IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, July 2025
Impact Factor: 7.67

A Machine Learning Framework for Vehicle Road Safety Risk Prediction and Identification

Dr. Puneet Garg
Associate Professor, Department of CSE-AI
KIET Group of Institutions, Delhi NCR, Ghaziabad
puneet.garg@kiet.edu

Abstract: *In an era of rapidly evolving transportation networks, the complexity of ensuring road safety demands more than traditional reactive measures. This study introduces a data-driven machine learning framework designed to proactively predict vehicle accident risks using structured traffic and environmental datasets. By integrating advanced preprocessing techniques—including feature engineering, one-hot encoding, SMOTE-based balancing, and normalization—the system prepares data for high-accuracy modeling. Four machine learning architectures were explored, with the Random Forest model demonstrating the most effective performance, achieving 96% accuracy, precision, recall, and F1-score. These consistent results across all metrics affirm its capability to handle imbalanced, multi-class accident severity prediction with reliability. The model draws on a range of features such as lighting conditions, vehicle types, road class, and time of day to identify risk-prone scenarios. Its interpretable design and balanced performance make it a practical asset for policymakers and traffic planners aiming to implement real-time, preventive strategies. The research validates the value of AI-powered safety frameworks in transforming reactive safety protocols into predictive, automated systems capable of safeguarding modern transport ecosystems.*

Keywords: Road Safety Prediction, Random Forest, SMOTE Balancing, Intelligent Transportation Systems.

I. INTRODUCTION

Worldwide, road accidents are one of the top causes of death and serious injury, and the problem is of enormous proportions. Accidents involving motor vehicles kill about 1.35 million people annually and injure another 20–50 million; many of these victims live with disabilities for the rest of their lives, putting a heavy strain on local communities, state and federal healthcare systems, and families [1]. There is a constant flow of incidents in both industrialized and developing nations, proving that road transit is both the most complicated and riskiest mode of transportation.

Road safety is not a by-product of chance but of concerted, systemic attempts by many parts of society. The development of consistent, long-term policies and initiatives to promote and sustain road safety is an essential public benefit that must be recognized and addressed by both governmental and non-governmental organizations [2][3]. Effective road safety management requires collaboration across various disciplines, including transport engineering, urban planning, law enforcement, public health, education, and vehicle manufacturing.

Accident typically happens because of a multifactorial mix of person and environs, and automobile disorders. Pedestrians and other road users would not be aware of traffic rules and regulations and do not observe them hence would lead to dangerous situations [4]. Alternatively, technical failures due to poor maintenance of vehicles are also a major threat to road safety e.g. brakes giving off, tires worn out, lights out etc. These risks are worsened by unfavourable weather conditions, inefficient road network infrastructure and low levels of traffic control solutions [5][6]. Collectively, this gives the multi-dimensional and dynamic net of risks and hazards that conventional safety management approaches tend to be reactive and fragmented cannot tackle appropriately and on a proactive basis.

The techniques available today are based on the drivers of historical accident data, occasional manual surveys, or using

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, July 2025 Impact Factor: 7.67

the fixed risk models that cannot be adaption in real time with already fast changing road and traffic conditions. The issue of road accidents can only be solved in a multidimensional solution in which the use of sophisticated technological answers be employed [7]. Over the past years, accident detection and prevention have become possible with machine learning and data analytics. Machine learning algorithms can learn about past patterns and tendencies related to crashing, by analysing the history of traffic data [8]. Through analysing other factors like traffic amount, traffic jams, weather conditions as well as time differences, predictive models can be used to foresee probable accident hotspots as well as risky areas [9][10]. Using sensors and analysis of velocities in real time through the use of GPS data makes it possible to identify the speeding vehicles and signs of aggressive driving [11][12]. Machine learning algorithms provide surveillance of vehicle speed, acceleration and trajectory to warn anything anomalous that could have resulted in an accident or unsafe driving behaviour. The video capture on the cameras installed on roads is in real time, and goes through computer vision and image processing [13][14]. The computer vision algorithms can be used to help machine learning to identify objects, spot cars, and record patterns related to accidents, such as sudden lane switching, accidents, or people crossing the road [15][16][17]. The implementation of data-driven strategies, machine learning and prevention of accidents in road safety have a huge potential.

A. Motivation and Contribution of the Study

The motivation behind this study stems from the urgent need to enhance road safety through intelligent and data-driven prediction systems. Traditional approaches to accident risk analysis often rely on manual interpretation or isolated statistical methods, which may not fully capture the complex, nonlinear patterns underlying road incidents. With the increasing availability of large-scale accident datasets and advancements in machine learning, there is a strong opportunity to build predictive frameworks that can proactively identify risk factors and prevent accidents. This study proposes a machine learning-based framework that leverages structured accident data to predict and identify potential road safety hazards. The key contributions of this research are as follows:

- Using a traffic accident dataset hosted by Kaggle to train and evaluate models thoroughly.
- Systematic data pre-processing involving cleaning, feature engineering, encoding, SMOTE balancing, and scaling.
- Creation and evaluation of several prediction models, such as XGBoost, Random Forest, Neural Network, and LSTM.
- Evaluation of performance utilizing commonly used classification measures including F1-score, Accuracy, Precision, and Recall.
- Proving that deep learning and machine learning can help with road safety planning and reducing risks.

B. Justification and Novelty of the paper

The increasing need for data-driven road safety systems that can foresee and intervene in real-time in the event of a risk justifies this study. Due to the complex and ever-changing nature of road accidents, traditional analytical methods are inadequate. This study introduces a novel integration of SMOTE-enhanced pre-processing with a Random Forest-based prediction system, tailored for structured accident datasets. The framework's strength lies in its practical feature engineering and deployment feasibility, making it suitable for smart traffic environments. Its novelty stems from combining systematic pre-processing with multi-model evaluation to deliver a scalable, interpretable, and accurate solution for identifying road safety hazards.

C. Organization of the paper

This is the structure of the paper: In Section II, take a look at previous research that pertains to road safety prediction. The proposed methodology is explained in Section III. Assessments of the models and outcomes of the experiments are detailed in Section IV. In Section V, review the important points and provide some suggestions for where the field could go from here in terms of further research into improving ML -based predictive road safety systems.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, July 2025
Impact Factor: 7.67

II. LITERATURE REVIEW

This section reviews recent advances in ML frameworks for vehicle road safety risk prediction and identification, emphasizing techniques that analyze diverse traffic, vehicle, and environmental factors. These approaches integrate real-time data processing with predictive modelling to improve the accuracy of accident risk assessment, enhance the interpretability of key contributing factors. Key studies reviewed in this context include:

Kumar et al. (2025) proposed a ML-based predictive framework for road accidents, classifying severity into Low, Medium, and High. The dataset includes driver and environmental factors like age, gender, road surface condition, and light conditions. Various ML models are compared, with most models having an 84% accuracy. The system can provide perspectives on accident prevention measures and policy formulation [18].

Noonpakdee et al. (2025) examined how Explainable AI (XAI) methods, specifically SHAP (Shapley Additive Explanations) could be used to determine which factors contribute to the severity of road accidents. ML models were created based on the data of the Thailand-based Ministry of Transport. In the obtained results, it is shown that the Light Gradient Boosting Machine (LGBM) model demonstrated the highest accuracy of 0.85 and F1 score of 0.83. According to SHAP analysis, number of motor-cycles involved, road code and the number of vehicles and the number of people involved in the accident were found to be significant contributors [19].

Pei, Wen and Pan (2024) proposed a new predictive and analysis model of identifying the main risk factors in road accidents. The model CNN-Belts-Attention is utilized to predict the risk of an accident, whereas aged Deep SHAP resorts to interpreting the model to derive important factors. The model has a lower Mean Absolute Error (MAE) on the UK and the US dataset of 0.2475 and 0.2683, respectively. The findings of Deep SHAP are more reasonable and enlightening to determine factors of varying severities [20].

Wei et al. (2024) assessed the performance of six ML algorithms on a dataset, with consideration on accident injury severity. Light algorithm is selected as the most efficient one, having the accuracy of 92.6%, and 92.8% F1-score. The SHAP method also forms part of the study in examining the importance, main effect, and interaction effect of factors in each accident injury severity category. The three factors which are of importance are the kilometres travelled by the driver per year, the distance he throws, and the speed limit of the road. The risks of the motorcyclists getting fatal injuries are high in instances of throwing distance that exceeds 1,000cm [21].

S N (2023) studied the analysis of parameter importance on Extra Trees Regressor and XGBoost to determine the prediction of the risk of accidents. In the research, 2016-2023 Kaggle data is applied to establish web-based prediction methodology. The study determines factors that determine accident risk and this shows that the prediction accuracy is affected by 85%. The results give a layered picture of the issue and allow pointing out possible directions of the future research in the field. The paper ends with the foresight view of what the findings mean to road safety and prevention of accidents [22].

Obasi and Benson, (2023), modelled using ML techniques such as Naive Bayes, Random Forest, Logistic Regression and Artificial Neural Networks in predicting the severity of traffic accidents. They examined the data of the UK traffic accidents in 2005-2014 and found the Random Forest and Logistic Regression models have 87 percent accuracy which was better than the Naive Bayes and Artificial Neural Networks [23].

Table I summarizes prior studies on ML -based frameworks for vehicle road safety risk prediction and key factor identification, highlighting methodologies, datasets, key findings, limitations, and proposed future directions

Table 1: Comparative Summary of Related Works on Machine Learning Frameworks for Vehicle Road Safety Risk Prediction

Author	Methodology	Data	Key Findings	Limitation	Future Work
Kumar et al. (2025)	ML models (RF, GB, LR, DT, SVM); severity classification into Low, Medium, High	Road accident dataset	Achieved ~84% accuracy; multiple models compared; user interface for real-time prediction	Lacks explainability; does not deeply analyze key factor impacts	Could integrate explainable AI (XAI); expand to larger/real-time datasets

Noonpakdee et al. (2025)	LGBM + SHAP for explainability; campaign design for behavior change	Thailand's Ministry of Transport accident data	LGBM highest accuracy (85%); SHAP found top features: motorcycle involvement, road code, total vehicles/people	Limited to a single country context; no advanced deep learning	Expand social marketing; adapt model to wider regions and real-time monitoring
Pei, Wen and Pan, (2024)	CNN-BiLSTM-Attention + DeepSHAP; spatiotemporal modeling	UK & US real-world datasets	Achieved MAE: 0.2475 (UK) & 0.2683 (US); identified key risk factors via DeepSHAP; verified factor rationality	High model complexity; hard to deploy in resource-constrained settings	Improve spatiotemporal data quality; deploy in real-time systems
Wei et al. (2024)	LightGBM + SHAP; factor interaction effect analysis	China In-depth Accident Study (CIDAS) data on vehicle-motorcycle crashes	Accuracy: 92.6%; SHAP identified top factors: km/year, throwing distance, speed limit; clear injury severity relations	Specific to motorcyclists; limited generalizability to all vehicle types	Extend to other vulnerable road users; test in mixed traffic conditions
S N (2023)	Extra Trees Regressor & XGBoost; parameter importance; web-based system	US accidents (2016–2023, Kaggle)	85% accuracy; key determinants highlighted; complete system architecture outlined	Lacks robust interpretability methods (like SHAP); not tested for real-time use	Incorporate XAI; expand to more dynamic traffic inputs
Obasi and Benson, (2023)	Naive Bayes, RF, LR, ANN; feature importance with RF	UK traffic data (2005–2014, ~2 million records)	RF & LR achieved 87% accuracy; identified sensitive factors: engine capacity, age of vehicle/driver, road class, maneuver	Limited model interpretability; older dataset	Use recent data; integrate explainable models; include real-time risk prediction

III. METHODOLOGY

A structured pipeline is used to gather data from Kaggle, using a comprehensive road accident dataset, as part of the suggested technique for road safety risk prediction. Data pre-processing is the subsequent step that encompasses necessary operations like data cleansing, feature engineering, categorical variable encoding, dataset balance via SMOTE, and feature scaling. An effective way to create and evaluate models is to divide the dataset into training and testing subsets after pre-processing. In order to assess trends and forecast danger levels linked to vehicle accidents, a number of ML and DL models are utilised, including LSTM, XGBoost, Neural Network, and RF. Accuracy, precision, recall, and F1-score are some of the most important performance metrics used to evaluate models in the last stage. As shown in Figure 1, this comprehensive framework guarantees the systematic processing and modelling of data related to road accidents. This, in turn, allows for predictions that are both reliable and easy to understand, which can aid in making informed decisions regarding road safety management.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, July 2025
Impact Factor: 7.67

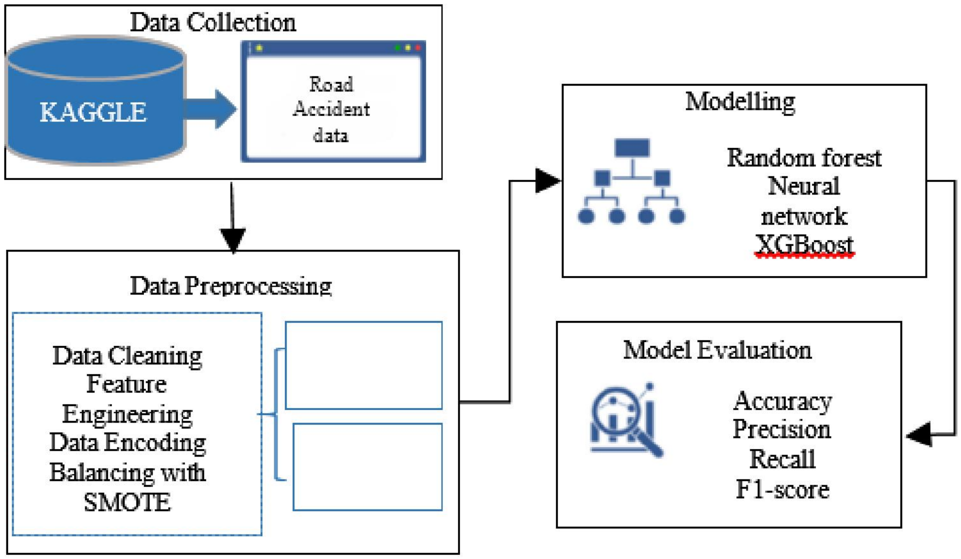

Fig. 1. Proposed Machine Learning Framework for Road Safety Risk Prediction

A. Dataset Collection

This dataset, sourced from Kaggle, contains comprehensive information on individual road accidents, including spatial details (Easting, Northing), number of vehicles involved, date and time of occurrence, road classification, and surface conditions. Additionally, it records incident-specific environmental factors like lighting and weather. Casualty-related information—such as class, severity, sex, age, and vehicle type—is also included. The visualizations of the data are given below:

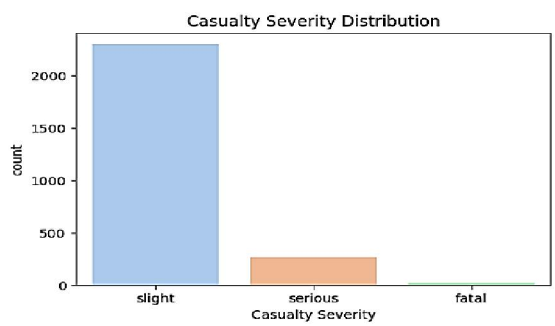

Fig. 2. Casualty Severity Distribution Bar Chart

Figure 2 provides a visual representation of the frequency of different levels of casualty severity. The most prominent category is "slight" casualties, which vastly outnumber the others with a count well over 2000, indicating that minor injuries are the most common outcome. Following far behind are "serious" casualties, which occur in a much smaller proportion, with their count approximately around 250. The least frequent category is "fatal" casualties, which register a remarkably low count, barely visible above the x-axis, suggesting that fatalities are a very rare occurrence compared to slight or serious injuries.

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 2, July 2025

Impact Factor: 7.67

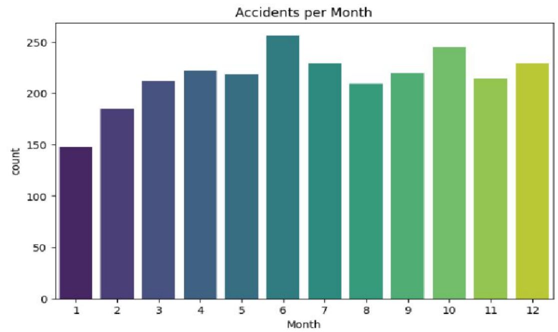

Fig. 3. Accidents Per Month Within the Data

Figure 3, provides a comprehensive visualization of accident frequency throughout the year. The initial months, January and February, record a relatively lower number of accidents, with counts hovering between approximately 150 and 180. A noticeable upward trend begins around March, leading to a significant peak in June, which registers the highest number of accidents, exceeding 250. Following this mid-year surge, there's a slight dip in July and August, but accident numbers remain generally high. The latter part of the year, particularly October, also shows a substantial number of accidents, reaching close to the June peak, suggesting that the warmer months and early autumn might be periods of increased accident occurrences.

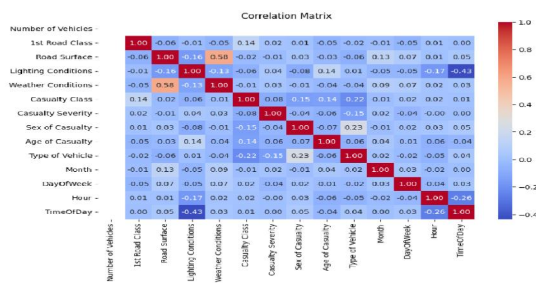

Fig. 4. Heatmap of the Correlation Matrix

Figure 4 shows the heat-mapped correlation matrix, which illustrates the linear correlations between the variables "Number of Vehicles" to "Time Of Day." On a scale from -1 (very negative, blue) to +1 (very positive, red), the numerical value and colour of each cell represent the intensity and direction of the link. There are other pairs with weak correlations close to zero, but one stands out: "Lighting Conditions" and "Time Of Day" (-0.43). This suggests, intuitively, that later in the day, there is worse lighting.

B. Data Preprocessing

Data pre-processing is the process of transforming raw data into a more reliable and high-quality format for analytical and predictive modelling. Following these steps guarantee that your data is ready to show you the relationships and patterns you're looking for. Data preparation is outlined in the following steps:

C. Data Cleaning

Every ML or data analysis effort must begin with data cleaning. Preparing the data for analysis involves making sure it is correct and complete. Results from predictive models are improved and data insights are more reliable when data is cleaned properly:

- Data entry, transfer, or collecting mistakes might cause missing values. One way to deal with missing data is to remove entire rows or columns if they include an excessive number of blanks. Another is to use statistical methods to fill in the blanks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, July 2025 Impact Factor: 7.67

- Duplicate entries can skew analysis results because they give repeated information too much weight. Locating duplicates usually entails looking for rows with the same data or columns with the same values in important identifiers. merge or eliminate duplicates to keep data pure.

D. Feature Engineering

The goal of the feature engineering method was to enhance the performance of the model by transforming the raw data into more easily understood categories. Road conditions were grouped as Safe or Adverse, weather as Normal or Adverse, and lighting as Well-lit or Poorly-lit. Road types were consolidated into High-Speed, Primary Roads, and Secondary Roads. Vehicle types were grouped into High-Risk, Passenger, and Heavy Vehicles. Age was binned into Young, Adult, Middle-aged, and Elderly.

E. Encoding Categorical Variables

In this method, the pandas library's pd.get_dummies() function is used to encode category variables in the dataset using one-hot encoding upon import. The mentioned categorical variables, which include Season, Time Of Day, Lighting Category and so on, are converted into binary columns, denoting each category. With drop first set at True, the initial category in each feature will be removed to make sure not to fall in the dummy variable trap, which may prevent multicollinearity in regression models. This is the process that translates categorical data into numeric format which can be used in a ML algorithm so that the ML model can use and understand it to be more accurate in its predictions.

F. Handling Class Imbalance (SMOTE)

to solve the problem of imbalance of the class in the dataset, Synthetic Minority Over-Sampling Technique (SMOTE) was used. In SMOTE, samples of the minority class are synthesized by relying on existing samples used and interpolating between them, in effect augmenting the representation of under-represented categories. The method does not simply duplicate data so there is less of a tendency towards overfitting. By increasing the representation of the minority class, the model acquires a capability to learn more pattern and subtleties so that the ability to differentiate between classes increases as well.

G. Feature Scaling

Several or most ML algorithms benefit from the numeric input variables that are rescaled to be within a typical range. Standard Scaler is an important step of data pre-processing in ML models [24]. It makes sure that numerical characteristics are converted to a normal scale, with a mean of zero and standard deviation of Equation (1).

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

where, x is the original value, μ is the mean value, and σ is the standard deviation.

H. Data Splitting

The dataset was divided into a training set (80%) and a testing set (20%) to build the model and assess its performance, respectively. This split helps provide a reliable estimate of how well the model perform on new, unseen data.”

I. Proposed Random Forest Model

Data mining tools like RF can fix issues with regression and classification. The accuracy of classification has been greatly enhanced by growing an ensemble of trees and determining the class type by voting. In order to train these ensembles, random vectors are generated. Every tree is created using a different vector at random. Classification and regression trees make up RF [25]. Examining the results of trees allows us to resolve classification issues. In order to determine the RF prediction, the majority of the class votes. With the addition of more trees to RF, the generalization error approaches a limiting amount; this is because over-fitting is not a problem in big RFs [26]. An additional feature ensures that the many decision-making pathways from the node to the last leaf remain uncompromised. Each feature and bias (the average value of the uppermost region covered by the training set) contribute to the final prediction. The

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 5, Issue 2, July 2025 Impact Factor: 7.67

prediction function for decision trees is described by Equation (2)

$$f(x) = Cfull + \sum_{m=1}^{M} contrib(x,k) \quad (2)$$

where k is the number of characteristics and M is the number of leaves in the tree. The root node value is C full. Take the k-th feature from feature vector x and return it as Contribute (x, k). Here is the definition of the RF prediction function in Equation (3):

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} f_j(x) \quad (3)$$

The average of the bias and the average of each contributing feature set is what makes up the RF prediction, as is evident. Equation (4) provides a definition for:

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} C_j full + \sum_{k=1}^{k}(\frac{1}{J}\sum_{j=1}^{J} + contrib_j(x,k) \quad (4)$$

A RF black box processes data by calculating the contributions of individual features and pursuing decision routes via a tree, as described in the previous expression.

J. Performance Assessment Metrics

ML classifiers' efficacy in predicting risks to road safety in vehicles was assessed using important metrics such F1-Score, Accuracy, Precision, and the Confusion Matrix [27]. The confusion matrix shows which risk categories were right and which were wrong by comparing the two sets of data. To further understand the model's generalizability and classification performance, ROC curves were used to display its capacity to differentiate between safe and unsafe driving behaviors. To better evaluate the model's prediction performance, the confusion matrix summarises four potential outcomes, as illustrated in Figure 5. In order to calculate a variety of performance measures, such as:

- **True Positive (TP):** A beneficial outcome, as anticipated by the model, came to pass.
- **True Negative (TN):** A bad event, as anticipated by the model, came to pass.
- **False Positive (FP):** The model incorrectly predicted a positive outcome, i.e, the actual outcome was negative. It is also known as a Type I error.
- **False Negative (FN):** The actual result was positive, whereas the model had projected a negative one. Another name for it is a Type II mistake.

Accuracy

The accuracy rate is defined as the percentage of forecasts that were correct, including both positive and negative outcomes. It gives you a feel for how well the model is doing overall, but it could be deceiving in datasets where one class is much more prominent than the others [28]. Equation (5) is used to determine the precision.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (5)$$

Precision

The evaluation takes into account the success rate of the projected positive cases. Its emphasis on reducing false positives makes it indispensable in spam detection and other situations where wrongly classifying negatives can have a significant financial impact. When the model accurately predicts a favourable outcome, it is considered dependable. Equation 6 shows the result of the computation.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall (sensitivity)

The accuracy rate of the model in detecting real positive cases is quantified by this metric. In settings like disease screening, where failing to detect a real positive could have devastating effects, the emphasis on reducing false negatives is crucial. A high recall rate shows that, as shown in Equation (7), the model successfully captures the majority of positive cases:

$$Recall(Rc) = \frac{TP}{TP+FN} \quad (7)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, July 2025 Impact Factor: 7.67

F1 score

It uses the harmonic mean of precision and recall to bring them into harmony with one another. When there is an uneven distribution of classes and a need to reduce the number of false negatives and positives, it is used most often. When making a trade-off between recall and precision, as in Equation (8), an F1 value of 0.8 or higher does indicate satisfactory performance.

$$F1\ score(F1) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (8)$$

These benchmarks of performance were used to evaluate the reliability of the model on the test set where the model was asked to make its predictions on the test set.

IV. RESULT ANALYSIS AND DISCUSSION

In this paper, the authors assess the capability of a ML structure using RF to forecast and detect car road hazards. The research was conducted on a Python programming language with the usage of the Scikit-learn library, the experiments took place on a traffic safety analytics workstation with a 64-bit version of the Windows 11 operating system, powered by an AMD Risen 9 processor, 64 GB RAM, and 64 GB RAM, and high-performance SSDs were used to process massive datasets involving sensors and traffic. Due to its high accuracy (96%) and all the other parameters precision, recall, F1-score being the same (96%), the RF model was found to be very reliable and balanced in identifying a risky driving behavior (Table II). The findings confirm the effectiveness of the model and its usefulness to road safety risk assessment in real time, since it is able to examine complex sensor signals and environmental conditions.

Table 2: Evaluation Metrics Of Random Forest Model For Road Safety Risk Prediction

Performance Metrics	Random Forest Model
Accuracy	96
Precision	96
Recall	96
F1-score	96

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       459
           1       0.94      0.94      0.94       459
           2       0.94      0.93      0.94       460

    accuracy                           0.96      1378
   macro avg       0.96      0.96      0.96      1378
weighted avg       0.96      0.96      0.96      1378
```

Fig. 5. Random Forest Model Road Safety Risk Classification Report

The Figure 5 shows a typical result of a classification report in ML. It provides the performance indicators of a multi-classification model (three classes 0, 1, 2). Precision, recall, f1-score and support are present per class. The report also includes overall accuracy, macro average, and weighted average, all showing a high performance of 0.96, indicating a well-performing model with 1378 total instances.

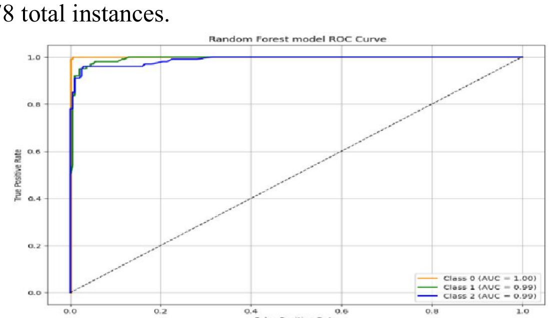

Fig. 6. ROC Curve of the Random Forest Model for Road Safety Risk

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, July 2025
Impact Factor: 7.67

A RF model's Receiver Operating Characteristic (ROC) curve is displayed in Figure 6. For the three categories (0, 1, 2), it displays the True Positive Rate vs the False Positive Rate. All of the class curves are quite near to the upper left corner, which means that they are doing very well. An extremely accurate multi-class classification model is indicated by the high Area Under the Curve (AUC) values: 1.00 for Class 0 and 0.99 for both Class 1 and Class 2.

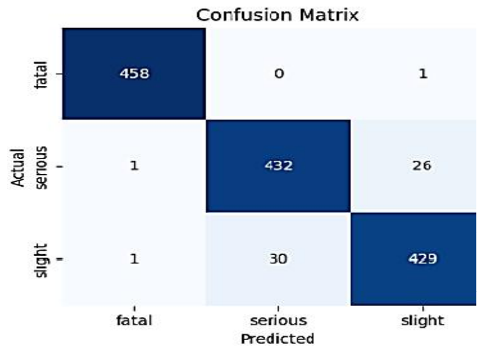

Fig. 7.Confusion Matrix of the RF Model for Road Safety Risk

Figure 7 displays a confusion matrix for a classification model, illustrating its performance across three classes: 'fatal', 'serious', and 'slight'. The diagonal values (458, 432, 429) represent correct predictions, showing high accuracy for each class. Off-diagonal values indicate misclassifications, which are notably low, suggesting the model distinguishes well between different severity levels. The heatmap color bar further emphasizes the count distribution.

A. Comparative Analysis

ML methods used to forecast and detect dangers to vehicle road safety are compared in this section. In Table III, can see how many architectures fared when accuracy was the measure of choice. As a result of its superior capacity to manage intricate feature interactions and deliver trustworthy risk categorization, the RF model attained the maximum accuracy of 96%. Once processed, the Long Short-Term Memory (LSTM) network achieved an accuracy of 91.74%, indicating that the network did an excellent job of learning the temporal patterns in driving data stream. Despite its simplistic design, the basic Neural Network model managed a respectable 91% accuracy rate in its predictions. At 89% accuracy, the XGBoost model also showed results that were competitive. This study's findings highlight the RF model as the top choice for predicting and assessing vehicle road safety risk levels, an essential step in developing effective preventative measures and reducing accident occurrences.

Table 3: Performance Comparison Of ML Models For Vehicle Road Safety Risk Prediction

Model	Accuracy
Random Forest model	96
LSTM[29]	91.74
Neural network[30]	91
XGBoost[31]	89

The suggested machine learning model successfully predicts road safety risks based on structured dataset on accidents, and the RF model shows an excellent result. The method makes it reliable as the pre-processing steps, such as feature engineering, SMOTE balancing, and normalization are performed in an organized manner. To the strength of the RF model, the framework allowed consistency in classification metrics that include: Accuracy, Precision, Recall, and F1-score of 96% respectively. Such findings confirm the high potential of the model to identify the risks of accidents and assist in pro-active treatment. The research proves the efficiency of the usage of the ML methods to create data-powered intelligent solutions to road safety improvement and wise transportation planning.

V. CONCLUSION AND FUTURE SCOPE

Road accident prevention is not limited to past analysis but it needs prospective that comes with intelligent systems. The study provides an integrated ML system to identify the road safety hazards before translating into actual accidents.

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 5, Issue 2, July 2025 Impact Factor: 7.67

Out of the tested models, RF algorithm was found to be the most competent, with the result being 96 per cent concerning accuracy, precision, recall and F1-score. These findings highlight the capability of the framework to identify intricate trends in traffic, environmental, and time series without imbalances and inconsistent performance to various levels of accident severities. The given model is not only accurate but also flexible and upgraded with the use of class balancing after the method of SMOTE and the intuitively chosen feature engineering. In spite of these successes the model works on historical data, and it is not able to adapt itself to live, changing very fast road conditions. This is something will work on in the future to combine live data sources like GPS, weather feeds and vehicle telemetry to add more power to the model. Further, the application of explainable AI (XAI) methods will introduce greater clarity and trust of stakeholders, particularly when applied in the regulatory or safety markets. With the evolution of urban infrastructures, the framework has a potential of becoming a building block of smart mobility systems- not only prediction, but preferably smart prevention of road accidents at a global, scalable level.

REFERENCES

[1] Y. Maqbool, A. Sethi, and J. Singh, "Road safety and Road Accidents: An Insight A Comprehensive Study on the Competent Traffic Management of Roads to Reduce Road Accidents in Srinagar, Jammu & Kashmir, India View project Road safety and Road Accidents: An Insight," *Int. J. Inf. Comput. Sci.*, vol. 6, no. 4, pp. 93–105, 2019.

[2] A. Borucka and S. Sobczuk, "The Use of Machine Learning Methods in Road Safety Research in Poland," *Appl. Sci.*, vol. 15, no. 2, p. 861, Jan. 2025, doi: 10.3390/app15020861.

[3] F. A. Mansuri, A. H. Al-Zalabani, M. M. Zalat, and R. I. Qabshawi, "Road safety and road traffic accidents in Saudi Arabia," *Saudi Med. J.*, vol. 36, no. 4, pp. 418–424, 2015, doi: 10.15537/smj.2015.4.10003.

[4] V. Singh, D. Pathak, and P. Gupta, "Integrating Artificial Intelligence and Machine Learning into Healthcare ERP Systems: A Framework for Oracle Cloud and Beyond," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 2, pp. 171–178, 2023.

[5] M. S. Shahriar, A. K. Kale, and K. Chang, "Enhancing Intersection Traffic Safety Utilizing V2I Communications: Design and Evaluation of Machine Learning Based Framework," *IEEE Access*, vol. 11, pp. 106024–106036, 2023, doi: 10.1109/ACCESS.2023.3319382.

[6] G. Modalavalasa, "Machine Learning for Predicting Natural Disasters: Techniques and Applications in Disaster Risk Management," *Int. J. Curr. Eng. Technol.*, vol. 12, no. 6, pp. 591–597, 2022.

[7] S. Kabade and A. Sharma, "Securing Pension Systems with AI-Driven Risk Analytics and Cloud-Native Machine Learning Architectures," *Int. J. All Res. Educ. Sci. Methods*, vol. 12, no. 4, pp. 5405–5416, 2024.

[8] B. K. M, A. Basit, K. MB, G. R, and K. SM, "Road Accident Detection Using Machine Learning," in *2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, Jul. 2021, pp. 1–5. doi: 10.1109/ICSCAN53069.2021.9526546.

[9] N. Malali, "Predictive AI for Identifying Lapse Risk in Life Insurance Policies : Using Machine Learning to Foresee and Mitigate Policyholder Attrition," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 5, no. 6, pp. 195–205, 2025, doi: 10.48175/IJARSCT-25328.

[10] R. Patel, "Automated Threat Detection and Risk Mitigation for ICS (Industrial Control Systems) Employing Deep Learning in Cybersecurity Defence," *Int. J. Curr. Eng. Technol.*, vol. 13, no. 06, pp. 584–591, 2023, doi: 10.14741/ijcet/v.13.6.11.

[11] M. Yousuf and S. Alhashmi, "Using Machine Learning for Road Accident Severity Prediction and Optimal Rescue Pathways," 2024.

[12] S. R. Sagili and T. B. Kinsman, "Drive Dash: Vehicle Crash Insights Reporting System," in *2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA)*, IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ICISAA62385.2024.10828724.

[13] X. Shi, Y. D. Wong, C. Chai, and M. Z.-F. Li, "An automated machine learning (AutoML) method of risk prediction for decision-making of autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7145–7154, 2020.

[14] Bhushan Chaudhari and Santhosh Chitraju Gopal Verma, "Synergizing Generative AI and Machine Learning for Financial Credit Risk Forecasting and Code Auditing," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 2, pp. 2882–2893, Apr. 2025, doi: 10.32628/CSEIT25112761.

[15] D. Patel, "Enhancing Banking Security: A Blockchain and Machine Learning- Based Fraud Prevention Model," *Int. J. Curr. Eng. Technol.*, vol. 13, no. 06, pp. 576–583, 2023, doi: 10.14741/ijcet/v.13.6.10.

[16] R. Q. Majumder, "A Review of Anomaly Identification in Finance Frauds Using Machine Learning Systems," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 5, no. 10, pp. 101–110, Apr. 2025, doi: 10.48175/IJARSCT-25619.

[17] S. R. Sagili, S. Chidambaranathan, N. Nallametti, H. M. Bodele, L. Raja, and P. G. Gayathri, "NeuroPCA: Enhancing Alzheimer's disorder Disease Detection through Optimized Feature Reduction and Machine Learning," in *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, IEEE, Jul. 2024, pp. 1–9. doi: 10.1109/ICEEICT61591.2024.10718628.

[18] V. A. Kumar, G. H. Radhan, R. S. Amshavalli, R. Raja, and N. Shafana, "Machine Learning-based Predictive Framework for Road Accident Severity Classification and Real-Time Risk Assessment," 2025, pp. 1038–1044. doi: 10.1109/ICVADV63329.2025.10961401.

[19] W. Noonpakdee, M. Satitsamitpong, P. Senachai, and K. Napontun, "Leveraging Machine Learning to Enhance Road Safety: A Social Marketing Approach," *ABAC J.*, vol. 45, no. 1, pp. 82–102, Jan. 2025, doi: 10.59865/abacj.2024.70.

[20] Y. Pei, Y. Wen, and S. Pan, "Road Traffic Accident Risk Prediction and Key Factor Identification Framework Based on Explainable Deep Learning," *IEEE Access*, vol. PP, p. 1, 2024, doi: 10.1109/ACCESS.2024.3451522.

[21] T. Wei, T. Zhu, M. Lin, and H. Liu, "Predicting and factor analysis of rider injury severity in two-wheeled motorcycle and vehicle crash accidents based on an interpretable machine learning framework," *Traffic Inj. Prev.*, vol. 25, no. 2, pp. 194–201, Feb. 2024, doi: 10.1080/15389588.2023.2284111.

[22] D. K. S N, "Accident Risk Prediction through Machine Learning: A Comprehensive Study," *Interantional J. Sci. Res. Eng. Manag.*, vol. 07, no. 08, pp. 1–8, 2023, doi: 10.55041/ijsrem25419.

[23] I. C. Obasi and C. Benson, "Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents," *Heliyon*, vol. 9, no. 8, Aug. 2023, doi: 10.1016/j.heliyon.2023.e18812.

[24] M. S. Reddy, E. M. Reddy, P. R. S. Varshith, and S. K. Kumari, "Predicting traffic flow in real time using machine learning," in *Hybrid and Advanced Technologies*, vol. i, London: CRC Press, 2025, pp. 525–530. doi: 10.1201/9781003559139-78.

[25] S. S. Yassin and Pooja, "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach," *SN Appl. Sci.*, vol. 2, no. 9, p. 1576, Sep. 2020, doi: 10.1007/s42452-020-3125-1.

[26] L. Hickman and M. Akdere, "Developing intercultural competencies through virtual reality: Internet of Things applications in education and learning," in *2018 15th Learning and Technology Conference (L&T)*, IEEE, Feb. 2018, pp. 24–28. doi: 10.1109/LT.2018.8368506.

[27] C. Shyamala, P. Sabarish, T. Vignesh, S. Yogeendran, and A. Professor, "Walmart Sales Prediction Using Machine Learning Algorithms," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 1, pp. 7872-7882–7872 – 7882, 2021.

[28] A. K. Mengistu *et al.*, "Predicting car accident severity in Northwest Ethiopia: a machine learning approach leveraging driver, environmental, and road conditions," *Sci. Rep.*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-08005-2.

[29] X. Wei, "Enhancing road safety in internet of vehicles using deep learning approach for real-time accident prediction and prevention," *Int. J. Intell. Networks*, vol. 5, pp. 212–223, 2024, doi: https://doi.org/10.1016/j.ijin.2024.05.002.

[30] M. S. Butt and M. A. Shafique, "A literature review: AI models for road safety for prediction of crash frequency and severity," *Discov. Civ. Eng.*, vol. 2, no. 1, May 2025, doi: 10.1007/s44290-025-00255-3.

[31] A. R. Mussah and Y. Adu-Gyamfi, "Machine Learning Framework for Real-Time Assessment of Traffic Safety Utilizing Connected Vehicle Data," *Sustainability*, vol. 14, no. 22, 2022, doi: 10.3390/su142215348.