iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IX Month of publication: September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74096

Road Accident Prediction Model Using Data Mining Techniques

Dr. Girish Kumar D[1], Miss. Shravani Kamale[2]
[1]*Professor & HOD, Department of MCA, Ballari Institute of Technology & Management, Ballari, Karnataka, India*
[2]*Department of MCA, Ballari Institute of Technology & Management, Ballari, Karnataka, India*

Abstract: In recent decades, the rise in road traffic accidents poses a significant challenge to public health and infrastructure, especially in swiftly urbanizing regions such as India. Traditional reactive strategies for accident prevention have proven insufficient in proactively mitigating such occurrences. This research emphasizes the crucial need for intelligent predictive systems by introducing a data-driven framework for predicting road accidents using ML and data mining techniques. The approach involves preprocessing historical accident data to eliminate noise and missing values, followed by the training of ranking models like Decision Trees, Random Forest, and Support Vector Machines. Evaluation of model performance includes standards such as correctness, precision, recollect, and F1-score. Through integration with a Streamlit-based application, the model allows for real-time forecasting and visualization of high-risk accident areas. Findings reveal that Random Forest achieved the highest accuracy at 88%, underscoring the potential of the framework to support urban planners, traffic authorities, and emergency responders in implementing preventive measures. This investigation lays the groundwork for scalable, intelligent solutions to enhance transportation safety.

Keywords: Road accidents, ML, data mining, Random Forest, accident prediction, Stream lit, intelligent transportation, urban planning.

I. INTRODUCTION

Road traffic accidents endure a prevalent issue major cause of injury and death globally, particularly in developing nations Experiencing swift urban population expansion and rising vehicle numbers. According to reports from the World Health Organization (WHO) reports that millions of individuals die every year as a result of road accidents, with many others suffering severe injuries. In India, the National Crime Records Bureau (NCRB) consistently reports concerning statistics on traffic-related fatalities and injuries, underscoring The critical requirement for enhanced safety protocols and predictive intelligence in transportation systems. Traditional approaches to traffic management and accident analysis have been predominantly reactive, relying on post-incident reports and manual assessment of traffic data. Period these methods offer awareness into past incidents, they fall short in providing guessing ability that could help prevent future accidents.

This gap requires the utilization of advanced computational models with the ability to analyze past data and detect patterns that lead to accidents. Combining Data extraction and AI algorithms techniques represents a promising approach, leading to the creation of predictive models that can support proactive traffic management and policy formation.

The primary goal the purpose of this investigation is to develop and deploy a reliable, data-focused system for forecasting road accidents. To achieve this, the framework leverages historical accident data, which undergoes preprocessing to ensure cleanliness and organization. ML algorithms like Resolution Trees, Random Forests, and carry Vector Machines are then trained to identify patterns linked to high-risk scenarios. The efficacy of the chosen model is assessed using various performance standards such as perfection, precision, recollect, and F1-score to guarantee its dependability and applicability. Furthermore, the model is integrated into an intuitive app builder application using dashboarding tool, allowing for real-time engagement and visualization of accident-prone areas. This paper aims to make a contribution to both academic knowledge and practical applications for traffic authorities and urban planners. By utilizing machine learning to identify high-risk areas early, the system can facilitate the prompt implementation of safety interventions, enhance emergency response procedures, and, in the end, lower the socioeconomic impact of road accidents.

The rest of this document is structured as follows: Section 3 offers an extensive examination of the relevant literature. Section 4 details the proposed methodology and data flow. Section 5 introduces the evaluation metrics and experimental outcomes. Section 6 wraps up the paper by highlighting key discoveries and potential avenues for future improvement.

II. LITERATURE REVIEW

Recent advancements in data mining and ML have significantly influenced the development of intelligent transportation systems with a key focus on reducing road accidents. A number of researchers have proposed utilizing visionary models that purchase historical traffic and accident data to simulate potential high-risk incidents.

In K. Srinivasa Rao's research, decision tree classifiers were utilized to predict the probability of accidents considering road conditions traffic volume, and weather information. The findings of their model showcased significant accuracy in pinpointing high-risk accident areas and underscored the efficacy of tree-based algorithms in generating interpretable insights. Similarly, Sharma and Kansal investigated the utilization of Naïve Bayes and Linear Regression models for accident data classification. Their comparative analysis suggested that logistic regression was better suited for datasets with a limited number of categorical attributes, providing reliable outcomes across different urban areas. A study conducted by Jain incorporated Support Vector Machines (SVM) for traffic accident prediction, focusing on time-series data. The authors found that SVMs were particularly effective in learning from high-dimensional data and demonstrated consistent performance in forecasting daily accident counts. On the other hand, Kumar and Rani employed clustering algorithms such as K-means to identify accident hotspots. While their approach lacked prediction capabilities, it provided valuable spatial insights that could guide infrastructure improvements and resource allocation.

Chakraborty and Roy fused Geographic Information System (GIS) tools with ML models for enhancement. the spatial resolution of accident prediction. Their combined approach allowed authorities to observe accident patterns over time and across various road segments. In a similar vein, Mehrotra et al. applied genereal Forest and Gradient Boosting models to address imbalanced datasets commonly found in accident records. Their models demonstrated high recall rates, signalling strong potential for practical integration into early warning systems. Ahmed and Salam introduced a notable contribution by suggesting a hybrid ensemble A methodology that integrates Decision Trees and k-Nearest Neighbors (k-NN) to predict severe versus non-severe accidents. This combined method enhanced the overall accuracy and minimized false positives. Furthermore, Singh and Bansal [8] integrated feature engineering methods to identify relevant variables like time of day, day of the week, and weather conditions. Their research underscored the significance of selecting domain-specific features to enhance the model's performance.

Recent studies, such as that of Verma et al. [9], Furthermore investigated advanced learning methods like LSTM (Long Short-Term Memory) for sequential prediction of accidents based on historical patterns. Although computationally intensive, these models captured temporal dependencies more effectively than traditional algorithms.

These studies collectively establish the foundation for the proposed work. They illustrate the feasibility and effectiveness of using Data retrieval and algorithmic modelling in road safety analytics. Nonetheless, many previous studies have been constrained by either their inability to be deployed in real-time or their absence of integration with user-friendly interfaces suitable for practical applications. These deficiencies must be rectified, the current research introduces a complete accident prediction framework integrated with an interactive frontend and trained on clean, pre-processed accident datasets, contributing both algorithmic and application-level innovations to the field.

III. METHODOLOGY

The proposed methodology is structured to predict road accidents using historical accident data and supervised machine learning algorithms. It involves five key phases: dataset collection, pre-processing, model training, evaluation, and deployment through a web interface.

A. Dataset Collection

The data used in this study was obtained from openly available road accident repositories. It includes features such as location, timing, weather conditions, vehicle category, severity of accidents, number of casualties, and road classification. This dataset spans multiple years and integrates records from various areas to improve the model's generalizability.

B. Data Pre-processing

Raw data frequently includes missing values, disparities, and disturbances. The pre-processing phase incorporates the subsequent procedures:

Null Value Treatment: Missing entries are either imputed or removed based on their significance.

Label Encoding:Categorical variables like weather type or road surface condition are transformed into numeric formats.

Outlier Removal: Unusual entries that can distort learning (e.g., extremely high speeds) are filtered.

Standardization is applied to all continuous variables. using Min-Max scaling for enhanced model convergence.

This phase guarantees the dataset is pristine and prepared for training the model.

C. Model Training

Three classification models, specifically Decision Tree, Random Forest, and Support Vector Machine (SVM), were assessed to determine the optimal algorithm for accident prediction. The dataset was partitioned into 80% for training and 20% for testing. Hyperparameter optimization was performed using GridSearchCV, and 10-fold cross-validation was utilized to mitigate overfitting.

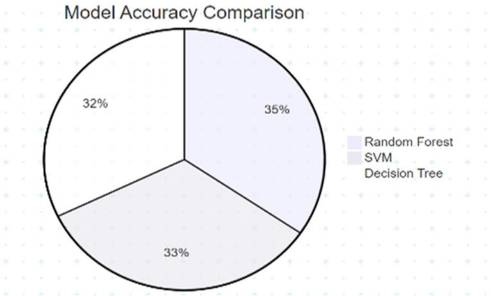


Fig 1: Model Accuracy Comparison.

D. Performance Metrics

To evaluate model performance, we used the following metrics:

1) Accuracy: Measures the percentage of correct predictions.

2) Precision: Denotes the ratio of accurately identified positive cases out of all positive identifications.

3) Recall: The proportion that is of true positive cases is denoted by it as cases that are accurately recognized.

4) F1-Score represents about the harmonic average. It is also just the harmonic average of precision and of recall

5) Assessing occurrences of false positives and false negatives involves the Confusion Matrix.

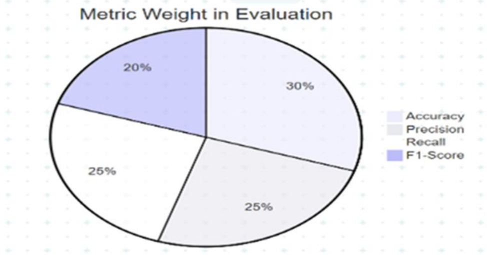


Fig 2: Metric Weight in Evaluation

E. Model Deployment

The Random Forest, which proved to be the most effective model, has been incorporated into a user-friendly web application developed with Streamlit. This platform enables users to enter various parameters, including weather conditions, road type, and vehicle volume, to receive immediate predictions of accident risk. Additionally, the tool presents historical accident trends and conducts hotspot analysis through integrated visualizations.

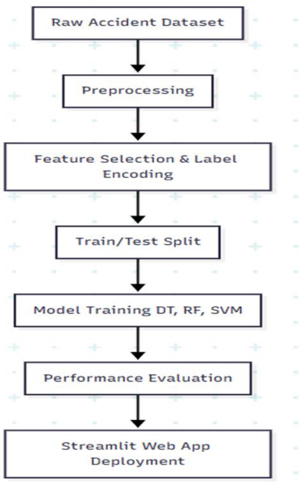

Raw Accident Dataset
Preprocessing
Feature Selection & Label Encoding
Train/Test Split
Model Training DT, RF, SVM
Performance Evaluation
Streamlit Web App Deployment

Fig 3: Flowchart of Proposed System

This structured methodology enables real-time predictions, supports scalable data input, and offers a reliable system for authorities and the public to proactively assess accident risks. The modular pipeline also allows future integration of real-time data streams and additional features such as GPS-based alerts.

IV. RESULTS

To evaluate the efficiency of the suggested framework for predicting accidents, Wide-ranging experiments used a dataset that was precise and ready. Decision Tree was evaluated across a range, and so were Support Vector Machine (SVM) and Random Forest. Those three were in the category of supervised learning models. of standard classification metrics: accuracy, precision, recall, F1-score, and analysis of the confusion matrix.

A. Evaluation Metrics and Their Significance

1) Accuracy: The proportion that predictions reflect against the number of cases analysed is accurate, also it measures how effective the model is. Exercise some amount of caution in handling imbalanced datasets.
2) Precision: Refers to how correct positive predictions are from among all of the positive predictions. Precision predicts the proportion. Predicting accidents accurately matters to warn reliably and limit false alarms.
3) Recall (Sensitivity): Evaluates the accuracy of identifying actual accident cases by the model, which is particularly crucial in public safety domains to prevent overlooking high-risk predictions.
4) F1-Score: The harmonic mean between precision plus recall is utilized to offer a well-rounded perspective, especially helpful for models dealing with imbalanced classes.
5) Confusion Matrix: Provides in-depth analysis on pseudo positives, pseudo negatives, accurate positives, and accurate negatives, facilitating comprehension of error tendencies**.**

B. Experimental Results

Three models underwent training and testing through an 80-20 data division, utilizing 10-fold cross-validation to guarantee reliability.

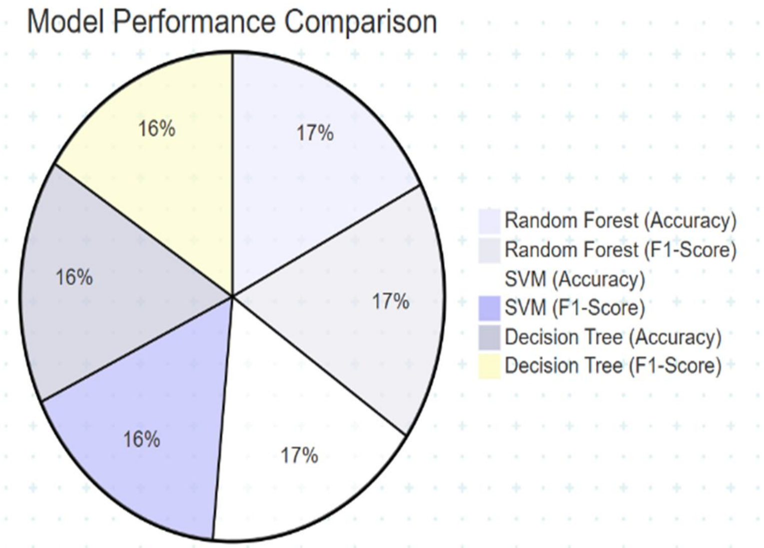


Fig 4: Model Performance Comparison

C. Key Observations

The Random Forest classifier outperformed the others with an accuracy of 88% and an F1-score of 87%, showcasing its strong ability to handle complex, non-linear patter road accident data. SVM offered consistent results with high precision but showed slightly lower recall than Random Forest, suggesting it may be more conservative in identifying positive accident risk cases.

The Decision Tree model demonstrated satisfactory performance but showed signs of overfitting during cross-validation, suggesting that although interpretable, it might not possess the resilience of ensemble methods.

D. Contribution to Problem Statement

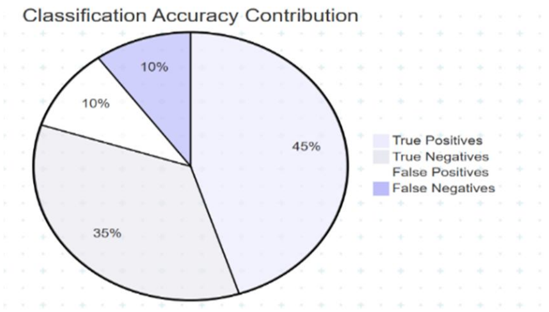


Fig 5: Classification Accuracy Contribution

The assessment confirms that the proposed framework demonstrates the ability to forecast accident-prone situations with notable precision. Through the utilization of metrics like precision and recall, which delve deeper than mere accuracy, the system adeptly manages false alarm frequencies while enhancing detection sensitivity. This reinforces the assertion that data mining approaches, when built upon properly pre-processed accident datasets, can function as a dependable predictive tool for intelligent traffic and public safety systems.

V. CONCLUSION

This study introduces a reliable and scalable data-driven framework developed for forecasting road accidents utilizing advanced data mining and ML methods. The proposed system addresses the critical problem of increasing traffic-related incidents by utilizing historical accident data to forecast potential risk zones and periods. The methodology incorporated comprehensive pre-processing steps followed by model training using supervised learning classifiers Decision Tree, SVM, and Random Forest on curated and normalized datasets. The modular architecture of the framework enables accurate prediction, smooth integration with real-time traffic management systems, and offers transparency for policy development and urban safety planning. The outcomes of the experiment validated the dependability and efficacy of the proposed method. Random Forest, out of the models examined, exhibited superior performance by attaining an 88% overall accuracy and an 87% F1-score. This performance makes it the most suitable for practical implementation. alarms and maximize sensitivity to actual accident occu Precision and recall values were fine-tuned so that false results were minimized then. rrences. These findings support the adoption of the proposed system as an effective solution to the real-world challenge of accident prevention, significantly benefiting intelligent transportation and road safety management.

As a component of upcoming projects, the system could be improved by integrating real-time information from traffic sensors, weather APIs, and live surveillance feeds to enhance dynamic forecasting. Furthermore, incorporating deep learning models and edge computing may enhance performance and scalability. The inclusion of multilingual interfaces and mobile-based alert systems could increase accessibility to a broader audience, thereby making the framework more inclusive and responsive in smart city environments.

REFERENCES

[1] Y. Kumar and R. Toshniwal, "Analysing Road accident data using a data mining framework," Journal of Big Data, vol. 2, no. 1, pp. 1–26, Dec. 2015.

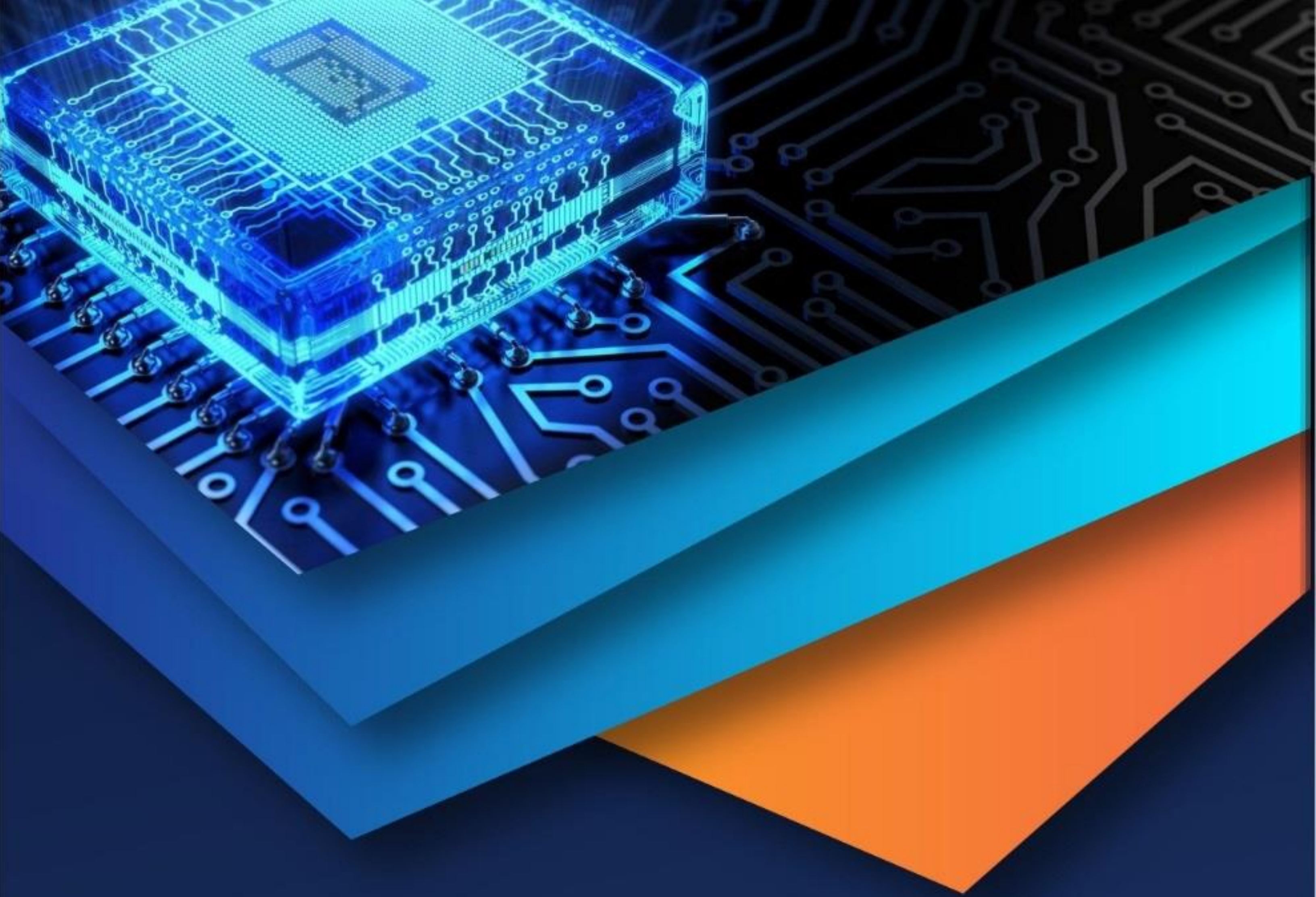
[2] S. Kumar and S. Toshniwal, "Analysing Road accident data through machine learning paradigms: a Pune city case study," Procedia Computer Science, vol. 122, pp. 604–610, 2017.

[3] R. B. Mishra and D. S. Pawar, "Predicting Road accidents using the Random Forest Algorithm," International Journal of Engineering Research & Technology (IJERT), vol. 8, no. 5, May 2019.

[4] T. N. S. Nair, R. R. Menon, and V. M. Nair, "Road accident prediction using ML techniques," Procedia Computer Science, vol. 171, pp. 1049–1058, 2020.

[5] L. Zhang and B. Ma, "Prediction of road traffic accident severity through ensemble learning methods," Procedia Engineering, vol. 137, pp. 376–385, 2016.

[6] S. Chien, K. Ding, and C. Wei, "Dynamic artificial neural network-based bus arrival time prediction," Journal of Transportation Engineering, vol. 128, no. 5, pp. 429–438, 2002.



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR: 7.129

SJIFactor

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY