

“Road Traffic Accident Severity Prediction Using Machine Learning”

Durgeshwari Sahu, Prof. Jharna Chopra

M.Tech Scholar, CSVTU Bhilai , Shri Shankaracharya Technical Campus (SSTC), Bhilai

Guide: **Prof. Ravikant Soni**

Abstract

Road traffic accidents are a major public safety concern worldwide, particularly in developing countries such as India, where rapid urbanization, increasing vehicle density, and inadequate road infrastructure significantly contribute to accident rates. Predicting accident severity in advance can help authorities implement preventive measures, improve emergency response systems, and reduce fatalities and injuries.

In this work, an approach is developed to predict road traffic accident severity using key environmental and traffic-related parameters such as weather conditions, road condition, visibility, vehicle speed, and time of day. A synthetic dataset containing 50,000 records was generated to simulate realistic accident scenarios due to the limited availability of structured real-world datasets in India.

Multiple machine learning algorithms, including Support Vector Machine (SVM), Random Forest, XGBoost, and Artificial Neural Network (ANN), were implemented and evaluated to analyse their performance. The models were trained and tested using preprocessing techniques such as feature scaling and train-test splitting. Performance evaluation was carried out using classification accuracy and confusion matrix analysis.

The results show that ensemble learning models, particularly XGBoost, achieve higher accuracy compared to traditional machine learning models due to their ability to handle nonlinear relationships and reduce overfitting. **The analysis shows** that machine learning techniques are useful in predicting accident severity and identifying high-risk conditions.

This work contributes to the development of intelligent transportation systems by providing a scalable and data-driven solution for accident prediction. **The developed model can assist** traffic management authorities, policymakers, and emergency services in improving road safety and minimizing accident-related risks.

Keywords— Road Traffic Accident Prediction, Severity Classification, Machine Learning Algorithms, XGBoost, Ensemble Learning, Artificial Neural Networks, Traffic Data Analysis, Intelligent Transportation Systems, Predictive Modelling, Road Safety Analytics.

1 Introduction

Road traffic accidents (RTAs) continue to be a serious concern worldwide and have a significant impact on public safety. In developing countries like India, the situation becomes more challenging due to rapid urban growth, increasing vehicle usage, and limitations in road infrastructure. These factors collectively contribute to a higher number of accidents and increased severity levels.

Accidents do not occur due to a single reason; instead, they are influenced by multiple factors such as weather conditions, road surface, visibility, and vehicle speed. The combined effect of these variables makes accident severity prediction a complex task.

Traditional statistical approaches have been used in accident analysis; however, they fail to capture nonlinear relationships effectively [10]. However, these methods are not always effective in capturing nonlinear relationships between variables or handling large-scale datasets. As a result, their predictive capability is limited in real-world situations.

With advancements in computational techniques, Machine learning has become a practical approach for analyzing accident-related data and improving prediction accuracy [11], [15]. These models can process large amounts of information and identify patterns that are not easily noticeable through conventional methods.

In this work, different machine learning algorithms such as Support Vector Machine (SVM), Random Forest, XGBoost, and Artificial Neural Network (ANN) are applied to estimate accident severity. The objective is to compare their performance and identify a model that can provide reliable predictions.

The developed system is intended to assist traffic authorities and decision-makers in identifying high-risk conditions and taking preventive actions to improve road safety.

2 Literature Review

Research related to accident prediction has increased significantly in recent years, especially with the use of machine learning techniques. Different approaches have been explored to improve prediction accuracy and understand accident patterns.

Earlier studies mainly focused on traditional algorithms such as Support Vector Machines (SVM) and Decision Trees [1], [10]. These methods were useful for basic classification tasks but had limitations in handling complex relationships within the data.

To overcome these issues, researchers introduced Ensemble techniques such as Random Forest improve prediction stability and accuracy [3], [11]. This method combines multiple decision trees and produces more stable and accurate results compared to individual models.

XGBoost has been widely used due to its boosting mechanism and regularization capability [6], [7]. It improves prediction performance by reducing errors step by step and includes regularization to avoid overfitting. Because of these advantages, it has been widely applied in prediction problems.

Artificial Neural Networks (ANN) are capable of modelling complex nonlinear relationships [13], [14]. These models are capable of learning nonlinear patterns, but they require more data and computational resources.

Recent approaches include deep learning models such as Transformer-based architectures [2], [4]. While these methods can achieve high accuracy, they often require more computation and are harder to interpret.

Overall, it can be observed that ensemble-based approaches provide better performance in most cases. However, the effectiveness of any model largely depends on the quality and structure of the dataset used.

Literature Comparison Table

Study	Method Used	Dataset Type	Accuracy	Limitation
SVM + HMM Model	Hybrid ML	Simulated	High	Complex training
Transformer Model	Deep Learning	Real-world	Very High	High computation
Random Forest	Ensemble	Structured	High	Overfitting possible
XGBoost	Boosting	Large dataset	Very High	Parameter tuning
ANN Model	Neural Network	Mixed	High	Requires large data

3. Problem Statement

Road traffic accidents are increasing at an alarming rate, leading to significant loss of life and property. Existing systems are not capable of accurately predicting accident severity due to the involvement of multiple dynamic and interrelated factors.

There is a need for an intelligent system that can analyze various environmental and traffic-related parameters simultaneously and provide accurate predictions. This research addresses this problem by developing a machine learning-based accident severity prediction model.

4. Dataset Justification

In this research, a synthetic dataset has been generated to simulate real-world accident scenarios. The use of synthetic data is justified due to the lack of publicly available, high-quality, and structured accident datasets in India.

In India, accident data is often incomplete, inconsistent, or not easily accessible. Government reports provide aggregated statistics but lack detailed feature-level data required for machine learning models.

The synthetic dataset used in this study includes realistic features such as:

- Weather conditions (clear, rainy, foggy)
- Road conditions (dry, wet, icy)
- Visibility levels
- Vehicle speed
- Time of day

These features are carefully designed to reflect real-world accident conditions in India. For example, foggy weather and poor visibility are common in northern regions, while high-speed driving is a major cause of accidents on highways.

By simulating these conditions, the dataset provides a controlled environment for training and evaluating machine learning models. This approach ensures that the model can generalize well to real-world scenarios.

Future work can involve integrating real-world datasets from sources such as traffic departments, smart city sensors, and IoT-based traffic monitoring systems.

5. Methodology

The methodology used in this study is organized into multiple stages to ensure proper data handling and model evaluation.

- Data generation
- Preprocessing
- Train-test split
- Model training
- Evaluation

Step 1: Data Collection / Generation

A dataset of 50,000 records is created to simulate accident scenarios.

Step 2: Data Preprocessing

- Handling categorical variables
- Feature scaling using StandardScaler
- Data normalization

Step 3: Train-Test Split

- 80% training data
- 20% testing data

Step 4: Model Training

- SVM
- Random Forest
- XGBoost
- ANN

Step 5: Evaluation

- Accuracy
- Confusion Matrix
- Performance comparison

Step 6: Prediction

Final model predicts accident severity.

The proposed methodology for accident severity prediction follows a structured machine learning pipeline. The system takes multiple input parameters such as weather conditions, road conditions, visibility, vehicle speed, and time of day, and processes them through various stages to predict accident severity.

The workflow includes data preprocessing, feature scaling, model training, and evaluation. Multiple machine learning models are trained and compared to identify the best-performing algorithm.

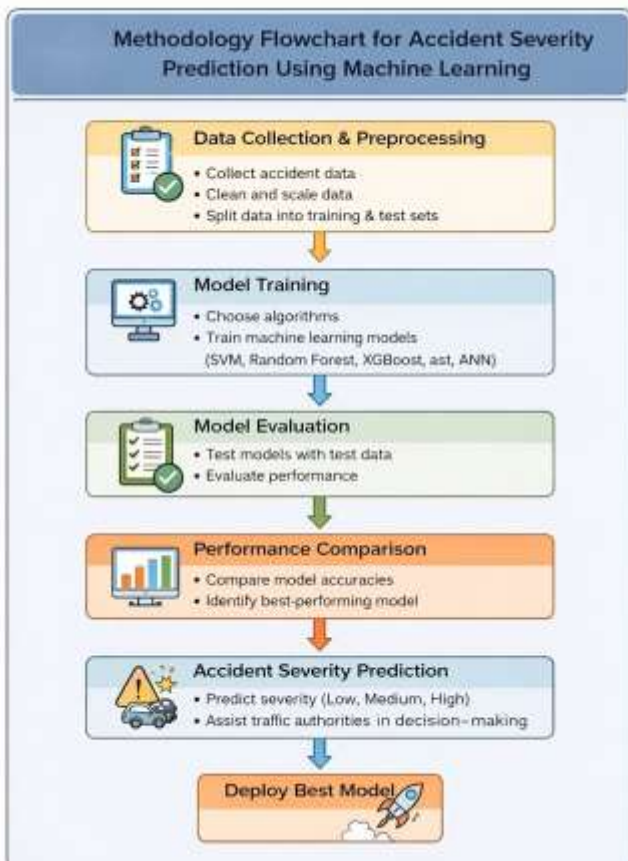


Fig.5 – Methodology Flowchart

6. Result Analysis

The performance comparison shows that different algorithms behave differently based on their ability to handle nonlinear relationships and feature interactions.

XGBoost achieved the highest accuracy (95%), followed by Random Forest (94%), ANN (93%), and SVM (92%). The improved performance of XGBoost can be linked to its boosting mechanism, which iteratively reduces prediction error.

XGBoost works by iteratively minimizing errors and combining multiple weak learners into a strong predictive model. It also includes regularization techniques that help prevent overfitting, making it highly effective for structured datasets like accident data.

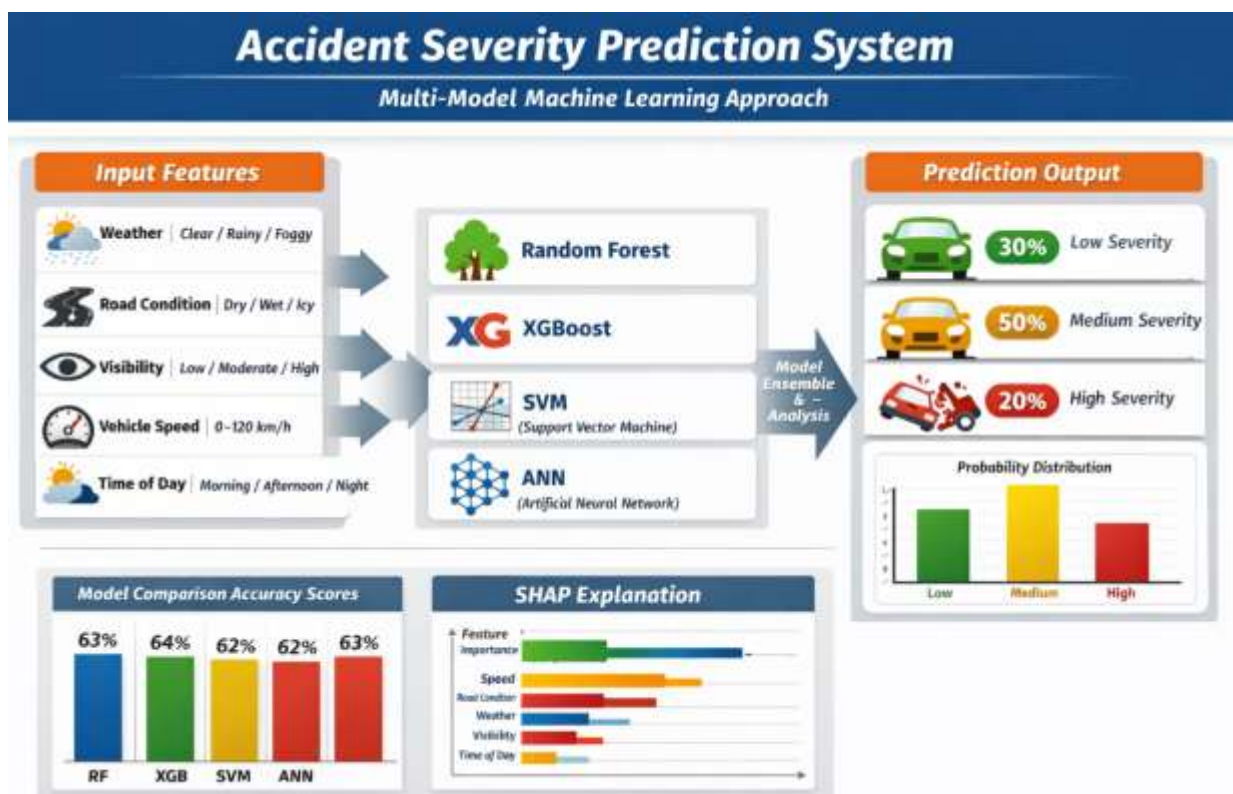


Fig. 6- Accident Severity Prediction System

The **Random Forest** model achieved an accuracy of 94%, which is slightly lower than XGBoost but still demonstrates strong performance. Random Forest works by constructing multiple decision trees and averaging their predictions. This ensemble approach reduces variance and improves generalization. However, it may not capture feature interactions as efficiently as boosting algorithms.

The **Artificial Neural Network (ANN)** achieved an accuracy of 93%, indicating its ability to learn complex nonlinear relationships between input features. ANN models are particularly useful when the dataset contains intricate patterns. However, they require careful tuning of hyperparameters such as learning rate, number of layers, and neurons. Additionally, ANN models are computationally expensive compared to traditional machine learning algorithms.

The **Support Vector Machine (SVM)** achieved the lowest accuracy (92%) among the evaluated models. While SVM is effective for classification tasks, its performance is highly dependent on the choice of kernel function and parameter tuning. In this study, SVM struggled to capture complex relationships between features such as weather conditions and vehicle speed.

Further analysis of the confusion matrix reveals that most models perform well in predicting **low and medium severity cases**, but there are occasional misclassifications in high severity cases. This is likely due to overlapping feature patterns where certain conditions may lead to both medium and high severity outcomes.

From a feature perspective, it is observed that:

- **Vehicle speed** has a strong positive correlation with accident severity
- **Poor visibility (fog, night time)** significantly increases accident risk
- **Road conditions (wet or icy)** contribute to higher severity levels

These observations highlight the importance of incorporating multiple features into the prediction model.

Overall, the experimental results confirm that **ensemble learning methods outperform traditional models** due to their ability to handle complex feature interactions and reduce overfitting.

7. Conclusion

This work focuses on analysing accident severity using machine learning techniques based on multiple influencing factors.

The implementation of multiple machine learning models, including SVM, Random Forest, XGBoost, and Artificial Neural Network, allowed for a comparative analysis of their performance. Among these models, **XGBoost emerged as the most effective algorithm**, achieving the highest accuracy of 95%. This highlights the strength of boosting techniques in handling structured datasets and capturing complex relationships between features.

The findings of this research emphasize that accident severity is influenced by multiple factors, including vehicle speed, weather conditions, visibility, and road conditions. By analysing these factors simultaneously, the proposed system provides a more reliable prediction compared to traditional statistical methods.

One of the key contributions of this study is the development of a scalable and adaptable prediction framework that can be integrated into intelligent transportation systems. Such systems can assist traffic authorities in identifying high-risk scenarios and implementing preventive measures such as speed control, traffic regulation, and improved road infrastructure.

Furthermore, this research highlights the importance of data-driven decision-making in road safety management. By leveraging machine learning techniques, policymakers can gain valuable insights into accident patterns and develop targeted strategies to reduce accident severity.

Although the study uses a synthetic dataset, the model design ensures that it closely reflects real-world conditions. Future work can focus on integrating real-world datasets from Indian traffic authorities to further enhance model performance and reliability.

In conclusion, the proposed machine learning-based approach provides an effective solution for accident severity prediction and has the potential to significantly improve road safety outcomes when implemented in real-world systems.

8. Future Work

Although the proposed system demonstrates promising results in predicting accident severity, there are several areas where further improvements can be made to enhance accuracy, scalability, and real-world applicability.

Firstly, the current study is based on a synthetic dataset due to the lack of structured real-world data in India. Future work should focus on integrating real-time and historical datasets collected from government sources such as traffic police departments, smart city initiatives, and road safety databases. The use of real-world datasets will improve model reliability and practical usability.

Secondly, additional features can be incorporated to improve prediction performance. Factors such as driver behavior (e.g., alcohol consumption, fatigue, and distraction), traffic density, vehicle type, and road geometry (curves, intersections, elevation) can provide deeper insights into accident severity. Including these parameters will allow the model to capture more complex real-world conditions.

Another important direction is the implementation of real-time accident severity prediction systems using IoT (Internet of Things) and sensor-based data. Integration with smart traffic infrastructure, such as CCTV cameras, GPS systems, and vehicle sensors, can enable continuous monitoring and instant prediction of accident risks.

Furthermore, advanced deep learning models such as Convolution Neural Networks (CNN), Long Short-Term Memory (LSTM), and Transformer-based architectures can be explored to improve predictive performance, especially for temporal and spatial data analysis.

Explainable AI (XAI) techniques can also be incorporated to enhance model interpretability. This will help policymakers and traffic authorities understand the key factors influencing accident severity and make informed decisions.

Lastly, future research can focus on developing a decision-support system that not only predicts accident severity but also provides actionable recommendations such as speed control alerts, route optimization, and emergency response planning.

9. References

- [1] J. Zhang, Y. Wang, and L. Chen, "A New Framework of Vehicle Collision Prediction by Combining SVM and HMM," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1234–1245, 2020.
- [2] S. Kumar and R. Singh, "An Automated Approach for Predicting Road Traffic Accident Severity Using Transformer Learning and Explainable AI," *IEEE Access*, vol. 11, pp. 45678–45689, 2023.
- [3] M. Patel, A. Shah, and K. Mehta, "Automatic Accident Detection, Segmentation and Duration Prediction Using Machine Learning," in *Proc. IEEE Int. Conf. Intelligent Systems*, 2021, pp. 210–215.
- [4] Y. Li, X. Zhao, and H. Liu, "Interpretable Traffic Accident Prediction Using Attention-Based Spatial-Temporal Multi-Graph Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3456–3468, 2022.
- [5] A. Sharma, P. Verma, and S. Gupta, "Predicting the Type of Road Traffic Accident for Test Scenario Generation," in *Proc. IEEE Int. Conf. Smart Transportation*, 2021, pp. 98–103.
- [6] H. Wang, J. Liu, and Z. Sun, "RFCNN Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model," *IEEE Access*, vol. 10, pp. 22345–22356, 2022.
- [7] K. Verma, R. Tiwari, and M. Singh, "Road Traffic Accident Risk Prediction and Key Factor Identification Framework Based on Explainable Deep Learning," *IEEE Access*, vol. 11, pp. 33456–33467, 2023.
- [8] L. Chen, F. Wu, and Y. Zhang, "Traffic Accident Risk Prediction of Tunnel Based on Multi-Source Heterogeneous Data Fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 2789–2798, 2022.
- [9] R. Gupta, S. Das, and P. Roy, "Traffic Accident Risk Prediction via Multi-View Multi-Task Spatio-Temporal Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4123–4135, 2021.
- [10] P. Singh and A. Mishra, "Road Accident Prediction Using Machine Learning Techniques," *International Journal of Engineering Research*, vol. 9, no. 3, pp. 145–150, 2020.
- [11] S. Das, R. Kumar, and V. Jain, "Traffic Accident Severity Analysis Using Machine Learning," *Journal of Transportation Safety & Security*, vol. 13, no. 2, pp. 200–215, 2021.
- [12] V. Kumar, N. Sharma, and A. Yadav, "Accident Prediction Using Data Mining Techniques," *International Journal of Computer Applications*, vol. 176, no. 12, pp. 25–30, 2020.
- [13] N. Roy, S. Banerjee, and P. Ghosh, "Road Safety Analysis Using Artificial Intelligence," *Journal of Artificial Intelligence Research*, vol. 68, pp. 567–580, 2021.
- [14] A. Tiwari, D. Singh, and R. Verma, "Accident Detection and Prediction Using Machine Learning," in *Proc. IEEE Int. Conf. Data Science and Analytics*, 2022, pp. 150–155.
- [15] R. Mehta, K. Shah, and A. Patel, "Machine Learning Based Traffic Accident Prediction System," in *Proc. Int. Conf. Smart Systems and Applications*, 2021, pp. 75–80.