MODELLING OF REAR END CONFLICT RISK AT UNCONTROLLED INTERSECTION USING SURROGATE SAFETY MEASURES

PROJECT REPORT

Submitted by

PARVATHY S

TKM21CETE11

to

the APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the Degree

of

Master of Technology in

Transportation Engineering

DEPARTMENT OF CIVIL ENGINEERING

TKM College of Engineering, Kollam

May 2023

DECLARATION

I, Parvathy S hereby declare that, this project report entitled “Modelling of Rear End Conflict Risk at Uncontrolled Intersection using Surrogate Safety Measures” which I submitted to partially satisfy the requirements for the award of a Master of Technology degree from the APJ Abdul Kalam Technological University in Kerala is a genuine work that I completed under the guidance of Prof. Meenu Tomson. This submission represents my ideas in my own words and where ideas or words of others have been included; I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Kollam

10-05-2023 PARVATHY S

DEPARTMENT OF CIVIL ENGINEERING

TKM COLLEGE OF ENGINEERING, KOLLAM

CERTIFICATE

Certified that this report entitled **"MODELLING OF REAR END CONFLICT RISK AT UNCONTROLLED INTERSECTION USING SURROGATE SAFETY MEASURES"** is the report of project presented by **PARVATHY S, TKM21CETE11** during 2022-2023 in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Transportation Engineering of the APJ Abdul Kalam Technological University.

Guide:	Project coordinator:	Head of the Department:
Prof. Meenu Tomson	**Dr. Adarsh S**	**Dr. Sajeeb R**
Assistant Professor	Associate Professor	Professor
Dept. of Civil Engg.	Dept. of Civil Engg.	Dept. of Civil Engg.
TKMCE, Kollam	TKMCE, Kollam	TKMCE, Kollam

ACKNOWLEDGEMENT

I would like to use this opportunity to express my sincere gratitude to everyone who helped and assisted me in completing the project successfully.

First of all, I thank almighty for giving me strength and courage to do the project. I extend my sincere gratitude to my respected Principal, **Dr. T.A Shahul Hameed** for his support towards the successful accomplishment of this course.

I avail this opportunity to express my gratitude to my guide **Prof. Meenu Tomson**, Assistant Professor, TKM College of Engineering for her inspiring assistance, encouragement and useful guidance.

I'm grateful to **Dr. Jithin Raj P. V.**, Assistant Professor, TKM College of Engineering and **Prof. Anandu V. G.,** Assistant Professor, TKM College of Engineering, for serving as my evaluators for the project.

I also thank **Dr. Sajeeb R**, Head of the Department of Civil Engineering, TKM College of Engineering, Kollam for the immense support.

I also thank my coordinator **Dr. Adarsh S**, Associate Professor, TKM College of Engineering for his timely guidance.

Finally, I express my sincere appreciation and thanks to my parents, family, and friends who have supported me throughout this journey, providing encouragement and motivation to help me reach the finish line.

PARVATHY S

ABSTRACT

Road traffic accidents are a global concern, with intersections being particularly vulnerable to collisions. Identifying the factors contributing to conflicts is crucial for predicting conflict risk and implementing effective measures to enhance vehicle safety. However, prior research on conflict severity has been based on aggregated data, overlooking individual vehicle dynamics. This study addresses this gap by utilizing trajectory data from individual vehicles to create a conflict risk classification model and identify the determinants of traffic conflicts. The study involves several stages, including cluster analysis of traffic conflict indicators, implementation of five machine learning classification models and interpretation of feature importance. Traffic conflict indicators such as MTTC, DRAC, and PSD were used to identify and classify conflict risk. Three clustering algorithms - K-means, spectral, and agglomerative - were employed to classify traffic conflict indicators into four risk levels: low, medium, high, and critical conflicts. Five machine learning models were evaluated: Logistic Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine. The Random Forest algorithm outperformed all other models, achieving an accuracy of 91%, precision of 91%, recall of 92%, and an AUC score of 0.98. To address interpretability challenges in machine learning models, the SHAP analysis was employed to identify the significant variables and measure their impact on conflict risk. The study identified the top five features most likely to influence conflict risk, which included maximum deceleration of leader (MDL), standard deviation of spacing between vehicles (SDSP), maximum acceleration of follower (MAF), standard deviation of speed follower (SDSF), and mean longitudinal spacing between the vehicles (MSPA). Further analysis of the beeswarm and dependency plots for each risk class indicates that certain features, such as SDSF, MAF and MDL exhibit an increase in risk levels with increasing values, while a decrease in MSPA and headway (HW) is associated with increased conflict risk. These findings provide valuable insights for developing effective countermeasures to mitigate conflict risk and improve traffic safety, particularly for connected and automated vehicles utilizing advanced driver assistance systems.

Keywords: *Conflict risk modelling, Machine Learning Models, Traffic Safety, Traffic Conflict Indicators, SHAP*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ADAS	Advanced Driver Assistance System
AUC	Area Under the Curve
CAV	Connected and Automated Vehicle
DRAC	Deceleration Rate to Avoid Collision
FN	False Negative
FP	False Positive
FNR	False Negative Rate
FPR	False Positive Rate
MI	Mutual Information
MORTH	Ministry of Road Transport and Highways
MTTC	Modified Time to Collision
PET	Post Encroachment Time
PSD	Proportion of Stopping Distance
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	Shapely Additive Explanations
SSM	Surrogate Safety Measures
SVM	Support Vector Machine
TCI	Traffic Conflict Indicators
TCT	Traffic Conflict Technique
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TTC	Time to Collision
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Transportation has become an integral aspect of every individual's life, with individuals interacting with roads in one way or another, whether as drivers, passengers, or pedestrians. Although the current transportation system has made long-distance travel more convenient, it has also heightened the likelihood of severe injuries or fatalities. Every year, millions of individuals become involved in road accidents, resulting in an incalculable loss of lives and numerous injuries. India alone witnessed approximately eighty thousand fatalities annually due to road crashes, accounting for a significant eleven per cent of the world's total (Ministry of Road Transport and Highways [MoRTH], 2021). Also, numerous research studies have highlighted that intersections represent a high-risk area for road crashes, with T-intersections and staggered intersections emerging as the primary culprits (Bonela & Kadali, 2022). Intersections, as traffic conflict zones, involve intricate traffic movements such as turning, weaving, and through traffic, influenced by various parameters such as driver behaviour, vehicle type, and speed. These parameters are critical in defining the conflict intensity, which can escalate rapidly and pose a significant threat to road safety.

Analysing traffic safety can improve the safety of road users by identifying factors that contribute to crashes, such as high-risk road sections, hazardous driving behaviours, and inadequate infrastructure design. This information can then be used to develop and implement evidence-based interventions, such as engineering solutions, education and awareness campaigns, and enforcement programs that target these specific risk factors. These interventions can be implemented at different levels, including the individual, community, and policy levels, and may involve a combination of approaches tailored to specific needs and circumstances. By addressing these factors, the likelihood of crashes will be reduced, and the severity of crashes that do occur will be minimized.

The effectiveness of traffic crashes as a leading indicator of road safety is constrained due to the difficulties involved in collecting and verifying crash data, which undermines its accuracy and dependability as a metric. The difficulties of crash-based analysis have led to the development of substitute safety measures based on traffic conflicts, which are

situations in which more than two road users are at danger of colliding because of their near temporal and physical proximity. Several surrogate safety measures are commonly employed, such as time to collision, change in velocity, post-encroachment time and others. Therefore, in this study, the traffic conflict technique is utilized to identify conflicting interactions between vehicles, and to assess the factors that influence these interactions.

In the past, safety studies using proactive and reactive techniques were based on aggregated data without taking into consideration the motion of each individual vehicle (Islam & Abdel-Aty, 2023; Shangguan et al., 2023; Yu et al., 2021). This approach has created a research gap that this study seeks to address by utilizing trajectory data from individual vehicles to create a conflict risk classification model and identify the determinants of traffic conflicts.

1.2 TRAFFIC SAFETY IN GLOBAL SCENARIO

The World Health Organisation (WHO) states that traffic accidents result in 1.35 million fatalities and 20 to 50 million non-fatal injuries each year, a situation that in effect causes substantial financial losses for people, families, and countries. Road accidents account for 3% of an average country's GDP, with pedestrians, cyclists, and motorcyclists accounting for over half of all traffic accident deaths. Developing nations, which represent just 48% of the world's vehicles, bear over 90% of the burden in terms of road accident fatalities. Ten countries, namely India, China, the United States, Russia, Brazil, Iran, Mexico, Indonesia, South Africa, and Egypt, have the highest recorded road traffic fatalities, accounting for 56% of the global population (World Health Organisation [WHO], 2018). Speeding, driving while inebriated or under the influence of drugs, not using protective gear, preoccupied driving, hazardous road conditions, unsafe vehicles, poor post-crash care, and lax enforcement of highway regulations are all examples of reckless driving behaviours. Traffic accidents are the leading cause of death for kids and adolescents between the ages of 5 and 29.

1.3 TRAFFIC SAFETY IN INDIAN SCENARIO

According to the WHO’s Global Status Report, India has the most road traffic deaths, which make up around 11% of all worldwide fatalities. This poses a critical road safety concern for India, given its vast and expansive road network, one of the largest in the

world. The issue is further compounded by the rapid pace of motorization and urbanization driven by the country's high economic growth rate. As per the MoRTH accident report, in India an estimated 150,000 individuals lose their lives every year due to road traffic accidents, which translates to an average of 1,130 accidents and 422 fatalities each day, or 47 accidents and 18 fatalities per hour. At the national level, the data on accident or collision types reveal a rise in incidents during 2021 as compared to the previous year. Notably, the category of 'Hit from Back' accounted for the highest proportion of both total accidents (21.2%) and total number of fatalities (18.6%) during the current year, followed by 'Head on Collision' at 18.5% and 17.7% for total accidents and total number of fatalities, respectively (MoRTH, 2021). A rear-end collision, commonly known as a hit from the back, transpires when one vehicle strikes another from behind. Driver distraction or inattention, tailgating at intersections, abrupt panic stops, and decreased traction due to precipitation or degraded pavement conditions are some of the common factors that often lead to such incidents.

1.4 NEED FOR THE STUDY

The majority of current research on traffic safety bases itself on crash data, which has a number of drawbacks. One such limitation is their exclusive focus on accidents resulting in injuries or fatalities, ignoring minor incidents that could cause harm in the future. Additionally, the reliability of crash-based studies is subject to the accuracy and completeness of police reports, leading to potential underreporting of certain types of accidents. Moreover, crash-based studies do not account for near-miss incidents, which could indicate dangerous driving behaviour or hazardous road conditions. Therefore, traffic conflict-based studies are crucial for a proactive approach to road safety, identifying potential hazards before they result in crashes. Additionally, research gaps often arise by the use of accumulated data in safety studies that rely on reactive and proactive techniques without taking individual automobile dynamics into account. To address this gap, this study proposes using trajectory data from individual vehicles to create a conflict risk classification model and identify the determinants of traffic conflicts. By analysing these incidents and their underlying causes, policymakers, transportation planners and engineers can develop targeted interventions and improvements to prevent accidents and create a safer and more efficient transportation for all road users.

1.5 SIGNIFICANCE OF THE STUDY

Research has revealed that traffic accidents may result from various factors, including distracted driving, driving under the influence of drugs or alcohol, reckless driving, speeding, poor road design, adverse weather conditions, and inadequate vehicle maintenance. It's crucial to assess the safety of other road users in real-time with the goal to reduce the likelihood of crashes. The old conventional automobiles are being quickly replaced by self-driving and connected automobiles nowadays. These vehicles are equipped with advanced technology that enables them to communicate with other vehicles, infrastructure, and devices. They also make use of advanced driver assistance system (ADAS) to assist drivers in various tasks, such as braking, accelerating, and steering. ADAS systems are designed to reduce the risk of accidents and provide a more comfortable driving experience.

The development of a conflict risk classification model can be a valuable tool in enhancing the safety of connected and automated vehicles. This model can help anticipate potential hazards and predict conflicts before they arise by analysing real-time data from various sources, such as sensors, communication devices, and traffic patterns. The model classifies conflicts based on their level of severity, from minor incidents that do not require immediate intervention to critical situations that demand immediate action. This information can then be transmitted to the connected vehicles involved, which can take appropriate action to avoid the potential conflict. Additionally, ADAS, which help drivers avoid conflicts or lessen their severity, can be developed with the help of the conflict risk classification model. Examples of ADAS features like collision avoidance systems, speed adaptation systems, and lane-keeping assistance all of which work to reduce the risk of conflicts on the road. Thus, this study results can serve as a valuable asset in identifying the factors that contribute to conflicts and developing more advanced driver assistance systems to improve road safety.

1.6 GAPS IDENTIFIED

In safety studies, aggregated vehicle data is commonly used to identify patterns and trends in crashes and to establish risk assessment models. Aggregated data provides a summary of the behaviour of all vehicles in a given area over a certain period of time, such as average speed or traffic volume. Aggregated vehicle data is typically collected

through various traffic sensors and detectors installed along the roadway network. However, aggregated data does not provide information on the behaviour of individual vehicles, which may be necessary to understand the causes and mechanisms of crashes (Yuan et al., 2022a; Yu et al., 2021).

Therefore, there is a need for studies that use individual vehicle data (Hu et al., 2022; Yuan et al., 2022b). Individual vehicle data provides information on the behaviour and interactions of individual vehicles, such as their speed, acceleration, lane position, and distance to other vehicles. This information can be used to identify the contributing factors to crashes, such as driver behaviour, vehicle characteristics, and road design.

Furthermore, with the advent of new technologies such as connected and autonomous vehicles (CAVs), individual vehicle data can be collected in real-time, providing more accurate and timely information on traffic conditions and potential hazards (Islam & Abdel-Aty, 2023; Shangguan et al., 2023) . This data can be used for real-time crash risk assessment and mitigation, improving overall road safety. Also, most of the conflict or safety studies use a binary approach, categorizing crashes or conflicts as either severe or non-severe without considering the multiple levels of injury that can occur.

In the field of traffic safety, identifying traffic conflicts is an important step in understanding and mitigating potential hazards on the roadway. Typically, a single conflict indicator is used to identify traffic conflicts, such as the time-to-collision (TTC) or post-encroachment time (PET). However, using only one indicator may not capture all types of traffic conflicts or provide a complete picture of the traffic situation (Arun et al., 2021; Wang et al., 2021).

By using multiple conflict indicators, researchers can identify traffic conflicts more accurately and comprehensively. For example, using both the TTC and PET indicators can provide complementary information about the timing and severity of traffic conflicts. Additionally, using multiple indicators can help identify conflicts that may not be detected by a single indicator, such as conflicts that occur over longer time periods or involve complex manoeuvres (Arun et al., 2022). Furthermore, the use of multiple conflict indicators is less explored in mixed traffic conditions. Most of the studies conducted in India are based on a single conflict indicator. One of the other gaps identified is the lack of research on conflicts in unsignalized crossroads and roundabouts (Bonela & Kadali, 2022; Arun et al., 2021).

Most of the machine learning based conflict severity models does not provide a proper interpretation of the model results. This limitation can be addressed with the use of SHAP (SHapley Additive exPlanations) values (Islam & Abdel-Aty, 2023; Yuan et al., 2022b).

This study aims to address the existing research gaps in the field of traffic safety analysis for uncontrolled intersections under mixed traffic conditions, by employing a novel approach that combines multiple traffic conflict indicators, individual vehicle kinetics data, and SHAP for model interpretation. The research gap addressed is listed as follows

- Only a limited number of conflict based studies are conducted in unsignalized intersections and roundabouts
- The majority of safety studies use aggregated data and disregard the unique dynamics of each vehicle.
- Only a limited number of machine learning based traffic safety studies provide a clear explanation of how different input variables influence various risk levels
- Most of the conflict based studies considered only one traffic conflict indicator for classifying conflicts

1.7 OBJECTIVES OF THE STUDY

- ❖ To identify the traffic conflicts using surrogate safety measures
- ❖ To classify the conflicting vehicle interaction into different risk levels using clustering technique
- ❖ To develop conflict risk classification models using machine learning algorithms that incorporate individual vehicle trajectory data and then compare the performance of various machine learning algorithms in classifying the conflict risk using performance metrics
- ❖ To utilize SHAP (SHapley Additive exPlanation) to interpret the significance of different features in classifying conflict risk

1.8 ORGANIZATION OF REPORT

This report is organized into five chapters as follows:

- Chapter 1 Introduction: This section aims to present a comprehensive overview of the research work, emphasizing the need and significance of the study. Furthermore, it highlights the identified gaps in the existing literature and outlines the objectives of the research

- Chapter 2 Literature Review: This chapter presents a comprehensive review of the relevant literature pertaining to three key areas: traffic safety, surrogate safety measures, and crash severity modelling

- Chapter 3 Methodology: This chapter provides a comprehensive description of the methodology employed in the study, encompassing several key components such as data collection, data extraction, cluster analysis, classification modelling, and SHAP feature importance

- Chapter 4 Results and Discussions: This chapter presents the results derived from the cluster analysis, the performance of five classification models, and an exploration of feature importance using SHAP

- Chapter 5 Summary and Conclusions: This chapter provides a comprehensive summary of the entire study, including key findings, conclusions drawn from the analysis, application, limitations of the study and suggestions for future research

- References

CHAPTER 2
LITERATURE REVIEW

2.1 GENERAL BACKGROUND

Intersections represent a crucial component of road networks, where opposing streams of traffic contend for limited space and time. Based on data from million plus cities, various types of intersections accounted for a significant proportion of road accidents, fatalities, and injuries in 2020. Un-signalized intersections are more common in suburban and rural areas, where high vehicle speeds and unpredictable driver behaviour pose additional risks. In particular, collisions between through and turning traffic from fast-moving vehicles at un-signalized junctions can result in severe outcomes, jeopardizing the safety of all road users.

Traditional approaches to safety estimation based on historical crash data entail a long waiting period to gather sufficient data, prompting safety experts to seek alternative methods. The Traffic Conflict Technique (TCT) is receiving attention because it assesses the security of road infrastructure, particularly junctions, using a number of proximal indications (Uzondu et al., 2018). By recording near-crash scenarios (conflicts) and analysing them, TCT anticipates accident risk and identifies contributing factors to crash situations. The chance of a collision is significant when multiple road users are near to one another in both space and time, unless their motions are changed.

2.2 TRAFFIC SAFETY STUDIES

Traffic safety is extremely important because it involves protecting human lives and reducing the risk of injury and death caused by accidents on the roadways. Vehicle crashes are the world's greatest source of mortality and fatalities, and they may have a profound effect on individuals, families, communities, and even entire countries. In fact, with the growth of urbanization, motorization, and globalization, traffic safety has become even more critical in today's world.

Causes of road traffic accidents and the parameters that influence them can be determined through crash-based safety studies, which are commonly referred to as crash analysis or crash investigation studies. In these studies, crash-related data is often gathered and

analysed in order to further comprehend the situations that contributed to the occurrence. Crash-based safety studies can assist in pinpointing the precise causes of an accident, such as driver behaviour, road design, vehicle characteristics, and environmental factors. There are two approaches to crash-based studies; one is crash frequency modelling and other is crash severity modelling.

Crash frequency modelling is a statistical technique employed to predict the anticipated number of collisions at a particular place over a certain period of time. It involves the application of mathematical models to determine the correlation between the frequency of collisions and numerous pertinent aspects or variables, including traffic volume, road geometry, weather, and driver behaviour.

Modelling crash severity is becoming increasingly popular due to its capacity to examine the correlation between the severity of injuries sustained and the occurrence of crashes. Unlike crash frequency studies, which yield integer outputs, severity studies typically yield injury scales or other categorical variables as outcomes. During modelling, it is essential to evaluate the goodness of fit and interpret the model results (Bhuiyan et al., 2022; Lord et al., 2021).

Crash severity modelling can be carried out through various approaches, including statistical modelling and data mining techniques. According to several research, the Multinomial Logit Model, Nested Logit Model, Random Parameter Logit Model and Ordered Probit Model are some typical statistical models used for this purpose (Das et al., 2023; Ye et al., 2023; Lee et al., 2021; Ye et al., 2021; Zhang et al., 2018; Islam et al., 2016). In some studies, econometric modelling frameworks have been used to tackle heterogeneity issues. For instance, Das et al. (2023) examined the extent of cycle injury using a random parameter multinomial logit model with heterogeneity in means and variance. The study found that distracted driving behaviour was a significant factor affecting injury severity. Shangguan et al. (2023) examined the development of driving threat while vehicle-following and the variables influencing various risk patterns using vehicle kinematic data and the Rear End Conflict Risk Index. Their research showed that driving risk was significantly influenced by the density of traffic, vehicle speed, separation between the vehicles, and fluctuation in headway distance, and that the random parameter multinomial logit model performed better than the traditional method. Ye et al. (2023) used modified time to collision indicator and a random parameters multinomial

logit model to evaluate the factors influencing traffic conflicts in the motorway divergence region at the time of construction periods. The investigation found that shifting lanes were more likely to result in traffic conflicts than rear-end collisions, and it also identified several important factors that affect traffic conflicts, such as vehicle velocity, acceleration, type, position, the proportion of heavy vehicles, and the overall number of automobiles around the vicinity.

The trend towards utilizing data mining and machine learning techniques for a "data-driven safety analysis" has increased in recent years, as more individuals aim to identify novel, precise, and valuable patterns. Many recent studies have used various machine learning algorithms like Random Forest (Hu et al., 2022; Yassin & Pooja, 2020), Boosting algorithms (Dong et al., 2022; Kashifi & Ahmad, 2022), Deep Learning models (Yuan et al., 2022; Yu et al., 2021; Hu et al., 2020; Formosa et al., 2020) and Support Vector Machine Algorithm (Salas et al., 2022; Hosseinzadeh et al., 2021). Islam and Abdel-Aty (2023) used connected vehicle data and a long short-term memory model to predict traffic conflicts. Their results showed that conflicts can be predicted with an accuracy rate of 72%, a recall of 81%, and a false alarm rate of 28%, by identifying variables such as acceleration rate, deceleration rate, and speed. Lu et al. (2021) used a simulated corridor to evaluate crash risk in a mixed traffic flow environment consisting of human-driven, autonomous, and connected vehicles. They employed advanced technology, such as Vehicle-to-Infrastructure, Vehicle-to-Vehicle, and Global Navigation Satellite System devices, to gather traffic safety and traffic data. The study used the Gini Importance method to select the top 10 features and Kernel Logistic Regression to predict traffic flow, with results indicating strong predictive capability comparable to SVM models. Kashifi and Ahmad (2022) used the Efficient Gradient Boosting Decision Tree with Histogram-based approach to predict the severity of accidents using data collected from the French Open-source dataset. The results showed that the HistGBDT model outperformed seven other baseline models, with safety equipment being the most important factor in predicting the severity of an accident.

Some of the recent studies used SHAP values for overcoming the interpretability limitation of machine learning models (Islam & Abdel-Aty, 2023; Dong et al., 2022; Yuan et al., 2022b). Dong et al. (2022) developed a predictive model for road traffic injury severity using four boosting-based ensemble learning models and SHAP values. The study found that LightGBM showed the highest accuracy in classification, with the

most important variables affecting injury severity being month, cause of accident, age of driver, and type of collision, with young drivers having the most chances of fatal injuries.

Ijaz et al. (2021) developed a machine learning model to predict the severity of crashes involving three-wheeled rickshaws using data from Rescue 1122 in Rawalpindi, Pakistan. The Decision Jungle algorithm outperformed the Decision Tree and Random Forest algorithms, with an overall accuracy of 83.7%, and identified several driver-related and roadway-related factors that were positively associated with increased crash severity. Jamal et al. (2021) compared the performance of traditional machine learning algorithms with the XGBoost algorithm in predicting the multinomial target variable for crash injury severity with three possible levels. The XGBoost algorithm outperformed decision trees, random forest, and logistic regression, with an overall prediction accuracy of 93% with most important variables being identified as type of collision, weather conditions, pavement conditions, type of vehicle, number of lanes, and cause of crash.

2.3 TRAFFIC CONFLICT TECHNIQUE

The Traffic Conflict Technique (TCT) is a method of crash analysis that utilizes information on critical incidents, which may or may not involve crashes, and is known as traffic conflicts. The first definition of traffic conflict was based on the evasive action taken by the drivers to avoid collision. The next definition of traffic conflict is “an observable circumstance in which two or more road users are close enough to one another in location and time that a collision is possible if their motions don't change” (Lord et al., 2021).

In TCT, Surrogate Safety Measures (SSMs) or Traffic Conflict Indicators (TCIs) are commonly used to evaluate the risk of a traffic crash based on microscopic traffic characteristics (Johnsson et al., 2021; Zheng & Sayed, 2019). These are measures used in traffic safety analysis to identify situations where there is a high risk of a collision or other safety-critical event. TCIs typically involve the observation and recording of certain traffic behaviours or events that are indicative of potential safety issues, such as abrupt braking, swerving, or close following. These SSMs can be categorized into three groups: temporal proximity, distance, and deceleration based. The temporal proximity SSMs measure how close the conflict parties are in time, while the distance SSMs measure the physical distance between them (Bonela & Kadali, 2022; Mohanty et al., 2021; Orsini et

al., 2021; Johnsson et al., 2018). The kinematic SSMs assess the participants' kinematic characteristics, which determine their reaction to the developing conflict. A pictorial representation of car following event is shown in figure 2.1.

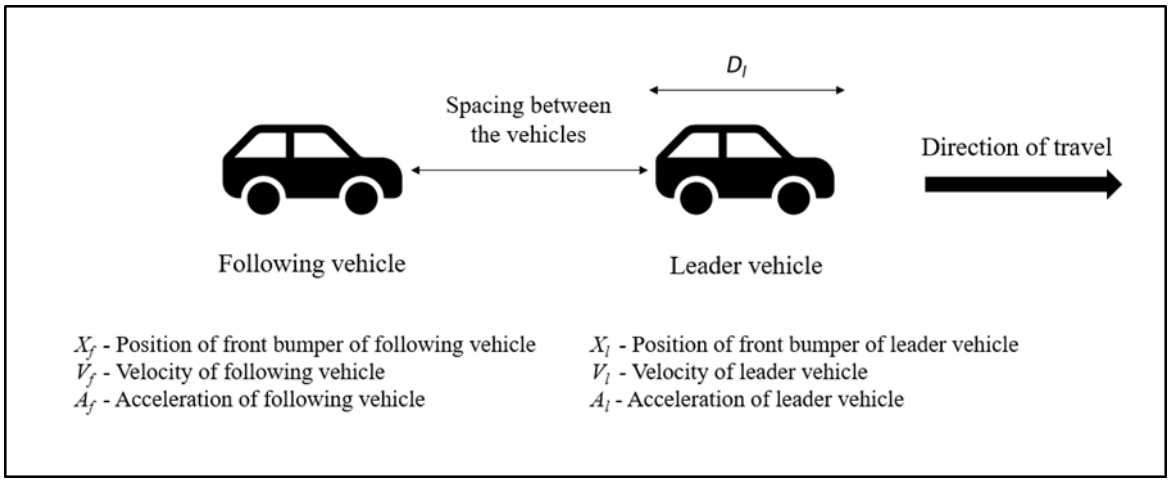

Figure 2.1 Car Following Event

Some of the commonly used traffic conflict indicators for rear end conflicts are given below:

2.3.1 Time to Collision (TTC): TTC is a crucial concept in traffic safety that estimates the time it will take for two vehicles, one leading and the other following, to collide if they maintain their current speeds and directions (Behbahani et al., 2016). The TTC value is determined based on the difference in speed between the two vehicles and is considered risky if it falls below a certain predefined threshold. However, it is important to note that TTC calculations rely on the assumption that all vehicles will continue to move in the same direction and at the same speed, and this assumption must be taken into account when evaluating potential collision risks on the road to ensure accurate TTC estimation. The equation for TTC (Wang et al., 2021) is as follows:

$$TTC = \frac{x_l - x_f - D_l}{v_f - v_l} \quad \text{..... (2.1)}$$

Where, x is the position of front bumper of vehicles (f = following vehicle; l = leading vehicle), D_l is the leading vehicle length, v is the vehicle velocity

2.3.2 Modified Time to Collision (MTTC): Although TTC is commonly used in traffic safety, it has limitations. For instance, it cannot account for scenarios in which there is still a risk of collision, even if the two vehicles are moving at the same speed, or when the

following vehicle is traveling slower than the leading vehicle. The TTC solely relies on the current speed difference between the two vehicles and neglects acceleration. In an attempt to overcome this limitation, Ozbay et al. (2008) introduced MTTC. The MTTC considers not only the relative speed and distance between the two vehicles but also their acceleration. Compared to TTC, MTTC was found to be a more accurate predictor of the probability of a collision. If the MTTC falls below a predetermined threshold, it indicates a hazardous condition, and there may be a risk of collision. The equation for MTTC (Wang et al., 2021) is as follows:

$$MTTC = \frac{-\Delta s \pm \sqrt{\Delta s^2 + 2\Delta a(x_l - x_f - D_l)}}{\Delta a} \quad \text{..... (2.2)}$$

Where, x is the position of front bumper of vehicle (f= following vehicle; l = leading vehicle), Δs is the difference in the speeds of the vehicles, Δa is the difference in their accelerations, D_l is the leading vehicle length

2.3.3 Proportion of Stopping Distance (PSD): PSD is a crucial concept in traffic safety that refers to the ratio of the distance available for a driver to manoeuvre their vehicle to the remaining distance to the projected location of a collision. As Arun et al. (2022) explain, this ratio provides an estimate of the driver's ability to avoid a collision given the available stopping distance. Specifically, if PSD is high, the driver has more space and time to manoeuvre their vehicle and potentially avoid a collision. On the other hand, if PSD is low, the driver's ability to evade a collision is limited. Therefore, PSD can be a useful metric for assessing collision risks and informing decisions related to driving behaviour and safety. The equation for PSD (Wang et al., 2021) is as follows:

$$PSD = \frac{x_l - x_f - D_l}{\frac{v_f^2}{2\mu g}} \quad \text{..... (2.3)}$$

Where, $x_l - x_f - D_l$ is the longitudinal gap between the vehicles, V_f is the velocity of following vehicle, μ is the frictional coefficient of road surface, g is the acceleration due to gravity.

2.3.4 Deceleration Rate to Avoid Collision (DRAC): DRAC is an important traffic safety concept that quantifies the speed reduction required for a following vehicle to evade a potential collision with a leading vehicle (Zheng & Sayed, 2019). To compute the DRAC, the difference between the speeds of the two vehicles is divided by their closing

time. This method considers the approaching speed and the speed difference between the two vehicles. The DRAC is useful for determining the required deceleration rate to avoid a collision and can provide guidance for driving practices and safety decisions. If the DRAC is high, a driver would need to decelerate rapidly to avoid a collision, which may not always be feasible or safe. Zheng and Sayed (2019) further explain that DRAC is a valuable tool for assessing collision risks and improving overall driving safety. The equation for DRAC (Paul & Ghosh, 2021) is as follows:

$$DRAC = \frac{(\Delta s^2)}{2*[x_l - x_f - D_l]} \quad (2.4)$$

Where, x is the position of front bumper of vehicle (f = the following vehicle, l = the lead vehicle), D_l is the length of leading vehicle, Δs is the relative speed of vehicles

Due to the limitations of crash-based analysis, recent studies are using surrogate safety measures as an alternative (Amini et al., 2022; Song et al., 2022; Orsini et al., 2021; Shangguan et al., 2021; Zheng & Sayed, 2019). Nadimi et al. (2022) developed a statistical method for comparing surrogate safety measures (SSMs) using the Collision Probability index, which is a combination of SSMs indicating the likelihood of a rear-end collision. The study utilized microscopic traffic data from the NGSIM website and emphasized the importance of selecting and grouping appropriate SSMs for safety evaluations. Beauchamp et al. (2022) utilized surrogate safety measures to assess the safety of automated shuttles and found that they have higher post-encroachment time and time-to-collision than motorized vehicles, indicating improved safety. However, the study also identified concerns regarding the higher speed difference and smaller time headway between automated shuttles and control vehicles. A study by Arun et al. (2021a) utilized indicators including MTTC, TTC, and ΔV to model crash frequency by severity. The authors employed a bivariate extreme value model with (MTTC and ΔV) and (TTC and ΔV), which outperformed the univariate peak over threshold model in terms of accuracy.

Arun et al. (2021b), in the review paper on traffic conflict-based safety measures discussed the various categories of indicators like temporal proximity indicator, kinematic indicator, spatial proximity indicators and their applications in various context. This paper also presents the future advancements in safety studies, the need for validation techniques and also the need for indicators for vulnerable road users. Bonela and Kadali (2022) reviewed the use of SSMs in evaluating the safety of T intersections under heterogeneous

traffic conditions. The review found that many previous studies failed to account for driver behaviour, which can impact the selection and performance of SSMs. The study recommends combining TTC and PET to better evaluate traffic safety, and highlights the need for new composite indicators to capture the complexity of driver behaviour. While previous studies on conflict indicators have mostly utilized one or two indicators (Pawar et al., 2022; Gastaldi et al., 2021; Goyani et al., 2021; Wang et al., 2021), Arun et al. (2022) developed an extreme value copula model to investigate the use of multiple conflict indicators in finding the probability of crash occurrence. This study used MTTC, DRAC, PSD and Delta-V to identify the rear end crashes at a signalized intersection. The results revealed that the performance of this model depend on the choice of the conflict indicators.

2.4 SUMMARY

The use of traffic conflict-based safety measures has become increasingly popular in recent years. These measures aim to identify and analyse situations where road users are at a high risk of collision, even if no collision has occurred. Various types of conflict indicators, such as temporal proximity, kinematic, and spatial proximity indicators, have been developed and used in research. The study has highlighted important research areas that need to be addressed in the future for the evaluation of safety based on conflicts. One of the key gaps identified is the lack of research on conflicts in unsignalized crossroads and roundabouts. To improve our understanding of highway safety through conflict-based analysis, it is crucial to use conflict measures that are appropriate for the context and their respective thresholds. Certainly, many safety studies in the past have focused on driver behaviour, environmental factors, and traffic flow characteristics to model crash risk. However, there is a growing need to incorporate individual vehicle trajectory data and interactions between vehicles into these models. This will enable a better understanding of the underlying mechanisms of crashes and help to identify effective countermeasures. Additionally, recent studies have highlighted the importance of using multiple conflict indicators and surrogate safety measures to better evaluate and improve highway safety. Addressing these gaps in the literature and utilizing appropriate measures can provide valuable insights for future research and help to reduce the incidence of traffic crashes.

CHAPTER 3

METHODOLOGY

3.1 GENERAL

This study focuses on developing a machine learning model to predict rear-end conflict risk using vehicle trajectory data. The research involves various stages, such as data collection, extraction, cluster analysis, pre-processing, feature selection, implementation of five machine learning models, data analysis using performance metrics, and interpretation of feature importance using SHAP. The subsequent sections will provide a detailed explanation of each stage. The complete methodological flowchart is presented in the figure 3.1.

3.2 SELECTION OF THE SITE AND DATA COLLECTION

The research is proposed to be conducted at an uncontrolled intersection due to the significant number of vehicle interactions such as merging and diverging that occur there. Initially, the plan was to select intersections based on accident records to identify those with the highest number of accidents. This led to shortlisting three intersections, including Palathara, Thattamala, and Pallimukku. However, these intersections presented major problems, as the intersection influence area was blocked by trees, creating occlusion that would hinder tracking of vehicles. Additionally, the widening activities through the intersections would disrupt normal traffic flow. Thus, the top three accident-prone intersections could not be selected. As an alternative, the Chemmamukku intersection was chosen, as it experiences a high volume of traffic on both major and minor roads. According to data on traffic accidents, this staggered intersection has one of the top occurrences of accidents. Furthermore, the intersection was chosen due to the convenient presence of a foot over bridge located at an average height of 5.5 meters, which facilitates uninterrupted recording of traffic videos. The study location is depicted in figure 3.2. The data was gathered from the junction during the busiest times of the day (9-11 a.m. and 4-6 p.m., respectively).The camera used for the study is Sony Handycam with maximum resolution 2 MegaPixels. The camera was placed on the footbridge at a height of approximately 6m from the ground level to record the conflicts between through moving vehicles in the major road.

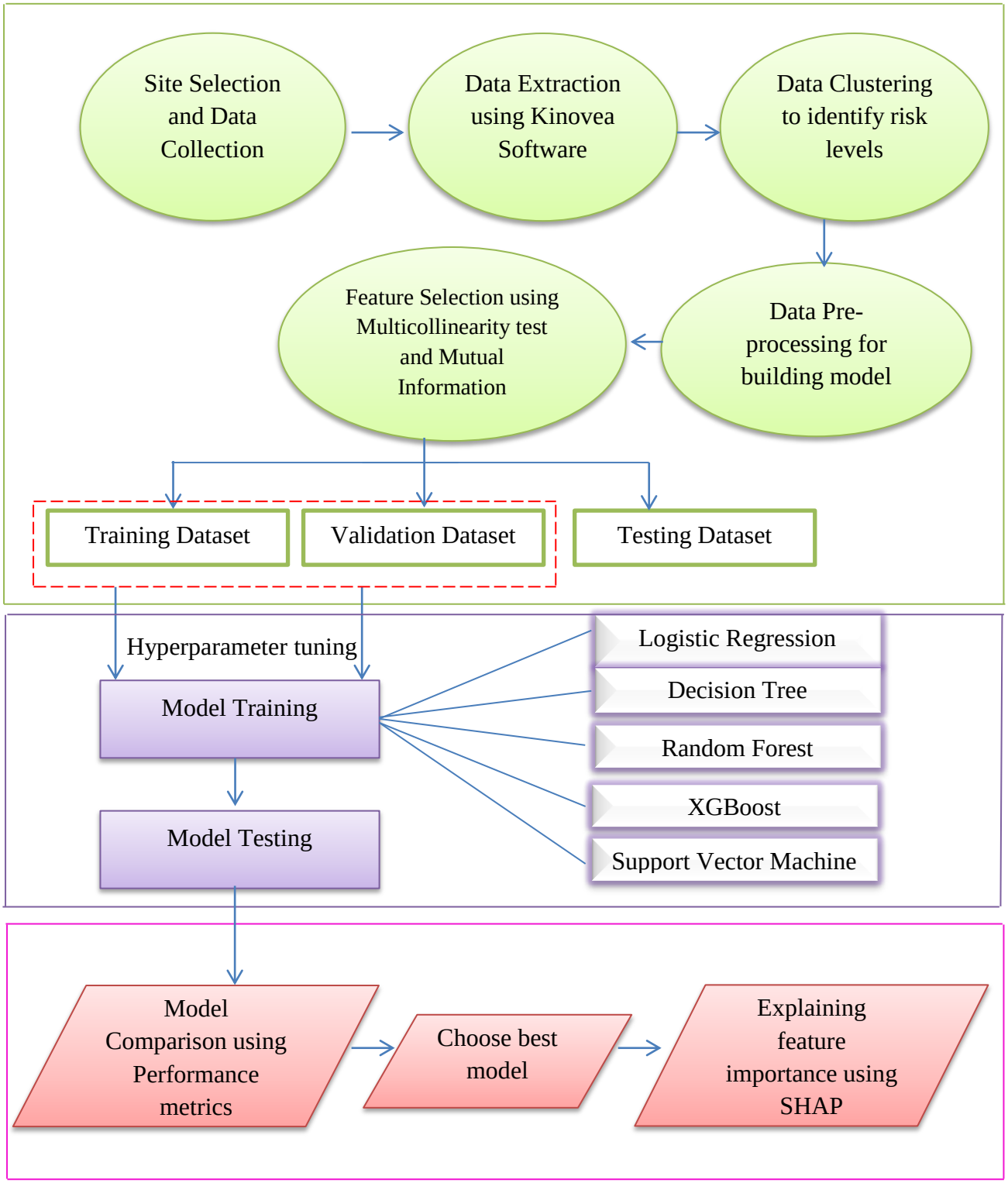

Figure 3.1 Complete methodological framework

Figure 3.2 Study Area

3.3 SELECTION OF CONFLICT INDICATORS

Based only on rear-end collisions as the form of conflict under consideration in this study, conflict indicators were chosen accordingly. As a result, the metrics to be employed were MTTC, DRAC, and PSD. Although TTC is frequently used as a measure of conflict intensity, MTTC was selected for this study because it more accurately captures rear-end conflicts. This is due to the fact that MTTC addresses a TTC constraint by accounting for the mutual separation, velocity, and acceleration/deceleration of the involved vehicles (Paul & Ghosh, 2021). TTC was used initially for identifying the safety relevant interactions, and the classification into risk levels was performed on those safety relevant interactions using MTTC, DRAC and PSD (Arun et al., 2022).

Many recent studies have advocated for the use of multiple indicators to evaluate safety, as relying on a single indicator only provides a partial image of the overall security picture. All security implications of a traffic interaction can be captivated by using a number of conflict indicators. Temporal proximity indicators refer to the degree of closeness in time between the conflicting participants, while spatial proximity indicators

measure the physical separation between them. Kinematic indicators, on the other hand, capture the relevant properties that dictate how the involved parties respond to the conflict. As they collectively take into account the factors related to the motorist, the automobile, and the environment, these various types of indicators can be combined to provide a more thorough assessment of a road safety issue (Arun et al., 2022). Therefore, for this study, three conflict indicators were employed, including one temporal proximity indicator (MTTC), one spatial proximity indicator (PSD), and one kinematic indicator (DRAC). These indicators were deemed appropriate for assessing rear-end conflicts, as reported by Wang et al. (2021).

3.4 EXPLANATORY VARIABLES

The dependent variable for this study is the risk level which is obtained by clustering the three traffic conflict indicators. The independent variables considered are

3.4.1 Variables related to leader vehicle

- Average speed (ASL)
- Standard deviation of speed (SDSL)
- Maximum longitudinal acceleration (MAL)
- Maximum longitudinal deceleration (MDL)
- Standard deviation of acceleration (SDAL)
- Type of vehicle (LEAD)
- Occupancy time (LOCC)

3.4.2 Variables related to follower vehicle

- Average speed (ASF)
- Standard deviation of speed (SDSF)
- Maximum longitudinal acceleration (MAF)
- Maximum longitudinal deceleration (MDF)
- Standard deviation of acceleration (SDAF)
- Type of vehicle (FOLL)
- Occupancy time (FOCC)

3.4.3 Variables related to the interaction between leader and follower vehicle

- Average speed difference (ASD)

- Standard deviation of speed difference (SDSD)
- Average longitudinal spacing (MSPA)
- Standard deviation of longitudinal spacing (SDSP)
- Time headway (HW)
- Interaction time (IT)

3.4.4 Other variables

- Presence of right turn vehicles from major road (Nominal) (RTMA)
- Presence of right turn vehicles from minor road (Nominal) (RTMI)
- Time of day (Nominal) (TIME)
- Number of vehicles in the major stream during vehicle interaction (Continuous) (FLOW)

3.5 EXTRACTION AND CALCULATION OF CONFLICT INDICATORS

The method for extracting and estimating conflict indicators from video recordings of a junction is described in this section. The use of the Kinovea software made it possible to extract the location coordinates, velocity, and acceleration of every vehicle inside the research area that had been determined to be a leader-follower pair. The study area was defined as a region extending 55 meters upstream and downstream of the intersection, as this area was found to represent the zone where vehicles began to decelerate and subsequently regain speed (Chauhan et al., 2021).

To extract the necessary data, a camera calibration was initially performed to convert image coordinates into real-world coordinates. This involved setting up a rectangular grid of known dimensions within the study area. Vehicles were then tracked within the study area, with a bounding box appearing around each vehicle as it entered the area. The screenshot of trajectory data extraction using Kinovea Software is shown in figure 3.3. Leader-follower pairs were identified as those vehicles that had a partial or complete overlap with the leader vehicle laterally (Chauhan et al., 2021). Once the vehicle left the study area, tracking was stopped, and the speed, acceleration, and position coordinates were exported to a separate spreadsheet.

The other explanatory variables, such as right turn vehicles from the minor and major roads, and the number of vehicles in the major stream during each interaction, were manually counted from the video.

Figure 3.3 Data Extraction using Kinovea Software

To focus on the most safety-relevant events, an initial filtering step was performed by calculating TTC for each interaction. Interactions with TTC less than 3 seconds were considered to be safety-relevant events (Arun et al., 2022; Paul & Ghosh, 2021). Three traffic conflict indicators, namely MTTC, DRAC, and PSD, were then calculated for these interactions. A total of 500 vehicle interactions were used for further analysis in this study.

For each interaction between a leader and a follower vehicle, individual speed, acceleration, and position coordinates were processed to determine the relative velocity, longitudinal spacing, and relative acceleration. The length of the leader vehicle was necessary to determine the longitudinal distance between the front end of the following vehicle and the back end of the leader vehicle. The representative vehicle dimensions were adopted from Joshi and Vagadia, 2013 and Chandra, 2004. The length of leader vehicle was subtracted from the difference between the position coordinates, as the position coordinate of the centroid of the front bumper was obtained during tracking.

Using these measures, the MTTC, DRAC, and PSD were then calculated using their standard equations. These indicators were continuous measures, generating values for each second. The smallest MTTC, least PSD, and highest DRAC were taken from each interaction between a leader and a follower vehicle. The vehicle trajectory data has also been employed to calculate the other explanatory variables, such as standard deviation of speed, acceleration and others.

3.6 CLUSTERING OF TRAFFIC CONFLICT INDICATORS

After the data extraction, next step is cluster analysis because the severity levels are unable to be established from individual measurements of conflict indicators. Therefore, cluster analysis was carried out with the objective of having an accurate representation of the degree of intensity of conflicts (Kumar et al., 2019). A form of unstructured learning approach called clustering consists of organising data points that are linked into groups based on some sort of similarity metric. Finding natural groups or patterns in the data without any prior knowledge of the class labels or classifications is the aim of clustering.

In this study, the silhouette score approach is utilized to determine the ideal number of clusters. When performing clustering analysis on a dataset, an approach called the silhouette score method is applied to figure out the ideal number of clusters. The cohesion and separation metrics are used to produce the silhouette score, which assesses the extent to which each data point matches with the cluster to which it has been assigned. Separation indicates how distinct an observation is from observations in other clusters; whereas cohesion measures how closely connected an observation point is to the other observation within the same group. An observation's silhouette score can vary from -1 to 1, with 1 denoting that the observation is well suited to its cluster and -1 denoting that it would be appropriate for a different cluster. The average silhouette score for all available data is computed to get the silhouette score for a certain clustering outcome. The most effective number of clusters is the one that yields the greatest average silhouette score, demonstrating that the data points are well-matched to the clusters to which they were allocated and that there is a fair degree of separation between the clusters.

This study employs three different clustering techniques, namely K-means clustering, Spectral Clustering, and Agglomerative Clustering. K-means clustering is a commonly used unsupervised learning algorithm that divides data into K groups based on

resemblance across observations. The algorithm iteratively allocates each data point to the closest centroid and updates the centre point on the basis of the average of all the data points given to it. The algorithm repeats this process until convergence, where the data points' assignments to clusters no longer change. K-means clustering initially requires the selection of K cluster centres, which are randomly chosen from the dataset. The algorithm then assigns data points to their nearest centroids based on a distance metric, such as Euclidean distance, and groups the closest data points together to form a cluster. Once all data points have been assigned to their nearest centroids, the centroids are revised by calculating the average of all data points in each cluster, which becomes the new centroid. The procedure is then repeated until convergence.

Next is the unsupervised machine learning method called spectral clustering, that combines data points according to their similarity. It is a graph-based algorithm that employs the spectral decomposition of the similarity matrix to identify the underlying data structure. The first step in this algorithm is to construct a similarity graph by defining a similarity metric between each pair of data points. While the similarity metric is often based on the Euclidean distance, other similarity measures are also available. The structure of the similarity network is represented by a Laplacian matrix, and its computation is the second step. The spectral decomposition method is then used to obtain the eigenvalues and eigenvectors of the graph Laplacian. The eigenvectors are used to embed the data points into a lower-dimensional space, where the clusters can be easily separated. The number of clusters in the data will determine how many eigenvectors are needed for embedding. Finally, the data points are partitioned into clusters using a clustering algorithm such as K-means, which is done in the lower-dimensional space obtained by embedding the data points using the eigenvectors.

Agglomerative clustering is a popular hierarchical clustering algorithm that aims to group similar data points together by merging smaller clusters into larger ones. Initially, each data point is considered as a separate cluster, and the algorithm calculates the distance or similarity between each pair of data points using a distance metric such as Euclidean distance or cosine similarity. Next, the two most similar clusters are selected based on the distance or similarity between them and merged into a single cluster. Repeating this procedure results in the formation of a pre-set number of clusters or the grouping of all data points into a single cluster. As the algorithm progresses, the clusters become larger and more diverse, and the distance between them increases. The resultant dendrogram,

which resembles a tree, depicts the clusters' hierarchical connections. The dendrogram's height of each branch shows how far apart the merged clusters are from one another. These three clustering technique helps in providing labels to vehicular interactions, thus aiding in better analysis and understanding.

3.7 DATA PRE-PROCESSING

The dataset consists of dependent variable i.e., the risk level and 24 independent variables. These independent variables include numerical and categorical data. But the machine learning models cannot work with categorical data; the data needs to be transformed into a numerical format before it can be used by most machine learning algorithms. This process is called encoding, and it involves converting each categorical variable into a numerical variable that can be understood by the machine learning model.

When dealing with categorical data, there are several techniques available for encoding them. One of the most popular methods is called one-hot encoding, which involves creating a binary column for each category present in the original variable. For instance, if the original categorical variable is "LEAD" with four categories (two-wheeler, three-wheeler, four-wheeler, heavy vehicles), one-hot encoding would create four binary columns: LEAD_two-wheeler (LEAD_0), LEAD_three-wheeler (LEAD_1), LEAD_four-wheeler (LEAD_2) and LEAD_heavy vehicles (LEAD_3).

Label encoding, on the other hand, assigns a numerical value to each category in the original variable. For example, two-wheeler may be assigned the value 0, three-wheeler the value 1, four-wheeler the value 2 and heavy vehicles the value 3. However, label encoding may introduce a sense of ordinality to the variable that may not actually exist, which may affect the performance of the machine learning model. In the example given, the label encoded values suggest that three-wheeler is "greater" than two-wheeler, and four-wheeler is "greater" than both two-wheeler and three-wheeler. However, this ordering may not be applicable to the original categorical variable. For example, it may not make sense to say that a three-wheeler is "greater" than a two-wheeler in terms of any inherent characteristic or feature.

Therefore, in this study one-hot encoding was employed to convert all the categorical variables into numeric value.

3.8 FEATURE SELECTION

Feature selection is a critical step in building classification models as it can improve model accuracy, reduce overfitting, speed up model training, and increase model interpretability. In classification problems, feature selection can help identify the most relevant features that contribute to predicting the target variable and reduce the number of irrelevant or redundant features. There are multiple approaches to feature selection, including filter methods, wrapper methods, and others (Tang et al., 2014).

Filter methods, in particular, do not rely on a machine learning algorithm and assess each feature independently. This feature independence allows filter methods to be used with any machine learning algorithm. Additionally, filter methods are easy to interpret because they use statistical techniques to measure the relationship between each feature and the target variable. As a result, this study employs the filter method for feature selection.

Before applying filter methods for feature selection, it is essential to perform multicollinearity testing. When multiple independent variables in a dataset have a high degree of correlation, this phenomenon is known as multicollinearity, which can result in unreliable parameters, make it challenging to understand model findings, and increase model complexity (Ye et al., 2023). To detect multicollinearity, a correlation matrix of all independent variables in the dataset is calculated, and one of the correlated features is removed. Since the dataset consists of both continuous and binary variables, different correlation analysis techniques are used. Spearman Correlation Coefficient is used to evaluate the association between continuous variables, Cramer's V coefficient for the association between binary variables, and the Point Biserial Coefficient for the association between continuous and binary variables (Shangguan et al., 2023; Ye et al., 2023).

After eliminating the features that showed multicollinearity, the next step is to perform feature selection using mutual information (MI) analysis. A statistical tool for analysing the connection between two variables is the MI. In the case of feature selection, MI analysis entails calculating the MI between each independent variable and the target variable, and selecting the features with the highest MI values. MI analysis can identify features that are highly relevant to predicting the target variable, even if they are not strongly correlated with other features.

Both of these methods for feature selection can be effective in enhancing model accuracy and reducing overfitting.

3.9 SPLITTING DATA AND HYPERPARAMETER TUNING

The training set and the testing set are two subsets that are frequently separated from the dataset in the field of machine learning. This split's objective is to assess the model's performance using hypothetical data. In contrast to the training set, which is used for tutoring the algorithm, the testing set is used to test the model's efficacy on data that it hadn't encountered previously. In this study, a 70-30 split was used, where 70% of the data was used for tutoring the algorithm and 30% was used for testing the quality of the model.

In machine learning, it is essential to pre-process the input data before training a model to ensure that each feature contributes equally to the model's performance. This step involves scaling the features, which can be achieved using techniques like StandardScaler or MinMaxScaler. StandardScaler transforms the data to have zero mean and unit variance, while MinMaxScaler scales the data to a given range, usually between 0 and 1. The choice of scaling technique depends on the distribution of the data and the performance of the model. In this study, MinMaxScalar was used to scale data for Logistic Regression, and StandardScalar was used for the remaining classifiers.

After scaling the data, the next step is to select a suitable supervised learning algorithm from a range of options, such as logistic regression, decision tree, random forest, support vector machine, or XGBoost. Once an algorithm is chosen, the next step is to optimize its hyperparameters. Hyperparameters are set by the user and affect the behaviour of the algorithm. Examples of hyperparameters include the maximum depth of tree in random forest, solver algorithm in logistic regression, learning rate in boosting algorithm, and kernel type in support vector machine. Defining a search space for hyperparameters involves specifying the range or set of possible values for each hyperparameter.

Grid search is a technique to find the best combination of hyperparameters for a machine learning model (Ijaz et al., 2021). It evaluates all possible combinations of hyperparameters in a search space. K-fold cross-validation, which divides the set of training data into k equally sized portions or folds, is used to assess the models. The model is tested on the remaining fold after being trained on k-1 folds. Each fold acts as

the validation set once for the subsequent k iterations of this operation. The performance metric for the model is the average validation score over the k-folds (Hosseinzadeh et al., 2021). The model that performs best is selected in accordance with the performance measure after considering all hyperparameter combinations. This model is then evaluated on a testing set to estimate its performance on new, unseen data. The testing set simulates the real-world scenario where the model encounters new data that it has not been trained on.

3.10 CLASSIFICATION MODEL

In this study, we employed five classification models to accurately classify risk levels. They are:

3.10.1 Logistic regression

Logistic regression is a classification algorithm used to predict the probability of an event occurring based on input variables. Multiclass classification can be achieved using multinomial logistic regression or softmax regression. The latter works by calculating a score for each class based on input features and transforming the scores into probabilities using the softmax function. During training, the model's parameters are learned by minimizing a loss function, such as cross-entropy, between predicted probabilities and true labels, through optimization techniques like stochastic gradient descent. The trained model can predict new input data by calculating scores, applying the softmax function to obtain probabilities, and selecting the class with the highest probability as the predicted class. The logistic regression plot separating three classes is shown in figure 3.4.

The benefits of logistic regression include ease and comprehensibility which make it simple to interpret and articulate the model's predictions. It works well when variables have linear correlations to one another and is computationally effective when dealing with big datasets. However, because logistic regression implies linearity, it may have trouble with interactions or non-linear correlations between variables. It is susceptible to outliers and multicollinearity. In summary, the logistic regression model for multiclass classification is a powerful tool for predicting the probability of an event belonging to multiple classes. By utilizing the softmax function and minimizing a loss function during the training process, the model can learn to make accurate predictions for new input data.

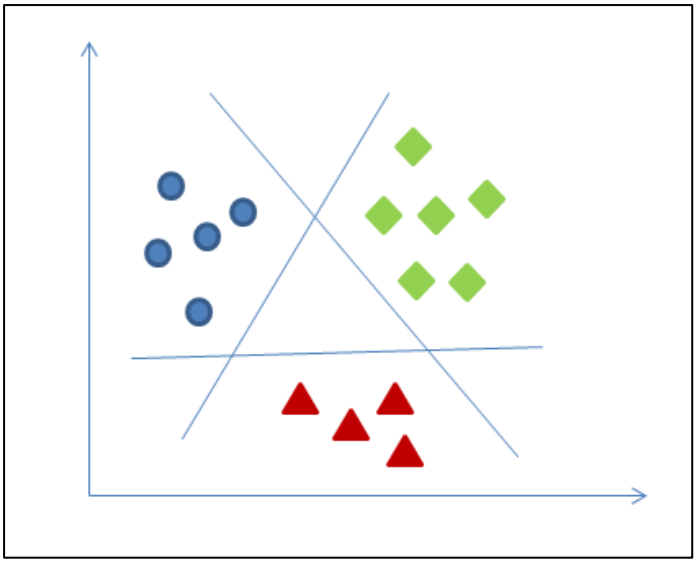

Figure 3.4 Logistic Regression

3.10.2 Decision Tree

Decision trees are a type of classification algorithm that use a tree-like model to make predictions. The procedure begins by iteratively breaking up the dataset into smaller subgroups until a halting requirement is satisfied. The split attribute is chosen based on criteria like information gain or Gini impurity, and each node in the tree reflects a decision depending on one or more input characteristics (Jamal et al., 2021). Once the model is built, it can be utilized to forecast the class for future samples, by advancing the tree from the initial node to a leaf node. Decision trees have advantages such as interpretability, the ability to handle both categorical and numerical data, and the ability to form ensemble methods. However, they can be sensitive to small variations in the data, which can lead to overfitting. To avoid overfitting, pruning techniques can be used to remove unnecessary branches from the decision tree. Overall, decision trees are a useful and powerful tool for classification tasks, but care must be taken to avoid overfitting and ensure that the model is accurate and interpretable. The sample decision tree classification is shown in figure 3.5.

3.10.3 Random Forest

Random forest is an ensemble learning algorithm for classification tasks that builds multiple decision trees on randomly selected subsets of the input data and input features (Shangguan et al., 2021). By aggregating the predictions of these trees, the algorithm can improve the accuracy and robustness of the model. During prediction, the random forest

model takes a majority vote of the predicted class by each decision tree. This algorithm can handle high-dimensional data , deal with missing data and outliers, and provide an estimate of feature importance. However, it can be computationally expensive to build a large number of decision trees, and the model may be less interpretable than single decision trees. The schematic representation of random forest classification is shown in figure 3.6.

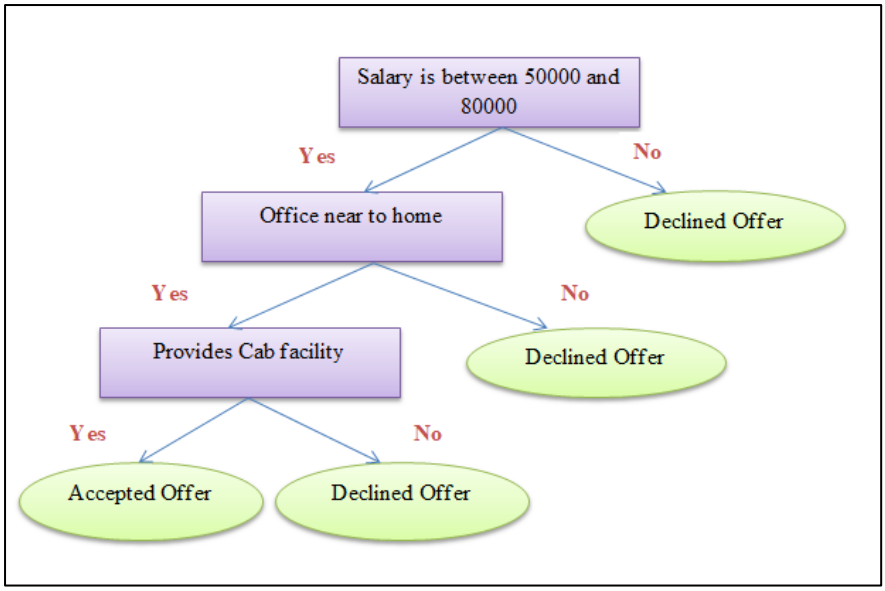

Figure 3.5 Decision Tree Classifier

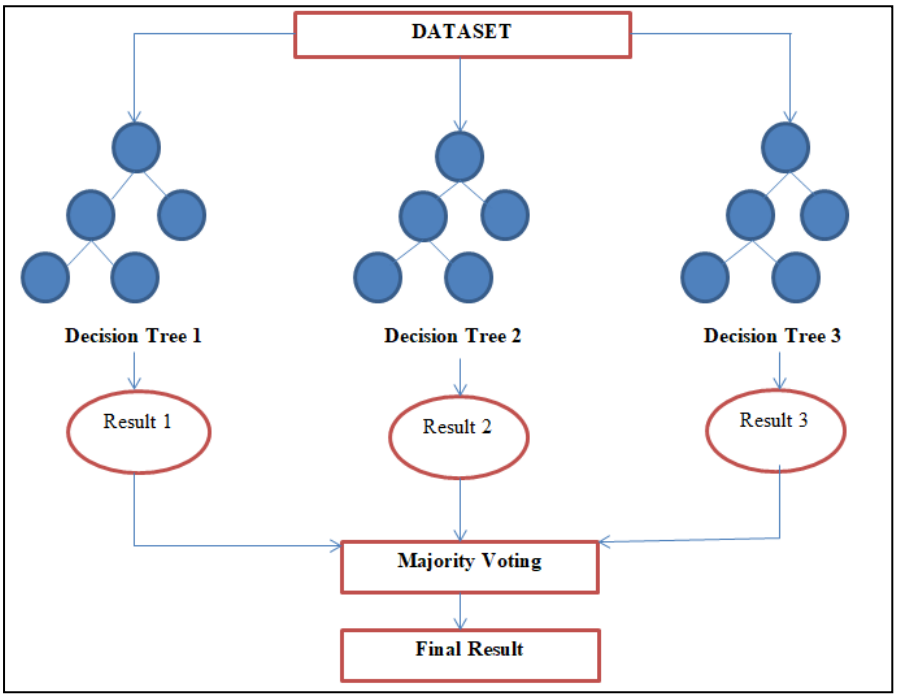

Figure 3.6 Random Forest Classifier

3.10.4 XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a popular supervised learning algorithm used for classification and regression tasks. It fits decision trees on the training data and iteratively updates them to minimize the loss function (Shangguan et al., 2021). XGBoost starts with a shallow tree and calculates the difference between the predicted value and actual value to train the next decision tree. It repeats this process iteratively, optimizing the model by adding new decision trees and updating the weights of the training examples. XGBoost also includes regularization and parallel processing to prevent overfitting and speed up the training process. Overall, XGBoost yields high accuracy and precision by using gradient descent and boosting techniques to improve predictions. The schematic representation of XGBoost classification is shown in figure 3.7.

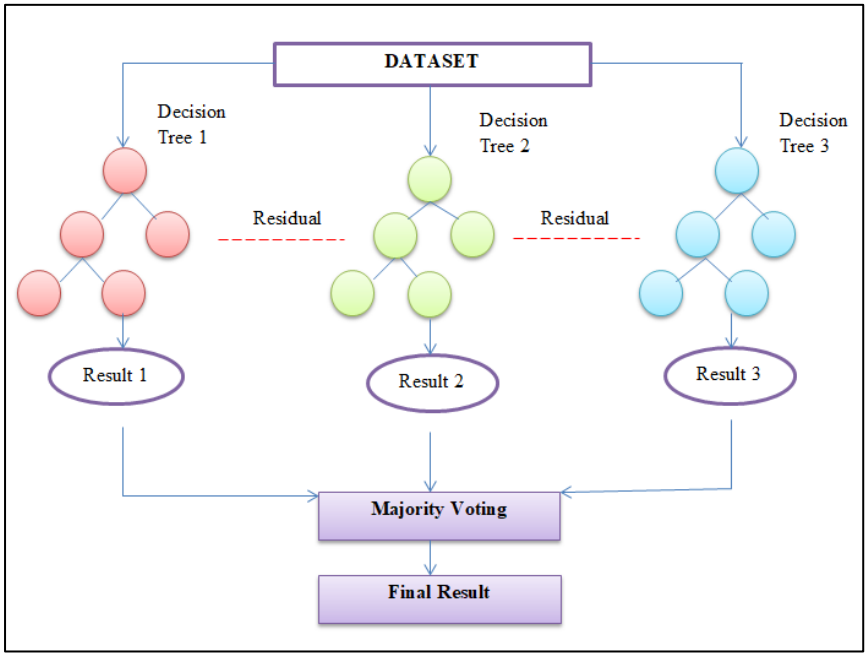

Figure 3.7 XGBoost Classifier

3.10.5 Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm used for classification tasks. It finds the hyperplane that best separates classes in the data. For multiclass classification, SVM uses a one-vs-all (OvA) approach by training multiple binary classifiers. SVM calculates the hyperplane that maximizes the margin between classes and uses a decision rule to determine the class of a new input data point (Hosseinzadeh et al., 2021). SVM can handle high-dimensional data and outliers, but has limitations such

as sensitivity to kernel function choice and high computational complexity. Modifications and extensions have been proposed to address these limitations, including optimizing kernel functions and speeding up training. SVM remains a widely used algorithm for classification tasks. The basic idea behind support vector machine classification is shown in figure 3.8.

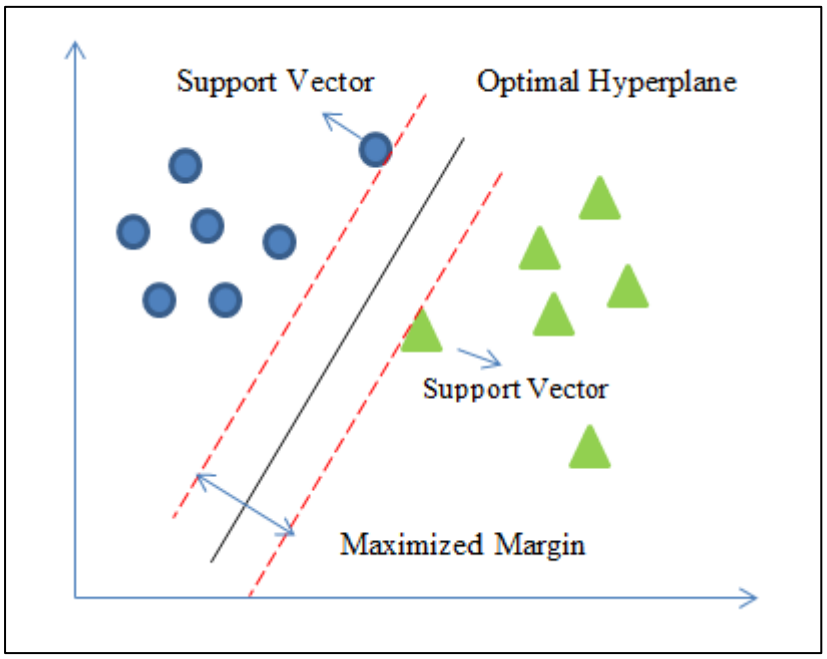

Figure 3.8 Support Vector Machine Classifier

3.11 MODEL COMPARISON USING PERFORMANCE METRICS

In order to evaluate the effectiveness of the model, performance metrics is used. These metrics provide a quantitative measure of how well the model is performing and can help identify areas for improvement. A confusion matrix, shown in figure 3.9, is used to assess the effectiveness of a classification model by displaying the number of accurate and inaccurate predictions generated by the model in comparison to the actual results (Tamim et al., 2022). The confusion matrix is typically a 2x2 matrix, but it can be larger for multi-class classification problems. The matrix displays True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) generated by the model on the testing dataset.

- True Positive (TP): The number of cases where the actual class is positive (1) and the predicted class is also positive (1).

- True Negative (TN): The number of cases where the actual class is negative (0) and the predicted class is also negative (0).

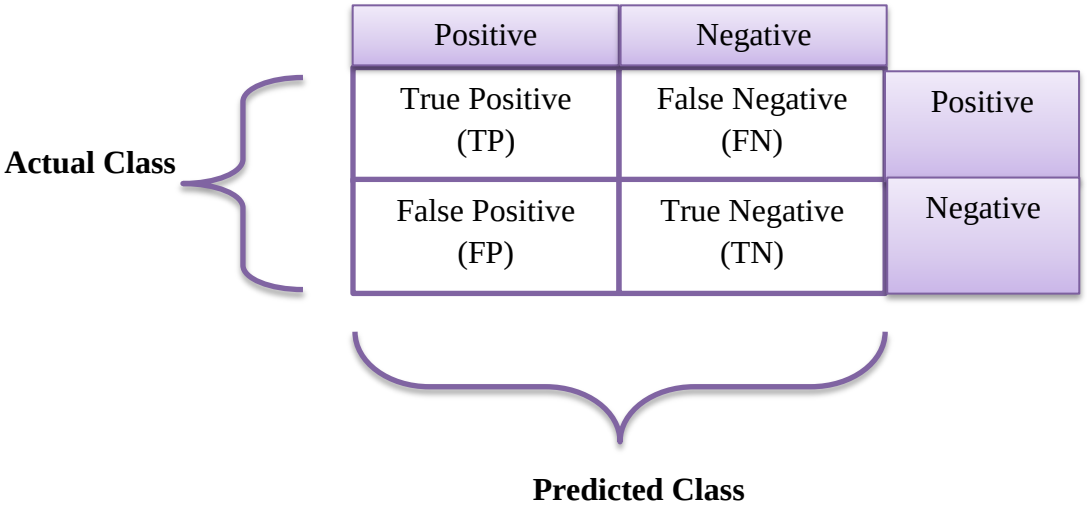

Figure 3.9 Confusion Matrix

- False Positive (FP): The number of cases where the actual class is negative (0) but the predicted class is positive (1). This is also known as a Type I error or a false alarm.

- False Negative (FN): The number of cases where the actual class is positive (1) but the predicted class is negative (0). This is also known as a Type II error or a miss.

In machine learning, we aim to optimize our models to minimize false positives and false negatives, and maximize true positives and true negatives. However, the relative importance of each metric can vary depending on the specific problem and associated costs for each type of error.

For multiclass classification problems, the confusion matrix is used to compute these metrics. TP represents the number of correctly predicted instances for a specific class. FN is computed by summing up the values in the corresponding row, except for the TP value (Zhang et al., 2022). FP is computed by summing up the values in the corresponding column, except for the TP value (Zhang et al., 2022). TN is computed by summing up the values in all other rows and columns, except for the values of the class for which we are calculating the metrics.

Some of the commonly used performance metrics are accuracy, precision, recall, F1 score, false positive rate, and false negative rate.

- **Accuracy:** Accuracy is the ratio of the correctly classified samples to the total number of samples (Ijaz et al., 2021). It is the most commonly used metric for evaluating the performance of a classification model.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (3.1)$$

- **Precision:** Precision is the ratio of the true positives to the total number of positive predictions. It measures the ability of the model to correctly classify the positive samples. A high precision score indicates that the model is making fewer false positive predictions (Ijaz et al., 2021).

$$Precision = \frac{TP}{TP+FP} \quad (3.2)$$

- **Recall:** Recall is the ratio of the true positives to the total number of actual positive samples. It measures the ability of the model to identify all the positive samples. A high recall score indicates that the model is making fewer false negative predictions (Ijaz et al., 2021).

$$Recall = \frac{TP}{TP+FN} \quad (3.3)$$

- **F1 score**: F1 score is the harmonic mean of precision and recall (Dong et al., 2022). It is a good metric to use when the dataset is imbalanced.

$$F1\ Score = \frac{1}{Precision} + \frac{1}{Recall} \quad (3.4)$$

- **False positive rate**: False positive rate is the ratio of false positives to the total number of actual negative samples (Ma et al., 2021). It measures the proportion of actual negatives that are incorrectly classified as positive.

$$FPR = \frac{FP}{FP+TN} \quad (3.5)$$

- **False negative rate**: False negative rate is the ratio of false negatives to the total number of actual positive samples (Ma et al., 2021). It measures the proportion of actual positives that are incorrectly classified as negative.

$$FNR = \frac{FN}{FN+TP} \quad \text{..... (3.6)}$$

In addition to the performance metrics, ROC AUC (Receiver Operating Characteristic Area Under the Curve) curve is used for evaluating the performance of model and it can be particularly helpful for comparing different machine learning models. This is due to the fact that it offers a single value that sums together the performance across all possible classification thresholds (Komol et al., 2021). This makes it a helpful indicator for examining the effectiveness of various models, especially when you have multiple evaluation metrics to consider. A sample ROC AUC curve is shown in figure 3.10. The ROC AUC curve plots the true positive rate (TPR) against the false positive rate (FPR) for different classification thresholds, which allows you to visualize and compare the trade-offs between sensitivity and specificity for different models (Kashifi & Ahmad, 2022). For random performance, the curve crosses over the straight diagonal and achieves an AUC of 0.5. AUC-ROC is close to 1 if the ROC plot passes towards the upper left corner; a value of AUC = 1 denotes flawless performance (Zhang et al., 2022).

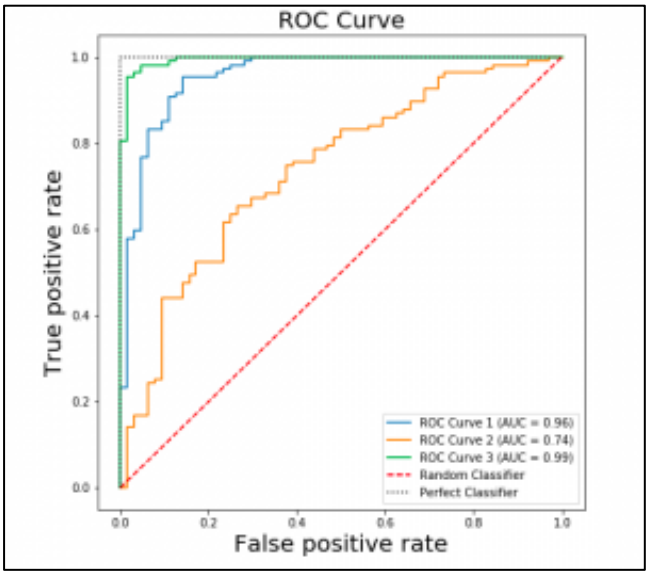

Figure 3.10 ROC AUC Curve (Source: https://www.medium.com)

3.12 FEATURE IMPORTANCE USING SHAP

SHAP (SHapley Additive exPlanations) is a strategy for the better understanding of the results of machine learning models. It assigns importance scores to the input features that contribute to each prediction (Islam & Abdel-Aty, 2023; Yuan et al., 2022b). The Shapley value of a feature calculates the average marginal contribution of that feature to the model's output, by comparing the model's output with and without that feature. It takes the average of these differences to determine the feature's contribution.

SHAP (SHapley Additive exPlanations) is a framework used to interpret the output of machine learning models by assigning importance scores to input features. The framework utilizes the Shapley value, a concept from cooperative game theory, to determine the contribution of each feature to the model's prediction. In cooperative game theory, the Shapley value measures the contribution of each player to the total payoff of a coalition. In the context of machine learning, the input features are the "players," and the model's prediction is the "payoff."

The Shapley value was first introduced by Lloyd Shapley in the 1950s as a way to fairly distribute the profits from a cooperative game among the players. In a cooperative game, the players work together to achieve a common goal, and the payoff depends on the contributions of all players. The Shapley value is a way to divide the payoff among the players in a fair and efficient manner, based on their individual contributions.

In the context of machine learning, the idea of using the Shapley value to assign feature importance scores was first proposed by Saabas in 2014. However, the SHAP framework introduced by Lundberg and Lee in 2017 extended this idea to work with a wider range of machine learning models, including decision trees, neural networks, and others.

The SHAP framework utilizes the Shapley value to assign importance scores to input features in a machine learning model (Dong et al., 2022). These importance scores can be used to interpret the output of the model for a particular instance or to rank the features by their overall importance across all instances. To calculate the Shapley value for a specific feature, we must determine its marginal contribution to the model's output for all possible combinations of the remaining features. These marginal contributions are then averaged, with weightings based on the number of possible combinations, to obtain an unbiased and precise measure of the feature's contribution to the model's output.

Although the computation of Shapley values requires the evaluation of the model's output for every possible feature combination, there exist efficient algorithms that can provide approximations with reasonable accuracy. One such algorithm is TreeSHAP, which approximates the calculation using decision trees. Apart from being a fair way to measure feature importance, Shapley values have various other benefits. They are model-agnostic, meaning they can be applied to any type of machine learning model. Additionally, they provide local explanations for the model's prediction on a specific instance, which can enhance trust and comprehension of the model. Furthermore, Shapley values can help uncover and mitigate bias and fairness issues by revealing how different features contribute to the disparities in the model's predictions.

The summary plot in SHAP provides a visualization of the overall feature importance, ranking the features based on their impact on the model's output. The dependence plot showcases the relationship between a specific feature and the model's predictions, illustrating how changes in the feature influence the output. In the SHAP beeswarm plot, each instance in the dataset is represented as an individual point for a given variable, allowing for a detailed examination of the contribution of each instance to the feature's impact.

3.13 SUMMARY

The aim of this study is to develop a rear-end conflict risk prediction model using machine learning techniques and vehicle trajectory data. The research encompasses several stages, including data collection, extraction, cluster analysis, pre-processing, feature selection, implementation of five machine learning models, comprehensive data analysis, and interpretation of feature importance using SHAP.

In the data collection stage, vehicle trajectory data is obtained from video cameras. The conflict indicators and other independent variables are then extracted using Kinovea software and analysed using cluster analysis techniques to identify different risk levels.

After the data extraction and cluster analysis stage, the data is pre-processed and cleaned to make them compatible for machine learning modelling. Feature selection techniques using multicollinearity and mutual information are applied to identify the most relevant features that contribute to the rear-end conflict risk.

Five machine learning models, including Random Forest, Support Vector Machine, Decision Tree, XGBoost, and Logistic Regression, are implemented to predict the rear-end conflict risk based on the selected features. Comprehensive data analysis using performance metrics and ROC AUC curve is performed to evaluate efficiency of each model and identify the model that performs best.

Finally, the SHAP method is used to interpret the feature importance and provide insights into the factors that contribute to the rear-end conflict risk.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 GENERAL

The following sections present and discuss the results obtained from cluster analysis, classification modelling, and SHAP interpretation.

4.2 CLUSTER ANALYSIS

The silhouette score plot was used to identify the optimum number of clusters and is shown in figure 4.1. In this plot, average silhouette coefficient is plotted against each number of clusters. The optimum number of cluster corresponds to the highest peak in the plot which in the present case is identified as four. In practical terms, this means that the data can be segmented into four groups based on their similarities and differences in the measured variables related to conflict severity.

Figure 4.1 Silhouette Score Plot

According to the definition and characteristics of traffic conflict indicators, the likelihood of a conflict occurring is high when the MTTC and PSD indicators are low, and the severity of a conflict is greater when the DRAC indicator is high. By examining the range of traffic conflict indicators in each cluster, we can gauge the level of severity. The table 4.1 displays the range of values in each cluster, revealing that the MTTC and PSD values are low in Cluster 1, while the DRAC value is high. In contrast, as we progress through the other clusters, the MTTC and PSD values increase, while the DRAC value decreases.

This suggests that Cluster 1 represents the most severe conflict, while Cluster 4 represents the least severe conflict. Cluster 1, 2, 3, and 4 can be categorized as critical conflicts, high conflicts, medium risk conflicts, and low-risk conflicts. Out of a total of 500 data points, 118 are classified as low-risk conflict, 119 as medium-risk conflict, 154 as high-risk conflict, and 109 as critical conflict. The threshold value of MTTC for critical conflicts is 0.4, 7.13 for DRAC, and 0.47 for PSD. Interestingly, the range of each cluster remained consistent across all three clustering techniques. The validation of the clusters was accomplished using the Silhouette coefficient, which was above 0.7 for each clustering method. Previous research has shown that silhouette values ranging from 0.7 to 1.0 create a strong cluster (Kumar et al., 2019; Mohanty et al., 2021).

Table 4.1 Cluster range in different clustering technique

Cluster Technique	**K-means**	**Spectral**	**Agglomerative**
Cluster 1	MTTC: 0.01 - 0.4 DRAC: 7.13 - 9.65 PSD : 0.01 - 0.47	MTTC: 0.01 - 0.4 DRAC: 7.13 - 9.65 PSD : 0.01 - 0.47	MTTC: 0.01 - 0.4 DRAC: 7.13 - 9.65 PSD : 0.01 - 0.47
Cluster 2	MTTC: 0.42 – 1.29 DRAC: 5.00 – 7.11 PSD : 0.50 – 1.29	MTTC: 0.42 – 1.29 DRAC: 5.00 – 7.11 PSD : 0.50 – 1.29	MTTC: 0.42 – 1.29 DRAC: 5.00 – 7.11 PSD : 0.50 – 1.29
Cluster 3	MTTC: 1.32 – 2.32 DRAC: 3.23 – 4.97 PSD : 1.32 – 2.20	MTTC: 1.32 – 2.32 DRAC: 3.23 – 4.97 PSD : 1.32 – 2.20	MTTC: 1.32 – 2.32 DRAC: 3.23 – 4.97 PSD : 1.32 – 2.20
Cluster 4	MTTC: 2.38 – 3.85 DRAC: 0.02 – 3.18 PSD : 2.32 – 3.27	MTTC: 2.38 – 3.85 DRAC: 0.02 – 3.18 PSD : 2.32 – 3.27	MTTC: 2.38 – 3.85 DRAC: 0.02 – 3.18 PSD : 2.32 – 3.27

The scatter plot for the K-means clustering technique is depicted in the figure 4.2. If the clusters overlap or are challenging to differentiate, it may suggest that the clustering algorithm is not functioning effectively. However, in this case, the scatter plot shows distinct separation between the various clusters. This indicates that the clusters are well-defined and can be easily distinguished from one another. Since the cluster labels

provided by the different clustering techniques for each vehicle interaction were identical, they are utilized as the dependent variable in this study (Mohanty et al., 2021).

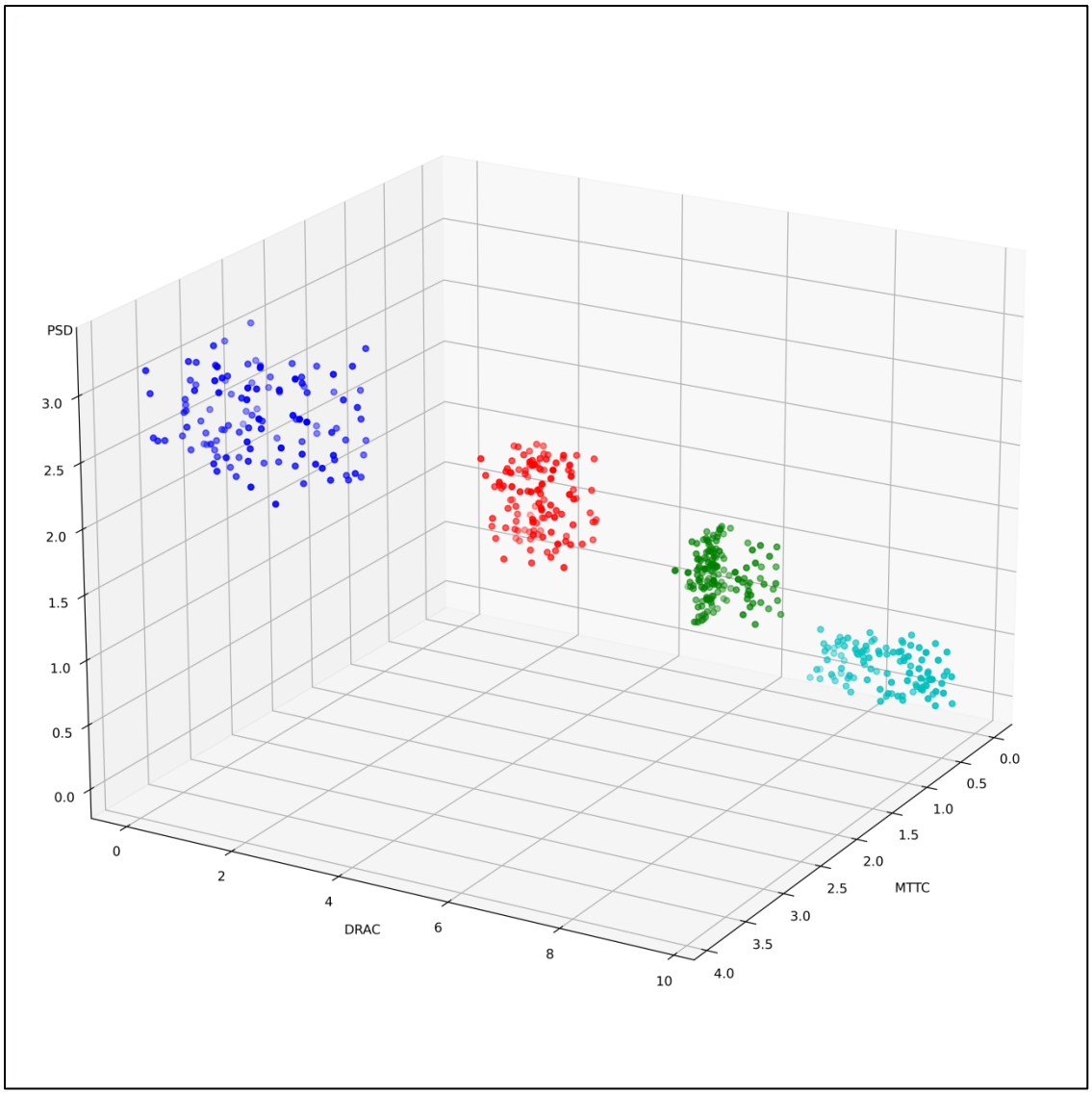

Figure 4.2 Scatter Plot in K-means Clustering Technique

4.3 FEATURE SELECTION

The feature selection was performed using multicollinearity test and mutual information. In this study, three correlation coefficients were used. A statistical indicator of the magnitude and course of the association between two continuous variables is the Spearman correlation coefficient. It is a non-parametric method that assesses the

monotonicity of the relationship, which means that it can detect both linear and nonlinear relationships between variables. The Spearman correlation coefficient ranges from -1 to 1, where -1 indicates a perfectly negative association, 0 indicates no association, and 1 indicates a perfectly positive association. Figure 4.3 represents the correlation matrix of continuous variables.

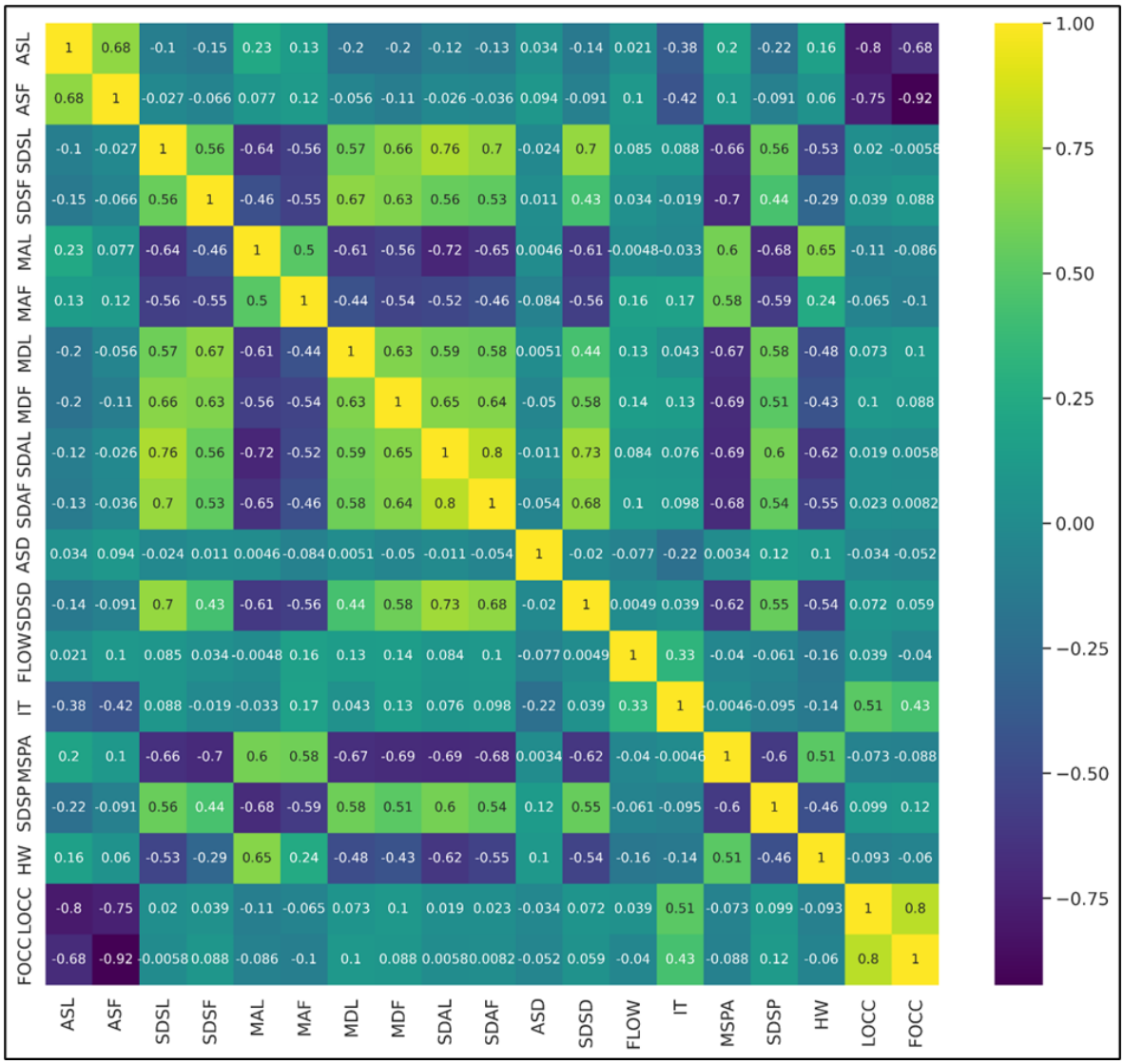

Figure 4.3 Correlation matrix of continuous variables.

A statistical indicator of the relationship between two category (dummy) variables is the Cramer's V coefficient. It has a range of 0 to 1, with 0 denoting no relationship and 1 denoting a perfect association. Cramer's V is based on the chi-squared statistic, which measures the difference between the observed frequencies and the expected frequencies under the null hypothesis of no association between the variables. Cramer's V is commonly used in contingency tables to assess the association between two categorical

variables. Figure 4.4 represents the correlation matrix between categorical or dummy variables.

A statistical indicator of the link between a continuous variable and a binary (dummy) variable is the Point-Biserial correlation coefficient. To calculate the correlation, we first compute the mean of the continuous variable for each category of the binary variable (i.e., the mean for the 0 category and the mean for the 1 category). We then compute the standard deviation of the continuous variable across the entire dataset. Finally, we calculate the correlation between the binary variable and the continuous variable using the corresponding formula.

The Point-Biserial correlation coefficient has a value ranging from -1 to 1, where -1 denotes perfect opposing relationship between the variables, 0 denotes no association, and 1 denotes perfect positive association. Figure 4.5 represents the correlation matrix between continuous and binary variable.

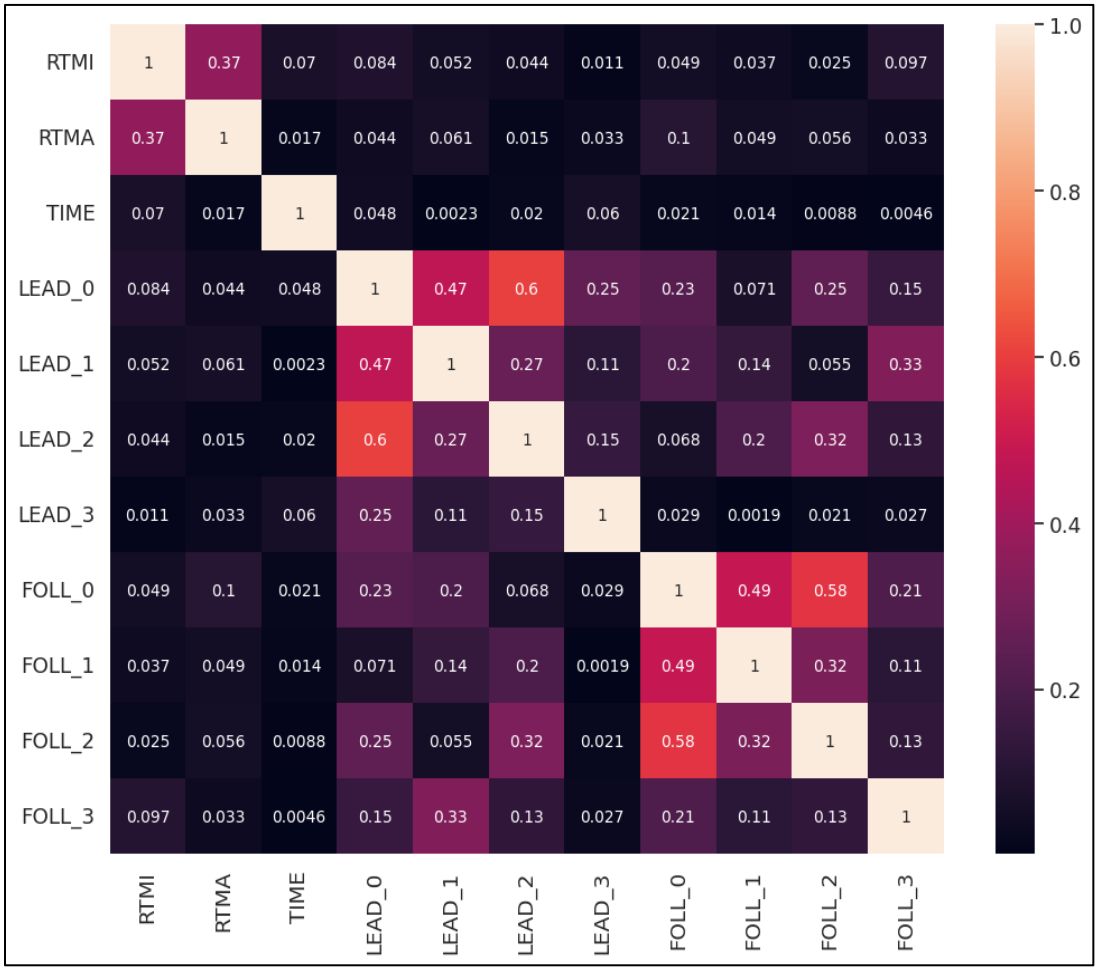

Figure 4.4 Correlation matrix of Dummy Variables

There is a strong correlation between variables when the correlation coefficient is greater than 0.7 (Shangguan et al., 2023; Ye et al., 2023). There is a perfect correlation between several variables; hence one of the correlated variables was removed from the study. As a result, LOCC, FOCC, SDSL, MAL and SDAL were eliminated from the study.

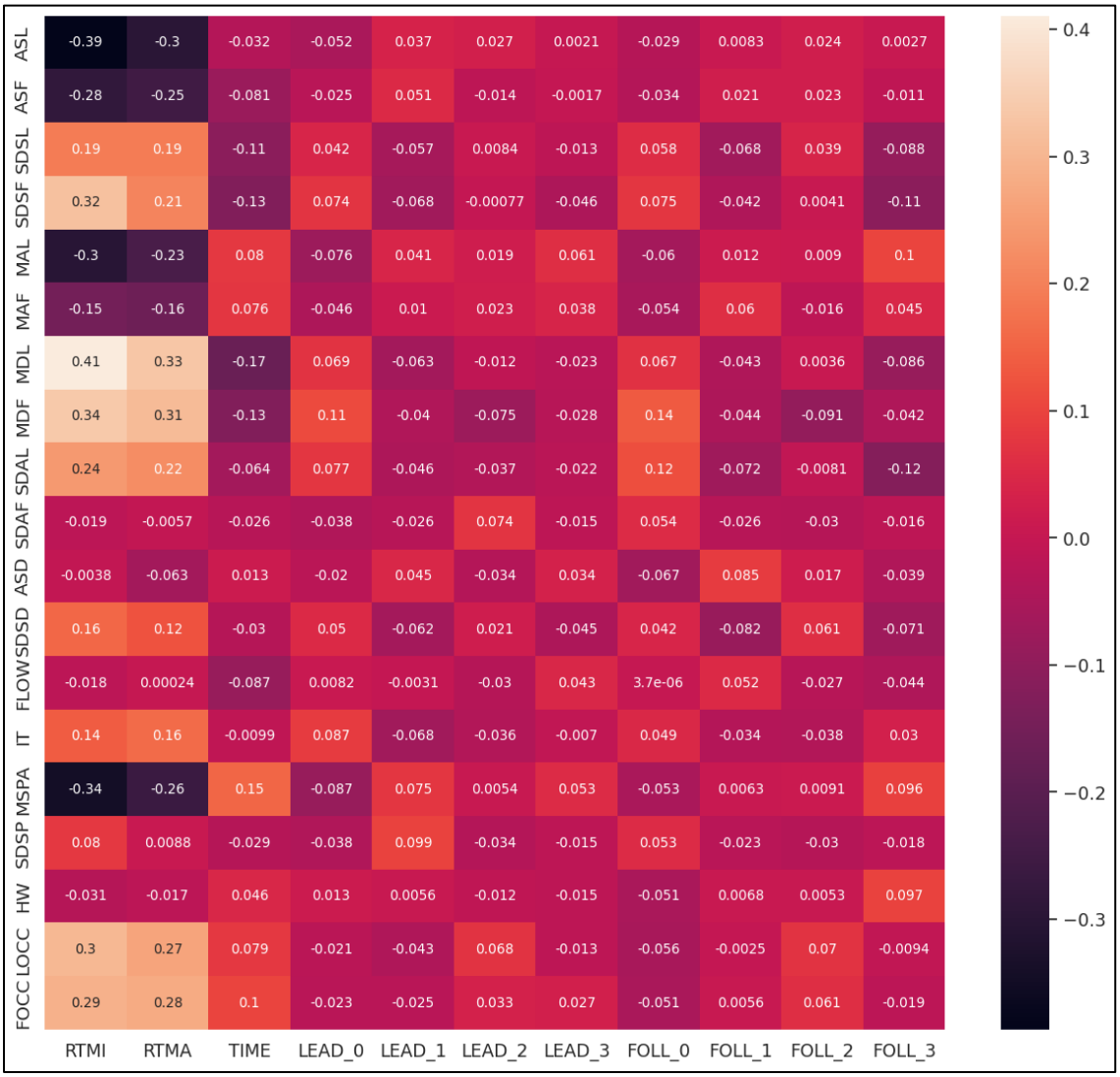

Figure 4.5 Correlation matrix between Dummy and Continuous variables

Further, we implemented mutual information to find importance of each variable. We first determine the mutual information between every attribute and the dependent variable in order to determine the feature significance utilizing mutual information. The characteristics are then ranked according to the mutual information scores, with those with higher scores being given greater weight. Figure 4.6 indicates the mutual information between features and target variable. The maximum deceleration of leader is the highest influencing variable. The top k features with the greatest mutual information

scores are often chosen using a heuristic technique when employing mutual information for feature selection or feature engineering. The value of k can be chosen based on domain knowledge or by using a validation set to evaluate the model's performance with different numbers of selected features. In this study, top 11 features were selected based on how well the model performs with each combination of variables.

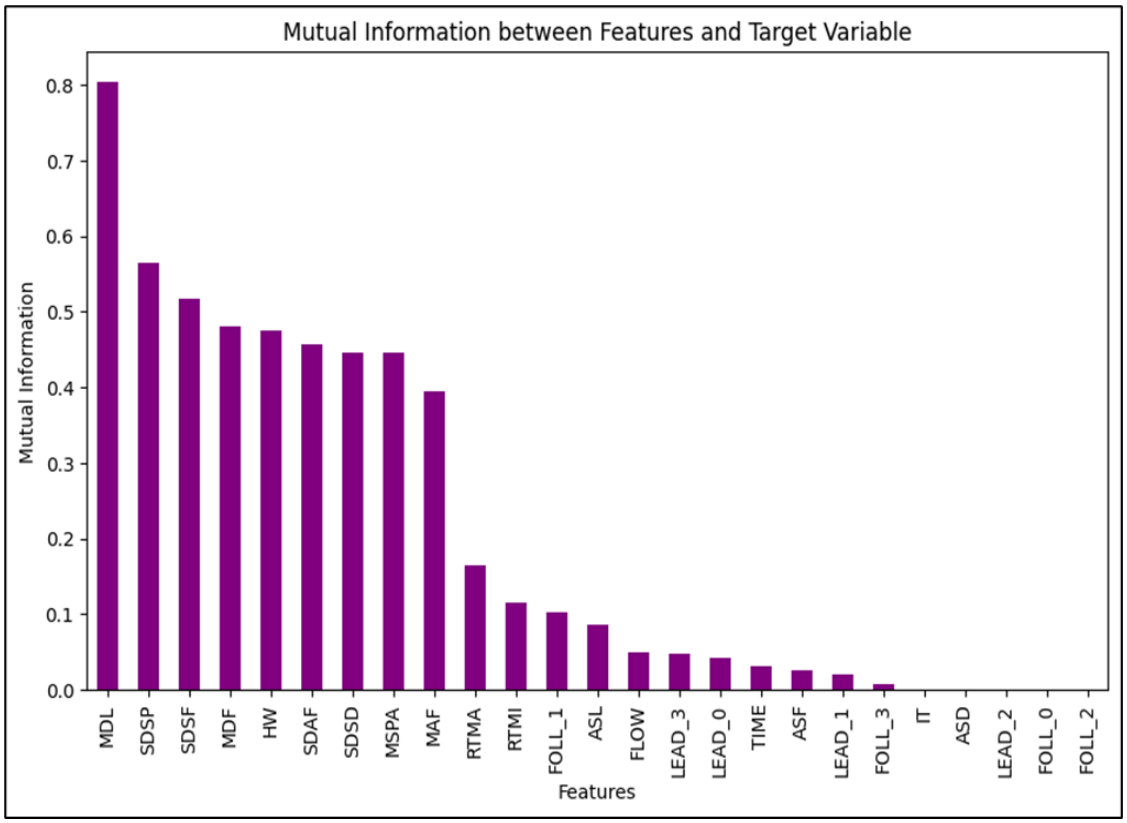

Figure 4.6 Mutual Information between features and target variable

4.4 CLASSIFICATION MODELLING

Based on the selected features, five machine learning algorithms were utilized to build classification models. In these models, the classes were defined as follows: class 0 represented low risk conflict, class 1 represented medium risk conflict, class 2 represented high risk conflict, and class 3 represented critical conflict.

The classification report and confusion matrix of logistic regression model is shown in table 4.2 and figure 4.7 respectively. The resulting logistic regression model demonstrated satisfactory performance for all four classes. However, the model

misclassified some critical conflicts as high risk conflicts, and correctly classified 29 out of 42 high risk conflicts. Furthermore, the model showed high accuracy in classifying low risk conflicts, with over 82% of them being correctly classified, while 14% of medium risk conflicts were misclassified. The model achieved a high accuracy of 78%, with a precision of 79%, recall of 78%, and F1 score of 78%. Notably, low risk conflict exhibited the highest precision and F1 score compared to all other classes, while medium risk conflict exhibited the highest recall.

Table 4.2 Classification report of Logistic Regression model

	Precision	**Recall**	**F1-Score**
Low risk conflict (0)	0.92	0.82	0.87
Medium risk conflict (1)	0.73	0.85	0.79
High risk conflict (2)	0.71	0.69	0.70
Critical conflict (3)	0.80	0.74	0.77
Macro Average	0.79	0.78	0.78
Weighted Average	0.79	0.78	0.78
Accuracy	0.78		

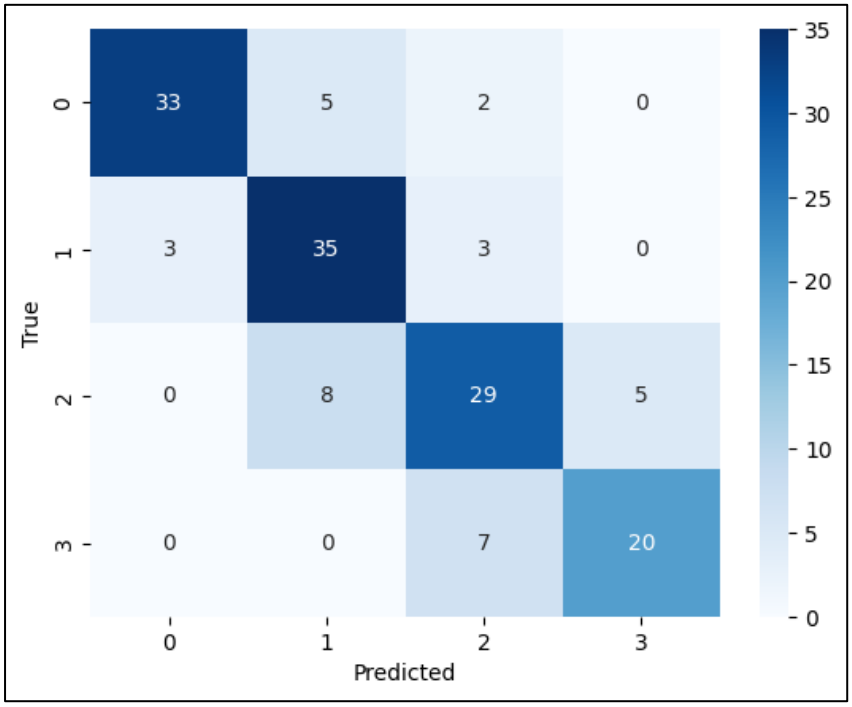

Figure 4.7 Confusion matrix of Logistic Regression model

The confusion matrix and classification report for the decision tree model presented in figure 4.8 and table 4.3 respectively demonstrated its efficacy in identifying different levels of conflict risk. Specifically, the model achieved a high level of accuracy in accurately classifying 25 out of 27 critical conflicts, indicating its robustness in identifying instances of critical conflict. However, the model did misclassify some high risk and medium risk conflicts as critical conflicts. Notably, the model achieved a high level of accuracy in identifying low risk conflicts, correctly classifying 90% of instances, while misclassifying 27% of medium risk conflicts. Additionally, the precision and f1-score were highest for low risk conflicts, indicating the model's ability to accurately classify instances in this category. However, the model's performance was relatively weaker in identifying high risk conflicts, which demonstrated the lowest precision and f1-score among all classes. Furthermore, the model demonstrated a high recall for critical conflicts; in contrast, the recall was lowest for medium risk conflicts, indicating an area of improvement in the model's ability to accurately identify instances in this class. Overall, the decision tree model achieved an accuracy of 82%, precision of 82%, recall of 83%, and f1-score of 82%.

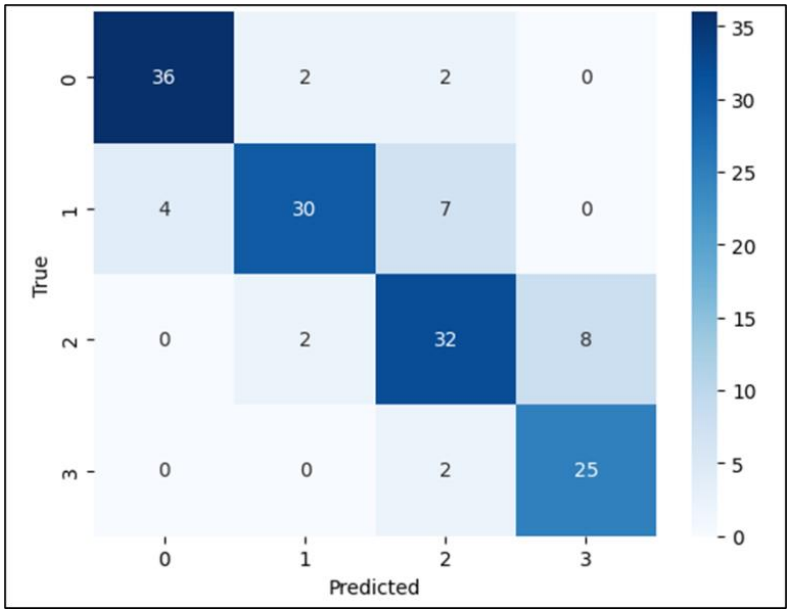

Figure 4.8 Confusion matrix of Decision Tree Classifier

The classification report and confusion matrix for the random forest classifier shown in table 4.4 and figure 4.9 respectively indicated its performance in accurately identifying conflict risk levels. The random forest model demonstrated a high level of accuracy in identifying all critical conflicts correctly.

Table 4.3 Classification report of Decision Tree Classifier

	Precision	**Recall**	**F1-Score**
Low risk conflict (0)	0.90	0.90	0.90
Medium risk conflict (1)	0.88	0.73	0.80
High risk conflict (2)	0.74	0.76	0.75
Critical conflict (3)	0.76	0.93	0.83
Macro Average	0.82	0.83	0.82
Weighted Average	0.83	0.82	0.82
Accuracy	0.82		

However, the model did misclassify some instances of high risk conflicts as medium risk and critical conflicts. The model also achieved a high level of accuracy in identifying instances of low risk conflicts, correctly classifying 93% of instances. Furthermore, the model achieved the highest precision for high risk conflicts and the highest f1 score for low risk conflicts. Critical conflicts demonstrated the highest recall of 100%, indicating the model's capability to correctly identify all instances in this class. Additionally, the model misclassified only two out of the total medium risk conflicts. Overall, the random forest classifier achieved an accuracy of 91%, precision of 91%, recall of 92%, and f1 score of 91%. These results demonstrate a robust performance of the model in predicting conflict risk levels, particularly in accurately identifying instances of critical conflicts and low risk conflicts.

Table 4.4 Classification report of Random Forest Classifier

	Precision	**Recall**	**F1-Score**
Low risk conflict (0)	0.95	0.93	0.94
Medium risk conflict (1)	0.89	0.95	0.92
High risk conflict (2)	0.97	0.79	0.87
Critical conflict (3)	0.82	1.00	0.90
Macro Average	0.91	0.92	0.91
Weighted Average	0.91	0.91	0.91
Accuracy	0.91		

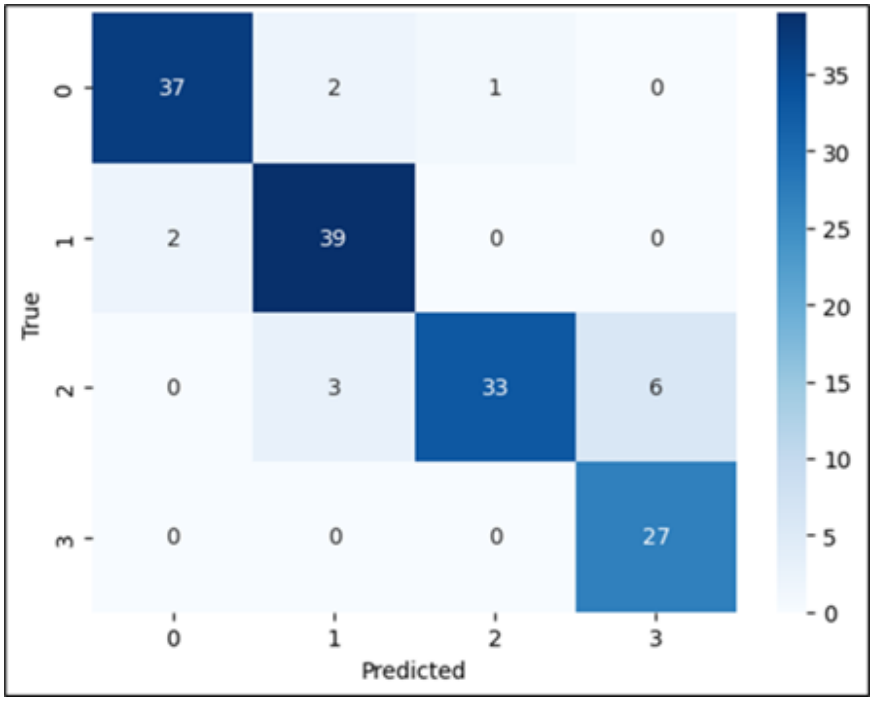

Figure 4.9 Confusion matrix of Random Forest Classifier

The classification report and confusion matrix of the XGBoost classifier shown in table 4.5 and figure 4.10 respectively provide insights into the model's ability to predict conflict risk levels. The model showed an excellent performance in correctly classifying all critical conflicts, which is a crucial class to identify accurately. However, it did misclassify some high risk conflicts as medium risk and critical conflicts. Furthermore, the model also misclassified medium risk conflicts as low and high risk conflicts. In addition, the model correctly classified 85% of the total low risk conflicts, which is lower compared to other models. The low and medium risk conflicts exhibited the f1 score of 88%, indicating that the model's performance in identifying instances in these classes is relatively good. The model showed the highest precision for high risk conflicts, indicating its ability to identify correctly a significant number of instances in this class. Moreover, critical conflicts showed the highest recall, with the model correctly identifying all instances in this class. The XGBoost classifier achieved an accuracy of 87%, a precision of 88%, a recall of 88%, and an f1 score of 87%. Although these results are impressive, there is room for improvement, especially in correctly identifying instances of high risk conflicts and low risk conflicts.

Table 4.5 Classification report of XGBoost Classifier

	Precision	Recall	F1-Score
Low risk conflict (0)	0.92	0.85	0.88
Medium risk conflict (1)	0.84	0.93	0.88
High risk conflict (2)	0.97	0.76	0.85
Critical conflict (3)	0.77	1.00	0.87
Macro Average	0.88	0.88	0.87
Weighted Average	0.89	0.87	0.87
Accuracy	0.87		

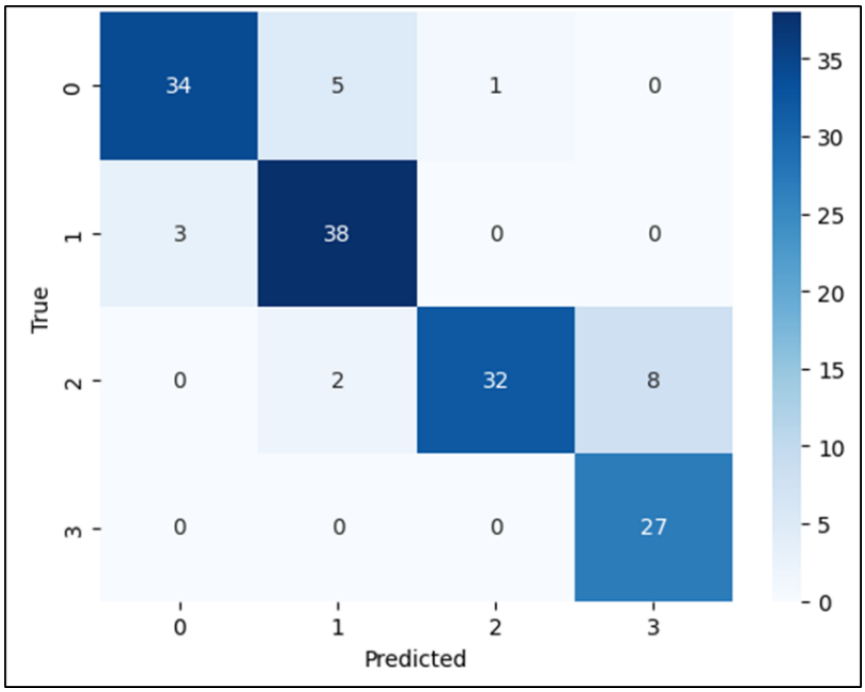

Figure 4.10 Confusion matrix of XGBoost Classifier

The Support Vector Machine (SVM) model's classification report and confusion matrix provide given in table 4.6 and figures 4.11 respectively provide insights into its performance in predicting conflict risk levels. The model accurately identified 22 out of 27 critical conflicts, a crucial class for conflict risk prediction. However, the model did misclassify some high risk conflicts as critical conflicts and medium risk conflicts. Furthermore, the model correctly classified 85% of the low risk conflicts and only misclassified 17% of the medium risk conflicts. The low risk conflicts exhibited the highest precision, recall, and f1-score, indicating the model's excellent performance in

identifying instances in this class. However, high risk conflicts exhibited the lowest precision, recall, and f1-score, indicating that the model struggled to identify instances in this class accurately. Overall, the SVM model achieved an accuracy of 81%, a precision of 81%, a recall of 81%, and an f1-score of 81%. Although these results are satisfactory, the model's performance can be improved, particularly in identifying instances of high risk conflicts.

Table 4.6 Classification report of Support Vector Machine

	Precision	**Recall**	**F1-Score**
Low risk conflict (0)	0.87	0.85	0.86
Medium risk conflict (1)	0.81	0.83	0.82
High risk conflict (2)	0.78	0.76	0.77
Critical conflict (3)	0.79	0.81	0.80
Macro Average	0.81	0.81	0.81
Weighted Average	0.81	0.81	0.81
Accuracy	0.81		

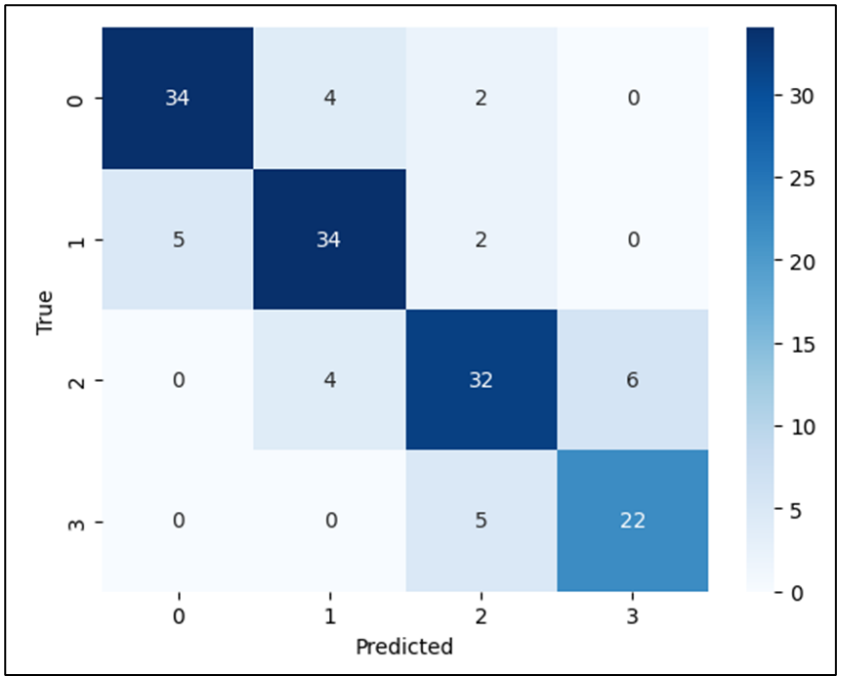

Figure 4.11 Confusion matrix of Support Vector Machine

The table 4.7 depicting the performance metrics of all five classifiers enables comparison of their effectiveness in predicting conflict risk levels. The results indicate that the random forest classifier surpasses all other models with its highest accuracy, precision, recall and F1 score. Additionally, the false positive and false negative rates are the lowest for this model, indicating that it makes the least amount of incorrect predictions. These results suggest that the random forest classifier is the most effective model for this classification task.

Table 4.7 Comparison of performance of various models

Classification model	**Performance Metrics**					
	Accuracy	***Precision***	***Recall***	***False Positive Rate***	***False Negative Rate***	***F1 Score***
Logistic Regression	0.78	0.79	0.78	0.074	0.23	0.78
Decision Tree	0.82	0.83	0.82	0.06	0.17	0.82
Random Forest	0.91	0.91	0.92	0.03	0.08	0.91
XGBoost	0.87	0.88	0.88	0.04	0.115	0.87
Support Vector Machine	0.81	0.81	0.81	0.06	0.19	0.81

The figure 4.12 to 4.16 represents the ROC AUC curve of different classification models. In machine learning, the ROC curve and AUC are frequently used to compare the effectiveness of various classification algorithms. The AUC is a summary statistic that assesses the overall performance of the model over all potential thresholds, and the ROC curve is a plot of true positive rate (sensitivity) vs false positive rate (1-specificity) for different model thresholds. A higher AUC value indicates better discrimination between the positive and negative classes, and it is often considered to be a key performance metric while comparing classification models. This is because the AUC provides a single summary statistic that considers the entire ROC curve, instead of just a single point on the curve. For multilabel classification, the assumption that all classes are equally important

is often untrue. Therefore, the macro-averaging method was applied to calculate overall AUC of model (Song et al., 2021).

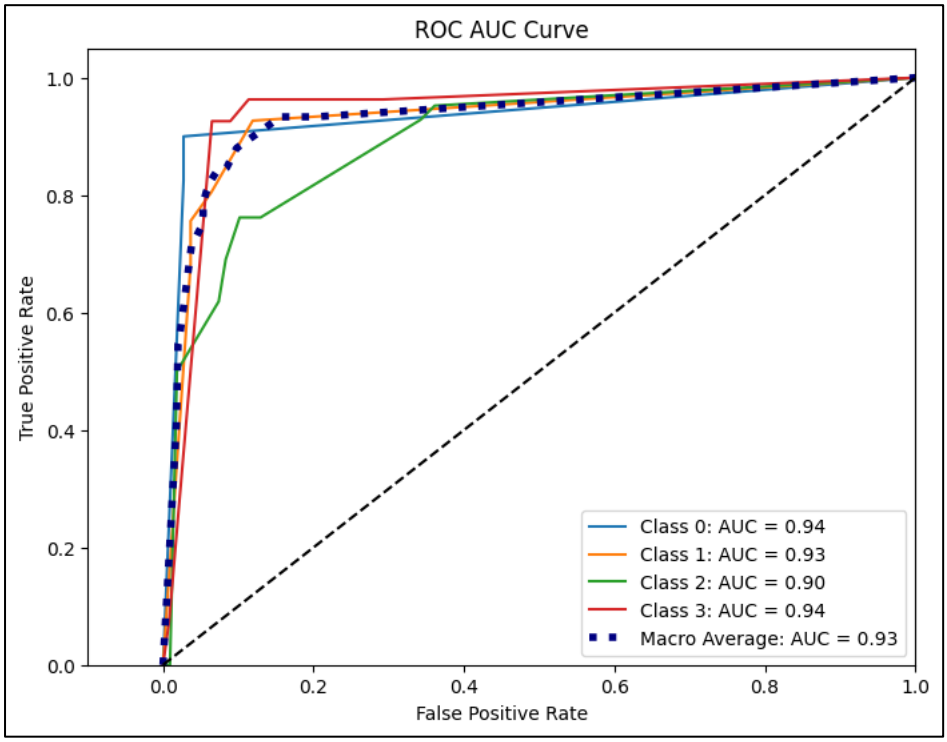

Figure 4.12 ROC AUC curve of logistic regression model

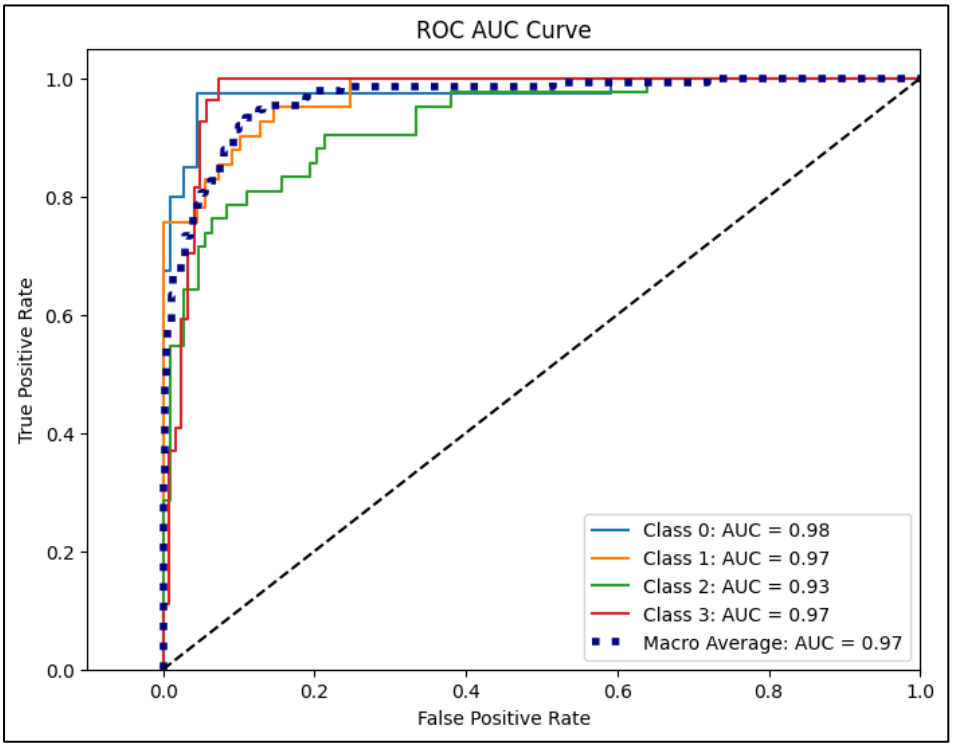

Figure 4.13 ROC AUC curve of Decision Tree Classifier

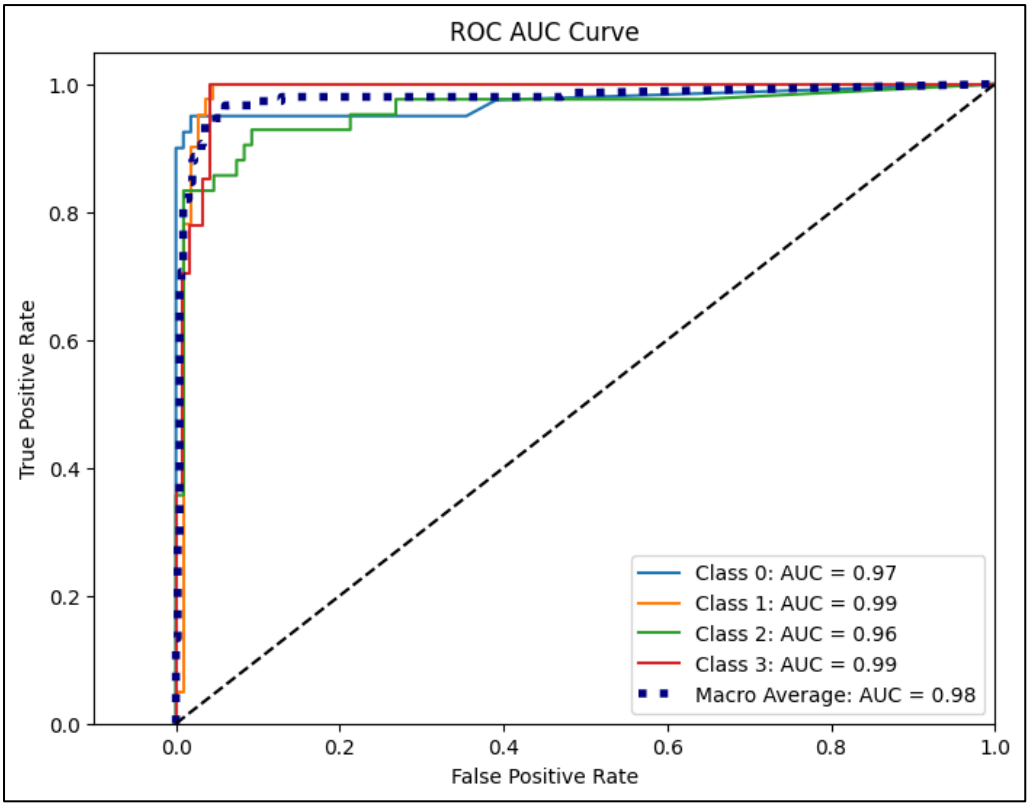

Figure 4.14 ROC AUC curve of Random Forest Classifier

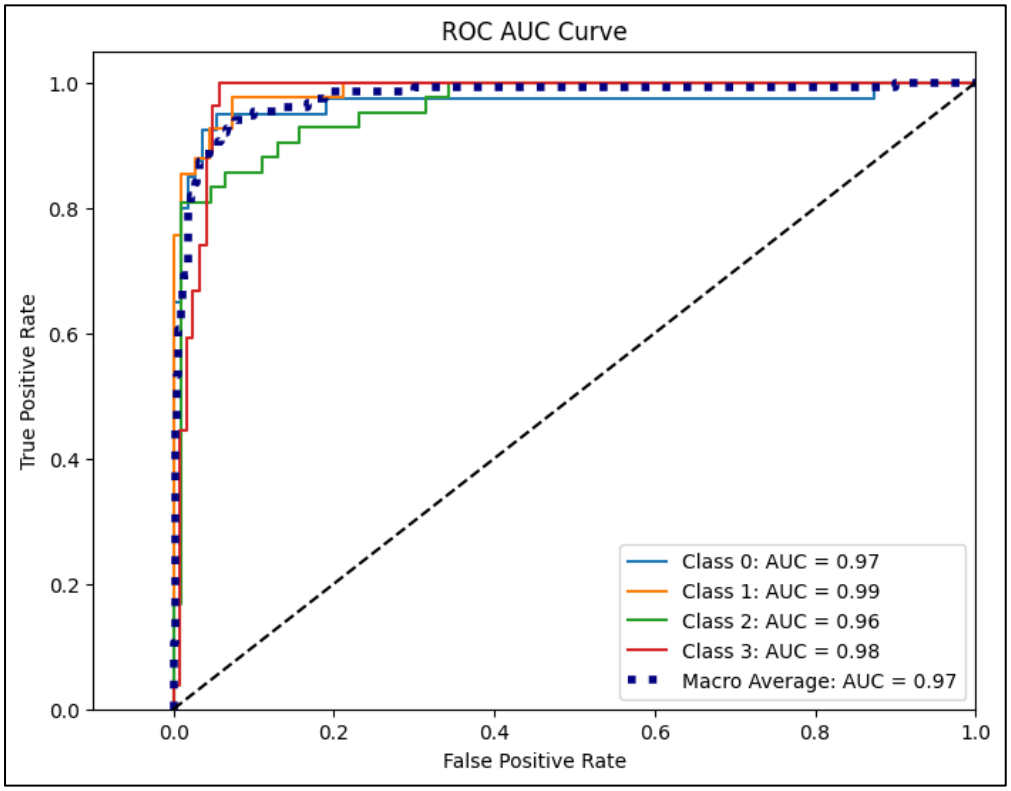

Figure 4.15 ROC AUC curve of XGBoost Classifier

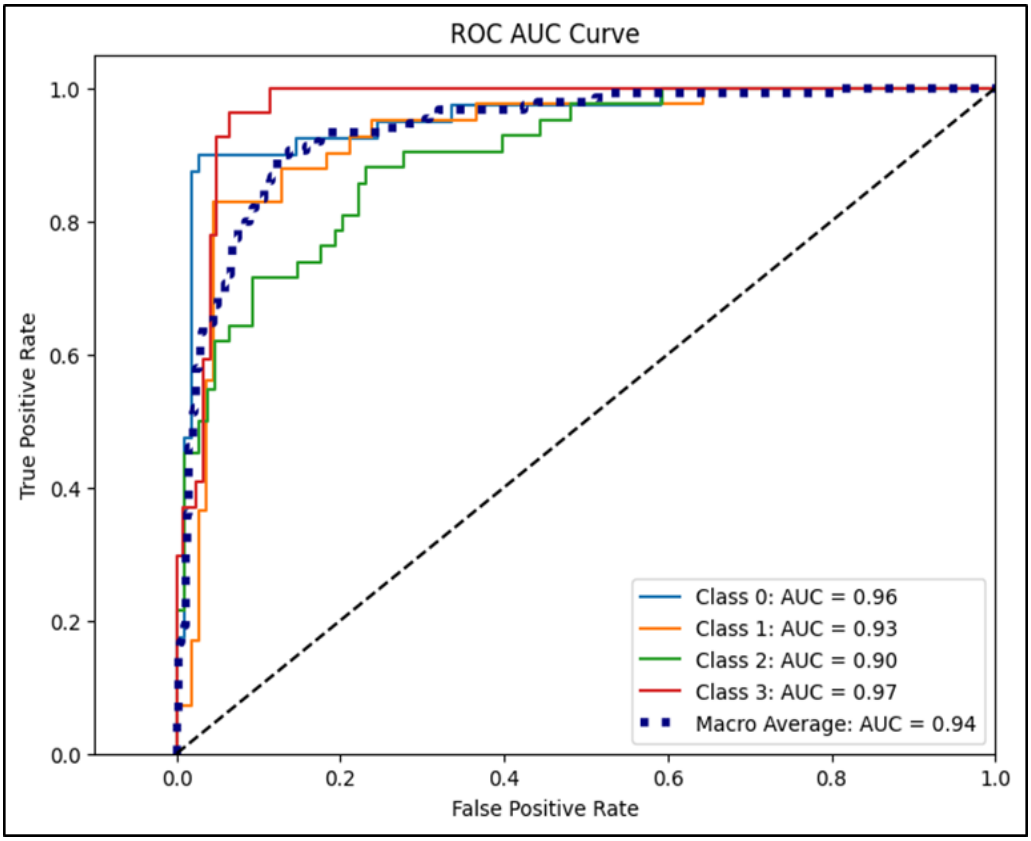

Figure 4.16 ROC AUC curve of Support Vector Machine

In the given context, it is clear that the Random Forest (RF) classification model exhibits the greatest AUC score across all classes, as well as the highest average AUC score of 0.98. Since the AUC score represents the optimal balance between true positive rate and false positive rate, a higher AUC value is typically desirable while evaluating a machine learning algorithm (Ijaz et al., 2021). Based on the analysis of the ROC curves, it can be concluded that the RF classification model excels over the other four classifiers in the context of the current study.

4.5 FEATURE IMPORTANCE

As previously discussed, SHAP values offer a way to gauge how much each attribute influences a model's prediction. To utilize SHAP values, a trained machine learning model is required as input. In this case, the best-performing model, random forest, was used. The global importance of input variables is represented in figure 4.17 using a SHAP summary plot, which is a bar chart that displays the average impact of each attribute on the outcome of the model in terms of its SHAP value. The features are ranked in order of their mean absolute SHAP value across all instances in the dataset. The output of the

model is more affected by features with longer bars than by features with shorter bars for a given class label.

The SHAP summary plot shows that maximum deceleration leader (MDL) is the most important feature, while right turn vehicles from major (RTMA) is the least important. However, some features have similar bar lengths across different classes, which may have resulted in misclassification or confusion between classes. For example, MDL and standard deviation of speed follower (SDSF) have nearly identical bar lengths for classes 1 and 2, indicating that these features may have contributed to confusion between the two classes.

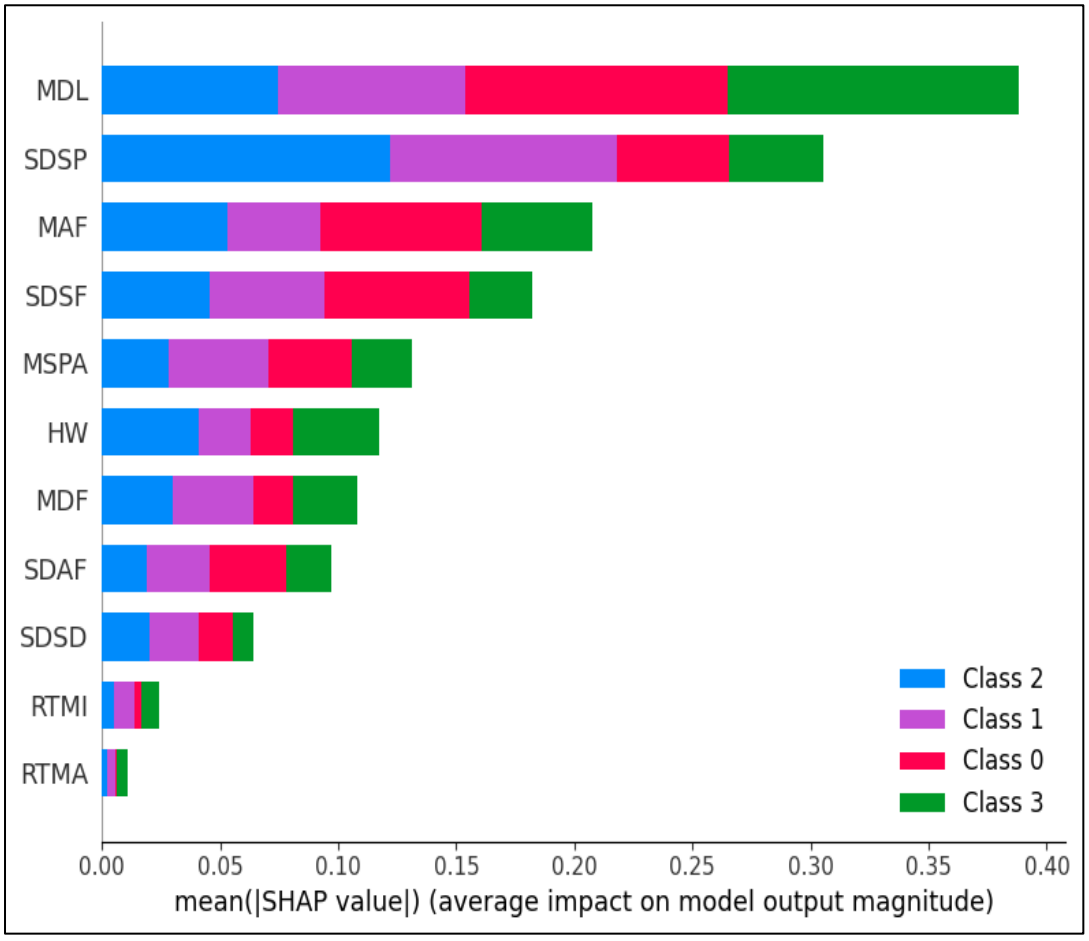

Figure 4.17 SHAP Summary Plot

In SHAP beeswarm plot, for each variable, every instance (i.e.row) of the dataset appears as its own point. Depending on their SHAP value, the points are dispersed laterally along the x axis. The points are piled vertically in regions with a high concentration of SHAP values (Islam & Abdel-Aty, 2023; Yuan et al., 2022b). The colour bar corresponds to the raw values (not to be confused with the SHAP values) of the variables for each instance (i.e. point) on the graph. The colour red indicates high values while blue indicates low values. For categorical variables like the presence of right turn vehicles from minor road (RTMI) taking binary values 0 and 1, the presence is coded as red while the absence is coded as blue. For continuous variables like SDSF with values ranging from 0 to 8, smaller values are coded as blue and larger values are coded as red. Each variable's x-axis colour distribution offers information on how its natural values and SHAP values relate to one another (Mangalathu et al., 2022). In the SHAP beeswarm plot, instances with points to the right of the midline contribute to an increase in the model's predicted probability, while those on the left correspondingly decrease it (Kannangara et al., 2022; Meshoul et al., 2022). Notably, the degree of impact on the model's predictions increases with distance from the midline.

Upon examination of the beeswarm plot for low risk conflict shown in figure 4.18, it becomes evident that the MDL feature holds the greatest importance, while RTMA feature holds the least importance. In particular, lower values of MDL, maximum acceleration follower (MAF), SDSF, standard deviation of speed leader (SDSL), standard deviation of acceleration follower (SDAF), and maximum deceleration follower (MDF) tend to contribute towards low risk conflicts, while higher values of mean spacing between the vehicles (MSPA) and headway (HW) also play a role in reducing risk levels. The absence of right turn vehicles from both minor and major road increases the chances of low risk conflicts. To provide further clarity, SHAP dependency plots shown in figure 4.19 can be utilized. Specifically, for SDAF, values below 3 are associated with an increased likelihood of low risk conflicts. Meanwhile, for MAF and MDL, accelerations below $4m/s^2$ are more likely to result in low risk conflicts. For MDF, values below $2m/s^2$ also contribute more heavily to low risk conflicts. Additionally, for SDAF and standard deviation of spacing between vehicles (SDSP), values below 4 have a greater association with low risk conflicts. Finally, for MSPA, values beyond 8m can increase the probability of low risk conflicts.

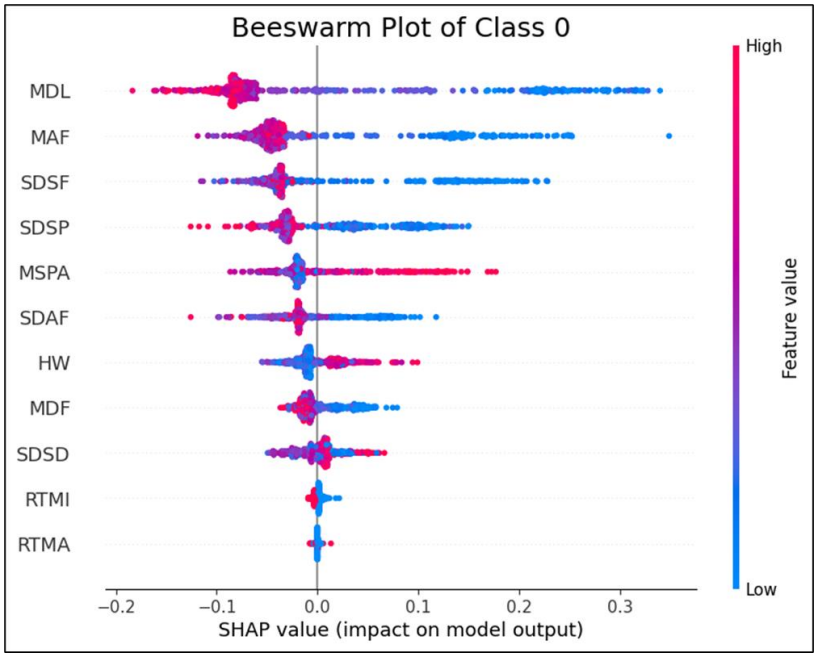

Figure 4.18 SHAP beeswarm plot for low risk conflicts

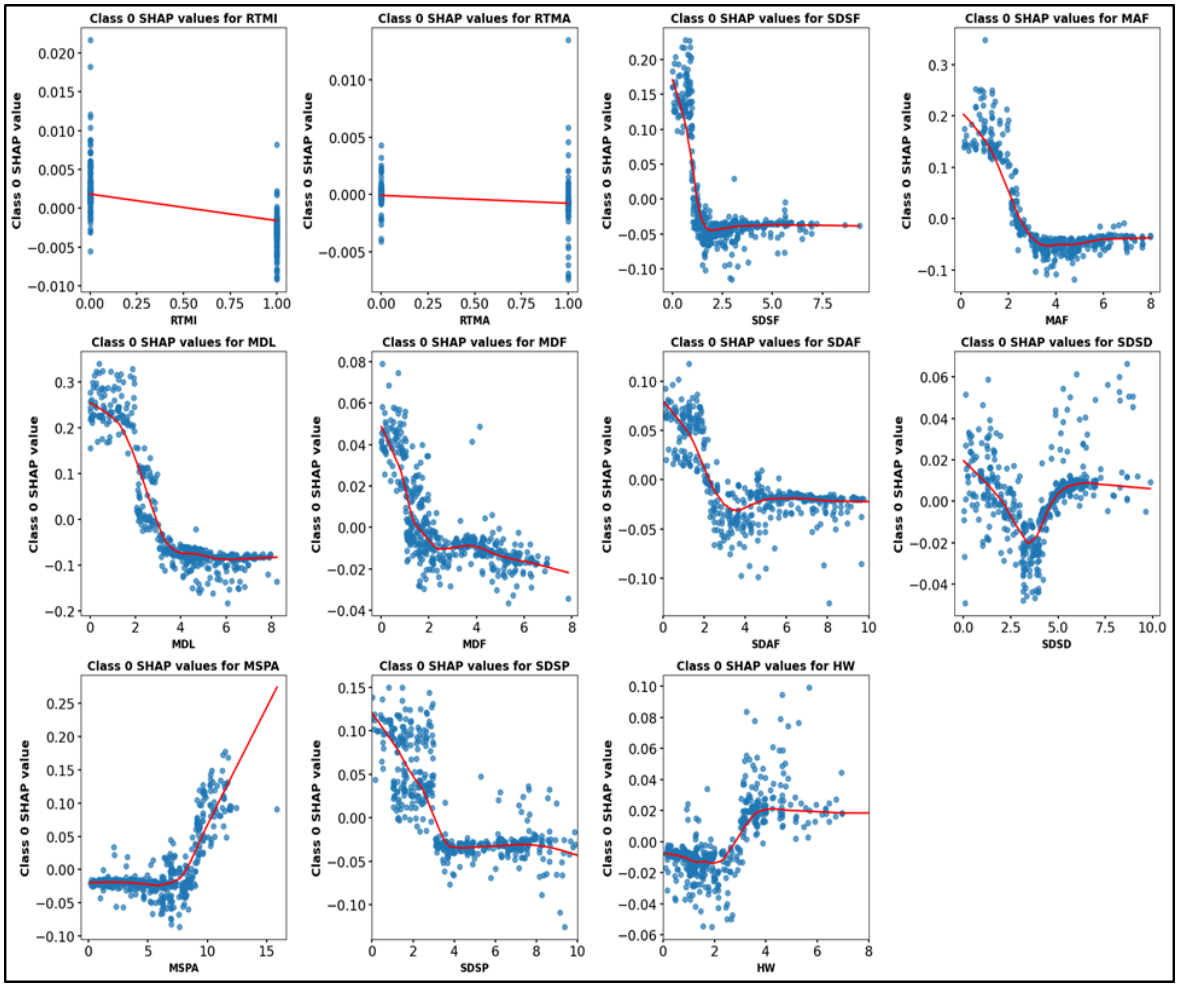

Figure 4.19 SHAP Dependency plot for low risk conflicts

For medium risk conflict, the beeswarm plot and dependency plot is shown in figure 4.20 and 4.21 respectively. From the plots, it is revealed that SDSP is the most important feature, whereas RTMA is the least important. Additionally, the lower values of SDSP, SDSF, and MDF were found to increase the probability of medium risk conflicts. Furthermore, it was observed that for the remaining features, neither the highest nor the lowest values had a positive impact on medium risk conflict. Instead, it was the middle range values that played a significant role in contributing to this type of conflict, as demonstrated in the dependency plot. Specifically, for the continuous variables, MDL and MAF, with values ranging from 0 to 8, it is observed that the highest positive SHAP value is around the value of 4, which is the middle range for these features. For SDAF, a continuous variable with values ranging from 0 to 10, the highest positive SHAP value is between 3 and 4. Regarding SDSF, the peak positive SHAP value was observed for values between 1 and 2.5. For MSPA, a continuous variable ranging from 0 to 15, values around 7 were found to contribute more towards increasing the risk of medium conflict. For, SDSP, the values below 4 has the highest positive contribution.

For high risk conflict, the beeswarm plot and dependency plot is shown in figure 4.22 and 4.23 respectively. In the case of high risk conflict, the contribution towards increasing the risk is not observed in the highest or lowest values of the features, but rather in the middle range values. This observation is applicable for most of the features such as SDSP, MDL, MAF, SDSF, MDF, and SDAF. However, for features such as HW and MSPA, the values in proximity to the lowest value have been found to have a higher contribution towards high risk conflict. The dependency plot provides a clearer insight into this finding. For continuous features such as MAF, MDL, MDF, and SDSP, the peak SHAP score is observed around the range of 4, which is the middle range as these features have values ranging from 0 to 8. For SDAF, which also has a continuous range of values from 0 to 10, the peak positive SHAP value is obtained around the range of 6. On the other hand, for SDSD, values above the range of 5 are observed to negatively impact the probability of high risk conflict. For HW, the values around 2s lead to high risk conflict. Right turn vehicles from minor road have a marginally favourable effect on the probability, whereas right turn vehicles from major roads have a marginally negative effect.

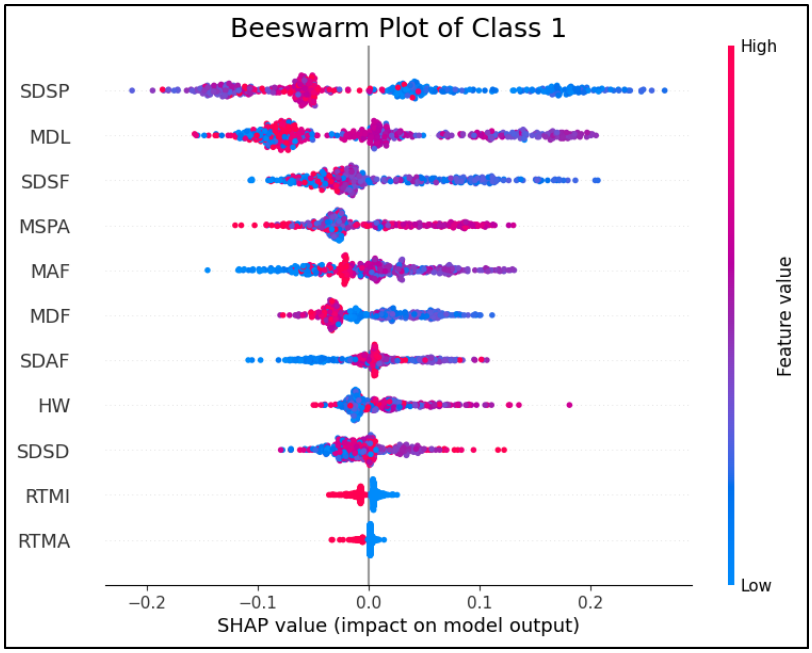

Figure 4.20 SHAP Beeswarm Plot for medium risk conflicts

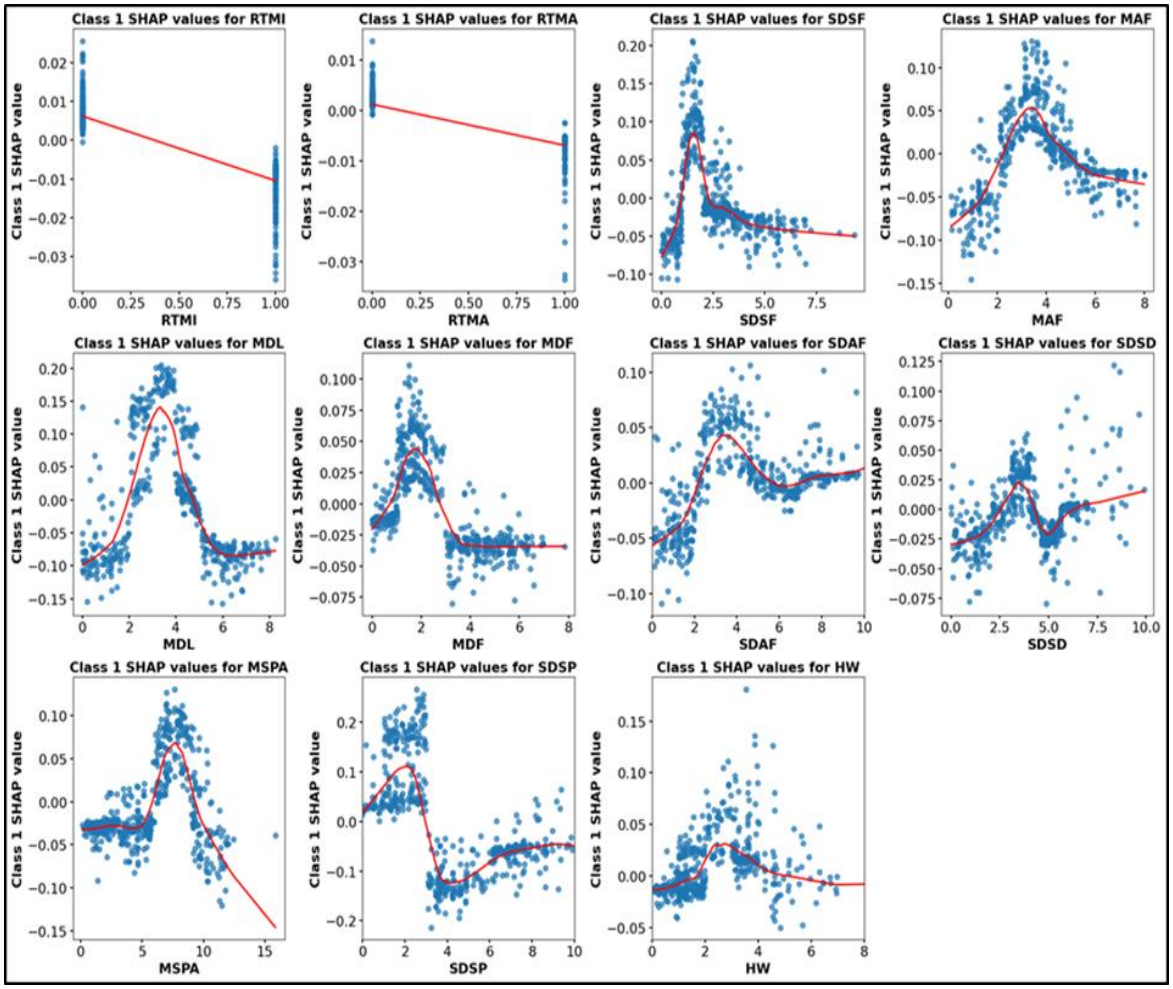

Figure 4.21 SHAP Dependency plot for medium risk conflicts

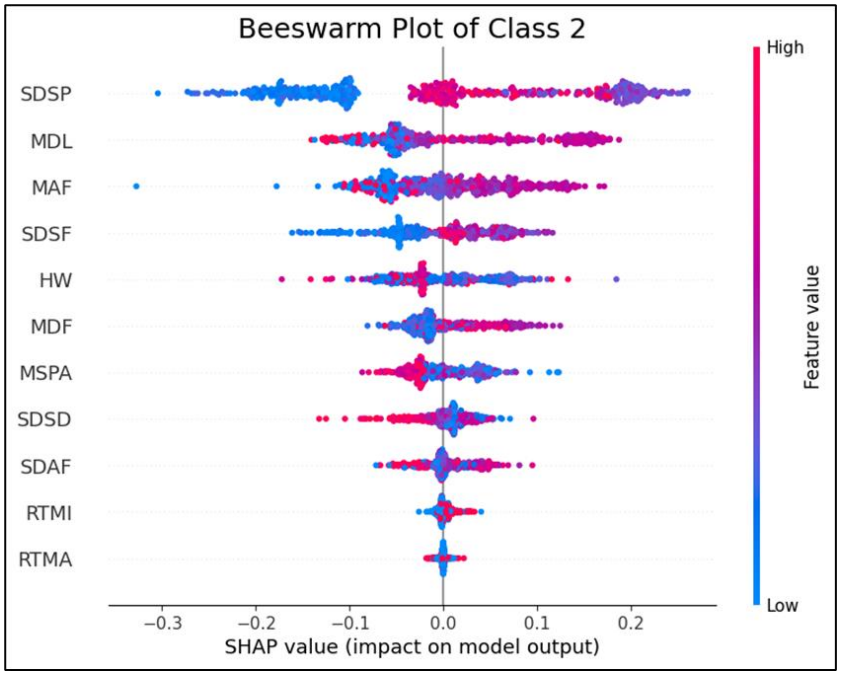

Figure 4.22 SHAP Beeswarm Plot for high risk conflicts

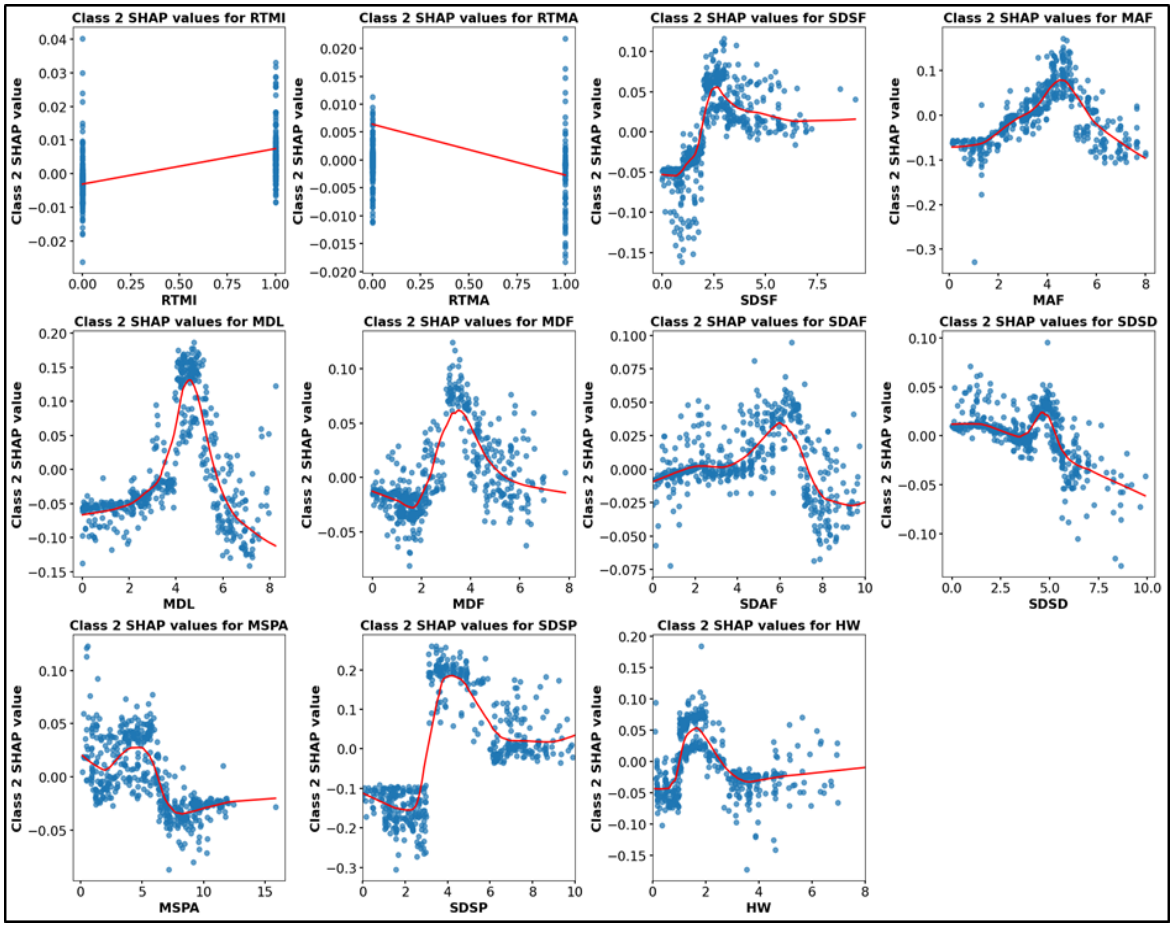

Figure 4.23 SHAP Dependency plot for high risk conflicts

For critical conflict, the beeswarm plot and dependency plot is shown in figure 4.24 and 4.25 respectively. In the case of critical conflicts, which are classified as class 3, it is the highest values of features like MDL, MAF, MDF, SDSF, SDAF, and SDSP that contribute towards an increase in the probability of occurrence. Additionally, the presence of right turn vehicles from both major and minor roads, denoted by RTMI and RTMA respectively, has a positive impact on the predicted probability. Conversely, lower values of HW and MSPA contribute to critical conflicts. In the case of SDSF, the predicted probability of critical conflicts increases beyond a value of 2.5. Similarly, for MAF, MDL and MDF, values beyond 4 m/s^2 increase the likelihood of critical conflicts. For SDAF, an increase in predicted probability is observed beyond a value of 6. Furthermore, for SDSD, the peak positive SHAP value is observed around a value of 6, indicating an increase in the probability of critical conflicts. Additionally, the probability of critical conflicts is higher when MSPA is less than 6m, while for HW; conflicts are more likely when the time gap is less than 2s. Finally, the peak positive SHAP value for SDSP is attained beyond a value of 6.

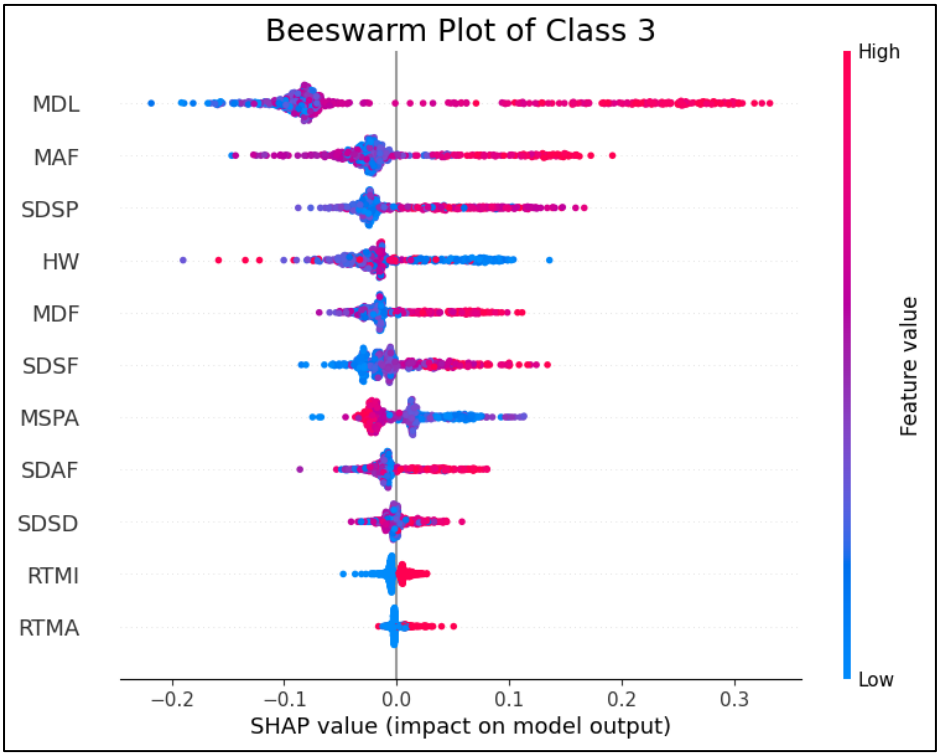

Figure 4.24 SHAP Beeswarm Plot for critical conflicts

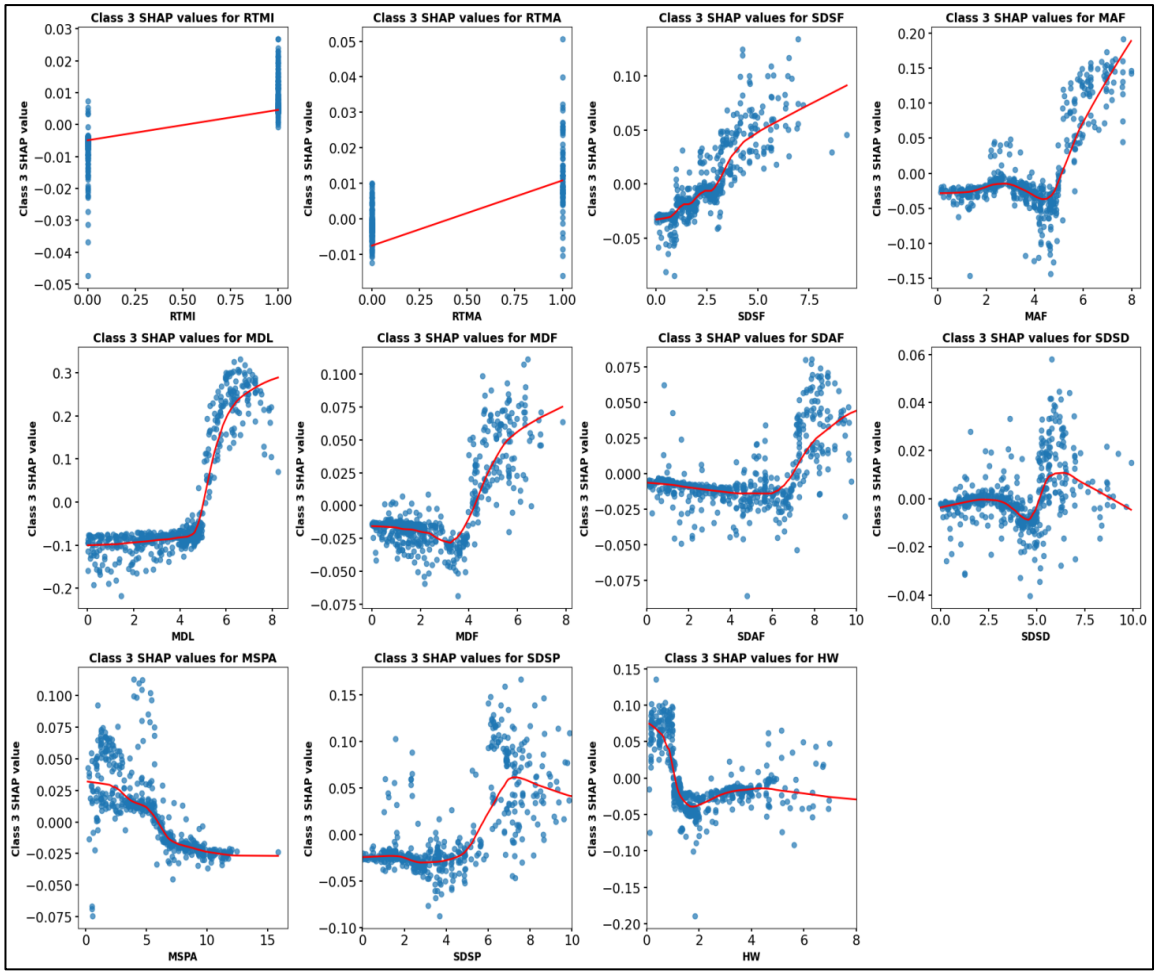

Figure 4.25 SHAP Dependency plot for critical conflicts

After conducting an analysis of the beeswarm and dependency plots for all four classes, it was observed that some features show an increase in risk levels while others show a decrease. Features such as SDSF, MAF, MDL, MDF, SDAF, SDSD and SDSP exhibit an increase in risk levels with an increase in their values. Conversely, a decrease in the values of MSPA and HW leads to an increase in the risk of conflict.

A higher value of standard deviation of speed follower indicates more speed variation, which can lead to erratic driving behaviour and increase the likelihood of conflicts. For example, if a driver is constantly accelerating and decelerating due to speed variation, it may increase the chances of rear-end collisions or other types of conflicts. Additionally, a high standard deviation of speed follower can indicate that the following driver is having difficulty maintaining a safe following distance from the lead vehicle, which can increase the risk of collisions (Hu et al., 2022). High values of maximum acceleration of following vehicle indicate that the followers are accelerating too fast which can lead to sudden

changes in speed and potential rear-end collisions. On the other hand, high values of maximum deceleration of leader and maximum deceleration of follower indicate that the leaders and followers are decelerating rapidly which can cause sudden stops and also contribute to the likelihood of rear-end collisions (Yu et al., 2021). Therefore, these features indicate aggressive driving behaviour and a lack of safe distance between the vehicles, which increases the risk of conflicts. Standard deviation acceleration follower, standard deviation of speed difference, and standard deviation of spacing are all measures of variability in the behaviour of vehicle on the road. As these values increase, it indicates a higher degree of unpredictability and inconsistency in the way vehicle is moving. This can lead to a higher likelihood of conflicts, particularly in situations where vehicles are closely spaced or moving at different speeds. For example, a high value of standard deviation of spacing between vehicles might indicate that vehicle is not maintaining a consistent following distance, which could increase the risk of rear-end collisions (Caird et al., 2014; Peng et al., 2017). Similarly, a high value of standard deviation of acceleration of following vehicle might indicate that some drivers are frequently accelerating and decelerating rapidly, which could increase the risk of sudden stops or collisions. A high value of standard deviation of speed difference might indicate that there is a wide range of speed differences between vehicles, which could make it difficult for drivers to anticipate each other's movements and avoid collisions. Mean spacing between vehicles is a measure of the average distance between vehicles on the road. A smaller mean spacing between the vehicles indicates that vehicles are closer to each other, which increases the probability of collisions and conflicts. This is because in case of sudden changes in speed or direction of a vehicle, there is less space for other vehicles to maneuver or brake, leading to a higher chance of collisions (Shangguan et al., 2023). Similarly, headway is a measure of the time gap between two consecutive vehicles on the road. A smaller headway indicates that vehicles are closer to each other in terms of time, which again increases the probability of conflicts. This is because a smaller headway means that there is less time for vehicles to react to sudden changes in speed or direction of other vehicles, leading to a higher chance of collisions.

For the local interpretation of each instance within the dataset, SHAP force plot can be used. The force plot provides a comprehensive understanding of a single model prediction. It allows us to identify the specific features that contributed to the model's prediction for a particular observation, making it a useful tool for error analysis or gaining

a deeper insight into a specific case. It is worth noting that the machine learning model's predictions for each instance can be obtained by adding a fixed base value to the sum of the SHAP values. In regression models, the base value is equivalent to the mean of the target variable whereas for classification models, the base value is equivalent to the prevalence of the positive class in a given dataset. The value of each feature is shown by an arrow on the plot; red arrows indicate characteristics that improved the model's score, while blue arrows indicate elements that made it worse. The size of the arrow represents how much the characteristic has an effect on the result (Dong et al., 2022). The force plot for random instances of each class is given below.

The figure 4.26 displays a force plot for the low risk conflict with a base value of 0.239, and the model's predicted probability for the class is 1. By comparing the dependency plot and force plot, it is evident that lower values of various features such as SDSF, MAF, MDL, MDF, SDAF, and SDSD are associated with low risk of conflict. As observed in the dependency plot, higher values of HW (more than 4s) are associated with lower risk of conflict. Interestingly, the force plot indicates that the most significant feature contributing towards the low risk of conflict is headway, with a value of 4.28.

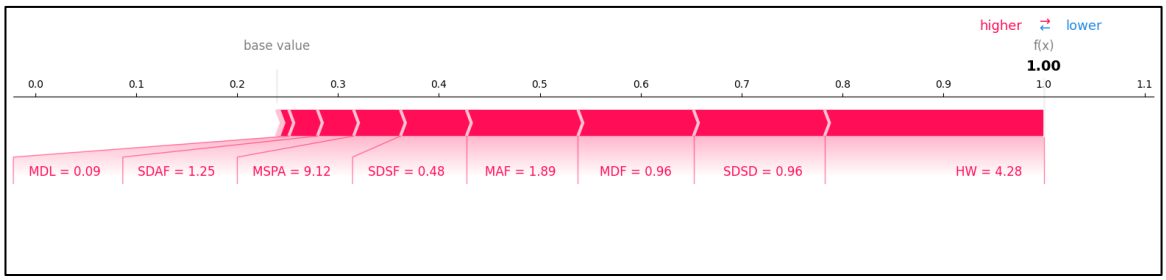

Figure 4.26 SHAP Force Plot for one particular observation from low risk conflict

The force plot for medium risk conflict shown in figure 4.27 displays a base value of 0.239, with a predicted probability of 0.98 for the instance in consideration. According to the plot, MDL is the most significant feature that increases the predicted probability, followed by SDSF and SDSD. As per the dependency plot, a value of MDL around 4 has a greater impact on increasing the probability, and in the force plot, we observe that MDL feature value of 4.62 contributes towards medium risk conflict. Similarly, SDSF exhibits a peak positive SHAP value when its value is approximately 2, and in this instance, we can observe a high and positive contribution towards the predicted probability when its value is 1.86.

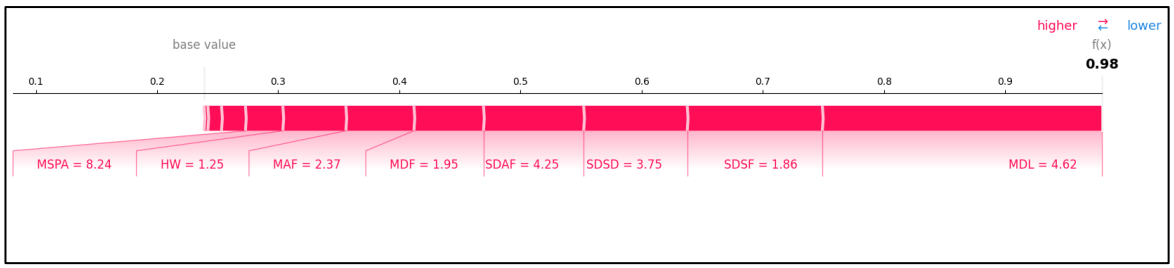

Figure 4.27 SHAP Force Plot for one particular observation from medium risk conflict

The force plot displayed in the figure 4.28 pertains to a specific instance of high risk conflict. The class has a base value of 0.306 and the predicted probability for this instance is 0.68. The most significant feature is SDSD, followed by MDL and SDAF. As observed in the dependency plot, the probability for this class increases when SDSD value is around 5, which is also evident in this instance as the value of SDSD is 5.42. The blue arrow indicates the feature that reduces the predicted probability, which in this case is SDSP. According to the dependency plot, the probability increases when the value of SDSP is around 4, but in this instance, the value is 6.98, which explains the negative impact.

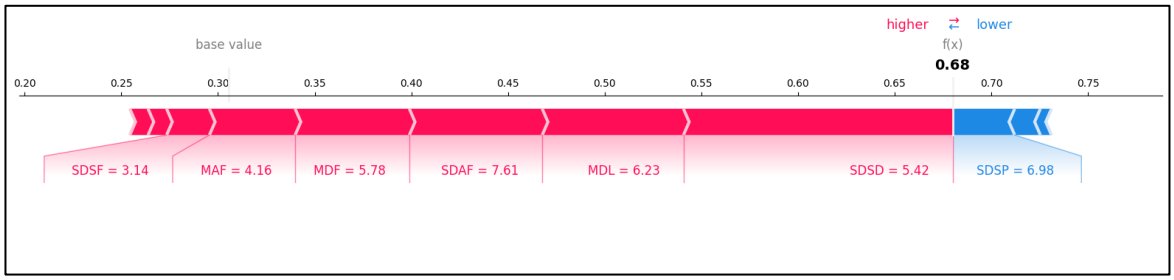

Figure 4.28 SHAP Force Plot for one particular observation from high risk conflict

The force plot for critical conflicts is displayed in figure 4.29. For the critical conflicts, the base value is 0.217 and the predicted probability for this instance is 0.86. The predominant feature for this instance is MSPA. The plot shows that when MSPA value is 3.64, the SHAP value is high and positive, indicating a significant positive impact on the predicted probability. This is consistent with the findings from dependency plot. Additionally, the plot shows that the feature SDAF has a positive impact when its value is 7.13, which aligns with the behaviour observed in the dependency plot. The other features also exhibit similar behaviour as explained in the dependency plot.

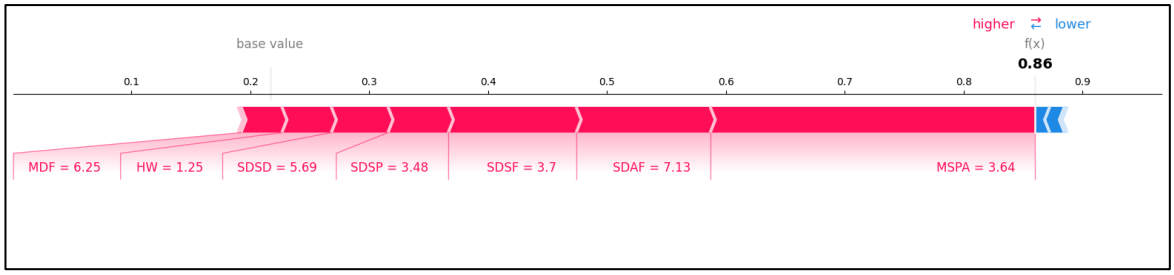

Figure 4.29 SHAP Force Plot for one particular observation from critical conflict

Overall, the force plot provides a clearer understanding of how the features contribute towards the predicted probability of conflicts for a particular instance.

4.6 SUMMARY

After the cluster analysis, it is found that there are four conflict risk levels recognized as low risk conflict, medium risk conflict, high risk conflict and critical conflict. The results of classification model reveal that random forest modelling is the best performing model out of the five classifiers. The SHAP interpretation provided insights into the relative importance of each feature in predicting the different risk levels. Specifically, the maximum deceleration of the leader, standard deviation of spacing between vehicles, maximum acceleration of the follower, standard deviation of speed follower, and mean longitudinal spacing between the vehicles were identified as the most critical features for accurate risk level classification.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 SUMMARY

In this study, a machine learning based rear end conflict risk modelling using vehicle trajectory data is performed. This study underwent through several stages including data collection, data extraction, cluster analysis, data pre-processing, feature selection, implementation of five machine learning models, comprehensive data analysis, and explanation of feature importance using SHAP.

5.2 CONCLUSIONS

After implementing and analysing, the following conclusion can be drawn:

- A silhouette score plot was used to identify the ideal number of clusters, which identified four as the ideal number of clusters. Thus the conflicts can be classified into four risk levels; low risk conflicts, medium risk conflicts, high risk conflicts and critical conflicts
- To categorise the risk levels, three distinct clustering techniques namely K-means clustering, spectral clustering, and agglomerative clustering were employed. Interestingly, the range of each cluster was found to be consistent across all three clustering techniques
- To mitigate the issue of multicollinearity, five variables were eliminated from the modeling process
- Furthermore, a feature selection technique based on mutual information was applied to identify only the most relevant features for classification modeling. This resulted in only the top 11 features being selected for further analysis
- Subsequently, various classification models, including logistic regression, decision tree, random forest, support vector machine, and XGBoost, were developed and their performance was compared using various performance metrics

- The results of the analysis indicated that the Random forest algorithm outperformed all other models, achieving an accuracy of 91%, precision of 91%, recall of 92% and AUC score of 0.98

- To address the challenge of interpretability in machine learning models, the SHAP (SHapley Additive exPlanations) analysis was used to find the relevant factors and evaluate their influence on the conflict risk

- From this analysis, the top five features that are most likely to influence the conflict risk were identified as maximum deceleration of leader, standard deviation of separation between automobiles, maximum acceleration of follower, Standard deviation of speed follower, and Mean longitudinal spacing between the vehicles

- By comparing the beeswarm plots for each class, valuable insights can be gained into which features are most crucial for each risk level and how they vary across different classes

- Furthermore, by analyzing the SHAP dependency plot, it is possible to visualize how the feature's value affects the model's prediction for each risk level. This contributes to a better comprehension of the connection between the model's input variables and output

- By analysing the SHAP force plot for a particular sample, we can see which features had the most significant impact on the model's output for that specific instance

The findings of this study provide insight into the risk variables that contribute to different levels of conflict and emphasise the significance of comprehending risk behaviours and their root reasons. Traffic accidents can be avoided and kept under control if driving hazards are properly addressed and managed.

The results indicate that unsafe driving behaviours are more likely to occur after that leading car has decelerated to its maximum. Safety countermeasures such automated emergency braking systems can be used to increase longitudinal safety while vehicle-following. These systems enable the following vehicle to brake forcefully when the preceding vehicle abruptly decelerates, preventing rear-end collisions. Moreover, it is crucial to maintain a safe distance between the leader and follower vehicles to allow for

sufficient reaction time in case of sudden stops or slowdowns. Lastly, avoiding distracted and aggressive driving is paramount as these behaviours significantly increase the risk of accidents. By adhering to these preventive measures, drivers can reduce the likelihood of rear-end crashes and promote road safety.

The development of a conflict risk classification model can be an important step towards enhancing the safety of connected and automated vehicles. This model works by analysing real-time data from various sources, such as sensors, communication devices, and traffic patterns, to identify potential conflicts and predict their severity. By classifying conflicts based on their level of severity, from minor incidents to critical situations, the model can help anticipate potential hazards and predict conflicts before they occur. The information provided by the conflict risk classification model can be transmitted to the connected vehicles involved, which can take appropriate action to avoid the potential conflict. This can include adjusting their speed or position on the road, or communicating with other vehicles to coordinate their movements. By taking these actions, connected and automated vehicles can help prevent accidents and improve overall safety on the road.

Additionally, the conflict risk classification model can be used to develop ADAS that help drivers avoid conflicts or lessen their severity. These systems, such as collision avoidance systems, speed adaptation systems, and lane-keeping assistance, can work in tandem with the conflict risk classification model to reduce the risk of conflicts on the road. By analysing the factors that contribute to conflicts and developing more advanced driver assistance systems, we can further improve road safety.

Although this study provides valuable insights, it has some limitations. Two notable limitations of this study are the small dataset, which exclusively focuses on rear-end conflicts, and the heavy reliance on manually extracted data. Future research could expand on this by incorporating trajectory data from connected and automated vehicles or using automated trajectory extraction through computer vision analytics. Additionally, the study could be extended to investigate conflicts at roundabouts and other types of unsignalized intersections to provide a more comprehensive understanding of conflict risk in different traffic scenarios.

5.3 SCOPE FOR FURTHER WORK

- *Integration of additional data sources*: The study focuses on using vehicle trajectory data to model rear-end conflict risk. However, integrating data from other sources like climatic conditions, road infrastructure, and driver behaviour can provide a more comprehensive understanding of risk factors.
- *Integration of CAV (Connected and Automated Vehicles) data in risk modelling*: CAVs generate a vast amount of data on their driving behaviour, which can be used to augment risk modelling efforts. The findings from this study can be extended to integrate CAV data and develop more robust risk models.
- *Safety countermeasures for CAVs:* CAVs are equipped with advanced sensors and automated systems that can enable them to avoid collisions. The study's findings can be applied to CAVs to develop more sophisticated safety countermeasures, such as predictive collision avoidance systems.

REFERENCES

Amini, R. E., Yang, K., & Antoniou, C. (2022). Development of a conflict risk evaluation model to assess pedestrian safety in interaction with vehicles. *Accident Analysis & Prevention*, *175*, 106773.

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., Sayed, T., & Mannering, F. (2022). How many are enough?: Investigating the effectiveness of multiple conflict indicators for crash frequency-by-severity estimation by automated traffic conflict analysis. *Transportation research part C: emerging technologies*, *138*, 103653.

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021a). A bivariate extreme value model for estimating crash frequency by severity using traffic conflicts. *Analytic methods in accident research*, *32*, 100180.

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021b). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. *Accident Analysis & Prevention*, *153*, 106016.

Beauchamp, É., Saunier, N., & Cloutier, M. S. (2022). Study of automated shuttle interactions in city traffic using surrogate measures of safety. *Transportation research part C: emerging technologies, 135*, 103465.

Bhuiyan, H., Ara, J., Hasib, K. M., Sourav, M. I. H., Karim, F. B., Sik-Lanyi, C., & Yasmin, S. (2022). Crash severity analysis and risk factors identification based on an alternate data source: a case study of developing country. *Scientific reports*, *12*(1), 21243.

Behbahani, H., Nadimi, N., & AlNoori, H. (2016). Proposing an aggregated model for surrogate safety measures. *Research in Civil and Environmental Engineering*, *4*(1), 1-17.

Bonela, S. R., & Kadali, B. R. (2022). Review of traffic safety evaluation at T-intersections using surrogate safety measures in developing countries context. *IATSS research*. *46*(3), 307–321.

Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention*, *71*, 311-318.

Chandra, S. (2004). Capacity estimation procedure for two-lane roads under mixed traffic conditions. *J. Indian Roads Congress, 165(*1), 139–170.

Chauhan, R., Dhamaniya, A., & Arkatkar, S. (2021). Spatiotemporal variation of rear-end conflicts at signalized intersections under disordered traffic conditions. *Journal of transportation engineering, Part A: Systems, 147*(11), 05021007.

Das, S., Tamakloe, R., Zubaidi, H., Obaid, I., & Ashifur Rahman, M. (2023). Bicyclist injury severity classification using a random parameter logit model. *International Journal of Transportation Science and Technology*.

Dong, S., Khattak, A., Ullah, I., Zhou, J., & Hussain, A. (2022). Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley Additive exPlanations. *International journal of environmental research and public health*, *19*(5), 2925.

Gastaldi, M., Orsini, F., Gecchele, G., & Rossi, R. (2021). Safety analysis of unsignalized intersections: a bivariate extreme value approach. *Transportation letters*, *13*(3), 209-218.

Goyani, J., Paul, A. B., Gore, N., Arkatkar, S., & Joshi, G. (2021). Investigation of crossing conflicts by vehicle type at unsignalized T-intersections under varying roadway and traffic conditions in India. *Journal of transportation engineering, Part A: Systems, 147*(2), 05020011.

Hosseinzadeh, A., Moeinaddini, A., & Ghasemzadeh, A. (2021). Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model. *Journal of Safety Research*, *77*, 151–160.

Hu, Y., Li, Y., Huang, H., Lee, J., Yuan, C., & Zou, G. (2022). A high-resolution trajectory data driven method for real-time evaluation of traffic safety. *Accident Analysis and Prevention*, *165*, 106503.

Hu, J., Huang, M. C., & Yu, X. (2020). Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. *Accident Analysis & Prevention*, *144*, 105665.

Ijaz, M., Zahid, M., & Jamal, A. (2021). A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. *Accident Analysis & Prevention*, *154*, 106094.

Islam, S., Hossain, A. B., & Barnett, T. E. (2016). Comprehensive injury severity analysis of SUV and pickup truck rollover crashes: Alabama case study. *Transportation Research Record, 2601*, 1–9.

Islam, Z., & Abdel-Aty, M. (2023). Traffic Conflict Prediction Using Connected Vehicle Data. *Analytic Methods in Accident Research, 39*, 100275.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., & Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International journal of injury control and safety promotion*, *28*(4), 408-427.

Johnsson, C., Laureshyn, A., & Dágostino, C. (2021). A relative approach to the validation of surrogate measures of safety. *Accident Analysis and Prevention*, *161*, 106350

Johnsson, C., Laureshyn, A., & De Ceunynck, T. (2018). In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators. *Transport Reviews*, *38*(6), 765–785.

Joshi, G., & Vagadia, D. (2013). Dynamic vehicle equivalent factors for characterisation of mixed traffic for multilane metropolitan arterials in India. *J. Indian Roads Congress*, *74*(2), 205-219.

Kannangara, K. P. M., Zhou, W., Ding, Z., & Hong, Z. (2022). Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(4), 1052-1063.

Kashifi, M. T., & Ahmad, I. (2022). Efficient Histogram-Based Gradient Boosting Approach for Accident Severity Prediction With Multisource Data. *Transportation Research Record*, *2676*(6), 236–258.

Komol, M. M. R., Hasan, M. M., Elhenawy, M., Yasmin, S., Masoud, M., & Rakotonirainy, A. (2021). Crash severity analysis of vulnerable road users using machine learning. *PLoS one*, *16*(8), e0255828.

Kumar, A., Paul, M., & Ghosh, I. (2019). Analysis of pedestrian conflict with right-turning vehicles at signalized intersections in India. *Journal of Transportation Engineering, Part A: Systems*, *145*(6), 04019018.

Lee, J., Li, X., Mao, S., & Fu, W. (2021). Investigation of Contributing Factors to Traffic Crashes and Violations: A Random Parameter Multinomial Logit Approach. *Journal of Advanced Transportation*, *2021*, 2836657.

Lord, D., Qin, X., & Geedipally, S. (2021). *Highway safety analytics and modeling*. Elsevier.

Lu, Q. L., Yang, K., & Antoniou, C. (2021). Crash risk analysis for the mixed traffic flow with human-driven and connected and autonomous vehicles. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2021-September*, 1233–1238.

Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis & Prevention*, *160*, 106322.

Mangalathu, S., Hwang, S. H., & Jeon, J. S. (2020). Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Engineering Structures*, *219*, 110927.

Meshoul, S., Batouche, A., Shaiba, H., & AlBinali, S. (2022). Explainable Multi-Class Classification Based on Integrative Feature Selection for Breast Cancer Subtyping. *Mathematics*, *10*(22), 4271.

Ministry of Road Transport and Highways. (2022). *Road accidents in India 2021*. https://morth.nic.in/sites/default/files/RA_2021_Compressed.pdf

Mohanty, M., Panda, B., & Dey, P. P. (2021). Quantification of surrogate safety measure to predict severity of road crashes at median openings. *IATSS research*, *45*(1), 153-159.

Nadimi, N., Amiri, A. M., & Sadri, A. (2022). Introducing novel statistical-based method of screening and combining currently well-known surrogate safety measures. *Transportation Letters*, *14*(4), 385-395.

Orsini, F., Gecchele, G., Rossi, R., & Gastaldi, M. (2021). A conflict-based approach for real-time road safety analysis: Comparative evaluation with crash-based models. *Accident Analysis and Prevention*, *161*, 106382.

Ozbay, K., Yang, H., Bartin, B., & Mudigonda, S. (2008). Derivation and validation of new simulation-based surrogate safety measure. *Transportation research record*, *2083*(1), 105-113.

Paul, M., & Ghosh, I. (2021). Development of conflict severity index for safety evaluation of severe crash types at unsignalized intersections under mixed traffic. *Safety science*, *144*, 105432.

Pawar, N. M., Gore, N., & Arkatkar, S. (2022). Examining crossing conflicts by vehicle type at unsignalized T-intersections using accepted gaps: a perspective from emerging countries. *Journal of transportation engineering, Part A: Systems*, *148*(6), 05022004.

Peng, Y., Abdel-Aty, M., Shi, Q., & Yu, R. (2017). Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. *Transportation research part C: emerging technologies*, *74*, 295-305.

Salas, P., De la Fuente, R., Astroza, S., & Carrasco, J. A. (2022). A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. *Expert Systems with Applications*, *193*, 116253.

Shangguan, Q., Fu, T., Wang, J., & Luo, T. (2021). An integrated methodology for real-time driving risk status prediction using naturalistic driving data. *Accident Analysis & Prevention*, *156*, 106122.

Shangguan, Q., Wang, J., Fu, T., & Fu, L. (2023). An empirical investigation of driver car-following risk evolution using naturistic driving data and random parameters multinomial logit model with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *38*, 100265.

Song, Y., Kou, S., & Wang, C. (2021). Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach. *Journal of safety research*, *76*, 64-72.

Song, P., Sze, N. N., Zheng, O., & Abdel-Aty, M. (2022). Addressing unobserved heterogeneity at road user level for the analysis of conflict risk at tunnel toll plaza: A correlated grouped random parameters logit approach with heterogeneity in means. *Analytic Methods in Accident Research*, *36*, 100243.

Tamim Kashifi, M., & Ahmad, I. (2022). Efficient histogram-based gradient boosting approach for accident severity prediction with multisource data. *Transportation research record*, *2676*(6), 236-258.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, *37*.

Uzondu, C., Jamson, S., & Lai, F. (2018). Exploratory study involving observation of traffic behaviour and conflicts in Nigeria using the Traffic Conflict Technique. *Safety science*, *110*, 273-284.

Wang, C., Xie, Y., Huang, H., & Liu, P. (2021). A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. *Accident Analysis & Prevention*, *157*, 106157.

World Health Organization. (2018). *Global Status Report on Road Safety*. https://www.who.int/publications/i/item/9789241565684

Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, *2*(9), 1-13.

Ye, Y., He, J., Wang, H., Zhang, C., Yan, X., & Wang, C. (2023). Research on Influencing Factors of Traffic Conflicts in Freeway Diverging Area during the Maintenance Period. *Journal of Transportation Engineering, Part A: Systems*, *149*(2), 04022149.

Yu, R., Han, L., & Zhang, H. (2021). Trajectory data based freeway high-risk events prediction and its influencing factors analyses. *Accident Analysis & Prevention*, *154*, 106085.

Yuan, C., Li, Y., Huang, H., Wang, S., Sun, Z., & Li, Y. (2022a). Using traffic flow characteristics to predict real-time conflict risk: A novel method for trajectory data analysis. *Analytic methods in accident research*, *35*, 100217.

Yuan, C., Li, Y., Huang, H., Wang, S., Sun, Z., & Wang, H. (2022b). Application of explainable machine learning for real-time safety analysis toward a connected vehicle environment. *Accident Analysis & Prevention*, *171*, 106681.

Zhang, X., Akber, M. Z., & Zheng, W. (2022). Predicting the slump of industrially produced concrete using machine learning: A multiclass classification approach. *Journal of Building Engineering*, *58*, 104997.

Zheng, L., & Sayed, T. (2019). Comparison of traffic conflict indicators for crash estimation using peak over threshold approach. *Transportation research record*, *2673*(5), 493-502.