National College of Ireland

Uncovering the Causes and Preventive Measures of Road Traffic Accidents in Kerala

MSc Research Project
Data Analytics

Jebitta Joseph
Student ID: x23151196

School of Computing
National College of Ireland

Supervisor: Prof.Abdul Qayum

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jebitta Joseph
Student ID:	x23151196
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Prof.Abdul Qayum
Submission Due Date:	12/12/2024
Project Title:	Uncovering the Causes and Preventive Measures of Road Traffic Accidents in Kerala
Word Count:	7360
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Uncovering the Causes and Preventive Measures of Road Traffic Accidents in Kerala

Jebitta Joseph
x23151196

Abstract

Road accidents is the current issue that occurred on a large scale and lead to high fatalities, and injuries and economic losses on an international level. This paper discusses the possibility of using a machine learning approach to forecast road accidents on roads as the traditional statistical measures like linear regression,moving averages etc have relatively fail to capture non-linear behaviour and trends in the accidents.The current systems lack comprehensive use of creative technologies such as machine learning (ML) and large language models (LLMs) for predicting and preventing RTAs, even though such technologies can easily analyze large databases and identify concealed patterns To improve the forecast precision and provide results that can be put into use, the study uses established techniques including Gradient Boosting Regressor, Random Forest, SARIMA. The presented model uses feature engineering, dimensionality reduction and ensemble for detecting most significant predictive characteristics and making prognosis of accidents.These findings may be useful to design effective preventive programmes tailored on students' needs to enhance road safety policies, in an attempt to mitigate the high socio-economic burden associated with road traffic accidents and support a evidence-based approach to solve traffic-related safety issues at the global level.
Keywords: Road Accident Analysis, Machine learning, LLMs, Data Analysis, Exploration of Data, Streamlit

1 Introduction

Road accidents are a pressing global issue, responsible for a substantial number of fatalities, injuries, and economic losses each year. The World Health Organization (WHO) estimates that road traffic accidents are among the leading causes of death worldwide, disproportionately affecting young people and economically productive age groups.Kumar and Harikrishna (2023) Beyond the tragic loss of life, road accidents impose significant societal and financial burdens, including healthcare costs, loss of productivity, and long-term physical and emotional trauma for victims and their families.Nassir et al. (2024) As urbanization accelerates and vehicle ownership grows, the complexities of traffic systems continue to increase. Factors such as inadequate infrastructure, human error, and non-compliance with traffic regulations exacerbate the risks of accidents. Despite efforts to mitigate these challenges through policies and safety measures, the rising trends in accident occurrences underscore the need for innovative, data-driven approaches to address this issue effectively .Understanding the dynamics of road accidents involves analyzing a wide range of factors, including vehicle types, accident severities, temporal patterns, and

geographic distributions.Bisht and Tiwari (2023) Traditional statistical methods, while useful, often fall short in capturing the intricate and non-linear relationships inherent in accident data. Machine learning techniques, with their ability to handle complex datasets and uncover hidden patterns, offer a promising alternative. This research seeks to harness the power of machine learning to analyze and predict accident trends. By leveraging advanced algorithms and data-driven methodologies, the study aims to provide actionable insights that can inform policymakers, stakeholders, and urban planners. These insights have the potential to guide the development of targeted interventions, enhance road safety measures, and ultimately reduce the incidence and impact of road accidents.Ajaykrishnan et al. (2024)

1.1 Research Problem

Despite the availability of extensive datasets on road accidents, identifying meaningful patterns and accurately predicting accident outcomes remains a significant challenge. The inherent complexity and non-linear nature of accident-related data hinder the effectiveness of traditional statistical methods, which often fail to capture intricate interactions and dependencies among variables. This limitation results in suboptimal predictive accuracy and inadequate insights for policymakers and urban planners. Therefore, there is a pressing need for advanced analytical frameworks and methodologies capable of addressing these complexities to enhance the understanding of accident dynamics and support evidence-based decision-making for road safety improvementThomas et al. (2024)
The motivation behind this study lies in the alarming rise in road accidents globally and the devastating consequences on human lives and economies. Developing accurate predictive models and actionable insights can significantly contribute to reducing accidents and improving road safety measures. Machine learning techniques offer a promising avenue for uncovering hidden patterns and relationships in accident data, which can drive targeted interventions.Choudhary et al. (2024)

1.2 Research Background

The increasing availability of accident datasets from various sources provides an opportunity to leverage data-driven approaches for road safety analysis. Previous studies have focused on using statistical methods and basic machine learning models, but these approaches often fail to address the complexity and high dimensionality of accident data.Vinoth et al. (2024) This research builds on existing work by applying advanced ensemble learning methods and time-series forecasting to enhance predictive accuracy and interpretability.

1.3 Research Solution

This study proposes a comprehensive framework utilizing machine learning models, including Gradient Boosting Regressor, Random Forest, and time-series models like Seasonal AutoRegressive Integrated Moving Average, to analyze and predict road accident trends. By incorporating techniques such as feature engineering, Principal component analysis for dimensionality reduction, and ensemble methods, the research aims to improve model performance and generate actionable insights. The solution also emphasizes

the importance of deploying results through interactive dashboards and Large Language Model based assistants to bridge the gap between analysis and real-world application.

1.4 Research Questions

The research seeks to address the following key question:

RQ1: Which machine learning models provide the more reliability in forecasting accident trends across varying datasets and regions?

RQ2: How can advanced machine learning techniques be utilized to analyze and predict road accident trends, providing actionable insights for improving road safety and reducing accident-related impacts?

RQ3: What factors contribute most significantly to road accident occurrences, and how can these factors be effectively mitigated through predictive analytics?

In this report the section 1 is an introduction to the reserch.Section 2 gives the related works .In section 3 a detailed methodology implemented in the reserach is expalined.Section 4 gives the design Specification and Section 5 explains implementation and section 6 and 7 gives Results and Evaluation and Conclusion and Future Works respectively.

2 Related Work

Understanding the dynamics of road accidents involves analyzing a wide range of factors, including vehicle types, accident severities, temporal patterns, and geographic distributions. Traditional statistical methods, while useful, often fall short in capturing the intricate and non-linear relationships inherent in accident data. Machine learning techniques, with their ability to handle complex datasets and uncover hidden patterns, offer a promising alternative.

2.1 Analysis and Factors Contributing to Road Traffic Accidents (RTAs)

Road Traffic Accidents (RTAs)Chand et al. (2024) article are hazards that threaten public health because of the being so indecipherable. The year 2022 witnessed 4,61,312 such RTAs occurring in India alone causing injures for 4,43,366 people and the deaths of 1,68,491.This research study also aims to analyze available data on RTAs through visualization techniques to know the possible causes, understand the effects concerning various vehicle types, and propose practical measures to influence the accidents in question. All the data from 2005 to 2022 were carefully collected in two stages, which utilized the yearly data about accidents taken from the Kerala State Crime Records Bureau along with First Information Sheets (FIRs). The first phase of this research established that fault of the driver played a major role (70 - 90%) for accidents among the various types of vehicles. The number of individuals involved was fairly high per accident in heavy vehicle RTAs (Avg. 1.42), which also had higher fatality rates. In the second phase these from analysis of FIRs brought to notice some priority areas for immediate intervention aimed at reducing the distraction of drivers.The paper's limitations include reliance on potentially biased historical data, a regional focus limiting generalizability, and an emphasis on visualization over predictive or actionable insights.

In Kerala, one would find that road transportation dominates the scene Beevi and Arya (2023). Planning urbanization, industrialization, and population explosion increases the compulsory use of motor vehicles and causes increasing accidents. It creates against traffic congestion and environmental pollution, etc. It has been found that every year at least 3881 major fatal accidents occur due to multi-vehicle accidents in Kerala. This study is an effort at analyzing the trend in Kerala with regard to the growth of motor vehicles and the associated motor vehicle accidents. It aims also to analyze the prime causes behind many accidents and suggest measures to reduce the number of motor vehicle accidents and so on.This paper by is its focus on correlation without fully addressing causal relationships between vehicle growth and accidents, and it lacks detailed implementation strategies for sustainable transportation solutions.

As for the recent Road Accidents Statistics (MORTH, 2020) Ajaykrishnan et al. (2024), it has stated that among the total road accidents reported in 2020, 15.8 percent involved pedestrians, while the remaining 65.1 percent occurred on straight road sections. Clearly indicates the difficulty and lack of safety for pedestrians.This study analyzes traffic safety of crossing pedestrians through proactive approach that would bring out a potential observable non-crash event that should have led to a conflict. Post-Encroachment Time (PET) would be the most popular time-based measure least time-consuming for an accurate estimation of surrogate safety. This paper attempts to identify the factors influencing the PET of pedestrian crossings at mid-block sections with designated or undesignated bus stops. Videographic surveys were carried out at two mid-block sections with bus stop in kerala where accidents to pedestrian occur frequently. Kinovea software was used for data extraction, and the information then analyzed with IBM SPSS software to outline the major determinants of PET.Lack of actionable frameworks or detailed strategies to implement sustainable transportation solutions effectively is the main drawback of this paper.

412,432 accidents 153,972 deaths 384,448 injuries happened on Indian roads in 2021 Vimalathithan et al. (2024) . India, having the highest number of road deaths, accounts for 11 percent deaths worldwide in road fatalities. Therefore, it becomes essential to find out the reasons behind accidents on Indian roads. This study aims to find out the factors involved in accidents in India using clustering analysis based on self-organizing maps (SOM) and, further provides some countermeasures based on the identified factors. The study used the accident data for India, collected by the members of ICAT-ADAC, developing and managing the entire operation of the ICAT-RNTBCI joint-approved project by the Ministry of Heavy Industries, Government of India for data collection related to accident statistics. (ii) From micro analysis - National Highway: Absence of underride guard bars/non-standard guard bars lead to serious rear-end cases, non-wearing of seat belts in bigger vehicles increases chances of fatal crashes. Thelimitation of the study involve using secondary data from MORTH which may contain inconsistencies.

2.2 Infrastructure and Emergency Response

The socio-economic development of a society creates an extra pressure on road infrastructure with the increasing number of vehicles which in turn crash on roadsNassir et al. (2024). Identifying road sections with frequent accidents is the first step for any successful road safety management process as resources are always limited. Emergency Vehicle Services (EVS) is part of the road crash management. Based on severity of accidents weightage estimated using PTV Visum Safety software, a total of sixty-two black

spots were generated in the study area. The locations of these black spots were then migrated to QGIS along with the locations of the 24-h-working hospitals and ambulance services in Ernakulam district. KANIV-108 ambulances were also mapped into QGIS, and a geo-spatial analysis was conducted to determine the fastest routes between ambulance service station locations to and from black spots and hospitals. The sixteen black spots that cannot be accessed within platinum time are recommended for new emergency vehicle services at suitably identified locations. This work will ease out effective emergency vehicles deployment as regards minimizing response time to road crashes in the area. The limitation includes that the analysis used past sales data, which might possess certain degrees of biases that is it has a regional context focusing on only Ernakulam district hence the findings may not be generalized to other setting and also no real life assessment of proposed interventions have been made.

The increased traffic load a place has in intersections is an increase in the rise of road accidents and confrontations and congestion among road users Nair and Raju (2024). The study concentrated on Puthuppally junction, one of the few main towns in Kerala's Kottayam district. Because of the closeness of the bus stops to this junction, it leads to added traffic from adjacent on-street parking, auto stands, and taxi stands. Thus, this junction needs a traffic signal installed to solve this issue. The paper presents how a traffic signal was designed for this particular junction by traffic volume studies in which morning and evening traffic movements were analyzed.Subsequently, a traffic signal design utilizing Webster's method was proposed. According to the analysis, the passenger automobile, pickup truck, and auto-rickshaw contribute the most (55%), while the pedal cycle contributes the least (about 0%) to the traffic flow in the junction. The busiest route is from Karukachal to Kottayam.The work has some drawbacks like being specific to a certain place, depending on some restricted data on traffic intensity.

The manoeuvres of merging and diverging also involve changes to the characteristics of traffic flow at diverging sections of roads when compared to mid-block sections. This study Kumar and Harikrishna (2023) was conducted to determine traffic characteristics at diverging sections of four-lane divided urban roads. The aim of this study is to investigate variations in speed, flow, and headways of vehicles at diverging sections. The study site is one of the four-lane divided urban roads in Kozhikode district, Kerala, India. Data was collected via video recording and TIRTL at diverging and mid-block sections for the observation stretch of 100 m. Speeds, headways, and traffic volume data were extracted and analysed. During peak and offseason, speeds during these periods reduced by about 20 to 30% or decreased by around 15 to 25% respectively. Various scenarios were analysed in order to know the changes in vehicle speed and headways at the diverging section.The research is conducted only at a selected segment of the four-lane urban road and, therefore, conclusions cannot be extended to other types of roads or geographical areas.

In addition, telecommuting and flexible work hours Radhakrishnan et al. (2024) can relieve the problem of traffic at peak hours. Flyovers would facilitate passage of vehicles from one point to another without having to use congested areas leading to free flow of traffic. Pollution can be indirectly attributed from traffic congestion, emphasizing the need for catalytic converters in exhaust fumes of vehicles. Therefore, a comprehensive solution aims to develop seamless and sustainable transport systems while meeting congestion and pollution challenges. This study is restricted to using site-specific data; several assumptions have been made during the simulations using VISSIM; comparison of results before and after the implementation of VM is not accompanied by a temporal variation analysis.

2.3 Technological and Policy Innovations

The article Chirakkal (2024) outlines the problems and answers for implementing smart mobility with respect to public transportation in India within the ambit of the Kerala State Road Transport Corporation (KSRTC). The determinants of smart adoption mobility include physical infrastructure, affordability constraints, and even the sociocultural context, which had been discussed in the research. It also adopts a quantitative research approach-Google Forms,-to examine KSRTC's performance while eliciting customer perceptions. The results reveal dependence on KSRTC, timeliness, and cost-based challenges. In specifying that successful smart mobility efforts have to include all-round development plans and top-level technology investment, and government partnerships, the article concludes the need to tackle some critical problem areas in making the transport systems more efficient and competitive.

Many governments Thomas et al. (2024) have mandated using a helmet for riding motorcycles and a set of traffic laws for their riders to lessen the rate of fatalities in accidents. The developed system is expected to assist law enforcing officials in the identification of motorbike riders violating traffic rules and in increasing their consciousness toward following the rules of the road. This system accepts a real-time traffic video input, and it identifies, using the Yolov5 model, objects from each frame. Based on a confidence score threshold, selection of one out of all frames showing the motorcycle and rider was carried out.While a craft model has been used to identify the location of the licence plate from the motorcycle ROI. The identification of the licence plate helps the authorities in identifying the offender. A user interface shows the results from classifier and text detection models to the user. The entire system has shown a 90% accuracy in different scenarios when evaluated.

This paper Vinoth et al. (2024) analyzes and predicts road traffic accidents (RTAs) in India using linear and non-linear regression analysis and Multilayer perceptron Neural Networks (MLPNNs), which can be used by policymakers. MLPNN is an advanced technology that has shown great promise in studying ancient data and forecasting forthcoming inclinations. A large number of highway accident forecasting models rely on 26 years of data, spanning 1994 to 2019, into accident counts on freeways across India. In this case, the most effective non-linear regression models and MLPNN model were designated considering model variables like years and total number of road accidents (in numbers). Performance of the model was contrasted against the performance of a model built using linear regression and non-linear regression techniques with the same goal. The results prove that MLPNN outperforms the traditional models regarding the forecasting ability under consideration and produces forecasts that come very close to predicting future highway traffic.The study seeks to identify the model that provides superior accuracy and generalization capabilities for road traffic accident prediction, thereby contributing to the development of more effective safety measures and traffic management strategies.

The ever-increasing length of road networks Bisht and Tiwari (2023), particularly expressway networks in India, happens to be associated with an increasing trend of RTCs. Hence, it is significant to study the crash pattern and identify hotspots on intercity expressways in India. The primary objective of this study is to identify the fatal crash hotspots using geospatial methods on the selected intercity expressways. First, hotspot sections are identified using ordinary kriging (OK) and kernel density estimation (KDE), as well as network kernel density estimation (NKDE) methods. Finally, a comparative analysis is conducted concerning the methods used to know their predictive performance

in determining the hotspot. The study of the selected 165 km intercity expressway used the fatal crash data from August 2012 to October 2018. The output of the geospatial methods was that some of the common hotspots were identified by both methods. The comparative analysis indicated that the NKDE method is more effective in identifying the hotspots in smaller segments than the other two methods. These results, therefore, provide a usable and readily applicable hotspot identification methodology for agencies owning intercity expressways in LMICs.

Table 1 indicates research Summary for All the Research Papers Studied

Table 1: Research Summary for All the Research Papers Studied

Author(s)	**Title**	**Dataset**	**Model**	**Result**
Chand et al., 2024	Contributing Factors of Road Traffic Accidents: Exploration Through Data Visualization	Accident Data, FIRs	Visualization Models	Predictive visualization aiding accident prevention and data completion.
Beevi et al., 2023	A Comparative Study of Growth of Motor Vehicles and Motor Vehicle Accidents in Kerala	Motor Vehicle Dept., Govt. of Kerala	No Models	Highlights the correlation between vehicular growth and accidents.
Ajaykrishnan et al., 2024	Factors Influencing Post-Encroachment Time of Pedestrians Near Bus Stops	Road Accidents Statistics	Post-Encroachment Time (PET)	Proposes a model to compute PET values for pedestrian safety.
Vimalathithan et al., 2024	Study of Indian Road Traffic Accident Characteristics Using Clustering Analysis	Ministry of Road Transport and Highways (MORTH)	SOM Clustering	Insights for policymakers to improve road safety.
Nassir et al., 2024	Spatial Analysis of Road Crash Black Spots: A Case Study of Ernakulam	Ministry of Road Transport and Highways (MORTH)	QGIS	Efficient deployment of emergency vehicles to minimize crash response time.
Nair et al., 2024	Design of Traffic Signal at Puthuppally Junction Using Webster Method	Not Specified	Webster Method	Analyzes traffic flow; autorickshaws contribute most to congestion.
Kumar et al., 2023	Traffic Flow Characteristics at Diverging Sections of Urban Roads	Video Recording, TIRTL	VISSIM	Models site traffic with geometric details, speeds, etc.
Radhakrishnan et al., 2024	Unscrambling Traffic Congestion and Increasing Sustainability	Accident Data	Not Specified	Proposes a comprehensive approach for sustainable transport systems.
Chirakkal et al., 2024	Improvement of Smart Mobility in Public Transportation Systems: KSRTC Case Study	KSRTC Data	Not Specified	Identifies challenges and solutions for smart mobility adoption.
Thomas et al., 2024	Automated Detection of Traffic Rule Violation Using Deep Learning	Not Specified	Yolov5, VGG16	90% accuracy in detecting helmet usage and rule violations.
Vinoth et al., 2024	Road Traffic Accident Prediction in India Using Machine Learning	26 Years of Highway Data	MLPNNs	MLPNNs outperform traditional models for accident prediction.
Bisht et al., 2023	Identification of Road Traffic Crash Hotspots on Expressways	Fatal Crash Data (2012–2018)	Kriging, KDE, NKDE	NKDE effectively identifies hotspots in smaller segments.

3 Methodology

The study's main objectives are to examine accident data in order to spot trends ,assess the severity of accidents by year, and forecast future trends for focused interventions. This

required gathering, preparing, and evaluating historical data before using statistical and machine learning models to extract useful information.Figure 1 shows the methododology diagram.

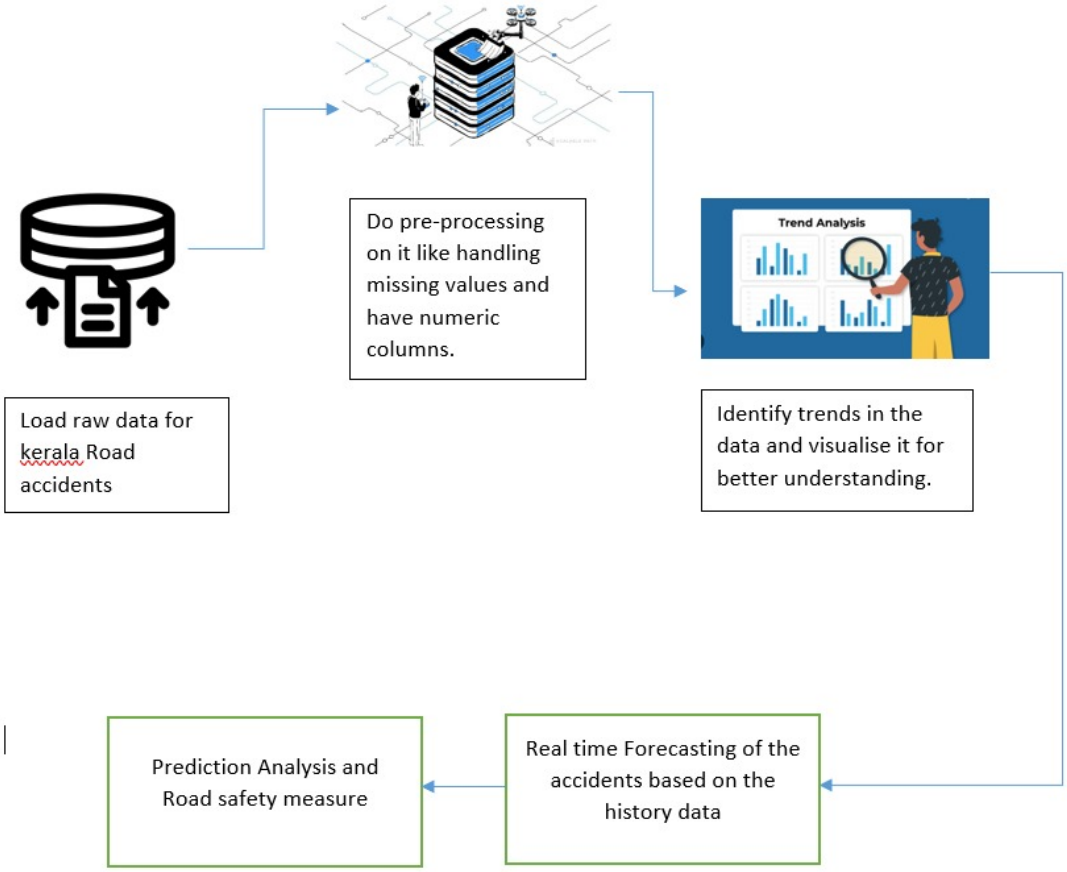

Figure 1: Methodology diagram

3.1 Dataset Description

The dataset utilized in this study includes comprehensive records of road accidents sourced from official government repositories and publicly available data sets. It captures key attributes such as the total number of accidents recorded annually, severity classifications (fatal, grievous injury, minor injury, non-injury), and the distribution of accidents across different vehicle types. In addition, it includes the number of people affected, including fatalities and injuries, allowing a detailed understanding of the impact of these accidents. The temporal coverage of the dataset spans multiple years, allowing for a robust analysis of trends and patterns over time, which is crucial for predictive modeling.

3.2 Data Preprocessing

The preprocessing phase was integral to preparing the raw accident data for machine learning tasks. The process began with data loading and inspection using the Pandas

library, enabling a structured examination of the dataset's contents. Missing values were handled using forward filling (ffill)Bisht and Tiwari (2023) to retain critical data points, and redundant rows or irrelevant columns were removed to maintain data relevance. Feature engineering played an important role in this phase. Additional metrics, such as year-over-year (YoY) growth in accident counts and accident-to-fatality ratios, were computed to derive actionable insights. Numeric columns were standardized, and the data was aggregated by year and vehicle type to facilitate trend analysis. The road accident dataset underwent systematic preprocessing to ensure data quality and readiness for analysis:

Loading and Cleaning: The dataset was loaded into a DataFrame, with redundant rows removed, columns renamed for clarity, and empty rows dropped.
Data Type Conversion: The Year column was converted to integers, and numeric columns like Total_Accidents and Persons_Killed were coerced into numeric types to handle inconsistencies.
Aggregation: Data was grouped by year, summing key metrics like Fatal_Accidents and Total_Persons to provide an annualized view of accident trends.Choudhary et al. (2024) This preprocessing ensured the dataset was clean, structured, and ready for analysis.

3.3 Feature Engineering

Temporal features, such as accident counts by severity, fatality ratios, and YoY growth percentages, were engineered to enhance the dataset's predictive capabilities.Nair and Raju (2024) These features encapsulated the behavioral and temporal patterns in accidents, making them critical for modeling. Here Dependent Variable is Total_Accidents, (total number of accidents that occurred in each year)and the Independent Variables are Fatal_Accidents, GI_Accidents and Persons_Killed fall under features influencing accidents.

3.4 Exploratory Data Analysis (EDA)

EDA validated the relevance of the engineered features through visualizations such as histograms and scatter plots. This step ensured that the structure of the data set was aligned with the modeling objectives. Feature distributions and inter-variable correlations were analyzed to identify patterns and potential outliers, enabling the refinement of data for subsequent analysis.

3.4.1 Annual Trends in Total Accidents with YoY Growth

The graph highlights total accidents and Year-over-Year (YoY) growth. There is a sharp increase from 2017 to 2018, stabilization in 2019, a decline in 2020, and a subsequent rise post-2020. The decline in 2020 reflects reduced mobility during the COVID-19 pandemic. However, the upward trend after 2020 indicates a need for renewed safety measures as mobility increases.The Figure 2 represents Annual Trends in Total Accidents with YoY Growth

3.4.2 Bar Chart

The bar chart identifies road types most associated with accidents. "Other Roads" have the highest number of accidents, followed by National Highways and State Highways.

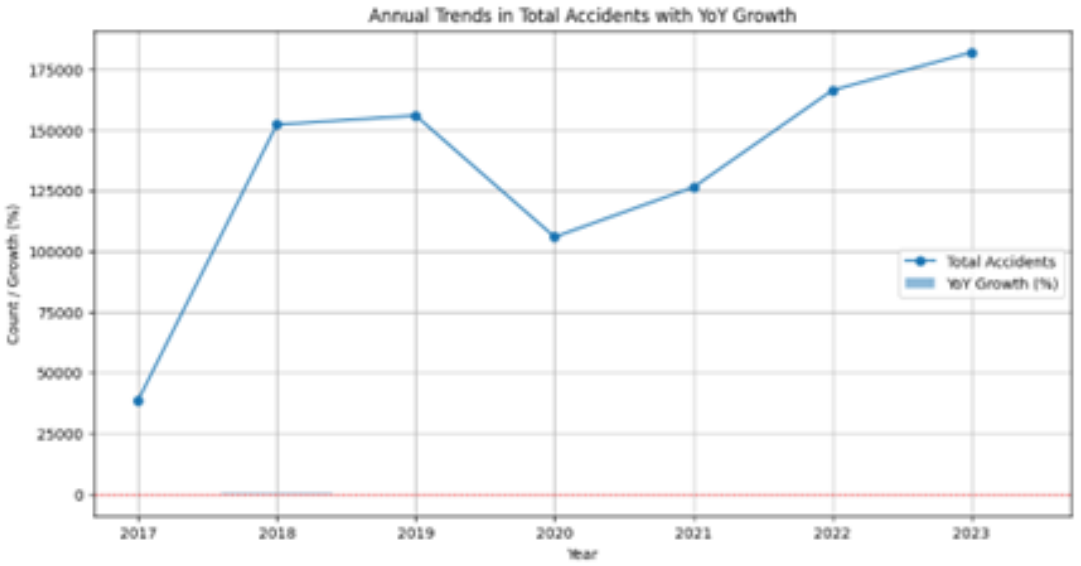

Figure 2: Annual Trends in Total Accidents with YoY Growth

Accidents on specific time slots like 9:00–12:00 hrs, 12:00–15:00 hrs, and 6:00–9:00 hrs are minimal in comparison. Targeted safety measures should be implemented on National and State Highways as well as "Other Roads." Highways need better infrastructure and enforcement of traffic regulations to reduce accident rates.The Figure 3 represents Most Accident-Prone Roads This chart highlights the most dangerous times for accidents.

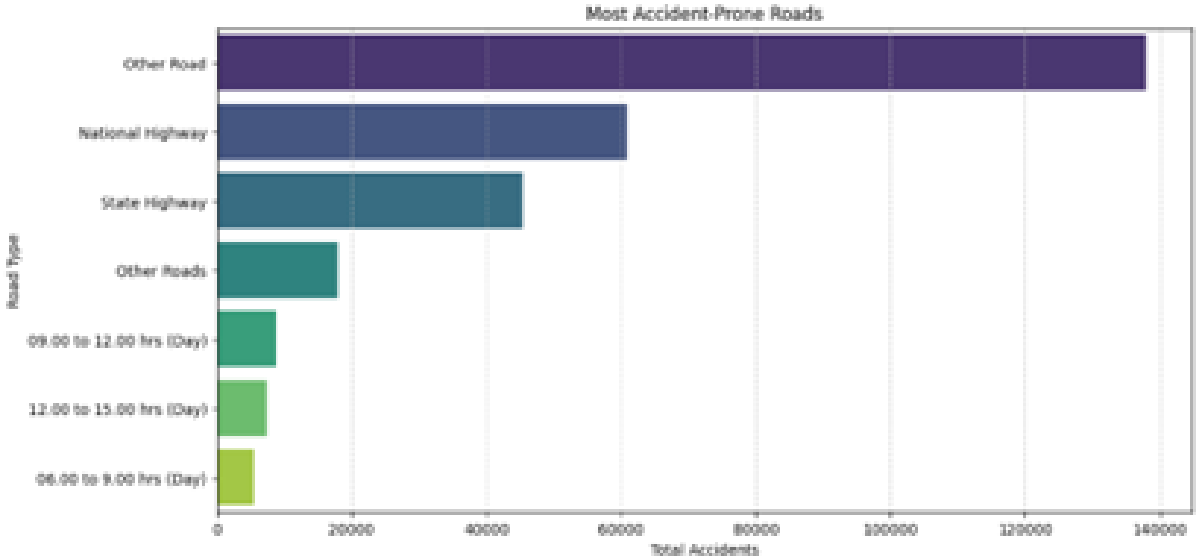

Figure 3: Most Accident-Prone Roads

Night hours (18:00–21:00 hrs) and afternoon hours (15:00–18:00 hrs) exhibit the highest accident counts, followed by late mornings (09:00–12:00 hrs). Interventions like improved lighting on roads and enhanced traffic monitoring during peak hours could mitigate accidents during these critical time frames.The Figure 4 represents Most Accident-Prone Times of Day The stacked bar chart shows the proportion of accident severity types from

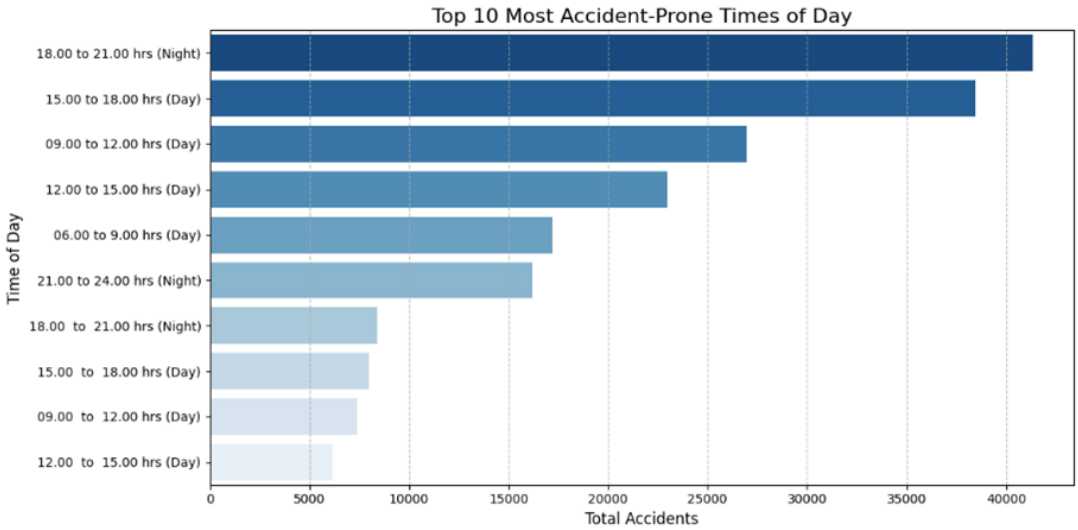

Figure 4: Most Accident-Prone Times of Day

2017 to 2023. Grievous injury accidents (GI_Accidents) dominate, followed by minor injuries. Fatal accidents and non-injury cases form a smaller proportion but are consistently present. The dominance of grievous injuries suggests the need for better trauma care systems. Stricter vehicle safety standards and law enforcement could address the steady proportion of fatal accidents.The Figure 5 represents Proportion of Accident Severity Types Over Years

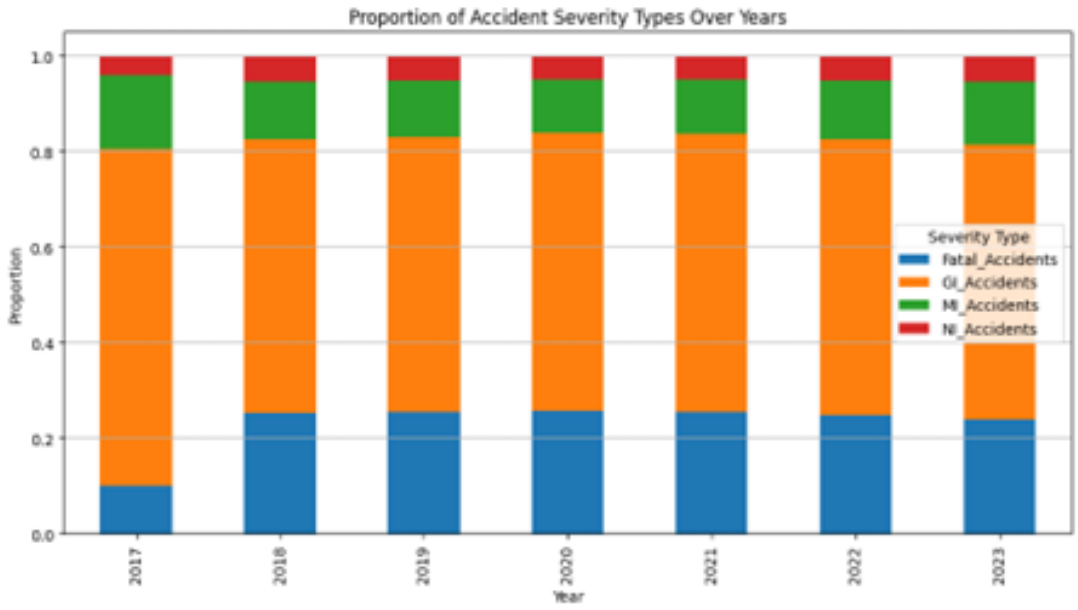

Figure 5: Proportion of Accident Severity Types Over Years

3.4.3 Stacked area chart

The stacked area chart represents accident severity trends over the years. Grievous injuries and minor injuries dominate, with fatal accidents showing a steady increase. Grievous injuries highlight the need for improved emergency care, while the rise in fatal accidents calls for stricter safety laws and better vehicle standards.The Figure 6 represents Trends in Accident Severity.

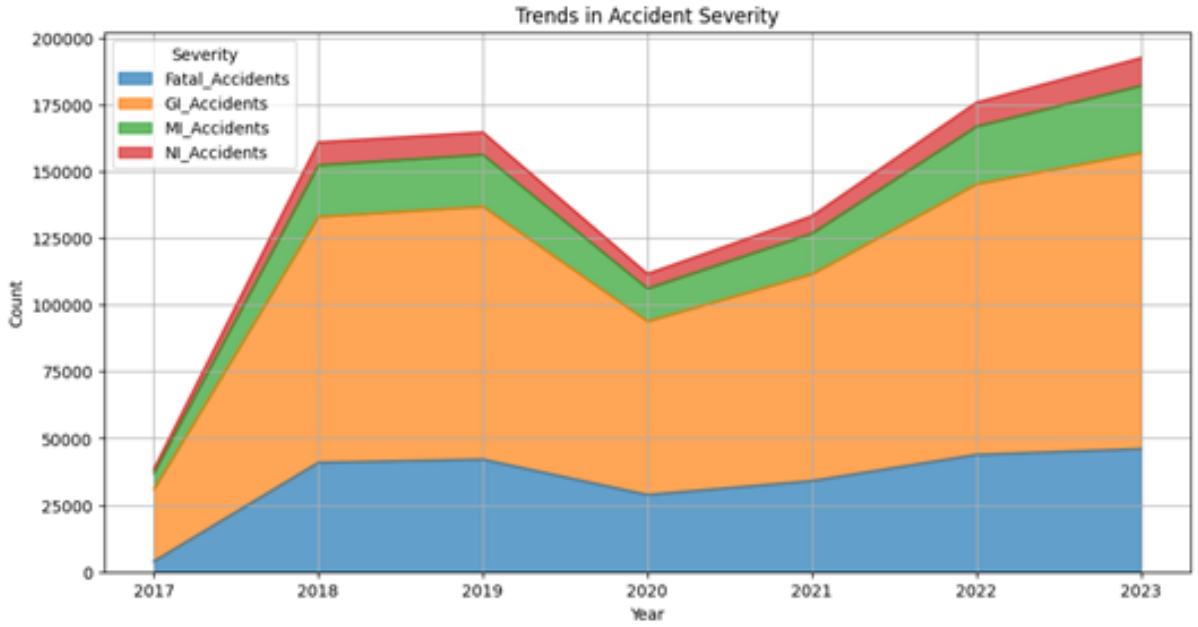

Figure 6: Trends in Accident Severity

3.4.4 Correlation Analysis

The heatmap illustrates the relationships between accident-related attributes, such as accident severity, total accidents, and persons affected. High correlations are observed among related variables, such as "Fatal_Accidents" with "Persons_Killed" and "GI_Accidents" with "Total_Accidents." Grievous injuries and fatal accidents significantly contribute to the total number of accidents. These attributes should be prioritized in safety interventions, and highly correlated features can aid in predictive modeling.The Figure 7 represents Correlation Heatmap of Accident Data.

3.4.5 Line graph

The line graph shows the number of persons affected over time, including "Persons_Killed," "GI_Persons" (Grievous Injuries), "MI_Persons" (Minor Injuries), and "Total_Persons." There is a consistent upward trend, particularly post-2020. Grievous injuries dominate the affected population, followed by minor injuries, while fatalities remain lower but steadily increase. Targeted safety measures and enhanced trauma response systems are necessary to curb these trends.The Figure 8 represents Trends in Persons Affected by Accidents.

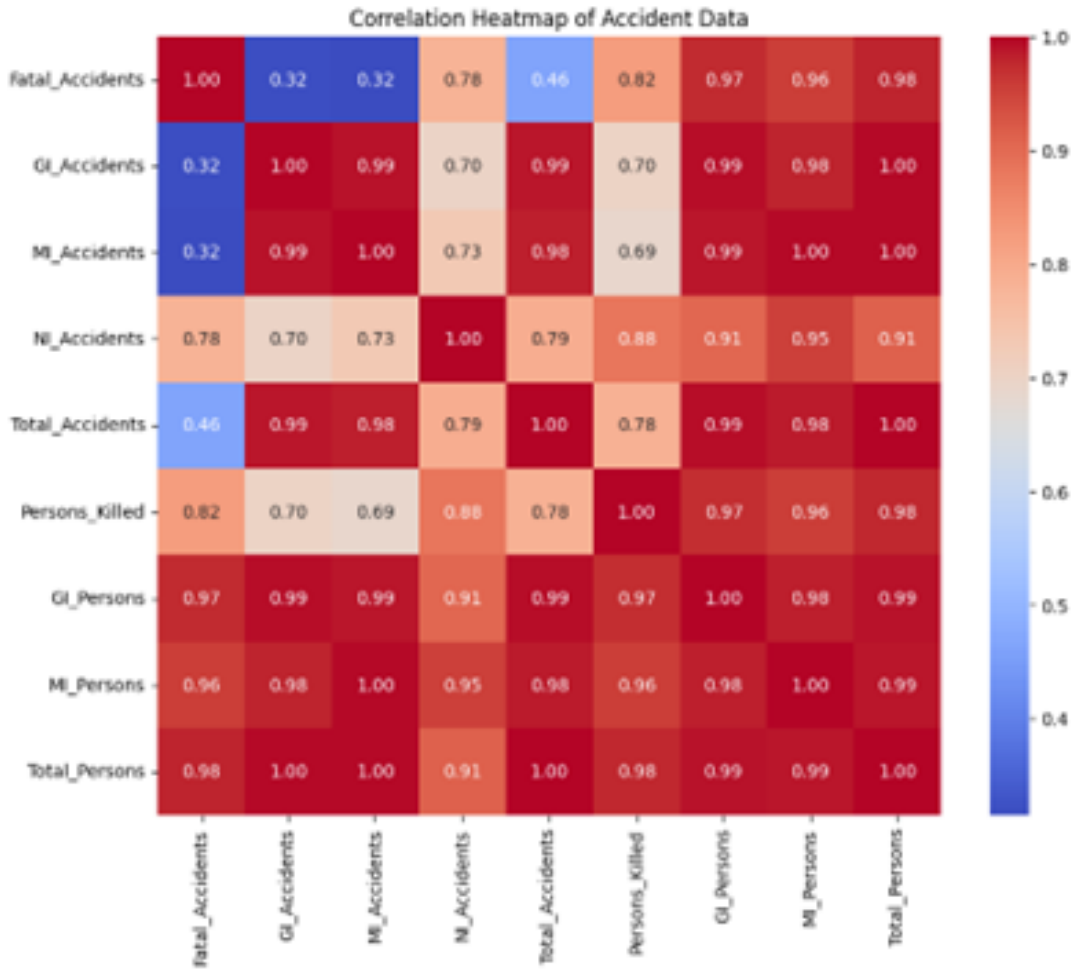

Figure 7: Correlation Heatmap of Accident Data

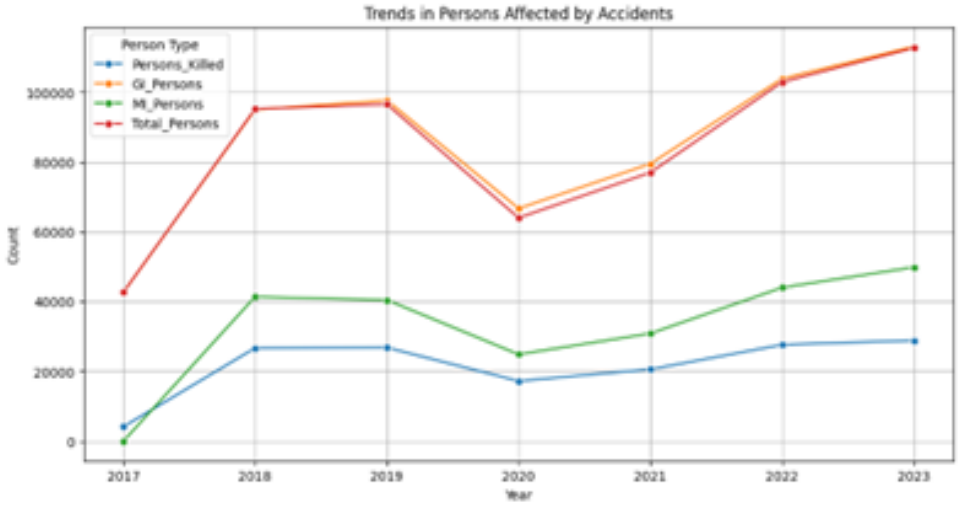

Figure 8: Trends in Persons Affected by Accidents

3.5 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique used to address high correlations among features by transforming them into uncorrelated principal components. It helps retain most of the data's variance while reducing redundancy and computational complexity. Principal Component Analysis (PCA)Thomas et al. (2024) used to mitigate dimensionality of the independent variables whereby the new components retain 95% of variance of the dataset, so as not to lose important information. When using PCA, the data is first normalized so that all features have an equal influence on the entire analysis process and none of them influences the results due to their large scales. The pca_df DataFrame hold the principal components that are extracted by conducting the PCA on the standardized numerical data. Numerical columns that are selected are accident related i.e., Fatal_Accidents, GI_Accidents, MI_Accidents, NI_Accidents, and Total_Accidents and people related i.e., Persons_Killed, GI_Persons, MI_Persons, Total_Persons. These features were first normalized using StandardScaler in order to bring everyone to the same scale which is important when doing PCA. The columns of pca_df as below are called principal components, which could be described as a directions in the space of data, instead of the original input variables; for example, PC1, PC2.The pca_df is the dataset that was used for modeling after the reduction of its dimensionality to improve efficiency..

3.6 Handling missing values and Outliers

To deal with the missing values in the time-series data for Total_Accidents another data preprocessing technique, forward fill or carry-forward is applied to fill missing entries with the last valid observation in order to sustain the data continuity. The few extreme observations in the transformed feature space and the target variable are eliminated with the IQR method where bounds at 1.5 * IQR below Q1 and above Q3 are set.Nassir et al. (2024) This step also avoids situations where extreme values are given undue consideration in the determination of the model performance while retaining the general variability of the dataset. Handling missing values and unit values are as important as handling train and test cases as they set the ground for having a good model for prediction in the subsequent stages of the data analysis process.

3.7 Hyperparameter Tuning

Hyperparameter tuning is done through the use of GridSearchCV as this is common way of optimizing hyperparameters in machine learning task for every machine learning model. For example, the experiments involve the tuning of randomly forest's hyper parameters such as the number of trees (n_estimators), the maximum depth of the trees (max_depth), minimum number of samples required to split an internal node (min_samples_split) or for creating a leaf node (min_samples_leaf). In the same manner, the learnable parameters for the Gradient Boosting are being tuned and these consist of the learning rate, the number of boosting stage (n_estimators), as well as the maximum depth of trees exists. GridSearchCV, it conducts cross validation on all features of the specified hyperparameter and chooses the configuration that yields the smallest error (for example, negative mean square).

3.8 Data Scaling and Splitting

The dataset was split into training and testing sets using an 80:20 ratio. Standardization was applied to numeric features using StandardScaler, ensuring consistency across all machine learning models.

3.9 Modeling

In this analysis we will use forecasting models for the accidents prediction and regression models for identifying the reasons for the accidents. Since we want to make a robust models for which we want to check in the real time about the impact of each of feature, we need those regression models which can help us with the coefficients values.The log transformed version of Total_accidents serves as the target variable.The feature variables are the PCA transformed Components(pca_df) which are obtained from the original people related and accident metrices

3.9.1 Linear Regression

Linear Regression is one of the simplest machine learning algorithms used for regression tasks. It assumes a linear relationship between the dependent variable (target) and independent variables (features).The algorithm minimizes the residual sum of squares (the difference between observed and predicted values) to find the best-fit line. While it's efficient and interpretable, it's limited in handling non-linear relationships, as seen in complex datasets like accidents data.Chirakkal (2024)

3.9.2 Decision Tree Regressor

A Decision Tree is a hierarchical model used for both classification and regression tasks. It splits the dataset into subsets based on feature thresholds, forming a tree-like structure. Each internal node represents a test on a feature, and branches represent decision outcomes, ending in leaf nodes that provide the final prediction. While effective for capturing non-linear patterns, it is prone to overfitting, which is mitigated in ensemble methods like Random Forest.Nair and Raju (2024)

3.9.3 Random Forest Regressor

The Random Forest Regressor (Vowels et al., 2022). is an ensemble learning model that builds multiple decision trees and combines their outputs by averaging the predictions. Each tree is trained on a random subset of the data and features, ensuring robustness and reducing overfitting. In the context of accident prediction, Random Forest can capture non-linear interactions between features and provide highly accurate predictions. It also ranks the importance of features, which helps identify the most influential factors.Kumar and Harikrishna (2023)

3.9.4 Gradient Boosting Regressor

Gradient Boosting is an ensemble technique that builds models iteratively, where each new model corrects the errors of the previous ones. It minimizes loss using a gradient descent algorithm, making it excellent for capturing complex, non-linear relationships. Gradient Boosting Regressor, as shown in your study, achieved the best performance,

balancing bias and variance effectively. Its success lies in its iterative approach and the ability to optimize weak learners to create a strong predictive model.Chand et al. (2024)

3.9.5 Support Vector Regressor

SVR is an extension of Support Vector Machines (SVM)Nair and Raju (2024)for regression. It uses a margin of tolerance (epsilon) to fit the data and employs kernel functions to model non-linear relationships. SVR excels in handling high-dimensional data, but its performance depends heavily on parameter tuning and feature scaling, which can lead to suboptimal results without careful optimization.

3.9.6 k-Nearest Neighbour (kNN)

K-Nearest Neighbors (KNN) is a simple and versatile machine learning algorithm used for classification and regression. It predicts outcomes by finding the 'k' closest data points in the dataset based on a distance metric and making decisions based on their majority class or average value. Commonly applied in tasks like image recognition, recommendation systems, and anomaly detection, KNN is easy to implement but can be computationally intensive for large datasets.(Jahin et al., 2024).Vimalathithan et al. (2024)

3.9.7 Seasonal AutoRegressive Integrated Moving Average (SARIMA)

SARIMA Krishna et al. (2023) is a time series forecasting method that extends ARIMA by incorporating seasonal components. It models data with trends, seasonality, and noise by applying differencing (to make the series stationary), autoregressive terms, and moving averages. The seasonal components enable SARIMA to capture periodic fluctuations, making it effective for predicting long-term accident trends and seasonal patterns.Before implementing SARIMA we have done ADF test to ensure the stationarity of the data.

3.10 Evaluation Criteria

The evaluation of the study employed a multi-faceted approach:

Regression Models: Assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) Choudhary et al. (2024)scores to evaluate predictive accuracy. Models with extremely low MAPE or RMSE scores, such as 1.0, were scrutinized for overfitting tendencies.

RMSE: Root Mean Squared Error (RMSE) evaluates prediction accuracy, with lower values indicating better model performance.

MAPE: Mean Absolute Percentage Error (MAPE) measures error as a percentage, providing an intuitive view of prediction accuracy.

Time Series Models: Validated through out-of-sample forecasting to ensure generalizability. Here the minimum RMSE or MAPE is selected.

All the models goes through the greedy selection measure in which the minimum metric in terms of loss or error is selected

3.11 Large Language Models

These models are used for the real time response generation for the road safety for the users in the application. When the users asks for the road safety related queries the LLMs

(Large Language Models) will generate the results based on the insights and data present in the database.The accident prevention Assistant starts by loading a knowledge base from a text file which contains information regarding the accident prevention measures.Here we have used a pretrained BERT model and it is consisted of the bert-large-uncased-whole-word-masking-finetuned-squad. The knowledge base is divided into smaller portions that should be easier to manage and perform. The answer_question method is the most crucial one, which also compares each chunk of the knowledge base and predicts the answers. The assistant also accepts upgraded query processing through the query_topic method which enables the user ask topic-specific questions to get particular results. And it elaborates the result of every query by giving the topic of the query, estimated confidence value, and a comprehensive recommendation. The assistant also works interactively and in batch mode: users enter questions directly and can pose a number of questions at once.

4 Design Specification

The design process fits with the ideas outlined to solve the major gaps observed in the management of RTAs employing ML, LLMs, and time-series analysis. It starts with the outlined areas of concern, which include, among others, the absence of the ability to forecast, intervene in real-time, and spatially. . As mentioned in the data pre-processing phase, raw data is cleaned and transformed for dealing with the problems like missing value and standardization LLMs extracting insights from crash report and geospatial data made normalized for the compatibility of ML.Bisht and Tiwari (2023) The algorithms in this step focus on what technique should be applied such as the ML model for making accident prediction, the time series for the prediction of future occurrences of the accident, and geospatial approaches in mapping the areas prone to accidents. Discrimination of these algorithms includes identifying patterns from previous events and current data about the drivers and traffic flows. Cross-validation results for accuracy, MSE and MAPE are utilized to assess the duration and effectiveness of the predictions while further feedback loops can be added to fill in the gaps, using changes like feature engineering or in hyperparameters. This systematic procedure allows immediate interactions, forecast, and rational usage of data with public participation through safety measures and promotions. In applying the workflow's systematized planning guide, using new technologies addresses research deficiencies, changes the paradigm of RTA to preventative, and promotes the sustainable development of a scientific evidence-based road safety environment.

5 Implementation

In Figure 9 the implementation diagram is shown. Analysis of road accident data involves selecting the specific problem to solve, within the broad context of road accidents, specifically, analysis of trends, future accidents and gaining insights into practical safety measures. Information is then gathered from different sources including governmental bodies' records of accidents and public data-sets to form data for analysis and modelling. Pre-processing is performed on the collected data to clean raw data, remove any irrelevant data and values, and transform data to form a format that is readily usable by machine learning models such as normalization or standardization techniques. Depending on the problem that needs to be solved it is chosen what kind of Machine Learning method is to be used, whether it is a Regression, Classification or Forecasting problem we select the al-

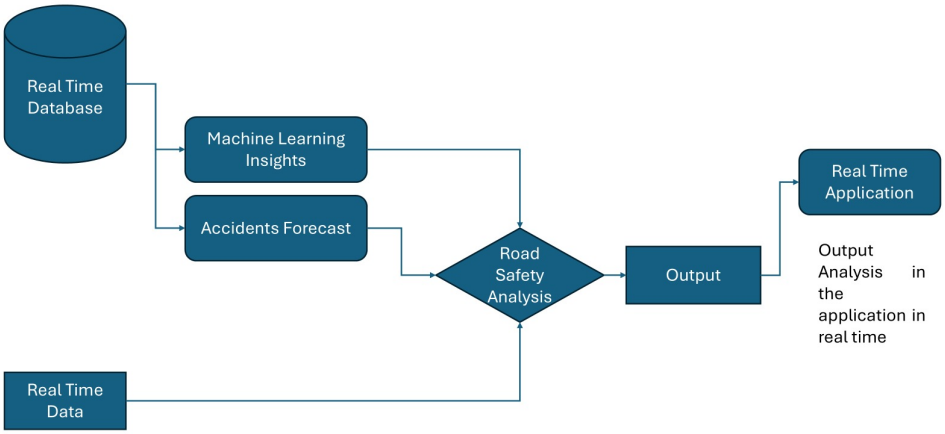

Figure 9: Implementation

gorithm to use from Linear Regression, Random Forest, Gradient Boosting or SARIMA. These are then used to train the processed data to identify patterns and associations and can involve feeding the training parameters of the learned data into the algorithm and weights. Mean Squared Error (MSE) Mean Absolute Percentage Error (MAPE) / accuracy is used to assess the trained models' efficiency on the task.Krishna et al. (2023) If the evalutation results are not good new corrections are fed back and new strategies is done and the model is retrained until the performance is optimized. Once the model gives satisfactory results, it gives out perceptions or forecasts okay usable for developments like estimating accident patterns or advising on safety measures.For accident prevention assistance we have implemented a pre-trained BERT model.The assistant also allows for advanced query processing through the query_topic method, through which the user can pose more refined topic-specific questions in order to elicit specific responses. However, it enriches the result of every query by providing topic of the query, estimated confidence value, and detailed recommendation.

6 Results and Evaluation

The performance of the machine learning models was evaluated using three primary metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics provided a comprehensive understanding of the models' predictive accuracy and error. The results of the evaluation are detailed below.

6.1 Machine Learning Model Performance

The Random Forest model demonstrated the best performance among the models tested. It achieved the lowest MSE of 2116.67, indicating a high level of predictive accuracy.In Figure 10 the model performance for MAPE is shown. Its RMSE was 46.01, further confirming its minimal prediction error. Additionally,the Random Forest model achieved a relatively low MAPE of 0.1541, highlighting its ability to make accurate percentage-

based predictions. The Decision Tree model, while effective, produced higher error values compared to Random Forest. It recorded an MSE of 2890.45 and an RMSE of 53.76, with a corresponding MAPE of 0.1836. These results suggest that the Decision Tree model may have been prone to overfitting or limited generalizability in its predictions, leading to higher error metrics. The Gradient Boosting model showed reasonable performance with an MSE of 2454.85 and an RMSE of 49.55. However, its MAPE was extremely high at 1.77×10^{13}, an anomaly that likely stemmed from outliers or issues in data preprocessing. This inflated MAPE value does not align with the relatively moderate MSE and RMSE values, warranting further investigation to identify potential data-related challenges or computational errors. The K-Nearest Neighbors (KNN) model delivered moderate performance, with an MSE of 2590.16 and an RMSE of 50.89. Its MAPE was 0.1625, which, although slightly higher than Random Forest, indicated better performance than Decision Tree and Gradient Boosting for percentage-based accuracy. These results position KNN as a reliable model with balanced predictive capabilities.In Figure 11shows Model performance for MSE and Figure 12shows Model performance for RMSE.We have used Random Forest model with least RMSE value for predicting the future accidents.

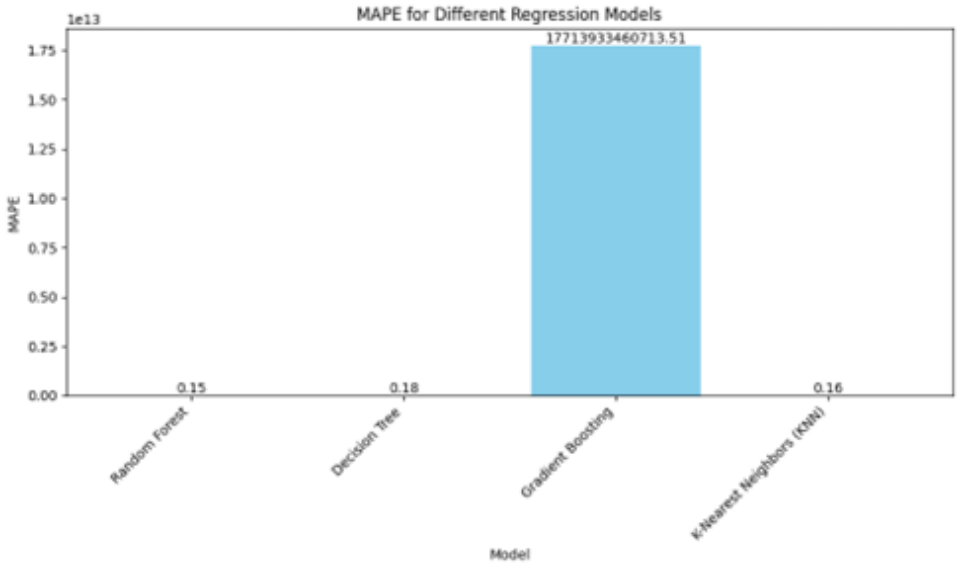

Figure 10: Model performance for MAPE

Table 2 shows the Model Performance Metrics

Table 2: Model Performance Metrics

Model	**MSE**	**RMSE**	**MAPE**
Random Forest	2116.67	46.01	0.1541
Decision Tree	2880.45	53.76	0.1836
Gradient Boosting	2454.86	49.55	17713933460713.51
K-nearest neighbors	2590.16	50.89	0.1625

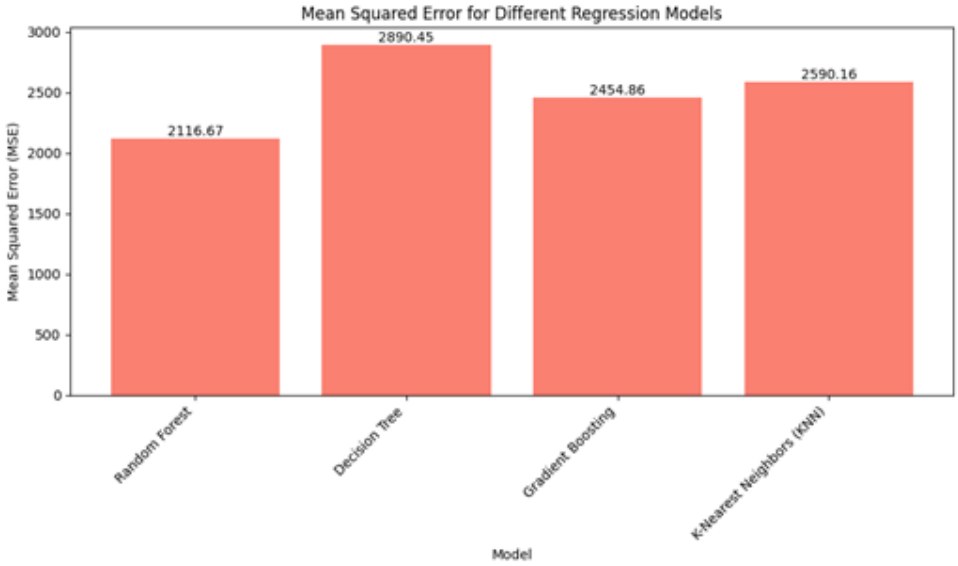

Figure 11: Model performance for MSE

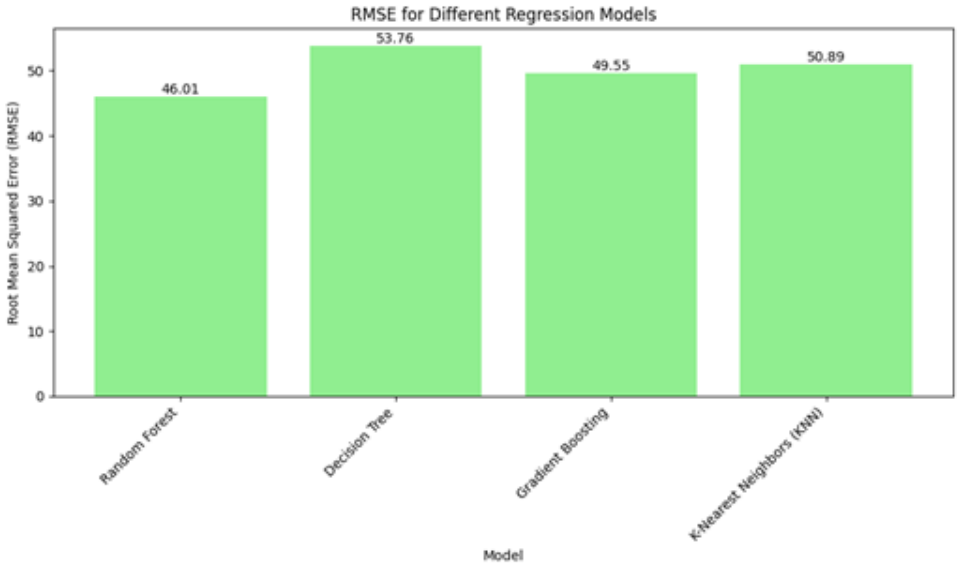

Figure 12: Model performance for RMSE

6.2 Time Series Forecasting Results

The "Improved SARIMA Forecast" visualization illustrates the predicted upward trend in total accidents over the forecasted period. The model's projections emphasize the critical need for preemptive road safety measures to mitigate potential risks. The SARIMA model's integration of seasonal components and trend analysis provides stakeholders with a reliable framework for long-term planning and resource allocation in accident prevention initiatives. Figure 13 indicates SARIMA model forecast. The graphs highlight an upward trend in accidents, with severe cases contributing significantly, indicating the need for targeted safety measures. Specific vehicle types and roads are at higher risk, requiring stricter enforcement and public awareness. The forecast emphasizes the urgency for proactive actions, such as better infrastructure, adaptive speed controls, and enhanced safety campaigns, to prevent further increases in accidents.

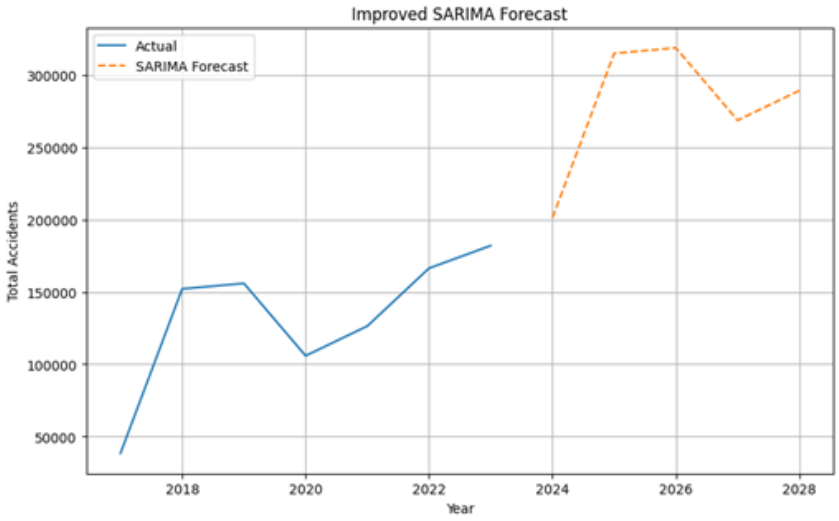

Figure 13: SARIMA model

6.3 LLM-Based Assistant for Accident Prevention

A LLM-driven assistant was used to provide on-demand guidance on the measures to observe to avoid an accident. Operating on top of the created knowledge base of road safety actions, the assistant was outstanding in terms of context awareness and relevance. The model contributed to answering user queries on various aspects including road infrastructural development, safety driving procedures, and traffic regulation.The Accident Prevention Assistant is supposed to accept a knowledge base that is read from a text file with extensive information on protection against accidents. For its question answering it uses a fine-tuned BERT model: bert-large-uncased-whole-word-masking-finetuned-squad. The knowledge base is divided into the smaller units for easier and more effective work with it during the processing.

The assistant's main feature is in the answer_question method, the responsibility of which is to analyze each chunk in the knowledge base itself and deliver the most relevant answers. Furthermore, the query_topic method improves the way of query interaction because users addressed topic-related questions and receive an appropriate answer. From

this feature, you get the topic of the query, an estimate of the confidence level of the system, and a detailed explanation or recommendation.

Table 3: Safety Recommendations

Query Topic	**Confidence Score**	**Recommendation Example**
Road Safety	0.34	Ensure proper road markings and lighting
Vehicle Safety	0.27	Conduct periodic vehicle inspections and maintenance
Driver Safety	0.47	Enforce strict zero-tolerance policies for impairments

The model was producing meaningful recommendations with high confidence values at all the time as seen in the LLM.Table 3 shows Safety Recommendations.For instance, it underlined implementing variable speed limit control measures in the black spots and also stressed the need to use periodic vehicle checks to improve safety. These outputs demonstrate that the model can be used for informing and raising awareness about road safety for the public as well as for providing recommendations for policymakers.The recommendations emphasize three critical areas: road safety, vehicle safety, and driver safety. The analysis shows driver safety demands immediate attention due to its 0.47 confidence rating indicating that human factors and zero-tolerance interfaces for impairment play a crucial role. Proper road markings and adequate lighting systems hold a critical position in road safety according to analysis with a confidence score of 0.34. Vehicle safety maintains importance despite its low confidence rating of 0.27 because inspections and maintenance work together to stop accidents from happening. Research findings demonstrate that improved driver conduct combined with better roadway conditions have the potential to deliver the highest safety gains. The usage of the LLM-based assistant shows that there is a scalable and practical approach possible to spreading essential safety information to the public and helping the stakeholders to apply proper accident prevention measures.

6.4 Discussion

The assessment of performance characteristics of the models indicated that the Random Forest model proved to be the most effective. It had the smallest mean square error (MSE) and the least prediction error (Root Mean Square Error, RMSE) and is thus useful when there are the most intricate relationships between data. Higher error values were observed in the Decision Tree model meaning overfitting of the data or low generalization. The Gradient Boosting model performed fairly satisfactorily but the MAPE was warranted; this could be due to data issues or calculation errors. The study also revealed that the reliability of K-nearest neighbors (KNN) model is good but not great when faces the highly complex KDD.Radhakrishnan et al. (2024)

This was evident by the total accidents, which were forecasted by the SARIMA model to increase in future, suggesting that road safety interventions ought to be proactively executed. Integrated safety initiatives, including the upgrade of road infrastructure, the installation of adaptive speed control systems, as well as the commencement of advanced safety campaigns, are highlighted through the light of the model as targeted safety measures. The pupils found useful instructions and high confidence values from an LLM-based assistant who offered on-demand recommendations to the classes regarding the prevention measures of accidents.

Nevertheless, certain limitations include Gradient boosting peculiarities, overfitting in Decision Tree, high computational complexity in KNN, and the necessity to ensure more thorough independent confirmation of the positive impact of the LLM assistant.

7 Conclusion and Future Work

Preventing RTAs is a major population health issue in Kerala, a state that has relatively complex and sensitive roads conditions such as high traffic volume, a blend of both urban and rural roads and predominant use of roads for transport. This review also shows that, despite the trend that has proposed groups of main causes of accidents such as driver distractions, infrastructural lapses, and unsuitable responses from the emergency center, there are still large knowledge gaps about how to employ the superior predictive models and how to develop various interlocking structures. Involvement of pedestrian-high and the recurrent incidence of crashes on straight road-sections in Kerala calls for specific measures. Deploying emerging technologies such as machine learning, large language models as well as time series forecasting provides a vast opportunity for definite accident hotspots location, continuous tracking and efficient working resource management. There is a need for future research to model the accident risks using factors peculiar to Kerala like; weather changes, road geometry, and traffic intensity patterns. Since FIRs contain different types of information about specific accidents that have already occurred in Kerala, LLMs can extract useful understanding to infer changes for the policies. Various applications including traffic control using Artificial intelligence, safety of pedestrians around bus stops, and advanced keeping an eye on drivers all in real-time can decrease accidents. To do this, geo-spatial improvement to the state's emergency response systems will guarantee a quicker response to areas that experience numerous emergent incidents.

However, Kerala needs to develop road safety initiatives, supported by statistical data and improved public erformance that will promote safe driving and adherence to the law. Engagement of state policymakers, urban planners and technology providers would greatly be useful in the execution of these strategies. Using such a wide-spectrum approach that utilizes technology addressing transportation issues in Kerala, the state stands a good chance of cutting down RTAs greatly leading to a safer transport environment. This vision dovetails with Kerala's strategic planning objectives of socio-economic growth and the public good.

References

Ajaykrishnan, M., Sethulakshmi, G. and Mohan, M. (2024). Factors influencing post-encroachment time of road crossing pedestrians near, *Technologies for Sustainable Transportation Infrastructures: Select Proceedings of SIIOC 2023* **529**: 197.

Beevi, S. and Arya, S. (2023). A comparative study of growth of motor vehicles and motor vehicle accidents in kerala with emphasis on the need for sustainable transportation, *NCSDSGE–2023* p. 259.

Bisht, L. and Tiwari, G. (2023). Identification of road traffic crashes hotspots on an intercity expressway in india using geospatial techniques, *IATSS Research* **47**(3): 349–356.

Chand, A., Jayesh, S. and Bhasi, A. (2024). Contributing factors of road traffic accidents: Exploration through data visualization, *Transactions of the Indian National Academy of Engineering* pp. 1–13.

Chirakkal, M. (2024). Improvement of smart mobility in public transportation system: A case study of india's kerala state road transport corporation (ksrtc), *Young Scientist, Conference/Jaunasis mokslininkas, konferencija* pp. 200–204.

Choudhary, A., Garg, R., Jain, S. and Khan, A. (2024). Impact of traffic and road infrastructural design variables on road user safety–a systematic literature review, *International Journal of Crashworthiness* **29**(4): 583–596.

Krishna, C., Singh, A. and Jha, K. (2023). Safety improvement on indian highways, *J Saf Eng* **12**(1): 1–12.

Kumar, C. and Harikrishna, M. (2023). Traffic flow characteristics at diverging section of four-lane divided urban roads, *International Conference on Transportation System Engineering and Management* pp. 189–208.

Nair, H. and Raju, G. (2024). Design of traffic signal at puthuppally junction using webster method, *AIP Conference Proceedings* **3059**(1).

Nassir, N., Sam, E., Thomas, J. and Mulleti, R. (2024). Spatial analysis of road crash black spots: A case study of ernakulam, *Recent Advances in Transportation Systems Engineering and Management–Volume 1: Selected Proceedings of CTSEM 2023* **1**: 307.

Radhakrishnan, R., Ajimon, M., Bose, S., Surya, S., Pillai, V. and Sandeep, U. (2024). Unscrambling traffic congestion and increasing sustainability in special urban intersection, *E3S Web of Conferences* **529**: 04011.

Thomas, R., Ajesh, F., John, A. and Sonia, K. (2024). Automated detection of traffic rule violation using deep learning techniques, *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* pp. 1–6.

Vimalathithan, K., KM, P., Vallabhaneni, P., Selvarathinam, V., Manoharan, J., Pal, C., Padhy, S. and Joshi, M. (2024). Study of indian road traffic accident characteristics using clustering analysis, *SAE Technical Paper* (2024-01-2754).

Vinoth, B., Prakash, V. and Shivakumar, B. (2024). Road traffic accident prediction in india using machine learning algorithm techniques, *2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)* pp. 1–5.