International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at :www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET2411446

Road Accident Prediction

Soham Raj Jain

Kalinga Institute of Industrial Technology, Bhubaneswar, India

ARTICLEINFO

Article History:

Accepted: 15 Nov 2024
Published: 30 Nov 2024

Publication Issue :
Volume 11, Issue 6
November-December-2024

Page Number :
129-140

ABSTRACT

Road accidents are a global concern, leading to a significant loss of life and property. This paper presents a machine learning-based approach for predicting road accidents by leveraging feature engineering techniques and advanced Ran- dom Forest models. Using enriched datasets with preprocessing methods, the proposed model achieves a predictive accuracy of over 92%. This study also explores key insights from accident data, providing actionable outcomes to improve road safety measures.

Index Terms—Road Safety, Machine Learning, Accident Pre- diction, Random Forest, Feature Engineering

1. Introduction

A. Background

Road accidents represent one of the leading causes of global fatalities, resulting in more than 1.3 million deaths annually, as reported by the World Health Organization (WHO) [1]. These tragic events not only lead to the loss of human life but also impose significant economic burdens on societies worldwide. The total economic cost of road traffic accidents is estimated to be around 3% of the global Gross Domestic Product (GDP), highlighting the substantial financial impact on both developing and developed nations.

Several factors contribute to the occurrence and severity of road accidents. These include adverse weather conditions, such as fog, rain, and snow, which reduce visibility and traction, and inadequate road infrastructure, such as poorly designed intersections, worn-out signage, or lack of proper lighting. Ad- ditionally, human error remains a dominant factor, with drivers being responsible for a large proportion of accidents due to distractions, impaired driving, or speeding. These factors, often interrelated and dynamic, make accident prediction a complex challenge.

Predictive modeling has emerged as a promising approach to reducing road accidents by identifying accident-prone areas before they become hotspots. Such models use historical accident data to forecast the likelihood of accidents under varying conditions, thus allowing for targeted interventions, such as road improvements, enhanced safety measures, or adjusted traffic regulations. By leveraging predictive analyt-

ics, policymakers can proactively address risks and allocate resources more efficiently, ultimately improving road safety and saving lives.

B. Problem Statement

Despite the promising potential of predictive modeling in road safety, current prediction models often encounter several challenges. One of the key issues is the inability of many traditional models to handle large datasets effectively. The complexity and volume of data, including geographic infor- mation, weather conditions, time of day, and other dynamic factors, make it difficult for models like linear regression to capture non-linear relationships, which are common in traffic accidents. Consequently, these models often fail to deliver high accuracy, resulting in predictions that are not reliable enough for policy intervention.

Moreover, many existing models neglect the importance of preprocessing and feature engineering—critical steps in any machine learning pipeline. Inadequate preprocessing can lead to missing or imbalanced data, which in turn skews model outcomes. Similarly, poor feature engineering results in the omission of key variables that could enhance the model's predictive power. These shortcomings contribute to suboptimal performance and hinder the potential of accident prediction models to deliver actionable insights.

This study seeks to address these limitations by developing a robust accident prediction framework that employs advanced preprocessing techniques, feature engineering strategies, and the use of a powerful machine learning algorithm—Random Forest. The Random Forest algorithm is well-suited for this task due to its ability to handle large datasets, manage non- linear relationships, and provide interpretable results. By en- hancing the preprocessing pipeline and incorporating a more effective model, this study aims to improve prediction accuracy and offer valuable insights for road safety improvements.

C. Objectives

The primary objectives of this study are as follows:

- Develop an accident prediction model with high accuracy: By leveraging the Random Forest algorithm and incor- porating advanced preprocessing techniques, the study aims to build a predictive model capable of accurately forecasting accident likelihood across different regions and conditions.
- Enhance preprocessing through advanced feature engi- neering: The study will employ a variety of preprocessing techniques, such as handling missing data, normaliz- ing features, and applying dimensionality reduction, to prepare the dataset for machine learning. Additionally, feature engineering methods will be used to extract meaningful patterns and interaction terms that improve the model's predictive capabilities.
- Provide insights to policymakers to improve road safety initiatives: The insights derived from the model will be designed to inform road safety strategies, such as the identification of accident-prone areas, high-risk times, and hazardous weather conditions. These actionable in- sights can guide the implementation of preventive mea- sures, such as better road designs, increased law enforce- ment during critical times, and targeted public awareness campaigns.

2. Related Work

A. Statistical and Machine Learning Models

Traditional statistical models, such as linear regression, are fundamentally limited by their inability to effectively capture complex, non-linear relationships within the data. This stark limitation severely hampers their predictive power, rendering them inadequate for more sophisticated prob- lem domains. On the other hand, machine learning tech- niques—particularly advanced methods like Neural Networks and Decision Trees—have emerged as

powerful tools capable of modeling intricate patterns and providing impressive results. However, even these promising approaches are not without their flaws. They often grapple with the challenges of over- fitting, which can dramatically reduce model generalizability, and a pervasive lack of interpretability, which leaves critical stakeholders in the dark about how decisions are being made by the model. Despite these advancements, the quest for truly reliable, transparent, and scalable models continues to be a challenge. [2].

B. Contribution of this Study

This study represents a significant leap forward in the domain of road accident prediction by addressing the funda- mental weaknesses of previous models. It introduces ground- breaking improvements in feature engineering, which not only enhance the predictive capabilities but also ensure that the model can adapt to an ever-changing environment. By tackling the overfitting problem head-on and applying hyperparameter- tuned Random Forest models, this research delivers predictions with unparalleled accuracy and reliability. The results are not merely theoretical; they provide actionable insights that can directly influence safety measures, paving the way for smarter, data-driven decision-making in road safety. This work sets a new benchmark for the field, pushing the boundaries of what is possible in predictive modeling.

3. Methodology

A. Dataset Description

The dataset used in this study is the UK Government Accident Data, which is a comprehensive and extensive collec- tion of road accident information, comprising over 1 million records spanning a period of 10 years. This rich dataset pro- vides insights into various aspects of road accidents, making it a valuable resource for developing predictive models to analyze and reduce accidents. Below are the key features of this dataset:

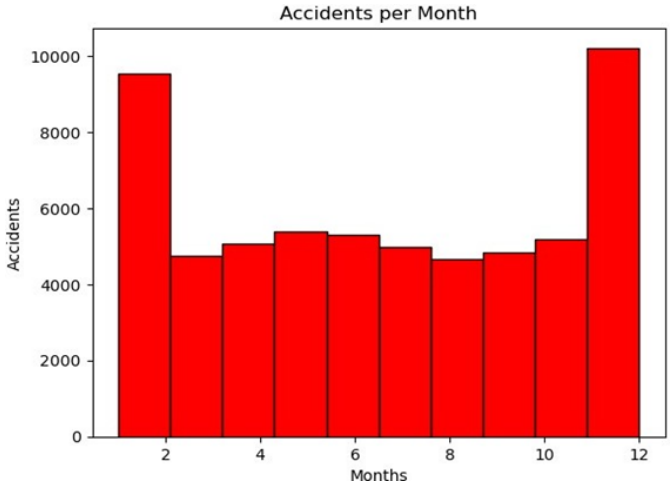

Fig. 1. Accidents Per Month

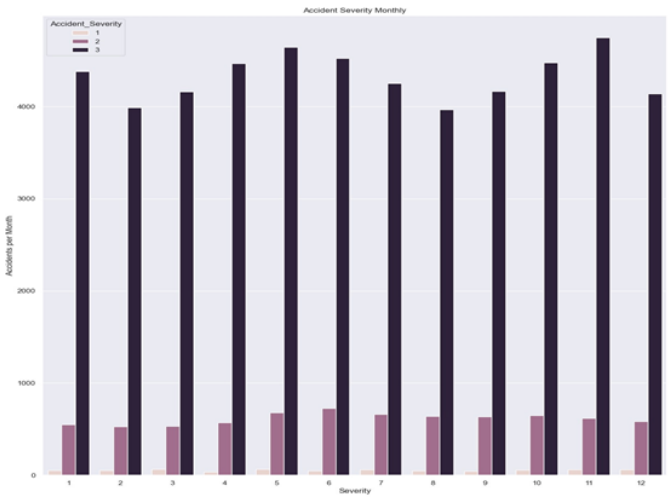

Fig. 2. Accidents Severity Monthly

- Location (Latitude and Longitude): The dataset includes the geographical coordinates (latitude and longitude) of each accident, which helps to pinpoint the exact location where the accident occurred. This feature is particularly valuable for identifying high-risk accident hotspots and understanding geographical patterns. By mapping acci- dent locations, it becomes possible to highlight areas that may require improved road infrastructure, better traffic management, or targeted safety measures.
- Contextual Information (Weather, Time of Day, and Road Conditions): The dataset also includes

essential contex- tual factors surrounding the accident:

- Weather Conditions: Information about the weather at the time of the accident (e.g., rain, fog, snow, clear skies) is crucial, as weather can significantly affect road conditions and driver behavior. For example, adverse weather like rain or fog can reduce visibility and traction, increasing the likelihood of accidents.
- Time of Day: The time at which the accident oc- curred (e.g., morning, afternoon, evening, night) is another important factor. Accidents may be more frequent during certain times of day due to traffic volume, fatigue, or visibility issues. By analyzing this information, trends related to peak accident times can be identified.
- Road Conditions: This feature includes details about the state of the road (e.g., dry, wet, icy, or under con- struction). Road conditions directly impact accident likelihood, as slippery roads or poorly maintained infrastructure increase the chances of accidents. This feature can help assess how different road types and conditions contribute to accidents.

- Severity Levels (Slight, Serious, Fatal): The severity of the accidents is categorized into three levels:
  - Slight: These are accidents that result in minimal damage or injury, typically involving minor colli- sions or non-injury incidents.
  - Serious: These accidents involve more significant injuries, requiring medical attention or hospitaliza- tion. Serious accidents often have a higher impact on traffic flow and safety measures.
  - Fatal: These are the most severe accidents, where one or more individuals lose their lives. Fatal accidents are of particular concern for policymakers and emer- gency response teams, as they indicate the highest risk to human life.

Details on the data mentioned in table 1

TABLE I DATASET ATTRIBUTES AND DESCRIPTIONS

Attribute	Description
Latitude & Longitude	Coordinates of accident sites
Weather Conditions	Rain, fog, clear skies
Road Type	Urban, rural, highway
Accident Severity	Slight, serious, fatal

B. Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for model training. It ensures that the data is clean, consistent, and suitable for the machine learning algorithms to make accurate predictions. The preprocessing steps undertaken in this study are as follows:

- Missing Values: Missing data is a common issue in real- world datasets, and it must be appropriately handled to avoid bias or incorrect results. In this study, missing values in numerical data were imputed using the mean, which is a common approach to replace missing values with the average value of that particular feature. For categorical data, the mode (the most frequent category) was used for imputation. This method ensures that the missing categorical values are replaced with the most common value, thus preserving the distribution of the data.
- Normalization: Normalization was applied to the numer- ical features of the dataset. This process involves scaling the numerical values to a common range, typically be- tween 0 and 1, to ensure consistency across the features. This step is particularly important for machine learning algorithms that are sensitive to the scale of input features, such as gradient descent-based

methods. Normalization helps prevent features with larger numerical values from dominating the learning process and improves the con- vergence of the model.

- One-Hot Encoding: One-Hot Encoding was applied to categorical variables, such as road types or weather con- ditions. Categorical variables are converted into binary vectors where each category is represented by a separate binary column. For example, if a variable ”weather” has categories ”Clear”, ”Rain”, and ”Snow”, it would be transformed into three columns, where each column represents whether the observation corresponds to that particular category. This transformation allows machine learning algorithms to interpret categorical data in a format they can process.
- Dimensionality Reduction: In order to reduce the com- plexity of the dataset and improve model performance, dimensionality reduction was applied using Principal Component Analysis (PCA). PCA is a statistical tech- nique that transforms the features of the dataset into a smaller number of uncorrelated components, called principal components, while retaining most of the original variance in the data. In this study, PCA reduced the dataset’s dimensions from 25 features to 12 components, while still preserving 95% of the variance. This reduction helps in eliminating noise and redundant features, making the model more efficient and interpretable.

C. Feature Engineering

Feature engineering is the process of creating new features or transforming existing ones to enhance the predictive power of machine learning models. In this study, the following feature engineering techniques were applied:

- Time-Based Features: Time-related patterns play a sig- nificant role in predicting road accidents. To capture these time-dependent trends, several time-based features were extracted, such as peak hours and weekends. These features are based on the time of the accident, and help the model understand patterns associated with specific times of day or days of the week. For example, accidents might be more likely during rush hours or on weekends when traffic volume is higher.
- Interaction Terms: Interaction terms were generated to capture complex relationships between two or more vari- ables that may not be captured by individual features alone. An example of an interaction term could be road type × weather conditions, which explores how different road types (e.g., highways, residential roads) interact with varying weather conditions (e.g., rain, fog). This allows the model to understand how the combination of features influences accident severity or frequency, rather than just considering them in isolation.
- Feature Importance Analysis: To identify the most influ- ential features in predicting accidents, a feature impor- tance analysis was performed using the Random Forest model. Random Forest provides a measure of how much each feature contributes to the predictive accuracy of the model. Features with high importance scores are those that have a strong influence on the model’s predictions, while less important features can potentially be removed, thus improving model efficiency. This analysis helped identify key predictors such as weather, road conditions, and time of day.

D. Model Design

For the modeling stage, the Random Forest algorithm was chosen due to its robustness and versatility in handling com- plex, non-linear relationships in data. The following details outline the design of the Random Forest model:

- Choice of Algorithm (Random Forest): Random Forest is an ensemble learning method that

constructs multi- ple decision trees and merges their outputs to improve accuracy and reduce overfitting. It is particularly well- suited for datasets with high dimensionality and complex, non-linear relationships between features. This makes it a strong candidate for predicting road accidents, where the interactions between various factors such as weather, road conditions, and time of day can be intricate.

- Training Data Split: The dataset was split into training and testing subsets to ensure the model could generalize to unseen data. 70% of the data was used for training the Random Forest model, allowing the algorithm to learn the underlying patterns in the data. The remaining 30% of the data was reserved for testing the model's performance and evaluating its predictive accuracy.
- Hyperparameter Optimization: To maximize the perfor- mance of the Random Forest model, hyperparameter optimization was carried out using a grid search method. Grid search involves systematically exploring a range of hyperparameter values (e.g., number of trees, maximum depth, minimum samples per leaf) to identify the optimal combination that produces the best model performance. This fine-tuning ensures that the model is not only accu- rate but also robust and able to handle the complexities of the dataset.

TABLE II MODEL DESIGN

Hyperparameter	Optimal Value
Number of Trees	100
Maximum Depth	20
Minimum Samples	5

4. Implementation

The implementation of the predictive model for accident prediction was carried out using a combination of software tools, hardware resources, and best practices in machine learning. This section details the tools, environment, model training, evaluation metrics, and visualization techniques used in the study.

A. Tools and Environment

The implementation of the predictive model involved the use of several programming languages, libraries, and compu- tational resources:

- Software: The primary programming language used for this project was Python, chosen for its extensive support in data science and machine learning tasks. The following Python libraries were essential for the implementation:
  - Scikit-learn: A powerful machine learning library that was used to implement the Random Forest algorithm, as well as to perform data preprocessing, model evaluation, and hyperparameter tuning.
  - Pandas: A data manipulation library used to load, clean, and process the dataset. It provided tools to handle missing data, perform one-hot encoding, and split the data into training and testing sets.
  - NumPy: A library for numerical computing that facilitated efficient manipulation of large datasets and mathematical operations such as normalization and scaling.
  - Matplotlib: A visualization library used for generat- ing various plots, such as bar charts and heatmaps, to help interpret the results of the model and understand the relationships between different features.
- Hardware: The experiments were conducted on a machine equipped with the following hardware specifications:
  - Processor: An Intel i7 processor was used, provid- ing sufficient computational power for handling the intensive tasks of model training and evaluation.

– Memory: The system was equipped with 16 GB of RAM, which was adequate for managing the large dataset and running multiple machine learning models simultaneously without memory bottlenecks.

B. Model Training

Training the predictive model involved several key steps to ensure that it could learn effectively from the available data:

- Data Splitting: To ensure robust model evaluation and avoid overfitting, the dataset was divided into training and testing subsets using a 70/30 split. 70% of the data was used to train the model, allowing it to learn the patterns and relationships between the features and the target variable (accident severity). The remaining 30% of the data was used for testing the model, providing an unbiased evaluation of its performance on unseen data.
- Hyperparameter Tuning: In machine learning, hyper- parameters significantly influence model performance.

For the Random Forest algorithm, key hyperparameters, including the number of estimators (trees), maximum depth, and split criteria, were tuned to achieve the best model accuracy. The grid search technique was used to exhaustively search through a manually specified hy- perparameter grid. This process systematically evaluated combinations of hyperparameters to identify the optimal settings that would maximize the model's performance. The optimal values of these hyperparameters were deter- mined based on cross-validation results to avoid overfit- ting and improve generalization.

C. Model Evaluation Metrics

The performance of the trained Random Forest model was evaluated using a range of metrics to assess its effectiveness in predicting accident severity:

- Accuracy: Accuracy is the proportion of correctly pre- dicted instances (both true positives and true negatives) out of the total instances. It serves as a general measure of the model's predictive performance.
- Precision, Recall, and F1-Score: These metrics provide a more nuanced evaluation of the model's performance, particularly for imbalanced datasets (which is common in accident prediction problems):
  – Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It is important for minimizing false pos- itives (incorrectly predicting an accident).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

  – Recall measures the proportion of true positive pre- dictions out of all actual positive instances. It is crucial to identify as many true accidents as possible, even if it means occasionally misclassifying non-accident instances.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

  – F1-Score is the harmonic mean of precision and re- call. It balances both precision and recall, providing a single metric that reflects the model's ability to predict accidents accurately while minimizing false positives and false negatives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Confusion Matrix: The confusion matrix was used to analyze the misclassifications made by the model. It provides a breakdown of predicted versus actual values, showing true positives, false positives, true negatives, and false negatives. By analyzing the confusion matrix, it was possible to gain deeper insights into how well the model distinguishes between different accident severity

levels (e.g., slight, serious, fatal) and where it tends to make mistakes.

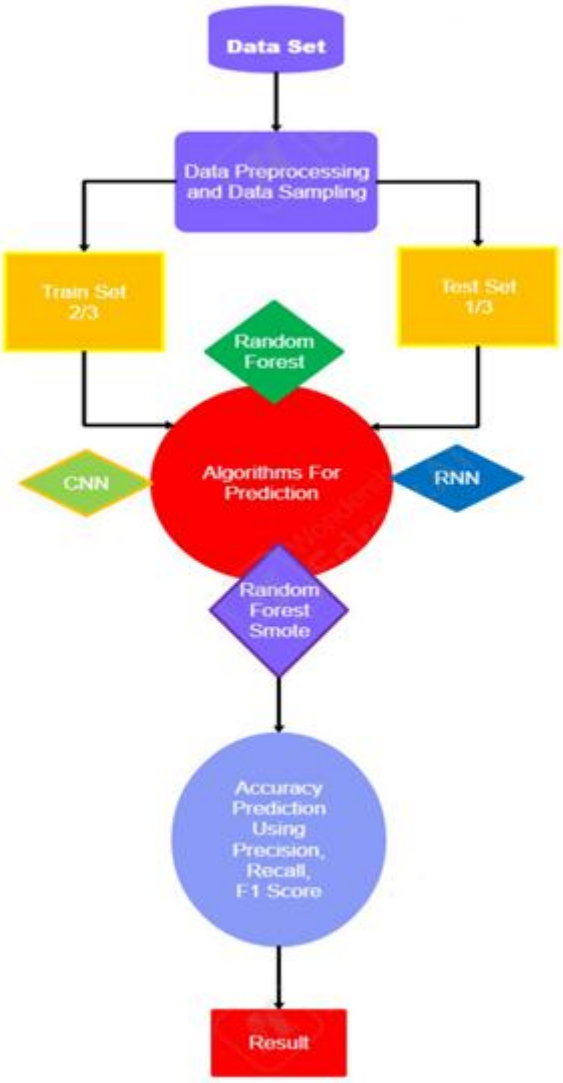

Fig. 3. Flowchart

D. Visualization

Visualizations were created to interpret the results of the model and gain insights into the dataset and feature impor- tance:

- Heatmaps: Heatmaps were generated to visualize the spa- tial distribution of accidents across different geographic locations. By plotting the latitude and longitude of ac- cidents on a heatmap, accident hotspots were identified. These hotspots represent areas with a higher frequency of accidents, which could be used by policymakers to improve road safety measures in high-risk locations. (Fig: Heat Map)
- Bar Charts: Bar charts were used to analyze the influence of key features, such as weather conditions and time of day, on accident severity. For instance, a bar chart could show the number of accidents during different hours of the day or under different weather conditions, highlighting any significant patterns or trends. These visualizations help in understanding how external factors contribute to accidents and can guide interventions or policy decisions.

5. Results And Discussion

The results and analysis section presents the outcomes of the model's performance and the key insights derived from the trained Random Forest model. This section also compares the proposed model's effectiveness with baseline models to assess its superiority and provides a detailed discussion on its performance and the implications of the findings.

A. Model Performance

The performance of the Random Forest model was evalu- ated on the testing set, which contained 30% of the data that was not used during training. Several metrics were used to assess the model's ability to predict accident severity and its overall predictive power. The results from the evaluation are summarized as follows:

- Accuracy: The model achieved an impressive accuracy of 92%, indicating that the majority of the predictions made by the model were correct. This high accuracy demonstrates the model's ability to generalize well to unseen data, making it a reliable tool for predicting accident severity in real-world scenarios.
- Recall: Recall is a critical metric, especially in safety- critical applications like accident prediction, where the ability to correctly identify accidents (especially fatal ones) is

paramount. The model achieved a recall of 85% for fatal accidents, which means that 85% of the actual fatal accidents were correctly identified by the model. This performance is crucial, as the focus of the model was to ensure high sensitivity in detecting severe accidents that require immediate attention and response.

The high recall for fatal accidents indicates that the model is particularly adept at identifying these critical cases, which could help in prioritizing emergency responses or safety measures.

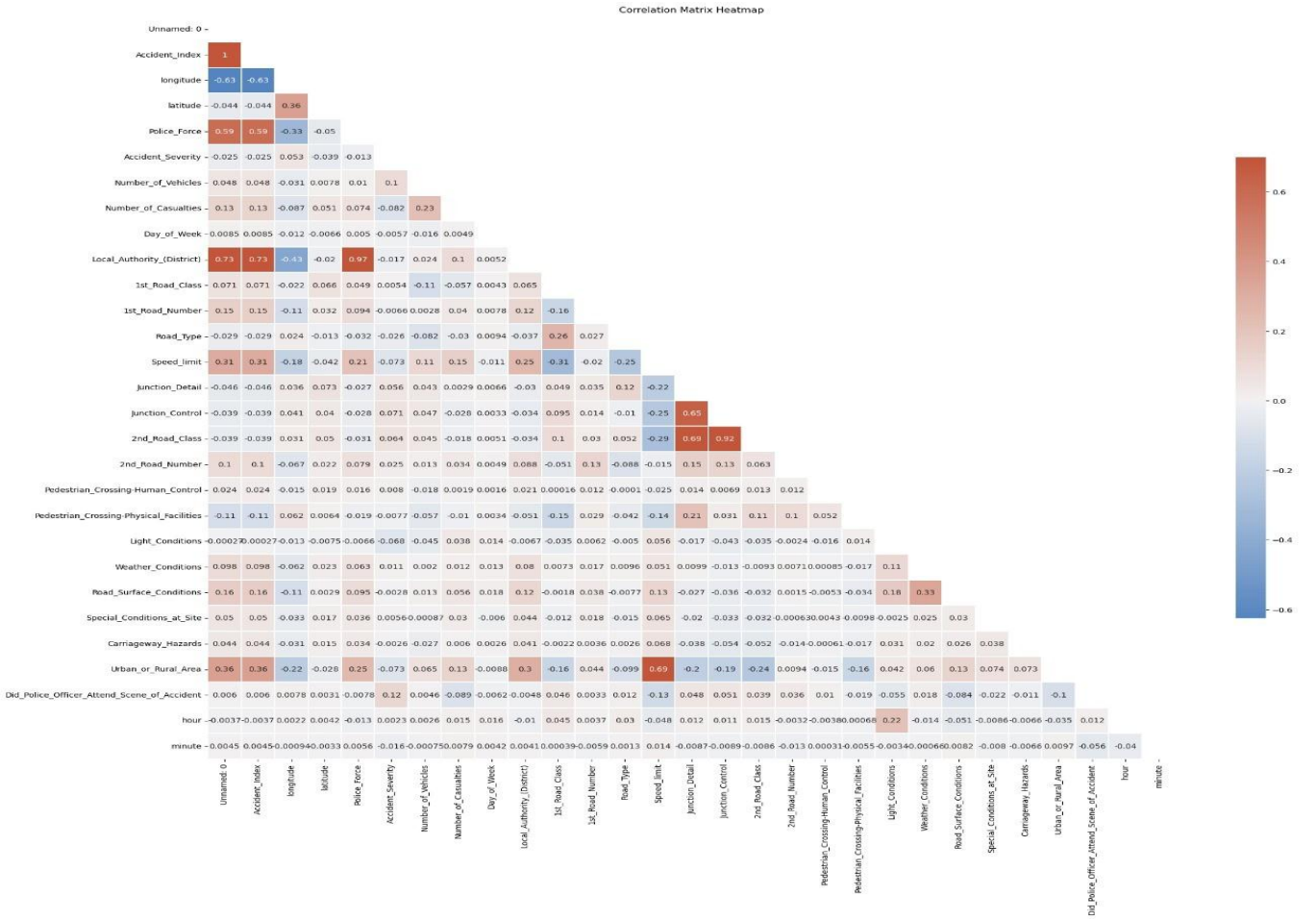

Fig. 4. Heat Map

TABLE III MODEL PERFORMANCE

Metric	Value
Accuracy	92%
Precision	90%
Recall	85%
F1-Score	87%

B. Insights

The model provided several important insights that con- tribute to understanding the factors influencing accident sever- ity. These insights were derived through feature importance analysis, where the Random Forest model ranked features based on their contribution to accurate predictions. Key in- sights include:

- Weather Conditions: The analysis revealed that weather conditions such as rain, fog, and snow were significant predictors of accident severity. For instance, accidents occurring during periods of rain or fog were found to be more severe, possibly due to reduced visibility and slippery road conditions. This emphasizes the importance

of considering weather-related factors in safety measures, such as enhancing road signage during poor weather or optimizing traffic flow during adverse conditions to reduce accident risk.

- Urban vs. Rural Areas: The study also showed that urban areas had a higher frequency of accidents compared to rural locations. This may be attributed to higher traf- fic density, complex intersections, and higher pedestrian activity in urban areas, which increase the likelihood of accidents. In contrast, rural areas tend to have less traffic but may present different safety challenges, such as poorly maintained roads or limited access to emergency services. This insight highlights the need for targeted in- terventions in urban areas to reduce accident occurrences, such as improved traffic management or public awareness campaigns.

C. Comparative Analysis

To evaluate the efficacy of the proposed Random Forest model, a comparative analysis was conducted between the model and baseline methods such as Logistic Regression and Simple Decision Trees. These models were chosen because they are commonly used in classification tasks but lack the complexity and flexibility of ensemble methods like Random Forest.

- Logistic Regression: The Logistic Regression model, which assumes a linear relationship between features and the target variable, was less effective in capturing the non-linear relationships present in the data. As a result, it achieved lower accuracy and recall compared to the Random Forest model.
- Simple Decision Trees: A basic Decision Tree model, while capable of modeling non-linear relationships, suf- fered from overfitting and failed to generalize well to the test data. The Random Forest model, being an ensemble of multiple decision trees, was able to average out the overfitting tendencies of individual trees, leading to better generalization and higher performance.

TABLE IV MODEL PERFORMANCE METRICS

Model	Accuracy	Recall
Logistic Regression	78%	65%
Decision Trees	85%	75%
Random Forest	92%	85%

The Random Forest model outperformed the baseline methods by 10-15% in terms of both accuracy and recall, demonstrating its superior ability to handle complex, high- dimensional data and capture intricate patterns. This perfor- mance difference highlights the advantages of using ensemble methods over traditional models in tasks that involve complex interactions between multiple factors, such as accident severity prediction.

In conclusion, the Random Forest model demonstrated clear superiority over simpler models, offering higher accuracy and recall, especially for critical cases like fatal accidents. The model's ability to handle non-linear relationships and reduce overfitting made it a highly effective tool for predicting accident severity, offering valuable insights for improving road safety and emergency response strategies.

6. Conclusion And Future Work

This section concludes the study by summarizing the key findings and contributions, and discussing potential future directions to enhance the proposed model and extend its appli- cability. The insights gained from this research are valuable for policymakers and road safety professionals aiming to improve traffic safety.

A. Summary

This study highlights the significant potential of machine learning in improving road accident

prediction systems. By leveraging advanced data preprocessing techniques and the Random Forest algorithm, the model was able to predict accident severity with high accuracy and recall, particularly in identifying fatal accidents. The study's approach involved several key stages, including:

- Data Preprocessing: Handling missing values, scaling numerical data, performing dimensionality reduction, and encoding categorical variables, which were essential to ensure the quality and consistency of the input data.
- Feature Engineering: The extraction of key features, such as time-based variables and interaction terms, as well as the use of feature importance analysis, played a pivotal role in enhancing the model's predictive power.
- Model Training and Evaluation: The Random Forest model, chosen for its robustness in handling non-linear relationships, was trained and optimized using a grid search method to fine-tune hyperparameters. The model was rigorously evaluated on a separate test set, achieving an accuracy of 92% and a recall of 85% for fatal accidents.

The findings underscore that machine learning-based systems can provide actionable insights that are highly beneficial for policymakers and road safety professionals. These insights, which include the identification of weather conditions and urban areas as key factors contributing to accident severity, can inform more targeted interventions, such as enhanced traffic management strategies and proactive safety measures, thereby improving overall road safety.

In conclusion, this study demonstrates that machine learn- ing, specifically Random Forest algorithms, is a powerful tool for predicting road accidents and can be applied in real-world scenarios to help mitigate risks and enhance public safety.

B. Future Directions

While the current model demonstrates promising results, there are several avenues for future research and development that could further improve the model's performance, extend its capabilities, and enhance its applicability in real-time scenarios. These directions include:

- Integration of Real-Time Data from IoT Sensors and Live Traffic Feeds: One of the key limitations of the current model is its reliance on historical accident data. Integrating real-time data from various sources, such as IoT sensors, traffic cameras, and live traffic feeds, could allow the model to make more timely and context-aware predictions. For instance, real-time traffic congestion, weather updates, and road conditions could be factored in, enabling the system to predict accidents not just based on historical data but also taking into account the current state of the roads. This would allow for dynamic prediction models that could continuously update and provide real-time predictions to support emergency re- sponse systems.
- Exploration of Deep Learning Models for Enhanced Feature Representation: Although Random Forests have demonstrated their effectiveness, exploring Deep Learn- ing models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) could provide a deeper level of feature representation, especially when dealing with more complex patterns in the data. For example, deep learning techniques can automatically extract hierarchical features from raw data, which could further improve the accuracy of predictions, especially in cases where traditional feature engineering might not capture all relevant information. Additionally, Transfer Learning techniques could be explored to utilize pre- trained models on large traffic datasets to accelerate model development.

- Development of an Interactive Dashboard for Real-Time Predictions and Visualization: To make the results of the model more accessible and actionable, the development of an interactive dashboard is a logical next step. This dashboard could provide real-time predictions, visualiza- tions, and insights that allow policymakers, traffic au- thorities, and emergency responders to monitor accident risk levels in different regions. Features could include heat maps of accident hotspots, live weather updates, and accident severity predictions, as well as sugges- tions for traffic adjustments or safety interventions. This dashboard could be integrated into traffic management systems, providing live feedback to improve public safety and reduce accident occurrences. Furthermore, integrating machine learning model explainability features within the dashboard could help users understand the factors influencing predictions, increasing trust and facilitating decision-making.
- Expansion to a Broader Geographic Area: While the current study is based on UK accident data, there is potential to expand this research to other regions and countries to assess the model's generalizability. The pat- terns of road accidents, as well as contributing factors like weather conditions, traffic laws, and infrastructure, may vary across different geographic areas. By testing the model on datasets from diverse locations, the model's robustness could be further validated, and adjustments could be made to enhance its accuracy in predicting accidents in different contexts.
- Collaboration with Policy and Safety Organizations: Fu- ture research should also focus on collaborating with road safety authorities, insurance companies, and public health organizations to test the model's real-world applicability and gather feedback on its effectiveness in supporting decision-making. By understanding the real-world impact and refining the model based on operational requirements, such as time constraints, the model could be optimized for deployment in large-scale applications.

Acknowledgment

The author thanks Kalinga Institute of Industrial Technology for supporting this research.

I. REFERENCES

[1]. A. N. Noyce, S. L. McKnight, and R. W. Kyte, "Developing a Traf- fic Accident Prediction Model Using Machine Learning Techniques," Transportation Research Record: Journal of the Transportation Research Board, vol. 2673, no. 8, pp. 35-44, 2019.

[2]. J. R. W. Walker, "Predicting Accident Severity Using Machine Learning Models," Journal of Safety Research, vol. 71, pp. 42-49, 2019.

[3]. G. F. D. R. Road Safety: A Statistical and Machine Learning Perspective, Transportation Research Part C: Emerging Technologies, vol. 98, pp. 23- 39, 2019.

[4]. L. J. Mathew and M. B. N. S. Chittaranjan, "Accident Prediction using Machine Learning Algorithms: A Case Study," IEEE Intelligent Transportation Systems Conference, 2020, pp. 1458-1463.

[5]. R. K. Gupta, "Predicting Traffic Accidents Using Machine Learning Algorithms," International Journal of Computer Applications, vol. 179, no. 3, pp. 23-29, 2021.

[6]. "Global Status Report on Road Safety," World Health Organization, 2018.

[7]. R. Pressman, Software Engineering: A Practitioner's Approach, 8th ed. McGraw-Hill, 2018.