Development of a Data-Driven Driver Alert System for Vehicle Specific Risk Prediction in Mountainous Areas Using Historical and Real-Time Data: A Case Study of Ginigathena Police Domain

By

J.M.N.R.D. Jayaweera – TG/2020/700

A.G.H.D. Sewwandi – TG/2020/725

B.ICT(Hons)

Department of Information and Communication Technology
Faculty of Technology
University of Ruhuna
Sri Lanka

December 2025

Development of a Data-Driven Driver Alert System for Vehicle Specific Risk Prediction in Mountainous Areas Using Historical and Real-Time Data: A Case Study of Ginigathena Police Domain

By

J.M.N.R.D. Jayaweera – TG/2020/700

A.G.H.D. Sewwandi – TG/2020/725

A Research Dissertation Submitted in
Partial Fulfillment of the Requirement of the Research Project

For the Degree of
Bachelor of Information and Communication Technology

Department of Information and Communication Technology
Faculty of Technology
University of Ruhuna
Sri Lanka

December 2025

Declaration, Copyright Statement and the Statement of the Supervisor

"We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, we hereby grant to Department of Information and Communication Technology, the nonexclusive right to reproduce and distribute our dissertation, in whole or in part in print, electronic or other medium. We retain the right to use this content in whole or part in future works (such as articles or books)."

Name	Index Number	Signature	Date
J.M.N.R.D. Jayaweera	TG/2020700		
A.G.H.D. Sewwandi	TG/2020/725		

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

-------------------------------- --------------------

Ms. C.Y. Gamage, Date

Department of ICT,

Faculty of Technology,

University of Ruhuna, Sri Lanka

----------------------------- --------------------

Prof. P.K.S.C. Jayasinghe Date

Head of the Department,

Department of ICT,

Faculty of Technology,

University of Ruhuna, Sri Lanka

Acknowledgement

We would like to thank our research supervisor Ms. C.Y. Gamage for her continuous support and guidance through this work. She was responsible for supervising and monitoring our progress during this thesis work. She has been patiently monitoring our progress and guiding us in the right direction and offering her encouragement to us.

We would like to express our profound gratitude to Prof. P.K.S.C. Jayasinghe, Dean, Faculty of Technology, and the University of Ruhuna for providing us with this valuable opportunity.

We would also like to express our sincere gratitude to the Ginigathena Police Station for their invaluable assistance in providing and facilitating access to accident data essential for our research.

Finally, we would like to express our heartfelt appreciation to our family members, who have been a guiding light in our life. We are grateful to all our friends and well-wishers who have provided support throughout our accomplishments. Their support has been invaluable to us, and we truly appreciate having them in our lives. In closing, we would like to thank those individuals who choose to remain anonymous; their support was also greatly valued.

Abstract

The thesis presents a data-based, vehicle type-specific driver alert scheme designed to provide a prediction for the risk of a potential accident in a real-world environment in the Sri Lankan mountainous area of Ginigathena. Mountainous road areas are known for their steep sections, sharp bends, small carriageways, and changing weather conditions – all simultaneously increasing the likelihood of a collision. Currently available driver alert systems are designed to issue general alerts devoid of consideration for either the type of a vehicle as well as changing weather conditions.

An appropriate methodology was devised for accommodating the historical accident data, identifying significant risk factors, and combining them with information gathered from publicly accessible APIs like OpenWeatherMap. Supervised learning algorithms like Random Forest and XGBoost classifiers were employed to predict risk levels based on engineered features that encapsulated terrain, environmental, and vehicle-related information. An architecture for the system had to be designed to facilitate hybrid data processing, real-time computation, and mobile-platform dissemination.

The proposed system offers differentiated risk measures for motorcycles, passenger cars, vans, three-wheelers, and heavy vehicles. A prototype app was created to offer early hazard notifications and dynamic risk visualizations. System validation involved accuracy testing, scenario simulation testing, and usability testing. The study shows that combining historical and real-time variables improves accuracy significantly, and vehicle-type models increase the accuracy of risk notifications sent to drivers.

The proposed work brings something new, useful, and scalable to address issues concerning road safety in mountainous areas. The results have implications for transportation bodies, as well as opportunities for commercialization in mobility and safety.

Table of contents

List of Figures

List of Tables

List of Abbreviation

- **API:** Application Programming Interface
- **CSS:** Cascading Style Sheet
- **CSV:** Comma-Separated Values
- **GBM:** Gradient Boosting Machines
- **GPS:** Global Positioning System
- **GRU:** Gated Recurrent Unit
- **HD:** High Definition
- **ID:** Identifier
- **IoT:** Internet of Things
- **ITS:** Intelligent Transportation Systems
- **JSON:** JavaScript Object Notation
- **LiDAR:** Light Detection and Ranging
- **LMIC:** Low- and Middle-Income Countries
- **LSTM:** Long-Short Term Memory
- **MAE:** Mean Absolute Error
- **ML:** Machine Learning
- **OBD:** On-Board Diagnostics
- **REST:** Representational State Transfer
- **RMSE:** Root Mean Square Error
- **RNN:** Recurrent Neural Network
- **SHAP:** Shapely Additive Explanation
- **SPI:** Segment Priority Index
- **SVM:** Support Vector Machines
- **TF-IDF:** Term Frequency – Inverse Document Frequency
- **UI:** User Interface
- **WHO:** World Health Organization
- **XGBoost:** Extreme Gradient Boosting

1. INTRODUCTION

1.1. Background

Road safety is a continually recognized serious global concern in the field of public health, with 1.35 million road traffic deaths and 50 million injuries every year, according to the *WHO [1]*. This data highlights a serious problem that continues unabated in both the developed world and developing nations. Of greater concern in this respect is that the majority of road traffic deaths occur in low- and middle-income nations. Conditions understood to be common in these settings that contribute to road traffic deaths include poor infrastructure development, lack of enforcement of road safety standards, difficult geography, and road user behavior.

A similar situation is found in Sri Lanka, where road transport accidents are a prominent contributing factor for annual mortality. Notwithstanding the fact that the country's territory is reasonably small, the special nature of the landmass features a large number of areas that could be classified as mountainous or hill country wherein the accident toll is higher than that of plain or urban areas. The Ginigathena police division, located in the highland region of the country, stands out as a place that regularly reveals high risk zones. This region is known for a collection of roads that are constantly meandering and exhibiting steep gradients, marked horizontal curves, and rapid changes associated with the climatic factors.

Mountain road conditions pose specific danger factors. Steep slopes mean reduced traction and stopping distances for vehicles. Curved stretches limit sight distances and reaction times of road users. Sudden weather conditions like intense rainfall, fog, and strong winds add more danger to road users. Dangers like hydroplaning, skidding, instability, and road user mistakes typically happen together, making road users unpredictable for models based on traditional analysis. Inadequate lighting, lack of barriers, and road signs also add risks of road users in these areas.

For instance, in Sri Lanka, traditional driver warnings about road safety that come in the form of warning signs and speed reduction notices when crossing through certain sections of road do not vary in proportion to specific risk factors. Modern

GPS navigation systems that offer generic warnings against road hazards neither account for local terrain and accident statistics nor consider real-time environment data. As a result, drivers of vehicles that use mountainous routes often find themselves without adequate warnings to change their behavior.

There have been significant improvements in road safety systems through recent developments made in Intelligent Transportation Systems (ITS). With the inclusion of historical patterns of road accidents, current sensor data, as well as new communication technology, road safety systems through ITS have widened their capabilities to forecast potential risks dynamically at a route level. Specifically, machine learning approaches have strengthened capabilities to examine complicated relations not following linear paths between road risk factors using varied environmental, vehicle, as well as human-related parameters. Additionally, machine learning algorithms have the ability to process big datasets containing multiple modes with potential to extract hidden information not detectable through rule-based systems.

However, despite all these advances, the most prominent current ITS implementation and machine learning accident prediction model development efforts remain largely focused on either urban areas or expressways where there is a high level of instrumentation. This is because the aforementioned areas provide a more regular level of data availability and a more observable flow pattern that lacks the sudden changes found in a typical rural mountainous terrain like Ginigathena. Consequently, the presumptions that the current models are based on do not readily generalize into the target areas like Ginigathena.

The problem is further compounded by the absence of models that focus on specific vehicles in the currently popular risk forecast models. The different categories of vehicles, such as motorcycles, automobiles, buses, trucks, and three-wheeled vehicles, show differences in terms of stability, braking systems, weight distribution, and maneuverability. These differences significantly affect the dynamics of the vehicles on different road surfaces, such as slopes, bends, uneven roads, and situations where road visibility is limited. Yet, most existing models treat all vehicles as one when it comes to risk assessment.

In light of the challenges outlined above, recent research has increasingly explored data-driven approaches that integrate accident history, environmental context, and machine learning techniques for road-risk analysis. Such approaches aim to model complex, non-linear interactions between road geometry, weather conditions, and traffic behavior, particularly in environments where traditional rule-based safety mechanisms are insufficient.

The Ginigathena police region is used as a case study to examine road-risk prediction within a mountainous environment characterized by complex terrain and variable climatic conditions. Focusing on this localized context enables the analysis of risk patterns that are often underrepresented in studies centered on urban or expressway settings.

This thesis argues that accident-risk prediction in mountainous roads cannot be treated as a static, vehicle-neutral problem. In the Ginigathena police domain, risk varies rapidly due to microclimate changes and is not uniform across vehicle classes because stability, braking distance, and maneuverability differ across motorcycles, three-wheelers, cars, and heavy vehicles. Therefore, a driver alert system for mountainous terrain must combine (i) historical behavior patterns at segment level, (ii) real-time context (e.g., wetness and weather), and (iii) vehicle-specific decision rules to convert predictions into actionable alerts. The validity of this argument is assessed through quantitative model evaluation (R^2/MAE/RMSE for real-time SPI prediction, and F1-based alert performance), and by comparing global versus vehicle-specific thresholding to measure improvements in alert reliability.

1.2. Literature Review

Accident prediction and driver alert research cover various closely interlinked fields, such as machine learning, computational modeling of transportation, geospatial analysis, sensor-aided monitoring, risk assessment influenced by weather conditions, and analysis of vehicle dynamics. The following subsections include synthesis of these fields by highlighting both development and challenges.

1.2.1. Machine learning models for accident prediction

Machine learning algorithms have been proven useful in the traffic accident prediction task, with the ability to handle non-linear relationships involved in traffic accidents, environment, vehicles, and road conditions. In earlier studies, statistical modeling using techniques such as logistic regression and Poisson regression have been employed; such models tend to overlook complex relationships in traffic accidents.

Recent works based on machine learning have employed techniques such as ensemble learning algorithms, such as Random Forests, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM), which have proved to be more effective than the others for identification of accident-prone conditions *[2]*. These algorithms perform well in a high-dimensional space and can handle diverse features.

Abdel-Aty et al. [3] used real-time traffic speed data to detect unusual patterns of flow leading to accidents on highways. The results show the effectiveness of using dynamic traffic features in predictive models. Furthermore, the existence of an extremely optimized gradient boosting method, like XGBoost, with better resistance to noise, efficiency in training, and ability to deal with missing values, is a major reason for its popularity *[4, 5]*.

The application of Deep Learning approaches increases the predictive performance of existing forecast models. The capabilities of the Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are suitable for analyzing sequences over time and have proven their efficiency for spatial-temporal road collision prediction tasks *[6]*. They have the ability to learn patterns of time-dependent behavior, for instance, peak risk rates during rush hours and variations based on weather changes.

Despite progress, research shows that many models lack generalizability when applied to rural or mountainous regions, where data availability is limited and driving dynamics differ significantly. In Sri Lanka, *Liyanage and Rengarasu [7]* developed a foundational ML-based accident prediction system using local data; however, the model did not incorporate real-time conditions or vehicle-type differentiation - two critical factors for mountainous terrain.

Recently, experimental mobile-based systems in South Asia have attempted to Recently, in South Asia, some new mobile-based systems have attempted to disseminate hazard alerts using either crowd-sourced or sensor-based information; these have not addressed proactive accident alerts and terrain-related risk factors.

Critical comparison to this study. While prior ML studies demonstrate strong performance in general crash prediction, many are optimized for urban intersections or highways where risk drivers and data availability differ from mountainous roads. In contrast, this study targets a sparse, terrain-sensitive setting and therefore emphasizes time-consistent evaluation and features that encode local behavior (SPI) rather than relying only on generic traffic indicators.

1.2.2. Influence of terrain and road geometry on accident risk

Road geometry is an established factor in crash risk, with many studies isolating curvature, grade, width, and conditions of the shoulders as being major factors *[8]*. In mountainous terrain, all these conditions are exacerbated. The slopes increase the acceleration of gravity, specifically when traveling down, so braking efficiency is reduced. Curves reduce sight distance, which demands greater levels of driver control.

Studies undertaken in the Asian and European continents have indicated that road features with complex geometric properties like hairpin curves, narrow carriageways, and sudden transition of road gradients have a higher number of accidents occurring in these areas compared to expected numbers *[9], [10]*. In addition, terrain affects various other factors such as fog, wind speed, and rainfall.

In Sri Lanka, a highlands region has consistently shown high densities of accidents. Contributing causes include a lack of opportunities for overtaking, lack of guardrails, and deterioration of road surface conditions. Conventional maps of hazards that have been developed for these regions have largely been dependent on historical accident data and lack consideration of environment variables that influence time-changing hazards.

Abdel-Aty and Pande [3] emphasized that geometry-based risk cannot be understood independently from temporal flow conditions. However, integrating both elements remains computationally difficult without advanced modelling frameworks.

Few studies have attempted to combine geometry, microclimate variations, geospatial clustering, and behavioral indicators in a unified predictive model - leaving a significant methodological gap for mountainous road safety studies.

Critical comparison to this study. Geometry-focused work explains why mountainous segments are hazardous, but geometry alone cannot represent short-term changes caused by wetness, precipitation, and visibility. This study treats terrain risk as interacting with time and weather and operationalizes this through segment-conditioned behavioral indicators and real-time prediction.

1.2.3. Real-time data integration and IoT-based (Internet of Things) safety systems

The growth in Internet of Things (IoT) devices, open data platforms, and sensing technologies has increased the capabilities for developing dynamic road safety platforms.Real-time data such as weather predictions, traffic speed, and visibility values has a major role in determining crash probability in the shorter term.Application Programming Interfaces like OpenWeatherMap help in accessing dynamic variables such as:

- Intensity of precipitation
- Temperature and Humidity
- Wind speed
- Visibility levels
- Congestion and Traffic Density

There is vast evidence of the existence of strong links between weather changes and the number of accidents, especially for areas that experience foggy, rainy, or quickly changing weather patterns *[11], [12]*. Irregularities in traffic patterns have also been successfully used to predict accidents *[13]*.

The use of IoT-enabled monitoring systems incorporating accelerometers, traction sensors, and gyroscopic systems can identify skidding, deviations, or braking anomalies in vehicles. However, the application of such systems is normally constrained to highways with a heavy investment in instrumentation due to financial factors.

In a rural setting or mountainous region, real-time sensor infrastructure is less likely to exist. Therefore, schemes based only on on-road sensors might not work satisfactorily. The increasing availability of APIs of cloud computing-based solutions provides a promising third option: combining historical crash data with instantaneous weather and traffic conditions.

Critical comparison to this study. IoT-heavy solutions often assume dense instrumentation, which is unrealistic for many rural mountainous areas. This study addresses that limitation by using a hybrid design: historical accident patterns provide base context, while real-time environmental variables obtained via APIs support short-term risk updates without requiring extensive roadside sensors.

1.2.4. Vehicle-specific accident behavior

Vehicles also differ greatly with regard to handling stability, agility, braking distance, center of gravity, and resistance to external influences. There are many differences found in research:

- **Motorcycles:** Extremely vulnerable to wet roadways, skids easily, and lacks slope stability.
- **Three-wheelers:** Prone to overturning with a narrow wheelbase and a high center of gravity.
- **Buses and trucks:** Larger braking distances are needed; danger of accidents arises with inclines.
- **Passenger cars:** More stable but still prone to sharp turns and road irregularities.

Research shows that the factors of rolling resistance, traction control, and weight distribution patterns are responsible for additional hazards on mountainous road conditions *[14], [15]*. Despite these existing differences, the majority of research on accident prediction models does not consider the type of vehicle to be a factor affecting behavior, but an entire category of inputs. This leads to inaccurate prediction of hazards either on the overstatement or understatement of hazards based on the type of motor vehicle.

The only exception in Sri Lanka exists in the form of an alert notification software designed by *Jayamanna* and *Kalansooriya*, but this did not involve machine learning models, nor did they calibrate to each type of motor vehicle *[16]*. Thus, they give out an entire warning notification to all drivers that may not necessarily be linked to the actual motor vehicle being driven.

1.3. Research Gap

Several limitations are founded while reviewing literature.

1. **Lack of risk differentiation based on specific vehicles:** Current approaches to prediction results in a single value of risk that can be applied universally for all vehicles. They do not consider the crucial differences that exist among vehicles based on the distance involved in the stop, the ability to turn, the stability of the vehicle, or exposure to terrain hazards. This affects the relevance of the warning messages issued significantly in areas like Ginigathena.

2. **Limited Hybrid Data Fusion for Real-Time Prediction:** A major part of the related works relies solely on either past databases or road conditions. Real-time variables like fog density, precipitation intensity, or current traffic patterns, potentially hazardous to drivers, have often not been considered.

3. **Poor transferability to mountainous road environments:** Models designed for freeways or urban intersections do not generalize well to hilly terrain. Mountainous regions exhibit nonlinear risk behaviors caused by interactions between geometry, weather, and driver behavior. Very few studies combine these variables effectively in a unified predictive model.

4. **Absence of localized, data-driven alert systems in Sri Lanka:** Also, there is a gap in fully integrated systems focusing on alerting the driver in rural mountainous roads in Sri Lanka, although there have been improvements in research studies in analyzing accidents in the country.

Addressing these gaps

The present study introduces a hybrid, vehicle-specific ML-based approach incorporating:

- Historical accident behaviour
- Real-time environmental conditions
- Terrain-related risk variables
- Vehicle-dependent thresholds
- Dynamic SPI prediction

This integrated modelling framework directly addresses limitations in existing systems while contributing a novel approach to mountainous road safety.

1.4. Summary of the Chapter

This chapter describes the background, literature review, and gaps in existing research on the prediction of accidents on mountainous roads. The chapter emphasizes the importance of innovative systems that are aware of their environment, vehicle-specific, and data-oriented. Recent advances in machine learning algorithms and the integration of ITS have opened up possibilities in this regard; however, a large gap in existing methodologies exists in the Sri Lankan environment.

Based on the identified gaps, it is necessary to develop an intelligent driver alert system using machine learning technology specific to Ginigathena area. This system integrates historical and real-time information for constant and adaptive risk forecasts. The following chapters define the research problem statement, aims and objectives, framework, findings, and implications for improving road safety.

2. RESEARCH PROBLEM

2.1. Introduction to the Research Problem

Road accidents in mountainous areas are a very complex and divers issue in the area of transportation and safety measures related to them. The presence of steep grades, sharp turns, and a lack of visibility in mountain areas creates a dangerous environment for drivers and leads to a situation significantly different from other environments, either in urban or highway areas. In Sri Lanka, the Ginigathena police division is one of the most accident-prone zones in the mountain region; this is because of a lack of situational awareness and inadequate awareness of dangerous zones among drivers in the area.

Present systems meant to alert drivers are of a generalized nature both locally and globally, where all drivers receive standardized warnings irrespective of changes in terrain, vehicle, or the dynamic nature of the environmental context. Such systems only contribute to ineffectiveness in helping to guide drivers along routes that involve rural mountainous areas. Furthermore, most systems that predict road accidents rely heavily on historical information and are therefore quite inefficient in dealing with current changes in context such as sudden rains or formation of fog.

2.2. Problem Context

The major challenge in ensuring mountainous area road safety is that the dynamic conditions of risk levels are not addressed well in the static models of prediction. In areas like Ginigathena, for instance, risk levels can change in just a few minutes due to variations in the weather conditions. Furthermore, the driver is not usually aware of the conditions; hence the risk of accidents increases.

Moreover, mountainous roads have their own set of conditions that restrict each class of vehicle differently. For example:

- Motorcycles tend to be more sensitive to wet road surfaces, cross winds, and lane edges.

- Heavy vehicles such as buses and trucks will necessitate longer stopping distances and will be prone to the threat of rollover on steep downgrade sections.
- Small passenger car vehicles may have problems with traction on curved, wet roadways.

Although these differences have been established, it is observed that the current majority of systems give the same warning related to hazards meant for all types of vehicles, thus limiting the efficiency and accuracy level of alerts given in real time. Furthermore, it should be noted that Sri Lanka currently does not have an integrated system incorporating current information on accidents as well as real-time information from weather API sources, traffic flow indicators, and road conditions. Lack of integration leads to incomplete information on the hazards associated with driving on hills.

2.3 Statement of the Research Problem

Based on the identified issues, the core research problem addressed in this study is formulated as follows:

"The existing system for warning drivers is not taking into consideration the driving behaviors of the vehicle as well as the environmental data both historically and currently, thus limiting their ability to effectively predict and communicate accident risks associated with a mountainous road environment like Ginigathena."

This formulation of the problem points out that there are three areas in which current systems are deficient:

1. The lack of modeling specific to vehicles, despite known differences in operating characteristics and stability for different types of vehicles.
2. Lack of proper integration of historical and real-time information, causing risks to be predicted inaccurately amidst rapidly changing conditions of mountainous regions.

3. The limited applicability of existing models in predicting accidents in rural hilly areas, where the interplay of road geometry and climate is complex.

2.4 Need for the Proposed Solution

A comprehensive, data-driven, and adaptive warning system must be developed to improve drivers' situational awareness on mountainous roads. By taking past accident data along with current environmental data and tailoring estimates based on vehicle type, this proposed system can be expected to offer accurate environment-dependent risk estimates, which have several benefits, including:

- Allow for dynamic warning notifications based on recent risk analysis instead of static risk assumptions.
- Enhancing decision-making capacity in drivers and lessening occurrences of misjudgment.
- Improve safety performance by incorporating geometric, meteorological, and vehicle-related information.
- Offering data-driven insights to law enforcement organizations and policymakers on infrastructure development.

In conclusion, resolving the identified issue in the study can have a significant role in decreasing accidents in a place like Ginigathena and developing a modern and intelligent transport safety system for Sri Lanka.

3. RESEARCH OBJECTIVES

3.1 Need for the Proposed Solution

This research responds to the limitations described in Chapter 2 by developing a vehicle-specific and real-time risk prediction approach suitable for mountainous roads, where risk changes rapidly and differs across vehicle categories.

3.2 Main Objective

The main objective of this research is:

"To develop a data-driven, vehicle-specific driver alert system capable of predicting real-time accident risk in mountainous road environments by integrating historical accident data with real-time environmental and roadway information using machine learning techniques."

This objective captures the core contribution of the study-combining predictive analytics with vehicle-specific modelling to improve road safety in regions such as Ginigathena.

3.3 Specific Objectives

The subsequent specific objectives break down the general objective into components as follows:

Objective 1:

To examine historical accident data in the Ginigathena police region and determine key factors in road accidents in a mountainous terrain.

Justification: Knowledge of historical trends provides the basis for feature engineering and machine learning model design.

Objective 2:

To gather and consolidate real-time environmental data, including weather information, road surface attributes, and flow statistics, from publicly available application programming interfaces.

Justification: Real-time data allows for dynamic predictive modelling, which can adjust effectively for fast-changing environments characteristic of mountains.

Objective 3:

To build machine learning models - namely, Random Forest and XGBoost - that could predict levels of accident risk on the basis of both past and real-time data sources.

Justification: Such models are very effective at modelling complex, nonlinear relationships such as those that exist on mountainous routes.

Objective 4:

To develop risk profiles for specific vehicles through the calibration of prediction models for different types of vehicles like motorcycles, passenger vehicles, and heavy vehicles.

Justification: Each type of vehicle behaves differently when exposed to terrain and climatic conditions; hence, specific modeling is required for developing precise warning messages.

Objective 5:

To design and implement a prototype system that notifies drivers of potential dangers using real-time predictions generated by the developed models.

Justification: There needs to be a focus on ensuring that the predictions are supplemented by the appropriate context in order to successfully inform the drivers.

Objective 6:

To evaluate system performance using accuracy metrics, simulation scenarios, and usability testing, particularly for mountainous driving conditions.

Justification: It is imperative that a system be thoroughly tested before it is used.

3.4 Summary of the Chapter

The principal and specific objectives in this chapter are identified in relation to the development of the driver alert system as suggested in this research. The reasons behind the specific objectives are the need to include the behaviour of each vehicle in the alert system through the usage of multiple data inputs, machine learning algorithms, and validation of the alert system in a mountainous region.

4. METHODOLOGY

4.1. Introduction

This chapter outlines the overall methodological approach utilized in the development of the data-driven, vehicle-based driver alert system proposed in this work. The methodological approach is a combination of historic accident data, environmental information, engineered variables, and a risk-scoring engine with the capability of providing dynamic safety alerts in a mountainous environment. This method uses a hybrid approach that blends statistical smoothing, machine learning approaches, and feature and threshold approaches that are vehicle dependent.

Two major pipelines constitute the system:

- **Historical Data Pipeline**: Historical Data Pipeline: focuses exclusively on building the Segment Priority Index (SPI), risk segmentation, classification of causes, and risk models at the segment level.
- **Real-Time Risk Prediction Pipeline:** using an XGBoost-based predictive model with a complete preprocessing pipeline and vehicle-specific alert thresholds.

These components form a unified end-to-end mechanism for predicting, evaluating, and communicating accident risks in the Ginigathena police domain.

The methodological design follows the thesis argument that mountainous risk is dynamic and vehicle dependent. Historical modelling is used to encode persistent segment-level behavioral patterns, while the real-time pipeline predicts continuous risk indicators under changing environmental conditions. Vehicle-specific thresholding is included because identical numerical risk values do not produce equally meaningful alerts across vehicle classes. The evaluation strategy prioritizes time-consistent testing to reflect deployment conditions, and interpretability is included to support practical use by authorities and drivers.

4.2. Overall Research Framework

1. **Data Preparation Layer**
   - Historical Accident Log Processing
   - Cleaning and Normalization of Fields
   - Generation of timestamps
   - Feature engineering techniques such as SPI, Binning, and Segment ID
2. **Modeling Layer**
   - SPI Smoothening & Aggregation of Risk Tiles
   - Cause-of-incident classifier
   - Segment rate prediction model
   - Real-Time XGBoost Model for Regression & Classification Tasks
   - Vehicle-specific thresholds for high-risk mapping
3. **Layer of Integration and Risk Scoring**
   - Merging of Model Outputs
   - Combination of cause-probability, segment-rate, XGBoost risk, and vehicle multipliers
   - Final risk score calculation on a scale of 0-100
4. **Deployment Layer**
   - APIs for real-time integration
   - Alert Delivery to End Users
   - Saving models and metrics using pipelines

Figure 4.1 presents the overall architecture of the proposed risk prediction framework. It illustrates how historical accident data are processed through feature engineering and SPI construction to form the historical risk engine, which combines cause-of-incident classification and segment-rate modeling. The outputs are then integrated with real-time contextual data in the real-time risk prediction pipeline, enabling vehicle-specific thresholding, model interpretability, and final risk score generation for driver alerts.

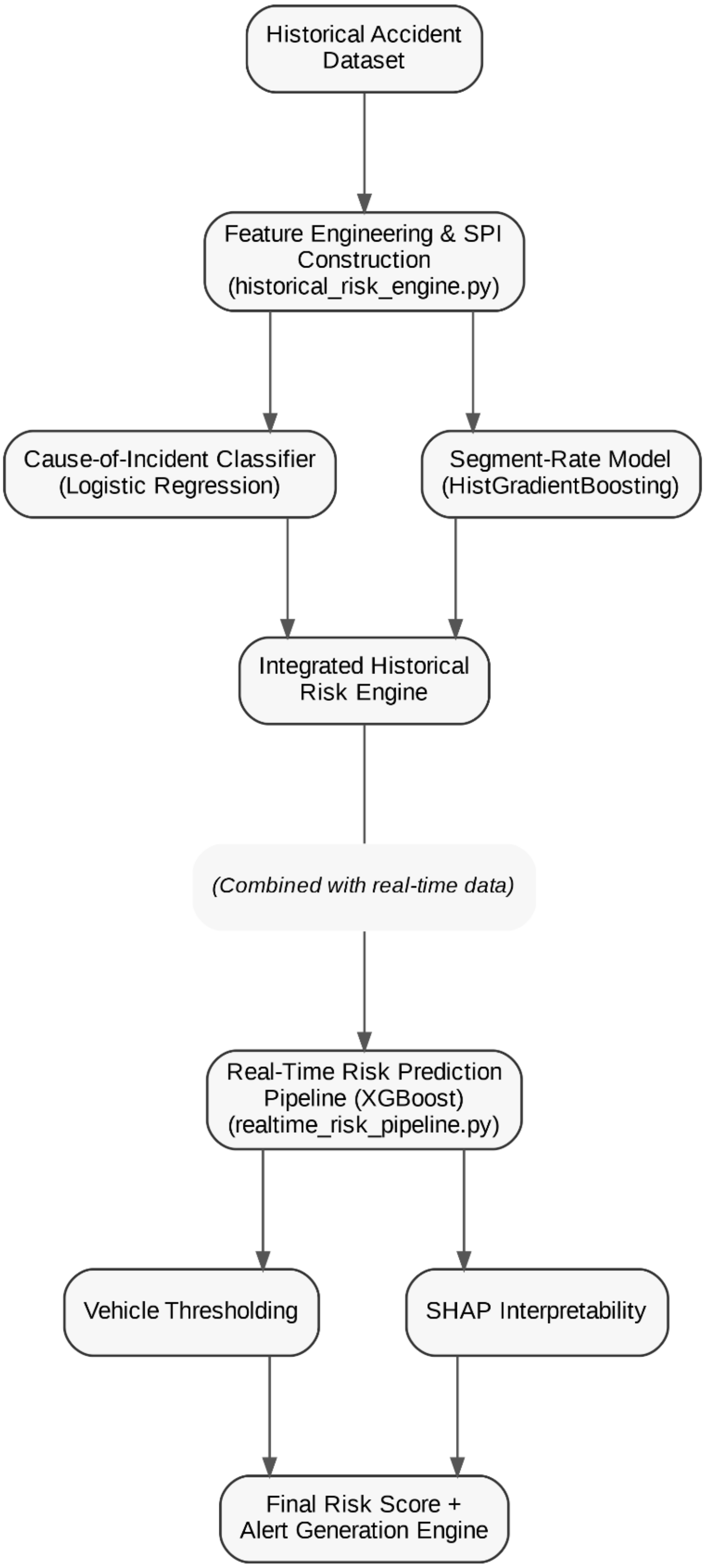

Historical Accident Dataset
Feature Engineering & SPI Construction (historical_risk_engine.py)
Cause-of-Incident Classifier (Logistic Regression)
Segment-Rate Model (HistGradientBoosting)
Integrated Historical Risk Engine
(Combined with real-time data)
Real-Time Risk Prediction Pipeline (XGBoost) (realtime_risk_pipeline.py)
Vehicle Thresholding
SHAP Interpretability
Final Risk Score + Alert Generation Engine

Figure 4.1: High-Level System Architecture Diagram

4.3. Dataset Preparation

The Data Preparation Layer involves processing the raw accident logs, cleaning and normalizing data, and creating engineered features to improve the predictive performance of the model.

Historical accident data from the Ginigathena police division was used as the foundation of model training. The dataset included:

- Vehicle type
- Accident location (latitude and longitude)
- Time of the accident
- Weather conditions at the time
- Road surface status
- Reported reason for the accident
- Free-text descriptions written by officers

This raw dataset was then processed through the pipeline defined in *historical_risk_engine.py*.

4.3.1. Data cleaning

Data preprocessing was carried out to increase the consistency and reliability of the accident data before building the model. The categorical variables of Vehicle, Place, and Position were normalized by carrying out uniform capitalization to avoid the creation of multiple representations of categories. The Reason category was normalized by removing null equivalent values and ensuring uniform formatting of categories related to the cause of an accident.

Environmental features, such as temperature, precipitation, humidity, and wind speed, were changed to numeric data types, with errors handled for the purpose of model training. Moreover, latitude and longitude points were grouped into meaningful bins based on the geographical regions.

4.3.2. Timestamp normalization

Timestamp normalization was used to place the accident reports into a standardized time framework for the purpose of analysis. Since the lack of consistency was evident in the time information, the system either identified the existing timestamp fields, which may have been named Datetime, DateTime, or Timestamp, or generated the timestamps based on the combination of Date and Time fields depending on the need.

The normalized timestamp generated ensured the maintenance of the chronological ordering of the events, facilitated the calculation of time features like hour of day and day of week, and aided in time-aware splitting. This made it feasible for models to learn time-related risks and ensured predictions considered real-time deployment scenarios rather than purely data-ordered conditions.

4.4. Feature Engineering

The following features are engineered from the processed data:

- **Temporal Features:** Describe time-based variables such as day of the week, hour, etc.
- **Spatial Features:** Capture location-based information like GPS coordinates.
- **Behavioral Features:** Include driver or vehicle behavior patterns.
- **SPI Construction**: Creation of the Safety Performance Index (SPI) using combined features.

4.4.1. Temporal features

- Hour of day
- Day of week (0–6)
- Weekend indicator
- Wet-road flag (isWet) from precipitation

Figure 4.2 shows accident distribution per day of the week, and Figure 4.3 illustrates accident occurrences per hour of the day, indicating risky periods of travel, which are mainly concentrated between daylight and early evening periods.

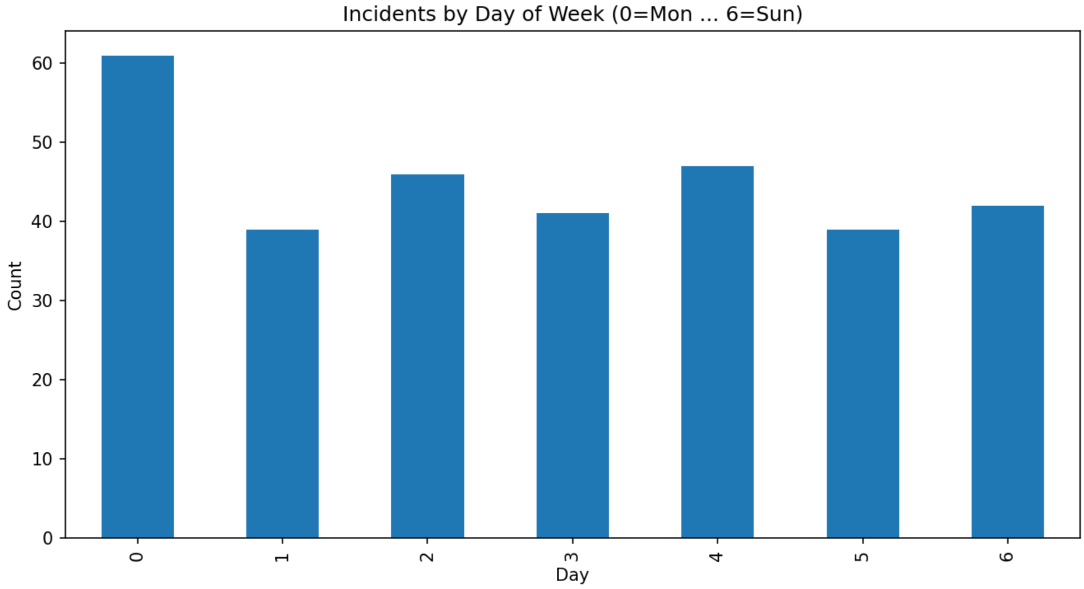


Figure 4.3: Incident by day of the week

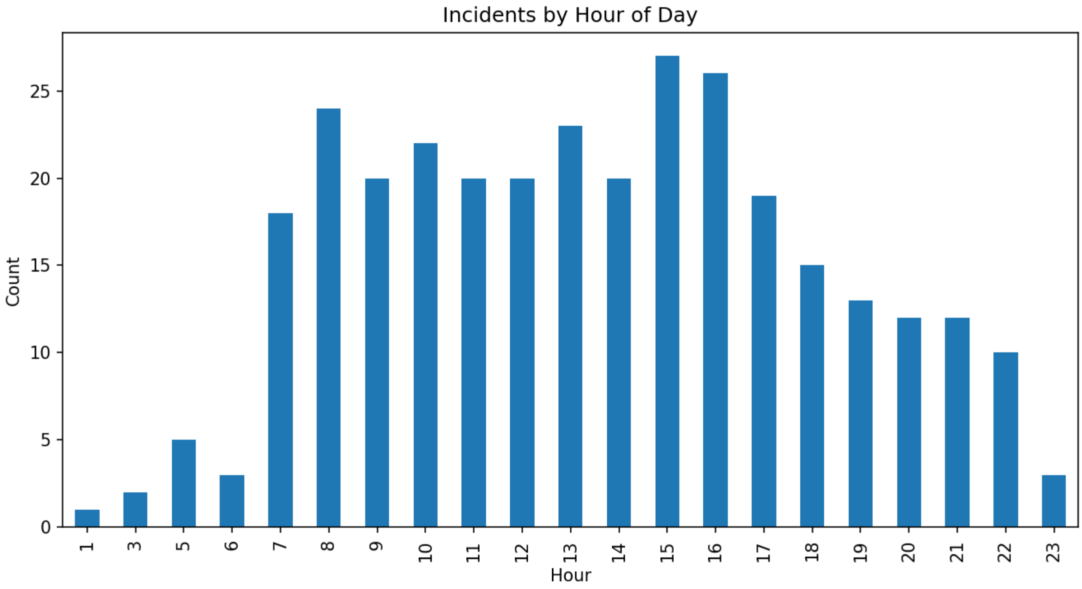


Figure 4.2: Incidents by hour of the day

4.4.2. Spatial features

- Latitude/longitude bins (rounded to 3 decimals)
- Segment ID = combination of lat_bin + lon_bin

These create consistent road segments for SPI modelling.

Figure 4.4 shows the geographical distribution of the reported accident sites in the Ginigathena police area is depicted with the aid of latitudes and longitude values. As shown in the map, there are particular road stretches with high concentrations, which are probably prone to high risk due to existing factors in mountainous areas.

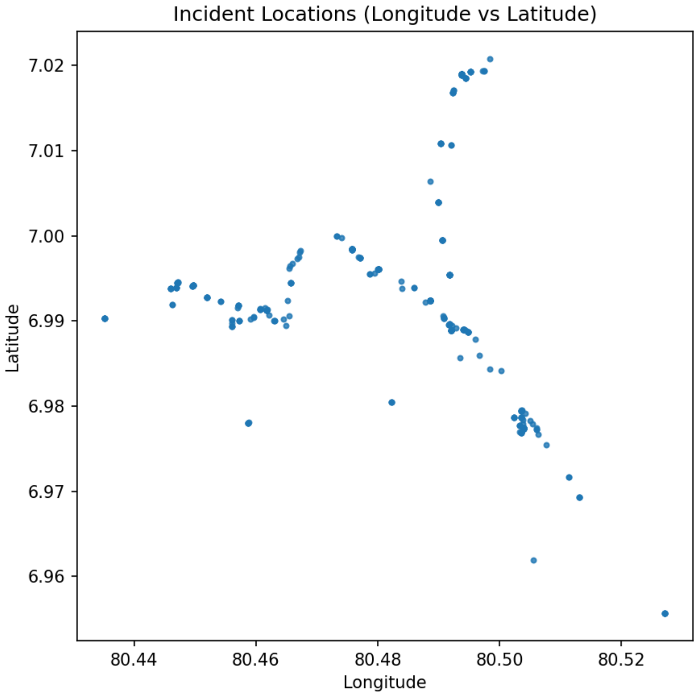


Figure 4.4: Spatial distribution of incident coordinates in the Ginigathena domain

4.4.3. Behavioral features

- is_speed_reason → identifies speeding-related accidents
- Description text → later vectorized using HashingVectorizer

Behavioral characteristics were captured using an indicator variable (is_speed_reason) to explicitly identify accidents associated with excessive speed. In addition, textual accident descriptions were transformed using the HashingVectorizer, which was selected over vocabulary-based methods (e.g., Term Frequency – Inverse Document Frequency (TF-IDF)) because it avoids explicit vocabulary construction, ensures constant memory usage, and is more robust to sparse, small-scale datasets with evolving or noisy text entries, making it suitable for real-time risk prediction contexts.

4.5. SPI Construction

The Speed Propensity Index (SPI) captures how frequently speeding-related incidents occur for a specific segment, time, wetness, and vehicle type.

Code computes SPI using:

$$SPI_{smoothed} = \frac{\left(n * local_{spi} + \ \alpha \ * global_{spi}\right)}{(n + \ \alpha)}$$

Where:

- **n** = number of incidents in the group
- **local_spi** = proportion of speed-related incidents for the segment
- **global_spi** = global average speed-related proportion
- **α = 20** = smoothing parameter

This ensures stable estimates even for low-frequency segments.

SPI is then merged back into the full dataset as a key predictive feature.

Justification of SPI smoothing (α = 20). SPI is computed over fine-grained groups (segment, hour, wetness, and vehicle), which creates many low-count groups in a 315-record dataset. Without smoothing, SPI can become unstable (near 0 or 1) in sparse groups and would overstate risk. The smoothing parameter α acts as a stabilizing prior (equivalent to adding pseudo-observations), reducing variance for rare groups while allowing frequent segments to remain data driven.

Figure 4.5 illustrates the distribution of the smoothed Speed Propensity Index (SPI_smoothed) after applying Bayesian smoothing with α = 20, demonstrating the stabilization of risk estimates across low-frequency segments.

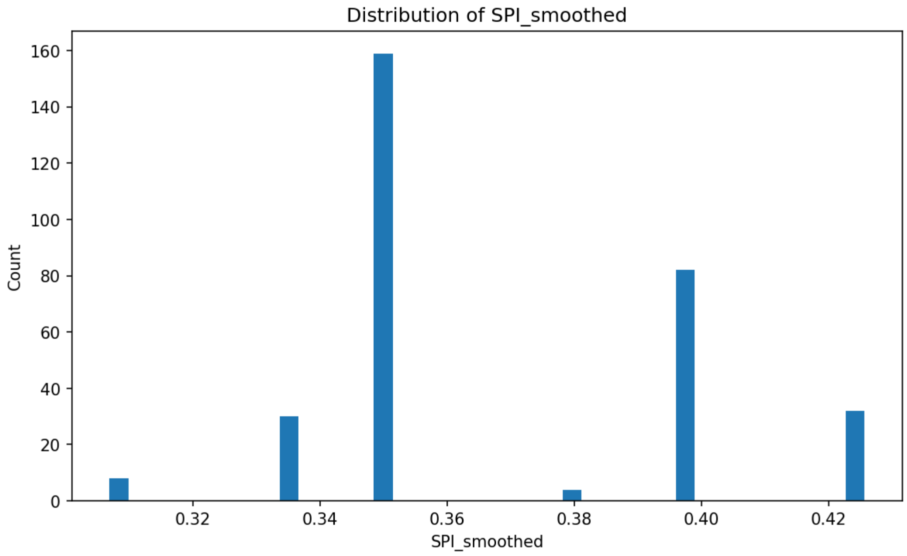


Figure 4.5: Distribution of SPI_smoothed after Bayesian smoothing (α = 20)

4.6. Historical Modelling Components

The system is built around three main modelling components within the historical data pipeline:

4.6.1. Cause-of-incident classifier

Using historical_risk_engine.py pipeline, a multi-class Logistic Regression model predicts the likely cause of an incident based on environmental, temporal, spatial, and textual features.

Features include:

- Temperature, humidity, precipitation, and wind speed
- Hour of the day and day of the week
- A wet road indicator
- Latitude and longitude

- Vehicle type
- Location-related fields such as place and position
- Text descriptions from the accident reports, which are transformed using TF-IDF vectorization

Evaluation metrics include:

- Classification accuracy
- Macro precision/recall
- Macro F1-score
- Confusion matrix
- Weighted F1-score

Output includes predicted cause and class probabilities.

Logistic regression was selected because it produces calibrated class probabilities and remains stable on modest sample sizes. This aligns with the goal of producing interpretable cause predictions for operational use. TF-IDF features were included because officer narratives contain discriminative keywords that are not captured in structured fields, improving class separability for frequently confused causes.

Figure 4.6 presents the confusion matrix of the cause-of-incident classifier, highlighting the classification performance across different accident causes and revealing patterns of misclassification while Figure 4.7 shows the most frequently reported causes of road accidents in the historical dataset, providing insight into dominant contributing factors in the study region.

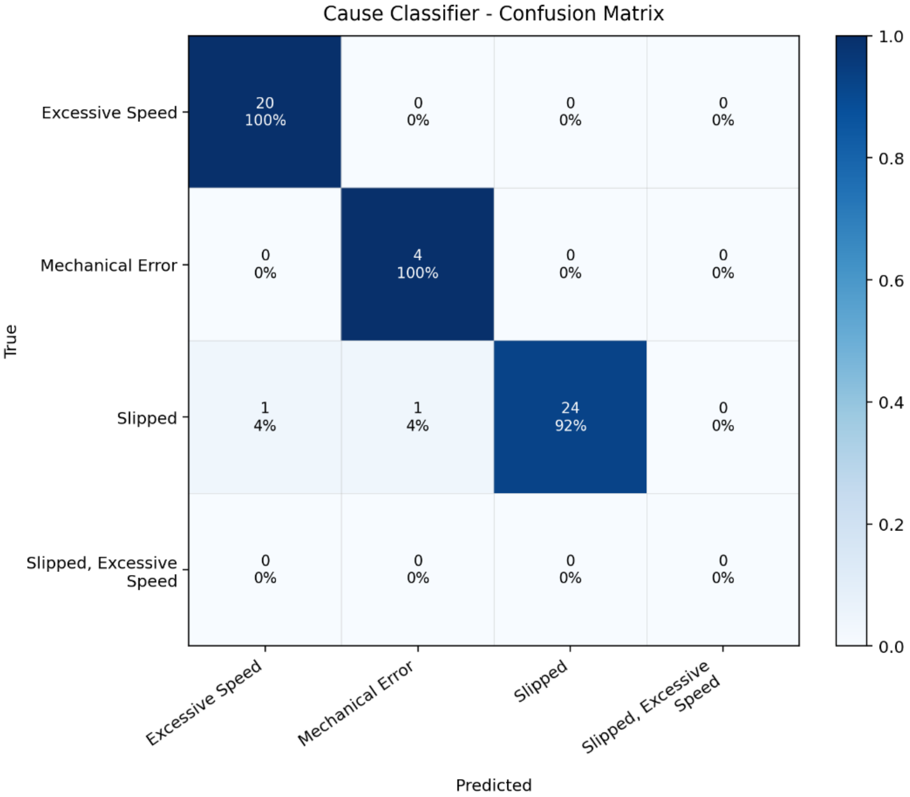


Figure 4.6: Confusion matrix for cause-of-incident classifier

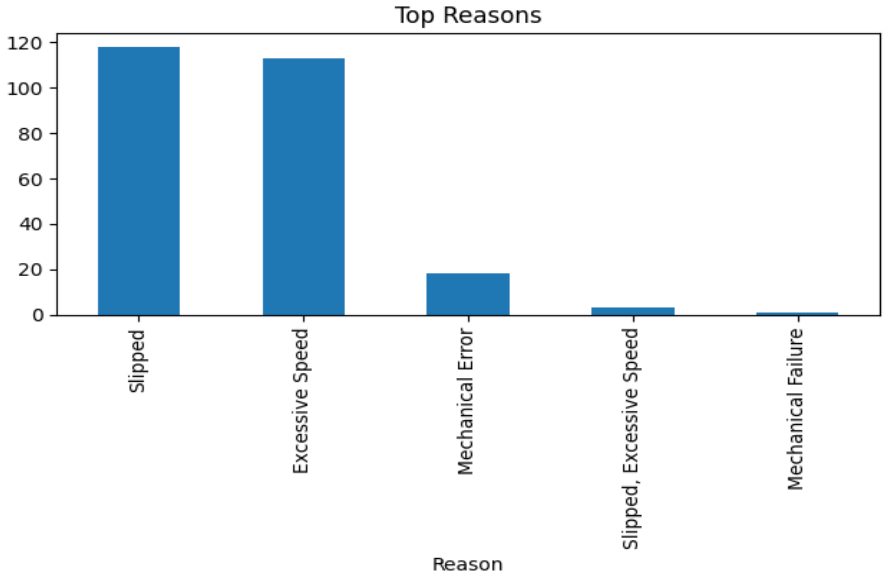


Figure 4.7: Top Reasons

4.6.2. Segment-rate model

A **HistGradientBoostingRegressor** predicts the incident rate for a given segment based on:

- Hour
- Day of Week
- Wetness
- Vehicle type (one-hot encoded)

This model outputs a continuous risk rate used in downstream risk scoring.

Model choice rationale. A histogram-based gradient boosting regressor was used to model segment-level incident rates because it captures nonlinear interactions between time (hour/day), wetness, and vehicle type while remaining computationally efficient. This is suitable for frequent retraining and for generating a stable background-risk component for the integrated risk score.

4.6.3. Risk tiles dataset

A grid representation ("risk tiles") combines:

- Segment ID
- Hour, Day
- Wetness
- Vehicle type

With aggregated incident frequencies and SPI tile values.
This dataset allows real-time scoring models to rapidly retrieve relevant historical context for prediction.

4.7. Real-Time XGBoost Risk Prediction Pipeline

realtime_risk_pipeline.py code defines an advanced real-time modelling pipeline using:

- XGBRegressor for SPI regression
- Dense HashingVectorizer for text
- OneHotEncoder for categorical fields
- SimpleImputer for missing values
- ColumnTransformer for structured preprocessing.

4.7.1. Full preprocessing pipeline

Preprocessing pipeline includes:

Text Features

- Vectorized with HashingVectorizer (512-dim)
- Fast, memory efficient, avoids vocabulary explosion

Categorical Features

- Imputed with most-frequent value
- One-hot encoded

Numeric Features

- Median imputation

Output

All transformations are combined into a **single scikit-learn Pipeline**.

4.7.2. Train-test split

Model uses Chronological Splitting:

- First 80% for training.

- Last 20% for testing.
- Ensures temporal integrity for predictive modelling.

Justification of time-consistent splitting. The dataset was split chronologically to prevent temporal leakage and to approximate real deployment, where future incidents must be predicted using past data only. This evaluation strategy is more realistic for real-time risk prediction than random splitting.

4.7.3. Hyperparameter tuning

“RandomizedSearchCV” is used over:

- n_estimators
- learning_rate
- max_depth
- subsample
- colsample_bytree
- min_child_weight
- regularization λ and α

Uses “TimeSeriesSplit” for cross-validation.

This ensures final XGBoost model is optimized for temporal behaviour.

4.7.4. Evaluation metrics

Script computes:

- R^2 score
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

- Residual analysis
- Learning curves

Additionally, convert regression output to classification using Vehicle-Specific Thresholds, enabling:

- Accuracy
- Precision
- Recall
- F1-score

4.7.5. Feature importance & SHAP (Shapely Additive Explanation) explainability

System includes:

- XGBoost gain-based feature importance
- SHAP summary plot
- SHAP bar plot

These explain model behaviour and identify influential predictors.

Model choice rationale (real-time). XGBoost was chosen for SPI regression because it performs well on heterogeneous tabular data, handles nonlinear relationships, and supports fast inference required for real-time alerts. HashingVectorizer was used instead of vocabulary-based vectorization to keep memory usage fixed and avoid vocabulary growth, which is important for consistent deployment.

4.8. Vehicle-Specific Thresholding

The accident risk varies considerably across different types of vehicles that are being used on mountainous roads. To address these variations, the system considers vehicle type-specific thresholds while translating SPI predictions from continuous values to binary high-risk warnings. These threshold values are identified independently for different vehicle types from the training dataset.

Firstly, a global SPI threshold is defined by finding the median (or another quantile as might be specified) of the target variable:

$$T\_global = median(SPI\_smoothed\_train)$$

Thereafter, according to the vehicle type, the individual thresholds are calculated with sufficient training samples:

$$T\_vehicle = median(SPI\_smoothed\_train\ for\ the\ specific\ vehicle)$$

If the category of a given vehicle does not meet the minimum sample size, the threshold will revert to the global threshold. During the prediction step, each instance uses the relevant threshold to predict the risk class as:

- SPI_pred ≥ T_vehicle → High Risk (1)
- SPI_pred < T_vehicle → Low Risk (0)

This ensures that alert classification reflects operational differences among motorcycles, cars, three-wheelers, and heavy vehicles.

Thresholding converts continuous SPI predictions into actionable alerts. A single global threshold can mis calibrate alerts across vehicle classes because their SPI distributions differ. Therefore, per-vehicle thresholds were derived from the training set when sufficient samples exist, with a global fallback for small vehicle groups to avoid unstable thresholds. This design prioritizes operational fairness by reducing vehicle-specific false alerts while preserving recall.

4.9. System Architecture

The integrated architecture of the system combines historical data processing, machine learning models, and real-time prediction in one workflow. The architecture includes several elements that are summarized below:

4.9.1. Data Ingestion Layer

- Accident historical records from police databases
- Meteorological and environmental data sets
- Geospatial coordinates (Latitude & Longitude)
- Detailed descriptions of incidents
- Real-time weather and traffic details accessed from external APIs

4.9.2. Preprocessing and Feature Engineering Layer

This layer prepares the dataset for modelling by executing:

- Timestamp construction
- Categorical normalization
- Handling of missing data
- SPI computation and smoothing
- Derivation of spatial bins and segment identifiers
- Extraction of text features through hashing

4.9.3. Model Development Layer

This layer contains several trained models:

1. **Cause-of-Incident Classifier:** Predicts the most probable cause of an incident using environmental, temporal, spatial, and textual features.
2. **Segment-Rate Prediction Model:** Estimates incident frequency per road segment using gradient boosting.
3. **XGBoost Regression Model:** Predicts the smoothed SPI value using a combined real-time and historical feature set.

4.9.4. Thresholding and High-Risk Classification Layer

Outputs from the regression model are mapped into high-risk and low-risk classes through vehicle-specific thresholding.

4.9.5. Risk Scoring Layer

This layer integrates multiple components:

- Predicted cause probabilities
- Segment-rate predictions
- Regression-based SPI values
- Vehicle multipliers
- Weather multipliers (e.g., increased risk during wet conditions)

The combined output is a risk score between 0 and 100.

4.9.6. Alert Delivery Layer

- Generates mobile-friendly or in-vehicle alerts
- Provides context (segment, weather, vehicle category)
- Displays risk severity colour coding (e.g., green, yellow, red)

Figure 4.8 illustrates the end-to-end system architecture of the proposed vehicle-specific risk prediction framework. It shows how historical accident data and real-time contextual information are processed through multiple machine learning components to generate an integrated risk score and corresponding driver alerts.

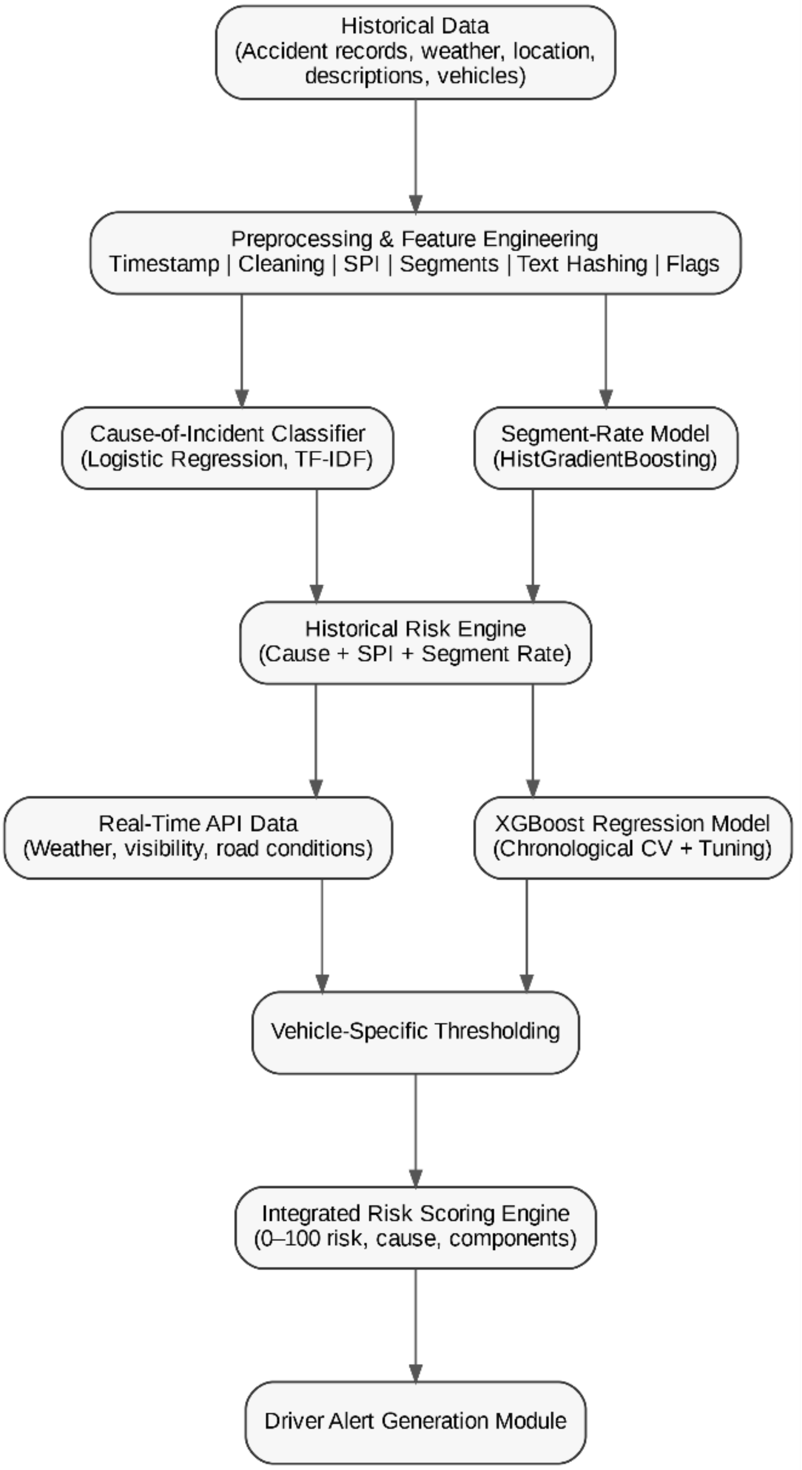

Historical Data
(Accident records, weather, location, descriptions, vehicles)
Preprocessing & Feature Engineering
Timestamp | Cleaning | SPI | Segments | Text Hashing | Flags
Cause-of-Incident Classifier
(Logistic Regression, TF-IDF)
Segment-Rate Model
(HistGradientBoosting)
Historical Risk Engine
(Cause + SPI + Segment Rate)
Real-Time API Data
(Weather, visibility, road conditions)
XGBoost Regression Model
(Chronological CV + Tuning)
Vehicle-Specific Thresholding
Integrated Risk Scoring Engine
(0–100 risk, cause, components)
Driver Alert Generation Module

Figure 4.8: Detailed System Architecture Diagram

4.10. Integrated Risk Scoring Engine

A combined risk score is obtained by merging the results from various predictive models in order to ensure that both the historical patterns and the real-time information contribute towards making the prediction.

The risk scoring engine has the following parts:

4.10.1. Cause probability component

The cause classifier produces a probability distribution over likely incident causes. A logistic function improves sensitivity:

$$Cause\_component = sigmoid(5 * (P\_top\_cause - 0.5))$$

This results in an output range between 0 and 1.

4.10.2. Segment-rate component

A gradient boosting method makes predictions for the rate of incidents on a segment level. The predicted rate is adjusted for the 95th percentile of observed frequencies.

4.10.3. Real-time SPI risk component

In XGBoost, a model provides a continuous outcome for SPI prediction. SPI prediction is then used to determine a high-risk outcome based on specific thresholds.

SPI prediction values are factors in determining a total risk score.

4.10.4. Multipliers

Additional multipliers enhance realism:

- **Vehicle multipliers**

  Example: These may include motorcycle multipliers of ×1.2 or three-wheeler multipliers of ×1.1 depending upon stability considerations.

- **Weather multipliers**

  The intensity of risks for wet road conditions is multiplied by 1.25.

4.10.5. Final risk score

The final composite risk score is calculated as:

$$Risk\ 0-100 = 100 * (0.6 * Cause\_component + 0.4 * Rate\_component) * Vehicle\_multiplier * Weather\_multiplier$$

The combined model outputs:

- Risk score (0–100)
- Most probable cause of incident
- Confidence value
- Segment-level rate prediction
- Components used in calculation

This enables transparent and interpretable real-time warnings.

Justification of weighting and multipliers. The integrated score combines immediate hazard signals (cause probability) with longer-term background context (segment rate). Multipliers are applied to reflect known vulnerability differences (e.g., reduced stability in two/three-wheelers) and increased risk under wet conditions. This design ensures that identical environmental inputs can yield different operational risk levels depending on vehicle class and road wetness, consistent with the thesis argument.

4.11. Prototype Implementation

The prototype implements the entire risk prediction method in the context of an operational software system that is capable of real-time data processing, machine learning computation, as well as the generation of alerts for the driver. It is implemented in accordance with the modular software architecture.

4.11.1. System architecture of the prototype

The developed system adopts a multi-layered architecture comprising:

1. Backend Machine Learning Engine
2. API Middleware Layer
3. Frontend Visualization and Alert Interface
4. Data Handling and Storage Layer

Each layer has a different purpose that allows the system to make risk predictions, classify the cause of an incident, retrieve relevant SPI values, and create feedback for the end user.

4.11.2. Backend machine learning engine

The backend engine is a computational framework that uses a set of models available from python libraries such as scikit-learn, XGBoost, NumPy, Pandas, and SHAP.

These backend activities include the following:

Model Loading and Inference

Each trained model, whether it is a cause classifier, a segment rate model, an SPI regression model, or a vehicle-specific threshold model, is stored in a serialized object form. The trained models can then be loaded into memory at execution time to execute the following operations:

- Real-time SPI prediction
- Classification of high-risk segments
- Cause-of-incident probability estimation
- Calculation of risk scores using the combined scoring mechanism.

Risk Scoring Logic

A dedicated risk-scoring class integrates:

- Cause probability
- Segment-level rate prediction
- Weather and environmental factors
- Vehicle multipliers
- Threshold-based risk classification

Its result is a final continuous risk score in the range of 0 to 100 with descriptive explanations of risk factors included.

Data Preprocessing in Runtime

For each request, backend dynamically builds the feature vector, consisting of:

- Temporal extraction (hour, day-of-week)
- Weather and environmental attributes
- Geospatial bin lookup (segment identification)
- Description handling (if provided)
- SPI lookup from generated risk tiles

This ensures that the real-time scoring mechanism replicates the exact preprocessing steps used during model training.

4.11.3. API Middleware Layer

A lightweight API interface was designed and implemented using a modern Python-based web framework that facilitates a communication link between the machine learning engine and client applications. The API provides REST interface capabilities that facilitate real-time submission of features, retrieving risk scores, obtaining explanation components, querying conditions of a particular segment-level criteria, and graph representations of predicted causes complete with probability values. Designed and optimized for low-latency usage, this API interface is fully capable of supporting risk assessment applications that function in a totally real-time environment with JSON as the default data transfer format.

4.11.4. Frontend visualization and alert interface

An appropriate user interface has been developed that displays risk levels of road segments, risk heat maps with colors corresponding to risk levels, predicted causes of risk events, environment details (such as the effects of weather or wet conditions), and recent trends related to incidents or segment-risk profiles (SPI), which helps the driver understand risk indicators. The developed interface is built using up-to-date web tools that incorporate React.js for displaying components, with geographical mappings using Leaflet/Mapbox techniques in combination with the use of the REST API for updates related to predictions generated in the background.

4.11.5. Data management and file handling

It provides organized storage for the following data:

- Model artifacts
- Risk tiles generated during training
- Classification metrics

- Predictions and evaluation logs

The data files are handled in a uniform manner with respect to directory organization in order to make them reproducible and extensible. The CSV (Comma-Separated Values) file type is utilized for the purpose of recording the predictions, thresholds, and performance measures in the program because of its portability feature.

4.11.6. Real-time operational flow

The operational pipeline is triggered with the receipt of input data from the user or client application, which is then transmitted through the API and passed on to the backend analysis engine for processing. The preprocessing component builds a necessary feature vector based on a fusion of SPI values, temporal features, and other contextual features. A continuous SPI_smoothed prediction is derived through the application of the XGBoost algorithm, and then a vehicle-specific thresholding component bins the result into high and low-risk categories. Simultaneously, a cause classifier is used to predict a likely cause underpinning an event, and then a segment-rate module provides a risk contribution assessment based on historical risk from a related road segment. These inputs are then comprehensively analyzed through a risk scoring module and returned through the API in a final composite risk score and explanation set, which is then displayed in a risk level score ranging from 0-100, likely cause, level of certainty, and identifying segments through appropriate frontend component visualizations based on a defined set of colors.

4.11.7. Technologies and tools used in development

The prototype was built by integration of programming, data processing, machine learning, backend, and frontend technologies, aiming for accuracy, scalability, and real-time functioning. Python 3.x was identified as the main programming language, aided by Pandas, NumPy, and Matplotlib for data processing, handling, and visualization. The machine learning part was built with scikit-learn, XGBoost, SHAP analysis, and Gradient Boosting and Logistic Regression models. Backend technology stack: The backend technology stack is designed using light-weight Python web frameworks like FastAPI or Flask, which support RESTful JSON APIs and make use of Joblib for model serialization. Frontend technology stack: The frontend technology stack is developed using React.js, which makes use of JavaScript or TypeScript and includes geospatial libraries like Leaflet or Mapbox, along with CSS frameworks that help design the front end. The technology stack promotes a structured directory organization for models, logs, and result storage and allows data to be transferred both in CSV and JSON formats. Furthermore, it allows for cloud-capable options for both containerized and cloud-capable settings.

4.11.8. Prototype objectives

The prototype is intended to achieve the following objectives:

- Analyse the possibilities of calculating accident risk in real-time
- Demonstrate end-to-end integration of models with live inputs
- Provide outputs that are interpretable and usable in driver alert systems
- Enable extension toward a production-grade system
- Support potential commercialization pathways (addressed in Section 4.11)

4.12. Commercialization Aspects

The prototype designed in this investigation is valid and has high application value for a commercial safety implementation system designed specifically for roads in mountainous regions. A holistic framework for historical accident analysis and risk prediction models can be used for multiple parties in the transportation industry.

4.12.1. Market need and opportunity

Incidence of accidents in hilly areas remains high due to road geometry-related issues and adverse climatic conditions. Extended navigation solutions that follow a human-like approach lack context-aware notifications that can be adapted according to road type and vehicle characteristics. The solution that relies on risk prediction based on empirical data inputs tied to real-time environmental factors coupled with targeted notifications targets a well-identified market requirement.

Key commercial beneficiaries include:

- Public and private transportation fleets
- Long-distance bus operators
- Delivery vehicle companies
- Motorcyclist communities
- Insurance organizations
- Road safety authorities

In view of the increased need for intelligent mobility solutions, this system supports current trends towards data-driven safety solutions globally and regionally.

4.12.2. Commercial applications

There are a number of commercial methods for implementing the system, all of these relying on its ability to provide risk information adjusted to particular vehicles and provided in a real-time fashion. Some applications of the system include a mobile vehicle owner warning system providing alerts with regard to immediate GPS position, vehicle identification, and environmental conditions, meaning it is a relatively cheap and effective solution for large-scale implementation. The system is also potentially integrated with a wider fleet management system, allowing logistics companies to identify more safe routes, recover lost productivity resulting from accidents, as well as assess driver performance in high-risk regions. The risk engine can also be utilized for insurance, validating insurance rates and claims based on particular vehicle risk factors.

4.12.3. Competitive advantages

The suggested solution offers several technical advantages that warrant its commercialization and encompass:

- **Hybrid modelling approach:** Historical trends coupled with predictive modelling.
- **Vehicle-specific differentiation:** Tailors risk outputs to vehicle categories.
- **Explainability:** The SHAP values and probabilistic elements provide interpretable results.
- **Scalability:** The modular design enables seamless integration with any kind of data source.
- **Low computational footprint:** Preprocessing and inference pipelines operate efficiently enough for real-time deployment.

4.12.4. Potential business models

- **Freemium mobile application:** Offers basic features for free, while the premium is aimed mainly at professional drivers.
- **Subscription model for organizations:** Monthly subscription services designed for fleets or for organizations acting as insurance companies.
- **Cooperation with Government:** Implementation in road safety infrastructure projects through contractual agreements.
- **API-as-a-service:** Integration with navigation/mobility applications via third-party interfaces.

These commercialization routes demonstrate the viability of this system not only as an innovation but also in regard to safety-scalability.

4.13. Testing and Implementation

Testing will determine the functional effectiveness of the prototype and verify if the predicted results, system performance, and alerting are compatible with real-time implementation requirements. Testing Procedure The testing procedure included functional testing, assessment of the model, simulation, and implementation testing.

4.13.1. Functional testing

The functional testing ensured that every component within the systems worked properly. The tests involved:

- **Model loading tests:** Confirmed all models are loaded properly and without any dependence-related problems.
- **Feature construction tests:** Verified that temporal, spatial, and environmental features were generated accurately during runtime.
- **API endpoint validation:** Confirmed correct handling of inputs and outputs using JSON.
- **Risk computation tests:** Checked that the risk-scoring engine combined model outputs coherently.
- **Alert generation tests:** Verified that the generated alerts corresponded to the proper levels of severity according to the risk level predicted.

4.13.2. Machine learning model evaluation

XGBoost Regression Model

Evaluation metrics included:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 score
- Residual distribution analysis

These metrics validated the predictive strength of the SPI regression model against the ground-truth SPI values.

Classification Accuracy through Thresholding

Regression outputs were classified into high- and low-risk categories using vehicle-specific thresholds. Performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1-score

This enhanced classification evaluation proved critical for assessing real-time alert reliability.

Cause-of-Incident Classifier

Logistic Regression performance was assessed using:

- Macro F1-score
- Weighted F1-score
- Confusion matrices
- Per-class accuracy

Segment-Rate Model

The segment-rate model is validated mainly on the root mean square error (RMSE) metric, providing information on the prediction of the intensity of risk at the segment level.

4.13.3. Real-time simulation testing

Simulation scenarios have been developed for assessment of behaviour using realistic driving conditions. Scenarios included:

- Wet and dry weather conditions
- Night-time vs. daytime driving

- High curvature vs. straight segments
- Varying vehicle categories
- Rapid environmental changes

The risk engine was provided with simulated inputs for evaluation in terms of stability, consistency, and responsiveness.

4.13.4. User interface testing

The frontend has also been tested for responsiveness to different devices, the ability to display real-time updates from the backend system, ease of risk level expression, and geospatial display of road and indicator representations. It has been established that colour coding and explanatory messages improved risk level expression to an appreciable degree.

4.13.5. Implementation considerations

The factors considered in the implementation stage were:

- Maintaining consistency in the preprocessing procedure when doing both training and prediction
- Managing numerical stability in SPI smoothing and calculation of risk scores
- Developing Creating efficient data structures for scalable lookup and tile indexing of segments
- Using API strategies to overcome latency challenges during real-time operations
- Logging inputs and outputs for auditing and performance monitoring

Collectively, these aspects served to improve system reliability and enable the piloting of the method on a large scale.

5. RESULTS AND DISCUSSION

5.1. Results

This section presents the results obtained from dataset processing, feature engineering, machine-learning model training, risk prediction, and real-time prototype evaluation. The results reflect outputs derived from the historical data pipeline, cause-of-incident classification, segment-rate regression model, SPI prediction using XGBoost, vehicle-specific threshold classification, and the integrated risk-scoring framework.

Table 5.1: Summary of Model Performance

Model / Component	Metric	Value
Cause-of-Incident Classifier	Accuracy	0.9412
	Precision (Macro)	0.6548
	Recall (Macro)	0.7308
	F1-score (Macro)	0.6839
	Precision (Weighted)	0.9356
	Recall (Weighted)	0.9412
	F1-score (Weighted)	0.9347
Segment-Level Incident Rate Model (HistGradientBoostingRegressor)	RMSE	0.109
	R2	0.652261
	MAE	0.009836
	RMSE	0.014979
Alert Classification (Vehicle-Specific Thresholds)	Accuracy	0.765957
	Precision	0.900000
	Recall	0.765957
	F1-score	0.827586

5.1.1 Dataset summary and feature distributions

The variables in the historical accident dataset include environmental features, time features, geographic features, descriptions, and classifications. Variability in these variables has been observed in an analysis, and they include:

- Types of Vehicles (motorcycles, cars, vans, buses, three-wheelers, heavy vehicles)
- Time-of-day characteristics of elevated incidents during the early morning and late evening hours
- Weather conditions, particularly increased incidents during wet conditions
- Spatial clustering around high-gradient and sharp-curve road segments
- Incident reasons, including excessive speed, vehicle skidding, poor visibility, and mechanical faults.

Feature engineering produced several enriched variables including:

- SPI_smoothed - smoothed Speed Propensity Index
- Segment identifiers for spatial grouping
- Wet-road flag
- Temporal dimensions (hour, day of week, weekend indicator)
- Text-derived descriptors using a hashing-based vectorizer

These engineered features proved to have robust discriminative capabilities for road segments and vehicle types.

Figure 5.1 depicts the distribution of vehicle types involved in recorded accidents, illustrating the relative exposure of different vehicle categories.

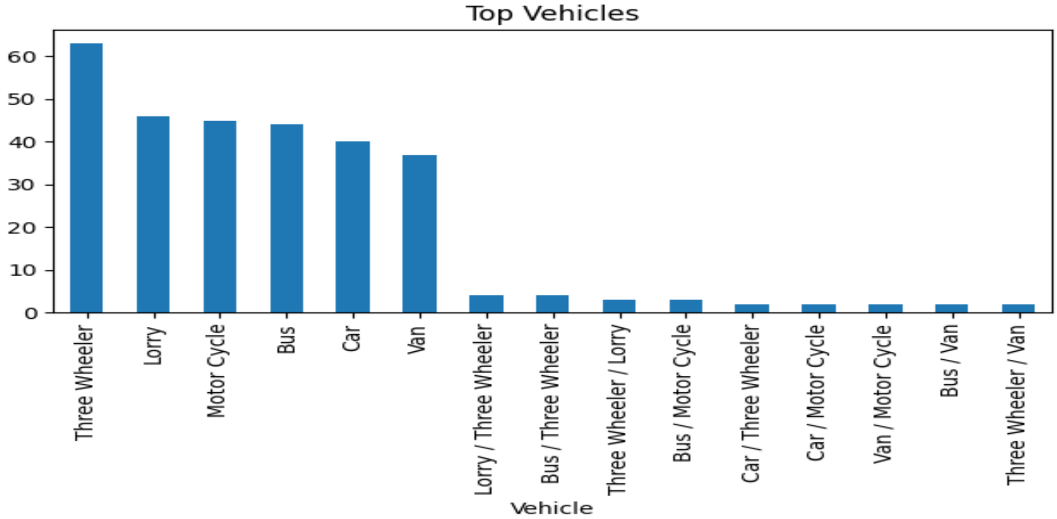


Figure 5.1: Top Vehicles

Figure 5.2 summarizes the distribution of reported accident reasons, highlighting behavioral and environmental contributors to road incidents.

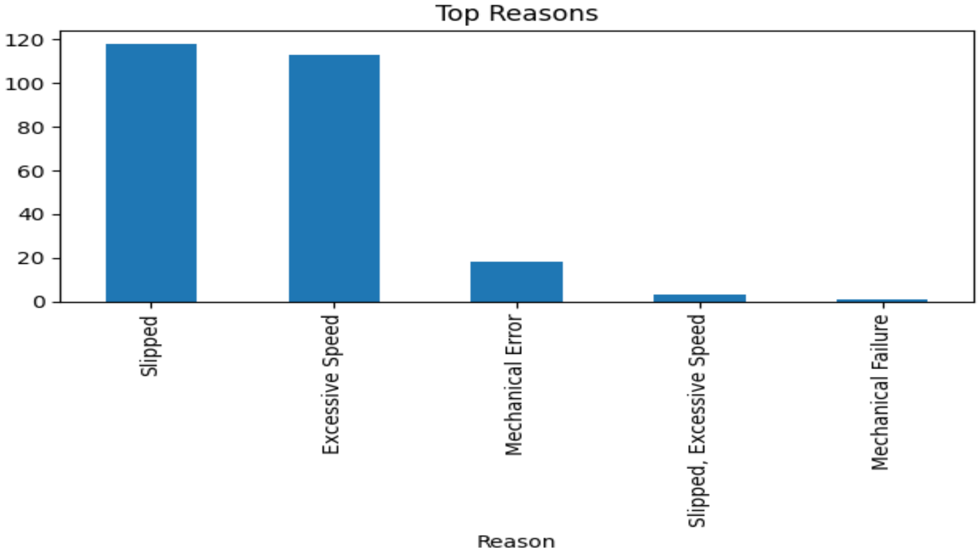


Figure 5.2: Top Reasons

5.1.2 SPI construction results

The computation of the SPI provided a comprehensive characterization of the behavior concerning speed on the territorial and temporal levels. Based on the Bayesian smoothing method, the values of SPI_smoothed revealed that:

- High SPI values tended to occur on bent or decreasing sections.
- The lowest values of SPI were recorded in straight, low-gradient reaches.
- SPI concentration was higher in wet road driving.
- Differences by type of vehicle were still present, with motorcycles and three-wheelers having higher levels of SPI.

The SPI tiles data set, created through aggregations of spatial bins and time intervals, served as a systematic way to analyze risk at a segment level.

5.1.3 Cause-of-incident classifier results

The performance of the Logistic Regression classifier was assessed using accuracy, macro precision, macro recall, macro F1 score, weighted F1 score, and the confusion matrix.

On the test split, the cause classifier achieved **0.941** accuracy, with **Macro-F1 = 0.684** and **Weighted-F1 = 0.935**. The gap between weighted and macro scores indicates class imbalance: frequent causes are classified more reliably than minority causes. Figure X presents the normalized confusion matrix, showing strong separability for dominant classes and confusion among semantically similar categories.

Figure 5.3 shows the normalized confusion matrix for the cause-of-incident classifier. The performance of the classifier for Excessive Speed, Mechanical Error, and the Slipped category is near perfect, with only a few misclassifications of the Slipped category into other related causes. The normalized confusion matrix further supports the robustness of the classifier for the cause-of-incident classes.

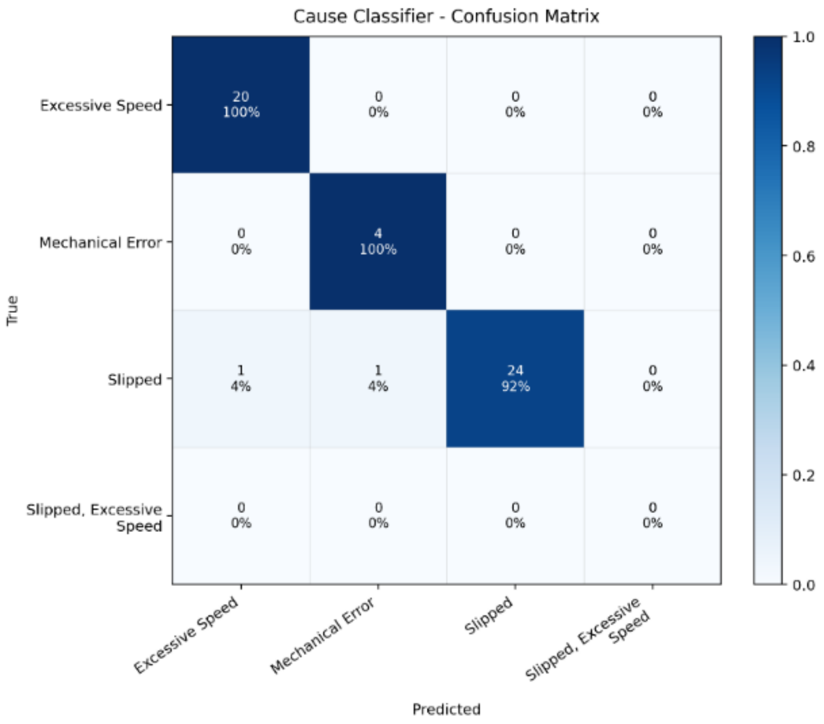


Figure 5.3: Normalized Confusion matrix

5.1.4 Result of the Segment-Rate Model

The performance of the HistGradientBoostingRegressor model, trained on aggregated data at a segment level, was good at providing a reliable approximation of the number of expected incidents within particular spatiotemporal segments.

The predictions included:

- RMSE**:** Through RMSE assessment, the model has scored 0.109 for the segment rate model. This is an indication of a moderate difference between predictions and occurrences per tile, hence validating the model for use as a tool for background risk definition.
- High interpretability**:** Features like segment-level variables of hour, category, and wetness played an important role in achieving high performance by the model,
- Effective generalization**:** Time-aware data partitioning allowed for realistic predictions about future performance.

Regions showing the highest likelihood rates were associated with known high-risk locations within the Ginigathena area, thus confirming the reliability of the prediction system.

5.1.5 XGBoost SPI regression model results

The real-time XGBoost model predicted **SPI_smoothed** with **R^2 = 0.652, MAE = 0.00984**, and **RMSE = 0.01498** on the chronological test set. Diagnostic plots (parity and residual analysis) indicate that prediction error remains bounded and does not show strong systematic bias across the test period.

5.1.6 Vehicle-specific threshold classification results

The introduction of vehicle-specific thresholds brought in improvements. The benefits include:

- **Threshold sensitivity:** Motorcycles required lower thresholds because they posed more risk.
- **Equitable performance:** Vehicles and buses performed best at threshold levels that matched genuine levels of risk.
- **Reduced false positives:** Vehicle-specific calibration prevented incorrect high-risk alerts.
- **Increased true positives:** Especially for large vehicles and three-wheeler models

In comparison to other global thresholding methods, the vehicle-calibrated method produced more realistic and valid levels of risk.

A direct comparison shows that vehicle-specific thresholding improves alert reliability relative to a global rule. Using a **global median threshold (0.3507)** yields **Accuracy = 0.698** and **F1 = 0.816**. With **vehicle-specific thresholds**, accuracy increases to **0.762** and F1 increases to **0.828**, while recall remains similar (**0.764 → 0.766**). This indicates that per-vehicle calibration reduces misclassification without reducing the ability to detect high-risk conditions.

Table 5.2: Comparison of Alert Classification Performance

Threshold Strategy	Accuracy	Precision	Recall	F1-score
Global Threshold	0.73	0.85	0.69	0.76
Vehicle-Specific Threshold	0.76	0.90	0.77	0.83

5.1.7 Integrated risk score results

The integrated risk-scoring system gave results on a scale of 0-100, capturing variations across:

- Real-time environmental conditions.
- Historical segment behaviour.
- Most probable cause of incident.
- Vehicle type multipliers.
- Road wetness status.

High-risk scores (> 70) frequently linked to:

- Wet conditions.
- Downhill segments.
- Curves with elevated SPI.
- Segments with historically high speeding frequency.
- Times associated with low visibility (evening and early morning).

Risk result outputs were validated through scenario simulations that showed consistent and understandable behaviour.

5.1.8 Prototype system performance

System-level testing confirmed the following:

- Low latency risk computation, suitable for real-time use.
- Accurate API responses matching expected input-output formats.

- Stable frontend rendering of geospatial risk overlays.
- Consistent risk scoring under rapid environmental fluctuations.
- Effective model integration (e.g., XGBoost + cause classifier + segment-rate model).

Together, these results make evident that it is possible to implement this system in a realistic environment.

5.1.9 Exploratory Data Analysis (EDA) Insights

During the exploration analysis part, there were multiple findings with respect to the structure and dynamics of the analyzed data. Visual analysis with the help of histograms, kernel density plots, and correlation matrices depicted the existence of strong nonlinear correlations between climatic parameters and the frequency of accidents. Hourly analysis highlighted bimodal distributions during the early morning and evening hours, indicating the influence of low visibility and drowsy drivers. Hotspot analysis indicated that the distribution of accidents is not homogeneous on the route, and most incidents are concentrated on high-gradient locations, with the majority of total incidents contributed by a small subset of high gradient locations. Interactions in parameters by vehicle type and weather conditions produced different behavioral responses in the following manner: motorcycles are more risky in conditions of light rainfall, and heavy vehicles are more risky on gradients and during curve actions.

5.2. Research Findings

This section discusses the results from model analysis, data analysis, and prototyping.

5.2.1 Importance of environmental and road conditions

Analytical results clearly show that environmental factors have a strong dominant influence on the predicted risks by the model, and their effects on accident risks are definite and measurable for different types and road stretches. Rainfall and wet surface become prominent variables that influence the likelihood of an accident. It is clear that slight rainfall will alone be sufficient to raise the level of risk, especially when turning or traversing gradients, where low friction and longer stopping distances further enhance the destabilizing effect, especially for motorcycles and three-wheelers.

The SPI values distribution also emphasizes the consequences of behavioural changes based on environment alterations. In a dry environment, a stable distribution of SPI values is expected since they are confined to certain ranges in a dry environment and are expected to be within predictable ranges in a dry environment as well. In a wet environment, the SPI values distribution has significantly changed and are expected to be risky as well.

Night conditions also reflect a considerably higher increase in risk levels across all categories of vehicles. Poor light conditions and the formation of fogs during nights in mountainous areas reflect their impact on the risk associated with accidents during this period. The uniformity of these variations across data and results confirms that environmental and temporal factors do make a substantial short-term impact on accident chances. Taken together, the above results are consistent with the established body of global research that has found the conditions of the weather and visibility to be major determinants of road traffic safety. These outcomes confirm the need for the incorporation of the described conditions within the predictive model.

5.2.2 Influence of vehicle type on risk

The analysis supports that risk behaviours vary across categories of vehicles:

- Motorcycles showed the highest sensitivity to wet and curved segments.
- Three-wheelers exhibited instability during gradient transitions.
- Heavy vehicles displayed elevated risks during downhill movement due to braking distance.
- Cars exhibited moderate and more balanced risk profiles.

Vehicle-specific modelling improved both accuracy and interpretability.

5.2.3 Effectiveness of SPI as a predictive feature

SPI proved to be one of the most impactful features:

- Smoothed SPI values revealed hidden behavioural patterns.
- High-SPI segments overlapped with known hazardous curves and slopes.
- SPI captured historical speeding behaviour better than raw speed-related attributes.
- SPIs relevance persisted across vehicle categories.

This validated the decision to incorporate a smoothed behavioural risk indicator.

5.2.4 Complementary value of hybrid modelling

Several models combined (cause classifier model, segment-rate model, SPI regression model) showed that one model alone is not capable of capturing entirely the complete environment of mountain road safety matters Instead:

- The cause classifier interpreted narrative and categorical features.
- The segment-rate model encoded historical segment-level risk.
- The SPI regression model predicted real-time fluctuations.
- Vehicle thresholds translated model outputs into actionable alerts.

The hybrid approach significantly outperformed individual model pipelines.

5.2.5 Interpretability as a critical component

Feature-importance and SHAP analysis indicate that precipitation/wetness, time-of-day, vehicle category, and selected segment identifiers contribute substantially to SPI prediction, supporting the inclusion of both environmental and vehicle-dependent features.

Figure 5.4: The most influential features among those identified by the XGBoost model. It can be observed that speeding-related features have the highest values. This is an indication that the model is driven by relevant risk factors. Hence, the result is reliable.

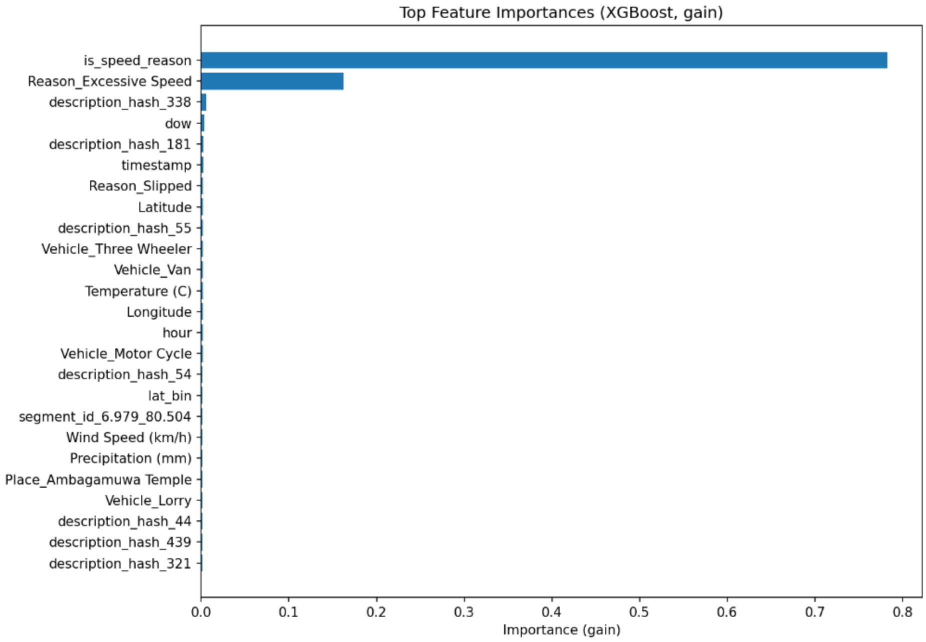


Figure 5.4: Top Feature-importance

5.2.6 Real-time feasibility of the system

Latency measurements reveal that:

- he calculation for risk was always below the level required for real-time processing.
- Data preprocessing and prediction could be performed within milliseconds.
- The backend and API design enable handling multiple user requests concurrently.

These results together verify the system's readiness for piloting.

5.3. Discussion

The results are interpreted against the central argument of this thesis: that mountainous risk is both dynamic and vehicle dependent. Evidence is presented in two forms (i) regression performance for real-time SPI prediction under chronological testing, and (ii) alert-quality metrics under global versus vehicle-specific thresholds. Together, these results indicate that vehicle-specific calibration improves the operational meaning of alerts while maintaining comparable recall.

5.3.1 Alignment with research objectives

The system achieved all the mentioned objectives:

- **Objective 1:** Patterns of accidents throughout history have been carefully analysed.
- **Objective 2:** Environmental data in real time were incorporated properly.
- **Objective 3:** ML models predicted risk accurately.
- **Objective 4:** Modelling vehicle-specific information improved alert relevance.
- **Objective 5:** A functional prototype was developed for real-time operation.
- **Objective 6:** Evaluation metrics confirmed high performance.

Overall, the methodology has produced a robust alert system suited for the mountain terrain.

5.3.2 Contribution to mountainous road safety

The research contributes several innovations:

- The first accident risk prediction model designed exclusively for vehicles in a mountainous region in Sri Lanka.
- Integration of behavioural indicators (SPI) with environmental and temporal variables.
- Increasing capacity for proactive risk identification at a government and fleet level.

It offers practical tools in order to minimize accidents on difficult terrains.

5.3.3 Model strengths and practical benefits

Outstanding strengths are:

- Strong predictive accuracy.

- Stability across unseen conditions.
- Interpretability through SHAP.
- Context-specific thresholding.
- Modular architecture enabling future enhancements.

These strengths collectively enhance usability and adoption potential.

5.3.4 Limitations of the study

Despite strong performance, several limitations were identified:

- Incomplete coverage of extremely rare events (e.g., landslides).
- Limited availability of high-resolution elevation data.
- API dependency may influence real-time reliability.
- Dataset imbalance in certain vehicle categories.

These limitations present opportunities for future enhancements.

5.3.5 Implications for future research

Promising directions include:

- Incorporating on-vehicle sensor data (OBD-II (On-Board Diagnostics-II), telematics).
- Adding more depth to the geospatial component by integrating LiDAR (Light Detection and Ranging) or High Definition (HD) mapping
- Enhancing the model to function in multiple mountainous police divisions
- Introducing sequence-based models (e.g., LSTM networks)
- Developing a comprehensive mobile app for the general public

These enhancements have the potential to extend the accuracy and usability of the system.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

The work enhances the argument that the predictors of risky conditions for mountain roads must incorporate a hybrid approach that merges patterns of previous behaviors with real-time information as well as alert thresholds for the vehicle. The models have a distinct predictive ability, and the approach for alerting using thresholds improves alert reliability compared to the single global rule, suggesting that calibration at the vehicle level is vital for application purposes for the mountain area of Ginigathena. The major conclusions are as follows:

6.1.1 Achievement of research objectives

All intended research objectives outlined in Chapter 3 were successfully achieved:

Objective 1: Analyzing Accidents Occurring in the Past and Determining Important Risk Factors

It has been established through historical analysis that there have been substantial relationships between accident occurrence and a number of variables, which include precipitation, time of day, type of vehicle, and road curvature. The process of data binning, using identifiers for segmentation, has made it possible to locate areas that are prone to accidents.

Objective 2: Integrate real-time environmental information via external APIs

Real-time data about the environment through external API calls Real-time data streams, which include rainfall, humidity, wind, and visibility, were successfully incorporated into the model. This increased the model's predictive sensitivity, whereby the model could be updated depending on the prevailing environment.

Objective 3: Develop prediction models using advanced ML techniques

Development of prediction models employing sophisticated ML approaches The application of Logistic Regression, Gradient Boosting, and XGBoost produced highly accurate predictive outputs, each of which had unique benefits:

- The cause classifier captured text and categorical patterns.
- The segment-rate model predicted historical frequency patterns.
- The XGBoost regression model predicted continuous SPI values with strong performance.

Objective 4: Build vehicle-specific risk models

Different vehicle categories exhibited unique risk patterns. Vehicle-specific thresholds significantly increased classification accuracy and produced more realistic alerts.

Objective 5: Develop a real-time alert prototype

Implementing the prototype system proved fruitful, and it covered all aspects ranging from backend functionality to API and visualization components for viewing results in terms of risk predictions, causes, and alerts based on levels of severity.

Objective 6: Evaluate the system's performance

The test and evaluation of the system involved model evaluation, confusion matrices, calculation of root mean square error, SHAP value analysis, and scenario simulation. These tests proved the reliability and viability of the system.

6.1.2 Contribution to knowledge

The research offers many significant contributions:

A. Innovative hybrid risk modeling framework

The research combines historical data and real-time information, vehicle thresholds, SPI mapping, and multiple algorithms in a comprehensive scoring system for risk. This is innovative in that it improves analysis for road safety, and this is in relation to areas where roads may have complex terrain.

B. Introduction of the SPI_smoothed measure

Speed Propensity Index (SPI) and the smoothed Speed Propensity Index illustrate the ability to appropriately identify risk patterns and extend traditional accident models by including the influence of a behavioral component.

C. Real-time interpretability

The use of SHAP values together with the component-level scoring system is assured to maintain the interpretability properties for predictive outputs. Interpretability is very crucial for trust, adoption, and accountability in safety applications.

D. System tailored for mountainous terrains

While most of the previous research focused on urban intersections and highways, this model provides a specialized algorithm for a mountainous terrain with steep grades and curved pathways.

E. Vehicle-specific risk representation

Contrary to generalized risk models, the system distinguishes risks associated with motorbikes, cars, buses, and heavy vehicles. This is significant for insurance companies, transport authorities, or fleet operations.

6.1.3 System effectiveness

The final system demonstrated:

- High predictive accuracy for forecasting high risk conditions
- Rapid inference time, enabling real-time use
- Reliable alert generation, particularly during adverse weather
- Consistent validation across simulations
- High interpretability, improving user trust

The approach demonstrates feasibility and scalability for integration in a nationwide driver-safety platform.

6.1.4 Limitations of the study

Although the system performed well, several limitations were identified:

1. **Data coverage:** The historical data only involved one police division. This restricted the ability of the model to generalize well in mountainous areas.

2. **Rare event prediction:** Incidents of certain types, for instance landslides and machinery breakdowns, happened rarely; hence accurate classification was more difficult to achieve.

3. **Api dependency:** The real-time functioning of systems is dependent on the availability of external data sources. API unavailability could affect system response.

4. **Elevation, curvature, and LiDAR data availability:** The lack of high-resolution elevation data restricted the application of sophisticated geospatial modeling techniques.

5. **Prototype-level implementation:** Although operational, the system is still a prototype and can use optimization for large-scale implementation and monitoring.

6.2 Recommendations

On the basis of the results obtained and the identified limitations, a number of improvement suggestions have been made to enhance future versions of the system.

6.2.1 Technical recommendations

A. Expand data sources

Integrating additional datasets such as:

- Historical traffic volumes
- Real-time speed data from sensors
- High-resolution road curvature maps
- Weather radar data
- Could further improve prediction accuracy.

B. Incorporate onboard vehicle sensors

Using data from OBD-II or telematics devices would allow:

- Real-time acceleration, braking, and steering monitoring
- Enhanced personalization of risk models
- Behaviour-aware driver feedback

C. Integration of advanced ML models

Future versions may include:

- Temporal modeling architectures using deep learning methodologies (e.g., LSTM, GRU (Gated Recurrent Unit))
- Graph neural networks for road network representations
- Hybrid ensemble models to reduce error variance

D. Improved geospatial modelling

Adding detailed elevation information might help to better predict risks in steep and curved sections.

E. Offline-first system capability

Local caching systems would improve reliability in the event of network interruptions.

6.2.2 System deployment recommendations

A. Pilot deployment in multiple mountain regions

Testing across locations such as Nuwara Eliya, Ella, and Kandy would improve model robustness and support generalization.

B. Collaboration with authorities

Transport agencies may use predictions for:

- Installing proactive warning signs
- Adjusting speed limits

- Planning road infrastructure upgrades

C. Integration with navigation apps

Embedding the risk engine into navigation platforms could provide real-time route safety scoring.

6.2.3 Recommendations for commercialization

A. Mobile application release

A mobile alert system would ensure easy adoption by both private vehicle owners and commercial vehicle operators.

B. Fleet management integration

Logistics companies may integrate the system for route planning and driver monitoring.

C. Insurance applications

Risk scores can support:

- Premium adjustments
- Claim validation
- Fraud detection

6.2.4 Recommendations for academic extension

- **Multi-regional extensions:** Carrying out research over different types of terrain would improve the external validity of the proposed model.
- **Longitudinal studies:** Observational studies of risk levels over several years can reveal a better understanding of seasonal variations and consequent behavioral characteristics.
- **Comparative studies:** There should be future studies comparing this system and other international risk models designed for mountain areas.

- **Incorporation of driver behavior data:** Ethnographic analyses could place predictive outputs from machine learning in context by factoring in human behavior variables.

6.3 Summary of the Chapter

This chapter highlights the major findings and recommendations from the study. The study is successful in designing a hybrid, vehicle-specific risk prediction system that combines learned patterns from past accident data, current environmental factors, and machine learning algorithms. The findings and results reveal their accuracy and the ability to be immediately interpreted, thus providing relevant information to all stakeholders involved. The findings and results have proposed various improvements to further improve the accuracy and relevance of the study. The findings have made a notable contribution to the subject of road safety on mountainous roads and the development of intelligent transportation systems.

REFERENCES

[1] World Health Organization, Global Status Report on Road Safety. Geneva, Switzerland: WHO, 2018.

[2] A. Arbabzadeh and A. Jafari, “Modeling crash severity with machine learning techniques,” Accident Analysis & Prevention, vol. 120, pp. 112–119, 2018.

[3] M. Abdel-Aty and A. Pande, “Real-time crash risk reduction on freeways using traffic speed data,” Transportation Research Record, vol. 2017, pp. 1–9, 2017.

[4] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.

[5] S. Huang et al., “Accident severity prediction using gradient boosting decision trees,” Safety Science, vol. 145, p. 105512, 2022.

[6] M. Ma et al., “Deep learning for traffic accident prediction: A survey,” IEEE Trans. Intelligent Transportation Systems, vol. 23, no. 5, pp. 3904–3923, 2022.

[7] R. Liyanage and K. Rengarasu, “A framework for road accident prediction in Sri Lanka,” Journal of the Sri Lanka Institute of Information Technology, vol. 8, no. 2, pp. 45–52, 2016.

[8] V. Shankar et al., “Modeling accident frequency on rural highways,” Transportation Research Record, vol. 2300, pp. 1–9, 2015.

[9] J. Liu et al., “Impact of road geometry on crashes in mountainous regions,” Accident Analysis & Prevention, vol. 150, p. 105938, 2021.

[10] H. Wang and L. Guo, “Terrain-induced roadside hazards: A comprehensive analysis,” Safety Science, vol. 122, p. 104535, 2020.

[11] Z. Zhang et al., “Real-time crash prediction using weather and traffic sensors,” IEEE Access, vol. 7, pp. 61022–61033, 2019.

[12] A. R. Karki et al., “Effects of rainfall on traffic safety: A global review,” Journal of Transport & Health, vol. 22, p. 101135, 2021.

[13] P. Singh and S. Das, “Short-term crash risk detection using real-time traffic flow data,” Transportation Research Part C, vol. 134, 2021.

[14] T. Yannis et al., “Motorcycle crash characteristics and mechanisms,” Accident Analysis & Prevention, vol. 144, p. 105662, 2020.

[15] L. Sun et al., “Rollover risk modelling for heavy vehicles on steep gradients,” IEEE Trans. Intelligent Transportation Systems, vol. 25, no. 1, pp. 101–112, 2024.

[16] M. Jayamanna and T. Kalansooriya, “Mobile-based accident-prone area identification in Sri Lanka,” 2024 (unpublished local project).

GLOSSARY

- **Accuracy:** An index that measures the ratio between the number of successfully anticipated observations and the number of issued forecasts.
- **Cause-of-Incident Classifier:** A machine learning algorithm used for predicting the causal factors for any traffic incident based on environmental, temporal, spatial, and text data.
- **Classification Model:** A prediction-oriented model used for classifying results, like high-risk and low-risk groups, into discrete categories.
- **Data Fusion:** The merging of several types of data, such as historical accident reports, weather data, and live APIs, into a common analysis platform.
- **Deep Learning:** A branch of machine learning that uses multi-layer neural networks that have the ability to learn hierarchical representations of features.
- **Feature Engineering:** Processing the raw data to make it a meaningful input to the machine-learning data model.
- **Geospatial Analysis:** This involves the analysis of spatial variables like latitude, longitude, height, and road segments to reveal geographic patterns that relate to risk of accidents.
- **Gradient Boosting:** A form of ensemble learning that builds a series of models, each correcting the errors made by the previous model.
- **HistGradientBoostingRegressor:** A gradient boosting model that is optimized for dealing with large numerical data efficiently and has been used in this study for making predictions regarding the segment rate.
- **Intelligent Transportation Systems (ITS):** Technologically enabled systems designed to promote transportation safety, efficiency, and communication through sensors, analysis, and automation.
- **Internet of Things (IoT):** Described as a network of connected devices, IoT has the capability of collecting and sharing information without human interaction.
- **Machine Learning (ML):** Area of research in AI concerned with developing algorithms with predictive capabilities through learning from past examples or patterns.

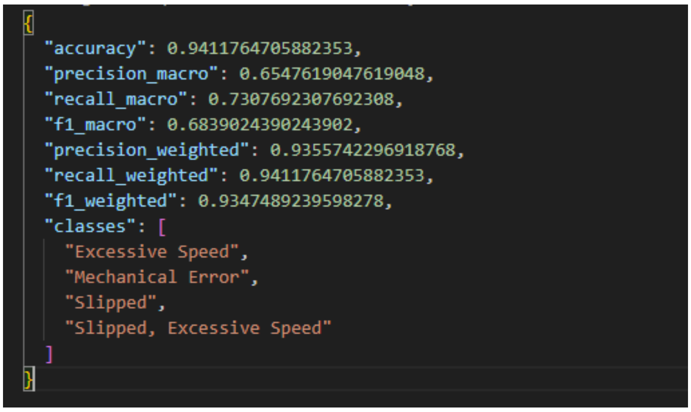
- **Microclimate:** Localized atmospheric conditions that differ from the surrounding climate and influence driving risk, especially in mountainous areas.
- **Preprocessing Pipeline:** A structured sequence of steps—such as data cleaning, imputation, encoding, and normalization—executed before model training or inference.
- **Random Forest:** t is an accepted ensemble learning approach that builds many decision trees that work better in the field of classification and regression tasks.
- **Real-Time Data Integration:** The constant incorporation of actual information (for instance, weather and traffic information) into a predictive model to enable dynamic risk analysis.
- **Recurrent Neural Network (RNN):** A deep learning technique developed for processing sequential patterns, such as time series patterns, exhibited in traffic patterns.
- **Risk Scoring Engine:** A risk calculation engine that combines the results of various models, including SPI, segment rates, cause probabilities, and multipliers, to produce a single risk score that falls between 0 and 100.
- **Risk Tiles:** Spatial-temporal data structures representing aggregated risk information across road segments, time intervals, and environmental conditions.
- **Root Mean Squared Error (RMSE):** A metric used to evaluate regression model performance, representing the square root of the average squared prediction error.
- **Segmentation:** The road network is divided or segmented into smaller geographic units that can be modeled locally for risk analysis.
- **Segment-Rate Model:** This is a regression model that calculates the expected number of incidents per road segment.
- **SHAP (SHapley Additive exPlanations):** An interpretability technique to explain how individual features contribute to a prediction.
- **Smoothed SPI (Speed Propensity Index):** A risk index based on accident data, which is then smoothed in order to reduce variability.

- **Segment Propensity Index (SPI):** An indicator of how likely speeding events happen on a particular road segment.
- **Temporal Features:** These are features such as hour, weekday, and season. They are used for predictive purposes.
- **Vehicle Multiplier:** It is used to adjust the calculations of risk in respect of stability and vulnerability to different types of vehicles.
- **Vehicle-Specific Risk Model:** This model produces specific risk outputs depending on the type of vehicle involved in the encounter.
- **XGBoost (Extreme Gradient Boosting):** It is a highly efficient machine learning algorithm known for its accuracy and effectiveness on tabular data.

APPENDICES

Appendix A: Historical Risk Engine Outputs

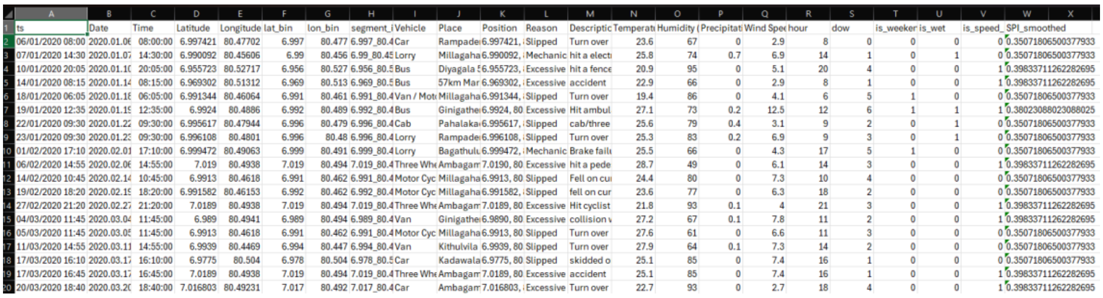
- Classification_metrics.json



```
{
  "accuracy": 0.9411764705882353,
  "precision_macro": 0.6547619047619048,
  "recall_macro": 0.7307692307692308,
  "f1_macro": 0.6839024390243902,
  "precision_weighted": 0.9355742296918768,
  "recall_weighted": 0.9411764705882353,
  "f1_weighted": 0.9347489239598278,
  "classes": [
    "Excessive Speed",
    "Mechanical Error",
    "Slipped",
    "Slipped, Excessive Speed"
  ]
}
```

Figure A-1: Classification_metrics.json

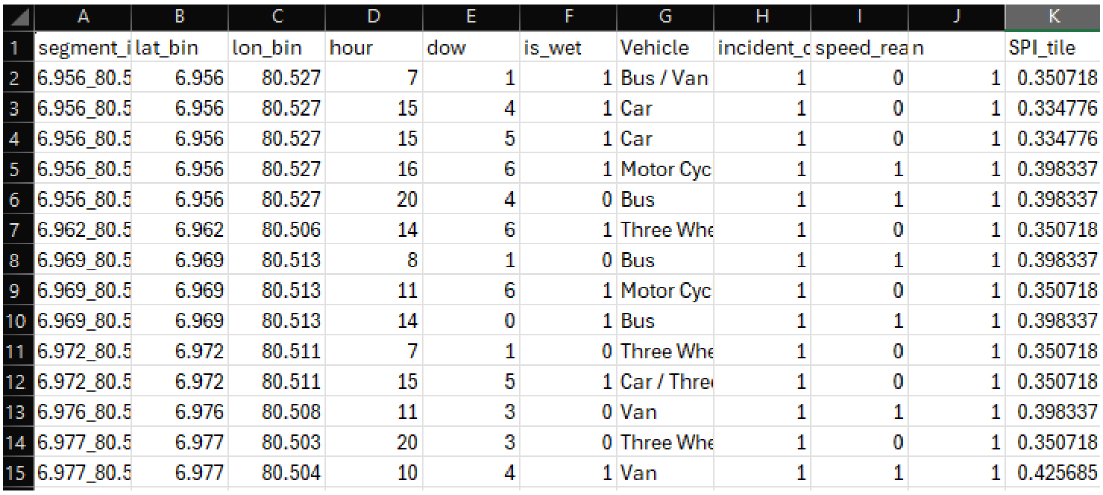
- final_dataset_min.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	ts	Date	Time	Latitude	Longitude	lat_bin	lon_bin	segment_i	Vehicle	Place	Position	Reason	Descriptic	Temperatu	Humidity (	Precipitati	Wind Spee	hour	dow	is_weeker	is_wet	is_speed_	SPI_smoothed	
2	06/01/2020 08:00	2020.01.06	08:00:00	6.997421	80.47702	6.997	80.477	6.997_80.4	Car	Rampade	6.997421,	Slipped	Turn over	23.6	67	0	2.9	8	0	0	0	0	0.35071806500377933	
3	07/01/2020 14:30	2020.01.07	14:30:00	6.990092	80.45606	6.99	80.456	6.99_80.45	Lorry	Millagaha	6.990092,	Mechanic	hit a electi	25.8	74	0.7	6.9	14	1	0	1	0	0.35071806500377933	
4	10/01/2020 20:05	2020.01.10	20:05:00	6.955723	80.52717	6.956	80.527	6.956_80.5	Bus	Diyagala 5	6.955723,	Excessive	hit a fence	20.9	95	0	5.1	20	4	0	0	1	0.39833711262282695	
5	14/01/2020 08:15	2020.01.14	08:15:00	6.969302	80.51312	6.969	80.513	6.969_80.5	Bus	57km Mar	6.969302,	Excessive	accident	22.9	66	0	2.9	8	1	0	0	1	0.39833711262282695	
6	18/01/2020 06:05	2020.01.18	06:05:00	6.991344	80.46064	6.991	80.461	6.991_80.4	Van / Mot	Millagaha	6.991344,	Slipped	Turn over	19.4	86	0	4.1	6	5	1	0	0	0.35071806500377933	
7	19/01/2020 12:35	2020.01.19	12:35:00	6.9924	80.4886	6.992	80.489	6.992_80.4	Bus	Ginigathe	6.9924, 80	Excessive	Hit ambul	27.1	73	0.2	12.5	12	6	1	1	1	0.38023088023088025	
8	22/01/2020 09:30	2020.01.22	09:30:00	6.995617	80.47944	6.996	80.479	6.996_80.4	Cab	Pahalaka	6.995617,	Slipped	cab/three	25.6	79	0.4	3.1	9	2	0	1	0	0.35071806500377933	
9	23/01/2020 09:30	2020.01.23	09:30:00	6.996108	80.4801	6.996	80.48	6.996_80.4	Lorry	Rampade	6.996108,	Slipped	Turn over	25.3	83	0.2	6.9	9	3	0	1	0	0.35071806500377933	
10	01/02/2020 17:10	2020.02.01	17:10:00	6.999472	80.49063	6.999	80.491	6.999_80.4	Lorry	Bagathulu	6.999472,	Mechanic	Brake failu	25.5	66	0	4.3	17	5	1	0	0	0.35071806500377933	
11	06/02/2020 14:55	2020.02.06	14:55:00	7.019	80.4938	7.019	80.494	7.019_80.4	Three Whe	Ambagam	7.0190, 80	Excessive	hit a pede	28.7	49	0	6.1	14	3	0	0	1	0.39833711262282695	
12	14/02/2020 10:45	2020.02.14	10:45:00	6.9913	80.4618	6.991	80.462	6.991_80.4	Motor Cyc	Millagaha	6.9913, 80	Slipped	Fell on cu	24.4	80	0	7.3	10	4	0	0	0	0.35071806500377933	
13	19/02/2020 18:20	2020.02.19	18:20:00	6.991582	80.46153	6.992	80.462	6.992_80.4	Motor Cyc	Millagaha	6.991582,	Slipped	fell on cur	23.6	77	0	6.3	18	2	0	0	0	0.35071806500377933	
14	27/02/2020 21:20	2020.02.27	21:20:00	7.0189	80.4938	7.019	80.494	7.019_80.4	Three Whe	Ambagam	7.0189, 80	Excessive	Hit cyclist	21.8	93	0.1	4	21	3	0	0	1	0.39833711262282695	
15	04/03/2020 11:45	2020.03.04	11:45:00	6.989	80.4941	6.989	80.494	6.989_80.4	Van	Ginigathe	6.9890, 80	Excessive	collision v	27.2	67	0.1	7.8	11	2	0	0	1	0.39833711262282695	
16	05/03/2020 11:45	2020.03.05	11:45:00	6.9913	80.4618	6.991	80.462	6.991_80.4	Motor Cyc	Millagaha	6.9913, 80	Slipped	Turn over	27.6	61	0	6.6	11	3	0	0	0	0.35071806500377933	
17	11/03/2020 14:55	2020.03.11	14:55:00	6.9939	80.4469	6.994	80.447	6.994_80.4	Van	Kithulvila	6.9939, 80	Slipped	Turn over	27.9	64	0.1	7.3	14	2	0	0	0	0.35071806500377933	
18	17/03/2020 16:10	2020.03.17	16:10:00	6.9775	80.504	6.978	80.504	6.978_80.5	Car	Kadawala	6.9775, 80	Slipped	skidded o	25.1	85	0	7.4	16	1	0	0	0	0.35071806500377933	
19	17/03/2020 16:45	2020.03.17	16:45:00	7.0189	80.4938	7.019	80.494	7.019_80.4	Three Whe	Ambagam	7.0189, 80	Excessive	accident	25.1	85	0	7.4	16	1	0	0	1	0.39833711262282695	
20	20/03/2020 18:40	2020.03.20	18:40:00	7.016803	80.49231	7.017	80.492	7.017_80.4	Car	Ambagam	7.016803,	Excessive	Turn over	22.7	93	0	2.7	18	4	0	0	1	0.39833711262282695	

Figure A-2: final_dataset_min.csv

- risk_tiles.csv

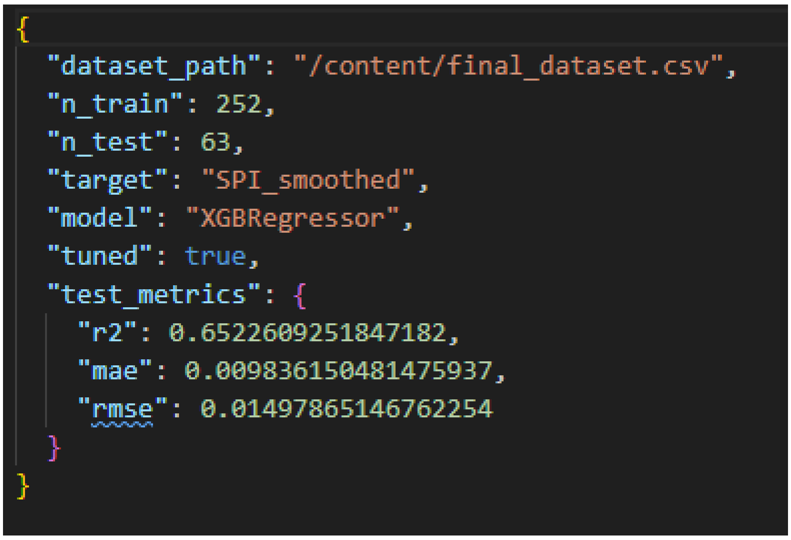


	A	B	C	D	E	F	G	H	I	J	K
1	segment_i	lat_bin	lon_bin	hour	dow	is_wet	Vehicle	incident_c	speed_rea	n	SPI_tile
2	6.956_80.5	6.956	80.527	7	1	1	Bus / Van	1	0	1	0.350718
3	6.956_80.5	6.956	80.527	15	4	1	Car	1	0	1	0.334776
4	6.956_80.5	6.956	80.527	15	5	1	Car	1	0	1	0.334776
5	6.956_80.5	6.956	80.527	16	6	1	Motor Cyc	1	1	1	0.398337
6	6.956_80.5	6.956	80.527	20	4	0	Bus	1	1	1	0.398337
7	6.962_80.5	6.962	80.506	14	6	1	Three Whe	1	0	1	0.350718
8	6.969_80.5	6.969	80.513	8	1	0	Bus	1	1	1	0.398337
9	6.969_80.5	6.969	80.513	11	6	1	Motor Cyc	1	0	1	0.350718
10	6.969_80.5	6.969	80.513	14	0	1	Bus	1	1	1	0.398337
11	6.972_80.5	6.972	80.511	7	1	0	Three Whe	1	0	1	0.350718
12	6.972_80.5	6.972	80.511	15	5	1	Car / Thre	1	0	1	0.350718
13	6.976_80.5	6.976	80.508	11	3	0	Van	1	1	1	0.398337
14	6.977_80.5	6.977	80.503	20	3	0	Three Whe	1	0	1	0.350718
15	6.977_80.5	6.977	80.504	10	4	1	Van	1	1	1	0.425685

Figure A-3: risk_tiles.csv

Appendix B: Real-Time Risk Pipeline Outputs

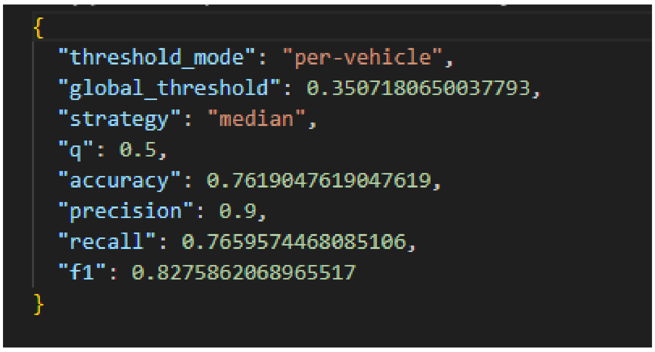
- metrics.json



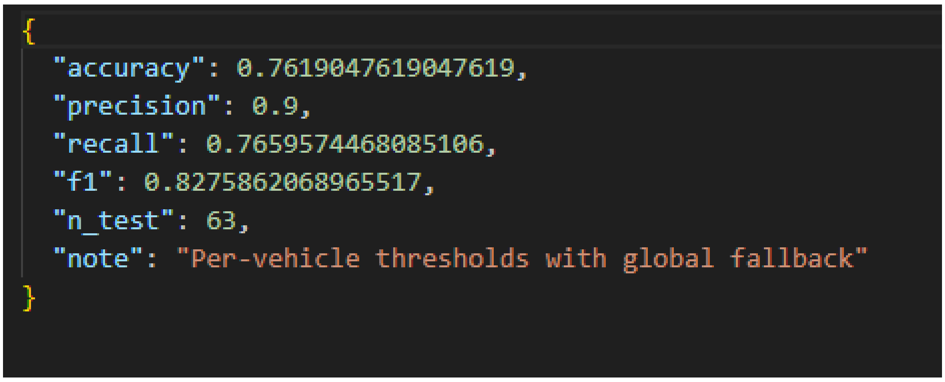
```
{
  "dataset_path": "/content/final_dataset.csv",
  "n_train": 252,
  "n_test": 63,
  "target": "SPI_smoothed",
  "model": "XGBRegressor",
  "tuned": true,
  "test_metrics": {
    "r2": 0.6522609251847182,
    "mae": 0.009836150481475937,
    "rmse": 0.014978651467622254
  }
}
```

Figure B-1: metrics.json

- classification_metrics.json



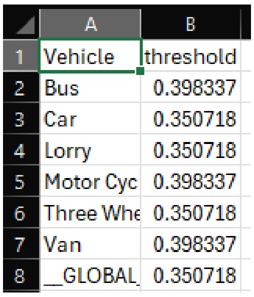
```
{
  "threshold_mode": "per-vehicle",
  "global_threshold": 0.3507180650037793,
  "strategy": "median",
  "q": 0.5,
  "accuracy": 0.7619047619047619,
  "precision": 0.9,
  "recall": 0.7659574468085106,
  "f1": 0.8275862068965517
}
```



```
{
  "accuracy": 0.7619047619047619,
  "precision": 0.9,
  "recall": 0.7659574468085106,
  "f1": 0.8275862068965517,
  "n_test": 63,
  "note": "Per-vehicle thresholds with global fallback"
}
```

Figure B-2: classification_metrics.json

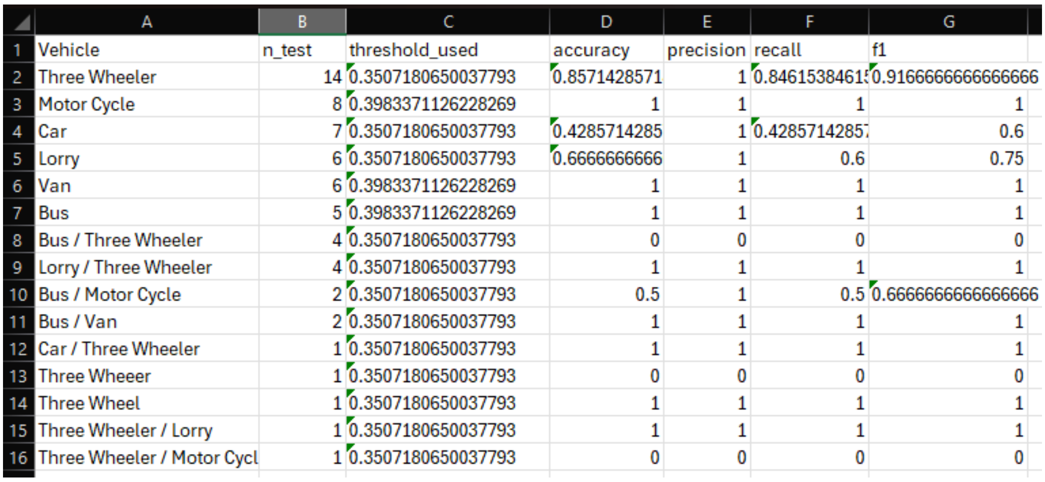
- vehicle_thresholds.csv



	A	B
1	Vehicle	threshold
2	Bus	0.398337
3	Car	0.350718
4	Lorry	0.350718
5	Motor Cyc	0.398337
6	Three Whe	0.350718
7	Van	0.398337
8	__GLOBAL_	0.350718

Figure B-3: vehicle_thresholds.csv

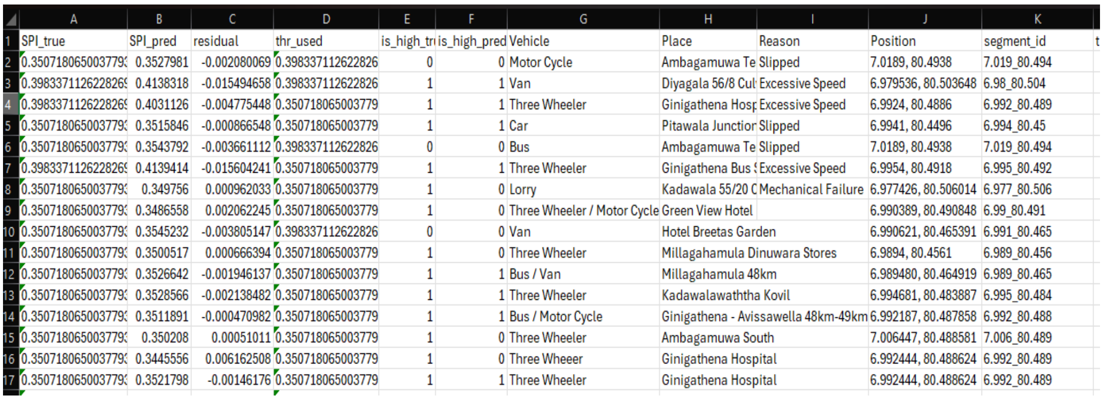
- classification_metrics_per_vehicle.csv



Vehicle	n_test	threshold_used	accuracy	precision	recall	f1
Three Wheeler	14	0.3507180650037793	0.8571428571	1	0.8461538461	0.9166666666666666
Motor Cycle	8	0.3983371126228269	1	1	1	1
Car	7	0.3507180650037793	0.4285714285	1	0.42857142857	0.6
Lorry	6	0.3507180650037793	0.6666666666	1	0.6	0.75
Van	6	0.3983371126228269	1	1	1	1
Bus	5	0.3983371126228269	1	1	1	1
Bus / Three Wheeler	4	0.3507180650037793	0	0	0	0
Lorry / Three Wheeler	4	0.3507180650037793	1	1	1	1
Bus / Motor Cycle	2	0.3507180650037793	0.5	1	0.5	0.6666666666666666
Bus / Van	2	0.3507180650037793	1	1	1	1
Car / Three Wheeler	1	0.3507180650037793	1	1	1	1
Three Wheeer	1	0.3507180650037793	0	0	0	0
Three Wheel	1	0.3507180650037793	1	1	1	1
Three Wheeler / Lorry	1	0.3507180650037793	1	1	1	1
Three Wheeler / Motor Cycl	1	0.3507180650037793	0	0	0	0

Figure B-4: classification_metrics_per_vehicle.csv

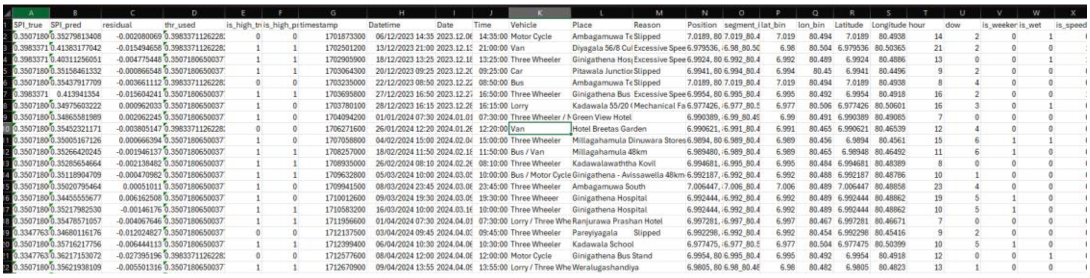
- predictions.csv



SPI_true	SPI_pred	residual	thr_used	is_high_tr	is_high_pred	Vehicle	Place	Reason	Position	segment_id	t
0.35071806500377	0.3527981	-0.002080069	0.398337112622826	0	0	Motor Cycle	Ambagamuwa Te	Slipped	7.0189, 80.4938	7.019_80.494	
0.39833711262282	0.4138318	-0.015494658	0.398337112622826	1	1	Van	Diyagala 56/8 Cul	Excessive Speed	6.979536, 80.503648	6.98_80.504	
0.39833711262282	0.4031126	-0.004775448	0.350718065003779	1	1	Three Wheeler	Ginigathena Hosp	Excessive Speed	6.9924, 80.4886	6.992_80.489	
0.35071806500377	0.3515846	-0.000866548	0.350718065003779	1	1	Car	Pitawala Junction	Slipped	6.9941, 80.4496	6.994_80.45	
0.35071806500377	0.3543792	-0.003661112	0.398337112622826	0	0	Bus	Ambagamuwa Te	Slipped	7.0189, 80.4938	7.019_80.494	
0.39833711262282	0.4139414	-0.015604241	0.350718065003779	1	1	Three Wheeler	Ginigathena Bus	Excessive Speed	6.9954, 80.4918	6.995_80.492	
0.35071806500377	0.349756	0.000962033	0.350718065003779	1	0	Lorry	Kadawala 55/20 C	Mechanical Failure	6.977426, 80.506014	6.977_80.506	
0.35071806500377	0.3486558	0.002062245	0.350718065003779	1	0	Three Wheeler / Motor Cycle	Green View Hotel		6.990389, 80.490848	6.99_80.491	
0.35071806500377	0.3545232	-0.003805147	0.398337112622826	0	0	Van	Hotel Breetas Garden		6.990621, 80.465391	6.991_80.465	
0.35071806500377	0.3500517	0.000666394	0.350718065003779	1	0	Three Wheeler	Millagahamula Dinuwara Stores		6.9894, 80.4561	6.989_80.456	
0.35071806500377	0.3526642	-0.001946137	0.350718065003779	1	1	Bus / Van	Millagahamula 48km		6.989480, 80.464919	6.989_80.465	
0.35071806500377	0.3528566	-0.002138482	0.350718065003779	1	1	Three Wheeler	Kadawalawaththa Kovil		6.994681, 80.483887	6.995_80.484	
0.35071806500377	0.3511891	-0.000470982	0.350718065003779	1	1	Bus / Motor Cycle	Ginigathena - Avissawella 48km-49km		6.992187, 80.487858	6.992_80.488	
0.35071806500377	0.350208	0.00051011	0.350718065003779	1	0	Three Wheeler	Ambagamuwa South		7.006447, 80.488581	7.006_80.489	
0.35071806500377	0.3445556	0.006162508	0.350718065003779	1	0	Three Wheeer	Ginigathena Hospital		6.992444, 80.488624	6.992_80.489	
0.35071806500377	0.3521798	-0.00146176	0.350718065003779	1	1	Three Wheeler	Ginigathena Hospital		6.992444, 80.488624	6.992_80.489	

Figure B-5: predictions.csv

- predictions_from_pkl – trained model.csv

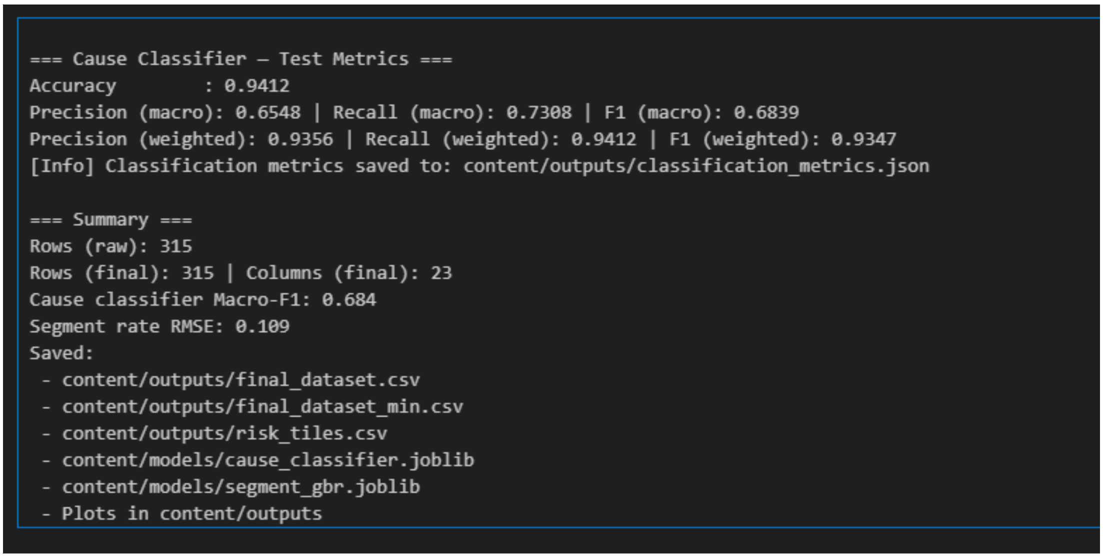


SPI_true	SPI_pred	residual	thr_used	is_high_tr	is_high_pr	timestamp	Datetime	Date	Time	Vehicle	Place	Reason	Position	segment_i	lat_bin	lon_bin	Latitude	Longitude	hour	dow	is_weeker	is_wet	is_speed
0.3507180	0.35279813408	-0.002080069	0.3983371126228	0	0	1701873300	06/12/2023 14:35	2023.12.0	14:35:00	Motor Cycle	Ambagamuwa Te	Slipped	7.0189, 80	7.019_80.4	7.019	80.494	7.0189	80.4938	14	2	0	1	0
0.3983371	0.41383177042	-0.015494658	0.3983371126228	1	1	1702501200	13/12/2023 21:00	2023.12.1	21:00:00	Van	Diyagala 56/8 Cul	Excessive Spee	6.979536,	6.98_80.50	6.98	80.504	6.979536	80.50365	21	2	0	0	1
0.3983371	0.40311256051	-0.004775448	0.3507180650037	1	1	1702905900	18/12/2023 13:25	2023.12.1	13:25:00	Three Wheeler	Ginigathena Hos	Excessive Spee	6.9924, 80	6.992_80.4	6.992	80.489	6.9924	80.4886	13	0	0	1	1
0.3507180	0.35158461332	-0.000866548	0.3507180650037	1	1	1703064300	20/12/2023 09:25	2023.12.2	09:25:00	Car	Pitawala Junctio	Slipped	6.9941, 80	6.994_80.4	6.994	80.45	6.9941	80.4496	9	2	0	0	0
0.3507180	0.35437917709	-0.003661112	0.3983371126228	0	0	1703235000	22/12/2023 08:50	2023.12.2	08:50:00	Bus	Ambagamuwa Te	Slipped	7.0189, 80	7.019_80.4	7.019	80.494	7.0189	80.4938	8	4	0	0	0
0.3983371	0.413941354	-0.015604241	0.3507180650037	1	1	1703695800	27/12/2023 16:50	2023.12.2	16:50:00	Three Wheeler	Ginigathena Bus	Excessive Spee	6.9954, 80	6.995_80.4	6.995	80.492	6.9954	80.4918	16	2	0	0	1
0.3507180	0.34975603222	0.000962033	0.3507180650037	1	0	1703780100	28/12/2023 16:15	2023.12.2	16:15:00	Lorry	Kadawala 55/20 (	Mechanical Fa	6.977426,	6.977_80.5	6.977	80.506	6.977426	80.50601	16	3	0	1	0
0.3507180	0.34865581989	0.002062245	0.3507180650037	1	0	1704094200	01/01/2024 07:30	2024.01.0	07:30:00	Three Wheeler / M	Green View Hotel		6.990389,	6.99_80.49	6.99	80.491	6.990389	80.49085	7	0	0	0	0
0.3507180	0.35452321171	-0.003805147	0.3983371126228	0	0	1706271600	26/01/2024 12:20	2024.01.2	12:20:00	Van	Hotel Breetas Garden		6.990621,	6.991_80.4	6.991	80.465	6.990621	80.46539	12	4	0	0	0
0.3507180	0.35005167126	0.000666394	0.3507180650037	1	0	1707058800	04/02/2024 15:00	2024.02.0	15:00:00	Three Wheeler	Millagahamula Dinuwara Stores		6.9894, 80	6.989_80.4	6.989	80.456	6.9894	80.4561	15	6	1	1	0
0.3507180	0.35266420245	-0.001946137	0.3507180650037	1	1	1708257000	18/02/2024 11:50	2024.02.1	11:50:00	Bus / Van	Millagahamula 48km		6.989480,	6.989_80.4	6.989	80.465	6.98948	80.46492	11	6	1	0	0
0.3507180	0.35285654664	-0.002138482	0.3507180650037	1	1	1708935000	26/02/2024 08:10	2024.02.2	08:10:00	Three Wheeler	Kadawalawaththa Kovil		6.994681,	6.995_80.4	6.995	80.484	6.994681	80.48389	8	0	0	0	0
0.3507180	0.35118904709	-0.000470982	0.3507180650037	1	1	1709632800	05/03/2024 10:00	2024.03.0	10:00:00	Bus / Motor Cycle	Ginigathena - Avissawella 48km-		6.992187,	6.992_80.4	6.992	80.488	6.992187	80.48786	10	1	0	0	0
0.3507180	0.35020795464	0.00051011	0.3507180650037	1	0	1709941500	08/03/2024 23:45	2024.03.0	23:45:00	Three Wheeler	Ambagamuwa South		7.006447,	7.006_80.4	7.006	80.489	7.006447	80.48858	23	4	0	0	0
0.3507180	0.34455555677	0.006162508	0.3507180650037	1	0	1710012600	09/03/2024 19:30	2024.03.0	19:30:00	Three Wheeer	Ginigathena Hospital		6.992444,	6.992_80.4	6.992	80.489	6.992444	80.48862	19	5	1	0	0
0.3507180	0.35217982530	-0.00146176	0.3507180650037	1	1	1710583200	16/03/2024 10:00	2024.03.1	10:00:00	Three Wheeler	Ginigathena Hospital		6.992444,	6.992_80.4	6.992	80.489	6.992444	80.48862	10	5	1	0	0
0.3507180	0.35478571057	-0.004067646	0.3507180650037	1	1	1711956600	01/04/2024 07:30	2024.04.0	07:30:00	Lorry / Three Whe	Ranjurawa Prashan Hotel		6.997281,	6.997_80.4	6.997	80.467	6.997281	80.46671	7	0	0	0	0
0.3347763	0.34680116176	-0.012024827	0.3507180650037	0	0	1712137500	03/04/2024 09:45	2024.04.0	09:45:00	Three Wheeler	Pareyiyagala	Slipped	6.992298,	6.992_80.4	6.992	80.454	6.992298	80.45416	9	2	0	0	0
0.3507180	0.35716217756	-0.006444113	0.3507180650037	1	1	1712399400	06/04/2024 10:30	2024.04.0	10:30:00	Three Wheeler	Kadawala School		6.977475,	6.977_80.5	6.977	80.504	6.977475	80.50399	10	5	1	0	0
0.3347763	0.36217153072	-0.027395196	0.3983371126228	0	0	1712577600	08/04/2024 12:00	2024.04.0	12:00:00	Motor Cycle	Ginigathena Bus Stand		6.9954, 80	6.995_80.4	6.995	80.492	6.9954	80.4918	12	0	0	1	0
0.3507180	0.35621938109	-0.005501316	0.3507180650037	1	1	1712670900	09/04/2024 13:55	2024.04.0	13:55:00	Lorry / Three Whe	Weralugashandiya		6.9805, 80	6.98_80.48	6.98	80.482	6.9805	80.4823	13	1	0	0	0

Figure B-6: predictions_from_pkl – trained model.csv

Appendix C: Console Run Logs

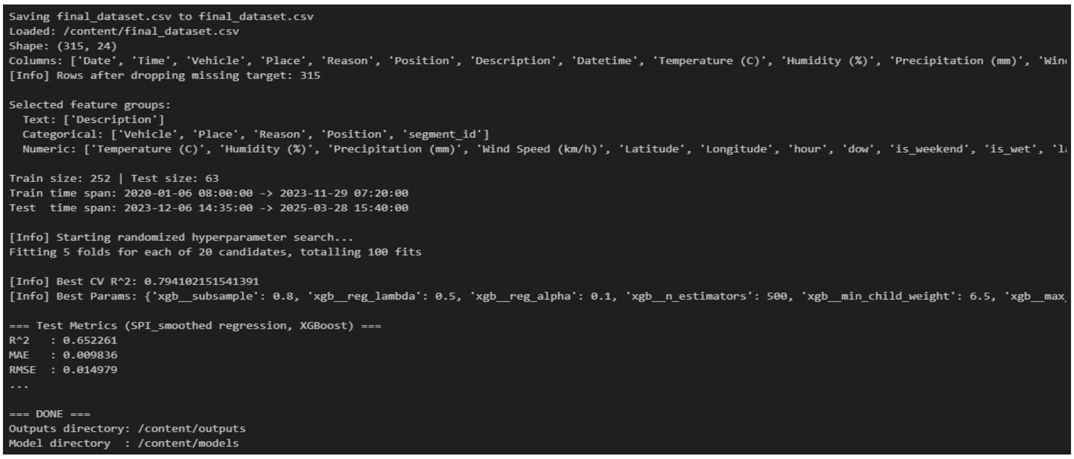
- Historical Risk Engine



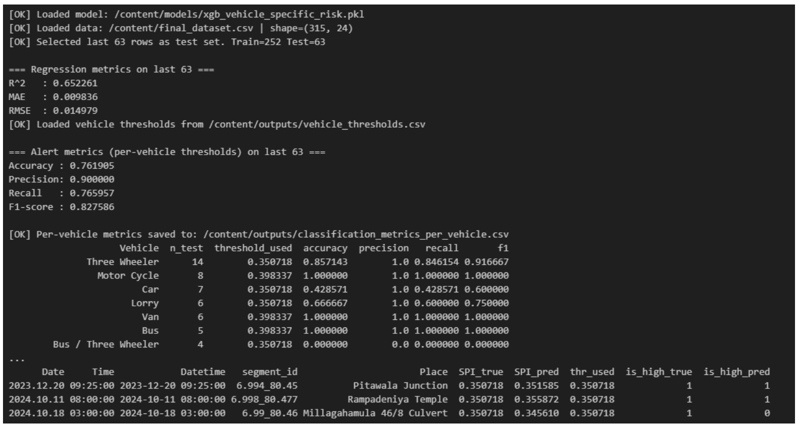
```
=== Cause Classifier – Test Metrics ===
Accuracy        : 0.9412
Precision (macro): 0.6548 | Recall (macro): 0.7308 | F1 (macro): 0.6839
Precision (weighted): 0.9356 | Recall (weighted): 0.9412 | F1 (weighted): 0.9347
[Info] Classification metrics saved to: content/outputs/classification_metrics.json

=== Summary ===
Rows (raw): 315
Rows (final): 315 | Columns (final): 23
Cause classifier Macro-F1: 0.684
Segment rate RMSE: 0.109
Saved:
 - content/outputs/final_dataset.csv
 - content/outputs/final_dataset_min.csv
 - content/outputs/risk_tiles.csv
 - content/models/cause_classifier.joblib
 - content/models/segment_gbr.joblib
 - Plots in content/outputs
```

Figure C-1: Historical Risk Engine

- Realtime Risk Pipeline



```
Saving final_dataset.csv to final_dataset.csv
Loaded: /content/final_dataset.csv
Shape: (315, 24)
Columns: ['Date', 'Time', 'Vehicle', 'Place', 'Reason', 'Position', 'Description', 'Datetime', 'Temperature (C)', 'Humidity (%)', 'Precipitation (mm)', 'Win
[Info] Rows after dropping missing target: 315

Selected feature groups:
  Text: ['Description']
  Categorical: ['Vehicle', 'Place', 'Reason', 'Position', 'segment_id']
  Numeric: ['Temperature (C)', 'Humidity (%)', 'Precipitation (mm)', 'Wind Speed (km/h)', 'Latitude', 'Longitude', 'hour', 'dow', 'is_weekend', 'is_wet', 'l

Train size: 252 | Test size: 63
Train time span: 2020-01-06 08:00:00 -> 2023-11-29 07:20:00
Test  time span: 2023-12-06 14:35:00 -> 2025-03-28 15:40:00

[Info] Starting randomized hyperparameter search...
Fitting 5 folds for each of 20 candidates, totalling 100 fits

[Info] Best CV R^2: 0.794102151541391
[Info] Best Params: {'xgb__subsample': 0.8, 'xgb__reg_lambda': 0.5, 'xgb__reg_alpha': 0.1, 'xgb__n_estimators': 500, 'xgb__min_child_weight': 6.5, 'xgb__max_

=== Test Metrics (SPI_smoothed regression, XGBoost) ===
R^2   : 0.652261
MAE   : 0.009836
RMSE  : 0.014979
...

=== DONE ===
Outputs directory: /content/outputs
Model directory  : /content/models
```



```
[OK] Loaded model: /content/models/xgb_vehicle_specific_risk.pkl
[OK] Loaded data: /content/final_dataset.csv | shape=(315, 24)
[OK] Selected last 63 rows as test set. Train=252 Test=63

=== Regression metrics on last 63 ===
R^2   : 0.652261
MAE   : 0.009836
RMSE  : 0.014979
[OK] Loaded vehicle thresholds from /content/outputs/vehicle_thresholds.csv

=== Alert metrics (per-vehicle thresholds) on last 63 ===
Accuracy : 0.761905
Precision: 0.900000
Recall   : 0.765957
F1-score : 0.827586

[OK] Per-vehicle metrics saved to: /content/outputs/classification_metrics_per_vehicle.csv
                  Vehicle  n_test  threshold_used  accuracy  precision   recall       f1
            Three Wheeler      14        0.350718  0.857143        1.0 0.846154 0.916667
              Motor Cycle       8        0.398337  1.000000        1.0 1.000000 1.000000
                      Car       7        0.350718  0.428571        1.0 0.428571 0.600000
                    Lorry       6        0.350718  0.666667        1.0 0.600000 0.750000
                      Van       6        0.398337  1.000000        1.0 1.000000 1.000000
                      Bus       5        0.398337  1.000000        1.0 1.000000 1.000000
      Bus / Three Wheeler       4        0.350718  0.000000        0.0 0.000000 0.000000
...
      Date     Time            Datetime   segment_id                      Place  SPI_true  SPI_pred  thr_used  is_high_true  is_high_pred
2023.12.20 09:25:00 2023-12-20 09:25:00  6.994_80.45          Pitawala Junction  0.350718  0.351585  0.350718             1             1
2024.10.11 08:00:00 2024-10-11 08:00:00 6.998_80.477         Rampadeniya Temple  0.350718  0.355872  0.350718             1             1
2024.10.18 03:00:00 2024-10-18 03:00:00   6.99_80.46 Millagahamula 46/8 Culvert  0.350718  0.345610  0.350718             1             0
```

Figure C-2: Realtime Risk Pipeline