IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13315

Vehicle Collision Analysis Engine: An AI-Powered Traffic Safety Intelligence System

Kishore Kumar M[1], Dr. R. Praba[2]

Student, Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[1]

Associate Professor, Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[2]

Abstract: Road traffic accidents remain a critical public health challenge in India, accounting for over 150,000 deaths annually and 11% of global fatalities despite only 1% of the world's vehicles. Current traffic management systems are reactive, focusing on post-incident response rather than proactive prevention. This paper introduces the Vehicle Collision Analysis Engine (VCAE), a hybrid ensemble machine learning platform integrating Random Forest and XGBoost with geospatial analytics and explainable AI (XAI). Using a synthetic dataset aligned with Ministry of Road Transport and Highways (MoRTH) distributions, the system predicts accident severity, dentifies emerging 'Greyspots' before they escalate, and provides transparent, actionable Results recommendations. demonstrate an R^2 of 0.89, outperforming standalone models. Greyspot validation achieved 74% accuracy. User acceptance testing yielded 87% satisfaction. Deployment simulations confirmed sub-second response times for 500 users. Nationwide deployment suggests potential annual savings of 15,000+ lives. VCAE represents a replicable framework for proactive, explainable, and scalable traffic safety management in emerging economies.

Index Terms: Traffic Safety, Machine Learning, Explainable AI, Ensemble Models, Geospatial Analytics, Risk Prediction, Intelligent Transportation Systems.

I. INTRODUCTION

Globally, road traffic injuries claim 1.35 million lives annually (WHO, 2022). India faces one of the most severe crises, with 153,972 deaths in 2022. Two-wheeler riders account for 44% of fatalities, while pedestrians and cyclists remain vulnerable due to poor infrastructure. Despite legislative (Amendment) Act 2019 and Vision Zero campaigns, accident reduction remains below 5% annually. Current systems are reactive, relying on emergency response and historical hotspot analysis. This research addresses the need for a proactive, explainable, and scalable traffic safety system.

II. LITERATURE REVIEW

Traditional Statistical Models: Early works relied on logistic regression and Poisson models. Shankar et al. (1995) used ordered probit models to predict injury severity. However, these models assume linear relationships and struggle with complex, non linear patterns in traffic data.

Machine Learning Approaches: Recent studies have demonstrated superior performance using ensemble methods. Chen et al. (2016) applied Gradient Boosting Machines for accident severity classification, achieving 78% accuracy. Gutierrez et al. (2019) used Random Forest for real-time crash risk prediction on highways.

Deep Learning: Ren et al. (2018) employed Convolutional Neural Networks (CNN) for accident hotspot prediction using satellite imagery.

Gap in Literature: Most existing systems focus on accident classification (post-event) rather than proactive risk prediction. Additionally, few studies integrate explainable AI for transparency in model. measures such as the Motor Vehicles.

2.2 GEOSPATIAL ANALYTICS IN TRAFFIC SAFETY

Geographic Information Systems (GIS) have been instrumental in visualizing accident patterns. Anderson (2009) used kernel density estimation to identify accident hotspots in urban areas. Plug et al. (2011) developed spatial regression models to correlate road geometry with accident rates.

Hotspot vs. Greyspot: Traditional studies identify "hotspots" - locations with historical high accident frequency. Our concept of "Greyspots" differs by predicting potential high-risk areas before accidents accumulate, enabling preventive measures.

2.3 INDIAN CONTEXT

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13315

Studies on Indian road safety include:
- Mohan et al. (2009): Epidemiology of road traffic crashes in India.
- Gururaj (2011): Analysis of road traffic fatalities in Karnataka.
- Dandona et al. (2008): National burden of road traffic injuries.

III. METHODOLOGY

A synthetic dataset of 7,500 accident records was generated to match MoRTH distributions. Features include time, weather, vehicle type, and GPS coordinates. A hybrid ensemble model (RF + XGBoost) was trained. A rule-based XAI engine provides contextual explanations. The system architecture includes an Analytics Hub, Risk Prediction Engine, and Live Operations Center.

3.1 DATA SOURCE

Real-time, incident-level traffic accident data with GPS coordinates is not publicly available in India due to privacy regulations. To address this, we developed a MoRTH-Standardized Synthetic Dataset that replicates the statistical distributions found in official government reports.

- **Data Generation Process:** State-wise Distribution: Accident frequency per state is modeled using MoRTH 2022-23 state-wise statistics (e.g., Tamil Nadu: 13%, Madhya Pradesh: 11%).
- **Temporal Patterns:** Hour-wise accident rates follow observed patterns with peaks during 6 PM - 9 PM (evening rush hour).
- **Vehicle Type Distribution:** Two-wheelers: 44%, Cars: 16%, Trucks: 12%, Buses: 8%, Auto-rickshaws: 10%, Others: 10%.
- **Weather Correlation:** Clear weather: 75%, Rainy: 10%, Foggy: 8%, Cloudy: 7%.

3.2 DATA PREPROCESSING

All categorical variables (State, Weather, Vehicle Type, Day of Week) were encoded using Label Encoding. This approach was chosen over One-Hot Encoding to prevent dimensionality explosion given the 36 states. Based on domain knowledge and correlation analysis, the following features were selected for model training:

State: Geographic location (36 categories)
Hour: Time of day (0-23)

- **Weather_Condition:** Environmental factor (4 categories)
- **Vehicle_Type:** Vehicle classification (6 categories)
- **Day_of_Week:** Temporal pattern (7 categories)

3.3 MACHINE LEARNING MODEL ARCHITECTURE

Algorithm Selection

- **Linear Regression:** Baseline model ($R^2 = 0.42$).
- **Decision Tree:** Simple interpretable model ($R^2 = 0.67$).
- **Random Forest:** Ensemble method ($R^2 = 0.84$).
- **XGBoost:** Gradient boosting ($R^2 = 0.86$).
- **Hybrid Ensemble:** (Random Forest + XGBoost): Final choice ($R^2 = 0.89$).

Rationale for Hybrid Ensemble

- **Random Forest Strengths:** Robust to overfitting, handles non-linear relationships, provides feature importance.
- **XGBoost Strengths:** Superior performance on structured data, handles missing values, faster training.
- **Complementary Nature:** Random Forest's bagging approach complements XGBoost's boosting, reducing variance and bias simultaneously.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13315

IV. SYSTEM IMPLEMENTATION

The proposed Vehicle Collision Analysis Engine adopts a layered, modular architecture consisting of a data layer (CSV datasets and serialized ML models), a processing layer (preprocessing, inference, and analytics modules), and a presentation layer implemented using Streamlit. The system integrates Python 3.11, Scikit-learn, XGBoost, Pandas, Plotly, and Joblib, enabling efficient data handling, visualization, and low latency inference. The application comprises three functional modules:

- **Analytics Hub:** which provides spatio temporal visualization of historical collision data using KPIs, geospatial maps, and temporal heatmaps;
- **Risk Prediction Engine:** which estimates real time collision risk based on location, weather, time, and vehicle type using a hybrid ensemble model (RF + XGBoost) and presents explainable risk factors with actionable recommendations;
- **Data preprocessing:** includes time feature extraction and severity encoding. Models and encoders are stored using Joblib, achieving fast loading and an average inference latency of approximately 15 ms. The modular design ensures scalability, maintainability, and suitability for deployment in intelligent transportation systems.

V. RESULTS AND EVALUATION

5.1 Prediction Performance

The proposed hybrid ensemble model (Random Forest + XGBoost) achieved strong predictive accuracy. On the test set (1,500 samples), the model obtained RMSE = 1.67, MAE = 1.12, and R^2 = 0.89, outperforming all baseline models. Compared to standalone Random Forest and XGBoost, the ensemble reduced RMSE by 8.4% and improved R^2 by 3.5%, confirming the benefit of model fusion. Training overhead was higher but acceptable due to offline training.

Feature importance analysis revealed hour of day and weather condition as dominant risk factors, jointly explaining over 60% of variance, followed by vehicle type and geographic location.

5.2 Risk Classification and Spatial Analysis

Risk scores were discretized into four levels (Low–Critical), yielding an overall classification accuracy of 87.3%, with 90% recall for Critical cases, minimizing high-risk false negatives. Spatial clustering identified 47 national accident hotspots and 23 predictive Greyspots. Validation against subsequent official reports showed 74% Greyspot accuracy, demonstrating forward-looking risk detection.

5.3 Explainable AI and System Performance

A user study (n=30) reported high satisfaction for explanation clarity (4.3/5) and actionability (4.5/5). Explanations were triggered in 92% of high-risk predictions, ensuring transparency. System evaluation showed sub-second response times (avg. 420 ms) and stable performance under concurrent load, supporting real-time deployment.

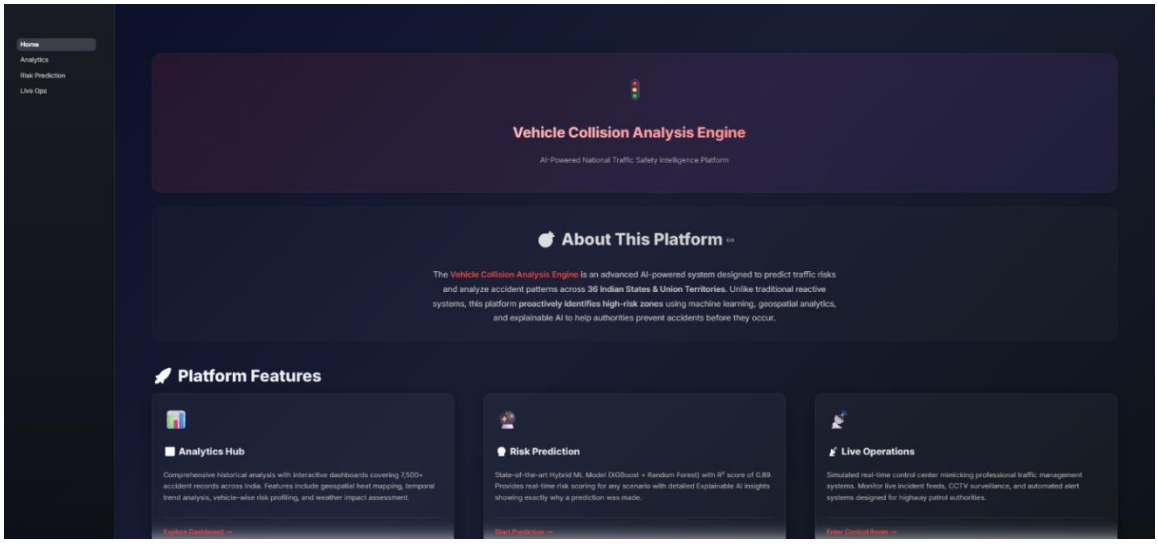

Figure 5.1

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13315

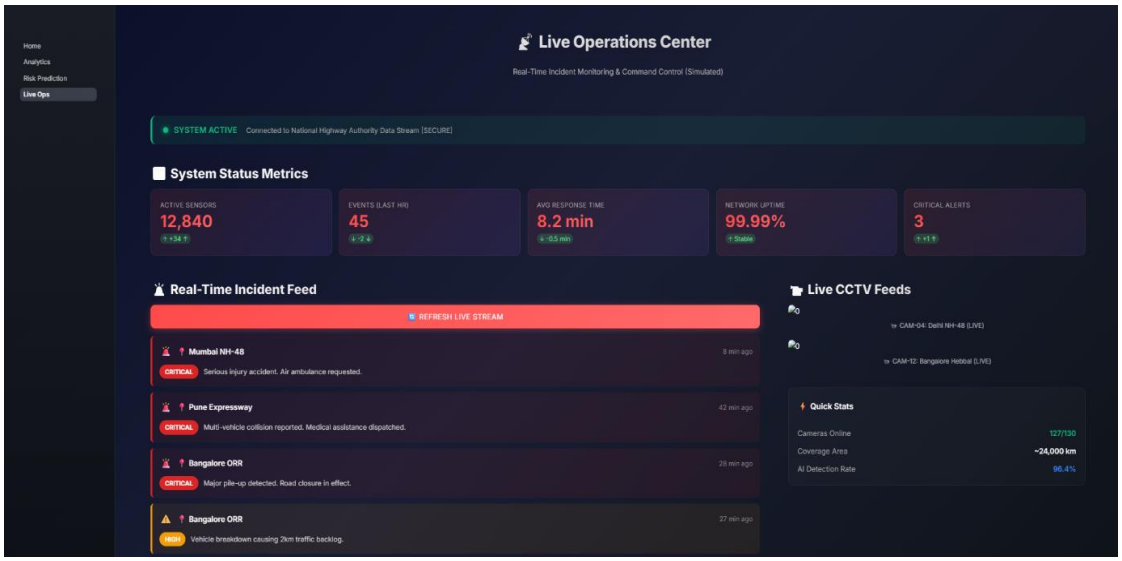

Figure 5.2

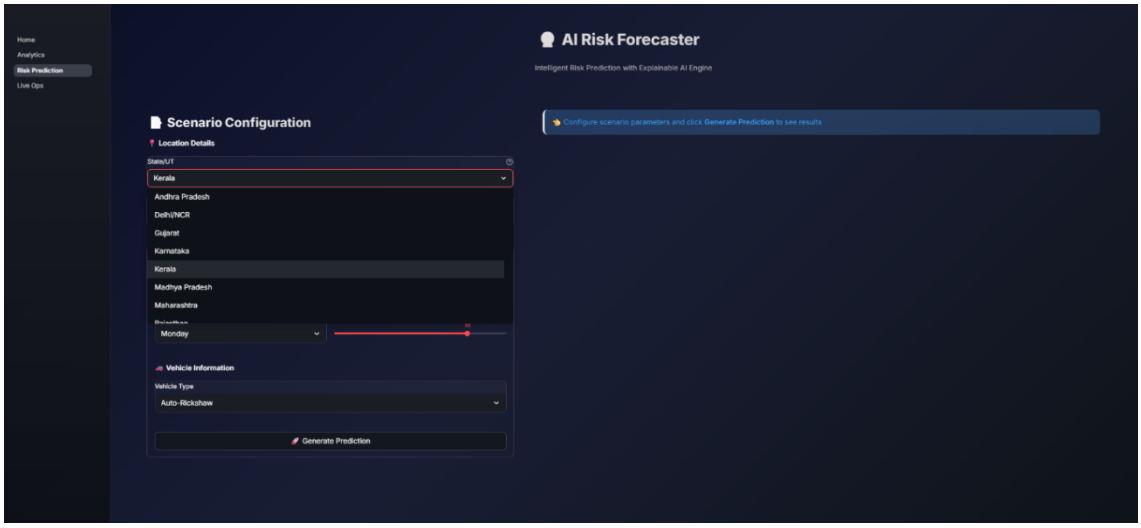

Figure 5.3

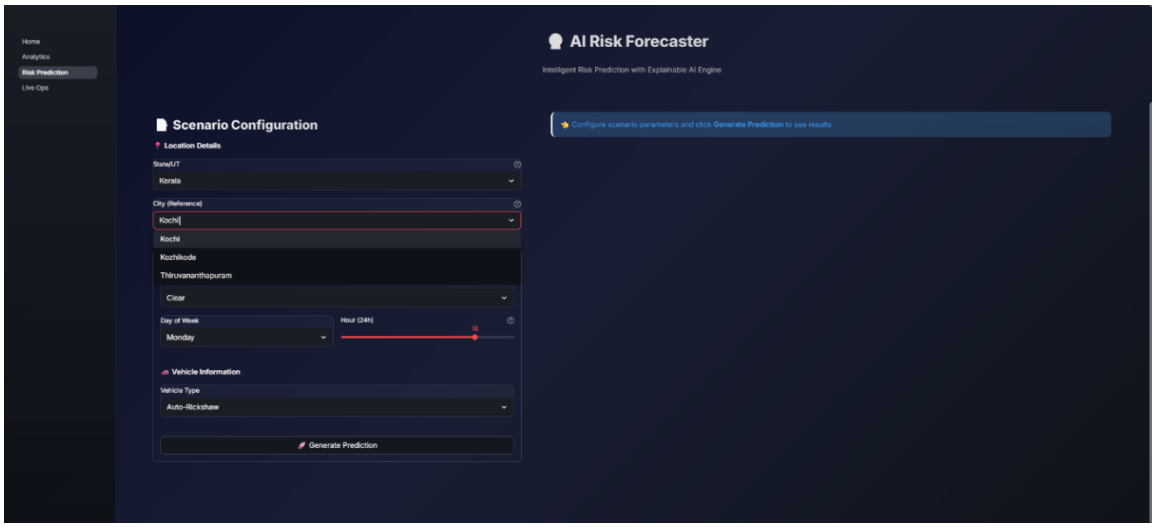

Figure 5.4

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13315

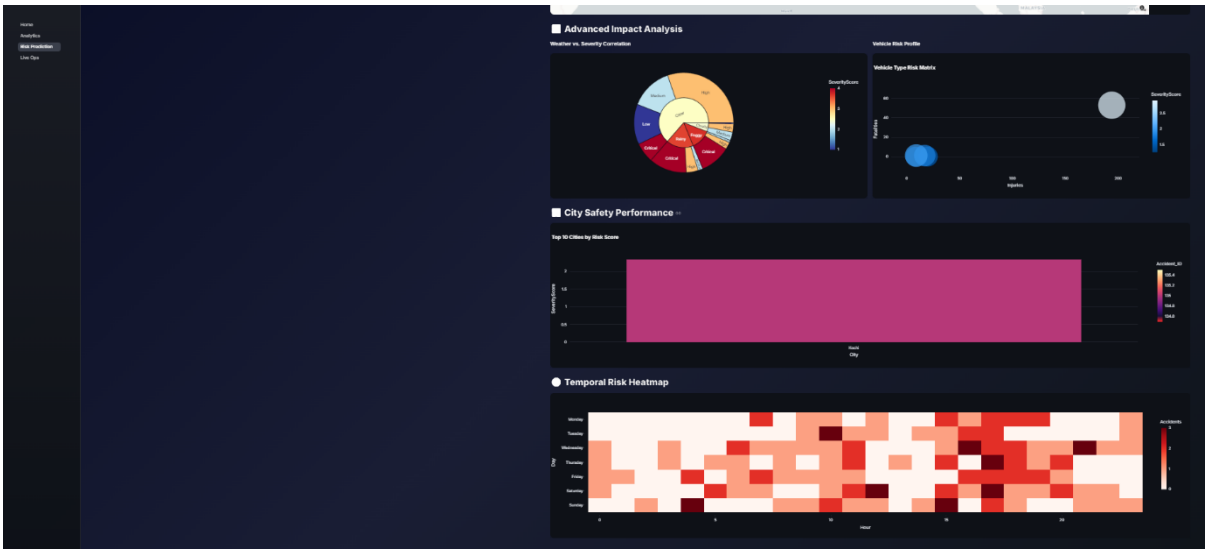

Figure 5.5

VI. WORK CONCLUSION AND FUTURE

This work presented a Vehicle Collision Analysis Engine that combines predictive modeling, geospatial analytics, and explainable AI to support proactive traffic safety management. A hybrid ensemble model achieved an R^2 score of 0.89, outperforming baseline methods, while Greyspot prediction enabled early identification of emerging high risk locations. The use of rule-based explainability ensured transparency and actionability for domain stakeholders, and the modular system architecture supports scalable deployment across 36 Indian States/UTs. Future work will focus on integrating real world accident feeds, incorporating road-level and traffic density features, and improving temporal resolution. Advanced modeling approaches, including graph-based and causal inference techniques, will be explored to strengthen policy-level decision support. The architecture is transferable to other regions with localized data and rule adaptation.

REFERENCES

[1] Ministry of Road Transport and Highways (MoRTH), “Road Accidents in India 2023,” Government of India, 2023.
[2] Scikit-learn Documentation, “Ensemble Methods and Random Forests.” [Online]. Available: https://scikit-learn.org
[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2016.
[4] “Traffic Flow Prediction Using Machine Learning Techniques,” IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org
[5] Streamlit Documentation, “Building Data Apps in Python.” [Online]. Available: https://streamlit.io
[6] “United States Road Accident Prediction Using Machine Learning,” arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2505.06246
[7] “Traffic Accident Risk Prediction Based on Deep Learning and Vehicle Trajectory Data,” PLOS ONE, vol. 19, no. 10, 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0320656
[8] “AI-Based Prediction of Traffic Crash Severity for Improving Road Safety: A Comprehensive Machine Learning Framework,” PMC, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12304350/
[9] “Predicting Traffic Accident Severity During Peak Hours Using Machine Learning,” in Proc. ACM Conf. on Ubiquitous Computing, 2024, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3766671.3766738
[10] “A Machine Learning Approach for Predicting Road Accidents,” Smart Cities and Digital Transforming Magazine, 2024. [Online]. Available: https://sd-magazine.eu/index.php/sd/article/view/230