A PREDICTIVE MODEL OF ROAD TRAFFIC ACCIDENTS USING SURROGATE SAFETY MEASURES

Prithviraj Gadgi[1], Dr. Ranju Mohan[2]

[1]Department of Data and Computational Science, Indian Institute of Technology, Jodhpur – 342030
[2]Department of Civil and Infrastructure Engineering, Indian Institute of Technology, Jodhpur – 342030

Abstract – This paper presents a comprehensive approach to develop a predictive model for road traffic accidents using Surrogate Safety Measures (SSM). The primary objective is to estimate the likelihood of accidents at both micro and macro levels, encompassing intersections and mid-blocks, as well as identifying accident-prone areas within city networks. SSM, derived from various traffic-related variables including road geometry, traffic flow characteristics, and driver behaviour, serve as indirect indicators for potential crash occurrences. A predictive model is developed using machine learning algorithms, adapting to the complexity of the data and desired accuracy. The performance of the model is evaluated through metrics like accuracy, precision, recall, and F1 score, which gauges its reliability in estimating accident risks. Furthermore, the model is deployed into practical applications, potentially integrated into traffic management systems, providing real-time accident risk predictions to drivers and relevant stakeholders. This research offers a holistic framework for predicting road traffic accidents, contributing to improved road safety and accident prevention.

Index Terms – Surrogate Safety Measures, Random Forest, Decision Tree, Logistic Regression, Hyperparameter Tuning, Road Safety.

I. INTRODUCTION

According to the death statistics released by the World Health Organization, the number of traffic accidents occurring annually in the world is alarming. The traffic accidents killed 1.2 million people each year and 50 million people were injured. Approximate 3,300 people were killed and 137,000 people were injured every day. Direct economic losses of 43 billion dollar, the frequent occurrence of traffic accidents directly threaten human life and property safety.

Road accident prediction is one of the most important research areas in traffic safety. The occurrence of road traffic accidents is mainly affected by geometric characteristics of road, traffic flow, characteristics of drivers and environment of road.

Many studies have been conducted to predict accident frequencies and analyse the characteristics of traffic accidents, including studies on hazardous location/hot spot identification, accident injury-severities analysis, and accident duration analysis. Some studies focus on mechanism of accidents. Other factors include weather and light conditions of the road.

Addressing road traffic accidents on a global scale is of paramount importance due to the substantial loss of life, injuries, and economic damages they cause. In pursuit of enhanced road safety, various predictive models have been developed to reduce the occurrence of accidents. This research project introduces an innovative approach to road safety, focusing around the creation of a predictive model that utilizes Surrogate Safety Measures (SSM) to gauge the likelihood of accidents.

Unlike conventional accident prediction models, along with data, this project incorporates SSM, encompassing traffic flow characteristics, road geometry, and driver behaviour. The aim is to proactively identify accident-prone areas, both at micro-levels such as intersections and mid-blocks, and at a macro-level within city networks. This approach has the potential to facilitate timely interventions and pre-emptive measures, thereby reducing the frequency of accidents.

The significance of Surrogate Safety Measures (SSM) in traffic safety evaluation arises from the lack of reliable statistical safety models in many scenarios. This is particularly relevant for transportation facilities with complex site characteristics and/or nontraditional traffic safety treatments, where historical crash data may be limited or unavailable for developing safety predictive models [6]. Research on SSM dates back to the early 1970s [3], and significant progress has been achieved in this field since then.

Lee et al [8] developed a probabilistic model relating significant crash precursors to changes in crash potential. Abdel [9] built a previous crash prediction model with the matched case-control logistic regression technique. No specific approach available for the traffic police to predict which area is accident prone at a specific time. The traffic accident prediction plays an important role in the integrated planning and management of traffic, the reason which with much randomness about the traffic accident include some nonlinear elements, such as people, car, road, climate and so on. The traditional way of linear analyses cannot reveal the really situation since the noise pollution and amount of data are too little, cause the result of prediction cannot satisfactory.

II. OBJECTIVE

The ultimate objective of this research is to contribute to road safety by providing a predictive model that offers real-time risk estimations for accidents. Such a transportation model can seamlessly integrate into traffic management systems, providing valuable insights to drivers, traffic authorities, and other stakeholders. This integration aims to facilitate safer road usage and prevent accidents. This research aspires to be a pioneering step towards establishing a safer and more efficient road system.

III. BACKGROUND

Road accidents, on a scale are a concern when it comes to public health and safety. According to the World Health Organization (WHO) [4] these accidents lead to millions of deaths and injuries every year. Not do they cause harm but they also have significant social and economic consequences for societies around the world. Fortunately, many road accidents can be. Identifying risks early on is crucial in minimizing their impact.

To understand the likelihood of an accident occurring researchers have turned to Surrogate Safety Measures (SSM). These measures act as indicators of crashes since collecting and analysing actual crash data can be challenging and time consuming. SSM consider factors such as traffic flow characteristics, road layout, weather conditions, driver behaviour, and other variables that influence the likelihood of road accidents.

Developing models for road accident prediction holds promise. It provides insights for traffic authorities, transportation planners and stakeholders in implementing targeted safety measures allocating resources efficiently and ultimately reducing the occurrence of road accidents. By embracing data analysis techniques predicting road accidents using safety measures has gained attention as an approach, to enhancing road safety and saving lives.

IV. PROBLEM STATEMENT

Addressing road accidents is a significant concern for public health and safety, and predicting accident risks early on is essential for minimizing their impact. However, the conventional reliance on actual crash data, such as First Inspection Reports (FIR), for accident prediction is not only time-consuming but also inefficient. There is a growing necessity to create predictive models capable of estimating the likelihood of road accidents using surrogate safety measures, indirect indicators that signal potential crash occurrences.

Within the realm of mobility, road safety emerges as a critical aspect with the primary goal of preventing accidents and mitigating their consequences. The problem statement encompasses key elements related to road safety, including accident prediction, accident prevention, and data analysis for safety improvement. In essence, the problem statement aligns with the broader objective of enhancing road safety by developing predictive models that leverage surrogate safety measures. This proactive approach aims to prevent road accidents, optimize resource allocation, and facilitate evidence-based decision-making to improve road safety.

V. EXISTING SYSTEM

No specific approach available for the traffic police to predict which area is accident prone at a specific time. The traditional Back propagation network has defects. It has a 17% lower accuracy than the proposed model. We propose the use of a machine learning technique. Machine learning has the ability to model complex non-linear phenomenon.

VI. PROPOSED SYSTEM

An ML powered web app which predicts accidents severity based on the current conditions. It is trained with huge volume of historical data and SSM. The purpose of such a model is to be able to predict which conditions will be more prone to accidents, and therefore take preventive measures. We will even try to locate more precisely future accidents in order to provide faster care and precaution service. According to the predicted severity, a message will be sent to the traffic police to take preventive measures.

VII. LITERATURE SURVEY

A literature survey in a software development process is a most significant part as it shows the various analyses and research made in the field of your interest including substantive findings, as well as theoretical and methodological contributions to a particular topic. It is the most important part of the report as it gives you a direction in the area of your research; it helps in setting up the goals for the analysis. The purpose is to convey to the reader what knowledge and ideas have been established on a topic, and what their strengths and weaknesses are.

A. *A Review on Surrogate Safety Measures in Safety Evaluation and Analysis*

- Author: Dungar Singh and Pritikana Das
- Year: 2022
- Objective: The aim of this article is to conduct a comprehensive study of existing research study on surrogate safety assessment using a systematic literature review by the research questions like:
  What are the different categories of surrogate safety measures used in road safety?
  What are the different categories of surrogate safety measures used in road safety?
  How surrogate safety measures (SSMs) are used to measure traffic conflict?
  How can SSMs be used to measure traffic conflict using a microsimulation approach?

B. *Traffic Conflict Prediction Using Connected Vehicle Data*

- Author: Zubayer Islam and Mohamed Abdel-Aty
- Year: 2023
- Objective: Use of individual vehicle trajectories and vehicle dynamics from connected vehicle probe data for conflict prediction. Network level vehicle conflict prediction using infrastructure-free probe vehicle data only. A LSTM model was developed to capture temporal data patters that can lead to potential conflict scenario. The LSTM model output was also carefully explained using SHAP.

C. *A Review of Surrogate Safety Measures and their Applications in Connected and Automated Vehicles Safety Modelling*

- Author: Chen Wang, Yuanchang Xie, Helai Huang, and Pan Liu
- Year: 2021
- Objective: Understand the state-of-the-art of SSM and their applications in CAV safety studies. Discuss the current issues of SSM used in CAV safety studies. Identify promising future research directions of SSM and their applications in CAV safety.

D. *A Model of Traffic Accident Prediction Based on Convolutional Neural Network*

- Author: Lu Wenqi Luo Dongyu Yan Menghua
- Year: 2017
- Objective: To predict the traffic accident severity by using convolution neural Network.

E. *The Traffic Accident Prediction Based on Neural Network*

- Author: Fu Huilin, Zhou Yucai
- Year: 2017
- Objective: Traditional way of linear analyses cannot reveal the really situation the result of prediction is not satisfactory. Compares traditional BP network with its proposed solution.

F. *On the Selection of Decision Trees in Random Forests*

- Author: Simon Bernard, Laurent Heutte and Sebastien Adam
- Year: 2017
- Objective: This paper presents a study on the Random Forest (RF) family of ensemble methods.

G. *Hyperparameter Tuning of a Decision Tree Induction Algorithm*

- Author: Rafael G. Mantovan, Ricardo Cerri, Joaquin Vanschoren
- Year: 2016
- Objective: This paper investigates how sensitive decision trees are to a hyperparameter optimization process. Four different tuning techniques were explored.

VIII. PREDICTION FACTORS

Day of Week: • Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday.
Latitude and Longitude: • Exact location of the vehicle
Light Conditions: • Daylight, Darkness – lights lit, Darkness – lights unlit, Darkness – no lighting.
Weather Conditions: • Fine no high winds, Raining no high winds, Snowing no high winds, Fine + high winds, Raining + high winds, Snowing + high winds, Fog or mist.
Vehicle Type: • Pedal cycle, Motorcycle 50cc and under, Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, Motorcycle over 500cc, Taxi/Private hire car, Car, Minibus (8 - 16 passenger seats), Bus or coach (17 or more pass seats), Tram, Truck(Goods), Electric motorcycle.
Road Surface Conditions: • Dry, Wet or damp, Snow, Frost or Ice, Flood, Mud.

Age of Driver: - Age of the driver.
Engine Capacity in cc: - The capacity of the engine in cc.
Age of Vehicle: - Age of the vehicle.
Gender: - Gender of the driver.
Speed Limit: - Speed limit of the vehicle.
Number of police officers present at the spot: - Numerical value.
Output: - Accident Severity – 1: Possibility of Fatal Accident - Accident Severity – 2: Possibility of Serious Accident - Accident Severity – 3: Possibility of Slight Accident

Table 1: Prediction Factors

IX. SURROGATE SAFETY MEASURES (SSM)

The assessment of safety performance in designs, countermeasures, or systems often relies on crash frequency and severity, crucial indicators of their effectiveness.

However, crashes are infrequent events, making it impractical and, to some extent, ethically questionable to rely solely on historical crash data for evaluating the performance of new safety strategies, such as the introduction of a new traffic sign. In response to this challenge, Surrogate Safety Measures (SSM) derived from traffic conflicts have gained popularity as an alternative.

Traffic conflicts represent observable non-crash events where interactions among multiple road users create a risk of collision unless their courses of movement are altered [1].

A conflict is deemed etiologically connected to a crash when a failure (e.g., human operator failure, road failure, vehicle failure) leading to the conflict cannot be adequately corrected [2] [7].

Due to this causal relationship, measures employed to identify traffic conflicts and assess their severities can be considered as SSM. Notably, traffic conflicts are more frequent compared to actual crashes.

It is important to highlight that numerous safety-related measures have been developed over time, but not all qualify as SSM. According to [6], two criteria for SSM qualification are:

- It must be derived from traffic conflicts directly linked to crashes.
- The relationship between traffic conflicts and the potential crash frequency and/or severity can be quantified using practical methods.

From this standpoint, traffic exposure/flow measurements such as Annual Average Daily Traffic (AADT), speed variation, and average operating speed do not meet the criteria for SSM, despite their proven associations with crash risk and occasional adoption as crash "surrogates." Here we consider the safety measures that satisfy the two qualifying criteria mentioned above, and the parameters considered are: Latitude, Longitude, Day of Week, Weather Conditions, Light Conditions, Road Surface Conditions, Age of Driver, Vehicle Type, Age of Vehicle, Engine Capacity in CC, Gender, Speed Limit, Number of Police Officers Present at the Scene.

A. Crash Potential Index (CPI),

It incorporates the parameters listed to assess the likelihood of crashes at a given location or roadway segment. Here's how CPI could be formulated to include the listed parameters:

CPI = $w1$ × Road Factor + $w2$ × Driver Factor + $w3$ × Vehicle Factor + $w4$ × Environmental Factor

Where:

- $w1$, $w2$, $w3$, $w4$ are weights assigned to each factor based on their relative importance.
- Road Factor includes parameters such as road type, speed limit, road curvature, and road surface conditions.
- Driver Factor incorporates driver-related parameters such as age, gender, and compliance with traffic laws.
- Vehicle Factor considers vehicle-related parameters including vehicle type, age, engine capacity, and safety features.
- Environmental Factor accounts for environmental conditions such as weather, light conditions, and the presence of law enforcement officers.

Each factor can be quantified based on empirical data, expert judgment, or statistical analysis of historical crash data. The weights $w1$, $w2$, $w3$, $w4$ are determined to reflect the relative influence of each factor on crash potential.

For example:

- If a location has a higher speed limit (Road Factor), it might receive a higher weight in the CPI calculation.
- Drivers with a history of traffic violations or involvement in crashes (Driver Factor) might contribute more to the overall CPI.
- Older vehicles with fewer safety features (Vehicle Factor) could increase crash potential.
- Adverse weather conditions (Environmental Factor) might significantly elevate the CPI for a particular location.

By combining these factors into a single index, the Crash Potential Index provides a comprehensive measure of crash risk, allowing transportation agencies to prioritize safety interventions and allocate resources effectively to mitigate potential hazards. Additionally, ongoing data collection and analysis can refine the CPI model over time to enhance its accuracy and relevance.

B. Multivariate Hazard Index (MHI):

- Formula: MHI = $\sum_{i=1}^{n}$ wi × Zi
- Description: The Multivariate Hazard Index combines multiple parameters to assess crash risk comprehensively. Each parameter (Zi) represents a specific aspect of safety, and weights (wi) are assigned based on their relative importance. Here's how the parameters listed could be incorporated:

  1. Location Factors:
     - Latitude and Longitude: These geographic coordinates could be used to identify the location of interest.
  2. Temporal Factors:
     - Day of Week: Represented as a categorical variable indicating different days of the week.
  3. Environmental Conditions:
     - Weather Conditions: Categorized into clear, rainy, snowy, etc.
     - Light Conditions: Categorized into daytime, nighttime, dawn, dusk.
     - Road Surface Conditions: Categorized into dry, wet, icy, etc.
  4. Driver and Vehicle Characteristics:
     - Age of Driver: Represented as categorical groups (e.g., young, middle-aged, elderly).
     - Gender: Categorized as male or female.
     - Vehicle Type: Categorized based on vehicle classification (e.g., passenger car, truck, motorcycle).
     - Age of Vehicle: Categorized into new, moderately aged, old vehicles.
     - Engine Capacity in CC: Represented as a continuous variable indicating vehicle power.
  5. Speed Limit and Enforcement:
     - Speed Limit: Represented as a continuous variable indicating the legal speed limit for the roadway segment.
     - Number of Police Officers Present at the Scene: Represented as a binary variable indicating presence or absence of law enforcement.

The Multivariate Hazard Index provides a holistic measure of crash risk by considering multiple factors simultaneously. It allows transportation agencies to prioritize safety interventions based on the combined influence of various parameters rather than considering them in isolation. The weights assigned to each parameter can be determined through expert judgment, statistical analysis, or optimization techniques to reflect their relative importance in contributing to crash risk.

C. Driver Risk Index (DRI):

The DRI aims to assess the risk of crashes by considering various parameters related to driver behaviour, vehicle characteristics, and environmental conditions.

DRI = $w1$ × P(Speeding) + $w2$ × P(Weather) + $w3$ × P(Light) + $w4$ × P(Road) + $w5$ × P(Driver) + $w6$ × P(Vehicle)

Where,

- P(Speeding): Probability of Speeding
- P(Weather): Probability of Adverse Weather Conditions
- P(Light): Probability of Low Light Conditions
- P(Road): Probability of Poor Road Surface Conditions
- P(Driver): Probability of Risky Driver Behaviour (e.g., based on age, gender)
- P(Vehicle): Probability of Vehicle-related Factors (e.g., vehicle age, engine capacity)

Parameters Incorporation:

- Probability of Speeding: Based on historical data, traffic studies, or enforcement data, the likelihood of speeding violations could be estimated for the specific location or roadway segment.
- Probability of Adverse Weather Conditions: Using weather data, such as precipitation and temperature, the probability of adverse weather conditions (e.g., rain, snow) affecting driving conditions can be determined.
- Probability of Low Light Conditions: Light conditions, such as daytime, nighttime, dawn, or dusk, could be categorized, and their impact on crash risk assessed based on visibility and driver reaction times.
- Probability of Poor Road Surface Conditions: Road surface conditions, including wet, icy, or uneven surfaces, can be quantified based on maintenance records or observational data.
- Probability of Risky Driver Behaviour: Demographic factors such as age and gender, along with behavioural factors like distracted driving or impaired driving, can be incorporated to estimate the likelihood of risky driver behaviour.
- Probability of Vehicle-related Factors: Vehicle characteristics such as age, type, and engine capacity can influence crash risk and can be quantified based on vehicle registration data or surveys.

Weighting Factors ($w1$, $w2$, …, $w6$): The weighting factors represent the relative importance of each parameter in contributing to crash risk and can be determined through expert judgment or statistical analysis of crash data.

The Driver Risk Index provides a comprehensive measure of crash risk by considering multiple factors related to driver behaviour, vehicle characteristics, and environmental conditions. It can help transportation agencies identify high-risk areas and prioritize interventions to improve road safety.

D. Speeding Risk Index (SRI):

This measure will integrate parameters such as latitude, longitude, day of the week, weather conditions, light conditions, road surface conditions, speed limit, and driver demographics.

SRI = $w1$ × Weather + $w2$ × Light + $w3$ × Road Condition + $w4$ × Day of Week + $w5$ × Driver Age + $w6$ × Speed Limit

Where:

- Weather: A categorical variable representing weather conditions (e.g., clear, rainy, snowy).
- Light: A categorical variable representing light conditions (e.g., daytime, nighttime, dawn/dusk).
- Road Condition: A categorical variable representing road surface conditions (e.g., dry, wet, icy).
- Day of Week: A categorical variable representing the day of the week (e.g., Monday, Tuesday, ..., Sunday).
- Driver Age: A continuous variable representing the age of the driver.
- Speed Limit: A continuous variable representing the posted speed limit for the roadway segment.

Description:

- Each parameter in the formula contributes to the overall Speeding Risk Index (SRI) based on its influence on the likelihood of speeding-related crashes.
- Weather, light conditions, and road surface conditions are categorical variables that reflect environmental factors affecting driving conditions and speed management.

- Day of the week captures variations in traffic patterns and enforcement levels that may influence speeding behaviour.
- Driver age accounts for differences in risk-taking behaviour and driving experience.
- The speed limit of the roadway segment serves as a reference point for evaluating speeding behaviour relative to legal limits.

Application:

- Transportation agencies can use the SRI to identify high-risk roadway segments prone to speeding-related crashes.
- Targeted interventions such as speed limit adjustments, enhanced enforcement efforts, road surface improvements, and driver education campaigns can be prioritized based on SRI values.
- Longitudinal analysis of SRI values over time can track the effectiveness of interventions and inform ongoing safety improvement efforts.

This example demonstrates how a surrogate safety measure like the Speeding Risk Index can integrate multiple parameters to assess the risk of specific types of crashes and inform targeted safety interventions.

X. DECISION TREE, RANDOM FOREST AND LOGISTIC REGRESSION

Decision Tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

A decision tree consists of three types of nodes:

- Decision nodes – typically represented by squares
- Chance nodes – typically represented by circles
- End nodes – typically represented by triangles

One purpose of Machine Learning is to design high performance classification systems from a set of representative samples of a population of data. An efficient way to tackle this kind of problematic is to combine an ensemble of individual classifiers to form a unique classification system, called Classifier Ensemble. This approach has been fed since the early 90's, by researches that have shown some combination principles to be particularly efficient, such as Boosting [8] (or Arcing [10]), Bagging [11], Random Subspaces [12], or more recently, Random Forests [13]. The efficiency in combining classifiers leans on the ability to take into account the complementarity between individual classifiers, in order to improve as much as possible the generalization performance of the ensemble. This ability is often defined through the diversity property. Although there is no agreed definition for diversity [14], this concept is usually recognized to be one of the most important characteristics for the improvement of the generalization performance in an ensemble of classifiers [15]. One can define it as the ability of the individual classifiers of an ensemble to agree mainly on good predictions and to disagree on prediction errors.

Random Forest is a family of ensemble methods. In a "classical" RF induction process a fixed number of randomized decision trees are inducted to form an ensemble. This kind of algorithm presents two main drawbacks: (i) the number of trees has to be fixed a priori (ii) the interpretability and analysis capacities offered by decision tree classifiers are lost due to the randomization principle. This kind of process in which trees are independently added to the ensemble, offers no guarantee that all those trees will cooperate effectively in the same committee.

Logistic Regression is a statistical method used for binary classification—that is, predicting the probability of one of two possible outcomes. It's widely used in various fields such as medicine, economics, and machine learning. Here’s an overview of the key concepts and steps involved in logistic regression:

Key Concepts

1. Binary Outcomes: Logistic regression is used when the dependent variable (target) is binary. For instance, predicting whether a patient has a disease (yes/no), or whether an email is spam (spam/not spam).
2. Logistic Function: The core of logistic regression is the logistic function (also called the sigmoid function), which maps any real-valued number into the range [0, 1]. The logistic function is defined as:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Where, z is the linear combination of input features.

3. Linear Combination of Features: Logistic regression computes a weighted sum of the input features (plus a bias term), which is then passed through the logistic function. Mathematically:

$$Z = β0 + β1x1 + β2x2 +…+ βnXn$$

Where, β0 is the intercept, β1, β2,…,βn are the coefficients for the input features X1, X2,…,Xn.

4. Probability Output: The output of the logistic function is the probability that the given input belongs to the positive class (e.g., probability of having a disease). For binary classification, we can define a threshold (commonly 0.5) to decide the final class label.

5. Cost Function: Logistic regression uses a cost function known as the log-loss or binary cross-entropy to measure the performance of the model. The goal is to minimize this cost function. The log-loss for a single training example is:

$$−y \log(y\text{^}) − (1−y) \log(1−y\text{^})$$

Where, y is the actual label (0 or 1) and y^ is the predicted probability.

XI. HYPERPARAMETER TUNING

Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the best set of hyperparameters for a machine learning model. Hyperparameters are parameters that are set before the learning process begins and control the behaviour of the training algorithm. Unlike model parameters, which are learned during training, hyperparameters are not directly updated during the training process.

The goal of hyperparameter tuning is to improve the performance of the model, typically measured by accuracy, precision, recall, F1 score, or other relevant metrics on a validation set. Here’s an overview of the key aspects of hyperparameter tuning:

A. Types of Hyperparameters

1. Model-Specific Hyperparameters: These include parameters specific to the model type, such as the number of layers in a neural network, the number of trees in a random forest, or the kernel type in a support vector machine (SVM).
2. Training Algorithm Hyperparameters: These include parameters that affect the training process, such as the learning rate, batch size, and number of epochs.

B. Common Hyperparameter Tuning Methods

1. Grid Search:
    - This method involves specifying a set of possible values for each hyperparameter and evaluating the model for every possible combination of these values.
    - It is exhaustive but can be computationally expensive, especially with a large number of hyperparameters or large value ranges.
2. Random Search:
    - Instead of trying every possible combination, random search selects random combinations of hyperparameters to evaluate.
    - It is often more efficient than grid search and can be more effective in finding good hyperparameters, especially when some hyperparameters are more important than others.
3. Bayesian Optimization:
    - This approach builds a probabilistic model of the objective function and uses it to select the most promising hyperparameters to evaluate.
    - It balances exploration of new areas of the hyperparameter space and exploitation of areas known to perform well.
4. Gradient-Based Optimization:
    - Techniques such as gradient descent can be used to optimize hyperparameters by treating the tuning process as a continuous optimization problem.
    - This method is more complex and typically used for hyperparameters that can be adjusted continuously.

C. *Practical Steps in Hyperparameter Tuning*

1. Define the Hyperparameter Space: Specify the range and distribution of values for each hyperparameter.

2. Choose a Tuning Strategy: Select a method such as grid search, random search, or Bayesian optimization.

3. Evaluate Model Performance: Train and validate the model using cross-validation or a separate validation set to evaluate the performance of each set of hyperparameters.

4. Select the Best Hyperparameters: Choose the hyperparameters that result in the best performance based on the evaluation metric.

5. Refinement (Optional): Refine the search by narrowing the hyperparameter ranges or using more advanced tuning methods.

D. *Challenges in Hyperparameter Tuning*

- Computational Cost: Evaluating multiple hyperparameter combinations can be computationally intensive.

- Overfitting: There is a risk of overfitting the hyperparameters to the validation set, leading to poor generalization on the test set.

- Dimensionality: The number of hyperparameters can lead to a high-dimensional search space, making optimization challenging.

XII. METHODOLOGY

This outlines the systematic process of developing a predictive model for road traffic accidents using Surrogate Safety Measures, starting from data collection and pre-processing, feature selection, model development, and evaluation, and finally, deployment and validation. It ensures that the model is accurate, reliable, and effective in preventing road accidents.

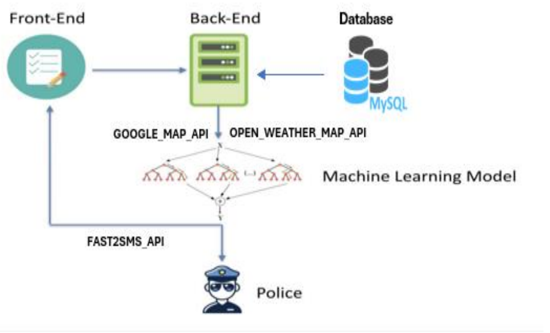


Figure 1: System Design

Front-End: Users input for the prediction factors are taken and sent to the backend server.

Back-End: The model is deployed here and the input data is fed into the Machine Learning model.

Database: MySQL DB is used for the user management to implement the user registration and login functionality. ML model will directly fetch the user data from the database, no need of manual entry.

Machine Learning Model: ML models used are decision tree, random forest and logistic regression and also applied hyperparameter tuning to increase its efficiency. Random Forest algorithm showed the highest accuracy of 89.50% and hence chosen for the model. The model runs and predicts the severity. The severity metrics are 1= Fatal, 2= Serious, 3= Slight. The output is sent back to the front-end and displayed to the user. A SMS containing the location coordinates and the severity of accident is sent to the police so that it can take preventive measures at the location.

A. *Data Collection*

Gathering relevant data related to road accidents, surrogate safety measures, and other relevant variables such as road characteristics, weather conditions, and traffic patterns. This data may come from various sources, including traffic cameras, accident databases, and road surveys. The dataset is taken from Kaggle The dataset contains 3 files - AccidentsBig.csv, CasualtiesBig.csv, and VehiclesBig.csv

The dataset is very huge and detailed based on the various parameters including location, weather conditions, road type, police force, number of vehicles, type of vehicle, speed limit, date, light conditions, accident severity, etc.

B. *Data Preprocessing and Data Visualization*

Cleaning and preparing the collected data for analysis, which may involve data validation, outlier detection, and data normalization. This step is crucial to ensure that the data used for modelling is accurate and reliable. vaex python library is used to load the data as pd dataframe. Visualizing the data based on total number of accidents on the day of the week, time of the day/night, age of people involved in the accidents, accident percentage in speed zone, co-relation between variables.

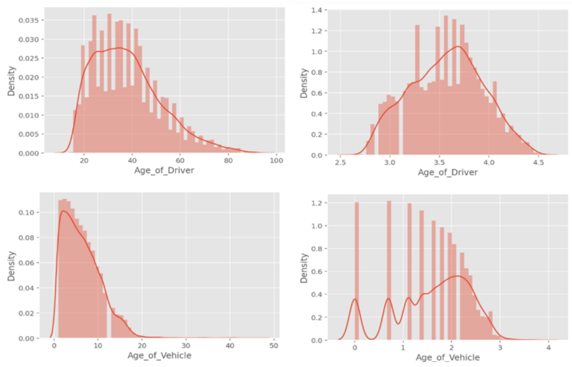


Figure 2: Data Normalization

C. *Feature Selection*

Identifying the most relevant surrogate safety measures that have a significant impact on road accidents. This may involve statistical analysis, feature ranking, and feature engineering techniques to select the most informative variables for the predictive model. Identify the most relevant surrogate safety measures (SSM) for accident prediction.

Potential SSM may include:

- Crash Potential Index (CPI)
- Multivariate Hazard Index (MHI)
- Driver Risk Index (DRI)
- Speeding Risk Index (SRI)

Using statistical analysis and feature ranking techniques to prioritize SSM and Conduct feature engineering to create new features if necessary.

D. *Model Development*

Building a predictive model that can estimate the likelihood of road accidents based on the selected surrogate safety measures. This may involve machine learning techniques such as:

- Decision tree
- Random forest
- Logistic regression

Then the data can be split into training and testing sets for model training and evaluation. Train the model using the training dataset and optimize hyperparameters. Finally, Validate the model's accuracy and performance through cross-validation techniques.

Here decision tree, random forest and logistic regression is used for ML model and also applied hyperparameter tuning to increase its efficiency. Random Forest algorithm showed the highest accuracy of 89.09% and hence chosen for the model. The model runs and predicts the severity. The severity metrics are 1= Fatal, 2= Serious, 3= Slight.

E. *Model Evaluation*

Assessing the performance of the developed model using appropriate evaluation metrics, such as:

- Accuracy: The proportion of correct predictions.
- Precision: The ability to make accurate positive predictions.
- Recall: The ability to identify all relevant instances.
- F1 Score: A combined metric of precision and recall.

Below data shows the accuracy level of each ML model which are used:

i. Random Forest: 89.50%

ii. Logistic Regression: 89.36%

iii. Decision Tree: 77.28%

iv. Logistic Regression with Hyperparameter tuning: 89.45%

v. Decision Tree with Hyperparameters tuning: 88.95%

vi. Random Forest with Hyperparameter tuning: 89.45%

F. *Model Deployment*

Implementing the developed model into a practical application, such as a traffic management system, which can provide real-time predictions of road accident risks to drivers, traffic authorities, or other relevant stakeholders. The model will be deployed in AWS Cloud.

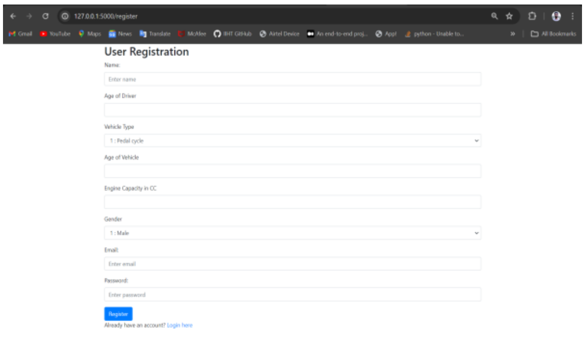
This involves, Integration with Traffic Management System, Realtime Predictions, Alert Generation, Visualization and User Interface, and Response Mechanism. The model runs and predicts the severity. The severity metrics are 1= Fatal, 2= Serious, 3= Slight. The output is sent back to the front-end and displayed to the user. A SMS containing the location coordinates and the severity of accident is sent to the police so that it can take preventive measures at the location.

G. *Model Validation*

Validating the predictive model using real-world data to ensure its accuracy and reliability in real-world scenarios. This may involve comparing the model's predictions with actual road accidents data to assess its performance and make necessary refinements.

XIII. RESULTS

The result of this project includes the development of a heat-map and a user interface for traffic authorities. These outcomes are crucial for providing actionable insights to enhance road safety.


127.0.0.1:5000/register

User Registration

Name: Enter name

Age of Driver

Vehicle Type: 1 : Pedal cycle

Age of Vehicle

Engine Capacity in CC

Gender: 1 : Male

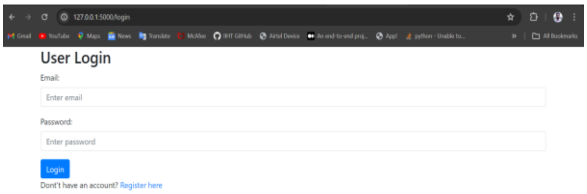
Email: Enter email

Password: Enter password

Register

Already have an account? Login here

Figure 3: User Registration


127.0.0.1:5000/login

User Login

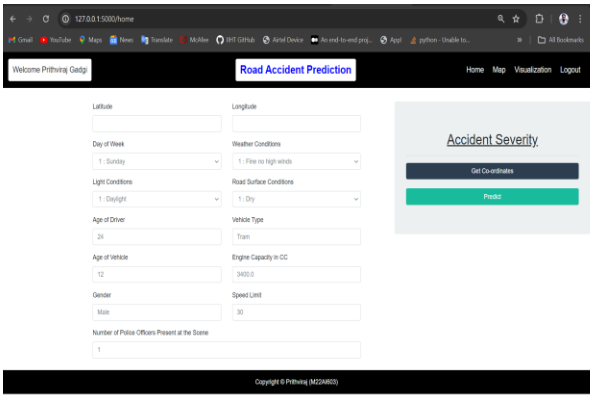
Email: Enter email

Password: Enter password

Login

Don't have an account? Register here

Figure 4: User Login


127.0.0.1:5000/home

Welcome Prithviraj Gadgi

Road Accident Prediction

Latitude:

Longitude:

Day of Week: 1 : Sunday

Weather Conditions: 1 : Fine no high winds

Light Conditions: 1 : Daylight

Road Surface Conditions: 1 : Dry

Age of Driver: 24

Vehicle Type: Tram

Age of Vehicle: 12

Engine Capacity in CC: 3400.0

Gender: Male

Speed Limit: 30

Number of Police Officers Present at the Scene: 1

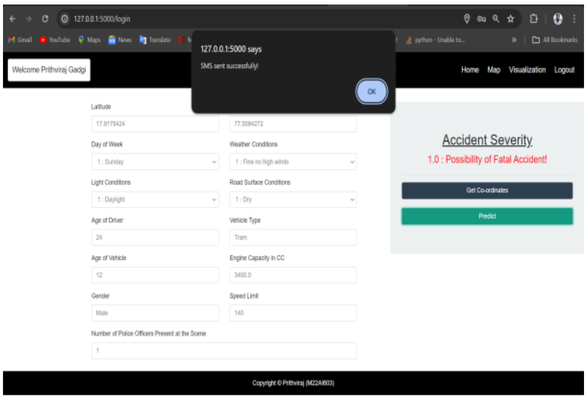
Accident Severity

Get Co-ordinates

Predict

Copyright © Prithviraj (M22AI603)

Figure 5: Home Page


127.0.0.1:5000/login

127.0.0.1:5000 says

SMS sent successfully!

OK

Welcome Prithviraj Gadgi

Latitude: 17.9175424

77.5094272

Day of Week: 1 : Sunday

Weather Conditions: 1 : Fine no high winds

Light Conditions: 1 : Daylight

Road Surface Conditions: 1 : Dry

Age of Driver: 24

Vehicle Type: Tram

Age of Vehicle: 12

Engine Capacity in CC: 3400.0

Gender: Male

Speed Limit: 140

Number of Police Officers Present at the Scene: 1

Accident Severity

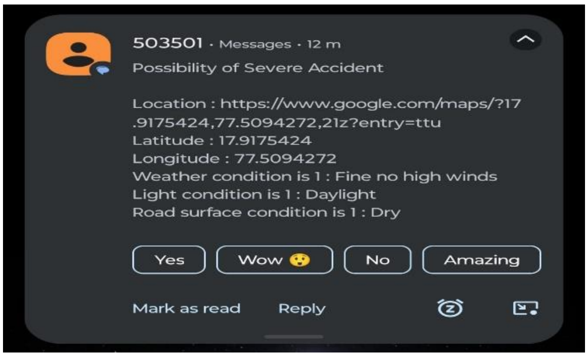
1.0 : Possibility of Fatal Accident!

Get Co-ordinates

Predict

Copyright © Prithviraj (M22AI603)

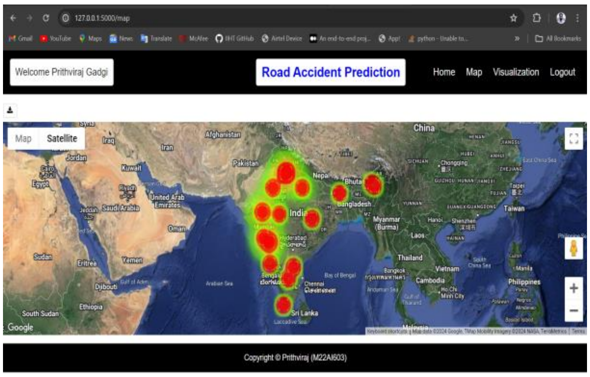
Figure 6: Accident Prediction


503501 · Messages · 12 m

Possibility of Severe Accident

Location : https://www.google.com/maps/?17.9175424,77.5094272,21z?entry=ttu
Latitude : 17.9175424
Longitude : 77.5094272
Weather condition is 1 : Fine no high winds
Light condition is 1 : Daylight
Road surface condition is 1 : Dry

Figure 7: SMS received with location and other details


127.0.0.1:5000/map

Welcome Prithviraj Gadgi

Road Accident Prediction

Copyright © Prithviraj (M22AI603)

Figure 8: Heat Map

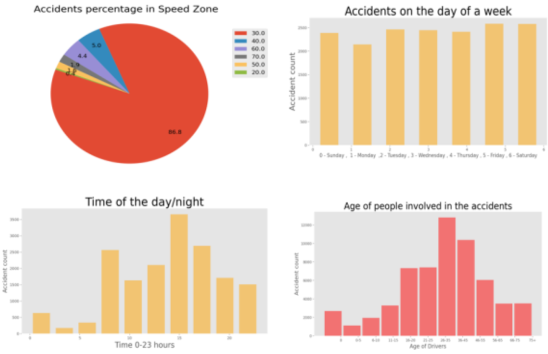


Figure 9: Data visualization

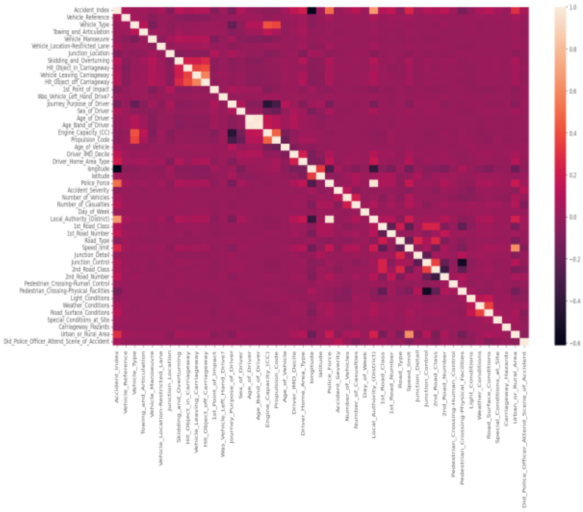
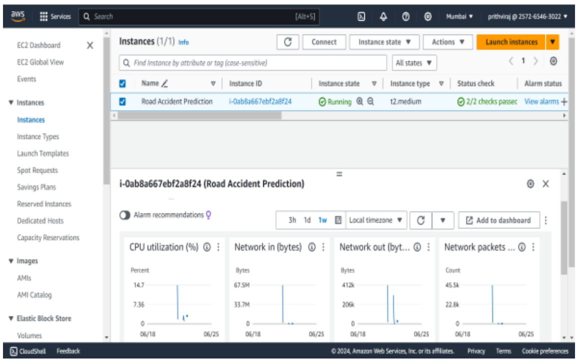


Figure 10: Co-relation between variables

The user needs to register by entering the necessary information in the registration form, the user data will be stored in MySQL database, and the password will base64 encrypted. In the login page the user can login by entering his/her email and password. Upon login the user will see home page of the web app where all the necessary information will automatically populated form the database, to update the co-ordinates user needs to click on Update Co-ordinates button, once the co-ordinates are updated the weather condition, light condition, day of week, road surface condition will be automatically fetched form the Open Whether Map API of that particular location. The user needs to click on Predict button, the ML model will predict the accident severity, and the application will send an SMS with all the necessary details using FAST2SMS API.



Instances (1/1) Info

☑	Name	Instance ID	Instance state	Instance type	Status check	Alarm status
☑	Road Accident Prediction	i-0ab8a667ebf2a8f24	Running	t2.medium	2/2 checks passec	View alarms

i-0ab8a667ebf2a8f24 (Road Accident Prediction)

Figure 11: Cloud Deployment

Finally, the application is deployed in AWS Cloud using Ubuntu 24.04 EC2 instance of instance type t2.medium which is having 2 CPU, 8GB RAM, 20GB of EBS Volume. The application is secured with HTTPS using a Self-Signed SSL Certificate.

Application Link: https://ec2-52-66-168-202.ap-south-1.compute.amazonaws.com:5000/

Git Repository: https://github.com/prithviraj-gadgi/road-accident-prediction.git

Data:https://www.kaggle.com/datasets/data125661/india-road-accident-dataset

Note: For the first time, the application link may prompt that the application is not secure. For cost cutting/demonstration purpose a Self-Signed SSL Certificate has been used to secure the application with HTTPS instead of running on HTTP. If it prompts, click on Advanced button and then click on Proceed Unsafe. For production servers, the SSL certificate from a trusted Certificate Authority (CA) will be used, some popular CAs include Comodo, DigiCert, GlobalSign, and GoDaddy.

XIV. CONCLUSION

In the face of escalating road traffic accidents and their devastating consequences, this project aimed to pioneer a proactive approach to road safety through the development of a predictive model utilizing Surrogate Safety Measures (SSM).

In conclusion, this project serves as a pioneering step toward safer and more efficient road transportation systems. By equipping traffic authorities and stakeholders with advanced tools and insights, it reinforces the mission to reduce road traffic accidents and create safer roadways for all. The project underscores the significance of data-driven innovation in addressing one of the most pressing challenges of our time — Road Safety.

XV. FUTURE ENHANCEMENT

- *Advanced Feature Engineering*: Continued advancements in feature engineering techniques. Exploring feature selection methods by considering temporal and spatial dependencies. Incorporating domain-specific knowledge to capture the complex dynamics of traffic accidents.

- *Multi-modal and Multi-source Data Fusion*: Integrating data from multiple sources and modes of transportation, such as pedestrian and cyclist data. By fusing data from various transportation modes and sources, predictive models can capture the interactions and interdependencies that contribute to accidents across different road user groups

- *Integration of IoT Technologies*: Integration of IoT can help in real-time data collection, real-time alerts and interventions. It can improve communication capabilities by using sensor networks and infrastructure such as connected vehicles and smart traffic signals.

- *Incorporation of External Factors*: Weather conditions, road construction, and special events, can enhance the predictive capabilities of the model. By integrating these external factors into the modelling process, the models can provide more accurate and context-aware predictions.

ACKNOWLEDGMENT

I would like to extend my deepest gratitude to Dr. Ranju Mohan for her invaluable guidance and support as my project supervisor. Her insightful feedback, unwavering encouragement, and extensive knowledge have been instrumental in the successful completion of this project. Dr. Mohan's dedication to excellence and her willingness to share her expertise have greatly enhanced my learning experience. I am truly grateful for her mentorship and the opportunity to work under her supervision.

REFERENCES

[1] Finn H Amundsen and Guro Ranes. 2000. Studies on traffic accidents in Norwegian road tunnels. Tunnelling and underground space technology 15, 1 (2000), 3–11.

[2] Gary A. Davis, John Hourdos, Hui Xiong, and Indrajit Chatterjee. 2011. Outline for a causal model of traffic conflicts and crashes. Accident Analysis Prevention 43, 6 (2011), 1907–1919. https://doi.org/10.1016/j.aap.2011.05.001

[3] John Hayward. 1971. Near misses as a measure of safety at urban intersections . Pennsylvania Transportation and Traffic Safety Center.

[4] Chinmoy Pal, Shigeru Hirayama, Sangolla Narahari, Manoharan Jeyabharath, Gopinath Prakash, and Vimalathithan Kulothungan. 2018. An insight of World Health Organization (WHO) accident database by cluster analysis with selforganizing map (SOM). Traffic injury prevention 19, sup1 (2018), S15–S20.

[5] Steven G Shelby et al . 2011. Delta-V as a measure of traffic conflict severity. In 3rd International Conference on Road Safety and Simulati. September. 14–16.

[6] Andrew P Tarko. 2018. Surrogate measures of safety. In Safe mobility: challenges, methodology and solutions. Vol. 11. Emerald Publishing Limited, 383–405.

[7] Andrew P. Tarko. 2020. Chapter 3 - Traffic conflicts as crash surrogates. In Measuring Road Safety Using Surrogate Events, Andrew P. Tarko (Ed.). Elsevier, 31–45. https://doi.org/10.1016/B978-0-12-810504-7.00003-3

[8] Lu Wenqi, Luo Dongyu & Yan Menghua, "A Model of Traffic Accident Prediction" INSPEC Accession Number: 17239218 DOI: 10.1109/ICITE.2017.8056908

[9] Abdel-Aty, M., N. Uddin, and A. Pande. Split Models for Predicting Multivehicle Collisions during High-Speed and Low-Speed Operating Conditions on Freeways. In Transportation Research Record: Journal of the Transportation Research Board, No. 1908, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 51–58.

[10] Thineswaran Gunasegaran Yu-N Cheah, "Evolutionary Cross validation" INSPEC Accession Number: 17285520 DOI: 10.1109/ICITECH.2017.8079960

[11] Simon Bernard, Laurent Heutte and Sebastien Adam, "On the Selection of Decision Trees in Random Forests" INSPEC Accession Number: 10802866 DOI: 10.1109/IJCNN.2009.5178693

[12] Rafael G.Mantovan,, Ricardo Cerri, Joaquin Vanschoren, "Hyper-parameter Tuning of a Decision Tree Induction Algorithm" INSPEC Accession Number: 16651860 DOI: 10.1109/bracis.2016.018

[13] Fu Huilin, Zhou Yucai, "The Traffic Accident Prediction Based on Neural Network", 2011

[14] Lin, L., Wang, Q., Sadek, A.W., 2014. Data mining and complex networks algorithms for traffic accident analysis. In: Transportation Research Board 93rd Annual Meeting (No. 14-4172).

[15] Gunasegaran, T., & Cheah, Y.-N. (2017). Evolutionary cross validation. 2017 8th International Conference on Information Technology (ICIT). doi:10.1109/icitech.2017.8079960