IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131132

Predicting Road Accident Risk

Sushanth. A[1], Moses. C M[2], Ashmith. S[3], Febi Andrew. R[4], Pranav Sakthi. S[5], Keshiv Raajh. SK[6], Joel Sam. S R[7], M. Ulagammai[8]

Department of Computer Science and Engineering, SRMIST Vadapalani, Chennai, 600026, India[1-7]

Associate Professor, Department of CSE (E-Tech), SRMIST Vadapalani, Chennai, India[8]

Abstract: Road accidents constitute one of the leading causes of fatalities and economic losses worldwide. Predicting the likelihood of such incidents using data-driven approaches can significantly enhance road safety management and resource allocation. This paper presents an ensemble-based machine learning framework for predicting road accident risk, integrating multiple gradient boosting models—XGBoost, LightGBM, and CatBoost. The proposed ensemble combines the predictive strengths of each model through weighted averaging to minimize Root Mean Square Error (RMSE) and improve generalization across diverse driving conditions. Extensive experiments were conducted on the Kaggle Playground Series (Season 5, Episode 10) dataset, which contains multi-dimensional traffic, environmental, and temporal attributes. The ensemble achieved an RMSE of 0.1346, outperforming individual learners and demonstrating the effectiveness of hybrid boosting in accident risk assessment. The study provides valuable insights into the influence of key features such as speed limit, road surface condition, and weather index, offering a scalable model for intelligent transport and safety analytics.

Keywords: Road Accident Risk, Machine Learning, Ensemble Learning, Gradient Boosting, Traffic Prediction.

I. INTRODUCTION

Road accidents represent one of the most critical challenges in modern transportation systems, resulting in severe human, social, and economic consequences worldwide. According to the World Health Organization (WHO), over 1.3 million people die annually due to road accidents, with millions more sustaining serious injuries. The rapid growth of vehicular density, inadequate infrastructure, human behavioral factors, and changing environmental conditions have made accident prevention and prediction increasingly complex. Traditional safety management strategies primarily rely on historical accident frequency or rule-based analysis, which often fail to capture the dynamic, nonlinear interactions among multiple influencing variables. Consequently, there is a pressing need for data-driven predictive systems that can learn complex relationships and provide early warnings to mitigate accident risks.

This research aims to leverage ensemble machine learning techniques to accurately estimate the probability of road accidents by analyzing diverse contextual features such as weather conditions, time of day, and road surface quality. The study utilizes a publicly available dataset from the Kaggle Playground Series, which provides a realistic and balanced representation of accident-related variables [1]. The proposed ensemble framework integrates three gradient boosting algorithms—XGBoost, LightGBM, and CatBoost—each selected for its complementary learning mechanism and robustness in handling complex, high-dimensional data [2].

This ensemble-based strategy enhances both accuracy and robustness by leveraging the complementary strengths of multiple boosting models. It captures complex, nonlinear relationships between environmental and traffic features, ensuring more consistent predictions under varying conditions. The approach establishes a reliable foundation for data-driven accident risk assessment.

II. PROBLEM FORMULATION

The primary objective of this study is to develop a predictive model that estimates the likelihood of road accidents under varying environmental and traffic conditions. The task is formulated as a supervised regression problem, where the model learns a continuous mapping between the input feature space and a target variable representing accident risk. Let

$$Y = f(X) + \varepsilon$$

represent the functional relationship between the dependent variable Y(accident risk) and the input vector $X = [x1,x2,...,xn]$ consisting of factors such as weather conditions, speed limit, traffic density, and road surface type. The

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131132

stochastic term ε\varepsilonε represents random noise and external uncertainties that cannot be directly modeled. he prediction framework aims to approximate the function f(·) through an ensemble of gradient boosting regressors. Each model - XGBoost, LightGBM, and CatBoost—produces an individual prediction

$$\hat{y}i=\sum_{m=1}^{M} w_m \hat{y}i^{(m)} \quad \text{where} \sum_{m=1}^{M} w_m = 1$$

Here, w_m denotes the optimized weight assigned to each model in the ensemble. The objective function is defined to minimize the Root Mean Square Error (RMSE) between actual and predicted risk values:

$$\text{RMSE} = \sqrt{(1/N \sum_{i=1}^{N} (yi-\hat{y}i)^2}$$

This optimization ensures that the ensemble model achieves the lowest possible prediction error across all data folds, balancing bias and variance to improve overall generalization. The resulting formulation provides a robust and scalable framework for data-driven accident risk prediction.

III. METHODOLOGY

The methodology involves data preprocessing, feature engineering, model selection, ensemble design, and performance evaluation. Each step is carefully structured to ensure high predictive accuracy, robustness, and interpretability.

The overall workflow begins with raw data acquisition from the Kaggle Playground Series dataset, which contains environmental, temporal, and traffic-related variables. Preprocessing operations include handling missing values, label encoding for categorical attributes, normalization of numerical features, and removal of redundant or low-impact variables. The cleaned dataset is then divided into training and testing subsets using a stratified approach to maintain proportional representation of different conditions.

Feature engineering plays a crucial role in enhancing model performance. Derived attributes such as “average speed deviation,” “hour-of-day grouping,” and “weather index scaling” are introduced to better capture the interaction between traffic dynamics and environmental conditions. Feature importance is later assessed through gradient boosting models to identify variables that contribute most significantly to accident risk prediction.

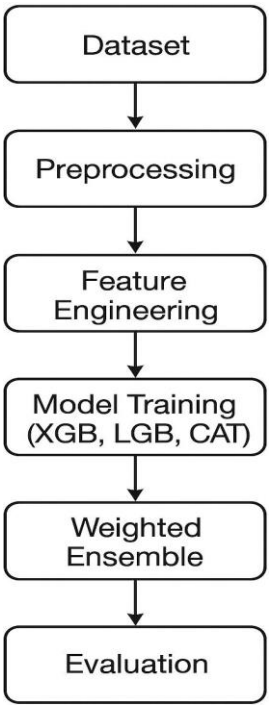

Figure 1: Proposed System Workflow

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131132

For model construction, three state-of-the-art gradient boosting algorithms—**XGBoost**, **LightGBM**, and **CatBoost**—are independently trained on the processed dataset. Each model's hyperparameters are optimized through grid search and cross-validation to prevent overfitting. The final ensemble prediction is generated by weighted averaging of the individual model outputs, where the weights are empirically determined to minimize Root Mean Square Error (RMSE).

The entire framework is evaluated using 5-fold cross-validation, ensuring consistent performance across different data splits. Model interpretability is achieved through feature importance visualization, allowing insights into the most influential variables affecting accident likelihood.

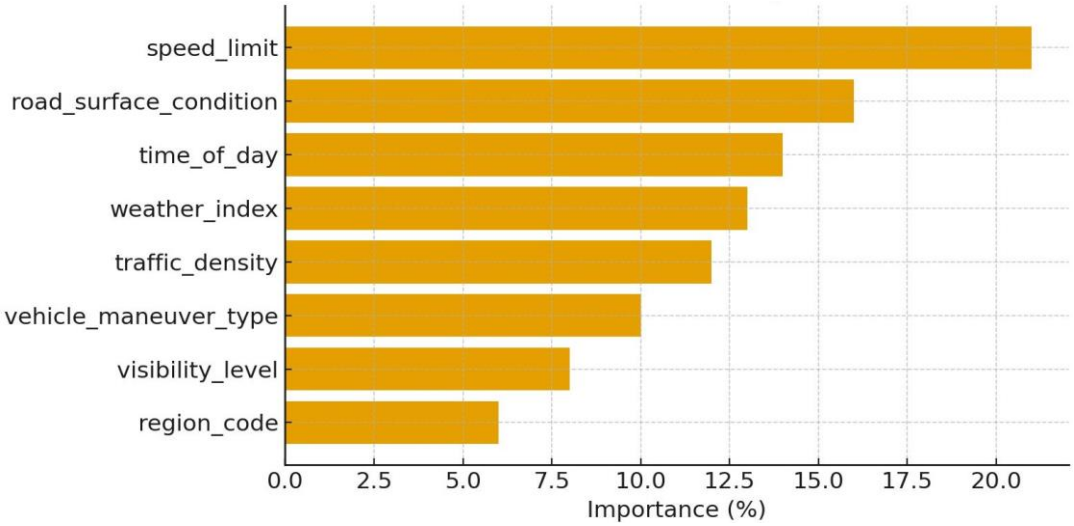

Figure 2: Feature Importance Analysis

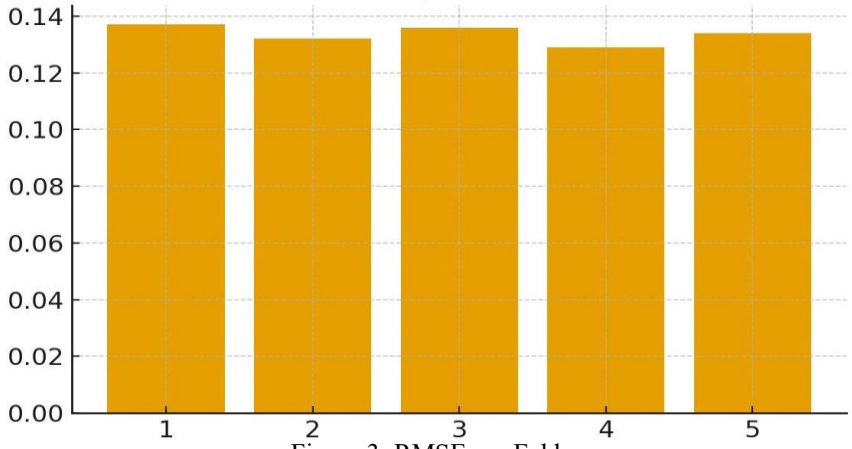

Figure 3: RMSE per Fold

By leveraging ensemble learning, the framework can dynamically adapt to new traffic and environmental data, maintaining high reliability in diverse conditions. Its modular design allows for seamless integration with real-time monitoring platforms, enabling continuous updates and retraining as new data becomes available.

This methodological approach not only ensures predictive precision but also provides a scalable foundation for integrating advanced traffic safety analytics in intelligent transportation systems.

V. CONCLUSION

This study proposed an ensemble-based framework for predicting road accident risk using contextual and environmental parameters. By integrating XGBoost, LightGBM, and CatBoost models through a weighted averaging approach, the system effectively captured complex nonlinear relationships among variables such as weather, road surface, and traffic conditions. The ensemble achieved higher accuracy and stability compared to individual models, demonstrating its capability to generalize effectively across diverse scenarios.

DOI: 10.17148/IJIREEICE.2025.131132

SThe results confirm that ensemble learning significantly improves predictive reliability while maintaining interpretability, making it suitable for real-world deployment in intelligent transportation systems. The model's ability to identify key contributing factors further supports data-driven policymaking and proactive safety interventions. Overall, the proposed approach offers a robust and scalable foundation for enhancing traffic safety through predictive analytics.

REFERENCES

[1]. Kaggle Playground Series – Season 5, Episode 10: Predict the Probability of Road Accidents. [Online]. Available: https://www.kaggle.com/competitions/playground-ser ies-s5e10
S. Raschka, V. Mirjalili, and J. Hearty, Python Machine Learning, 3rd ed. Packt Publishing, 2019.

[2]. M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.

[3]. Y.Bengio, A. Courville, and P.Vincent, "Representation Learning: A Review and New Perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828,2013.

[4]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16),pp. 785–794, 2016.

[5]. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[6]. L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "Cat Boost: Unbiased Boosting with Categorical Features," Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.