Vol 6, Issue 3, March 2026 www.ijesti.com E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

Machine Learning and GIS for Predicting and Analyzing Road Traffic Accident Severity

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

Road traffic accidents (RTAs) pose a significant global public health challenge, leading to substantial fatalities and injuries. Recent advancements in machine learning (ML) have facilitated better prediction and analysis of accident severity, addressing the limitations of traditional statistical methods. ML models, including Random Forest and deep learning techniques, outperform classical approaches by handling complex, nonlinear datasets. Moreover, GIS-based frameworks have enhanced the spatial analysis of accident hotspots, offering targeted intervention strategies. Despite these advances, challenges like real-time data integration and model interpretability remain, requiring further research to improve predictive accuracy and system adaptability.

Keywords: *Traffic Accidents, Machine Learning, GIS, Prediction.*

I. INTRODUCTION

Road traffic accidents (RTAs) represent one of the most significant global public health and transportation safety challenges, contributing to substantial human fatalities, injuries, and economic losses every year. According to global safety assessments, RTAs rank among the leading causes of death, particularly affecting low- and middle-income countries where rapid urbanization, increasing vehicle density, and inadequate road safety infrastructure intensify accident risks. The severity of these accidents is not only influenced by human error but also by a complex interaction of environmental, vehicular, infrastructural, and behavioral factors. In recent years, the growing availability of large-scale traffic datasets and advancements in computational intelligence have encouraged researchers to adopt machine learning (ML) and statistical modeling techniques for analyzing and predicting accident severity. These approaches provide data-driven insights that support policymakers, traffic engineers, and emergency response systems in developing effective accident prevention and mitigation strategies (Al-Mashagba et al., 2026; Aruna et al., 2026). Traditionally, road safety analysis relied on descriptive statistics and classical regression models that focused on identifying general trends and correlations among accident-related variables. However, these methods often failed to capture the nonlinear and complex interactions among contributing factors such as weather conditions, road geometry, driver behavior, traffic density, and vehicle characteristics. With the emergence of artificial intelligence, machine learning models have demonstrated superior capability in handling high-dimensional datasets, managing nonlinearity, and improving predictive accuracy. For instance, decision tree-based models, ensemble learning techniques, and neural networks have been widely used to classify accident severity into categories such as slight, moderate, severe, and fatal. Studies have shown that Random Forest, AdaBoost, and deep learning-based architectures consistently outperform traditional statistical methods in predicting accident outcomes due to their robustness and ability to handle imbalanced datasets (Manzoor et al., 2021; Wubineh et al., 2023). Recent advancements in machine learning applications for traffic accident analysis have also introduced hybrid and deep learning frameworks that integrate spatial, temporal, and contextual features. Li et al. (2025) emphasized that conventional models often overlook the spatiotemporal heterogeneity of traffic

accidents, limiting their predictive performance. To address this limitation, they proposed a Spatio-Temporal RiskFormer model that combines ConvLSTM networks with Transformer architectures to effectively capture both sequential dependencies and semantic information such as driver demographics and vehicle attributes. Their findings highlighted that incorporating spatial and temporal dependencies significantly enhances prediction accuracy and provides deeper insights into accident causation patterns. Similarly, GIS-based machine learning frameworks have been introduced to identify accident hotspots and spatial clustering of high-risk zones, enabling authorities to implement targeted road safety interventions (Al-Mashagba et al., 2026).

In addition to deep learning models, ensemble learning methods have shown significant promise in improving classification performance for accident severity prediction. For example, Random Forest and AdaBoost models have been widely applied due to their ability to reduce overfitting and improve generalization. Pourroostaei Ardakani et al. (2023) demonstrated that ensemble-based approaches outperform standalone classifiers such as Naïve Bayes and logistic regression when applied to large-scale traffic datasets. Similarly, Türker and Gündüz (2023) found that Decision Tree models achieved remarkably high accuracy levels in predicting crash severity, highlighting the importance of feature selection and dataset characteristics in determining model effectiveness. These findings suggest that while advanced deep learning models are gaining popularity, traditional machine learning algorithms still play a crucial role in interpretable and efficient accident prediction systems. Another important aspect of road traffic accident analysis involves addressing the issue of class imbalance in datasets, where fatal and severe accidents are often underrepresented compared to minor accidents. This imbalance can significantly affect model performance and lead to biased predictions. To overcome this challenge, researchers have widely adopted data preprocessing techniques such as Synthetic Minority Oversampling Technique (SMOTE), normalization, encoding, and outlier removal. Aruna et al. (2026) applied SMOTE in their study to balance accident severity classes, resulting in improved predictive accuracy across Random Forest, Decision Tree, and Logistic Regression models. Similarly, Rezashoar et al. (2025) highlighted that even simple probabilistic models like Naïve Bayes can achieve competitive performance when properly handling imbalanced datasets, demonstrating the importance of data preprocessing in machine learning pipelines. The integration of machine learning with Geographic Information Systems (GIS) has further enhanced the spatial dimension of accident analysis. GIS-based models allow researchers to visualize accident hotspots, identify high-risk corridors, and analyze spatial patterns of traffic incidents. Al-Mashagba et al. (2026) utilized kernel density estimation within a GIS framework to map severe and fatal crash locations, providing valuable insights for urban planners and transportation authorities. This spatial intelligence enables proactive decision-making in road design, traffic regulation, and emergency response planning. Moreover, the combination of spatial analytics with predictive modeling has opened new avenues for smart transportation systems and intelligent traffic management frameworks. Despite significant advancements, several challenges remain in the field of road traffic accident severity prediction. One major limitation is the lack of real-time data integration in most existing models, as many studies rely on historical datasets that may not fully represent current traffic dynamics. Additionally, the interpretability of complex deep learning models remains a critical concern, particularly in safety-critical applications where decision transparency is essential. Furthermore, differences in geographic, infrastructural, and behavioral conditions across regions often limit the generalizability of developed models. As highlighted by Li et al. (2025) and Manzoor et al. (2021), future research must focus on developing adaptive and explainable AI models that can generalize across diverse traffic environments while maintaining high predictive performance.

II. RESEARCH BACKGROUND

Al-Mashagba et al., (2026) investigated road traffic injuries as a significant public health issue in low- and middle-income countries, particularly Jordan. They developed a hybrid machine learning and Geographic Information Systems (GIS)–based framework to predict crash severity and identify behavioral, environmental, and infrastructural factors influencing injury outcomes. A dataset of 11,345 crashes from the Jordan Traffic Institute (2018) was analyzed, where data preprocessing included encoding, imputation, normalization, and outlier treatment. They applied a two-stage analytical approach: first, association rule mining (Apriori; minimum support 0.05, confidence 0.60) was employed to reveal dominant behavioral and environmental patterns associated with varying injury levels. Subsequently, Decision Tree, Random Forest, and AdaBoost models were trained using a 70/30 data split to classify crash severity into slight, medium, severe, and fatal categories. Model performance was evaluated through accuracy, precision, recall, F1-score, and confusion matrices, while GIS-based kernel density estimation was utilized to identify spatial hotspots of severe and fatal crashes.

Aruna et al. (2026) investigated road accidents as a significant public safety concern worldwide, highlighting their role in causing substantial loss of life, injuries, and economic damage annually. They proposed a data-driven system for predicting accident severity, aiming to assist authorities in implementing preventive measures, optimizing emergency response, and enhancing traffic management. The study analyzed historical accident data to identify patterns influencing severity, employing advanced data science and machine learning techniques to process factors such as weather conditions, road type, temperature, visibility, lighting, and speed limits. Data preprocessing methods, including cleaning, normalization, and categorical encoding, were applied to prepare the dataset, while the Synthetic Minority Oversampling Technique (SMOTE) was used to address class imbalance. Predictive models were developed using Random Forest, Decision Tree, and Logistic Regression algorithms, and feature importance analysis was conducted to identify critical severity determinants. The models were evaluated using accuracy, precision, recall, and F1-score, demonstrating the system's potential to inform traffic management strategies, reduce fatalities, and improve road safety planning through intelligent predictive analytics.

Rezashoar et al., (2025) investigated the factors affecting the severity of road traffic accidents on UK roads, noting their substantial societal impact and high annual fatality rates. The study examined the predictive capabilities of several machine learning models using a Kaggle dataset comprising approximately 252,000 records and 32 independent variables spanning 2010–2014, with all algorithms implemented in Python. The authors reported that the Naïve Bayes classifier achieved the highest accuracy at 75.10%, slightly outperforming the Support Vector Machine and Neural Network models, which recorded accuracies of 75.05% and 74.59%, respectively. They highlighted that the Naïve Bayes model handled data imbalance more effectively, producing an F1 score of 45.49% compared to 42.87% for both SVM and Neural Network. Furthermore, ROC-AUC analysis emphasized the model's superiority, yielding a score of 51.10% against 50% for the others. The training and testing process completed in three seconds, indicating the method's efficiency, and the study was suggested to provide valuable insights for data-driven traffic safety strategies.

Li et al., (2025) examined the challenges posed by traffic accidents to public safety and emphasized the importance of accurately predicting their causation and severity for traffic safety research. They noted that traditional machine learning approaches largely relied on direct traffic records, which neglected the spatial and temporal heterogeneity inherent in traffic accidents. The authors argued that although existing deep learning algorithms could extract spatiotemporal features effectively, they often overlooked

semantic factors such as driver demographics and vehicle attributes, and failed to distinguish between different severity levels. To overcome these limitations, they developed the Spatio-Temporal RiskFormer (ST-RiskFormer), which combined ConvLSTM for capturing spatiotemporal correlations with a Transformer-based module for processing semantic features. Grad-CAM was applied to visualize the model's outputs. Simulation results indicated that medium-term traffic data revealed critical spatiotemporal accident patterns, and ST-RiskFormer outperformed conventional machine learning models in predicting accident severity, offering insights for deeper-level traffic data mining.

Pourroostaei Ardakani et al., (2023) investigated the severe risks posed by traffic accidents, noting their significant contribution to global fatalities and emphasizing the importance of reducing incidents to promote sustainable urban communities. They argued that machine learning and data analysis techniques could be utilized to interpret accident causes and propose mitigation strategies, while also stressing the necessity of leveraging big data due to the increasing size and velocity of traffic accident information. The study explored patterns in road accident datasets and developed a predictive model by analyzing relevant features such as accident severity, casualty numbers, and vehicle involvement. A pre-processing framework was designed to clean and standardize the raw data through missing and irrelevant feature removal, attribute generalization, and outlier elimination using interquartile methods. Four classification algorithms—decision trees, random forest, multinomial logistic regression, and naïve Bayes—were applied and evaluated, revealing acceptable predictive accuracy for all methods except naïve Bayes. The authors concluded with data-driven discussions identifying key accident factors and suggested strategies to foster safer, community-oriented urban environments.

Wubineh et al., (2023) investigated road traffic accidents (RTAs), defined as vehicle accidents occurring on a route involving collisions with other vehicles, pedestrians, animals, or fixed obstructions. They highlighted that RTAs ranked as the sixth leading cause of death in Ethiopia, resulting in approximately 31,564 fatalities annually. The study aimed to analyze road accident data to predict accident severity using machine learning techniques. A dataset of 7,744 records with six attributes, including the final class, was employed, and the authors applied SMOTE to balance the data. The dataset was split into training (80%) and testing (20%) subsets. The study implemented artificial neural networks (ANN), random forest (RF), decision tree (DT), and support vector machine (SVM) algorithms. Reported accuracies were 90% for ANN, 89% for RF, 87% for DT, and 89% for SVM, with ANN slightly outperforming the others. The authors concluded that machine learning approaches showed promising capability in predicting accident severity.

Türker and Gündüz (2023) investigated the severity of traffic accidents, emphasizing its critical impact on human life and the importance of predicting accident severity for prevention and safe driving. They highlighted that identifying relevant risk factors through predictive methods could facilitate the implementation of targeted countermeasures. The study acknowledged that traffic accidents resulted from a wide array of variables, including road disturbances, weather conditions, and vehicle speed, and noted the challenge of obtaining real-time information about accident situations. To address this, they applied machine learning algorithms to a dataset of traffic accidents collected in US states between 2016 and 2023, aiming to predict accident severity. Comparative analyses were performed using Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes, with performance evaluated via Precision, Recall, and F1-Score metrics. The findings indicated that the Decision Tree algorithm achieved the highest classification accuracy at 99.6%, demonstrating the effectiveness of machine learning in mitigating traffic accidents through risk-informed interventions.

Manzoor et al. (2021) investigated traffic accidents on highways, noting that despite advancements in traffic safety measures, such accidents remained a leading cause of death, particularly burdening developing countries. They highlighted that multiple factors influenced accident severity, with some exerting stronger effects than others. The study employed data mining techniques to identify influential variables, using Random Forest to determine significant features correlated with accident severity, such as distance, temperature, wind chill, humidity, visibility, and wind direction. Furthermore, the authors proposed an ensemble model combining Random Forest and Convolutional Neural Network, termed RFCNN, to predict road accident severity. They evaluated the performance of RFCNN against several base learner classifiers, using accident records from the USA spanning February 2016 to June 2020. The results demonstrated that RFCNN improved decision-making processes and outperformed other models, achieving an accuracy of 0.991, precision of 0.974, recall of 0.986, and F-score of 0.980 when employing the twenty most significant features for severity prediction.

Jadhav et al. (2020) highlighted that road accidents had emerged as a significant global concern, ranking as the ninth leading cause of death worldwide. They observed that the high frequency of road accidents each year had made the issue particularly severe in their country, emphasizing the urgent need for intervention to prevent further casualties. To address this challenge, they argued that a detailed and precise analysis was necessary. Their study applied machine learning approaches to examine traffic accidents in depth, aiming to assess accident severity and identify the critical factors influencing such incidents. They reported that Deep Learning Neural Networks and AdaBoost, as supervised learning techniques, had been employed to classify accidents into four categories: Fatal, Grievous, Simple Injury, and Motor Collision. Furthermore, they suggested that the insights gained from their analysis could inform beneficial strategies and recommendations for mitigating the impact of road accidents at a national level.

Labib et al. (2019) highlighted that road accidents had emerged as a significant global issue, ranking as the ninth leading cause of death worldwide, with Bangladesh experiencing a particularly high incidence. They emphasized that the loss of lives due to traffic accidents in the country was both unacceptable and alarming, necessitating urgent measures for mitigation. To address this challenge, the study undertook a detailed analysis of traffic accidents to assess their severity and identify influential factors using machine learning approaches. The researchers applied four supervised learning techniques—Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and AdaBoost—to classify accident severity into Fatal, Grievous, Simple Injury, and Motor Collision categories. The study found that AdaBoost delivered the highest predictive performance among the methods tested. Furthermore, the research provided insights into the key determinants of road accidents and suggested practical recommendations aimed at reducing traffic-related fatalities and injuries in Bangladesh.

III. KEY FINDINGS FROM STUDY

Author (Year)	Study Focus	Methodology / Models Used	Dataset	Key Findings
Al-Mashagba et al. (2026)	Crash severity prediction and GIS-based spatial analysis	Apriori algorithm, Decision Tree, Random Forest, AdaBoost, GIS (Kernel Density Estimation)	11,345 crash records (Jordan Traffic Institute)	Behavioral, environmental, and infrastructural factors significantly influence severity; GIS identified accident hotspots effectively
Aruna et al.	Accident	Random Forest,	Historical	Weather, road type,

(2026)	severity prediction using data science	Decision Tree, Logistic Regression + SMOTE	accident dataset	visibility, and speed limits strongly affect severity; SMOTE improved class balance and accuracy
Rezashoar et al. (2025)	Comparative ML analysis for crash severity	Naïve Bayes, SVM, Neural Network	UK road accident dataset (2010–2014)	Naïve Bayes achieved highest accuracy (75.10%) and handled imbalance effectively
Li et al. (2025)	Spatio-temporal deep learning for severity prediction	ConvLSTM + Transformer (ST-RiskFormer)	Traffic accident datasets	Hybrid model outperformed traditional ML by capturing spatial, temporal, and semantic dependencies
Pourroostaei Ardakani et al. (2023)	ML-based accident pattern analysis	Decision Tree, Random Forest, Logistic Regression, Naïve Bayes	Traffic accident dataset	Ensemble models provided better accuracy; Naïve Bayes performed comparatively lower
Wubineh et al. (2023)	Accident severity prediction in Ethiopia	ANN, Random Forest, Decision Tree, SVM + SMOTE	7,744 accident records (Addis Ababa)	ANN achieved highest accuracy (90%), followed by RF and SVM
Türker & Gündüz (2023)	ML-based crash severity prediction	Decision Tree, Random Forest, Logistic Regression, SVM, KNN, Naïve Bayes	US accident dataset (2016–2023)	Decision Tree achieved highest accuracy (99.6%)
Manzoor et al. (2021)	Hybrid deep learning model for severity prediction	Random Forest + CNN (RFCNN ensemble)	USA accident dataset (2016–2020)	Proposed model achieved 99.1% accuracy with high precision and recall
Jadhav et al. (2020)	Accident severity classification	Deep Neural Networks, AdaBoost	Road accident dataset	Successfully classified accidents into severity levels; ensemble methods improved performance
Labib et al. (2019)	ML-based traffic accident severity analysis	Decision Tree, KNN, Naïve Bayes, AdaBoost	Bangladesh accident dataset	AdaBoost outperformed other models in predicting severity

IV. CONCLUSION

The reviewed literature clearly demonstrates that road traffic accident (RTA) severity prediction has evolved significantly with the integration of machine learning (ML), statistical modeling, and hybrid artificial intelligence techniques. Traditional statistical approaches, although useful in identifying basic relationships among accident variables, are limited in capturing the complex, nonlinear interactions between behavioral, environmental, infrastructural, and vehicular factors. In contrast, modern ML-based

approaches such as Decision Trees, Random Forest, Support Vector Machines, Artificial Neural Networks, and ensemble learning methods have shown superior capability in handling large, imbalanced, and high-dimensional datasets for accident severity classification. Across the reviewed studies, it is evident that ensemble learning techniques such as Random Forest and AdaBoost consistently outperform single classifiers due to their robustness, higher generalization ability, and reduced overfitting. Hybrid deep learning models, such as CNN-based ensembles and spatiotemporal architectures like ConvLSTM-Transformer frameworks, have further improved predictive performance by capturing hidden spatial, temporal, and contextual patterns in accident data. For example, studies such as Manzoor et al. (2021) and Li et al. (2025) demonstrated that integrating deep learning with traditional ML significantly enhances prediction accuracy and reliability. Furthermore, the incorporation of Geographic Information Systems (GIS) has strengthened the spatial understanding of accident distribution by identifying high-risk zones and crash hotspots, enabling more targeted road safety interventions. Similarly, data preprocessing techniques such as normalization, encoding, outlier removal, and SMOTE-based class balancing have been crucial in improving model performance and reducing bias toward majority classes. Overall, ML-based frameworks have proven highly effective in improving road safety analysis, enabling more accurate prediction of accident severity levels, and supporting evidence-based transportation planning and decision-making. Despite these advancements, the reviewed studies also indicate variability in model performance depending on dataset characteristics, feature selection techniques, and regional differences in traffic behavior. This highlights the need for more generalized, adaptive, and interpretable models that can perform consistently across diverse traffic environments. Additionally, while deep learning models provide high accuracy, their lack of transparency remains a limitation in safety-critical applications where explainability is essential.

V. FUTURE SCOPE

- **Integration of Real-Time Traffic Data Systems:** Future research should focus on integrating real-time data from IoT sensors, GPS devices, smart cameras, and connected vehicles. This will enable dynamic accident severity prediction and support proactive traffic management systems instead of relying solely on historical data.
- **Development of Explainable AI (XAI) Models:** Although deep learning models provide high accuracy, their interpretability is limited. Future studies should incorporate explainable AI techniques such as SHAP, LIME, and attention mechanisms to improve transparency and help policymakers understand decision-making processes.
- **Hybrid Spatio-Temporal Modeling Frameworks:** More advanced hybrid models combining spatial, temporal, behavioral, and environmental features should be developed. Integration of GIS, deep learning, and graph-based neural networks can significantly improve prediction accuracy and spatial understanding of accident patterns.
- **Cross-Regional and Generalized Model Development:** Most existing studies are region-specific, limiting their applicability. Future research should focus on developing generalized models that can be trained and tested across multiple geographical regions to ensure better scalability and robustness.
- **Improved Handling of Imbalanced and Noisy Data:** Advanced data augmentation techniques, synthetic data generation, and improved sampling strategies beyond SMOTE should be explored to handle imbalanced accident severity datasets more effectively.

- **Integration with Intelligent Transportation Systems (ITS):** ML-based accident prediction models should be integrated into intelligent transportation systems for real-time traffic control, emergency response optimization, and adaptive traffic signal management.
- **Incorporation of Human Behavior and Psychological Factors:** Future models should include driver behavior, fatigue, distraction, and psychological parameters, which are often overlooked but play a critical role in accident severity outcomes.

REFERENCES

1. Al-Mashagba, L., Sumari, P., Mashagba, M., Al Abri, F., Mashagba, H. A., Abd Aziz, A. B., & Al-Bawri, S. S. (2026). Machine learning and Geographic Information Systems (GIS)–based model for road crash severity prediction and behavioural pattern analysis. *Traffic Injury Prevention*, 1-10.
2. ARUNA, D., LALITHA, G., ISWARYA, V., KUMAR, J. V., & DEVANADH, P. (2026). ROAD ACCIDENT SEVERITY PREDICTION USING DATA SCIENCE. *International Journal of Data Science and IoT Management System*, *5*(1), 490-500.
3. Rezashoar, S., Kashi, E., & Saeidi, S. (2025). Comparison of machine learning algorithms for predicting traffic accident severity (case study: United Kingdom from 2010 to 2014). *International Journal of Crashworthiness*, 1-10.
4. Li, K., Guo, K., Liang, Z., Xu, H., & Duan, X. (2025). Traffic accident severity prediction and analysis via spatio-temporal deep learning: A ConvLSTM-transformer approach. *Chaos, Solitons & Fractals*, *200*, 117047.
5. Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability*, *15*(7), 5939.
6. Wubineh, B. Z., Asamenew, Y. A., & Kassa, S. M. (2023, December). Prediction of Road Traffic Accident Severity Using Machine Learning Techniques in the Case of Addis Ababa. In *International Conference on Robotics and Networks* (pp. 129-144). Cham: Springer Nature Switzerland.
7. Türker, G. F., & Gündüz, F. K. (2023). A study on traffic crash severity prediction using machine learning algorithms. *Uluslararası Sürdürülebilir Mühendislik ve Teknoloji Dergisi*, *7*(2), 152-161.
8. Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H. A., & Bisogni, C. (2021). RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. *IEEE Access*, *9*, 128359-128371.
9. Jadhav, A., Jadhav, S., Jalke, A., & Suryavanshi, K. (2020). Road accident analysis and prediction of accident severity using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *7*(12), 1-8.
10. Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019, June). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In *2019 7th international conference on smart computing & communications (ICSCC)* (pp. 1-5). IEEE.