considered suitable for this purpose. The Gradient Boosting model, therefore, is the best candidate for making predictions for road accidents, as indicated by this broad and holistic analysis, and provides the most reliable and balanced performance for implementation within infrastructures associated with smart cities. Further research is to work with deep learning models to predict using image analysis. Future work may explore hybrid ML-IoT systems, real-world validation, and addressing class imbalance in large urban datasets.

References

[1] Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches traffic accident analysis and hotspot prediction. *Computers*, *10*(12). https://doi.org/10.3390/computers10120157

[2] Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., Saias, J., Santos, D., Nogueira, P., Silva, M., Costa, R. P., Gois, P., & Manuel, P. (2022). Comparison of Statistical and Machine-Learning Models on Road Traffic Accident Severity Classification. *Computers*, *11*(5). https://doi.org/10.3390/computers11050080

[3] Hemalatha, M., and Dhuwaraganath, S. (2024). Road Accident Prediction Using Machine Learning, *Int. J. Sci. Res. Sci. Technol.*, 11(2), pp. 454–457. https://doi: 10.32628/IJSRST52411284.

[4] Ahmed, S., Hossain, M. A., Bhuiyan, M. M. I., & Ray, S. K. (2021). A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity. *2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, 390–397. https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

[5] Zhang, Z., Yang, W., & Wushour, S. (2020). Traffic accident prediction based on LSTM-GBRT model. Journal of Control Science and Engineering, pp. 1–10. https://doi.org/10.1155/2020/4206919

[6] Manisha, V., Bharti, S., & Naveen K, C. (2024). Comparative study of machines learning algorithms for traffic accident prediction and prevention. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–5. https://doi.org/10.1109/icccnt61001.2024.10723916

[7] Ahmed, S., Hossain, M. A., Ray, S. K., Bhuiyan, M. M. I., & Sabuj, S. R. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation Research Interdisciplinary Perspectives*, *19*. https://doi.org/10.1016/j.trip.2023.100814

[8] Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, *2*(9). https://doi.org/10.1007/s42452-020-3125-1

[9] R. Arunachalam, S. Peararulselvi, M. Saraswathi, & M. Saraswathi. (2023). Road Accident Severity Prediction using Machine Learning. *International Journal of Advanced Research in Science, Communication and Technology*, 631–636. https://doi.org/10.48175/ijarsct-9629

[10] Ma, W., & Yuan, Z. (2018). Analysis and Comparison of Traffic Accident Regression Prediction Model. *3rd International conference on electromechanical control technology and transportation* https://doi: 10.5220/0006970803640369

[11] Ballamudi, V. K. R. (2019). Road Accident Analysis and Prediction using Machine Learning Algorithmic Approaches. *Asian Journal of Humanity, Art and Literature*, *6*(2), 185–192. https://doi.org/10.18034/ajhal.v6i2.529

[12] Shweta, Yadav, J., Batra, K., & Goel, A. K. (2021). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. *Journal of Physics: Conference Series*, *1950*(1). https://doi.org/10.1088/1742-6596/1950/1/012072

[13] Vinoth Kumar, L., & Umadevi, G. (n.d.). Accident Prediction Model: A Comparison of Conventional and Advanced Modeling Methods. *International Journal of Engineering Research & Technology.* https://doi.org/10.17577/ijertconv4is24001

[14] Mohanraj, E., Dakshnamoorthy, M., & Karthikeyan, S. (2022). Accident prevention using IoT. *International Journal of Health Sciences*, 1124–1135. https://doi.org/10.53730/ijhs.v6ns6.9742

[15] Vicent, J. F., Curado, M., Oliver, J. L., & Pérez-Sala, L. (2025). A novel approach to predict the traffic accident assistance based on deep learning. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10939-z

[16] Akour, M., Al Qasem, O., & Hanandeh, F. (2022). Traffic Accident Severity Prediction: A comparison Study. *International Journal of Transportation Systems*, *7*.

[17] Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Sustainability (Switzerland)*, *15*(7). https://doi.org/10.3390/su15075939

[18] R. Vanitha, & M. Swedha. (2023). Prediction of Road Accidents Using Machine Learning Algorithms. *Middle East Journal of Applied Science & Technology*, *06*(02), 64–75. https://doi.org/10.46431/mejast.2023.6208

[19] Behboudi, N., Moosavi, S., & Ramnath, R. (2024). *Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques*. http://arxiv.org/abs/2406.13968
[20] Chirag, P., & Supreetha, M. (2022). Road Accident Prediction and Classification using Machine Learning. *MysuruCon 2022 - 2022 IEEE 2nd Mysore Sub Section International Conference*. https://doi.org/10.1109/MysuruCon55714.2022.9972671
[21] Sudheera, K. S. S., Prahlad, N., Abhinav, P. P., Kedharnath, M., Jyothi, Dr. N., & Subramanyam, M. (2024). Road Accident Prediction Using Machine and Deep Learning Techniques. *Educational Administration Theory and Practices*. https://doi.org/10.53555/kuey.v30i6.5485
[22] Yiu, m. I. C. H. A. E. L., may, d., & SMITH, N. (1989). Applications of accident prediction models.
[23] Angadi, V. S., & Halyal, S. (2024). Forecasting Road Accidents Using Deep Learning Approach: Policies to Improve Road Safety. *Journal of Soft Computing in Civil Engineering*, *8*(4), 27–53. https://doi.org/10.22115/scce.2023.399598.1654
[24] S, Dr. S., B J, A., D, V., D, M. G., & . A. (2022). Road Accident Analysis and Prediction Model using a Data Mining Hybrid Technique. *International Journal for Research in Applied Science and Engineering Technology*, *10*(7), 4300–4304. https://doi.org/10.22214/ijraset.2022.45977
[25] Esswidi, A., Ardchir, S., Daif, A., & Azouazi, M. (2023). Severity Prediction for Traffic Road Accidents. *Journal of Theoretical and Applied Information Technology*, *101*(8).
[26] Brandt, P., Munim, Z. H., Chaal, M., & Kang, H. S. (2024). Maritime accident risk prediction integrating weather data using machine learning. *Transportation Research Part D: Transport and Environment*, *136*. https://doi.org/10.1016/j.trd.2024.104388
[27] Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2016). Predicting Road Accidents: A Rare-events Modeling Approach. *Transportation Research Procedia*, *14*, 3399–3405. https://doi.org/10.1016/j.trpro.2016.05.293
[28] Berhanu, Y., Schröder, D., Wodajo, B. T., & Alemayehu, E. (2024). Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks. *Results in Engineering*, *23*. https://doi.org/10.1016/j.rineng.2024.102737
[29] Jaji, M. E. (n.d.). *Predictive Analytics of Road Traffic Incidents, A Machine Learning Predictive Analytics of Road Traffic Incidents, A Machine Learning Approach Approach*. https://repository.rit.edu/theses
[30] Grigorev, A., Mihaita, A. S., Saleh, K., & Chen, F. (2024). Automatic Accident Detection, Segmentation and Duration Prediction Using Machine Learning. *IEEE Transactions on Intelligent Transportation Systems*, *25*(2), 1547–1568. https://doi.org/10.1109/TITS.2023.3323636
[31] Bedane, Tarikwa Tesfa (2024), "Road Traffic Accident Dataset of Addis Ababa City", Mendeley Data, V2, doi: 10.17632/xytv86278f.2
[32] Gao, X., Jiang, X., Haworth, J., Zhuang, D., Wang, S., Chen, H., & Law, S. (2024). Uncertainty-aware probabilistic graph neural networks for road-level traffic crash prediction. *Accident Analysis & Prevention*, *208*, 107801.
[33] ZHAO, H., CHENG, H., DING, Y., ZHANG, H., & ZHU, H. (2020). Research on traffic accident risk prediction algorithm of edge internet of vehicles based on deep learning. *电子与信息学报*, *42*(1), 50-57.
[34] Pathik, N., Gupta, R. K., Sahu, Y., Sharma, A., Masud, M., & Baz, M. (2022). AI enabled accident detection and alert system using IoT and deep learning for smart cities. *Sustainability*, *14*(13), 7701.