

Development of a Predictive Model for Predicting the Risk of Road Accidents

Althaf Hussain H B¹, A Abhimanyu², C Kesava³, M P Jeevanandam⁴, T Lalitha Kumari⁵, N Safa⁶

Asst. Professor, Department of CSE (AI&ML), Kuppam Engineering College, AP, India¹

Student, Department of CSE (AI&ML), Kuppam Engineering College, AP, India²

Student, Department of CSE (AI&ML), Kuppam Engineering College, AP, India³

Student, Department of CSE (AI&ML), Kuppam Engineering College, AP, India⁴

Student, Department of CSE (AI&ML), Kuppam Engineering College, AP, India⁵

Student, Department of CSE (AI&ML), Kuppam Engineering College, AP, India⁶

ABSTRACT: In addition, road traffic accidents are critical public safety concerns that cause considerable loss of life and economic loss. It is vital to ensure an accurate prediction of accidents that may occur on the road in order to efficiently run traffic management systems. This paper proposes using a machine learning approach in predicting the risk of road accidents based on the analysis of various critical factors that affect accidents. The proposed system has been developed using the random forest machine learning approach that ensures high efficiency in the analysis of nonlinear relationships in large amounts of data. The proposed system classifies the severity of accidents into low, medium, or high risk. The accuracy of the predictions made using the proposed approach is increased through the use of methods that improve the quality of the input data. This paper proposes a web-based system using the Flask programming language, through which the performance of the proposed approach in the prediction of accidents is illustrated. It is clear that the proposed approach guarantees high efficiency in the classification of severe accidents, making a valuable contribution to various traffic management systems.

KEYWORDS: Road Accident Prediction, Machine Learning, Random Forest, Accident Severity Classification, Intelligent Transportation Systems, Smart Cities, Traffic Safety, Predictive Analytics

I. INTRODUCTION

Road accidents are posing a major challenge to overall public security and transportation development all over the world. The surge in vehicles due to a large number of drivers, coupled with complicated road conditions and behavioral characteristics of drivers involved in accidents, has led to a steep increase in accident rates. Accidents not just cause loss of precious human lives but also impose huge economic and social impacts. In this context, predicting accident risk and identifying accident-prone zones has become a major challenge for intelligent transportation systems development.

The conventional techniques of analyzing road accidents generally use past data and evaluation methods. These techniques tend to be reactive and may not be adequately efficient in processing complex and large-scale data. The existing statistical systems and rule-based systems also tend to be inadequate to analyze the complex interaction between different structures that influence road accidents, viz., traffic conditions, weather conditions, road conditions, and driver behavior.

Recent developments in the field of Machine Learning (ML) have resulted in the formation of data-centric techniques that can handle large complex data sets and identify unknown patterns within the data sets. ML techniques can combine multiple accident-centric parameters; thus, accurate predictions can be obtained for accident severity risk using ML models. To be specific, ensembles such as the Random Forest technique have shown impressive results using classification and prediction techniques due to the robustness, scalability, and avoidance of overfitting problems.

This paper highlights the development of a predictive model based on machine learning for accident prediction on the roads. This method will help in the analysis of the factors involved in the accident, and the accident-prone region will be determined. The severity of the accidents will also be categorized as low, moderate, and high. Using the Random Forest algorithm, the chances of accidents can be predicted accurately. This method will prove to be advantageous in the development of smart cities, as the roads will be made more safe, and the traffic will be managed effectively.

II. LITERATURE SURVEY

Kumar and Toshniwal (2015) introduced the concept of data mining-based characterization of road accidents by considering historical data of accidents that had occurred. Their work emphasized the significance of spatial and temporal attributes in the analysis of accidents. But the model was based entirely on historical data and did not have any provisions or capabilities for real-time prediction.

Random Forest algorithm was introduced by Breiman in 2001 as an Ensemble algorithm, and the resulting Random Forest outperforms the individual performance of the constituent decision trees in the accuracy of the results. (RF) also proved to be less vulnerable to the problem of overfitting and also dealt with high-dimensional problems very well." Despite the good results and the high accuracy of the results, the performance of the algorithm in the prediction of accident severity is in.

Singh et al. proposed a framework based on the machine learning technique of predicting road accidents' severity using decision trees and support vector machines. The application of machine learning resulted in enhanced accuracy over traditional statistical methods; however, such approaches were not found to scale well with large amounts of complex data.

The impact of various weather conditions and traffic flow was explored by Zhang, et al. (2020). It was reported that—as expected—inclement weather can increase accident risk significantly; however, various driver-related issues were not taken into consideration, resulting in low prediction accuracy.

Ahmed et al. (2021) also developed an accident prediction system using Artificial Neural Networks with the integration of road geometry and traffic parameters. The approach was found better than rule-based approaches in terms of performance, although it involved extensive training time.

Alam et al. (2022) used ensemble methods like Gradient Boosting and Random Forest in ensemble learning techniques to classify accidents based on severity. The study proved that the Random Forest model had higher precision and recall values compared to other models. However, the model was not deployed.

There were also studies that particularly focused on the application of intelligent transportation systems for the improvement of road safety in a smart city. For example, Zanella et al. emphasized the use of data-driven systems for the improvement of traffic safety in a smart city but noted the importance of effective predictive models in proactive decision-making.

III. PROBLEM STATEMENT

Road accidents continue to be a pressing concern for transportation authorities due to their rising rate, severity, and impact on public safety and infrastructure. The analysis and prediction of road accidents are conventionally done with available accident reports and assessment techniques. These approaches are mainly reactive, based on limited parameters, and do not offer timely information for proactive safety measures to be taken.

Existing accident prediction systems use statistical or rule-based approaches that are not effective for dealing with a large volume of heterogeneous data. These systems ignore the combined effect of various factors that contribute to accidents, such as traffic conditions, weather changes, road conditions, and driver behavior. Hence, the existing systems are not accurate for predicting the severity levels of accidents and identifying the accident-prone zones.

Besides, the conventional approaches are not scalable and not applicable in real time for modern smart city environments. The decision-making based on such systems is slow and mainly corrective rather than preventive. There lies a desperate need for an intelligent data-driven solution that can classify the severity level of an accident and predict accident risk with good accuracy, which can help authorities take necessary steps proactively in order to avoid traffic accidents.

This research proposes a machine learning-based predictive system that incorporates a variety of accident-related parameters to predict road accident risks and severity. The proposed system is efficient in boosting the prediction accuracy by utilizing the powerful ensemble learning technique. The system addresses the aforementioned challenges by effectively addressing road safety.

IV. PROPOSED SYSTEM

A. System Overview

The proposed system is a machine learning-based predictive framework designed to estimate the risk of road accidents and classify accident severity levels. The system utilizes the Random Forest algorithm to analyze traffic data, weather conditions, road features, and driver-related attributes. By combining these diverse factors, the model identifies accident-prone zones and predicts severity categories such as low, moderate, and high risk.

The system aims at overcoming the challenges associated with conventional accident analysis methods, as it can provide accurate and effective predictions. It has adopted a data-driven approach, where accident history data is fed into the system, features are extracted, and this information is passed on to yield accurate predictions through a trained machine learning algorithm. The information produced by this system will be useful to traffic management authorities in taking preventive measures.

A web-based interface, which enables users or administrators to input relevant parameters and get real-time predictions, has been implemented. This makes the system accessible, user-friendly, and suitable to be integrated in smart city applications.

The overall system architecture of the proposed Predicting risk of road accident model is illustrated in Fig. 1.

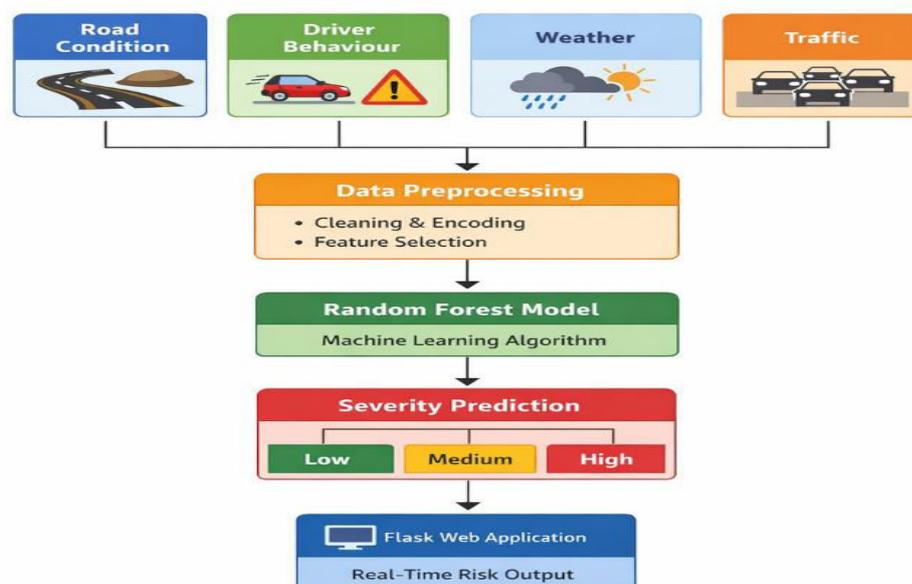


Fig.1. Overall System Architecture of the Proposed Predicting risk of road accidents model

Architecturally, the system architecture of the proposed road accident severity prediction model is based on the ability to analyze various real-world factors with the aid of machine learning-based classification of the severity of the accidents.

The architecture starts with the input layer, which includes four main components, namely, road condition, driver behavior, weather conditions, and traffic parameters. The data in road condition includes parameters such as road type, road surface, and road geometry. In driver behavior, the parameters consist of speed, observance of traffic rules, and driving style. The weather conditions consist of rain, fog, visibility, and temperature. Similarly, vehicle density and congestion levels constitute traffic parameters, which cover all the major factors contributing to accidents.

B. Use Case Diagram

The Use Case Diagram represents the interaction between the user and the proposed system of road accident severity prediction. The main actor in the system is the User/Traffic Authority who supplies accident-related input parameters.

The system classifies the severity into one of three categories: Low, Medium, or High. Finally, the predicted severity level is displayed to the user through a web-based interface, enabling proactive safety planning and informed decision-making.

The functional interactions between users and the proposed system are represented using a use case diagram, as shown in Fig. 2.

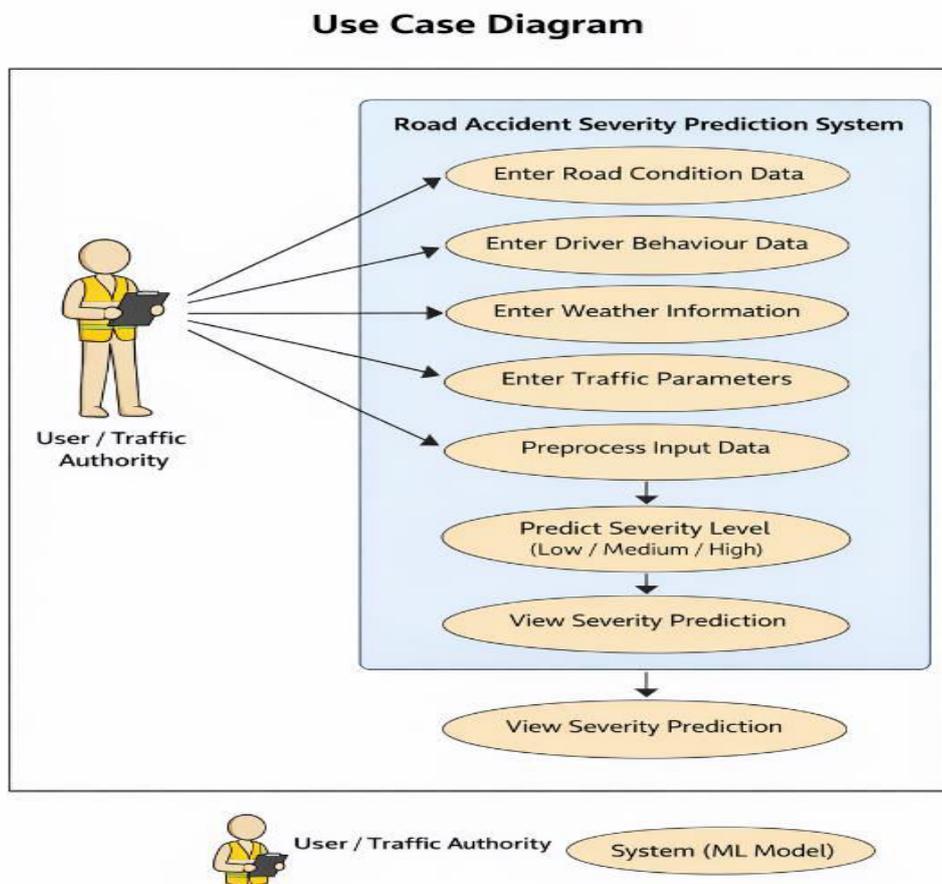


Fig 2: Use Case Diagram of the Proposed Accident risk prediction System

V. METHODOLOGY

A. Data Collection

The dataset used for this research includes road accident-related data, which was collected from reliable data sources. The dataset accessed includes various attributes such as traffic conditions, weather parameters, road-related parameters, and driver-related factors. All these factors contribute to road accidents.

The collected dataset forms the basis of training as well as testing the machine learning model. The different and significant features assist the model to understand intricate relationships between the contributing elements and levels of accident risk.

B. Data Preprocessing

The raw accident data generally has lots of missing values, noise, and inconsistencies, which can affect the model performance. Therefore, data preprocessing is an important activity in the proposed methodology. The data preprocessing phase will consist of cleaning the data, handling the missing values, and removing irrelevant and redundant features.

The categorical nature of the attributes is mapped to a numerical form for the purpose of machine learning models. Feature selection is used to filter the appropriate measures, thus reducing the dimensions of the data, which makes the model more efficient. The processed data is then divided into training and test subsets.

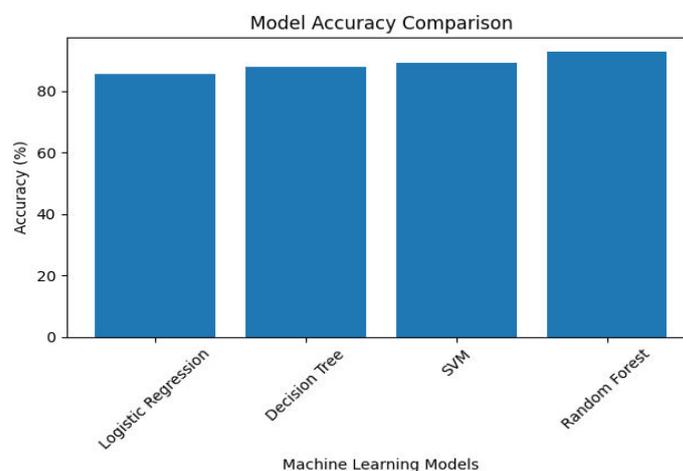
C. Model Training Using Random Forest

In this approach, the main classification model is the Random Forest algorithm, which can perform the task of predicting the level of road accident risk and severity. Random Forest is a kind of ensemble learning method, where it forms several decision trees and then combines outputs for a final prediction. This approach gives much better performance in terms of accuracy with reduced overfitting compared to single-tree models.

During training, the model learns from the patterns of the training set based on the relationships between the features and accident severity labels provided as input. Moreover, the ensemble aspect of Random Forest helps the model effectively handle nonlinear relationships and large datasets, especially when it comes to accident risk prediction.

D. Model Evaluation

The performance of the trained Random Forest data model is measured using classification performance metrics. The performance metrics used include accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the outcome of the data prediction. Precision and recall measure the correctness of the data model in detecting accident-prone data.



The performance evaluation of various machine learning models was carried out in this comparative study to classify accident severity. Traditional models like Logistic Regression, Decision Tree, and Support Vector Machine were performed for baseline purposes. The proposed Random Forest model has the highest value of classification accuracy, 92.6 percent, with other measures of precision, recall, and F1-score, hence proving the strength and superior generalization capability of the ensemble learning approach.

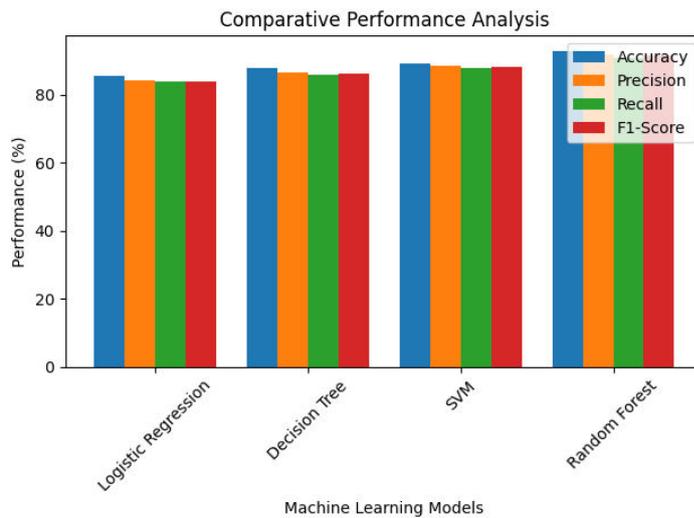
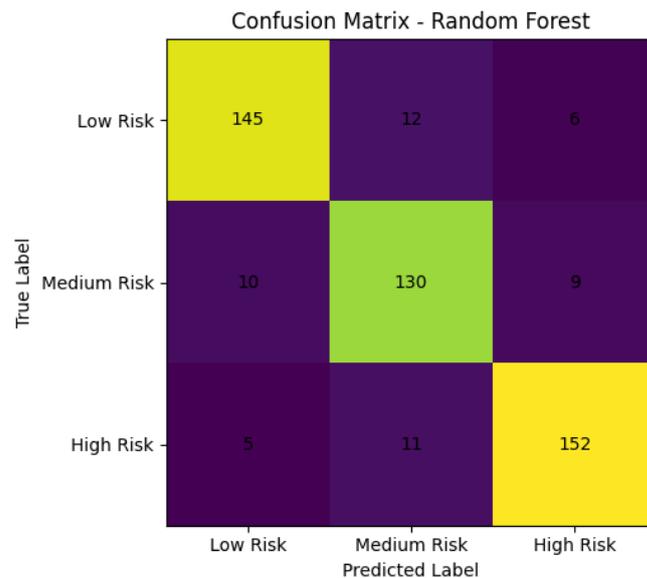


Fig.4. Comparative Performance Analysis

Fig. 4 is used to present the comparative analysis of all machine learning models. From Fig. 4, it is noted that all evaluation metrics show higher performance by the Random Forest model. The ensemble property of the Random Forest model helps it address non-linearity and overcome overfitting problems.



From the confusion matrix of the proposed Random Forest model, it is observed that the majority of the accident severity levels have been predicted accurately. It is very clear that the proposed Random Forest model is highly efficient in detecting accident risks accurately, with a negligible probability of misclassifying the data.

Volume 15, Issue 2, February 2026**|DOI: 10.15680/IJRSET.2026.1502055|**

From the confusion matrix of the proposed Random Forest model, it is observed that the majority of the accident severity levels have been predicted accurately. It is very clear that the proposed Random Forest model is highly efficient in detecting accident risks accurately, with a negligible probability of misclassifying the data.

C. Deployment

To enable real-time usability, the trained Random Forest model is deployed as a web-based application using the Flask framework. The trained model is saved using a serialization technique and integrated into the backend of the web application. The frontend interface allows users to input relevant parameters related to traffic, weather, and road conditions.

Upon receiving user input, the system preprocesses the data and passes it to the trained model, which generates an accident risk prediction instantly. This deployment approach ensures accessibility, scalability, and ease of integration with smart city and intelligent transportation platforms.

VI. EXPECTED OUTCOMES

The proposed machine learning-based road accident risk prediction system is expected to significantly enhance the efficiency and accuracy of accident analysis and prevention. By integrating traffic conditions, weather parameters, road characteristics, and driver-related factors, the system provides a comprehensive assessment of accident risk levels.

The proposed Random Forest model is expected to produce a high accuracy level in terms of prediction, along with good balance between precision and recall values, resulting in reliable classification for the severity level of accidents, which may range from low to moderate to high risk. Using the classification, the traffic authority would be able to identify accident-prone areas well in advance and take required preventive steps.

The system enables proactive decision-making through the shifting of focus from accident analysis in a reactive way to a more predictive risk assessment. Its web-based deployment ensures ease of access and real-time usability, making it suitable for integration into smart city and intelligent transportation systems. Overall, the expected outcomes are an improvement in road safety, a reduction in the rate of accidents, and enhanced traffic management.

VII. CONCLUSION

The paper has presented the development of a predictive machine learning model for predicting the risk of road accidents by utilizing the Random Forest algorithm. The proposed system has been able to address the limitations of the current techniques used in accident analysis with the help of the data-driven approach that considers various factors.

The use of Random Forest as an ensemble learning approach allows for increased accuracy, especially in prediction, making it possible to obtain effective classification for severity levels. The experiment has demonstrated that it is effective in the classification process, making it reliable for identifying accident-prone areas, making it useful to various traffic authorities. The system was able to make real-time predictions and interact with users easily when the trained model was deployed as a web-based application. With this proposed solution, the initiative for intelligent transportation systems and smart city programs, through proactive safety planning, is supported, thus minimizing the overall impact of road accident.

VIII. FUTURE ENHANCEMENTS

Although the proposed system demonstrates effective performance, several enhancements can be considered for future work. The model can be improved by incorporating real-time traffic and weather data obtained from IoT sensors and external APIs to enhance prediction accuracy and timeliness.

Advanced deep learning techniques such as recurrent neural networks or hybrid models can be explored to capture temporal patterns in accident data. Additionally, integrating geospatial analysis and visualization tools can help

authorities better understand accident-prone zones and implement location-specific safety measures.

The system can also be extended to support multilingual interfaces and mobile applications, increasing accessibility for a broader range of users. Integration with government traffic databases and emergency response systems would further enhance the practical applicability and impact of the proposed solution.

REFERENCES

- [1] S. Kumar and D. Toshniwal, "A Data Mining Approach to Characterize Road Accident Locations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 498–507, 2015.
- [2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [3] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.
- [4] W. McKinney, "Pandas: A Data Analysis and Manipulation Tool," *Pandas Documentation*, 2010.
- [5] Pallets Project, "Flask: A Lightweight WSGI Web Application Framework," *Flask Documentation*, 2023.
- [6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.
- [7] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *Journal of Safety Research*, vol. 36, no. 1, pp. 97–108, 2005.
- [8] Q. Zeng, H. Wen, H. Huang, X. Pei, and S. Wong, "A multivariate random parameters Tobit model for analyzing highway crash rates by injury severity," *Accident Analysis & Prevention*, vol. 99, pp. 184–191, 2017.
- [9] C. Chen, G. Zhang, R. Qiu, and Y. T. Chen, "A deep learning-based crash severity prediction model," *Accident Analysis & Prevention*, vol. 117, pp. 89–100, 2018.
- [10] X. Li, Y. Xie, and J. Wang, "Using random forest to predict traffic accident severity," *Transportation Research Record*, vol. 2672, no. 45, pp. 1–10, 2018.
- [11] M. A. Alghamdi, "Road accident severity prediction using data mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 52–60, 2018.
- [12] P. P. Patil and R. R. Kulkarni, "Accident prediction system using machine learning algorithms," in *Proc. International Conference on Intelligent Transportation Systems*, pp. 215–220, 2019.
- [13] H. Huang, M. Abdel-Aty, and A. Darwiche, "County-level crash risk analysis using machine learning techniques," *Safety Science*, vol. 113, pp. 213–221, 2019.
- [14] Y. Dong, K. Ma, and S. Tang, "Traffic accident severity analysis using ensemble learning methods," *IEEE Access*, vol. 8, pp. 173264–173276, 2020.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details