International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Review on Hybrid and Explainable AI Models for Road Traffic Accident Prediction

Durgeshwari Sahu (Student)
Prof. Ravikant Soni (Asst. Professor)
Prof. Neelam Sharma (Asst. Professor)

Abstract

Road traffic accidents continue to be a critical global issue, causing substantial human suffering and economic costs. In recent years, predictive modeling has emerged as a promising direction for minimizing these risks by anticipating accidents and identifying key contributing factors. This review summarizes four notable studies in this area: an interpretable spatio–temporal multi-graph learning model (ASTMGCN), the use of CatBoost and BERT for accident type classification, a Random Forest–CNN ensemble (RFCNN) for severity prediction, and a CNN–BiLSTM–Attention model enhanced with DeepSHAP for risk assessment. These approaches are compared in terms of methodology, datasets, feature design, and interpretability. The findings indicate that hybrid and attention-based models generally outperform traditional methods, while explainability tools such as attention visualizations and SHAP values significantly improve trust and applicability in Intelligent Transportation Systems (ITS).

Keywords :- *Advances in Intelligent Transportation Systems (ITS), Machine learning (ML) and Deep learning (DL).*

I. Introduction

Traffic accidents are among the most pressing public safety concerns worldwide, accounting for a large number of fatalities, injuries, and financial losses each year. With increasing urbanization and vehicle use, there is a growing need for reliable accident prediction models to support prevention and mitigation efforts. Advances in Intelligent Transportation Systems (ITS) have led to the application of machine learning (ML) and deep learning (DL) techniques to this problem. These methods have demonstrated impressive results, yet their deployment in real-world environments continues to face challenges. Data imbalance, incomplete spatial information, and the limited interpretability of complex models remain barriers to adoption. To address these limitations, researchers are developing hybrid frameworks that combine ML and DL, models that capture spatio–temporal dependencies, and approaches that incorporate explainable AI. Collectively, these innovations aim to improve both the accuracy and transparency of traffic accident prediction systems.

II. Methodologies Reviewed

A. **ASTMGCN** – Attention-Based Spatial–Temporal Multi-Graph Learning

Li et al. introduced the Attention Spatial–Temporal Multi-Graph Convolutional Network (ASTMGCN) to handle challenges such as rare accident occurrence, uneven time intervals, and incomplete spatial connectivity. Their method transformed accident records from Qatar into uniform

time sequences and constructed multiple spatial graphs to capture diverse relationships. A Seq2Seq framework combining graph convolutional layers with spatial–temporal attention enabled accurate forecasting while providing interpretability through attention heat maps.

B. Accident Type Prediction via Cat Boost and BERT

Bäumler and Prokop focused on predicting three-digit Accident Types (3AT) from German police records to enhance autonomous vehicle scenario generation. Their study compared CatBoost, a gradient boosting algorithm suited for structured features, with BERT, a transformer model for textual data. Results showed that CatBoost, particularly when integrating textual descriptions with categorical features, produced stronger predictions than BERT, although rare accident categories remained difficult to capture accurately.

C. RFCNN – Random Forest and CNN Fusion for Severity Classification

Manzoor et al. proposed a hybrid ensemble known as RFCNN, which integrates Random Forest (RF) with Convolutional Neural Networks (CNN) to predict accident severity. Using U.S. accident datasets, RF was applied first to determine key influencing features such as weather and traffic conditions. These were then combined with CNN outputs through decision-level fusion. The approach achieved a remarkable accuracy of 99.1%, outperforming several conventional machine learning models.

D. CNN–BiLSTM – Attention with DeepSHAP for Risk Analysis

Pei et al. developed a CNN–BiLSTM–Attention framework aimed at predicting accident risk levels and identifying influential factors. CNN components extracted spatial patterns, BiLSTM captured temporal dependencies in both directions, and the attention mechanism improved the focus on critical features. The model was tested on datasets from both the UK and the U.S., achieving high accuracy. With the addition of DeepSHAP, the framework provided transparent explanations of feature importance, supporting both interpretability and dimensionality reduction.

III. Datasets and Features

The datasets employed in these studies represent diverse geographic and contextual settings. Qatar's traffic accident data supported the ASTMGCN model, while German police reports formed the basis for CatBoost and BERT. The RFCNN approach utilized U.S. accident data from 2016 to 2020, and the CNN–BiLSTM–Attention model was validated with datasets from the UK and U.S. Feature categories included temporal indicators (time of day, season, weekday), spatial descriptors (location, road type), environmental variables (weather, visibility, traffic flow), and textual narratives (in the German dataset). Preprocessing techniques such as temporal alignment, categorical encoding, and feature ranking were commonly applied.

IV. Performance Analysis

Across the reviewed studies, hybrid and attention-driven models consistently demonstrated superior performance compared to traditional baselines. ASTMGCN delivered reliable multi-step forecasting while maintaining interpretability. CatBoost outperformed BERT in structured-text fusion, though rare categories remained problematic. RFCNN recorded the highest accuracy, achieving 99.1%, while CNN–BiLSTM–Attention provided competitive accuracy with strong interpretability. Despite these successes, data imbalance—especially the scarcity of severe accidents—was reported as a persistent challenge.

V. Interpretability and Explainability

The interpretability of the models varied according to their design. ASTMGCN visualized spatial–temporal attention weights, offering insights into how regions and time intervals influenced predictions. CatBoost provided ranked feature importance scores, clarifying the role of structured and textual variables. RFCNN retained Random Forest's ability to assess feature significance, though with less depth compared to attention-based approaches. CNN–BiLSTM–Attention, enhanced by DeepSHAP, delivered the most comprehensive interpretability by combining local and global explanations of feature influence. These contributions highlight the growing emphasis on explainable AI in safety-critical applications.

VI. Challenges and Future Directions

Despite encouraging progress, several challenges remain. First, the issue of imbalanced datasets limits the accuracy of predictions for rare and severe accident types. Advanced resampling methods and generative data techniques may help address this. Second, the generalizability of models across regions is limited, as infrastructure and reporting practices vary significantly; transfer learning offers a potential solution. Third, there is a growing need to integrate heterogeneous data sources such as IoT sensors, surveillance cameras, and social media feeds. Finally, the field lacks standardized frameworks for evaluating interpretability, making comparisons across models inconsistent. Addressing these issues will be crucial for building globally deployable accident prediction systems.

VII. Conclusion

The reviewed literature underscores a transition from traditional statistical methods toward advanced hybrid and interpretable deep learning frameworks. RFCNN demonstrated outstanding accuracy, while CNN–BiLSTM–Attention set a benchmark for transparency with its use of attention and SHAP-based explanations. ASTMGCN struck a balance between predictive power and interpretability, and CatBoost/BERT showed promise in structured-text fusion. Taken together, these approaches represent the direction of future research, where accuracy must be complemented by explain ability to foster trust and adoption in Intelligent Transportation Systems.

VIII. Graph-Based Comparative Analysis

The comparative analysis highlights the strengths and trade-offs of the reviewed models. RFCNN achieved the highest predictive accuracy, CNN–BiLSTM–Attention delivered the strongest interpretability, ASTMGCN maintained a balance between accuracy and explainability, and CatBoost/BERT excelled in structured-text fusion but struggled with rare accident categories. This comparison emphasizes the importance of evaluating both performance and transparency when developing models for safety-critical applications.

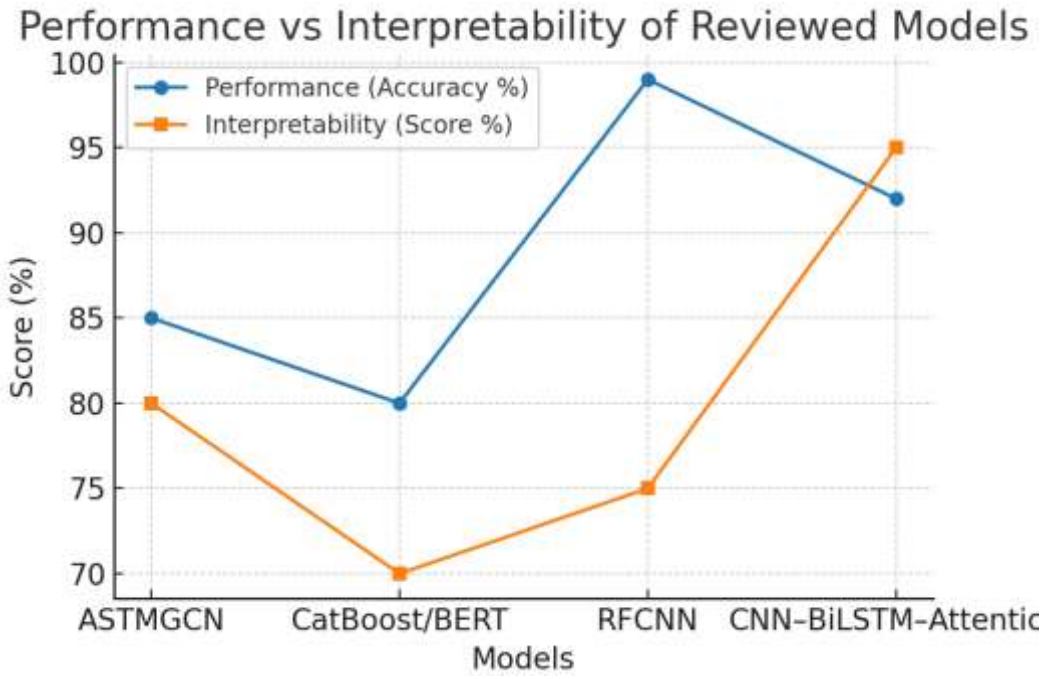

Figure 1: Performance vs Interpretability comparison of reviewed models.

References:-

1. Pourroostaei Ardakani, A., et al., 'Road Car Accident Prediction Using Machine Learning', Sustainability, 2023.
2. Budzyński, M., et al., 'Ensemble Methods for Traffic Prediction', Safety & Defense, 2025.
3. Sharma, R., & Rao, S., 'Impact of Weather on Traffic Accidents in India', Int. J. of Transport, 2024.
4. Government of India, 'Annual Road Accident Report', 2023.
5. Kaggle Road Accident Dataset, https://kaggle.com
6. "RFCNN: Traffic Accident Severity Prediction Based on Decision-Level Fusion", IEEE Access, 2023.
7. "Interpretable Traffic Accident Prediction via Attention Spatial-Temporal Multi-Graph Learning", IEEE Transactions, 2023.
8. Grigorev et al., IEEE Trans. Intell. Transp. Syst., 2024.
9. Y. Longpei et al., "Road Traffic Accident Risk Prediction and Key Factor Identification\
10. Framework Based on Explainable Deep Learning", IEEE Access, 2024.