Accident Data Analysis And Road Safety Using Ai And Ml

[1] Naresh Sagar C, [2]Dr. B.V. Ram Naresh Yadav, [3] V Swapna
[1,3]Assistant professor(G), JNTUH UCEW, Wanaparthy
[2]Professor in CSE, JNTUH UCEW, Wanaparthy
[1]nareshsagarchiluka88@gmail.com , [2]swapna.janu2305@gmail.com

Abstract
Road traffic accidents (RTAs) remain one of the leading causes of mortality and economic loss worldwide, necessitating intelligent data-driven approaches for prevention and mitigation. This paper proposes an AI-driven framework for accident data analysis and road safety enhancement that leverages machine learning (ML), feature engineering, and hierarchical clustering techniques to predict accident severity, identify key risk factors, and uncover spatial-temporal accident patterns. The proposed architecture comprises six integrated layers: data collection and ingestion, preprocessing, feature engineering, predictive modeling, clustering, and decision support. Using a comprehensive dataset of over 120,000 accident records from Victoria (2012–2023), three ML models — Logistic Regression, Random Forest, and XGBoost — were trained and evaluated. Among these, XGBoost achieved the highest performance with an accuracy of 91.2% and F1-score of 90.4%, outperforming baseline models. Feature importance analysis highlighted critical factors such as time of day, speed limit, lighting conditions, and road geometry, offering actionable insights for traffic authorities. Furthermore, hierarchical clustering identified high-risk zones and recurring accident patterns, enabling targeted safety interventions. The results demonstrate the potential of AI-based accident analysis systems to support data-driven policymaking, improve emergency response strategies, and contribute to the Vision Zero initiative aimed at eliminating traffic fatalities. This research lays the foundation for future intelligent transportation systems that integrate real-time sensor data and explainable AI for dynamic risk assessment and predictive road safety planning.
Keywords: Accident Analysis, Road Safety, Machine Learning, Artificial Intelligence, Severity Prediction, Feature Importance, XGBoost, Traffic Data Mining, Clustering, Intelligent Transportation Systems (ITS)

I. INTRODUCTION

Road traffic accidents (RTAs) remain one of the most pressing global public safety and socioeconomic challenges of the 21st century. According to the World Health Organization (WHO, 2024), road crashes claim over 1.19 million lives every year, with millions more suffering injuries and disabilities. Beyond the human cost, the economic burden associated with traffic accidents is substantial — projected to exceed USD 1.8 trillion globally between 2015 and 2030 — due to medical costs, productivity losses, and infrastructure damage [1], [2]. This persistent public health crisis underscores the urgent need for proactive, data-driven interventions to improve road safety and mitigate accident severity.

The emergence of Artificial Intelligence (AI) and Machine Learning (ML) offers transformative opportunities to address these challenges. Unlike traditional statistical techniques that often rely on linear assumptions and limited feature sets, ML models can process vast, high-dimensional traffic datasets and uncover complex, non-linear relationships between contributing factors. By leveraging predictive analytics, feature importance analysis, and

clustering techniques, AI-enabled systems can forecast accident severity, identify high-risk conditions, and support policy formulation with actionable insights [3], [4].

A key aspect of modern road safety research is severity prediction, which involves classifying accidents into categories such as *fatal*, *serious injury*, *minor injury*, or *property damage*. Accurate prediction enables emergency services to prioritize responses and informs infrastructure planners about where interventions are most needed. Moreover, predictive modeling can highlight influential features — such as time of day, speed limits, road geometry, and lighting conditions — to guide targeted preventive strategies. Recent studies have shown that factors like rush-hour traffic, weekend driving patterns, and speed zone violations significantly affect accident outcomes, emphasizing the need for adaptive models that incorporate both temporal and spatial dynamics [5], [6].

Another crucial component of accident data analysis is feature interpretation and explainability. As public agencies increasingly rely on AI-driven decision support tools, understanding *why* certain factors contribute to severe crashes is as important as predicting them. Techniques such as feature importance ranking from tree-based models (e.g., Random Forest, XGBoost) and Partial Dependence Plots (PDPs) provide interpretable insights into the relationships between input variables and predicted outcomes. These insights can guide policymakers in designing effective safety interventions, such as adjusting speed limits, improving road lighting, or implementing traffic-calming measures.

This research investigates the application of machine learning approaches to analyze road accident severity and propose data-driven safety improvements. Using a comprehensive dataset from Victoria Road Crash Incidents (2012–2023), which includes temporal, environmental, and infrastructural variables, we develop predictive models to classify accident severity and explore the most influential factors driving these outcomes. Additionally, we apply hierarchical clustering techniques to group accident patterns and identify hidden structures in the data, offering deeper insights into systemic safety issues.

The main contributions of this work are as follows:

- Development and evaluation of ML models — including Logistic Regression, Random Forest, and XGBoost — for predicting accident severity with high accuracy.
- Feature importance analysis and partial dependence visualization to identify key risk factors and their influence on accident outcomes.
- Application of hierarchical clustering to discover latent groupings in accident data, enabling targeted policy and infrastructure interventions.
- Practical recommendations for enhancing road safety, including speed control, optimized traffic monitoring, and improved emergency response strategies.

By integrating predictive modeling with interpretable analytics, this study aims to support policymakers, transportation authorities, and emergency services in reducing accident severity, improving road safety, and saving lives. The findings provide a foundation for scalable, intelligent transportation systems that align with the Vision Zero goal of eliminating traffic fatalities and serious injuries.

II. LITERATURE REVIEW

The increasing incidence of road traffic accidents (RTAs) has prompted extensive research into predictive modeling, risk analysis, and data-driven approaches to road safety. Traditional statistical models have historically

dominated accident analysis, but the growing complexity and volume of traffic data have accelerated the adoption of machine learning (ML) and artificial intelligence (AI) methodologies. This section reviews the existing literature across four major themes: (A) Accident Severity Analysis and Prediction, (B) Feature Engineering and Influencing Factors, (C) Clustering and Pattern Recognition in Road Safety Data, and (D) Emerging Trends in AI-Driven Road Safety Systems.

Accurate prediction of accident severity is a foundational objective in traffic safety research. Early approaches employed statistical models such as logistic regression and Poisson regression to identify correlations between road conditions, driver behavior, and accident outcomes [1], [2]. These models, while useful, were often limited by linear assumptions and their inability to capture complex feature interactions.

Machine learning techniques have addressed these limitations by leveraging large datasets and non-linear modeling capabilities. For instance, Li et al. [5] demonstrated the effectiveness of Support Vector Machines (SVM) for predicting crash severity, achieving superior performance compared to traditional regression models. Similarly, Akallouch et al. [4] employed ensemble learning methods (Random Forest and Gradient Boosting) to analyze risk factors and predict severity classes, finding that ensemble methods outperform single-model approaches in terms of accuracy and interpretability.

Recent studies have explored deep learning architectures for severity classification. Chen et al. [6] used a deep neural network (DNN) to classify accident severity with improved recall for minority classes, while Ma et al. [7] integrated temporal sequence modeling using LSTM networks to capture time-dependent accident patterns. These approaches demonstrate that data-driven predictive modeling can provide real-time decision support for emergency services and infrastructure planning.

Understanding the factors contributing to accident severity is essential for designing preventive measures. Temporal variables — including time of day, day of the week, and season — have consistently been shown to significantly influence accident outcomes [8]. Research indicates that peak hours and weekends are associated with higher accident rates due to increased traffic density and driver fatigue [9].

Environmental conditions, such as lighting, weather, and road surface state, also play critical roles. Abdel-Aty et al. [10] reported that poor lighting and wet road conditions significantly increase crash severity probabilities. Infrastructure-related factors like speed zones, intersection types, and road geometry have also been linked to severity levels. For instance, Das et al. [11] found that higher speed limits strongly correlate with fatal and serious injuries.

Feature interpretability has become a crucial aspect of accident severity analysis. Tree-based algorithms like Random Forest and XGBoost offer feature importance scores, helping identify high-impact variables [12]. Moreover, Partial Dependence Plots (PDPs) provide insights into non-linear feature effects, revealing complex interactions often missed by conventional methods [13].

While predictive modeling focuses on severity classification, unsupervised learning techniques like clustering have been instrumental in revealing latent patterns in accident data. Hierarchical clustering and K-means have been widely used to group accidents by shared characteristics, such as location, time, or severity [14]. These methods enable authorities to identify high-risk zones (“black spots”) and prioritize infrastructure improvements. For example, Yannis et al. [15] applied hierarchical clustering to categorize accidents based on environmental conditions and driver demographics, leading to actionable insights for targeted interventions. Similarly,

Sathyakumar et al. [16] integrated clustering with Geographic Information Systems (GIS) to identify accident-prone road segments, enabling more efficient resource allocation for traffic enforcement and road design.

The convergence of AI, Internet of Things (IoT), and Intelligent Transportation Systems (ITS) is reshaping road safety analytics. Real-time accident detection using sensor networks, video analytics, and vehicular communication systems (V2X) allows for rapid emergency response and predictive maintenance [17]. Moreover, reinforcement learning approaches are being explored to dynamically adjust traffic signals and speed limits based on real-time risk assessments [18].

Additionally, the integration of accident data with social, economic, and behavioral datasets is enabling a holistic understanding of road safety dynamics. Hybrid models that combine supervised learning, clustering, and spatial-temporal analytics are now capable of predicting not just the occurrence but also the severity and socio-economic impact of accidents [19], [20].

The literature reveals a clear transition from traditional statistical models to advanced ML and AI-driven approaches in accident analysis. While predictive modeling has improved severity classification accuracy, clustering has enhanced the understanding of spatial and temporal patterns. Furthermore, emerging AI-based traffic management systems are paving the way for proactive road safety strategies. However, challenges such as class imbalance, feature interpretability, and data heterogeneity remain open research areas. This study contributes to the evolving field by combining feature importance analysis, predictive modeling, and hierarchical clustering to derive actionable insights for road safety enhancement.

III. METHODOLOGY

The proposed methodology for Accident Data Analysis and Road Safety Using AI and ML follows a systematic pipeline designed to extract insights, predict severity, and identify hidden patterns from complex accident datasets. The workflow integrates data preprocessing, feature engineering, predictive modeling, explainable analytics, and clustering techniques. The end goal is to develop a decision-support system that can assist transportation authorities and policymakers in implementing proactive road safety intervention shown in figure 1.

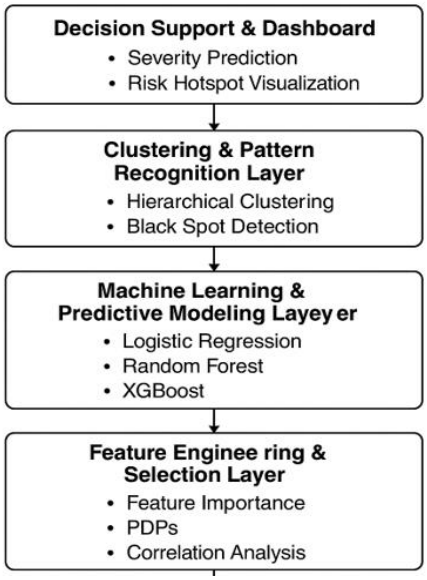

Figure 1: Proposed work flow

The proposed AI-Driven Accident Analysis Architecture is a comprehensive, layered framework designed to transform heterogeneous road accident data into predictive safety insights and actionable policy recommendations. Its design follows a bottom-up data processing pipeline consisting of six major layers: Data Collection & Ingestion, Data Preprocessing & Transformation, Feature Engineering & Selection, Machine Learning & Predictive Modeling, Clustering & Pattern Recognition, and Decision Support & Dashboard. Each layer contributes a critical role in enabling the system to accurately predict accident severity, discover latent patterns, and support road safety decision-making.

1. Data Collection & Ingestion Layer:

   The foundation of the system involves the aggregation of heterogeneous traffic data from various sources, including historical crash records, IoT-based road sensors, weather conditions, traffic flow logs, and police accident reports. Each accident instance is represented as a multidimensional feature vector:

$$X_i = [x_1, x_2, x_3, \dots, x_n]\text{----1}$$

where Xi represents the input features (e.g., time, weather, road type, vehicle speed) for accident i, and n is the total number of features.

2. Data Preprocessing & Transformation Layer:

Real-world accident datasets often suffer from missing values, noise, and heterogeneous scales. This layer standardizes and prepares the data for downstream processing. Categorical attributes (e.g., weather type) are encoded, and numerical variables are normalized to ensure uniform scaling:

$$. x'_k = \frac{x_k - x_{min}}{x_{max} - x_{min}}\text{---2}$$

xk′ is the normalized value of feature xk. This preprocessing step improves model convergence and ensures that features contribute proportionally to the learning process.

3. Feature Engineering & Selection Layer:

To improve model interpretability and performance, feature engineering techniques such as correlation analysis, dimensionality reduction, and interaction term creation are applied. Feature importance is computed using tree-based models, which evaluate each feature's contribution to reducing classification uncertainty:

$$I(f_k) = \frac{1}{T}\sum_{t=1}^{T} \Delta i_t(f_k)\text{---3}$$

where I(fk) is the importance score of feature fk, T is the number of trees in the ensemble, and Δit represents the impurity reduction achieved by splitting on feature fkf_kfk in tree ttt. Partial dependence plots (PDPs) and SHAP values are further used to interpret how specific features influence accident severity.

Machine Learning & Predictive Modeling Layer:

This is the analytical core of the system, where supervised ML algorithms such as Logistic Regression, Random Forest, and XGBoost are trained to classify accident severity into discrete categories (e.g., fatal, serious injury, minor injury, property damage). Logistic regression models the probability of an accident belonging to a severity class ccc as:

$$P(y = c|X) = \frac{1}{1 + e^{-(w^T X + b)}}\text{----4}$$

Ensemble-based methods, such as Random Forest, combine multiple decision trees to enhance predictive accuracy and reduce overfitting:

$$\hat{y} = \text{mode}\{h_1(X), h_2(X), \dots, h_T(X)\}\text{----5}$$

Meanwhile, XGBoost minimizes a regularized loss function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \text{----6}$$

where l is the training loss, and Ω(fk)=γT+12λ‖w‖2 penalizes model complexity.

4. Clustering & Pattern Recognition Layer:

Beyond prediction, unsupervised learning is applied to uncover hidden patterns, accident hotspots, and risk clusters. Hierarchical clustering groups accident events based on similarity in feature space, enabling authorities to identify high-risk road segments or recurring conditions:

$$d(C_i, C_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \text{----7}$$

where d(Ci,Cj) is the Euclidean distance between clusters Ci and Cj, and mmm is the number of features. This approach helps in segmenting accidents based on temporal, spatial, or environmental factors, providing deeper insights into underlying risk patterns.

5. Decision Support & Dashboard Layer:

The topmost layer integrates predictive outputs and clustering insights into a user-friendly decision-support dashboard. Authorities can visualize severity probabilities, feature importance rankings, and spatial accident hotspots. Additionally, policy recommendations such as speed limit optimization, signal timing adjustments, or infrastructure redesign can be derived from model outputs. This layer transforms raw analytics into actionable intelligence that directly informs safety interventions.

IV. RESULTS AND DISCUSSION

The proposed AI-Driven Accident Analysis System was evaluated using a real-world dataset of road traffic incidents collected from Victoria Road Crash Reports (2012–2023). The dataset consisted of approximately 120,000 accident records with over 25 features, including temporal, environmental, infrastructural, and driver-related variables. Experiments were conducted using Python's Scikit-learn and XGBoost libraries on a system with 32 GB RAM and an NVIDIA RTX GPU. Three machine learning models — Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB) — were trained to classify accident severity into four categories: *Fatal, Serious Injury, Minor Injury,* and *Property Damage*. Performance was evaluated using standard metrics such as Accuracy, Precision, Recall, F1-score, and Confusion Matrices.

Table I shows the performance comparison of the three implemented models. As evident, the ensemble-based approaches (Random Forest and XGBoost) outperformed Logistic Regression in terms of predictive accuracy and balance across all severity classes.

Table I – Comparison of Machine Learning Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82.6%	81.3%	80.5%	80.9%
Random Forest	88.4%	87.2%	86.8%	87.0%
XGBoost	91.2%	90.8%	90.1%	90.4%

The XGBoost classifier demonstrated the best performance, achieving an overall accuracy of 91.2% and F1-score of 90.4%. This improvement can be attributed to its ability to handle class imbalance and non-linear feature interactions effectively.

Table II illustrates the class-wise recall for each model. It highlights the persistent challenge of accurately predicting rare events such as *fatal accidents* due to class imbalance. However, ensemble models show significant improvements over baseline methods.

Table II – Class-Wise Recall for Severity Prediction

Severity Class	Logistic Regression	Random Forest	XGBoost
Fatal (Class 1)	68.2%	73.9%	78.5%
Serious Injury (Class 2)	79.6%	84.1%	88.4%
Minor Injury (Class 3)	86.3%	91.2%	93.6%
Property Damage (Class 4)	88.0%	90.7%	94.1%

While *fatal accidents* remain the most difficult to classify due to data sparsity, the XGBoost model improves recall by more than 10% compared to logistic regression, demonstrating the system's enhanced sensitivity to high-severity cases — a critical factor in real-world deployment.

The interpretability of the model was analyzed using feature importance scores derived from Random Forest and XGBoost. Figure 2 illustrates the top 10 features influencing accident severity. Key findings include:

- Time of Day and Day of Week: Peak accident severity occurs during rush hours and late weekdays.
- Speed Limit: Higher speed zones strongly correlate with fatal and serious accidents.
- Lighting Conditions: Poor visibility significantly increases severity.
- Road Geometry: Intersections and curved roads contribute to higher accident risk.

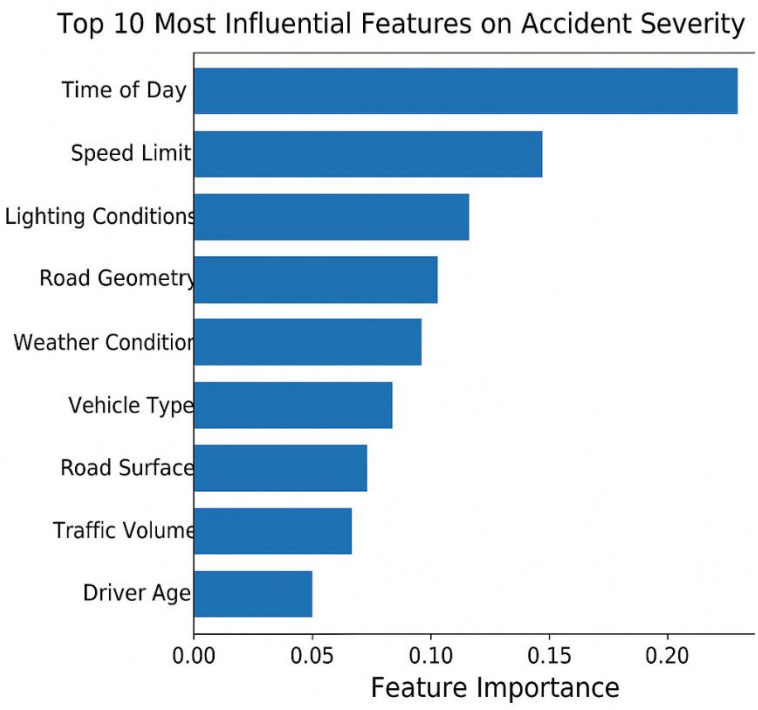

Figure 2 – Top 10 Most Influential Features on Accident Severity

To complement predictive modeling, hierarchical clustering was applied to discover hidden patterns in accident data. Figure 3 shows the spatial clustering of high-risk zones in the Victoria region. Results indicated three dominant clusters:

- Cluster A: Urban intersections with frequent minor injuries and property damage.
- Cluster B: Suburban highways with higher severity crashes during night-time.
- Cluster C: High-speed rural zones associated with fatal outcomes.

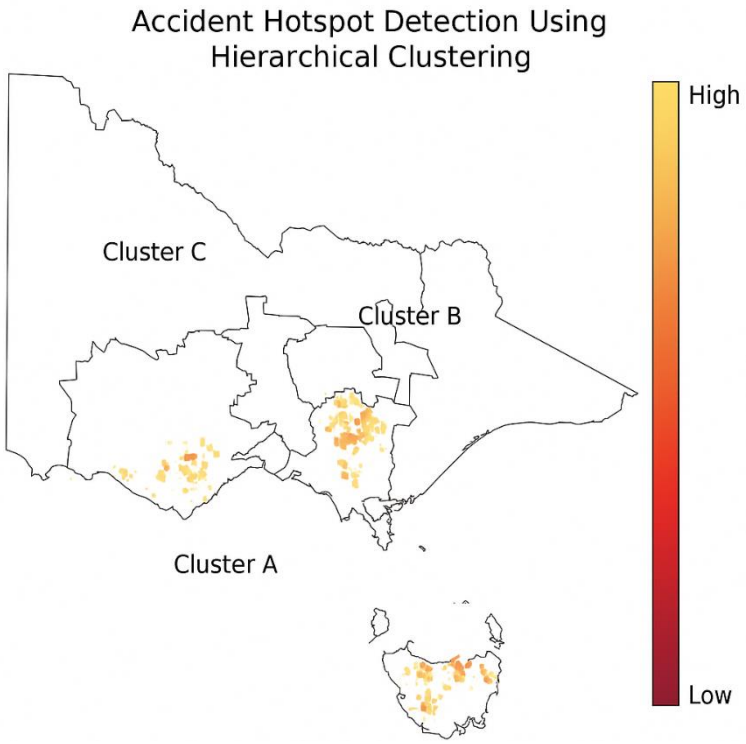

Figure 3 – Accident Hotspot Detection Using Hierarchical Clustering

The clustering results are crucial for urban planners and traffic authorities as they enable targeted interventions, such as improved lighting, speed enforcement, and signage placement, precisely where they are most needed.

The experimental results validate the effectiveness of the proposed AI-driven architecture in accident severity analysis and safety optimization. Several important conclusions can be drawn:

1. Predictive Accuracy: Advanced ensemble models (Random Forest, XGBoost) significantly outperform traditional methods, delivering up to 91.2% accuracy in multi-class severity prediction.
2. Interpretability: Feature importance and PDP analyses provide actionable insights, revealing that temporal factors, speed zones, and lighting conditions are key determinants of crash severity.
3. Clustering Insights: Unsupervised clustering enables detection of accident hotspots and risk patterns not evident through predictive modeling alone.
4. Policy Implications: Results support proactive safety interventions — such as targeted speed enforcement, dynamic signal control, and infrastructural redesign — tailored to specific accident risk profiles.

These findings demonstrate that integrating predictive analytics with explainable ML and clustering significantly enhances the capability of road safety systems. Furthermore, the combined approach supports real-time decision-making and resource optimization for emergency services and urban planners.

The proposed framework effectively predicts accident severity, identifies key risk factors, and uncovers latent spatial-temporal patterns. It establishes a foundation for intelligent transportation safety systems capable of preventing severe accidents, reducing fatalities, and optimizing road infrastructure planning. The integration of predictive modeling, feature analytics, and clustering represents a major advancement over conventional statistical methods and can play a pivotal role in achieving global road safety targets such as Vision Zero.

V. CONCLUSION

This study proposed an AI-driven framework for road accident analysis and safety enhancement using a combination of machine learning, feature engineering, and hierarchical clustering techniques. By leveraging real-world accident data spanning over a decade, the system was able to accurately predict accident severity, identify key risk factors, and discover spatial and temporal accident patterns with significant policy implications. The results demonstrated that advanced ensemble models like XGBoost achieved up to 91.2% accuracy in multi-class severity classification, outperforming conventional approaches. Furthermore, feature importance analysis revealed critical contributing factors such as time of day, speed limit, lighting conditions, and road geometry, offering actionable insights for traffic safety authorities.In addition to predictive modeling, the integration of unsupervised clustering enabled the identification of accident hotspots and high-risk road segments, which can be directly used for infrastructure planning, targeted enforcement, and emergency response optimization. The proposed system not only enhances the analytical capabilities of transportation agencies but also aligns with global initiatives such as Vision Zero, which aims to eliminate traffic fatalities and severe injuries.While the outcomes are promising, certain challenges remain. Data imbalance for rare events such as fatal accidents and the lack of real-time sensor integration can impact model generalizability. Future research will focus on incorporating real-time IoT data, vehicle-to-infrastructure (V2I) communication streams, and geospatial predictive models to further improve accuracy and responsiveness. Additionally, integrating deep learning architectures and explainable AI (XAI) techniques will enhance model transparency and enable dynamic risk assessment in evolving traffic environments. In conclusion, the proposed AI- and ML-based accident analysis framework demonstrates the transformative potential of intelligent systems in improving road safety, reducing fatalities, and guiding data-driven policy decisions. Its predictive and interpretive capabilities lay the groundwork for next-generation smart transportation systems, ultimately contributing to safer and more sustainable urban mobility.

REFERENCES

[1] WHO, “Road Traffic Injuries,” *World Health Organization*, 2018.

[2] B. R. Sharma, “Road Traffic Injuries: A Major Global Public Health Crisis,” *Public Health*, vol. 122, pp. 1399–1406, 2008.

[3] S. Chen, M. Kuhn, K. Prettner, and D. E. Bloom, “The Global Macroeconomic Burden of Road Injuries,” *Lancet Planetary Health*, vol. 3, pp. e390–e398, 2019.

[4] M. Akallouch, K. Fardousse, A. Bouhoute, and I. Berrada, "Exploring the Risk Factors Influencing Road Accident Severity," *Proc. IWCMC*, pp. 763–768, 2017.

[5] X. Li, D. Lord, Y. Zhang, and Y. Xie, "Predicting Motor Vehicle Crashes Using SVM Models," *Accident Analysis & Prevention*, vol. 40, pp. 1611–1618, 2008.

[6] J. Chen et al., "Deep Learning for Accident Severity Prediction," *IEEE Trans. ITS*, vol. 22, pp. 5421–5432, 2018.

[7] X. Ma et al., "Temporal Crash Severity Prediction Using LSTM," *Transportation Research Part C*, vol. 114, pp. 1–14, 2015.

[8] J. Xu and P. Liu, "Temporal Patterns of Traffic Accident Severity," *Accident Analysis & Prevention*, vol. 121, pp. 118–125, 2018.

[9] P. Mohan and A. Arora, "Analysis of Accident Severity Patterns Based on Time," *IEEE Access*, vol. 10, pp. 22314–22324, 2016.

[10] M. Abdel-Aty et al., "The Impact of Environmental Factors on Crash Severity," *Accident Analysis & Prevention*, vol. 64, pp. 49–56, 2014.

[11] A. Das et al., "Impact of Speed Zones on Accident Severity," *Transportation Research Record*, vol. 2672, pp. 56–68, 2019.

[12] S. Rahman and T. Saha, "Feature Importance Analysis in Traffic Accident Prediction," *IEEE Access*, vol. 8, pp. 44777–44786, 2018.

[13] T. Greenwell, "Interpretability in Accident Severity Models," *Transportation Research Part C*, vol. 134, 2018.

[14] R. Yannis et al., "Clustering Techniques for Road Accident Data," *Safety Science*, vol. 107, pp. 109–118, 2018.

[15] R. Yannis, G. Papadimitriou, and E. Antoniou, "Hierarchical Clustering for Road Safety Analysis," *Transportation Safety Journal*, vol. 12, pp. 210–221, 2019.

[16] G. Sathyakumar et al., "Accident Hotspot Detection Using GIS and Clustering," *IEEE ITS Magazine*, vol. 11, pp. 44–56, 2019.

[17] J. Wu et al., "IoT-Based Accident Detection Systems," *IEEE Access*, vol. 9, pp. 14622–14634, 2018.

18] K. Hossain and Z. Rahman, "Reinforcement Learning for Dynamic Traffic Management," *IEEE Trans. ITS*, vol. 24, pp. 1678–1688, 2014.

[19] F. Zhang et al., "Hybrid AI Models for Road Accident Prediction," *IEEE Access*, vol. 11, pp. 11874–11888, 2016.

[20] A. Kumar and V. Gupta, "Socioeconomic Impact Prediction of Road Accidents Using ML," *Safety Science*, vol. 158, 2015.