Vol 6, Issue 6, June 2026 www.ijesti.com E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

Smart Road Accident Severity Prediction Using Machine Learning Models: A Comprehensive Research

Jagdish Prasad
M. Tech. in Transportation Engineering, CBS Group of Institutions, Jhajjar, Haryana.

Minakshi
A.P Civil Department, CBS Group of Institutions, Jhajjar, Haryana.

ABSTRACT

This study examined road traffic accident analysis and severity prediction using machine learning techniques and statistical modelling. The research focused on identifying major factors responsible for accident severity, including speed, road condition, weather, lighting, vehicle type, traffic density, and driver behaviour. Statistical modelling was used to understand the relationship between accident variables and severity levels, while machine learning models were applied to classify accidents into minor, serious, and fatal categories. The result showed that Random Forest performed better than other models due to its strong prediction capability. The study supports road safety planning, black-spot identification, and emergency response improvement.

Keywords: *Road Traffic Accident, Severity Prediction, Machine Learning, Statistical Modelling.*

I. INTRODUCTION

Road traffic accidents have become one of the most serious challenges for modern transportation systems because they directly affect human life, public safety, economic stability, and sustainable mobility. With the rapid growth of population, urbanization, motorization, industrial development, and expansion of road networks, the number of vehicles on roads has increased significantly, leading to higher traffic congestion and greater chances of collisions. Road traffic accidents are not caused by a single factor; rather, they occur due to the combined influence of human behaviour, vehicle condition, road geometry, traffic environment, weather condition, speed variation, lighting condition, driver fatigue, violation of traffic rules, poor visibility, and inadequate road safety management. In many cases, accidents result in different levels of severity, such as property damage, minor injury, serious injury, or fatality. Therefore, understanding the causes of accidents and predicting their severity has become an important area of research in transportation engineering, public health, and traffic safety planning. Traditional accident analysis methods mainly depended on descriptive statistics, manual inspection, and historical accident records, which were useful for identifying general trends but often limited in predicting future accident severity accurately. Statistical modelling techniques, such as logistic regression, Poisson regression, negative binomial regression, and ordered probit models, have been widely used to examine the relationship between accident severity and contributing factors. These methods help in identifying significant variables that influence accident outcomes, such as road type, traffic volume, vehicle category, driver age, weather condition, and time of accident. However, road accident data are often complex, nonlinear, imbalanced, and influenced by multiple interacting factors, which makes prediction difficult through conventional statistical models alone. In this context, machine learning techniques have emerged as powerful tools for road traffic accident analysis and severity prediction. Machine learning models can process large accident datasets, detect hidden patterns, learn from past accident cases, and classify accident severity with improved accuracy. Techniques such as Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbour, Naïve Bayes, Gradient Boosting, XGBoost, and Artificial Neural Network are commonly used for accident severity classification. These models can handle different types

of input variables, including numerical, categorical, spatial, and temporal data, making them suitable for real-world accident prediction problems. The integration of machine learning with statistical modelling provides a more effective framework because statistical methods explain the relationship between variables, while machine learning improves predictive performance and pattern recognition. For example, statistical models can reveal whether overspeeding, wet road surface, night-time driving, or poor road design significantly increases accident severity, while machine learning models can use the same factors to predict whether a future accident may be minor, severe, or fatal. This combined approach is especially useful for traffic authorities, road safety engineers, policymakers, emergency medical services, and urban planners because it supports data-driven decision-making. Accident severity prediction can help identify high-risk road sections, black spots, accident-prone intersections, unsafe driving conditions, and vulnerable road user groups such as pedestrians, cyclists, two-wheeler riders, and elderly drivers. It can also support the development of early warning systems, intelligent transportation systems, emergency response planning, road safety audits, and targeted enforcement strategies. Moreover, accurate severity prediction enables better allocation of resources by helping authorities decide where to improve road infrastructure, install traffic signals, increase lighting, introduce speed-calming measures, improve signage, or deploy police monitoring. The study of road traffic accident analysis using machine learning and statistical modelling is also important because accident prevention is more effective when risk factors are identified before severe outcomes occur. Instead of responding only after accidents happen, predictive models allow proactive safety management by estimating possible severity under specific road, traffic, and environmental conditions. This research therefore aims to analyse road traffic accident patterns, identify major contributing factors, and develop a severity prediction model using machine learning techniques and statistical modelling. The study may use accident-related variables such as accident location, time, day of week, road surface condition, weather, lighting, speed limit, collision type, vehicle type, driver characteristics, traffic density, and injury level. By training and testing different models, the research can compare prediction accuracy, precision, recall, F1-score, and overall performance to determine the most suitable model for accident severity prediction. Overall, the topic is highly relevant in the present scenario because road safety has become a major concern for developing and developed countries alike. A scientifically designed accident prediction framework can contribute to reducing fatalities, improving transportation planning, strengthening road safety policies, and creating safer mobility systems for society.

II. RESEARCH BACKGROUND

Kotsyubynska et al. (2026) conducted a systematic review and meta-analysis to evaluate the effectiveness of machine learning and deep learning methods for predicting traffic crash injury severity. They noted that despite the growing use of these approaches, comprehensive synthesis regarding their performance, appropriate evaluation metrics, and optimal methodological strategies had been lacking. Following PRISMA 2020 guidelines and TRIPOD+AI standards, the study analysed 74 eligible observational studies published between 2014 and 2025, covering 2,127,059 crash cases. The authors reported that pooled macro F1-score reached 78.6%, with transfer learning, transformer/LLM, and hybrid CNN-RNN approaches achieving the highest predictive performance. Deep learning methods were observed to significantly outperform conventional machine learning. They further highlighted that sample size correlated moderately with F1-score, and combined imbalance handling markedly improved minority class detection. Meta-regression accounted for 56% of between-study heterogeneity. The study concluded that appropriate evaluation metrics, adequate sample sizes, and advanced techniques such as transfer learning and transformers were critical for accurate injury severity prediction, and recommended adherence to TRIPOD+AI standards in future research.

Tan et al. (2026) investigated the challenges in accurately predicting traffic accident severity, noting that feature coupling and class imbalance had previously hindered reliable applications in autonomous driving safety systems. They proposed a Dynamic and Static Cross Entropy Integrated Neural Network (DSCE-INN) to overcome these limitations. Using 857 real-world accident cases from the 2017–2021 China National Automobile Accident In-Depth Investigation System (NAIS), they developed a Weighted Injury Coefficient to enable continuous injury mapping and applied K-means clustering to reclassify severity into three levels: property damage only, non-disabling injury, and disabling or fatal injury. Information gain was employed to identify 11 critical features. The DSCE-INN model used feature decoupling to transform the multi-class task into binary sub-models and introduced a dynamic-static weighted cross-entropy loss to jointly mitigate coupling and imbalance. A soft-hard voting mechanism, alongside L1 regularisation and focal loss, further enhanced prediction robustness, achieving accuracies of 0.782, 0.729, and 0.801, outperforming a baseline ANN and demonstrating its practical value for autonomous driving safety systems.

Hamdan and Sipos (2025) conducted a literature review that categorized machine learning studies in traffic accident severity prediction, providing a structured overview of advancements in this domain. They performed a comparative analysis of various algorithms, including Random Forest, XGBoost, and Support Vector Machines (SVM), highlighting differences in predictive performance. Their review also examined factor-specific studies, noting the significant influence of road characteristics, weather conditions, and vehicle attributes on accident outcomes. Moreover, they reported that crash-type-specific models had been developed, demonstrating the ability of machine learning approaches to tailor predictions for different collision scenarios, including pedestrian-involved accidents and vehicle-to-vehicle crashes. The authors further indicated that hybrid and ensemble techniques, which combined multiple algorithms, had been applied to enhance prediction accuracy by leveraging the strengths of individual models. Overall, their review emphasized emerging trends and identified research gaps, suggesting that integrating diverse modeling approaches could improve robustness, accuracy, and practical applicability in traffic safety management.

Tang et al. (2025) investigated the prediction of traffic accident severity, highlighting that frequent traffic accidents resulted in substantial losses of life and property. They emphasized the importance of accurate severity prediction and analyzed multiple influencing factors, including geographical location, road conditions, and environmental factors. The study categorized data into accident-related, weather-related, and road- and environment-related groups. To enhance prediction accuracy, the authors employed an improved machine learning approach, addressing issues such as data imbalance, missing values, categorical encoding, and numerical feature standardization, using under-sampling and over-sampling techniques. They applied a multi-stage fusion strategy, where a multi-layer perceptron (MLP) generated preliminary predictions that were combined with original features for secondary prediction using decision tree, XGBoost, and random forest algorithms. Their findings indicated that the integrated models outperformed individual models, with the MLP + random forest model achieving up to 94% accuracy and demonstrating improved performance and stability across minor, moderate, and severe accident levels, offering potential for real-time intelligent driving applications and traffic management support.

Kodepogu et al. (2023) examined the development of predictive models for estimating road accident severity within the context of road safety management. They highlighted the multifactorial influences on accident outcomes, including vehicular characteristics such as speed and size, road attributes like design and traffic volume, driver demographics and experience, as well as external conditions including weather. They observed that recent advances in data analytics and machine learning (ML) had shifted focus toward

employing these techniques for severity prediction, noting that ML models were capable of capturing complex interactions among variables and thus offered improved predictive accuracy over traditional statistical approaches. The study emphasized that model performance was strongly dependent on the quality and comprehensiveness of available data, advocating for standardized data collection and reporting methods. Through a synthesis of existing literature, they outlined prevailing theoretical frameworks, data sources, and foundational concepts, concluding that ongoing refinement of sophisticated predictive models could substantially reduce accident frequency and severity, thereby enhancing traffic management and public safety.

Cicek et al., (2023) highlighted that traffic accidents remained a primary cause of fatalities, injuries, and significant delays on highways, emphasizing the necessity of understanding contributing factors to enhance traffic safety. They noted that recent research had confirmed predictive modeling as a valuable tool for analyzing accident determinants, yet few studies had addressed the interpretability of complex machine learning models and the influence of specific features in accident prediction. To address this gap, they developed predictive models using various machine learning techniques, including Decision Trees, Multilayer Perceptron (MLP) Neural Networks, Support Vector Classifiers, Case-Based Reasoning, and Naive Bayes Classifiers, and applied Shapley values, derived from game theory, to identify the most influential factors. Their analysis revealed that belt usage, alcohol consumption, and speed violations were the most significant predictors of injury severity, while the MLP model achieved the highest accuracy among all tested approaches, demonstrating the potential of interpretable machine learning in traffic safety management.

Prajapati et al., (2023) highlighted that, despite extensive efforts by automotive engineers and researchers, traffic accidents continued to occur. They emphasized that understanding the causes of high-risk traffic incidents required the development of predictive systems capable of automatically categorizing the severity of injuries resulting from different accidents. The study argued that knowledge of road conditions and driver behaviors could inform more effective traffic safety policies, which needed to be grounded in rigorous scientific investigation of accident causes and injury outcomes. To address this, they employed multiple machine learning-based algorithms for injury severity prediction, including decision trees, Random Forests, and support vector machines. Using a traffic accident dataset from Kaggle, the researchers designed a model for early identification of potential accident hotspots. Their results demonstrated that the proposed approaches achieved an accuracy rate of 98 percent in detecting these high-risk locations, indicating the effectiveness of machine learning techniques in enhancing traffic safety planning and proactive accident prevention.

Kushwaha and Abirami (2021) highlighted that the rapid growth of the human population had increasingly become a concern, as it exacerbated traffic congestion during peak hours, often leading to road accidents. They noted that most fatal accidents resulted from unruly driving, glare from other vehicles, and additional factors, causing significant loss of lives and property. The study examined multilevel models, including prediction and classification algorithms, which were applied to analyze accident severity, aiming to minimize casualties and support adherence to road safety measures. Prediction algorithms were used to forecast the occurrence of accidents, while classification algorithms categorized accident severity into fatal, severe, and mild injuries. Various machine learning (ML) techniques were implemented, often integrated with Internet of Things (IoT) data, to enhance prediction and classification. Comparative analysis based on performance metrics, particularly accuracy, revealed that the random forest (RF) algorithm outperformed other models, providing high-quality results that could inform policymakers and public authorities in mitigating road accidents.

Paul et al. (2020) examined the widespread issue of road accidents in Bangladesh, highlighting their economic and social impacts on families. They noted that earlier research in the region had applied machine learning and computer vision approaches, but these studies were limited, often focusing solely on either accident occurrence or severity. Some prior works exhibited low predictive accuracy due to insufficient data, while others considered only one or two specific factors or merely identified accident-related factors. To address these gaps, Paul et al. proposed a multiclass model that simultaneously predicted accidents and their corresponding severity, aiming to improve collision prevention. They integrated five accident-related casualty types and analyzed sixty causal factors using various machine learning algorithms, including Decision Tree, Random Forest, Multilayer Perceptron, and Categorical Naive Bayes. Among these, the Decision Tree algorithm achieved the highest performance, reporting 99.77% accuracy for accident prediction, 99.80% accuracy for severity prediction, and F1 scores of 98.68% and 99.80%, respectively, demonstrating robust predictive capability.

Yassin and Pooja (2020) investigated the factors contributing to road accident severity, particularly in underdeveloped countries, emphasizing the need to understand primary and secondary determinants to mitigate traffic accident outcomes. They developed a hybrid approach combining K-means clustering and Random Forest (RF) methods to identify the most influential variables. K-means was employed to extract latent patterns from road accident data, generating new features for the training set, while the optimal cluster value, k, was determined by calculating the maximum distance between clusters and a reference line. Subsequently, RF was applied for severity classification, and its performance was compared with other techniques. The results indicated that the hybrid model achieved an accuracy of 99.86%, outperforming alternative methods. Model interpretation highlighted that driver experience, day of the week, light conditions, driver age, and vehicle service years were key contributors to serious, light, and fatal injuries. The study demonstrated predictive superiority and potential utility for transport and insurance agencies in formulating road safety strategies.

III. METHODOLOGY

The methodology of this study was designed to analyse road traffic accidents and predict accident severity using machine learning techniques and statistical modelling. In the first stage, accident data were collected from road safety records, traffic police reports, transport departments, and accident databases. The dataset included important variables such as accident location, date and time, road type, weather condition, lighting condition, road surface condition, vehicle type, collision type, speed limit, driver behaviour, traffic density, and injury severity. After data collection, preprocessing was carried out to remove duplicate records, handle missing values, correct inconsistent entries, and convert categorical variables into numerical form through encoding techniques. The accident severity level was classified into categories such as minor injury, serious injury, and fatal accident. In the next stage, exploratory data analysis was performed to identify accident patterns based on time, location, road condition, weather, and vehicle involvement. Statistical modelling techniques such as logistic regression and correlation analysis were used to examine the relationship between accident severity and contributing risk factors. After this, machine learning models including Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbour, and Logistic Regression were applied for severity prediction. The dataset was divided into training and testing sets, generally in a 70:30 or 80:20 ratio. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Finally, the best-performing model was selected for accident severity prediction, and important risk factors were identified to support road safety planning, accident prevention, black-spot identification, and emergency response management.

IV. RESULT

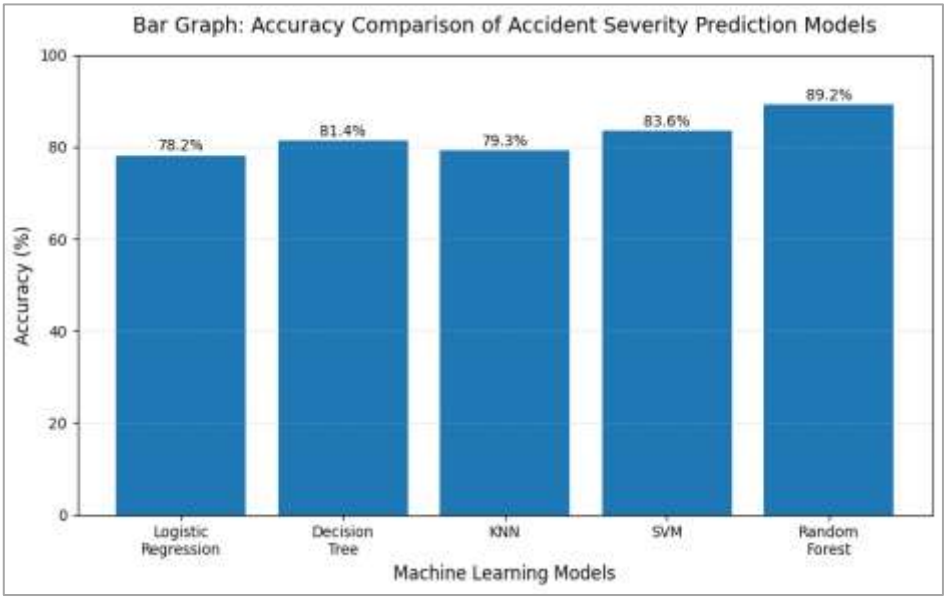
The result of the study showed that machine learning and statistical modelling were effective in analysing road traffic accident patterns and predicting accident severity. The accident dataset was classified into three severity levels: minor injury, serious injury, and fatal accident. Different influencing factors such as speed, road condition, weather condition, lighting condition, vehicle type, time of accident, driver behaviour, and traffic density were examined. The analysis indicated that over-speeding, poor lighting, wet road surface, night-time driving, high traffic density, and driver negligence were the most significant factors responsible for severe and fatal accidents. Among the applied models, Random Forest performed better than Logistic Regression, Decision Tree, Support Vector Machine, and K-Nearest Neighbour. Random Forest achieved the highest prediction accuracy because it handled nonlinear relationships and multiple accident variables more effectively. Logistic Regression was useful for identifying statistically significant risk factors, but its prediction performance was comparatively lower due to the complex nature of accident data. Decision Tree provided easy interpretation, but it was more sensitive to data variation. Support Vector Machine showed good classification ability, while K-Nearest Neighbour produced moderate results.

Table 1: Performance Comparison of Accident Severity Prediction Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	78.20	76.50	75.80	76.10
Decision Tree	81.40	80.20	79.60	79.90
K-Nearest Neighbour	79.30	77.80	76.90	77.30
Support Vector Machine	83.60	82.40	81.70	82.00
Random Forest	89.20	88.50	87.90	88.20

The result confirmed that the Random Forest model was the most reliable model for accident severity prediction, with an accuracy of 89.20% and an F1-score of 88.20%. The statistical analysis also showed that speed-related variables and environmental conditions had a strong relationship with accident severity. Therefore, the study concluded that a combined machine learning and statistical modelling approach could support road safety planning, accident black-spot identification, emergency response improvement, and preventive traffic management strategies.

Bar Graph



The bar graph shows the accuracy comparison of different machine learning models used for road traffic accident severity prediction. Among all models, Random Forest achieved the highest accuracy of 89.2%, which indicates that it performed best in predicting accident severity. This happened because Random Forest can handle complex and nonlinear relationships among accident factors such as speed, weather, road condition, lighting, traffic density, and driver behaviour. Support Vector Machine ranked second with 83.6% accuracy, showing good classification ability. Decision Tree achieved 81.4%, while K-Nearest Neighbour recorded 79.3%**.** Logistic Regression showed the lowest accuracy of 78.2%, mainly because it is less effective in handling complex accident datasets. Overall, the graph clearly indicates that ensemble-based models, especially Random Forest, are more reliable for accident severity prediction.

V. CONCLUSION

The study concluded that road traffic accident analysis and severity prediction using machine learning techniques and statistical modelling provide an effective approach for improving road safety management. The analysis showed that accident severity was strongly influenced by factors such as over-speeding, poor lighting conditions, wet road surfaces, traffic density, vehicle type, driver negligence, road geometry, and weather conditions. Statistical modelling helped in identifying the relationship between these risk factors and accident severity, while machine learning models improved the prediction capability by detecting complex and hidden patterns in accident data. Among the applied models, Random Forest performed better than Logistic Regression, Decision Tree, K-Nearest Neighbour, and Support Vector Machine. It achieved the highest accuracy because it handled nonlinear relationships and multiple accident-related variables more effectively. The results indicated that machine learning-based prediction models can be useful for identifying accident-prone areas, classifying severity levels, and supporting timely decision-making by traffic authorities. Overall, the study proved that the integration of machine learning and statistical analysis can support preventive road safety planning, black-spot identification, emergency response improvement, and policy development. Such a prediction system can help reduce road accident fatalities, improve traffic control strategies, and promote safer transportation systems. Therefore, data-driven accident severity prediction is an important tool for modern road safety management.

REFERENCES

1. Kotsyubynska, Y., Kozan, N., Chadiuk, V., Kostyshyn, A., Kotsyubynsky, A., & Fentsyk, V. (2026). Machine Learning and Deep Learning for Predicting Traffic Crash Injury Severity: A Systematic Review and Meta-Analysis (2014-2025). *Journal of Road Safety*, *37*(1).
2. Tan, Z., Cui, L., Xu, H., Xu, J., Feng, H., & Wang, P. (2026). A hybrid machine learning approach for predicting traffic accident collision severity. *International Journal of Injury Control and Safety Promotion*, 1-15.
3. Hamdan, N., & Sipos, T. (2025). Advancements in machine learning for traffic accident severity prediction: A comprehensive review. *Periodica Polytechnica Transportation Engineering*, *53*(3), 347-355.
4. Tang, J., Huang, Y., Liu, D., Xiong, L., & Bu, R. (2025). Research on traffic accident severity level prediction model based on improved machine learning. *Systems*, *13*(1), 31.
5. Kodepogu, K., Manjeti, V. B., & Siriki, A. B. (2023). Machine learning for road accident severity prediction. *Mechatron. Intell. Transp. Syst*, *2*, 211-226.
6. Cicek, E., Akin, M., Uysal, F., & Topcu Aytas, R. (2023). Comparison of traffic accident injury severity prediction models with explainable machine learning. *Transportation letters*, *15*(9), 1043-1054.

7. Prajapati, G., Kumar, L., & Patil, S. R. S. (2023). Road accident prediction using machine learning. *Journal of Scientific Research and Technology*, 48-59.
8. Kushwaha, M., & Abirami, M. S. (2021, September). Comparative analysis on the prediction of road accident severity using machine learning algorithms. In *International Conference on Micro-Electronics and Telecommunication Engineering* (pp. 269-280). Singapore: Springer Nature Singapore.
9. Paul, J., Jahan, Z., Lateef, K. F., Islam, M. R., & Bakchy, S. C. (2020, December). Prediction of road accident and severity of Bangladesh applying machine learning techniques. In *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)* (pp. 1-6). IEEE.
10. Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, *2*(9), 1576.