International Journal of Advanced Artificial Intelligence Research (IJAAIR)
eISSN: 3087-4270
Volume03 Issue07 | Page No: - 40-49

Received: 30 May 2026, Revised: 15 June 2026, Accepted: 28 June 2026, Published: 21 July 2026

DOI: https://doi.org/10.55640/ijaair-v03i07-04

Early Warning Systems for Traffic Accidents Using Predictive Machine Learning Models

Dr. Sunita Dixit [1]

[1] Professor, Department of Computer Science & Engineering, St. Andrews Institute of Technology & Management, Gurgaon, India

ABSTRACT: Accidents in transportation sector are a major public safety concern, particularly in high-traffic urban regions where increasing vehicle density and complex road conditions contribute to frequent crashes. The capacity to deliver timely early warnings based on predictive risk analysis is severely lacking in traditional traffic management systems, which are mostly reactive. This paper presents a solution to this constraint by leveraging the US Accidents (2016-2023) dataset, which includes 2,845,342 accident records with 46 characteristics, and proposing a prediction model based on Extreme Gradient Boosting (XGBoost) to provide early warning of traffic accidents. The proposed method incorporates exploratory data analysis, data preparation, label encoding, min-max normalization, SMOTE-based class balancing, and an 80:20 training-testing split, among a number of other things. The XGBoost model uses accuracy, precision, recall, and F1-score to evaluate the efficacy of its accident risk forecasting training. With an accuracy (ACC) of 95.9%, precision (PRE) of 96.0%, recall (REC) of 95.7%, and F1-score (F1) of 95.8%, the proposed model surpasses state-of-the-art machine learning (ML) methods, test results show, including FL, Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). According to these results, the XGBoost model is a solid and efficient way to anticipate when traffic accidents happen, which helps with IVSs by allowing for faster risk assessments and safer roads.

KEYWORDS: Predictive System, Real-time Traffic Monitoring, Emergency Response Optimization, Traffic Management, Real-Time Prediction, Crime Prediction.

1. Introduction

Road transportation is fundamental to economic development and human mobility. However, there has been a noticeable rise in traffic accidents globally due to the quick rise in car ownership and reckless driving practices. Every year, automobile accidents kill more than 1.35 million people and injure millions more, according to the World Health Organization (WHO) [1][2]. Human factors, including driver fatigue, distracted driving, speeding, and mobile phone usage, account for most accidents, although adverse weather, poor road conditions, and traffic congestion also contribute significantly [3][4]. Consequently, intelligent early warning systems capable of identifying accident risks before collisions occur have become an important research focus for improving road safety[5][6].

Through real-time monitoring and communication, recent advancements in Intelligent Transportation Systems (ITS), the Internet of Vehicles (IoV), linked cars, and ADAS have improved traffic safety. Modern vehicles utilize onboard sensors, cameras, radar, LiDAR, GPS, and wireless communication technologies to monitor driving conditions and exchange traffic information[7][8]. These technologies support collision prediction, traffic monitoring, and timely warning generation, enabling drivers to take preventive actions while improving traffic management and emergency response. However, connected transportation systems also introduce cybersecurity and privacy challenges, requiring secure and reliable data-sharing mechanisms[9].

The availability of extensive traffic accident datasets has led to a rise in the application of ML techniques for accident prediction and early warning[10]. ML algorithms examine intricate connections between variables like weather that are connected to accidents, road characteristics, traffic density, vehicle type, temporal information, and driver behavior to estimate accident risk. In addition, data mining techniques identify hidden patterns within accident data, improving the reliability of predictive models[11][12]. Compared with conventional statistical approaches, ML offers higher prediction ACC, better scalability, and greater adaptability for intelligent road safety applications. DL models including CNNs, RNNs, LSTM networks, and GNNs have further enhanced traffic accident prediction by automatically extracting characteristics from images, videos, and sensor data. These models enhance object detection, lane recognition, pedestrian detection, and traffic scene understanding, enabling accurate real-time collision prediction[13][14].

The adoption of DL techniques is limited in resource-constrained and real-time applications due to their substantial computing requirements, difficult model adjustment, and massive labeled dataset requirements. ML continues to be a viable option for practical accident prediction systems due to its efficiency and interpretability. This paper presents a framework for early-warning traffic accidents utilizing the XGBoost algorithm and predictive ML in response to these problems. By leveraging historical accident data and multiple traffic-related features, the proposed framework aims to accurately predict accident risk, identify influential features, and generate timely warnings to help drivers, transportation authorities, and intelligent transportation systems enhance road safety.

This research offers several key contributions as listed below:

- A predictive traffic accident early warning framework based on the XGBoost algorithm is proposed to accurately identify accident risks and support proactive road safety management.
- A comprehensive data preprocessing pipeline is developed, incorporating exploratory data analysis, missing value handling, train-test data partitioning, SMOTE-based class balancing, and min-max normalization to enhance model resilience and prediction performance.
- The usefulness of the suggested XGBoost model for accident risk prediction is thoroughly assessed utilizing a variety of performance measures, such as ACC, PRE, REC, F1, and ROC curve analysis.
- The most important accident-related factors are found and ranked using permutation feature importance analysis, providing improved model interpretability and insights into the key contributors to traffic accidents.
- An intelligent early warning framework is presented to enable timely accident risk prediction, thereby assisting drivers, traffic authorities, and Intelligent transport solutions to improve traffic safety and lower the probability of accidents.

The paper is organized as follows: Early warning methods for Predictive Traffic Accidents are discussed in Section II. In Section III, proposed early warning system is described. Section IV presents results obtained, and Section V draws the conclusions.

2. Literature Review

A comprehensive review of recent studies on Predictive Traffic Accident Early Warning was conducted to identify existing approaches, datasets, and research gaps. Table I summarizes proposed models, datasets, key findings, and limitations of previous studies, providing the foundation for the proposed research.

Kotsyubynska et al., (2026) a comprehensive evaluation and meta-analysis of DL and ML techniques for forecasting severity of traffic crash injuries, carried out in accordance with TRIPOD+AI criteria for prediction model reporting and PRISMA 2020 recommendations. Publications between 2014 and 2025 that satisfied following inclusion criteria were considered eligible: neural network-based observational studies that predict collision injury severity using reported data for the F1, G-mean, sensitivity, or confusion matrix. Over the course of 74 studies, 2,127,059 collision instances were analyzed. 78.6% (I^2=84%, 95% CI: 76.2-81.0%) was the pooled macro F1. With comprehensive imbalance treatment (SMOTE/ADASYN + class weighting), in comparison to 69.8% without handling, the fatal crash sensitivity increased from 42.1 to 73.6 percent, reaching 81.3% F1 (difference 11.5 percentage points, $p<0.001$). Half (56%) of the between-study heterogeneity was explained by meta-regression using study quality, algorithm type, sample size, and imbalance management[15].

Jabar and Hussain, (2025) examined driver behaviour, environmental variables, and traffic patterns. Key performance measures were used to assess these models. Python and RapidMiner were used in experiments, and RF classifier produced the best ACC of 94.8%. The knowledge gathered from this study helps create intelligent accident prevention systems, resulting in safer roads and fewer deaths[16]. Saranya et al., (2025) use ensemble classification methods and ML to forecast severity of a traffic collision. To increase classification ACC, AdaBoost with Random Forest (RF) combines the powerful prediction capabilities of RF with the boosting capability of AdaBoost. This model showed that it could handle imbalanced data and predict the severity of accidents with an ACC of 91.19%. The Flask framework was used to link trained ML models to the webpage, while HTML and CSS were used to construct the web interface[17].

Berhanu et al., (2024) suggested a new way to cut down on RTAs and traffic by using accident rates, an RF model, and spatial network analysis to show drivers the safest routes by examining the RF model and collision rate with accident data spanning 2014–2019, and able to forecast the frequency and occurrence of RTAs. The optimal path covers 28.6 kilometers in 41.58 minutes, whereas an alternate safe route covers 32.27 kilometres in 50.78 minutes. With a 78% success rate in predicting the target variable, the RF model has demonstrated its worth [18]. Alnashwan et al., (2023) suggest three ML methods for an Internet of Things (IoT) accident severity prediction system. The methods are RF, LightGBM, and XGBoost. Countrywide Traffic Accident Dataset, which covers 2.8 million vehicle accidents in the US from 2016 to 2021, was thus utilized. Finally, it seems that the method can accurately forecast how bad car accidents. With a prediction ACC of 72%, PRE of 70%, REC of70%, F1 of 70%, and area under the ROC curve (0.86), LightGBM surpassed the other two methods [19]. Kurian and Kumar Soori, (2023) suggest a way to reduce the hazards of traffic accidents, including sleepiness and attention, which have been shown to be the main causes of accidents globally. The designed system was subjected to different lighting conditions and evaluated in an experimental environment. According to the findings, the method can identify tiredness and distractions with 94.1% and 89% ACC, respectively, throughout day and night settings and issue warnings when necessary[20].

Yang, (2022) analyzes traffic accident data in the United States, this paper explores the practicability of cloud computing in data analysis and the possibility of using this model for traffic accident early warning. Then, traffic accident predictions are made using four ML algorithms in Spark MLlib. After comparing with the original data, LR, RF, and GB Trees all performed well, and despite the data's modest bias, the ACC rate can reach above 78%, indicating that this model can be utilized to forecast traffic accidents[21]. Mohd Shariff, Zainuddin and Amin Megat Ali, (2022) propose utilizing convolutional neural networks and auditory signals to identify wet road surfaces. The IDMT-Traffic database is the source of the data. The AlexNet is then trained using the scalogram created from the acoustic data. The Adam optimizer and low learning rate produce the greatest results, with an 89.9% validation ACC. In general, it is possible to identify wet road surfaces by using optimized AlexNet and auditory signals[22].

TABLE I. RECENT STUDIES ON PREDICTIVE TRAFFIC ACCIDENT PREDICTION USING MACHINE LEARNING TECHNQIUES

Author	Proposed Work	Dataset	Results	Limitations & Future Work
Kotsyubynska et al. (2026)	Systematic review and meta-analysis of ML/DL methods for traffic crash injury severity prediction following PRISMA 2020 and TRIPOD+AI guidelines.	Meta-analysis of 74 studies (2014–2025), 2,127,059 crash cases.	Transformer/LLM: 84.7%, Transfer Learning: 83.2%, Hybrid CNN-RNN: 81.8%, Overall Macro F1: 78.6%. DL outperformed conventional ML by 7.7%.	Review study only; does not propose a real-time early warning framework. Future work includes standardized datasets, external validation, explainable AI, and deployment in intelligent transportation systems.
Jabar & Hussain (2025)	Developed supervised ML models to predict road accidents using traffic, environmental, and driver behavior data for accident prevention.	Traffic accident dataset (processed using Python and RapidMiner).	Random Forest achieved 94.8% accuracy, outperforming other models.	Limited dataset diversity and geographical coverage. Future work suggests integrating real-time IoT/sensor data and DL for real-time prediction.
Saranya et al. (2025)	Road accident severity prediction using ensemble learning (AdaBoost + Random Forest) with SMOTE balancing and Flask-based web interface.	Road accident dataset with preprocessing, Label Encoding, and SMOTE.	91.19% accuracy using AdaBoost + Random Forest.	Focuses on severity classification rather than early warning. Future work includes real-time deployment, larger datasets, and IoT integration.
Berhanu et al. (2024)	Combined RF, crash-rate prediction, and GIS spatial network analysis for safe route recommendation and accident prevention.	Historical crash data (2014–2019).	78% prediction capability; identified safer routes with reduced crash probability.	Limited to historical crash records and regional data. Future work includes live traffic data, weather integration, and dynamic route optimization.
Alnashwan et al. (2023)	IoT-enabled accident severity prediction using Random Forest, XGBoost, and LightGBM based on weather conditions.	US Countrywide Traffic Accident Dataset (2.8 million accidents, 2016–2021).	LightGBM: Accuracy 72%, Precision 70%, Recall 70%, F1 70%, AUC 0.86.	Primarily weather-based prediction; lacks driver behavior and real-time sensor fusion. Future work suggests

				multimodal IoT data integration.
Kurian & Kumar Soori (2023)	Vision-based driver drowsiness and distraction detection using Raspberry Pi, eye tracking, and facial analysis for accident prevention.	Experimental driver monitoring dataset under different lighting conditions.	94.1% drowsiness detection accuracy and 89% distraction detection accuracy.	Focuses only on driver monitoring. Future work includes improved robustness, wearable sensors, and vehicle integration.
Yang (2022)	Cloud-based traffic accident prediction using Hadoop, Spark SQL, and Spark MLlib with ML algorithms.	US traffic accident dataset processed on Hadoop/Spark platform.	Logistic Regression, Random Forest, and GBT achieved over 78% accuracy.	Performance limited by data imbalance and offline analysis. Future work recommends real-time cloud analytics and larger datasets.
Mohd Shariff et al. (2022)	Wet road surface detection using acoustic signals and AlexNet CNN for safer driving.	IDMT-Traffic acoustic dataset.	89.9% validation accuracy using Adam optimizer.	Detects only wet road conditions rather than complete accident prediction. Future work includes multi-weather and real-time vehicle deployment.

Research gaps: Although previous studies have demonstrated effectiveness of ML and DL techniques for traffic accident prediction and severity analysis, many are limited by moderate prediction ACC, imbalanced datasets, or a focus on specific factors such as weather, driver behavior, or accident severity. Several existing approaches also rely on small or region-specific datasets and do not perform comprehensive feature selection or data balancing. Furthermore, few studies have developed an integrated predictive framework specifically for traffic accident early warning using large-scale real-world data. Therefore, there is a need for a robust and scalable XGBoost-based framework that combines effective preprocessing, feature importance analysis, and balanced learning to improve predictive traffic accident early warning performance.

2. Methodology

The aim of this study is to develop a reliable ML-based framework for early prediction of traffic accidents to support proactive road safety and intelligent transportation systems. To achieve this, an Extreme Gradient Boosting (XGBoost)-based predictive traffic accident early warning framework is proposed using the US Accidents (2016–2023) dataset. The methodology involves exploratory data analysis, data preprocessing, SMOTE-based class balancing, an 80:20 train-test split, and Min–Max normalization. The XGBoost model is trained to predict accident occurrence, and permutation feature importance is employed to identify most influential factors affecting prediction performance. ROC curve analysis, ACC, PRE, REC, and F1 are used to assess suggested model. The framework aims to improve prediction ACC while providing reliable early warning assistance for road safety applications. Fig. 1 illustrates the proposed workflow for predictive traffic accident early warning using ML.

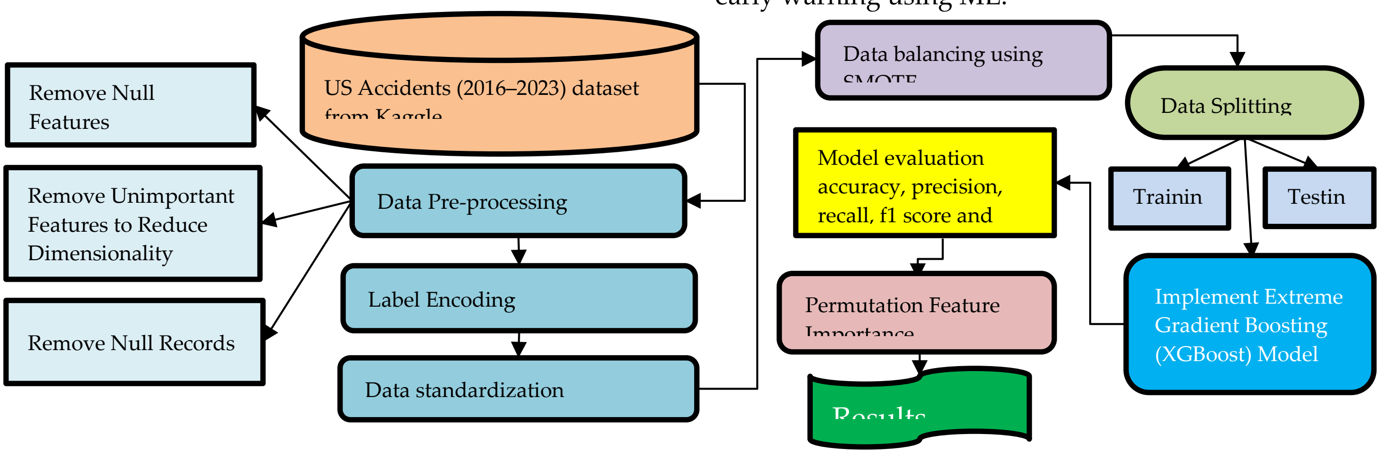

Fig. 1. Proposed flowchart for Predictive Traffic Accident Early Warning using machine learning

The following section provides a detailed explanation of each step in proposed methodology:

A. Data Gathering and Analysis

The US Accidents (2016–2023) dataset from Kaggle, which includes 2,845,342 traffic accident records with 46 characteristics, is used in this work for model training and assessment. The data was gathered from several sources, including APIs that compile reports from law enforcement, transportation agencies, and traffic cameras[23], and road sensors. Each record contains detailed information such as timestamps, geolocation, weather conditions, traffic-related attributes, surrounding points of interest, and accident severity. Class 1 (Low Severity), Class 2 (Moderate Severity), Class 3 (High Severity), and Class 4 (Very High Severity) are the four categories into which the accident severity is divided. Before developing a model, exploratory data analysis (EDA) and data visualization methods like bar graphs and heatmaps are used to look at feature distributions, find connections between variables, and comprehend the general features of the dataset.

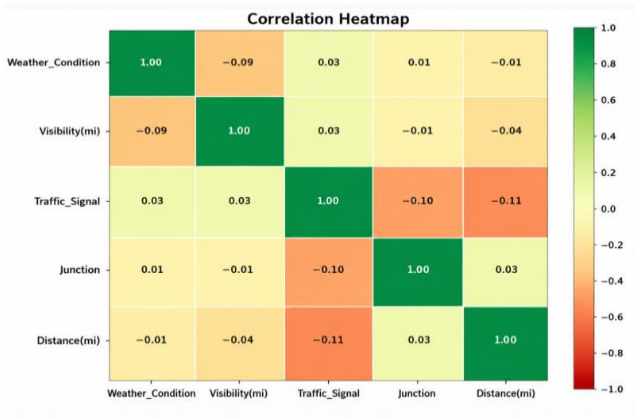

Fig. 2. **Correlation Matrix Heatmap**

The pairwise correlations between particular traffic accident variables are shown in Fig. 2's correlation matrix heatmap, where more favorable connections are shown by warmer hues, while negative correlations are indicated by cooler colors. The results demonstrate that most of the feature pairs have low correlation indicating low multicollinearity and thus each feature provides unique information for the predictive analysis.

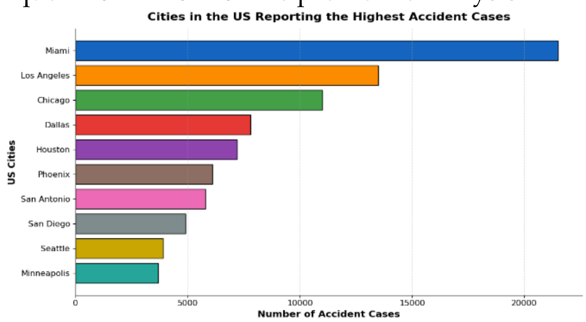

Fig. 3. **Bar Chart of Top U.S. Cities by Accident Cases**

In Fig 3 horizontal bar chart illustrates top U.S. cities with highest reported accident cases, where Miami records largest number of accidents, followed by Los Angeles and Orlando. The visualization shows the differences in the accident rate, in order to mark out the cities that are high risk for traffic safety measures and early warning system development.

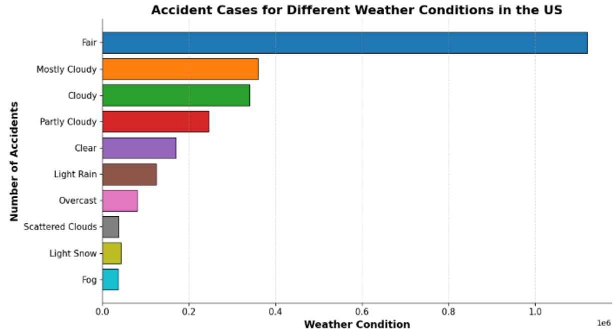

Fig. 4. **Bar Chart of Accident Cases for Different Weather Conditions in the US**

Fig. 4 shows weather conditions for traffic accident cases in the United States by bar chart. It shows that Fair weather records highest number of accidents, followed by Mostly Cloudy and Cloudy conditions, while Fog, Light Snow, and Scattered Clouds have the fewest reported accidents.

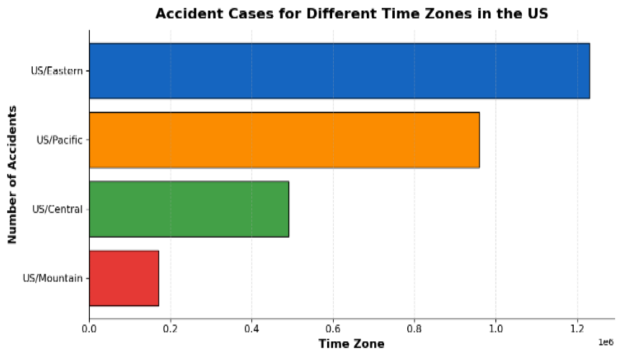

Fig. 5. Bar Chart of Accident Cases Across Different Time Zones in the US

In Fig 5, the bar chart illustrates distribution of traffic accident cases across major U.S. time zones. It indicates that the US/Eastern time zone records the highest number of accidents, followed by US/Pacific and US/Central, while the US/Mountain time zone reports the fewest accident cases.

B. Data Pre-processing

The US Accidents (2016–2023) Dataset [24]was used for data preparation, which involved concatenation, cleansing, and feature engineering. Missing values were handled, duplicates were removed, superfluous entries and noise were eliminated, and data labeling and normalization were applied as pre-processing procedures. The following is a summary of the primary preprocessing steps:

- Remove Null Features: A significant proportion of missing values (NULLs) may be present in some dataset attributes, rendering them unsuitable for efficient predictive modeling. Features that have more than 70% missing data are eliminated since their inclusion might cause noise and impair model performance.
- Remove Unimportant Features to Reduce

Dimensionality: High-dimensional datasets frequently include low-variance, redundant, or irrelevant characteristics that have little bearing on prediction performance.

- Remove Null Records: Once NULL characteristics have been removed, specific records (rows) with missing values are found. To preserve data integrity, a record is deleted if it contains crucial information that is absent from important features. This stage aids in avoiding errors in the ML model brought on by missing or inconsistent data points.
- Label Encoding: By allocating a distinct integer to every category, label encoding converts categorical variables into numerical labels. This enables ML algorithms to process categorical data efficiently.

C. Data standardization

The disparities in accident data feature scales need standardization of the data. By transforming feature values into range [0,1], min-max normalization can reduce impact of outliers and make data more consistent across datasets. Additionally, min-max normalization improves ACC and stability of several ML algorithms by scaling feature values to comparable ranges. This is how min–max normalization is as follows in Equation (1)

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X is original data value, X_{min} is the feature's minimum value, and X_{max} is the feature's highest value; X_{new} is the normalized data point, and its value normally ranges from 0 to 1.

D. Data balancing using SMOTE

Data balancing was performed to ensure an equal distribution of class labels, reducing prediction bias and improving model performance. This study employs SMOTE to establish a balanced class distribution and guarantee comparability with prior studies. The dataset is first split into training and testing sets, after which SMOTE is applied only to training data to balance the minority class while preserving an unbiased evaluation on the original test set. In particular, each minority-class instance's k-nearest neighbors are found, and the following formula is used to randomly pick a neighboring sample to a synthetic instance:

$$x_{new} = x + rand(0,1) * (x' - x) \quad (2)$$

where x is a randomly chosen sample from minor accident class, rand (0,1) is a randomly produced integer between 0 and 1, x′ is a randomly picked neighboring sample from minor accident class, and x_{new} is a freshly constructed synthetic sample.

E. Data Splitting

To guarantee a reliable and objective assessment of the ML models, for every stage of the experiment, the dataset is divided into 20% testing data and 80% training data. In order to improve model learning and lower model variation, 80:20 ratio was selected.

F. Proposed Extreme Gradient Boosting (XGBoost) Model

This work proposes a supervised ML–based boosting model, Extreme Gradient Boosting (XGBoost), for early predictive warning of traffic accidents. Decision trees are used by XGBoost, an ensemble learning system, to generate predictions. It may be used in regression scenarios by minimizing a loss function that determines the difference between the predicted and actual goal values. The XGBoost regression mathematical model is represented as (3):

$$y = f(x) \quad (3)$$

where $f(x)$is XGBoost model that forecasts, y based on x, y is expected property price, and x is a vector of input characteristics (such as square footage, number of bedrooms, etc.). XGBoost creates an ensemble of DTs trained to minimize mean squared error (MSE) loss function in order to compute $f(x)$. To provide a final forecast, the model integrates predictions from several DTs. The XGBoost regression model's general form is represented by (4):

$$y = \sum(k = 1 \; to \; K) \; fk(x) \quad (4)$$

where k is total number of DTs in ensemble and $fk(x)$is forecast of the kth DT. Each tree's forecast is weighted sum of its leaf values, which are discovered during training. Summing predictions of each DT in ensemble yields the XGBoost model's prediction for a given input x. The following hyperparameters were set up for suggested XGBoost model: Learning rate = 0.1, maximum tree depth = 6, number of estimators = 100, subsample = 0.8, colsample_bytree = 0.8, objective = binary, evaluation metric = logloss, random state = 42, and gamma = 0. These parameter values were selected to achieve stable learning, reduce overfitting, and improve predictive performance on the traffic accident dataset.

G. Evaluation metrics

The proposed design's performance is evaluated using a set of performance metrics derived from a confusion matrix that shows proportion of accurate and inaccurate predictions per class. This matrix was used to obtain key metrics such as True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These scores were then used to calculate important evaluation measures including ACC, PRE, REC, and F1:

Accuracy: It is defined as the ratio of properly predicted occurrences by the trained model to all dataset instances, as expressed in Equation (3)-

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

Precision: The percentage of all cases that the model projected to be positive that were accurately predicted to be positive. It is expressed in Equation (4)-

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall: This metric represents ratio of instances correctly predicted as positive to all TP instances. In mathematical form, it is expressed as Equation (5)-

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

F1 score: It provides a fair evaluation of both ACC and REC and is defined as the harmonic mean of both. Its value ranges from 0 to 1, and it is mathematically expressed as Equation (6)-

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

Receiver Operating Characteristic Curve (ROC): Plotting the percentage of instances correctly identified as positive against the percentage mistakenly labeled as positive for several decision cut-off points yields the ROC. While TPR is commonly referred to as sensitivity or REC, FPR is equal to 1-specificity.

4. Results and Discussion

This section describes experimental setup and presents performance evaluation of proposed model, highlighting its ACC and efficient processing. An ML Windows 11 with an Intel® Xeon® Gold 6226R CPU, 64 GB RAM, and a 24 GB NVIDIA RTX A5000 GPU was used to implement the suggested model. Python 3.9.0, PyTorch 1.8.0, and CUDA 11.1 were used in the development environment to train and assess the models.

A. Performance results of proposed models

The proposed model was trained and evaluated using the US Accidents (2016–2023) dataset utilizing key performance metrics as ACC, PRE, REC, and F1, as indicated in Table II. The XGBoost model showed exceptional predictive ability for early warning of traffic accidents, with an ACC of 95.9%. Additionally, it demonstrated a balanced classification capacity with PRE of 96.0%, REC of 95.7%, and an F1 of 95.8%. These results confirm the model's effectiveness in accurately identifying potential traffic accident risks while maintaining high overall performance.

TABLE II. CLASSIFICATION RESULTS OF PROPOSED MODEL FOR PREDICTIVE TRAFFIC ACCIDENT EARLY WARNING

Matrix	Extreme Gradient Boosting (XGBoost) Model
Accuracy	95.9
Precision	96
Recall	95.7
F1-score	95.8

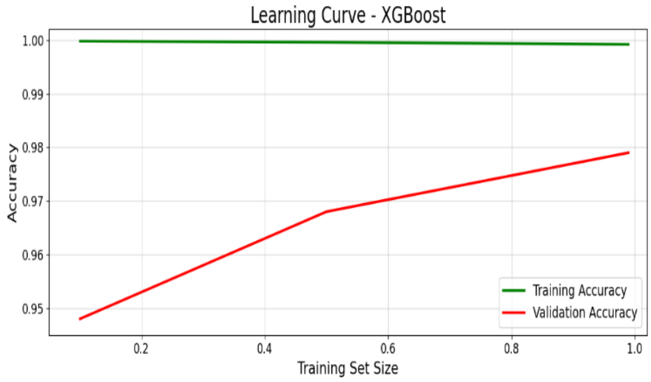

Fig. 6. Learning Curve for the XGBoost model

The XGBoost model's learning curve is depicted in Fig. 6, which demonstrates that as training set size grows, validation ACC continuously improves while training ACC stays consistently high. The robustness of the suggested model is demonstrated by narrow gap between two curves, indicating strong generalization with little overfitting.

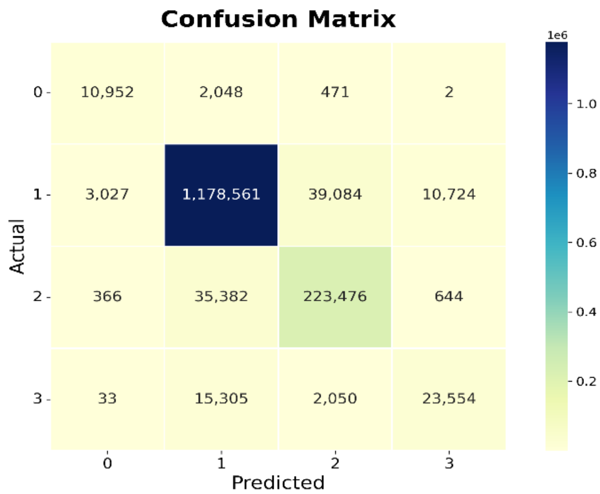

Fig. 7. Confusion Matrix for the XGBoost Model

Fig. 7 shows confusion matrix of the proposed model, with the actual class label shown in the column and the predicted class label in the row. The majority of samples are properly identified, as seen by the high diagonal values, whilst a few misclassifications are represented by the comparatively low off-diagonal values. The model's outstanding prediction capacity and balanced classification performance across all classes are demonstrated by the overall results.

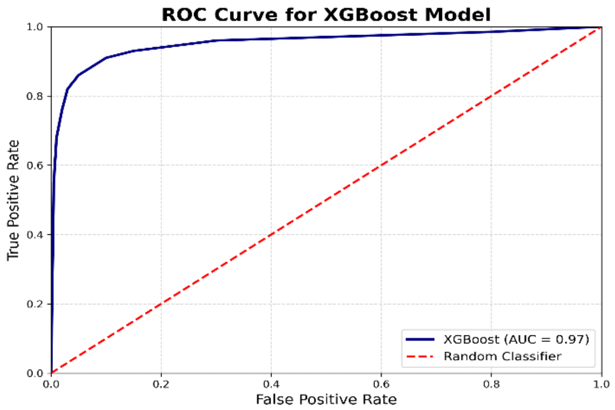

Fig. 8. ROC Curve for the XGBoost Model

The XGBoost model's classification performance is displayed by the ROC curve in Fig. 8, which is a trade-off curve between TPR and FPR at different threshold values. With an AUC of 0.97, the curve is around the upper-left corner and shows good discriminative ability and high prediction ACC. Overall, results confirm that model effectively distinguishes between target classes with minimal classification errors.

B. Feature Importance Analysis

The input features are ranked by feature importance analysis based on their contribution to prediction model, identifying most influential variables affecting its performance. Permutation importance was used to evaluate model's performance and confirm each feature's relevance after feature values were shuffled. By measuring the decrease in model performance brought on by shuffling, this strategy identifies crucial features since greater decreases indicate significant features. The most significant characteristics influencing the prediction model are Description, Source, and City, according to permutation importance analysis to determine the global feature importance.

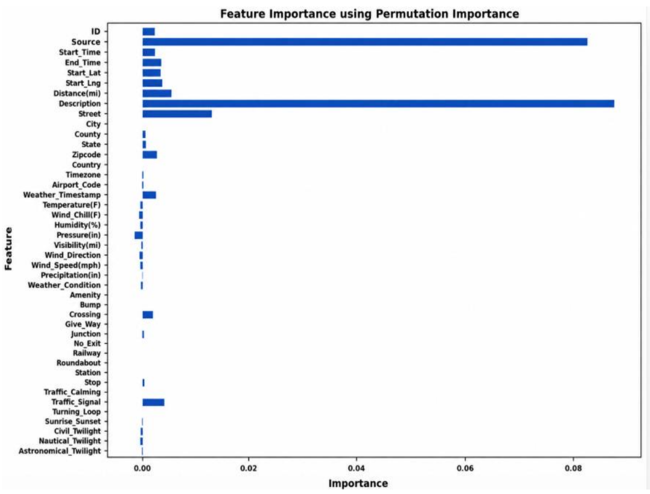

Fig. 9. Global Feature Importance Using Permutation Importance Analysis

Fig. 9 describes global feature importance obtained using permutation importance analysis, emphasizing each feature's role in prediction model. The findings show that, with greatest relevance ratings, Description, Source, and City are most crucial features, whereas other variables have a comparatively less impact on prediction performance of model. This analysis helps identify key factors that significantly impact the overall prediction ACC.

C. Comparative Analysis

In order to evaluate effectiveness of proposed model, comparative performance evaluation with the existing ML models is given in Table III. The proposed XGBoost model achieved highest ACC of 95.9%, outperforming Federated Learning (71.15%), DT (86.66%), RF (84%), and SVM (80%). It also achieved highest PRE (96.0%), REC (95.7%), and F1 (95.8%), showing good classification performance. These findings show that compared to other methods, suggested XGBoost model offers a more accurate and dependable solution for early predictive warning of traffic accidents.

TABLE III. COMPARISON OF DIFFERENT MACHINE LEARNING MODELS FOR PREDICTIVE TRAFFIC ACCIDENT EARLY WARNING

Model	Accuracy	Precision	Recall	F1-score
FL[25]	71.15	-	-	-
DT[26]	86.66	57.36	57.37	57.37
RF[27]	84	77	84	79
SVM[28]	80	80	77	78
Proposed XGBoost	95.9	96	95.7	95.8

The proposed XGBoost model achieved an ACC of 95.9%, outperforming existing models considered in this study. Its high ACC demonstrates model's effectiveness in identifying traffic accident risks with greater reliability. The results indicate that XGBoost provides improved predictive performance and is well-suited for traffic accident early-warning applications.

5. Conclusion and Future Study

Traffic safety research is becoming more crucial as traffic safety issues worsen and the risk of car accidents to individual safety increases. At same time, rapid development of Vehicular Ad Hoc Networks (VANETs) provides valuable real-time road information to support accurate traffic accident prediction and enable effective early warning systems. In this paper, an Extreme Gradient Boosting (XGBoost)-based predictive model for early warning of traffic accidents was proposed using the US Accidents (2016–2023) dataset. Based on experimental results, proposed XGBoost model achieved an ACC of 95.9%, outperforming existing ML models, including FL (71.15%), DT (86.66%), RF (84%), and SVM (80%). These findings demonstrate that proposed model provides a reliable and effective solution for traffic accident prediction and can support intelligent transportation systems by enabling timely accident risk assessment and improved road safety.

The limitation of this study is that proposed model is developed from data from the US Accidents (2016-2023) and has not been used with real-time traffic or real-time vehicle sensor data. Future research will include integrating real-time IoT and connected-vehicle data, studying hybrid models based on ML and DL, and validating the framework with data from diverse geographic areas to enhance its robustness and generalizability.

REFERENCES

1. V. K. Kukkala, J. Tunnell, S. Pasricha, and T. Bradley, “Advanced Driver-Assistance Systems: A Path Toward Autonomous Vehicles,” IEEE Consum. Electron. Mag., vol. 7, no. 5, pp. 18–25, Sep. 2018, doi: 10.1109/MCE.2018.2828440.
2. N. Adik, “The Rise of Open and Free Networks: A Community-Driven Paradigm for Decentralized Connectivity,” in 2025 IEEE First International Conference on Innovations in Engineering and Next-Generation Technologies for Sustainability (ICINVENTS), Coimbatore, India: IEEE, 2025, pp. 1–9, November. doi: 10.1109/ICINVENTS64613.2025.11402224.
3. J. Nidamanuri, C. Nibhanupudi, R. Assfalg, and H. Venkataraman, “A Progressive Review: Emerging Technologies for ADAS Driven Solutions,” IEEE Trans. Intell. Veh., vol. 7, no. 2, pp. 326–341, Jun. 2022, doi: 10.1109/TIV.2021.3122898.
4. P. H. Desai, P. Mahalle, and P. Chandre, “Intelligent Access Control Schemes for the Internet of Everything: A Survey of Techniques, Challenges, and Future Directions,” in Information Systems for Intelligent Systems, 2026, pp. 95–105. doi: 10.1007/978-3-032-13196-6_9.
5. S. Singamsetty, “AI-powered Satellite Image Processing for Global Air Traffic Surveillance Techniques Using NCNN–EGSA Optimization Techniques,” in Machine Learning Based Air Traffic Surveillance System Using Image Processing, Emerald Publishing Limited, 2026, pp. 179–197. doi: 10.1108/978-1-80592-062-520251010.
6. L. A. Yeruva, D. Singh, S. Suddala, N. Bhatt, and R. Uddin, “Augmented Data Management for Cache Performance, Cybersecurity, and Mobile Integration,” J. Comput. Mech. Manag., vol. 5, no. 3, pp. 280–293, May, Jun. 2026, doi: 10.57159/jcmm.5.3.26691.
7. S. Chatterjee, “A Data Governance Framework for Big Data Pipelines: Integrating Privacy, Security, and Quality in Multitenant Cloud Environments,” Tech. Int. J. Eng. Res., vol. 10, no. 5, 2023, doi: 10.56975/tijer.v10i5.158181.
8. T. Banerjee and H. Singh, “Securing Non-Human Identities in Industrial IoT a Blockchain-Based Trust Framework,” NIPES - J. Sci. Technol. Res., vol. 7, no. 4, pp. 228–246, Dec. 2025, doi: 10.37933/nipes/7.4.2025.1660.
9. M. Mittal, “The Role of Edge Computing in IOT: Enhancing Real-Time Data Processing Capabilities,” Int. J. Adv. Res. Electr. Electron. Instrum. Eng., vol. 06, no. 12, Dec. 2017, doi: 10.15662/IJAREEIE.2017.0612002.
10. S. Cheng, B.-B. Hu, H.-L. Wei, L. Li, and C. Lv, “Deep Learning-Based Hybrid Dynamic Modeling and Improved Handling Stability Assessment for Autonomous Vehicles at Driving Limits,” IEEE Trans. Veh. Technol., vol. 74, no. 4, pp. 5582–5593, Apr. 2025, doi: 10.1109/TVT.2024.3515209.
11. H. K. Yadav, “Design and Implement Machine Learning Based Predictive Analytics for Road Safety and Accident Prevention,” in 2026 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India: IEEE, 2026, pp. 1–6, March. doi: 10.1109/IATMSI68868.2026.11465965.
12. J. B. Mehta, “Autonomous Security Validation For Embedded Trust Chains In Connected Devices,” Int. J. Adv. SIGNAL IMAGE Sci., vol. 11, no. 6s, pp. 802–812, Dec. 2025, doi: 10.29284/1th3da51.
13. S. P. Ardakani et al., “Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis,” Sustainability, vol. 15, no. 7, Mar. 2023, doi: 10.3390/su15075939.
14. H. N. Dholariya, “Human-in-the-Loop AI for Cloud Data Engineering: The Collaborative Intelligence Architecture (CIRA) for Regulated Industries,” J. Inf. Syst. Eng. Manag., vol. 11, no. 1, pp. 883–898, Jan, 2026.
15. Y. Kotsyubynska, N. Kozan, V. Chadiuk, A. Kostyshyn, A. Kotsyubynsky, and V. Fentsyk, “Machine Learning and Deep Learning for Predicting Traffic Crash Injury Severity: A Systematic Review and Meta-Analysis (2014-2025),” J. Road Saf., vol. 1, no. 37, Feb. 2026, doi: 10.33492/JRS-D-26-1-2721386.
16. S. Jabar and M. Hussain, “Machine Learning-Based Framework for Road Accident Detection and Prevention,” Nov. 2025. doi: 10.36227/techrxiv.176366388.82802425/v1.

17. P. Saranya, G. R, N. A, and K. G, “Road Accident Prevention System using Machine Learning,” Int. J. Innov. Sci. Res. Technol., pp. 3616–3623, Jun. 2025, doi: 10.38124/ijisrt/25 May 1803.

18. Y. Berhanu, D. Schröder, B. T. Wodajo, and E. Alemayehu, “Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks,” Results Eng., vol. 23, p. 102737, Sep. 2024, doi: 10.1016/j.rineng.2024.102737.

19. R. Alnashwan, M. Mashaabi, A. Alotaibi, H. Qudaih, and L. Albraheem, “IoT-Based Accident Prevention System using Machine Learning techniques,” in 2023 6th Artificial Intelligence and Cloud Computing Conference (AICCC), New York, NY, USA: ACM, Dec. 2023, pp. 179–188. doi: 10.1145/3639592.3639617.

20. A. T. Kurian and P. Kumar Soori, “AI-Based Driver Drowsiness and Distraction Detection in Real-Time,” in 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, Mar. 2023, pp. 13–18. doi: 10.1109/ICCIKE58312.2023.10131730.

21. C. Yang, “Car accidents prediction based on cloud computing in data analysis,” in ISCTT 2022; 7th International Conference on Information Science, Computer Technology and Transportation, 2022, pp. 1–6.

22. K. K. Mohd Shariff, S. Zainuddin, and M. S. Amin Megat Ali, “Detection of Wet Road Surfaces from Acoustic Signals using Scalogram and Optimized AlexNet,” in 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE, May 2022, pp. 159–163. doi: 10.1109/ISCAIE54458.2022.9794556.

23. Y.-F. Zhou, K. Xie, X.-Y. Zhang, C. Wen, and J.-B. He, “Efficient Traffic Accident Warning Based on Unsupervised Prediction Framework,” IEEE Access, vol. 9, pp. 69100–69113, 2021, doi: 10.1109/ACCESS.2021.3077120.

24. S. Moosavi, “US Accidents (2016 - 2023),” Kaggle Contributor.

25. A. Ghaffari, H. Nguyen, A. Saleh, L. Lovén, and E. Gilman, “Traffic Accident Prediction and Warning System: Integration Use Case,” CEUR Workshop Proc., vol. 3894, pp. 108–118, 2024.

26. T. Hu and T. Hu, Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025), vol. 128, no. Iwadic 2025. In Advances in Computer Science Research, vol. 128. Dordrecht: Atlantis Press International BV, 2026. doi: 10.2991/978-94-6239-648-7.

27. M. E. Jaji, “Predictive Analytics of Road Traffic Incidents , A Machine Learning Approach Predictive Analytics of Road Traffic Incidents , A Machine Learning Approach by A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of,” 2024.

28. Z. Li, “Predicting Traffic Accident Severity During Peak Hours Using Machine Learning,” in Proceedings of the 9th International Conference on Electronic Information Technology and Computer Engineering, New York, NY, USA: ACM, Jun. 2025, pp. 375–380. doi: 10.1145/3766671.3766738.