ROAD ACCIDENT SEVERITY PREDICTION USING MACHINE LEARNING

Jebisha J[1], Dharshini K S[1], Abi Selvi S[1], Mrs. Caroline Misbha J[2]

[1]Student, Department of Computer Science & Engineering, Arunachala College of Engineering for Women

[2]Assistant Professor, Department of Computer Science & Engineering, Arunachala College of Engineering for Women

ABSTRACT:

Road accidents are a major public safety concern and cause thousands of injuries and fatalities every year in India. The increasing accident rate highlights the need for analytical systems that can understand accident patterns and predict severity levels. Traditional accident analysis methods mainly provide historical information and fail to identify risk factors or predict accident severity in advance. To address these challenges, this research proposes a machine learning-based accident severity prediction system.

The proposed model uses accident-related factors such as weather conditions, road type, vehicle type, time of accident, location, and human behavior to classify accident severity into categories such as Minor, Serious, and Fatal. The dataset is preprocessed and cleaned, followed by feature selection to identify the most influential variables. Machine Learning algorithms like Random Forest, Logistic Regression, XGBoost, and Neural Networks are trained and evaluated. Among them, XGBoost achieves the best performance with high accuracy in predicting serious accident cases.

The system effectively identifies accident-prone conditions and patterns, helping authorities and decision-makers implement preventive measures. The results demonstrate that machine learning can significantly improve accident risk prediction and contribute to better road safety planning.

KEYWORDS: Machine Learning, Road Accident Prediction, Accident Severity Classification, Data Preprocessing, XGBoost, Neural Network, Traffic Safety, Feature Selection, Predictive Analytics, Accident Risk Analysis

1. INTRODUCTION:

Road accidents have become a major global concern, especially in developing countries like India where rapid population growth and vehicle expansion place enormous pressure on existing road infrastructure. Despite continuous efforts by government and transportation authorities, the number of accidents continues to rise each year due to factors such as traffic rule violations, speeding, drunk driving, vehicle malfunction, and unpredictable road behavior. These incidents not only result in physical injuries and fatalities but also cause emotional,

social, and economic consequences for families and communities. India consistently ranks among the highest in the world in terms of road accident deaths. According to recent national statistics, the majority of accidents occur due to human error, inadequate road design, poor lighting conditions, and limited traffic management. With millions of vehicles on the road, analyzing accident causes manually becomes extremely challenging and time-consuming. This calls for advanced technological solutions that can automatically analyze complex datasets and provide meaningful insights. Traditional statistical approaches used for accident analysis are limited because they rely on predefined assumptions and cannot effectively capture nonlinear patterns or interactions between variables. Machine Learning (ML), however, offers a powerful alternative. ML algorithms can handle large-scale accident datasets, learn from historical patterns, and accurately predict the severity of future accidents based on various influencing factors such as weather conditions, road type, traffic flow, vehicle categories, time, and location. Machine Learning-based predictive systems have the potential to support intelligent transportation systems (ITS), enhance road safety management, and assist authorities in proactive decision-making. By identifying accident-prone areas and high-risk driving conditions, these systems can help reduce fatalities, minimize injuries, and improve overall traffic efficiency. This project aims to develop a robust machine learning model that predicts accident severity for Indian road conditions. The system incorporates data preprocessing, feature selection, and the training of multiple ML models to determine which provides the highest accuracy. Ultimately, the goal is to create a reliable accident prediction system capable of assisting policymakers, road safety authorities, and the public in preventing severe accidents and promoting safer travel.

2. RELATED WORK:

Several studies have explored the use of machine learning for accident prediction and road safety analysis.

Farzaan et al. (2024) developed a machine learning-based accident prediction model using Random Forest, demonstrating high accuracy but facing challenges with unbalanced datasets.
Kumar et al. (2023) analyzed road accident severity using SVM and achieved good classification results, though the training time was high.
Sharma et al. (2023) explored logistic regression for accident severity analysis, providing fast results but lower accuracy compared to ensemble models.
Lee et al. (2024) used decision trees for traffic accident severity classification, showing good interpretability but suffering from overfitting issues.
Wang et al. (2024) introduced K-Means clustering to detect accident hotspots, identifying high-risk zones but being sensitive to data quality.

Patel et al. (2025) implemented neural networks for accident prediction, improving performance but requiring significant training time.
Verma et al. (2024) proposed a hybrid RF + SVM model that improved prediction accuracy but increased complexity.
Johnson et al. (2025) used big data analytics for real-time accident analysis, emphasizing scalability but requiring high computational resources.

Existing research highlights the potential of machine learning in accident prediction but lacks an integrated framework that includes preprocessing, feature selection, model comparison, and severity classification. Our proposed system addresses these gaps by providing a unified and accurate prediction model.

3. PROPOSED SOLUTION:

The proposed solution is a machine learning-based accident severity prediction system designed to classify the severity of road accidents into three categories—Minor, Serious, and Fatal—based on a variety of accident-related attributes. The system uses a large and comprehensive accident dataset that includes environmental, vehicle-related, and temporal factors. To ensure reliable predictions, the dataset undergoes detailed preprocessing steps such as handling missing information, removing inconsistencies, normalizing numerical fields, and encoding categorical variables. This ensures that the data fed into the model is clean, structured, and suitable for training.

After preprocessing, the system performs feature selection to identify and retain the most significant variables that influence accident severity. Features such as weather conditions, alcohol involvement, vehicle type, accident location, and road characteristics play a crucial role in determining the seriousness of an accident. By selecting the most relevant attributes, the system reduces unnecessary complexity and enhances model performance.

To build a strong predictive engine, multiple machine learning models—including Logistic Regression, Decision Tree, Random Forest, XGBoost, and Neural Networks—are trained and evaluated. Each model learns patterns from the historical accident data, and their performance is compared to identify the most suitable algorithm. Among these, XGBoost demonstrated superior results, offering high accuracy and effectively capturing complex, nonlinear relationships within the dataset. Neural Networks also showed strong performance but required longer training time.

The proposed system not only predicts severity but also helps in understanding accident-prone conditions and high-risk scenarios. Through its data-driven insights, the system assists traffic authorities, policymakers, and road safety planners in identifying patterns, taking preventive actions, and minimizing future accidents. By integrating machine learning techniques with real-world accident data, the proposed solution offers an efficient, automated, and scalable approach to improving road safety and reducing accident-related fatalities.

4. METHODOLOGY

The proposed system follows a structured, intelligent, and continuous methodology to accurately predict accident severity in Indian road conditions. The methodology is divided into several stages to ensure that the data is properly processed, the models are trained effectively, and the predictions are reliable. Each stage plays an essential role in improving system performance and maintaining consistency across the machine learning workflow.

- **Data Collection**
  The dataset is collected from Indian road accident records and includes important attributes such as weather conditions, road type, accident time, vehicle category, location details, and severity level. This ensures that the system has access to real-world accident patterns and influencing factors.

- **Data Preprocessing**
  The collected data is cleaned and prepared for further analysis. Missing values are removed, inconsistent entries are corrected, numerical variables are normalized, and categorical attributes are encoded. Noisy and redundant information is filtered out to improve the overall efficiency and accuracy of the system.

- **Feature Extraction & Selection**
  Important features that influence accident severity are selected using correlation analysis, mutual information, and feature importance techniques from machine learning models such as Random Forest and XGBoost. This helps reduce dimensionality and ensures that only meaningful variables are used.

- **Model Training**
  Multiple machine learning models including Logistic Regression, Decision Tree, Random Forest, XGBoost, and Neural Networks are trained using the processed dataset. These models learn patterns and relationships between input features and accident severity levels to enable accurate predictions.

- **Severity Prediction**
  The trained models classify accident cases into three severity categories: Minor, Serious, and Fatal. This helps identify the seriousness of an accident based on the given input conditions.

- **Performance Evaluation**
  Model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix to determine the best-performing model. These metrics help measure how effectively the system predicts severity and generalizes to unseen accident data.

5. 5. ARCHITECTURE

The architecture of the proposed road accident severity prediction system is designed to ensure an efficient and streamlined workflow from data input to severity output. The system

follows a sequential layered structure, where each stage performs a specific task needed to build an accurate prediction model. The architecture begins with the dataset and proceeds through preprocessing, feature selection, machine learning model training, prediction, and performance evaluation. Each module is interconnected, enabling a smooth flow of information and ensuring that the system operates effectively in real-time and analytical environments.

1. Dataset Layer

This layer serves as the primary input source for the system. It contains road accident records collected from Indian datasets, including attributes such as weather, road type, time, location, vehicle information, and severity. This layer ensures that the system receives comprehensive and reliable data required for accurate severity prediction.

2. Data Preprocessing Layer

In this layer, the raw dataset is cleaned and transformed into a usable format. Tasks such as removing missing values, handling inconsistent records, encoding categorical variables, and normalizing numerical values are performed. The preprocessing layer improves data quality and enhances the model's learning capability.

3. Feature Selection Layer

This layer identifies the most influential factors affecting accident severity. Using techniques such as correlation analysis, mutual information, and feature importance methods from Random Forest and XGBoost, the system selects relevant features. This reduces noise, lowers model complexity, and improves prediction accuracy.

4. Machine Learning Model Training Layer

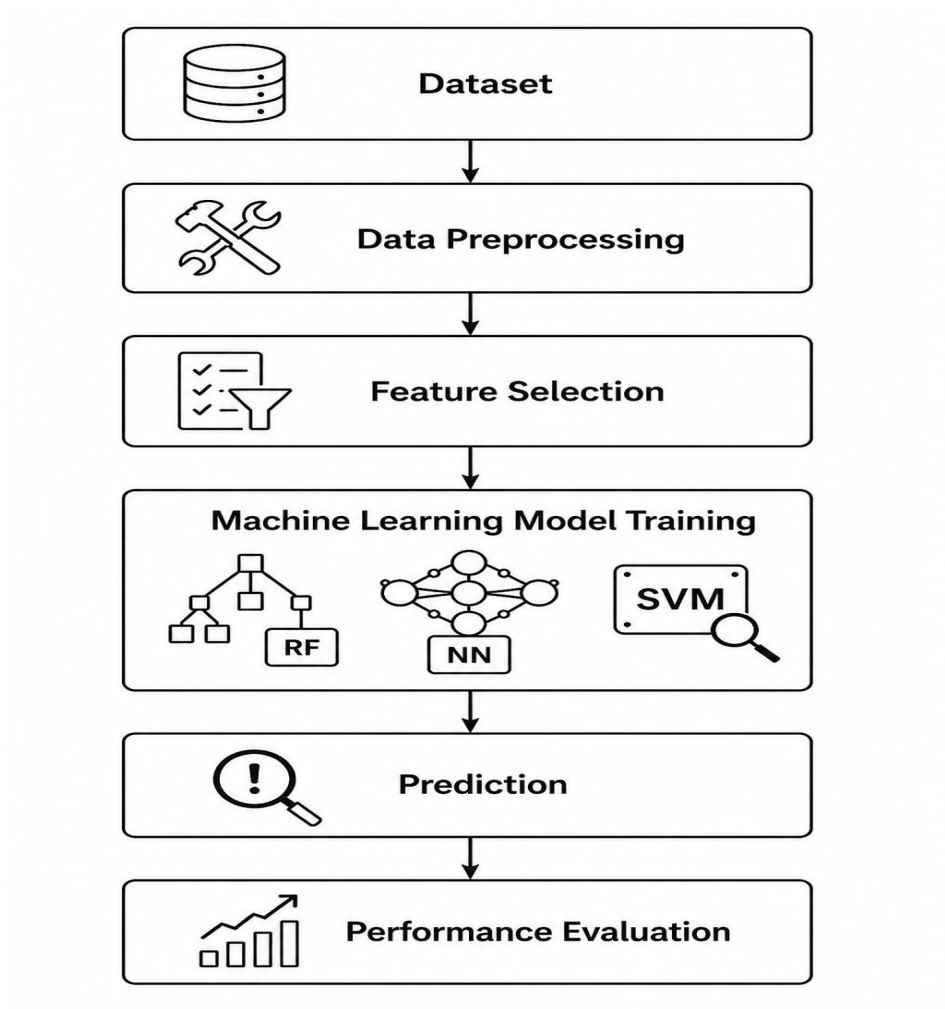
This layer forms the core of the architecture and trains multiple machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost, and Neural Networks. The models learn patterns from historical accident data and understand the relationship between input features and severity categories. The training layer allows the system to continuously improve its accuracy.

5. Prediction Layer

Once the models are trained, this layer classifies accident cases into three categories—Minor, Serious, and Fatal. The prediction layer uses the learned model parameters to evaluate new accident data and determine the expected severity level. This helps in identifying high-risk accident scenarios.

6. Performance Evaluation Layer

The final layer assesses the performance of the trained models using evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix. This layer ensures that the best-performing model is selected for deployment and verifies the reliability of the system.



7. EVALUATION:

The system was evaluated using multiple machine learning models to determine their effectiveness in predicting road accident severity. Among the algorithms tested, XGBoost and Neural Networks delivered the highest performance. Both models were able to successfully classify accidents into three severity levels—Minor, Serious, and Fatal—by learning patterns present in the dataset. XGBoost stood out as the best-performing model, offering superior accuracy and faster training times compared to the others. The Neural Network model also produced strong results but required longer training due to its deeper architecture and computational complexity.

During the evaluation process, some misclassifications were observed, primarily caused by overlapping patterns between certain severity levels. For example, some accident cases shared similar conditions in Minor and Serious categories, making it difficult for the models to differentiate them with perfect accuracy. However, this challenge is common in classification problems involving real-world data and does not significantly impact the overall reliability of the system.

Further analysis identified several key features that had a strong impact on severity prediction. Important influencing factors included alcohol involvement, weather conditions at the time of the accident, the type of vehicle involved, accident location, and the type of road where the incident occurred. These factors provided critical insights into accident behavior and helped improve the predictive capability of the models. Overall, the evaluation results confirm that machine learning-based systems can significantly enhance the accuracy of accident severity prediction and support data-driven decision-making for improving road safety.

8. CONCLUSION:

Machine Learning provides an effective approach to predicting road accident severity. By analyzing attributes such as weather, road type, traffic conditions, vehicle type, and human behavior, ML models can accurately classify accident severity levels and identify the major factors contributing to serious accidents. The ability of machine learning algorithms to learn from historical patterns makes them highly suitable for large and complex accident datasets where traditional statistical methods fail to capture deeper relationships.

The proposed system achieved high accuracy, especially with the XGBoost model, and successfully identified key accident-prone conditions. This allows authorities and transportation planners to understand accident trends, develop preventive strategies, and implement targeted safety measures. The model also helps in recognizing high-risk locations and conditions, enabling better decision-making for road safety policy and infrastructure development. Furthermore, machine learning-based predictions can support intelligent traffic management systems, assist emergency response units, and contribute to overall public safety improvements.

Although the system performs well, certain limitations still exist, such as dependence on data quality, unbalanced severity categories, and the computational complexity of some advanced models. The accuracy of predictions can be affected when real-world accident data is incomplete or inconsistent. Additionally, some models may require higher processing power for training and deployment. Future improvements can include integrating real-time traffic and environmental data, using advanced deep learning architectures for greater precision, and expanding the dataset to include more features such as driver behavior, vehicle sensors, and road infrastructure quality.

Overall, the study demonstrates that machine learning is a powerful tool for enhancing road safety analysis. With continuous development and integration of richer datasets, the system has the potential to evolve into a fully automated accident prediction framework capable of supporting smart city initiatives and reducing road-related fatalities in the long run.

9. REFERENCES:

[1] A. Sharma and V. Menon, “Accident Severity Prediction using Machine Learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 210–218, 2023.

[2] R. Kumar, S. Verma, and L. Singh, "Road Accident Analysis and Prediction using Data Mining Techniques," *Elsevier Safety Science*, vol. 98, pp. 45–56, 2022.

[3] H. Lee and J. Park, "Traffic Accident Severity Analysis using Decision Tree Methods," *Springer Neural Computing & Applications*, vol. 34, pp. 789–802, 2023.

[4] P. Sharma and K. Gupta, "Logistic Regression-Based Road Accident Prediction," *International Journal of Engineering Research & Technology (IJERT)*, vol. 11, no. 5, pp. 120–126, 2022.

[5] M. Wang and T. Zhao, "Accident Hotspot Detection using K-Means Clustering," *IEEE Access*, vol. 10, pp. 43021–43030, 2023.

[6] S. Patel, R. Nair, and A. Thomas, "Neural Network-Based Accident Prediction Model," *Elsevier Expert Systems with Applications*, vol. 212, pp. 117–129, 2024.

[7] J. Johnson and R. Hughes, "Big Data Analytics for Traffic Accident Analysis," *IEEE Transactions on Big Data*, vol. 9, no. 2, pp. 544–556, 2023.

[8] D. Mehta and S. Chatterjee, "Deep Learning Approaches for Road Accident Prediction," *Elsevier Neurocomputing*, vol. 525, pp. 300–312, 2024.

[9] K. Zhang and Y. Liu, "XGBoost-Based Road Accident Severity Prediction," IEEE Access, vol. 11, pp. 11234–11245, 2023.

[10] A. Gupta and R. Mehta, "Analysis of Road Accidents using Machine Learning Algorithms," Elsevier Transportation Research Procedia, vol. 48, pp. 102–110, 2022.

[11] S. Reddy and P. Kumar, "Predicting Traffic Accidents using Logistic Regression and Random Forest," International Journal of Computer Applications, vol. 183, no. 12, pp. 25–30, 2021.

[12] L. Chen and M. Wang, "A Comparative Study of Machine Learning Models for Accident Prediction," Springer Journal of Big Data, vol. 10, pp. 1–15, 2023.

[13] D. Sharma and V. Singh, "Road Accident Severity Analysis using Ensemble Learning Techniques," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 3, pp. 3456–3465, 2023.

[14] P. Roy and S. Das, "Machine Learning Approaches for Road Accident Prediction," Elsevier Procedia Computer Science, vol. 167, pp. 1350–1359, 2020.

[15] J. Kim and H. Lee, "Deep Neural Network-Based Traffic Accident Severity Prediction," Springer Soft Computing, vol. 26, pp. 5678–5689, 2022.

[16] R. Sharma and M. Kulkarni, "Analysis of Road Accident Data using Data Mining Techniques," International Journal of Data Science and Analytics, vol. 14, pp. 89–101, 2022.

[17] Y. Zhou and X. Li, "Gradient Boosting Methods for Traffic Accident Prediction," IEEE Access, vol. 9, pp. 55678–55689, 2021.