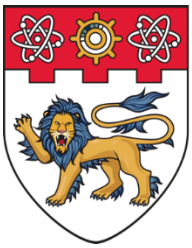


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ACCIDENT RISK ASSESSMENT AND PREDICTION USING SURROGATE INDICATORS AND MACHINE LEARNING

SHI XIUPENG

School of Civil and Environmental Engineering

2019

ACCIDENT RISK ASSESSMENT AND PREDICTION USING SURROGATE INDICATORS AND MACHINE LEARNING

SHI XIUPENG

School of Civil and Environmental Engineering

Thesis submitted to

the Nanyang Technological University

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

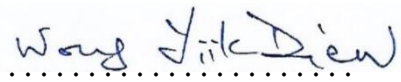
17 Jan 2019 ..... Date

Shi Xiupeng ..... Shi Xiupeng

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Date



Wong Yiik Diew

Authorship Attribution Statement

This thesis contains material from two papers published in the following peer-reviewed journals where I was the first and corresponding author.

Chapter 3 is published as X. Shi, Y.D. Wong, M.Z.F. Li and C. Chai. Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. *Accident Analysis and Prevention*, **117**, 346-356 (2018). DOI: 10.1016/j.aap.2018.05.007.

Chapter 5 is published as X. Shi, Y.D. Wong, M.Z.F. Li, C. Palanisamy and C. Chai. A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis and Prevention*, **129**, 170-179 (2019). DOI: 10.1016/j.aap.2019.05.005.

The contributions of the co-authors are as follows:

- Assoc Prof Wong provided the initial project direction and edited the manuscript drafts, and provided the resources needed to conduct the research.
- I wrote the drafts of the manuscripts, co-designed the study with Assoc Prof Wong and performed the modelling and analysis work at the School of Civil and Environmental Engineering.
- Assoc Prof Li refined the manuscripts.
- Mr Palanisamy contributed to data acquisition of accident video footages from the Land Transport Authority (LTA).
- Assoc Prof Chai assisted in the video-based data extraction of vehicle trajectory, provided guidance in the modelling, and revised the manuscripts.

17 Jan 2019 ..... Shi Xiupeng .....

Date Shi Xiupeng

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Assoc Prof Wong Yiik Diew, and my co-supervisor, Assoc Prof Li Zhi-Feng, Michael, for their valuable guidance and strong support.

I would like to express my appreciation to all the people who have helped me in my 3 years of PhD research. Thanks to Mr Chandrasekar from LTA for providing the accident data. Thanks to the committee members, Assoc Prof Wang Zhiwei, David, and Asst Prof Liu Fang, for their insightful suggestions. Thanks to those anonymous reviewers whose comments lead to a big improvement in my research papers. Many thanks also go to my fellow PhD students and colleagues in the Centre for Infrastructure Systems, for friendship and collaboration.

I am most grateful to the NGSIM Program for the availability of data, and to the contributors in the fields of machine learning and Python, among others.

My deepest gratitude is to my wife, Dr. Chai, and our beloved parents and son, thanks for your brilliant support and love in my life.

TABLE OF CONTENTS

SUMMARY

Road traffic accidents cause a great loss of lives and property damage. Reliable accident prediction and proactive prevention are undoubtedly of great benefit and necessity. This study focuses on the risk assessment and prediction of traffic accidents associated with vehicle conflicts, using machine learning and surrogate indicators to achieve vehicle-level risk rating and prediction based on instantaneous driving behaviours.

Accident events are generally unexpected and occur rarely. Pre-crash risk assessment by surrogate indicators is an effective way to identify risk levels, and thus boost crash prediction. Herein, the concept of Key Risk Indicator (KRI) is proposed, which assesses risk exposures using hybrid indicators. To evaluate the feasibility of indicator-based risk assessment, a typical real-world chain-collision accident (on Singapore's expressway) and its antecedent (pre-crash) road traffic movements are retrieved from surveillance video footage, and a grid remapping method is proposed for data extraction and coordinates transformation. Seven surrogate measures of traffic conflicts are assessed based on a temporal-spatial case-control comparison of which two surrogate measures are found to be more efficient in identifying pre-accident risk conditions, namely, Time Integrated Time-to-collision (TIT) and Crash Potential Index (CPI). Hence, the KRIs are formulated based on the hybrid of TIT and CPI, which hierarchically distinguish various risk levels. TIT enables the capture of risk signals (when TIT>0), while CPI further identifies the more severe ones (i.e. those conditions for near-crashes) (when CPI>0). Besides, the thresholds of risk levels in KRIs are more straightforward to define. For a rigorous validation, the results are examined by another independent real-world accident sample. Verified by real-world accidents, KRIs make a breakthrough in indicator-based risk assessment, and reveal new insights about pre-crash risk exposures.

From another perspective, indicator-based risk assessment is extended to general traffic streams. The unsupervised vehicle-level risk rating is achieved by clustering and the extraction of risk indicator features. The risk grading pertains to a distinctly imbalanced problem, with some intrinsic challenges. Based on the findings in KRIs,

a total of 12 risk indicator features are designed, which represent vehicle risk exposures in terms of temporal, kinematical and spatial aspects. To obtain reliable and robust partitioning of risk levels, an ensemble clustering model is built by majority voting of the risk labels produced by multiple clustering. The clustering is conducted on a large group of vehicles within a road segment. Based on pattern similarity, vehicles are clustered into distinct groups with graded risk labels. Clustering is performed in a progressive manner to obtain hierarchical partitioning of risk levels, which facilitates to identify the highest risk level. Moreover, label identification by classifiers is proposed to evaluate the clustering performance and determine the risk levels. Herein, vehicle trajectory data from the United States' NGSIM Program is used as a case study, and risk grading with six levels is established. The risk indicator features based on TIT and CPI are found with higher importance, according to feature ranking by random forest. Besides, a high-resolution risk mapping and positioning is demonstrated to delineate the risk potentials, including at-risk vehicles, locations and timestamps, as well as risk patterns (e.g. severity, frequency, trends). The proposed method is found to be effective to assess detailed risk potentials inherent to driving behaviour as exhibited by the general vehicle trajectory, and generate unsupervised data labelling of risk levels.

Furthermore, the linkages of risk levels and driving behaviours are explored, which empower behaviour-based risk prediction. An integrated feature learning framework is designed, to assess vehicle driving and predict risk levels. The framework integrates learning-based feature selection, unsupervised risk rating, and imbalanced data resampling. For each vehicle, about 1,300 driving behaviour features are extracted from trajectory data, which produce in-depth and multi-view measures on behaviours. To estimate risk potentials of vehicles driving on the roads, unsupervised risk rating is conducted using fuzzy C means (FCM), and four risk levels are used for data labelling. Besides, data under-sampling of the safe group is performed to reduce the risk-safe class imbalance. Afterwards, the linkages between behaviour features and corresponding risk levels are built using XGBoost, and key features are identified according to feature importance ranking and recursive elimination. The risk levels of vehicles in driving are predicted based on key features selected. As a case study,

NGSIM trajectory data are used in which four risk levels are clustered, 64 key behaviour features are identified, and an overall accuracy of 91.66% is achieved for behaviour-based risk prediction. Findings show that this approach is effective and reliable to identify important features for driving assessment, and achieve an accurate prediction of risk levels.

Finally, a domain-specific automated machine learning (AutoML) is built, which enables end-to-end learning from the driving behaviour data to detailed risk levels and corresponding key features. The AutoML assembles all necessary machine learning steps as an end-to-end pipeline and automates the pipeline to get the features, models, and hyperparameters that return the best performance as measured on validation sets. The AutoML platform has a self-learning and auto-optimisation mechanism, which can be easily updated by introducing the most advanced algorithms. Bayesian optimisation guides the self-learning of AutoML by effectively auto-tuning the hyperparameters and exploring the pipeline space for better performance. The identification of key features not only helps to produce better results with fewer computation costs, but also provides data-driven insights about system optimisation and sensing configuration. Application potentials are discussed, and the AutoML can be used in the risk decision-making and motion trajectory planning of autonomous vehicles (AVs) and ADAS (advanced driver assistance systems), pay-how-you-drive (PHYD) insurance, driving safety system under the connected vehicle environment, and short-term near-real-time crash prediction, among others. These studies contribute to traffic safety by providing a portfolio of techniques, ranging from vehicle-level crash risk prediction to personalised driving behaviour enhancement, which enables the development of effective measures and systems to reduce the likelihood of crashes.

LIST OF PUBLICATIONS

This thesis includes part of contents that have been published in the papers:

Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C. and Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis and Prevention, 129, 170-179. DOI: 10.1016/j.aap.2019.05.005.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2018). Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. Accident Analysis and Prevention, 117, 346-356. DOI: 10.1016/j.aap.2018.05.007.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2018). Accident risk prediction based on driving behavior feature learning using CART and XGBoost. Transportation Research Board 97th Annual Meeting, No. 18-06270.

This thesis includes part of contents in the following submitted journal manuscripts:

Shi X., Wong Y.D., Li M.Z.F., Palanisamy, C. and Chai C. (2018). Feature extraction and clustering for vehicle-level risk grading based on surrogate indicators. Accident Analysis and Prevention.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2019). An automated machine learning (AutoML) method of risk prediction for the decision-making of autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2019). Driving risk prediction and data-driven design for pay-how-you-drive (PHYD) insurance: An automated machine learning (AutoML) approach. Transportation Research Part A: Policy and Practice.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2019). Driving safety system under the connected vehicle environment based on automated machine learning

(AutoML). Transportation Research Part C: Emerging Technologies.

Other papers:

Chai C., Shi X. and Wong Y. D. (2017). Estimating safety effects of green-man countdown devices at signalized pedestrian crosswalk based on cellular automata. Journal of Advanced Transportation. DOI: 10.1155/2017/8391325.

Chai C., Shi X., Wong Y. D., Er M. J. and Gwee E. T. M. (2016). Fuzzy logic-based observation and evaluation of pedestrians' behavioral patterns by age and gender. Transportation Research Part F: Traffic Psychology and Behaviour, 40, 104-118. DOI: 10.1016/j.trf.2016.04.004.

Chai C., Shi X. and Wong Y. D. (2017). Safety evaluation of vehicle-to-vehicle (V2V) communications system on motorcycle-vehicle interaction based on fuzzy cellular automata (FCA). Transportation Research Board 96th Annual Meeting, No. 17-01099.

Chen, T., Shi, X. and Wong, Y. D. (2019). Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data. Accident Analysis and Prevention, 129, 156-169. DOI: 10.1016/j.aap. 2019.05.017.

LIST OF TABLES

TABLE OF FIGURES

LIST OF SYMBOLS

α	Deceleration rate to stop
$\delta_i(t)$	Switching variable with value 1 and 0
ε_m	In-cell length
d	Maximum acceptable deceleration rate
$d(P, l_m)$	Distance from a point P to a line l_m
F_i	Feature vector of instance i
$f(\cdot)$	Feature extractor
L_i	Vehicle length of vehicle i
$M_i(t_{a\to b})$	Movement trajectory of a vehicle i from time t_a to t_b
$P_i(t)$	Position of vehicle i at time t
τ_{sc}	A small time interval
Δt	Driver's reaction time
s_0	Initial distance
s_x	Conversion coefficient in length
s_y	Conversion coefficient in width
$v_i(t)$	Velocity of vehicle i at time instant t
(x_i, y_i)	Position coordinates of vehicle i in real-world units
(x_i^p, y_i^p)	Pixel coordinates of vehicle i in the image system
(x_i^c, y_i^c)	Cell address coordinates in the grid mesh system
(x_i^g, y_i^g)	Grid coordinates

CHAPTER 1
INTRODUCTION

1.1 Chapter Introduction

This chapter introduces the motivation, aims, research scope and significance of this research study, as well as the organisation of this thesis.

1.2 Motivation

Traffic accidents cause a great loss of lives and property damage, as well as severe negative social-economic impacts. Reliable accident prediction and proactive prevention are undoubtedly of great benefit and necessity.

Accident events are generally unexpected and occur rarely, and the accident occurrence is a complex mechanism, with many contributing factors. Numerous studies have been conducted on traffic accident prediction, risk assessment and causal analysis. Unsafe traffic conditions and risky driving behaviours have been explored to characterise accidents, including human errors, traffic speed and occupancy, weather and visibility. Statistical models and machine learning approaches are being widely applied to analyse the relationship between accidents and influencing factors, such as, Bayesian method (Pei et al., 2011), random forests (Abdel-Aty and Haleem, 2011). These studies are helpful to describe general linkages between accident numbers (e.g. occurrences, frequency rate) and coexisting factors or concurrent scenarios. Nevertheless, even under equivalent situations, actual accident occurrence remains unreliable to be assessed or predictable if merely relying on these trends and factors. Due to uncertainty and randomness, effective accident assessment and prediction has been found to be extremely difficult.

Risk assessment is essential when making any accident prediction. Pre-accident risk exposure is more meaningful for accident prediction and prevention. However, it is extremely difficult to obtain real-world pre-accident data of high quality, while noting that it is near impossible to pinpoint the precise time and location of an accident

before-hand. Although the occurrence of an accident is generally unexpected, for certain types of accidents, there is an accident-forming process, especially for accidents associated with traffic conflicts. Conflicts can improve the understanding of the accident mechanism and chain of events which may lead to a collision (Mahmud et al., 2017). Compared with actual accidents, incidences of traffic conflicts, with attendant collision risks of various degrees, are more frequent (Chin and Quek, 1997). Moreover, a strong relationship has been found between traffic conflicts and actual crashes in many studies. Herein, the scope of risk assessment should therefore focus on pre-accident traffic conflicts, as an alternative to actual accident occurrences. A clear and convincing grading scheme of risk levels is of great interest.

In addition, the identification of risk factors plays a key role in risk assessment and prediction. Generally, driver-centric factors could be found in most accidents, and driving behaviour assessment is an important aspect to enhance safety and reduce crashes. There is a perennial quest about assessing driving behaviour and predicting crash risk potentials in driving. Kinetics-related behaviour factors are commonly used to provide clues on crash risk (Perez et al., 2017). However, an in-depth and multi-view extraction and evaluation of driving behaviours remains lacking. For factor identification and risk prediction, machine learning has an advantage in complex systems, and works well on handling voluminous information and discovering complex patterns. Nevertheless, the machine learning for risk assessment and prediction has not been well investigated, as well as the identification of important behaviour factors that are useful in modelling. In particular, it remains unclear the system framework and predictive power of advanced machine learning in risk prediction based on behaviour.

1.3 Research Aims

This study focuses on the risk assessment and prediction of traffic accidents associated with vehicle conflicts, using machine learning and surrogate indicators, to achieve vehicle-level risk assessment and prediction based on driving behaviour. Three primary objectives of the research are as follows:

(1) Reliable surrogate indicators for crash risk assessment. Pre-crash risk assessment by surrogate indicators is an effective way to identify risk levels and thus boost crash prediction. The indicators are expected to reveal predictive insights about the potential of a crash, including pre-crash risk exposures, at-risk vehicles, locations and time. Furthermore, indicator thresholds should be straightforward and flexible to define. The indicators can enable one to evaluate risk severity with detailed levels, as well as the likelihood, and the feasibility should be evaluated based on real-world accidents.

(2) A powerful domain-specific machine learning system for risk assessment and prediction. The system is designed to achieve accurate crash risk prediction and targeted crash inference, and overcome the pertinent problems and challenges in crash prediction. Risk levels of vehicles in driving can be assessed in an unsupervised or semi-supervised manner. Targeted vehicles with higher risk levels can be identified proactively, and hence crash prevention countermeasures can be conducted prospectively. The system should be transparent, scalable and flexible for applications. Moreover, for potential real-world applications, an AutoML (automated machine learning) is to be developed, which achieves an end-to-end self-learning and auto-optimisation from general data to risk levels and key features.

(3) Effective identification of risk factors and unsafe behaviours. Pre-crash risk factors play an important role in crash prediction, especially factors in terms of unsafe driving behaviour. The deep mining of unsafe behaviour provides insightful predictors for reliable driving assessment and risk prediction, and early signals may also be found. In-depth and multi-view feature extraction is necessary, especially features with predictability. The identification of key behaviour features is an important aspect to enhance driving assessment and prediction performance.

1.4 Scope of Work

The foundations of this study are crash risk assessment based on traffic conflict techniques, and risk prediction based on machine learning integrated with feature engineering.

Hybrid indicators are developed to hierarchically assess pre-crash risk exposures, based on surrogate measures in traffic conflict techniques (TCT). Surrogate measures of accident risks conditioned on pre-crash vehicle movements are more practical and meaningful than actual accident occurrences, which provide sufficient motivation to capture early signals and trigger factors for the purpose of accident assessment and prediction. The reliability and validity of surrogate indicators are investigated based on real-world accidents. Two typical real-world chain-collision accidents and their antecedent (pre-crash) road traffic movements are retrieved from surveillance video footages, which are used to evaluate the feasibility and effectiveness of indicator-based risk assessment. The two accidents occurred during day-time on the expressways, as visibility is required to extract data. Day-time also means there is unlikely to be drink driving situation in the drivers.

Crash prediction emphasises on the prediction of risk levels based on driving behaviours. For machine learning aspect, the focuses are feature extraction and selection, and the integration of algorithms. A repertoire of machine learning algorithms is tested. To estimate risk potentials of vehicles in driving, clustering algorithms are applied to build a system of unsupervised risk rating, and risk indicator features for clustering are extracted. Risk levels are predicted based on driving behaviours as exhibited by general vehicle trajectory. Massive driving behaviour features are extracted to produce in-depth and multi-view measures on behaviours, and key features are identified based on supervised learning. The likelihood of a crash can be inferred based on predicted risk exposures. Vehicle trajectory data from the US NGSIM Program is used as a case study.

1.5 Organisation of the Report

The report is divided into seven chapters including this introductory chapter.

Chapter 2 reviews the key issues as reported in the literature about this research topic. The literature review highlights the fundamental concepts about traffic accident risk, risk assessment, and machine learning for prediction.

Chapter 3 assesses the surrogate measures of crash risk based on real-world accident

data. The feasibility of using surrogate indicators for pre-crash risk assessment is evaluated based on real-world accident cases, and Key Risk Indicators (KRIs) are developed to assess crash risks associated with vehicle conflicts. Accident data extraction and grid remapping method are described in this chapter.

Chapter 4 proposes the method of unsupervised risk rating for risk assessment, which extends indicator-based risk assessment from accident cases to normal traffic flow. Risk indicator features are extracted based on KRI as developed in the preceding task, and ensemble clustering by majority voting is proposed to create hierarchical risk grading of vehicles in driving. The risk levels are evaluated based on label identification by classifiers. The method also aims to generate unsupervised data labelling of risk levels.

Chapter 5 explores the linkages of risk levels and driving behaviours, which empower behaviour-based risk prediction. An integrated feature learning approach is designed, to identify important features for driving assessment, and achieve accurate prediction of risk levels. This approach combines learning-based feature selection, unsupervised risk rating (as developed in Chapter 4), and imbalanced data resampling. Massive driving behaviour features are extracted from vehicle movement trajectory, and key features are identified. The risk levels of vehicles in driving are predicted.

The shift from unsupervised learning (Chapter 4) to supervised learning (Chapter 5) has several practical implications: firstly, it is using a variable that is easy to measure, to predict the variable that is not easy to measure; secondly, the target shifts from a large number of vehicles to individual vehicles, which has more application scenarios; thirdly, it facilitates near-real-time prediction.

Chapter 6 develops an end-to-end AutoML based on the key findings and techniques developed in the preceding chapters. Besides, the systems and proof-of-principle prototype demonstration for downstream potential applications are discussed, including risk decision-making for autonomous vehicles (AVs), pay-how-you-drive (PHYD) insurance, driving safety system under the connected vehicle environment, and short-term crash prediction.

Chapter 7 covers a summary of accomplishments and key contributions in this research study, as well as the suggestions for future work.

CHAPTER 2
LITERATURE REVIEW

2.1 Chapter Introduction

This chapter reviews the key issues about this research topic. The literature review highlights the fundamental concepts about traffic accident risk, risk assessment, and machine learning for prediction. Firstly, the traffic accident situation in Singapore is introduced in Section 2.2. Current studies about traffic accident risk assessment and predictive analysis are discussed in Sections 2.3 and 2.4. In Section 2.5, machine learning approaches are discussed for crash prediction. Section 2.6 introduces feature extraction and selection. Ending section is the chapter summary, and the gaps in prevalent research and attendant opportunities are identified.

2.2 Traffic Accident Situation in Singapore

From the Traffic Police's 2013 to 2017 annual reports, the overall road safety situation is continuing to make improvements since 2011. There was a decrease in the number of fatal traffic accidents and fatalities, this being a continuation of a downward trend in fatalities since 2011. However, there was a slight increase in the number of accidents that resulted in injuries.

The traffic accident occurrences of different categories for the year 2014 are briefly shown in Table 2.1.

Singapore authorities harness technologies to improve road safety, such as Expressway Monitoring Advisory System (EMAS), i-Transport system, etc. Such smart traffic systems play extraordinary roles in managing traffic safety and congestion, and maintain the traffic problems within manageable levels. Along with existing systems based on real-time information, accurate information on vehicle-level risk prediction could add more values.

Table 2.1 Traffic accident situation in Singapore for the year 2014*

Accident classification	Slight injury	Serious injury	Fatal	Total
Accident occurrences	7,526	281	152	7,959
Percentage (%)	94.56	3.53	1.91	
Vehicle movement and conflict types				
Between moving vehicles	5,030	136	56	5,222
Percentage (%)	63.20	1.71	0.70	65.61
Single vehicle self-skidded	1,432	54	35	1,521
Percentage (%)	17.99	0.68	0.44	19.11
Vehicle against pedestrian	823	77	47	947
Moving vehicle and stationary vehicle	162	9	8	179
Single vehicle against object	16	1	3	20
Others	79	5	6	90
Road traffic types				
Two-way divided road	2,935	102	45	3,082
Percentage (%)	36.88	1.28	0.57	38.72
Expressway	1,750	52	34	1,836
Percentage (%)	21.99	0.65	0.43	23.07
Intersection	1,278	63	36	1,377
Two-way undivided road	325	15	17	357
One-way road	426	21	11	458
Others	812	28	9	849
Road surface conditions				
Dry	6,670	260	137	7,067
Percentage (%)	83.80	3.27	1.72	88.79
Wet	752	18	14	784
Others	104	3	1	108

* Data source: Singapore Police Force

Table 2.1 (continued) Traffic accident situation in Singapore for the year 2014*

Accident classification	Slight injury	Serious injury	Fatal	Total
Weather conditions				
Fine	6,942	264	142	7,348
Percentage (%)	87.22	3.32	1.78	92.32
Light rain	333	10	9	352
Heavy rain	156	3	1	160
Others	95	4		99

* Data source: Singapore Police Force

Annual statistics on traffic accidents are contained in Singapore Police Database from 2009 to 2014. The accident data is collected from the accident events, and the main information includes vehicle movement type, likely fault and cause, time and location, weather condition, road surface, vehicle type, damage and injury degree, etc. Accident reporting data is useful to analyse the links between accident events and concurrent risk factors, as well as coexisting scenarios associated with the accidents.

2.3 Traffic Accident Risk Assessment

2.3.1 Accident assessment

As mentioned in Chapter 1, accident occurrence is a complex mechanism, with many contributing factors (Mannering et al., 2016), including unsafe traffic conditions (e.g. traffic speed and occupancy, weather and visibility) and risky driving behaviours (e.g. human errors, aggressive driving) (e.g. Saifuzzaman and Zheng, 2014; Young, 2017). Many statistical models and machine learning have been explored to analyse the relationships between accidents and influencing factors, and make accident prediction and risk assessment, such as random forests (Abdel-Aty and Haleem, 2011), Bayesian method (Pei et al., 2011), support vector machine (Dong et al., 2015), among others. The general linkages between accident numbers (e.g. occurrences, frequency rate) are described, as well as coexisting factors and concurrent scenarios. Nevertheless, the risk assessment and prediction of actual

accident occurrence remain difficult and unreliable, if merely relying on these general trends and macro factors, even under equivalent situations. Besides, the uncertainty and randomness make accident assessment and prediction extremely difficult.

2.3.2 Accident risk hierarchy

The relationship between accident events and risk conditions is able to be described from severity hierarchy. Various severity hierarchy models were proposed to rank the traffic events in a safety continuum according to the severity and frequency, such as pyramid hierarchy model (Hydén, 1987), and diamond-shaped hierarchy model (Svensson, 1998). For example, the pyramid hierarchy model is shown in Figure 2.1. From the visual representations of safety continuum, a clear description is provided about the relationships among accident events and conflict risk conditions (Zheng et al., 2014).

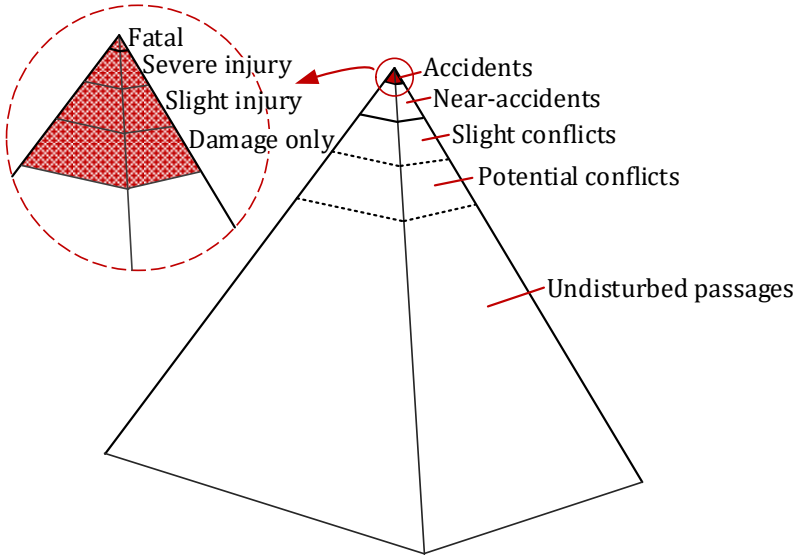


Figure 2.1 Hydén pyramid hierarchy model (Hydén, 1987)

The severity of risk condition is reflected by the severity of traffic conflicts (i.e. the proximity to a crash), while the severity of accident is typically defined by the possible consequences of the crash, such as damage only, slight injury, serious injury or fatal (Laureshyn et al., 2010; Bagdadi 2013). In addition, the accident severity is often correlated with the collision type (e.g. rear-end, run off road, right angle, sideswipe, etc.) and human factors (e.g. health conditions, age, protection measures,

etc.) (Zheng et al., 2014). Hence, accident severity is a more complex issue, and this study is mainly focused on the severity of risk conditions, especially for crash risk involving traffic conflicts.

2.3.3 Traffic conflicts

Conflict is an observed situation in which two or more road users approach each other in space and time to such an extent that there is a risk of collision if their movements remain unchanged (Amundsen and Hydén, 1977). Traffic conflict represents a transitional state between safety and near-crash state of a potential impending accident. The near-crash state is extremely useful in highlighting the main factors triggering or influencing crashes and their severity (Perez et al., 2017).

A near-crash is a circumstance that requires a rapid, evasive manoeuvre to avoid a crash; such manoeuvre is defined as steering, braking, accelerating, or any combination of control inputs that approaches the limits of the vehicle's capability (Guo et al., 2010). Herein, near-crash means a high likelihood of accident occurrence in a probabilistic sense, but the accident may or may not occur.

The traffic conflict precedes the evasive action, which can be either successful (leading to safe continuation) or not successful (leading to a collision) (Zheng et al., 2014). Although there is no agreement on the exact correlation between traffic conflicts and crashes, however, due to practical difficulties, the need for the validity of such correlation is overly exaggerated and even unnecessary (Chin and Quek 1997; Archer 2005).

2.3.4 Surrogate measures

Crash prediction is a proactive way for traffic safety, and it is important to identify the hazards and associated risks that may lead to accidents. This motivates the development of surrogate measures of accident and risks, and traffic conflict techniques (TCT) are the most prominent methods (Zheng et al., 2014). TCT analyses the risk conditions from the aspect of more observable traffic events than crashes, and there is a reasonable agreement on the usefulness of TCT as surrogate or

complementary measures (Dallat et al., 2017).

The severity of traffic conflicts is identified by either the intensity of evasive actions or the proximity in time and (or) space (Zheng et al., 2014). The temporal and/or spatial proximity provides a quantitative way to distinguish conflicts and non-conflict conditions, but the boundary is threshold sensitive and is not clearly defined. An important observation of the threshold variety is the unobserved heterogeneity (Svensson, 1998; Mannering et al., 2016). Moreover, the temporal and/or spatial proximity improves the understanding of the dynamics of a crash in a Newtonian mechanics view. Besides, the risk level (severity of conflict) and crash outcomes (severity of collision) are also estimated.

Surrogate measures are widely utilised in TCT as an alternative to actual crash data for safety assessment (e.g. Zheng et al., 2014). Mahmud et al. (2017) provides a comprehensive review on the developments and applications of 17 proximal surrogate indicators. The reliability and validity of surrogate indicators are well accepted for safety evaluation (Chai and Wong, 2015a; So et al., 2015), but such studies were based mostly on simulation and designed experiments (Chai and Wong, 2014). In practice, FHWA developed a Surrogate Safety Assessment Model (SSAM) as a post processor to determine the number and severity of conflicts obtained from simulation packages (Sobhani et al., 2013). Many advanced driving assistance systems (ADAS) have used surrogate indicators as important warning criteria (Wang et al., 2013).

Nevertheless, the effectiveness of surrogate indicators under real-world accidents has not been properly investigated. In particular, it remains unclear the extent to which the surrogate indicators are useful for pre-crash risk assessment. Besides, indicators are often designed under simplified assumptions, such as unchanged trajectory, constant speed and predefined deceleration rate. To represent complex crash mechanisms, integrated use of various indicators has been suggested (e.g. Laureshyn et al., 2010), since different measures provide different cues and underlying information of safety and risk. However, no consensus has been reached yet on which indicators should be selected and how to combine them (Guido et al., 2011). Besides,

the different indicators would have diverse suggestions in identifying risks, making it difficult to manage and interpret.

2.3.5 Risk indicators

Various surrogate indicators have been proposed to measure traffic conflicts and safety. Such an indicator is a mathematical calculation based on traffic-related variables that aims to measure traffic conflicts. Many indicators have been developed and new variants continue to be developed with the aim of getting better measurements. Surrogate indicators involving traffic conflict techniques are summarised in Table 2.2.

To measure vehicle conflicts and subsequent likelihood of a collision, surrogate indicators are well-recognised and frequently used in practice. Different surrogate indicators represent different aspects of crash risk associated with traffic conflict, including risk behaviour, risk condition, and risk avoidance.

Time to Collision (TTC) is prominent and widely used in microscopic traffic conflicts and safety simulation (Chai and Wong, 2016), and many ADAS systems have used TTC-based indicators as the main warning criteria. The integration relationship of TIT suggests a risk accumulation both in exposition time and severity values within a risk threshold.

In addition, indicators not only provide an interpretable assessment of risky driving behaviours, but also identify vehicles engaging in risky behaviours. The feasible combination of key indicators remains unclear for crash risk measurement and prediction.

Table 2.2 Surrogate risk indicators involving traffic conflict techniques

Indicator	Definition	Risk Measurement	References
Time to Collision (TTC)	The time until a collision between two vehicles would have occurred if the collision course and speed difference are maintained	Rear-end, head-on, turning, weaving, hit fixed or moving objects, crossing and hit pedestrian, etc.	Hydén, 1996
Time Exposed TTC (TET)	The length of time a TTC-event remains within a designated TTC-threshold	Same as TTC; Considers the elapsed time exposed under risk condition	Minderhoud and Bovy, 2001
Time Integrated TTC (TIT)	The integral of the TTC-profile during the time within the threshold	Same as TTC; Interprets the severity and the likelihood of a probable crash	Minderhoud and Bovy, 2001
Modified TTC (MTTC)	All potential longitudinal conflict scenarios due to acceleration or deceleration discrepancies	Multi-vehicle longitudinal conflict; Considers acceleration in an evolution process	Ozbay et al., 2008
Crash Index (CI)	Influence of speed on kinetic energy involved in potential collisions	Same as MTTC; Considers the elapsed time, severity and likelihood	Yang et al., 2010
Time Headway (H)	The elapsed time between two vehicles passing the same location	Mainly for conflicts related to follow up manoeuvre; Easy to measure	Evans, 1991
Post-Encroachment Time (PET)	The time difference between the moment of an offending vehicle passing out the potential collision area and the moment of conflicted vehicle’s arrival	Conflicts that occur at intersections; Right angle collision; Easy to measure; Represents driving behaviour	Allen et al., 1978

Table 2.2 (continued) Surrogate risk indicators involving traffic conflict techniques

Indicator	Definition	Risk Measurement	References
Proportion of Stopping Distance (PSD)	The ratio of the remaining distance to the potential collision point and the minimum acceptable stopping distance.	Single vehicle conflict, overturning, hit object; Provides more vehicles interaction and time exposure to conflict; Based on evasive actions	Allen et al., 1978
Potential Index for Collision with Urgent Deceleration (PICUD)	The distance between two consecutive vehicles when they abruptly break and stop completely	Mainly lane changing; Similar to TTC; More sensitive for vehicles with similar speeds and traffic condition change detection	Iida et al., 2001
Deceleration Rate to Avoid Crash (DRAC)	Differential speed between a following vehicle and corresponding lead vehicle divided by closing time	Rear-end, hit object/ pedestrian, merging and diverging manoeuvres; Explicitly considers the role of differential speeds and decelerations; Identifies potential traffic conflict situation	Cooper and Ferguson, 1976; Almqvist et al., 1991
Crash Potential Index (CPI)	The probability that a given vehicle DRAC exceeds its maximum available deceleration rate (MADR) or braking capability during a given time interval	Same as DRAC; Addresses vehicle braking capability for prevailing road and traffic conditions	Cunto and Saccomanno, 2007
Criticality Index Function (CIF)	The ratio of the speed squared and TTC	Right angle collision; Similar to TTC; Evaluates probability and severity of a collision	Chan, 2006

2.4 Crash Risk Prediction

2.4.1 Accident predictive analysis

Numerous studies have been conducted on traffic accident prediction and causal analysis. Most of the existing studies conduct the prediction based on historical crash analysis and data mining on accident information (e.g. time, location, weather conditions, driver behaviours such as alcohol or speeding). Examples of traffic accident predictive analysis are summarised in Table 2.3.

Table 2.3 Traffic accident predictive analysis

Prediction issue	Research focus	References
Accident occurrence and frequency prediction	Predict the number of accident occurrences according to specific area and period based on historical analysis (static)	Lord and Mannering, 2010; Ahmed et al., 2012; Roshandel et al., 2015; Huang et al., 2016
Accident severity prediction	Predict the consequences of an accident in terms of damage and time duration	Kunt et al., 2011; Xu et al., 2013; Yu and Abdel-Aty, 2014
Real-time accident (risk condition) prediction	Predict the accident likelihood (probability of accident occurrence) under certain conditions for a road segment	Ahmed and Abdel-Aty, 2013; Lin et al., 2015; Wang et al., 2015; Barraclough et al., 2016
Targeted accident (event) identification and prediction	Proactive prediction of accident likelihood, with targeted vehicles, location and time	/

A lot of predictive insights about traffic accidents are identified, such as high-frequency area and time (dangerous spots), accident-prone conditions, real-time prediction of accident likelihoods (risk conditions), etc. (Wang et al., 2013; Theofilatos and Yannis, 2014; Dezman et al., 2016). Such predictive studies improve

the understandings of crash prediction, but are basically reactive ways at the traffic flow level. Due to the rarity and randomness, the prediction of exact accident events is very difficult (Wang et al., 2013). Targeted prediction aims for the identification of traffic accident hot spots, including time, location and involved vehicles.

Aggregate data is used in most studies, which is extracted from loop detectors, accident reports, video surveillance system, automatic vehicle identification (AVI) system, etc. (Ahmed and Abdel-Aty, 2012; Roshandel et al., 2015). Under-reporting, over-dispersion, and within-period variation are likely to be produced when using such aggregate data (Ma, 2009; Lord and Mannering, 2010), which can result in biased and inconsistent model estimates and erroneous inferences. Moreover, aggregate data may result in important information loss and unobserved heterogeneity, and some factors may change significantly within the aggregated time period (Lord and Mannering, 2010). Hence, high-resolution data should be used for a better prediction, such as vehicle-level high-frequency data (Theofilatos and Yannis, 2014).

2.4.2 Accident influencing factors

Accidents are widely acknowledged to be a systems phenomenon (Dallat et al., 2017), with the compound effect of numerous factors (Salmon et al., 2017). Notable progress has been made in identifying factors contributing to accidents (Lord and Mannering 2010; Savolainen et al., 2011). Multiple risk factors have been explored to characterise accidents, and the impacts and combined effects are also investigated, as summarised in Table 2.4. In addition, according to the factors contributing to crash occurrence in the United States (US), the human factors were shown to influence 93% of all crashes (Lum and Reagan, 1995). The critical reasons for crashes investigated in the US national motor vehicle crash causation survey are shown in Figure 2.2, and about 94% is driver-related errors (Singh, 2015).

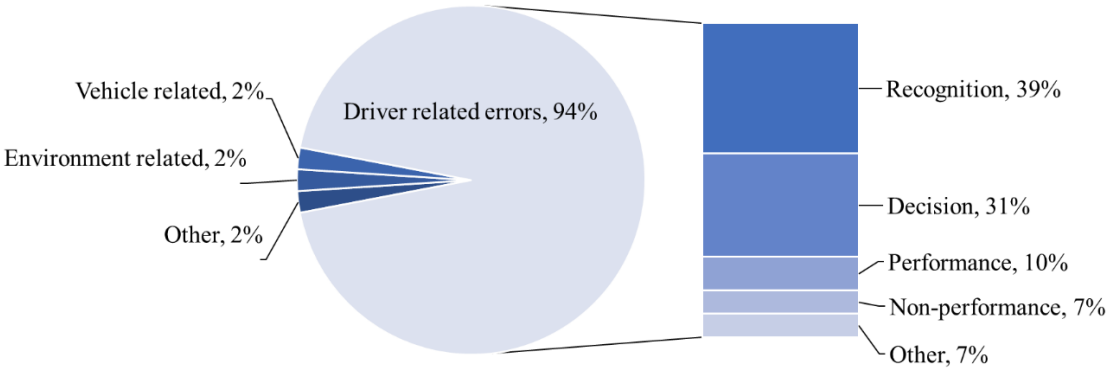


Figure 2.2 Critical reasons for crashes from causation survey

It is found that specific risk factors and combinations are relatively more accident-prone than others (Roshandel et al., 2015). However, some mixed effects and contradictory results were also founded (Theofilatos and Yannis, 2014). A reason is that aggregate data is used as the basis in most findings, and conclusions may vary with different data quality. Besides, microscopic factors and factors immediately prior to a crash remain lacking. However, the exact crash location and time are hard to obtain, leading to potential imprecision and ambiguity (Kockelman and Ma, 2010), and risk exposure over a long term (e.g. one hour or more) has a less obvious direct linkage to crashes (Roshandel et al., 2015). In addition, important behaviour-related factors are ignored, which may lead to inaccuracy and bias (Zheng, 2012). Moreover, drivers also have behaviour adaption, for example, being more alert under low visibility, being more cautious and reducing speed under adverse weathers (Bergel-Hayat et al., 2013). Such behavioural adaption and adjustment may counterbalance the adverse risk conditions (Bergel-Hayat et al., 2013; Theofilatos and Yannis, 2014). Without detailed behaviour information, the risk factor analysis may result in biased findings.

Risky driving behaviours play major roles in crash risk. Risky driving behaviour may result in vehicle collisions, conflict interactions, disturbance to traffic flow, and other risk conditions. Most risky driving behaviours are able to be reflected within the vehicle trajectory, such as excessive speeding, driving too close to the preceding vehicle, etc. It is an acceptable assumption that a severe vehicle conflict will result in a higher likelihood of a crash. Risky conditions are also important for risk assessment

and crash prediction. It is desirable that indicators are introduced to consider the risk level of external conditions. Unsafe conditions are compound effects of risky behaviours and unsafe traffic-related factors. A vehicle is subsumed in a risk condition if the surrounding vehicles are engaged in risky behaviours.

Table 2.4 Risk factors and variables

Risk factors	Factor variables	Risk conditions	References
Traffic and road characteristics	Traffic flow and volume, speed, density, level of congestion, traffic oscillations, road geometry (road curvature, surface condition, lane markings)	Higher speed variation and density (little headways); speed variation increase with an average speed decrease in downstream; stop and go driving in queued traffic	Zheng et al., 2010; Ahmed and Abdel-Aty, 2012; Quddus, 2013; Xu et al., 2013; Theofilatos and Yannis, 2014;
Weather conditions	Precipitation (rain, snow) intensity and duration, air temperature, visibility, wind speed	Extreme weather or sudden changes; low visibility due to fog	Koetse and Rietveld, 2009; Abdel-Aty et al., 2012; Yu et al., 2013
Behavioural factors	Driver mental state, fatigue, distraction, unsafe driving (tailgating, weaving in and out)	Excessive driving speed; aggressive driving; inappropriate speed for specific conditions	Young et al., 2011; Dong et al., 2011; Kang, 2013; Dingus et al., 2016

2.4.3 Unsafe driving behaviour

Generally, driver-centric factors could be found in most crash accidents, and driving behaviour assessment is an important aspect to enhance safety and reduce crashes. Many studies have been conducted to evaluate driving behaviour. Typical approaches include self-reported questionnaires, simulator-based experiments, and naturalistic driving studies (NDS) (Hong et al., 2014). A systematic survey on driving and driver

behaviour has been built into various questionnaires. Simulation experiments provide plentiful insights on specific risk behaviours and scenarios. In NDS, various characteristics and variables about both driving and drivers' behaviours are investigated, based on detailed information recorded using sensors, in-vehicle devices, and even smartphones (Eftekhari and Ghatee, 2018). A range of features have been proposed to describe unsafe behaviours, such as speeding, abrupt braking or jerk, tailgating, yaw, frequent and intense lane change, among others (Bagdadi, 2013; Wahlström et al., 2017). Generally, statistical profiles of movement-related variables are used as behaviour features. Besides, for a reliable driving assessment and risk prediction, in-depth and multi-view features are necessary, especially features with predictability.

2.4.4 Pre-crash risk levels

Accident investigation records and statistics have long been used as the basis to assess safety, but they are generally collected after the accident events. Compared with post-accident data, pre-crash information is much meaningful and informative. Kinetics-related factors are commonly used to provide clues on crash risk (Perez et al., 2017; Wu and Jovanis, 2013). Recently, in NDS (naturalistic driving studies), detailed information about near-crash incidents are collected, with the definitions of near-crash being typically based on driving behaviours, such as rapid evasive manoeuvres (Hankey et al., 2016). However, a quantitative, objective, external measure of graded risk levels remains lacking. Another essential issue is risk labelling. Moreover, detailed risk levels are inherently problematic to determine, since accurate thresholds are difficult to establish. Herein, a clear and convincing grading scheme of risk levels is of great interest.

High-quality data is necessary for risk measurement. Existing studies widely use accident data from police recording and self-reports, controlled experiments and simulation, loop detectors, etc. From such sources, it is extremely difficult to obtain pre-crash data of high quality (e.g. accurate, 1-second resolution or less, at the vehicle level) (Imprialou and Quddus, 2017). Besides, real-world accidents are generally unexpected and occur rarely, and purposive tracking is very costly (Hakkert and

Gitelman, 2014). Indeed, it is near impossible to pinpoint the precise time and location of an accident before-hand. Herein, a practical way to obtain pre-accident information is by retrieving the video footage that contains an accident event. Such video footage can be gathered by a surveillance camera system that continuously records traffic movements for the entire road network.

2.4.5 Methods for prediction

Methods for traffic accident prediction can be broadly classified into three categories, namely statistical models, machine learning approaches, and simulation-based methods.

(1) Statistical models

Statistical models have been used widely for accident predictive analysis. Typical methods include factor-based regression and case-control analysis. Regression models make prediction based on the relationships of influencing factors and potential accidents, such as logistic regression (Abdel-Aty et al., 2004), Poisson (Ma, 2009). Risk factors were considered as independent variables and regression models were developed to predict crash count (occurrence, frequency), accident probability, crash-prone conditions, etc. (Li et al., 2008; Lord and Mannering, 2010). However, such models are hard to capture the complex influencing mechanism and dynamic behaviour, which limits the application and reliability (Längkvist et al., 2014). Due to inconsistent performance and high prediction errors, such models are currently unsuitable for implementation at the real-world operational level (Roshandel et al., 2015).

In addition, case-control design is a predominant and efficient method for the study of rare events, and widely applied to investigate the link between risk factors and accident occurrences (Abdel-Aty and Pande, 2005; Zheng et al., 2010; Roshandel et al., 2015). Accident events are treated as cases, while the equivalent or corresponding scenarios are represented as controls, such as corresponding traffic regime, equivalent location and time. A control-to-case ratio of around 4:1 is recommended (Ahrens and Pigeot, 2014).

(2) Machine learning

Machine learning approaches have a better performance for complex systems, and produce reliable and precise results. Through learning from relationships and patterns from data, the hidden information, features and mechanism can be uncovered. The detailed review about accident prediction using machine learning, which is the approach adopted in this research, is described in Section 2.5.

(3) Simulation-based methods

Simulation approaches provide a framework to model complex systems, and are gaining growing acceptance in various fields. The fundamental is to model the behaviour, interaction and feedback of the elements and structures of the system.

For crash prediction, by mimicking vehicle movement, driving behaviour, and interactions in a traffic stream, the conflict process and risk factors are replicated, hence, the crash prediction is made as a probabilistic consequence (Young et al., 2014). The surrogate measures are widely used for risk assessment and crash identification. Agent-based modelling, fuzzy cellular automata, Markov Chain Monte Carlo (MCMC), and simulation packages (e.g. VISSIM, PARAMICS) are applied to conduct the simulation (Pei et al., 2011; Sobhani et al., 2013; Young et al., 2014; Essa and Sayed, 2015; Chai and Wong, 2015b).

The performance of simulation-based prediction relies on the quality of the data used to calibrate and validate, as well as an accurate representation of the behaviour and characteristics of each component in the systems (Young et al., 2014). However, some assumptions are too simple to represent realistic situations. The variability of realistic behaviour and real-world conditions should be considered, but are generally difficult to incorporate. In addition, the microscopic behaviour may enlarge the uncertainty and hence entail prediction errors.

(4) Hybrid models

Hybrid modelling is a promising solution for higher accuracy and wider reliability. Due to the fact that individual models perform well only for certain situations or data

sets, therefore, the model test should be conducted to find the most suitable one. Hybrid models provide great superiorities compared to the individual methods according to combined use, including weighting-based combination, model and methodology integration (Tascikaraoglu and Uzunoglu, 2014), model ensembles, among others. In addition, the combination is also performed as auxiliary processes, such as data processing (e.g. filtering), parameter selection and optimisation, post-processing (e.g. residual error evaluation) (Roshandel et al., 2015).

2.5 Machine Learning for Prediction

2.5.1 General modelling

Given the complex accident causes and diverse driver behaviour, it is rather impractical to predict traffic accident likelihood using simple historical data and pre-defined rules. Several studies have used machine learning and data mining to predict traffic accident. When the data is pre-processed, and the learning task is defined, a repertoire of machine learning algorithms and models is available to use.

Machine learning is a branch of Artificial Intelligence (Bishop, 2006), and has the advantages in solving inference problems which are impossible to be represented by explicit algorithms (Voyant et al., 2017). Machine learning is a big concept, and typical approaches include artificial neural networks (ANN, such as Deep Learning, self-organised map), support vector machine (SVM), decision tree learning, Bayesian inference, etc. The learning framework consists of two phases: (1) model training based on a given dataset, and obtain an estimate or approximation with optimal values of the specified objective function; and (2) predict the estimate of given new data by the trained model (Kourou et al., 2015).

Crash prediction is considered a multi-classification problem, including supervised classification and unsupervised pattern discovery. In supervised learning, a labelled set of training data is used to train a prediction model. The crash is predicted using the classification of a set of finite classes (risk levels). The performance strongly depends on the quality of designed features (Zeng et al., 2014), and feature selection and hyper-parameter tuning can help to improve modelling performance. However,

developing domain-specific features is not straightforward and requires the expertise of the data (Xue et al., 2016).

In contrast, under unsupervised learning, no labelled data is available and there is no notion of the output during the learning process. Clustering is a common unsupervised task, which is to discover the groups of the input data with similar feature patterns, and new data can be assigned into one of the clustered groups concerning pattern similarity. Besides, semi-supervised learning is a combination of supervised and unsupervised learning, with labelled and unlabelled data.

2.5.2 Typical algorithms

(1) Tree-based ensemble learning

Decision tree algorithms follow a tree-structured classification scheme, where the nodes represent the input variables and the leaves correspond to decision outcomes (Kourou et al., 2015). Tree-based ensemble learning methods use multiple decision tree algorithms to obtain better predictive performance, such as, random forest, GBDT (gradient boosting decision tree), XGBoost.

Based on the architecture of the decision trees, it is clear to interpret and quick to learn. The decisions resulting from the tree-based architecture allow for transparent and interpretable reasoning, which makes tree-based ensemble learning an appealing technique for crash prediction and prevention. A frequent pattern tree was used for real-time traffic accident risk prediction (Lin et al., 2015).

(2) ANN and deep learning

ANNs are widely used to handle a variety of classification-based prediction problems with outperformed performance. ANNs are trained to generate an output as a combination of the input. Multiple hidden layers are typically used for model fitting, which mathematically represent the neural connections (Chang, 2005). Generally, deep learning is constructed when multiple layers (> 5 layers) are used to extract higher level features from raw input progressively. Each layer learns to transform its input data into a slightly more abstract and composite representation.

Deep learnings have been found to be an efficient forecasting technique due to its capabilities such as easy implementation, simplified feature engineering, and nonlinear modelling with high accuracy. However, deep learnings usually require large amounts of data for training, and operate in a non-inferential way (black-box in nature), making the results extremely difficult to interpret or to assist in refining an underlying theory (Xie et al., 2007; Kunt et al., 2011).

(3) Bayesian inference

Classifiers based on Bayesian inference produce probability estimations rather than predictions. Bayesian inference obtains the conditional probability of some parameters given the observed data. Bayesian inference has been applied widely to several classification tasks as well as for knowledge representation and reasoning purposes (Kourou et al., 2015). Bayesian inference was used to predict mountainous freeway risks (Yu et al., 2013), crash injury severity (Yu and Abdel-Aty, 2014), real-time crash prediction conditioned on traffic speed (Sun and Sun, 2015), etc.

(4) Others

SVMs map the input feature into a high-dimensional space, and identify the best hyperplane that separates the classes well (Kecman, 2005; Li et al., 2008). SVMs achieve considerable generalisability and can therefore be used for the reliable classification-based prediction, for example, crash prediction at traffic-zone level (Dong et al., 2015). The identified hyperplane can be thought of as a decision boundary, and can be used for the detection of any misclassifications. However, SVMs also behave as a black-box, and the interpretable parameters are not provided.

2.5.3 Class imbalance

Generally, risk assessment pertains to a distinctly imbalanced problem, in which the risk conditions and safe conditions are highly imbalanced. The class imbalance issue has been discussed in various concepts, such as the safety pyramid model suggested by Hydén (1987). There are some intrinsic challenges in machine learning with imbalanced data. In supervised learning, there are several tactics to reduce the

impacts of imbalanced data. Such techniques are generally categorised into three major groups: data resampling, cost sensitive methods, and ensemble techniques (López et al., 2013).

Independent data preprocessing is performed by under-sampling and/or oversampling, to produce a more balanced class distribution (Díez-Pastor et al., 2015). Well-known methods include SMOTE (synthetic minority oversampling technique) and RUS (random undersampling) (Lin et al., 2017). In cost-sensitive methods, higher misclassification costs are defined for the minority classes so as to bias toward the minority class (López et al., 2013). Although ensemble techniques are not designed specifically for working with imbalanced data, they have been widely combined with data resampling, such as SMOTEBoost, and UnderBagging. Ensemble-based methods have achieved prominent performances for which a detailed review can be found in Díez-Pastor et al. (2015). However, the major drawbacks in the various methods are also obvious, for example, useful data might be potentially eliminated in undersampling; oversampling creates artificial instances, which may induce misleading data; and setting proper misclassification costs is difficult in cost-sensitive methods.

2.6 Domain-specific Feature Engineering

2.6.1 Feature extraction

A feature is a series of information that might be useful for modelling, and such information includes attributes, characteristics, and labels that describe and distinguish the observations, etc. The quality and quantity of the features have a great influence on prediction performance and problem-solving (Guyon and Elisseeff, 2003). Better features can produce simpler and more flexible models and yield better results as well. A feature vector is an n-dimensional vector of numerical features that represent the object (Liu and Motoda, 2012).

The feature learning is a set of data mining processes that extract, create and select understandable features from raw data, and supervised and unsupervised learning algorithm may be used to conduct these processes. Feature learning is motivated by

the desire of building features that are mathematically and computationally convenient for machine learning, and the most informative feature subset is chosen to perform a specific task. Feature learning starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent machine learning and generalisation steps, and in some cases leading to better human interpretations (Piramuthu and Sikora,2009).

Feature extraction is the transformation of raw data into a sequence of feature vectors for better fitting in a specific machine learning process. Features are extracted from raw data manually with prior knowledge, such as some pre-defined rules and functions. Manual feature extraction is a preliminary step for feature learning, and dominant features which can be easily interpreted by a human are obtained from raw data. High-level features are constructed based on already existent features being added to the feature vector, which is referred to as feature construction process (Liu and Motoda, 2012). A repertoire of constructive operators is applied to construct new features, and results in an increase in the size or dimension of the feature vector. Feature construction has long been considered a powerful tool for increasing both accuracy and understanding of the structure, particularly in high-dimensional problems (Bloedorn and Michalsi, 1998; Loh, 2011). Feature extraction and construction are helpful to bridge the gap between raw data and model input (the feature vector). The core of feature extraction and construction entails the creation of a set of features to describe the overall information and make a comprehensive understanding of raw data.

The performance of machine learning depends strongly on feature extraction and selection. The extracted features should contain enough information for accurate modelling. However, irrelevant and redundant features would result in misclassification and negative impacts of the high-dimension, as well as unstable and biased modelling. Therefore, reliable feature selection is introduced as an essential part of feature extraction, which is to evaluate and choose a small subset of features that is sufficient to describe the target concept (Liu and Yu, 2005).

2.6.2 Feature selection

Feature selection is the process of choosing a small subset of relevant features that ideally is necessary and sufficient to describe the target concept (Kira and Rendell, 1992). A goal of feature selection is to avoid selecting too many or too few features than is necessary (Sikora and Piramuthu, 2007). For certain target modelling, some of the extracted features could contain redundant or irrelevant information, which can be removed without incurring information loss.

Learning-based feature selection is performed using supervised learning algorithms to automatically discover and identify enough relevant and non-redundant features for a specific task, such as classification and regression. An eligible subset of the extracted features is selected from a lot of hand-crafted features, which is a dimensionality reduction process (Guyon and Elisseeff, 2006). Meanwhile, they are several benefits in reduction of the dimensions, such as modelling is simplified and easier to interpret, shorter training times, and enhanced generalisation by reducing over-fitting (Piramuthu and Sikora, 2009).

The basic processes of feature learning are: (1) ranking the feature quality for target modelling; (2) identifying the highly contributing ones that help to improve the modelling performance; and (3) discard the redundant and/or irrelevant features. A redundant feature would have high correlations with other features, which is not helpful to increase accuracy or reduce over-fitting. An irrelevant feature carries no discriminative information for target modelling (Bailly and Milgram, 2009; Fahad et al., 2014). After feature ranking and selection, only the highly contributing features are preserved. Various statistical models and supervised learning algorithms are applied for feature ranking and selection, such as principal component analysis (PCA), random forests, classification and regression tree (CART), etc.

For the feature selection procedures, there are typically three types: (1) filter, which employs a ranking criterion to score the features and remove the ones within a threshold, but the features are treated independently, without considering possible correlations, which may result in a suboptimal subset with redundant ones; (2) wrapper, which searches for feature subsets with the best performance and detects the

possible interactions, however, the optimisation search program is computationally intensive, and with risk of overfitting under insufficient number of instances (e.g. recursive feature elimination); and (3) embedded, which performs feature selection and fits learning simultaneously, but the results may depend on classifiers while hyper-parameter tuning is complicated and interrelated. An extensive summary of feature selection can be found in Hapfelmeier and Ulm (2013).

2.6.3 Data quality and mining

Data availability and quality are essential for reliable risk assessment and prediction. Compared with aggregate data, detailed vehicle-level high-resolution (e.g. one second or less) data can offer greater and more precise insights about conflict risk, hence crash likelihood. In addition, traffic crashes are complex events that involve a variety of human behaviours and interactions with surrounding vehicles, traffic-related factors, roadway and environmental conditions. Hence, various data is demanded for accurate risk assessment and crash prediction, including trajectory data and accident footage data. A summary about data collection and potential methods are given in Table 2.5.

Features involving driving behaviour are extracted from the vehicle movement trajectory. Vehicle movement trajectory is a reliable way to extract traffic conflicts and identify conflict risks (Jonasson and Rootzén, 2014). A trajectory is typically defined as an indexed data sequence of positions and velocities over a given time window (Sivaraman and Trivedi, 2013a). Vehicle movement trajectory is an overall reflection of the driver's behaviour during a given time interval and road space. Moreover, a series of relevant variables are able to be derived from vehicle movement trajectory, including vehicle's velocity, acceleration, deceleration, front gap, lateral position, etc. In addition, pre-crash traffic information is also able to be inferred from the trajectory of entire vehicle streams, such as traffic volume, inter-vehicular spatial-temporal relationship, chain of risk events, etc.

Table 2.5 Data sources and collection methods

Source	Data collected	Summary
Surveillance traffic recording system	Pre-accident data, vehicle stream data	Actual accident could be captured; but high requirement of video quality, and data extraction challenge
Onsite video recording	Vehicle stream data, derived driving behaviour data	Easy to conduct, sufficient conflict and risk conditions recorded; but high cost and labour intensive in data extraction
Vision-based in-vehicle driver monitoring	Driver physical and mental behaviour (e.g. distraction, drowsiness)	Based on facial expression (yawn, head pose, gaze direction), eye-related measures (closure, blinking); but with privacy issue
Vehicle On-Board Equipment (OBE)	Vehicle movement trajectory, manoeuvres, status	Using SWM (steering wheel movement), engine control unit (ECU), speedometer; reliable and efficient
High-resolution position system	Real time vehicle position	Easy and cost efficient; but limitations in precision, with privacy issue
Sensors (radar and lidar, etc.)	Vehicle movement and manoeuvre, external and internal information	Various techniques and functions, including radar and lidar (track external moving target), gyroscope (obtain accelerating, turning angular speed and vibration), etc.
Real-time information system	Real-time traffic condition and incidents information	Timely; but basic information about accident time, duration and location, such as Land Transport Authority (LTA) open APIs
Accident records and statistics	Reported accident information	Processed information, post-accident investigation, self-reporting, etc.

Pre-crash data contains the information during a certain time interval immediately before the accident, which is often measured and linked to accident likelihood (Roshandel et al., 2015). The information prior to an accident could provide insights about accident identification, surrogate measurement and causal analysis. The exact

time and location of accident occurrence are critical for accurate measurement, including the capture of the near-crash state (the situation in the immediate vicinity of an accident event). However, data extraction from video recording is also challenging. Existing methods of automatic vehicle detection and tracking are not sufficiently accurate nor efficient (e.g. Lai et al., 2010; Sivaraman and Trivedi, 2013b). Firstly, automatic vehicle detection based on image processing is useful for vehicle counting, traffic rule violation, automatic licence plate recognition (ALPR), etc. (Sivaraman and Trivedi, 2013; Wan et al, 2016), but is problematic for accurate trajectory measurement. Secondly, complex traffic interaction makes it hard to identify every vehicle correctly, especially with overlapping of vehicles' perspectives in dense traffic. Besides, image distortion compounds the difficulties in trajectory measurement, and a complex process is needed to reduce the error in the distortion. Therefore, there are many constraints in tracking vehicle trajectory. Chai and Wong (2013) developed and applied a technique of measuring vehicle trajectory by a projective transformation of video frames at first, and then indicate vehicle position by means of computer-aided annotation; this hybrid approach is flexible and easy to use, albeit involving certain tedium.

2.7 Chapter Summary

The literature review is undertaken to capture important research and development in the areas of traffic crash prediction and prevention, especially risk assessment and machine learning approaches. The literature findings are useful towards modelling design and approach development for crash risk assessment and prediction, ranging from data collection to algorithms testing.

CHAPTER 3
KEY RISK INDICATORS FOR ACCIDENT ASSESSMENT CONDITIONED ON PRE-CRASH VEHICLE TRAJECTORY

3.1 Chapter Introduction

Pre-accident risk assessment by surrogate indicators is an effective way to identify risk levels and thus boost crash prediction. This chapter assesses the feasibility of using surrogate indicators for risk assessment, based on real-world pre-accident data, and develops hybrid indicators, namely the Key Risk Indicators (KRIs), to evaluate pre-accident risk exposures (e.g. severity, likelihood) with graded levels. Section 3.2 develops the concept of KRI, and elaborates the selection of basic indicators. A typical real-world chain-collision accident and its antecedent (pre-crash) road traffic movements are retrieved from surveillance video footage, and in Section 3.3, a grid remapping method is proposed to extract and measure the vehicle trajectory data. Section 3.4 designs a temporal-spatial case-control to evaluate the feasibility of indicator-based risk assessment, and constructs the KRIs. In Section 3.5, another independent real-world accident event is examined for the validation of KRI. The final two sections cover the discussion and summary.

This chapter includes part of contents that have been published in the paper:

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2018). Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. Accident Analysis and Prevention, 117, 346-356. DOI: 10.1016/j.aap.2018.05.007.

3.2 Key Risk Indicator for Traffic Safety

3.2.1 Concept of KRI

The concept of KRI has important applications in several areas, such as operational risk management (Scandizzo, 2005) and enterprise risk management (ERM) (Hwang et al., 2010) and financing, among others.

As applied to road safety, KRIs are metrics capable of revealing risk exposure in traffic flow, and providing predictive signals of a potential accident. KRIs enable the analyst to identify risks that may lead to an accident, and grasp insights of an impending accident (such as at-risk vehicles, potential locations and time), thus the prevention strategy can be applied in a targeted and pre-emptive way. Hence, for accident assessment and prediction purposes, it is crucial that KRIs are designed effectively and reliably.

KRIs are developed based on a set of basic indicators that are effective and reliable in measuring risks. The design of KRIs starts with first shortlisting a set of existing metrics and then identifying the most critical ones that can serve as the basic indicators. The guiding principles for selecting metrics are outlined, such as meaningfulness, measurability, predictability (e.g. leading indicators), etc. Shortlisted metrics should offer useful insights about accident risks and be easy and clear to interpret. Complex metrics would make it difficult to track and manage. In addition, leading indicators should be included to offer predictive signals of a potential accident.

3.2.2 Risk behaviour indicators

Driving behaviour plays a major role in accident risk. High-risk driving behaviours may result in a high likelihood of an accident, such as excessive speeding, driving too close to the preceding vehicle, etc. As a result, temporal and spatial proximity can be used to evaluate such risk behaviours. In addition, indicators based on temporal proximity are relatively popular and objective, because they integrate both the spatial proximity and speed difference (Zheng et al., 2014). Moreover, a definition based on time and space proximity is easily understood and evaluated (Chin and Quek, 1997). Among time-based indicators, Time to Collision (TTC) is well-recognised and widely-used in practice, for theoretical and reliability reasons (Mahmud et al., 2017). Based on TTC, Time Exposed TTC (TET) and Time Integrated TTC (TIT) were further proposed (Minderhoud and Bovy, 2001), to measure risk duration and risk integration, respectively. These three time-based indicators are shortlisted.

(1) The basic TTC is defined by the time to a potential collision between two vehicles

(van der Horst, 1990), as follows:

$$TTC_i(t) = \begin{cases} \dfrac{x_{i-1}(t) - x_i(t) - L_{i-1}}{v_i(t) - v_{i-1}(t)} & v_i(t) > v_{i-1}(t) \\ \infty & v_i(t) \leq v_{i-1}(t) \end{cases}$$

where $x_i(t)$ and $v_i(t)$ are the position and velocity of targeted (following) vehicle (i) at timestamp t, and L_{i-1} is the length of preceding (leading) vehicle $(i-1)$.

Generally, risk behaviours and conditions are flagged for any vehicle pair with a TTC value less than a given threshold. TTC notion is illustrated with vehicle trajectories in Figure 3.1(a). Based on the velocity difference and gap at timestamp t, the two vehicles would be involved in a collision at the timestamp $t + TTC(t)$, if the collision course and velocity difference are maintained, as indicated by the red (short) dashed line (Time-based) in Figure 3.1(a).

TTC lower than the perception and reaction time should be considered as unsafe. The values of TTC threshold under various scenarios have been studied by many authors (e.g. Vogel, 2003). Some reported minimum or desirable critical TTC thresholds vary between 1.5 s and 4.0 s, as presented in Table 3.1. However, risk severity associated with TTC is not obvious, and it is not straightforward to define the threshold values that can be used to distinguish events between relatively safe and risky under various conditions.

(2) TET expresses the total time of a vehicle exposed to risk situations, as follows:

$$TET_i = \sum_{t=0}^{N} \delta_i(t) \cdot \tau_{sc}$$

$$\delta_i(t) = \begin{cases} 1 & \forall\ 0 \leq TTC_i(t) \leq\ TTC^* \\ 0 & otherwise \end{cases}$$

where, for a period $T = N \cdot \tau_{sc}$, there are N small time intervals, each interval is τ_{sc} (e.g. 0.1 seconds). $\delta_i(t)$ is a switching variable between 1 and 0, and value 1 indicates a signal of risk condition, when the TTC value is within threshold TTC*.

(3) TIT takes into account the accumulated impact of risk behaviour, using the

integration of TTC profile within a specified threshold, as follows:

$$TIT_i = \int_0^T [TTC^* - TTC_i(t)]dt \quad or, TIT_i = \sum_{t=0}^{N} [TTC^* - TTC_i(t)] \cdot \tau_{sc}$$

$$\forall\, 0 \le TTC_i(t) \le TTC^*$$

TIT suggests a risk accumulation in both risk severity and exposure duration. Hence, vehicle TIT is useful to express the risk level of driving behaviour and relative likelihood of conflicts involved. The graphical representation of TET and TIT is shown in Figure 3.1(b) (adapted from Minderhoud and Bovy, 2001).

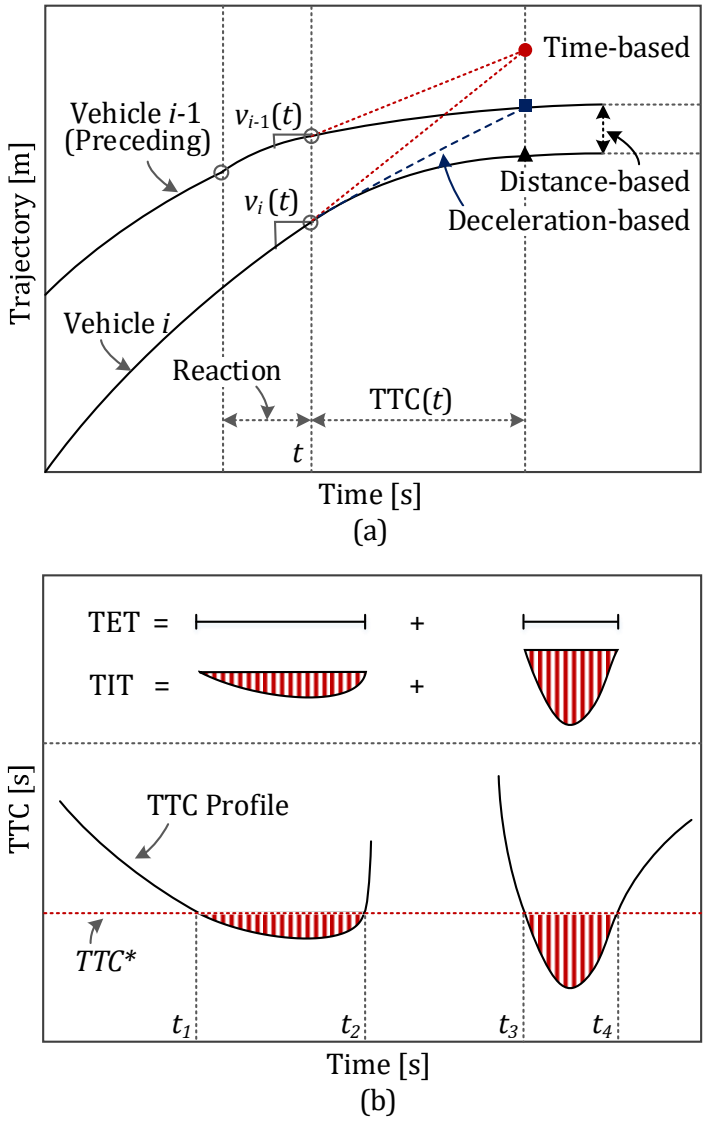


Figure 3.1 Graphical representation of the indicators

Based on the TTC curve, which is the time series data of TTC value, TET is the total length of time when TTC curve is under the defined threshold line (TTC^*), and TIT is calculated by the area between TTC curve and the threshold line.

Table 3.1 Thresholds and parameters involving the KRIs

Indicator	Condition	Threshold/ Parameters	References
TTC	Intersection	Minimum 1.0s; Desirable 1.5s	van der Horst, 1990
	Un-signalised intersection	2.0s for low level; 0.9s for high level	Sayed, Brown and Navin, 1994
	Intersection	Minimum 1.0s; Desirable 2.0s	Vogel, 2003
	2-lane rural road	3.0s	AASHTO, 2004
	Tunnel expressway	4.0s; Minimum 2.0s	Meng and Qu, 2012
	Urban roundabout	1.5s	Guido et al., 2010
PICUD	Deceleration rate	$3.3\mathrm{m/s^2}$	Uno et al., 2003
	Reaction time	1.0s	Uno et al., 2003
DRAC	Highway	$3.40\mathrm{m/s^2}$	AASHTO, 2004
	Urban intersection	$3.35\mathrm{m/s^2}$	Archer, 2005
	Urban roundabout	$3.35\mathrm{m/s^2}$	Guido et al., 2010
MADR	MADR 1: Different pavement surface condition (i.e. wet, dry, snow, etc.)	$g \cdot (f \pm i);\ g = 9.81 m/s^2;\ f = 0.4$	Ministero delle Infrastrutture e dei Trasporti (2001); Guido et al., 2010
	MADR 2: Truncated normal distribution	Average 8.45; S.D. 1.40; Lower limit 4.23; Upper limit 12.68 (unit in $\mathrm{m/s^2}$)	Cunto and Saccomanno, 2008

3.2.3 Risk avoidance indicators

The likelihood of an accident also depends on risk avoidance. A hazard condition that is hard to eliminate also indicates a high level of risk severity. The distinction between a crash and a near-miss depends on the evasive manoeuvres, such as a timely reaction and required braking. A vehicle collision could be avoided if the following vehicle timely decelerated to stop or matched the speed of the leading vehicle. Herein, two deceleration-based indicators are introduced to evaluate kinematic characteristics involving risk avoidance. Deceleration rate to avoid crash (DRAC) is recognised as a safety performance indicator, as it explicitly considers the role of differential speeds and decelerations in risk avoidance (Archer, 2005). Based on DRAC, an improved indicator was proposed, namely crash potential index (CPI), which considers many major factors in risk avoidance, including available braking capacity and time exposed to risk (Cunto and Saccomanno, 2008).

(1) DRAC evaluates the braking requirement during a vehicle conflict, which is represented by the minimum deceleration rate required to avoid the collision, as follows:

$$DRAC_i(t) = \begin{cases} \dfrac{[v_i(t) - v_{i-1}(t)]^2}{x_{i-1}(t) - x_i(t) - L_{i-1}} & v_i(t) > v_{i-1}(t) \\ 0 & v_i(t) \le v_{i-1}(t) \end{cases}$$

A higher DRAC value indicates a more dangerous (car-following) scenario. The American Association of State Highway and Transportation Officials suggested that a given vehicle is in conflict if its DRAC exceeds a threshold of 3.4 $\mathrm{m/s^2}$ (AASHTO, 2004).

(2) CPI measures the probability that a vehicle's DRAC exceeds its maximum available deceleration rate (MADR) or braking capacity during a given period, as follows:

$$CPI_i = \frac{\sum_{t=0}^{N} P(DRAC_i(t) > MADR^{[\alpha_1, \alpha_2, \cdots, \alpha_n]}) \cdot \tau_{sc}}{T}$$

MADR is specific for a given set of traffic and environmental attributes

$[\alpha_1, \alpha_2, \cdots, \alpha_n]$, and depends on factors such as pavement conditions (e.g. dry, wet, snow), vehicle weight (e.g. car, truck) and braking system (Cunto and Saccomanno, 2008). Two MADR methods are used to calculate the CPI probability, namely, MADR1 and MADR2. MADR1 is measured by the proportion of DRAC exceeding the defined MADR value, and MADR2 is assumed to follow a truncated normal distribution, with parameters given in Table 3.1.

3.2.4 Risk margin indicators

Some accidents are triggered under exceptional circumstances, such as an abrupt hazard in front, driver cognitive failures and driving errors (Chai et al., 2017). For spatial proximity in emergency risks, the time-based indicators are not sensitive, and maximum braking capacity is generally applied. Accident risk is high if the available space is smaller than the space needed for an evasive reaction. Herein, distance-based indicators are suitable to measure the accident risks associated with emergencies by spatial characteristics. Proportion of stopping sight distance (PSD) measures the risk margin to the potential point of an accident. Similarly, the potential index for collision with urgent deceleration (PICUD) evaluates the risk margin under emergency braking. Besides, PICUD is helpful to evaluate two vehicles with similar speeds and the effect of gap keeping.

(1) PSD is the ratio of remaining distance (RD) to the potential point of accident and acceptable minimum stopping distance (MSD) (Allen et al., 1978), as follows:

$$PSD = \frac{RD}{MSD}, \quad MSD = \frac{v_i(t)^2}{2d}$$

where d is acceptable maximum deceleration rate. A calibrated value of d is 3.92 $\mathrm{m/s^2}$ in Guido et al. (2010). PSD with a value less than 1.0 indicates that a collision scenario is hard to be avoided even if the maximum deceleration ability is applied.

(2) PICUD is defined as the distance between two consecutive vehicles when they stop completely under emergency braking (Uno et al., 2003), as follows:

$$PICUD = \frac{v_{i-1}(t)^2 - v_i(t)^2}{2\alpha} + s_0 - v_i(t) \cdot \Delta t$$

where α is deceleration rate, s_0 is the initial distance, Δt is driver's reaction time. Two predetermined parameters are required for estimation, namely the deceleration rate and reaction time, as listed in Table 3.1.

The above seven basic indicators are shortlisted. More information about them is described in Appendix A. The suitability and effectiveness of each indicator will be assessed based on real-world accident data, and the findings shall serve as a proof-of-concept in applying KRIs for accident assessment and prediction.

3.3 Pre-Accident Data Acquisition

3.3.1 Accident data retrieval

To evaluate the indicator performance in pre-accident risk assessment, a typical real-world traffic accident involving multiple-vehicle collision is selected for data extraction. The particular accident occurred on the fastest (right-most) lane of a straight, level segment of an expressway carriageway during morning peak hours. The accident was captured by the expressway surveillance system of Land Transport Authority (LTA) Singapore. The surveillance camera is on fixed settings. Video footage of the main accident process and 1-minute antecedent (pre-crash) road traffic movements is used. Images are extracted from the video footage at a sampling rate of one frame per 0.5 seconds. Each image has a resolution of 1,552 pixels in width and 1,200 pixels in height, as shown in Figure 3.2.

A practical way to gather accident data is by homing in on video footages that contain accident events as gathered by a surveillance camera system that continuously records traffic movements for the entire road length. In Singapore, Expressway Monitoring and Advisory System (EMAS) is an expressway incident management system that monitors and detects traffic incidents (e.g. accidents, vehicle breakdowns) along Singapore's expressways, thereby ensuring fast response to restore normal traffic flow. Hence, pre-accident data is able to be captured by such a surveillance system.

However, data extraction from video recording is also challenging. Existing methods in computer vision are useful for vehicle detection and tracking (e.g. vehicle/non-

vehicle classification, vehicle counting) (e.g. Sivaraman and Trivedi, 2013b), but they are problematic for accurate data extraction and measurement (e.g. vehicle trajectory, gap, speed). Additionally, there are many constraints in camera-based data acquisition, such as lens distortion from camera angles, object overlapping in dense conditions. The solution for exact measurement remains lacking. Chai and Wong (2013) developed and applied a technique of measuring a vehicle trajectory by a projective transformation of video frames at first, and then indicating vehicle positions by means of computer-aided annotation; this hybrid approach is flexible and easy to use, albeit involving certain tedium.

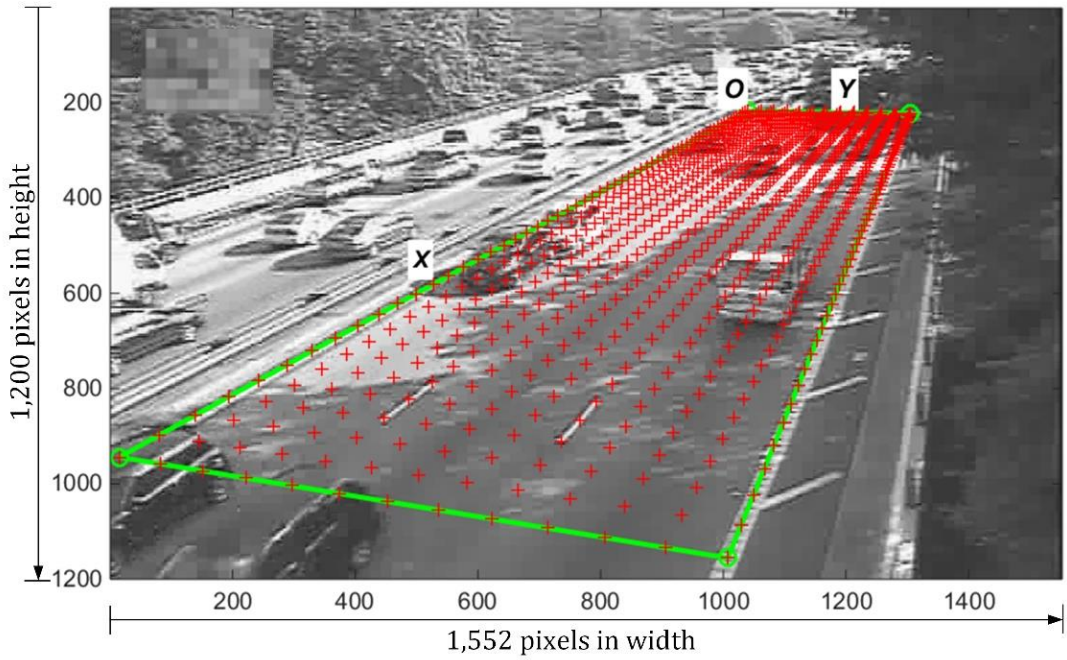


Figure 3.2 Surveillance camera coverage and grid remapping area

3.3.2 Image-based data extraction

The extraction of vehicle trajectory data from video images is conducted by computer-aided vehicle detection. This method relies on manually detecting the vehicle's trajectory point and automatically recording the pixel coordinates of the detected point. The pixel coordinates (x_i^p, y_i^p) locate a pixel in an image array, where x_i^p is the column index value and y_i^p is the row index value. The pixel indices are ordered from top to bottom and left to right, as illustrated by Figure 3.2.

A grid remapping area (indicated by the X-Y axes marked in Figure 3.2) is defined to measure trajectory by coordinates transformation, which is described in Section 3.3.3.

The centre of the vehicle's front number plate is used as the detected point. Vehicle length is measured by the distance between the headlight and corresponding rear-light. The data extraction from video images and vehicle trajectory labelling by pixel coordinates are illustrated in Figure 3.3.

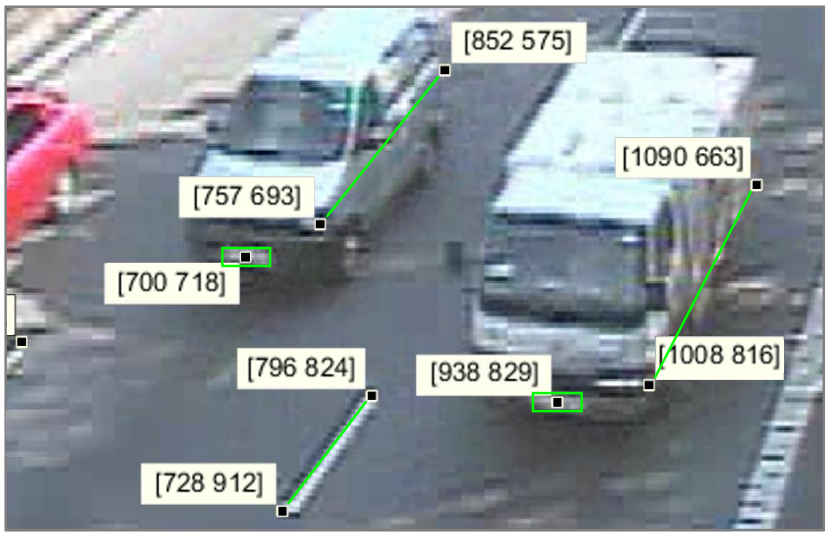


Figure 3.3 Data extraction from video images

3.3.3 Coordinates transformation

The extracted data is described by pixel coordinates in the image space, which needs to be transformed into coordinates described by real-world units (e.g. metres). Moreover, due to camera view, there is obvious lens distortion and irregular scaling. To accurately measure trajectory in real-world units, a grid remapping method is developed with three steps, namely grid mesh establishment, vehicle remapping and in-cell measurement.

(1) Grid mesh establishment based on camera calibration

The grid mesh is established by dividing the road segment on the image into many counting cells. Every cell is of equal size in real-world units, which represents an

equal-sized planar area on the actual road. The grid mesh with equal-sized cells is generated by implementing the camera calibration algorithm proposed by Zhang (2000, 2004), which is widely applied and flexible to use.

Firstly, a remapping area is identified from the real-world road segment. The remapping area is on a regular and conformal setting, as illustrated in Figure 3.4(a).

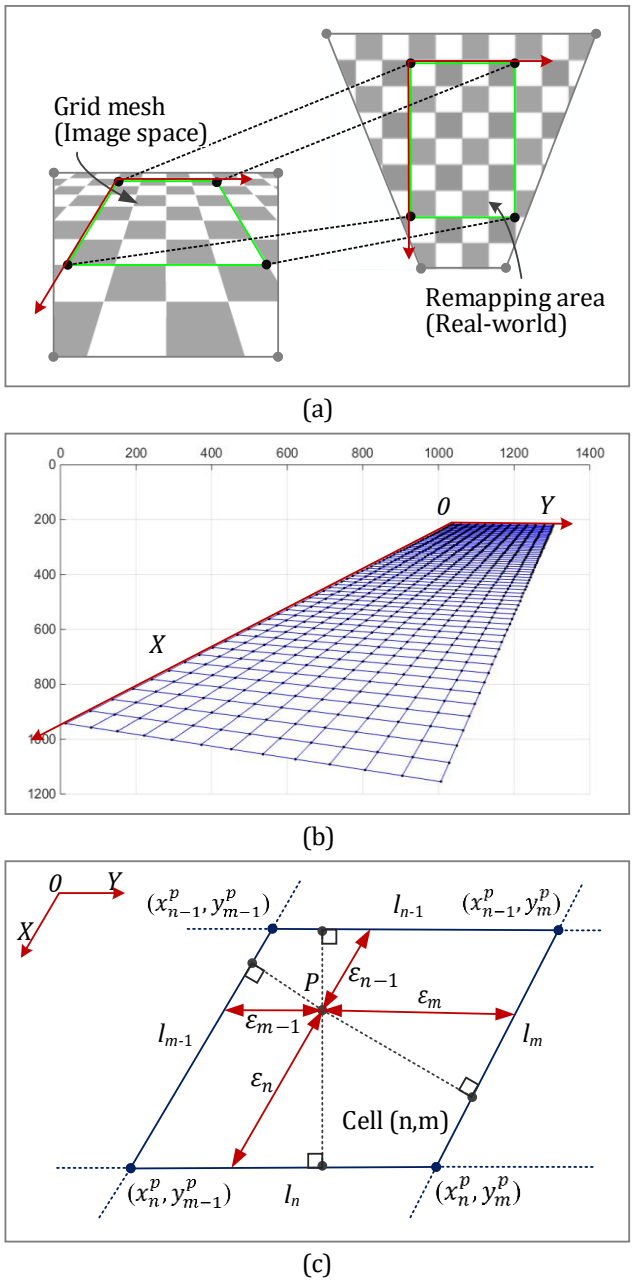


Figure 3.4 Coordinates transformation for trajectory measurement

Such setting is helpful to restore the parallel geometric relationship and consequently reduces the impact of lens distortion. In this case, a rectangular remapping area is selected based on parallel road markings, as illustrated in Figure 3.2.

Secondly, the vertices of the remapping area are marked in the image space. The four vertices are set up to determine the projection relationship between real-world and image space, as illustrated in Figure 3.2 and Figure 3.4(a).

Lastly, the grid mesh is generated based on the four vertices using Zhang's camera calibration algorithm. The grid mesh consists of equal-sized counting cells, as shown in Figure 3.4(a) and 3.4(b). Every cell has a corresponding distortion as the object in image space does. The configuration of the grid mesh is related to the layout of the observed road segment, as well as video camera position and angle of view.

(2) Vehicle remapping

This step is to estimate an approximate value of trajectory coordinates by mapping vehicles onto corresponding cells in the grid mesh system. Each grid cell has a unique address $(x_i^c(t), y_i^c(t))$, which is measured by the counting indices from the base point. The base point is set as the top-left vertex of the grid mesh system, thus the x-axis is along the right-most edge of the road section in the direction of vehicle travel, and the y-axis is along the baseline, as illustrated in Figure 3.2 and Figure 3.4(b). Hence, the cell address is a binary variable estimation of vehicle coordinates.

The vehicle remapping is a matching between the grid cell and the labelled vehicle, by judging whether the trajectory point $P: (x_i^p(t), y_i^p(t))$ is located within the cell $C: (n, m)$ or not, as shown in following:

$$F_{inpolygon}((x_i^p(t), y_i^p(t))|(n, m)) = \begin{cases} 1 & if\ pixel\ point\ in\ cell, \\ 0 & otherwise \end{cases}$$

$$\forall\ F_{inpolygon} = 1, \qquad \left(x_i^c(t), y_i^c(t)\right) = (n, m)$$

$$\Longrightarrow\ (x_i^p(t), y_i^p(t)) \leftrightarrow \left(x_i^c(t), y_i^c(t)\right)$$

The cell $C: (n, m)$ is a quadrilateral region whose four vertices are specified by pixel

coordinates (x_{n-1}^p, y_{m-1}^p), (x_{n-1}^p, y_m^p), (x_n^p, y_{m-1}^p), and (x_n^p, y_m^p), as illustrated in Figure 3.4(c). If the trajectory point is inside or on the edge of the quadrilateral cell region, $F_{inpolygon}$=1. The inpolygon function is based on the algorithm for point in polygon proposed by Hormann and Agathos (2001). Herein, an approximate value of vehicle trajectory is described by the cell address $(x_i^c(t), y_i^c(t))$ under the grid mesh system.

(3) In-cell measurement

This step is to measure more accurate trajectory coordinates by calculating detailed in-cell position. The cell address and in-cell position further form into detailed grid coordinates. The calculation of trajectory coordinates is shown in the following:

$$\underbrace{(x_i^g(t), y_i^g(t))}_{Grid\ coordinates} = \underbrace{(x_i^c(t), y_i^c(t))}_{Cell\ address} - \underbrace{(x_i^\varepsilon(t), y_i^\varepsilon(t))}_{In-cell\ position}$$

$$(x_i(t),\ y_i(t)) = (x_i^g(t) \cdot s_x,\ y_i^g(t) \cdot s_y)$$

$$(x_i^\varepsilon(t), y_i^\varepsilon(t)) = \left(\frac{\varepsilon_m}{\varepsilon_m + \varepsilon_{m-1}}, \frac{\varepsilon_n}{\varepsilon_n + \varepsilon_{n-1}}\right)$$

$$\frac{\varepsilon_m}{\varepsilon_m + \varepsilon_{m-1}} \cong \frac{D(P, l_m)}{D(P, l_m) + D(P, l_{m-1})}$$

where variables ε_m and ε_n are the in-cell proportions as illustrated in Figure 3.4(c). Since each cell is approximately a parallelogram, the in-cell proportions can be measured based on the point-line distance $D(P, l)$. Note that the trajectory coordinates $(x_i(t),\ y_i(t))$ in real-world units are obtained based on grid coordinates and conversion coefficients, s_x and s_y.

The conversion of the cell size into real-world units is determined by the relationship between grid mesh structure and corresponding real-world coverage. In this case, a 60×12 grid mesh structure is used to cover the remapping area with 720 counting cells. The remapping area is about 60 metres in length and with a width of 3 lanes. Each lane in this road section is 3.7 metres in width. Herein, the corresponding real-world size of the 60×12 grid mesh is 60 metres by 11.1 metres. Thus, the conversion coefficients are 1.0 metre/cell in length (s_x) and 0.925 metre/cell in width (s_y).

3.3.4 Calibration and verification

The proposed grid remapping method is feasible and practical to solve the problems of image-based data extraction and coordinates transformation. To ensure high-accuracy, the configuration of the rectangle remapping area should be convenient for real-world scale measurement, and with vertices to be clearly identified on image. Herein, an on-site survey of the remapping area would be helpful. To mitigate errors derived from lens distortion, smaller cell size is recommended, for instance, no greater than 1.0 metre/cell. Besides, the vehicles in far view are blurred for observation, therefore, only data from the fore-ground 50 metres road segment is used in subsequent analysis.

Verification is conducted by comparing the results with real-world targets. Targets with known size are used to check the accuracy, such as road surface markings and lines, known car length. In this case, the lane marking between lanes on the expressway is used for checking. According to LTA standard on the details of road elements, the length of the dashed white lane marking is 2.0 metres. For the lane marking in Figure 3.3, pixel coordinates (728, 912) and (796, 824) are converted to be (57.06, 7.95) and (54.98, 8.07), respectively. Thus, the result is 2.08 metres long, with an accuracy of 96%. To achieve a reliable measurement, an iterative process in terms of grid mesh estimation and remapping configuration is performed based on the discrepancies between the results and actual values, until the accuracy becomes acceptable. Compared with existing solutions, grid remapping method has a good performance in both accuracy and efficiency.

3.4 KRIs for a Real Accident Assessment

3.4.1 Pre-accident reconstruction

The pre-accident scenario is reconstructed to investigate factors that might cause/affect the accident. The four-vehicle head-to-rear collision occurred at the 43rd second on the fast lane. For the fast lane, there were 29 vehicles that passed through the 50-metre road segment over 45 seconds (43-second pre-accident and 2-second post-accident), therein, five vehicles were highly relevant to the accident. Initially,

several vehicles were moving fast with short front gaps, which reflected an unsafe condition existing in the dense traffic flow. At about the 39th second, a vehicle (Vehicle ID: 24) made an urgent deceleration, triggering a high-risk condition. The following vehicles (Vehicle IDs from 25 to 29) sequentially responded with deceleration, however, vehicle 27 failed in risk avoidance and hit the rear of vehicle 26, causing the accident. Furthermore, vehicles 28 and 29 also failed to stop before the accident point and were thus involved in this chain collision.

The pre-accident reconstruction reflects several major contributing factors to the accident, including unsafe driving behaviour (e.g. high speed with a short gap in dense flow), high-risk conditions (e.g. urgent deceleration, abrupt collision in front), failure of collision avoidance (e.g. braking capacity, stop before the accident point), etc.

3.4.2 Temporal-spatial case-control

A temporal-spatial case-control study is designed to evaluate the indicators for pre-accident risk assessment. The case-control design contains one accident case (group [a]), three temporal controls (groups [b], [c] and [d]), and two spatial controls (groups [e] and [f]), as shown in Table 3.2.

Case group [a] represents the near-accident event, which is defined by the time segment of 10-second before the accident, namely from the 33rd to 43rd seconds, and based on the fast lane. Five control groups are on the same stretch of road but at different periods and for different lanes. Each group has equal values in terms of time duration (10 seconds) and lane length (50 metres). Therein, three temporal controls show the risk conditions of 20-second pre-accident (group [b]), 30-second pre-accident (group [c]) and 40-second pre-accident (group [d]), for the fast lane. Two spatial controls cover the risk conditions on adjacent lanes, namely the middle lane (group [e]) and the slow lane (group [f]), during the near-accident period.

Table 3.2 Temporal-spatial case-control comparison

	Case	Temporal Controls			Spatial Controls	
	[a]	[b]	[c]	[d]	[e]	[f]
Accident	Yes	No	No	No	No	No
Lane	Fast	Fast	Fast	Fast	Middle	Slow
Time segment (s)	33-42.5	23-32.5	13-22.5	3-12.5	33-42.5	33-42.5
Traffic flow information						
Vehicle count#	8	9	9	8	9	6
Mean speed (km/h)	49.2	51.9	51.7	37.0	68.7	54.1
S.D. of speed	9.0	10.2	12.3	4.6	14.5	7.5
Mean front gap (m)	14.5	16.0	14.8	12.3	15.7	20.4
S.D. of front gap	8.8	6.3	8.0	5.2	11.4	10.7
TTC						
At-risk vehicles##	4	1	0	1	0	0
Exposure time (s)	4.0	0.5	0	0.5	0	0
Average TTC (s)	1.63	2.05	-	3.45	-	-
Min TTC (s)	0.50	2.05	6.60	3.45	4.19	5.79
TET						
TTC*=1.5 s	1.5	0	0	0	0	0
TTC*=2.0 s	3.0	0	0	0	0	0
TTC*=2.5 s	3.5	0.5	0	0	0	0
TTC*=3.0 s	3.5	0.5	0	0	0	0
TTC*=3.5 s	4.0	0.5	0	0.5	0	0
TTC*=4.0 s	4.0	0.5	0	0.5	0	0

Number of vehicles contained within the observation window of each group.

Number of at-risk vehicles measured based on TTC threshold value of 4.0, as a demonstration.

* Different TTC threshold values, for testing threshold sensitivity.

Table 3.2 (continued) Temporal-spatial case-control comparison

	Case	Temporal Controls			Spatial Controls	
	[a]	[b]	[c]	[d]	[e]	[f]
TIT						
TTC*=1.5 s	1.08	0	0	0	0	0
TTC*=2.0 s	2.20	0	0	0	0	0
TTC*=2.5 s	3.72	0.22	0	0	0	0
TTC*=3.0 s	5.47	0.47	0	0	0	0
TTC*=3.5 s	7.47	0.72	0	0.02	0	0
TTC*=4.0 s	9.47	0.97	0	0.27	0	0
DRAC						
No. of at-risk vehicles	3	0	0	0	1	0
Exposure time (s)	1.5	0	0	0	0.5	0
Average DRAC (m/s2)	6.24	-	-	-	3.83	-
Max DRAC (m/s2)	7.51	-	-	-	3.83	-
CPI						
MADR1	0.06	0	0	0	0.02	0
MADR2	0.33	0	0	0	0	0
PSD						
No. of at-risk vehicles	6	4	2	1	1	1
Exposure time (s)	5.5	2.5	1.5	1.5	0.5	0.5
Average PSD	0.68	0.79	0.90	0.88	0.80	0.85
PICUD						
No. of at-risk vehicles	8	6	4	3	5	2
Exposure time (s)	7.5	4.0	4.0	3.0	3.5	2.5
Average PICUD (m)	-18.2	-14.4	-9.6	-4.4	-6.8	-4.5

Seven basic indicators are calculated on the basis of each case-control group, as shown in Table 3.2. For instance, in the Case [a], the 10-second traffic volume has the vehicle count of eight on the fast lane (defined by the 50-metre segment), accordingly, risk assessment is based on the eight vehicles. Among the eight vehicles, one pair of vehicles crashed (Vehicle IDs: 26 and 27). Following vehicles after vehicle 27 are not taken into consideration, since at that juncture, the crash has already been initiated, and thus the pre-accident state is turned into a post-accident state.

One problem in existing indicators (e.g. TTC, DRAC) is how to determine the threshold values that are used to distinguish different risk levels. The definition of risk levels is threshold sensitive, determined by different thresholds. A serious exceedance of defined threshold indicates a high likelihood of an accident. However, standard threshold values have not been determined yet. Herein, a range of threshold values is used to test the impacts of different threshold values on risk assessment. As listed in Table 3.1, six TTC threshold values from 1.5 to 4.0 (seconds) are included for TET and TIT, benchmarked in previous studies (e.g. van der Horst, 1990; Mahmud et al., 2017). The DRAC threshold is 3.4 $\mathrm{m/s^2}$ (Archer, 2005; Guido et al., 2010). Regarding CPI, MADR1 (considering the proportion) and MADR2 (under the assumption of a truncated normal distribution) are applied (Cunto and Saccomanno, 2008; Guido et al., 2010). A deceleration rate of 3.3 $(\mathrm{m/s^2})$ and 1.0 second reaction time are assumed in PICUD (Uno et al., 2003).

3.4.3 Indicator based risk assessment

Pre-accident risks are assessed by the seven basic indicators, each of which is useful in distinguishing risk levels in certain ways. According to Table 3.2, it is found that: (1) the performance of each indicator is different in measuring risks, therein, TIT and CPI are better in identifying accident risk conditions from the traffic flow; (2) TIT can identify many pre-accident risk conditions and CPI is helpful in further identifying the most severe conditions (the near-accident) from among these risk conditions; (3) for the threshold issue, the impact of using different threshold values is not very critical in determining risk levels; and (4) PSD is helpful to describe the spatial proximity in emergency risks.

The lane-level risk assessment based on TIT and CPI is illustrated in Figures 3.5(a) and 3.5(b), respectively.

From the TIT-time curve in Figure 3.5(a), the lane-level pre-accident risk conditions are monitored, and TIT contributes to the identification of risk exposures from traffic flow, for example, around the 33rd and after the 37th seconds. A higher TTC threshold value (e.g. 4.0 seconds) is more sensitive in identifying risk conditions, while a critical threshold value (e.g. 1.5 seconds) is helpful to figure out the risk conditions with high severity. But the impact of different TTC threshold values is not very critical in determining the high-risk conditions.

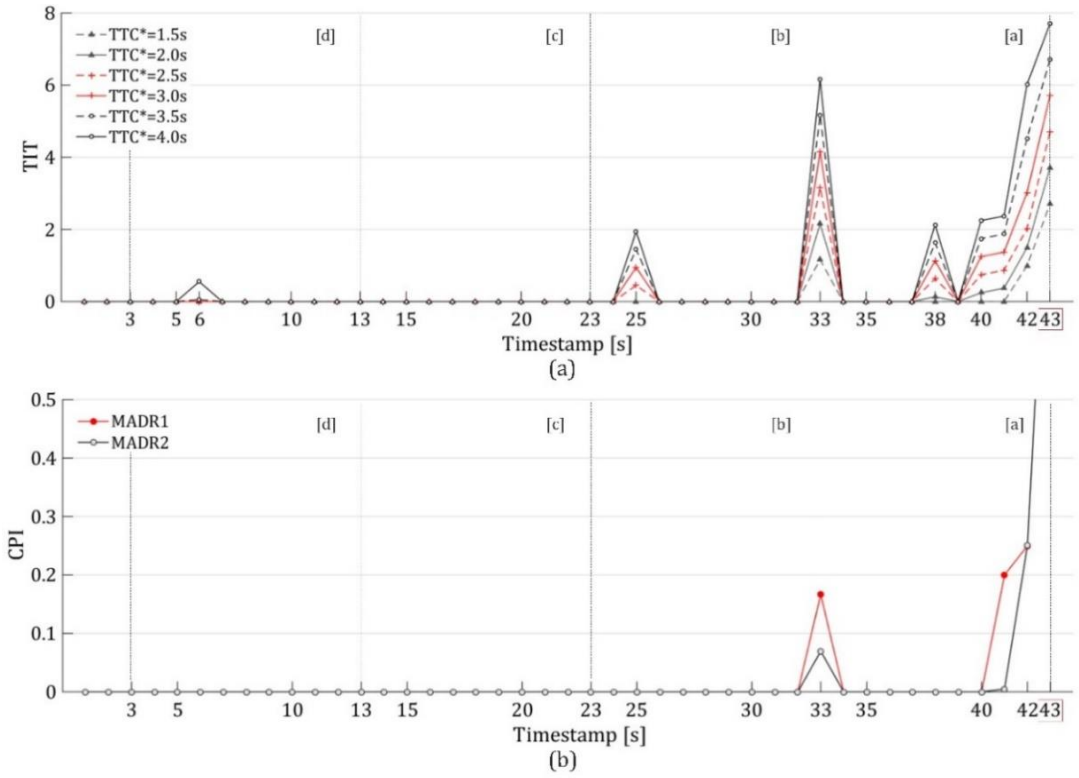


Figure 3.5 Pre-accident risk assessment for the fast lane

Figure 3.5(b) shows the serious risk exposures which are relatively difficult to avoid. Two serious risk conditions are identified by CPI, one of which closely matches with the actual accident. Two CPI measures (i.e. MADR1 and MADR2) are found to be similar in figuring out the risk conditions. The near-accident or near-miss events could be estimated by CPI, including potential locations and timestamps in exposure

to an accident.

Compared with the risk conditions suggested by TIT and CPI, there is an extensive identification of potential risks on the basis of distance-based indicators. PSD performs better in moderately identifying potential risks. Hence, PSD is introduced as a complementary measurement for accident risks associated with potential emergencies, such as a secondary accident in chain collisions.

3.4.4 Vehicle-level risk assessment

Risk exposures in terms of individual vehicles (vehicle pairs) are identified from vehicle TIT and CPI values, as shown in Figures 3.6(a) and 3.6(b), respectively.

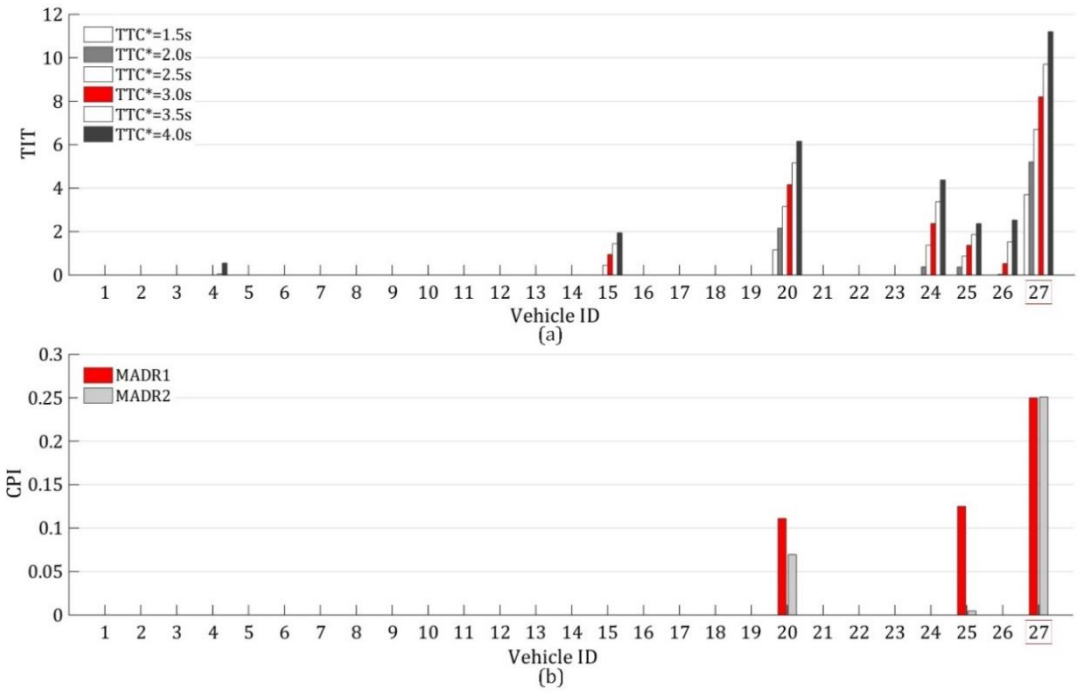


Figure 3.6 Vehicle-level risk identification by indicators

From vehicles' TIT, the at-risk vehicles are able to be identified, especially the vehicles 24-27, which were involved in the accident. The vehicle crash avoidance performance is measured by CPI. The CPI can identify the vehicles with high-risk behaviours and which specific vehicles are likely to cause an accident, as shown in Figure 3.6(b). Furthermore, TIT and CPI also characterise vehicle risks, which is potentially very useful, such as for driving behaviour analysis, insurance-based risk

pricing, and targeted safety enhancement, etc.

3.4.5 KRIs as hybrid indicators

Following these findings, the concept of KRI is further enhanced, which is using hybrid and hierarchical indicators to distinguish risk levels with simplified threshold measurements. It is clear that indicator-based risk assessment is quantifiable, especially using TIT and CPI. TIT shows a general collision risk by vehicle conflicts, and CPI exhibits the identification of higher risk severity, which tends to yield a higher likelihood of an accident. CPI and TIT are therefore used as the basic indicators to structure the hybrid and hierarchical KRIs, as well as PSD.

There are two main aspects of risk exposures, namely risk severity and the likelihood. The basic expression of KRI is proposed to measure the risk severity (S) with three levels, which are serious risk or high risk (SR), medium risk (MR) and low risk (LR), as follows:

$$KRI(S) = \begin{matrix} [LR] & [MR] & [SR] \\ \begin{cases} PSD & \leq 1 & \\ TIT & & > 0 \\ CPI & & > 0 \end{cases} \end{matrix}$$

Note that $KRI(S)$ is calculated with respect to an individual vehicle or the entire vehicle stream for a specific time and road space. The risk levels of LR and MR are primarily determined by TIT values, where TIT = 0 corresponds to risk with a low potential of an accident, especially when $PSD \leq 1$. CPI is then used to measure whether the risk is avoidable or not, which distinguishes the conflicts between MR conditions and SR events. The calculations of TIT and CPI are threshold-sensitive, which are evaluated based on the threshold values of TTC and DRAC, respectively. Since the risk levels are distinguished by the hybridised KRI, the impact of different threshold values is not very critical. It is a new insight in measuring risk levels, which is also flexible in defining straightforward thresholds in determining risk levels.

The risk likelihood (L) is measured by the total time of different severity levels with respect to the time period (T), defined as follows:

$$KRI(L) = \begin{cases} MR: & \sum_{i=1}^{M} \sum_{t=0}^{N} \alpha_i(t) \cdot \tau_{sc} \\ SR: & \sum_{i=1}^{M} \sum_{t=0}^{N} \beta_i(t) \cdot \tau_{sc} \end{cases}$$

$$\alpha_i(t) = \begin{cases} 1 & \forall\, TIT > 0 \;\&\; CPI = 0 \\ 0 & otherwise \end{cases}$$

$$\beta_i(t) = \begin{cases} 1 & \forall\, CPI > 0 \\ 0 & otherwise \end{cases}$$

where M is the entire vehicle stream in specific road space, τ_{sc} is a small-time interval ($T = N \cdot \tau_{sc}$), and $\alpha_i(t)$ and $\beta_i(t)$ are switching variables with value 1 and 0 for risk level judgement.

$KRI(L)$ measures the accumulation of vehicle-level risks and aggregation of flow-level risks for different severity levels. As a result, the likelihood can be used to describe risk exposures with different frequency levels, such as occasional, probable, frequent, etc. Therefore, $KRI(L)$ is meaningful to monitor the trends of risk exposures.

It is worthwhile to highlight that the risk assessment using KRI is based on the following assumptions that: (1) the higher severity of a risk condition or a vehicle behaviour, the more likely an accident will happen; (2) the harder a risk to be avoided by available evasive manoeuvres, the more likely an accident will happen; and (3) the more accumulation of vehicle-level risks or aggregation of flow-level risks, the more likely an accident will happen.

3.5 Additional Case for Validation

3.5.1 Validation procedure

To further validate the concept of KRI, a second real-world accident case is examined. Detailed analysis of this accident for validation of KRI is provided in the following subsections, including accident description, KRI-based risk assessment with predefined thresholds, risk exposures in terms of traffic flow and individual vehicles, etc. The findings of the second accident provide supporting evidence that there are certain pre-accident risk conditions for which KRI-based risk assessment is

helpful to determine such risk conditions at different levels. As consistent with the concept of KRI, the risk levels are distinguished by hybridised KRIs, and predefined threshold measurement is also workable. The results of the two accidents are in close agreements. Notwithstanding, for a proper and scientific evaluation of KRI, calibration and validation using more accident cases is still important, and this aspect is discussed in the discussion section.

3.5.2 Second accident case

An independent real-world accident is examined to further validate the concept of KRI. The second accident was also a head-to-rear collision that occurred on the fastest lane of an expressway carriageway during mid-afternoon. The road segment with grid remapping is shown in Figure 3.7(a). The traffic scenario at the time of the accident is provided in Figure 3.7(b).

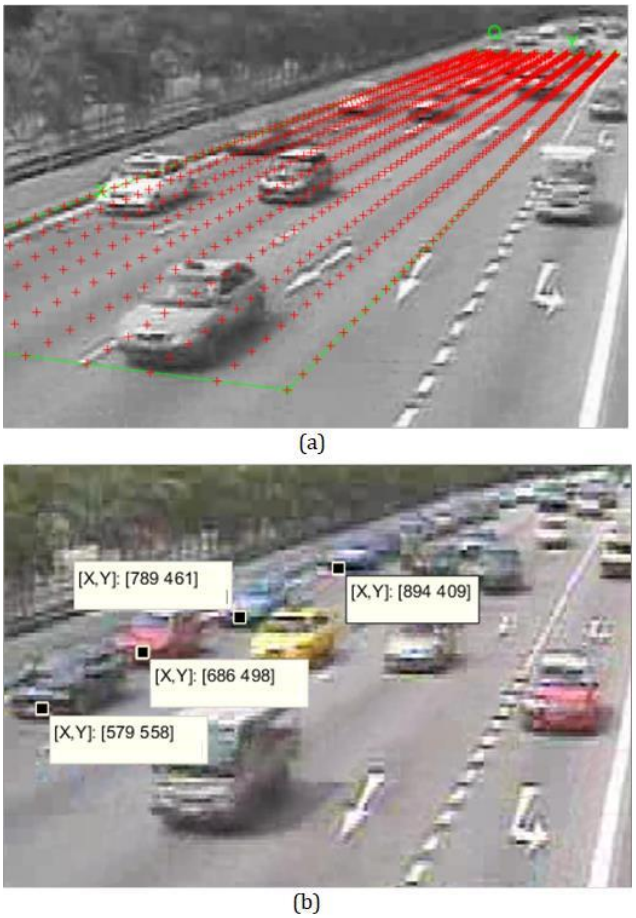


Figure 3.7 Road segment with grid remapping

At the 44th second, the vehicle with pixel coordinates (894, 409) was hit by a following vehicle, which thus led to a multi-vehicle collision. Similarly, the 40-second pre-accident vehicle trajectory data is extracted from video images and transformed into real-world units.

3.5.3 Analysis of validation of KRI

Based on the proposed KRIs, the risk levels of SR, MR and LR are determined. The risk conditions of MR-level and SR-level are figured out, using straightforward threshold measurement that CPI>0 for SR, and TIT>0 for MR. TTC threshold is 4.0 seconds for TIT calculation, and CPI value is the average of two results calculated based on MADR1 and MADR2 respectively. KRI-based risk assessment for traffic flow (defined by vehicle counts per 0.5 seconds for the fast-lane segment) is provided in Figure 3.8(a), and KRI-based vehicle-level risk assessment is provided in Figure 3.8(b).

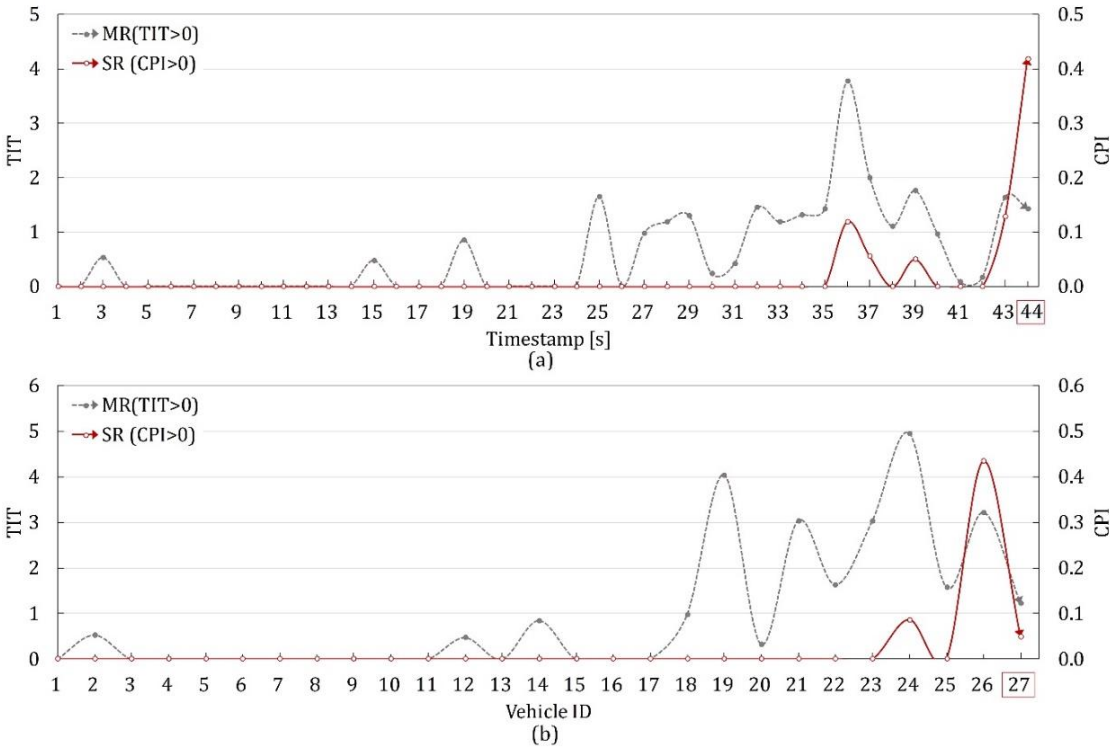


Figure 3.8 KRI-based pre-accident risk assessment

In Figure 3.8(a), the risk exposures in terms of traffic flow are identified, including risk severity and duration, the trends of risk aggregation, etc. There are two SR-level risk conditions (the 35^{th}-40^{th} seconds, and after the 42^{nd} second), which are determined by the aggregated CPI values. From a temporal comparison, most of the identified higher-risk conditions are presented during the period of 10-second before the time of accident occurrence. In addition, the trends in risk aggregation are also provided in Figure 3.8(a). The direction and range of risk trends can be used as early signals to infer accident likelihood. Herein, these findings provide supporting evidence that there are certain pre-accident risk conditions for which KRI-based risk assessment is helpful to identify such risk conditions.

The vehicle-level risk exposures are also identified by KRIs, as shown in Figure 3.8(b). The KRI-based vehicle risk assessment also reveals many insights about risk information in the vehicle stream, including at-risk vehicles (vehicle-pairs), risk accumulation within individual vehicles, and risk portrayal among vehicles, etc. From the vehicle with ID 17, the accumulations of MR-level risks have a growing trend and portrayal within vehicle platoons. Similarly, SR-level risk portrayal is also exhibited from the vehicle with ID 23. Therefore, the risk accumulation and portrayal can be used to infer the most-at-risk vehicles.

However, more details about near-accident events were not well captured in the second accident. The main reason is that the accident occurred at the far-ground of the remapping area, which is thus not conducive for detailed observation and data collection. In addition, the time span of the near-accident duration was relatively much shorter.

3.6 Discussion

KRIs can identify pre-accident risk exposures in terms of severity and likelihood. Higher risk levels and frequency have predictive values and can act as early signals of accidents, and constant monitoring of KRIs is helpful for accident assessment and prediction. Combined with the forecasting of vehicle movements, the at-risk vehicles, locations and timestamps of potential accidents can be inferred in real-time.

Therefore, targeted prevention could be applied pre-emptively. Furthermore, surrogate measures using KRIs are also helpful for risk-related analysis, such as risk data labelling, risk feature learning, and risk pattern recognition. For example, the proposed KRI has been applied to label risk levels of vehicle driving data, and using feature learning to identify risk behaviours and make accident prediction (Shi et al., 2018a).

This chapter provides a proof of concept of the KRI from an experimental standpoint. To get a stronger evaluation and validation, in-depth study using more accident cases is recommended. The scope of the technique can be extended by applying the method to additional real-world accidents cases, including various types of accidents and near-miss cases collected from expressways, intersections, arterial roads, etc. More accident cases can be analysed to allow examination of the sensitivity and specificity, the false positive rates and false negative rates, and d-primed value, etc. In addition, appropriate threshold values can be defined to further subdivide risk levels. For instance, a possible subdivision of the SR-level could be extra serious risk or near-accident level when CPI>0.5. Besides, there is a lack of leading indicators for accident prediction. Leading indicators should continue to be developed with the aim of proactive prevention with a lead time.

3.7 Chapter Summary

The risk assessment using hybrid and hierarchical indicators (the KRIs) offers new insights about risk exposures, which is helpful for accident assessment and prediction. In this chapter, the feasibility of using KRIs to measure pre-accident risk exposures conditioned on real-world accident data is assessed. Three main findings are summarised in the followings.

First, the concept of KRI is formulated to assess risk exposures using hybrid indicators. Seven individual indicators are selected as the basic indicators of KRIs in terms of risk behaviour, risk avoidance, and risk margin. A temporal-spatial case-control study is designed to investigate the feasibility of each indicator with the key findings that TIT can identify many pre-accident risk conditions and CPI can pick out

the most severe conditions (the near-accident) from among these risk conditions. In addition, it is found that the impact of different threshold values is not very critical in determining risk levels.

Second, KRI uses hybrid indicators to hierarchically distinguish various risk levels. The expressions of KRIs have been developed mainly based on TIT and CPI, to measure risk severity with three levels, as well as their likelihood. It is also flexible in defining straightforward threshold values for classification of risk levels. The KRIs and their threshold measurements are then further validated by another independent accident case.

Last but not the least, a grid remapping method has been developed to obtain vehicle trajectory data from surveillance video system, which can be applied for coordinates transformation from image pixels to real-world units with high quality. In addition, two real-world accident events and their antecedent (pre-crash) road traffic movements are retrieved, which unveils valuable insights of pre-accident risks.

CHAPTER 4
UNSUPERVISED LEARNING FOR VEHICLE-LEVEL RISK GRADING BASED ON SURROGATE INDICATORS

4.1 Chapter Introduction

This chapter focuses on the risk assessment of vehicles in driving for a general traffic condition. Unsupervised risk rating is proposed to group a large number of vehicles into distinct clusters based on risk pattern similarity, and identify detailed risk levels of each cluster.

Section 4.2 analyses the modelling framework and challenges of unsupervised risk grading, including basic clustering methods, class imbalance problems, lack of ground truth labels, etc. Section 4.3 elaborates the methodology and proposed solutions for imbalanced clustering, including ensemble clustering by majority voting, label identification by classifiers for evaluation, and risk indicator features as the inputs. In Section 4.4, NGSIM trajectory data is used as a case study. The results of hierarchical risk grading are presented, as well as the recommendations of risk levels. Section 4.5 describes the benchmark for risk estimation, and risk mapping and positioning. The ending two sections cover discussion and summary.

This chapter includes part of contents in the following paper and manuscript:

Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C. and Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis and Prevention, 129, 170-179. DOI: 10.1016/j.aap.2019.05.005.

Shi X., Wong Y.D., Li M.Z.F., Chandrasekar P. and Chai C. (2018). Feature extraction and clustering for vehicle-level risk grading based on surrogate indicators. Accident Analysis and Prevention. [under review]

4.2 Unsupervised Risk Grading

4.2.1 Risk grading model

Risk grading plays a key role in traffic safety, as it is important in a range of areas, such as, proactive crash prevention, identification of risk determinants, driving behaviour evaluation, among others. There is a perennial quest to assess and predict risk levels, for enhancing traffic safety.

Apart from measuring pre-accident risk signals (as discussed in Chapter 3), an alternative way to achieve risk grading is by assessing the driving behaviours of a large number of vehicles. The risk potentials are identified and grouped into graded levels. This can provide complementary insights on risk grading on a wide scale. Machine learning has an advantage in this task, since machine learning works well on handling voluminous information and discovering complex patterns without subjective intervention. Therein, clustering is a promising tool for discovering patterns from the data, which is to group similar instances into the same clusters, with each cluster containing similar pattern (Lin et al., 2017).

Interesting insights about risk grading have been presented by risk pattern identification and clustering (Guo and Fang, 2013; Sze et al., 2014; Nitsche et al., 2017). Notwithstanding, the clustering methods for risk grading have not been well investigated. Shortcomings of current clustering include instability, difficulty in deciding the number of clusters, and complexity in interpreting and decoding the clusters and patterns. Besides, the unavailability of ground truth labels is also a challenging problem.

Moreover, risk grading is a distinctly imbalanced problem. The frequency of risk conditions and accident cases (i.e. the minority class) is usually much smaller than the number of safety instances (i.e. the majority class). Assignment of risk instances into incorrect safety class entails a great misclassification cost (Elkan, 2001). Imbalanced data is challenging since standard learning processes are usually guided by global performance measures, such as accuracy rate. Learning algorithms are often biased towards the majority classes (known as the negative class), therefore, the

minority class (called the positive instances) might be ignored or wrongly discarded (treated as noise or outliers) (Díez-Pastor et al., 2015), hence, a remarkably high misclassification rate for the minority class instances would be produced (López et al., 2013). Moreover, there are observed challenges in class imbalance, related to data intrinsic characteristics, such as: (1) the presence of small disjuncts; (2) lack of density and information; (3) problem of overlapping between classes; (4) the significance of borderline instances to discriminate between positive and negative classes; and (5) the identification of noisy data, among others (Beyan and Fisher, 2015).

Furthermore, another key point is that the evaluation criteria for clustering with imbalanced data are still severely lacking. Imbalance problems make the clustering for risk grading much more complicated. Class imbalance problem is a hot topic being investigated recently in supervised learning and data mining, and some techniques (e.g. data resampling) are proposed to reduce class imbalance, which are discussed in Chapter 5. However, the unsupervised learning with imbalanced data is not well studied, since the detailed classes are unknown, instead, the risk clustering procedure is to find the potential imbalanced classes.

4.2.2 Clustering methods

Clustering-based risk grading of vehicles in driving entails using algorithms to group the vehicles (or driving behaviour) with similar risk patterns, and then estimate the risk level of each group by pattern decoding. This method can provide data-driven insights about risk exposures, and act as a procedure of labelling risk levels on moving vehicles, as well as the identification of risk conditions.

Cluster analysis is aimed at classifying elements into categories (or clusters, groups) on the basis of their similarity (Rodriguez and Laio, 2014). The basic mechanism and procedures are: (1) form a number of random cluster centres (prototypes), $C = \{c_1, c_2, \dots, c_k, \dots\}$; (2) estimate the proximity of instances to the respective cluster centres (e.g. Euclidean distance) and assign to the nearest cluster, $\arg\min D(x_i, c_k)$; and (3) generate a partition matrix with the maximised within-cluster similarity and

between-cluster dis-similarity, $x_i \in c_k \rightarrow (x_i, y_i),\ y_i = \text{label}\ (c_k)$.

Representative clustering algorithms include K-means, Fuzzy C-means (FCM), and Self-Organising Map neural network (SOM) (Kohonen, 2013; Lin et al., 2017). The main difference in the various clustering techniques pertains to cluster assignment. K-means measures the cluster scatter by minimising the within-cluster sum of squares, or variance, while FCM additionally introduces the membership and fuzzifier, namely, fuzzy belongingness. SOM produces the best matching unit by competitive learning and weight vectors. Details on these methods can be found, to name a few, in Izakian et al. (2015), Llanos et al. (2017), and Qin et al. (2017).

4.2.3 Feature extraction

Feature extraction and selection plays a key role in improving the quality and performance of machine learning. Feature extraction is a preliminary step, which serves to bridge the gap between raw data and algorithm input (Guyon and Elisseeff, 2003). Feature extraction and selection is concerned with selecting an optimal subset of features that are ideally sufficient and informative, and hence could improve performance and efficiency while reducing misleading information (Zarshenas and Suzuki, 2016). High interpretability is critical for sufficiently convincing and useful results, since risk patterns and underlying mechanisms can be highly complicated. Besides, strategies working at the feature level might be a potential path to deal with the imbalanced problem. However, the application of features for risk grading has been rarely examined.

Targeted for risk grading, feature extraction is expected to derive and construct information that is effective to distinguish between risk and safety. In addition, since risk grading is an imbalanced problem, functional features should be able to emphasise the learning on the minority classes, such as, improve the density of minority information, reduce class overlapping, concentrate the measures of the minority instances, etc. In this way, the minority will not be simply ignored or treated as outliers.

In traffic conflict techniques, some risk surrogate indicators are reliable to distinguish

between risk and safety (as discussed in Chapter 3). Such indicators offer insights about conflicts, hence crash risk potentials. Among them, some indicators emphasise on the risk instances by setting a proper threshold range. On the other hand, by virtue of data availability, vehicle movement trajectory data is suggested to be used for the risk indicator calculation. Herein, the risk indicator features are extracted from vehicle trajectory. More analysis of trajectory-based feature extraction is provided in Chapter 5. The data acquisition for practical applications is discussed in Chapter 6.

4.2.4 Evaluation metrics

It is problematic to define good clustering, since there is no ground truth data (i.e. labels, partitions) available for comparison or validation. Instead, the attempt is to find a convincing partition based on the data-driven outcomes. Besides, there is no consensus on what good clustering should be, and the clustered labels generally only indicate one possible grouping of the data set.

In clustering, some internal validity measures have been defined to give quality estimates of the clustering performance. The silhouette index (SI) and the Calinski-Harabasz score (CH) are two well-known metrics (van Craenendonck and Blockeel, 2015). They are a ratio of cluster compactness (instances in the same cluster should be similar) and separation (instances in different clusters should be dissimilar). The main difference is the definitions of compactness and separation. SH and CH provide a comparison of partitions with different numbers of clusters from the properties intrinsic to the data set. However, the validity for imbalanced data is not clear, and relying only on internal metrics might lead to erroneous conclusions.

4.3 Methodology

4.3.1 Progressive ensemble clustering

Ensemble clustering is essentially a stacking of outcomes by multiple runs, with each run being built using diverse algorithms and settings (e.g. random initial points, tuneable hyper-parameters), and in the end, all the outcomes (i.e. grouped risk levels) are summarised (e.g. by majority voting). A more robust and reliable result is

achieved by integrated modelling. Ensembles combine multiple learner systems which typically produce better results (Zhang and Suganthan, 2014). For a practical reason, ensembles generate a portfolio of outputs by diverse algorithms and settings, and the integrated result is more reliable and closer to reality. For a theoretical reason, the performance of ensemble learning is better, on average (Yu et al., 2016). Hybridised modelling has better convergence and generalisation.

In view of the class imbalanced problem and lack of ground truth labels in grading driving behaviour, a progressive ensemble clustering is designed with multiple numbers of clusters (i.e. various C values). A range of prospective values of C is predefined, which are evaluated to determine the most appropriate one. It is noted that generally with the increase of C values, a hierarchical partitioning is formed, and the minority classes are separated at high-resolution levels. The pseudo-code of the proposed progressive ensemble clustering is described in Algorithm 4.1.

Algorithm 4.1 Progressive ensemble clustering

1. Define C ($N = k$) (potential numbers of clusters);
2. Extract features of instance i, generate feature vector $x_i = \{f(\cdot)\}$;
3. Select S numbers of clustering algorithms as base learners;
4. For each base learner, do:
    a. Run T times of clustering on $X = \{x_i\}$ independently:
        1. Split X into k groups;
        2. Produce partition matrix $x_i \in c_j(t|s), j \in [1, k]$;
        3. Assign the label $y_i(t|s) \leftarrow$ label $c_j(t|s)$.
5. Majority voting $y_i = \max_{count}\{y_i(t|s)\}$;
6. Result $C^{N=k}(X) = \{x_i, y_i\}$ and then find the best-suited k.

Majority voting is employed as the fusion rule of the outcomes, which means the most frequently occurring value is assigned as the final label (Hapfelmeier and Ulm, 2013). To increase diversity, three qualified clustering algorithms are ensembled in the bucket, which are FCM, SOM, and K-means++ (an improved K-means). Besides, each algorithm runs equal times repeatedly to: (1) level off the uncertainty in

clustering; (2) determine the best performing component; and (3) fairly vote with equal weights. Since no ground truth label is available to guide the modelling, diagnostic checks are run for algorithm selection and hyper-parameter tuning. To build a good ensemble, the diversity of the base models is essential (Díez-Pastor et al., 2015), which is improved by using random initial points and various hyper-parameters.

4.3.2 Extraction of risk indicator features

Risk indicator features are built based on surrogate measures of vehicle conflicts, which are effective and reliable to evaluate risk potentials. More information about risk surrogate indicators is described in Chapter 3, especially the feasibility of using risk indicators to assess pre-accident risk conditions.

The process of constructing risk indicator features is elaborated as follows. Firstly, a series of variables are derived from raw trajectory data, including velocity, acceleration, the preceding and following vehicles, etc. Secondly, the risk surrogate indicators are calculated, which returns indicator time-series data. Afterwards, the third step is to capture the diagnostics on risks. Some operations are defined to present the key information of indicator data series, such as descriptive statistics, threshold-based filtering, aggregated or accumulated values, etc. Finally, the most relevant features for risk grading are selected as the clustering inputs.

For indicator time-series data (e.g. TTC, DRAC, PSD), descriptive statistics are used to represent the main characteristics, including values of minimum, maximum. For threshold-sensitive indicators, several features are generated based on different threshold values. For example, the thresholds of TTC ranging from 1.5s to 4.0s are considered (Shi et al., 2018a).

Besides, two additional features are proposed to describe relative risk signals, namely, the risk safety ratio (RSR) and high-risk rate (HRR). RSR is to describe the relationship of risk conditions and safety, which is measured by the time under risk conditions divided by the non-risk time. HRR is to compare the degree of high-risk, which is measured by the proportion of the occurrence time of high-risk conditions

with respect to the time under all risk conditions. Herein, the risk condition is measured by TIT>0, and the high-risk condition is defined by CPI>0, as discussed in Chapter 3. RSR and HRR are more flexible, since they are relative values defined at given time intervals.

$$RSR = \frac{\sum t_{TIT>0}}{T - \sum t_{TIT>0}}$$

$$HRR = \frac{\sum t_{CPI>0}}{\sum t_{TIT>0}}$$

According to potential usefulness demonstrated in Chapter 3, some risk indicators are suitable for feature construction, such as TIT and CPI. Herein, a set of 12 risk indicator features are built based on different settings of threshold values or threshold measurement, as shown in Table 4.1.

Table 4.1 Risk indicator features selected for clustering

Abbreviation	Explanation
TTC.Min	Minimum value of vehicle TTC(t)
TET	Vehicle TET for a specific scope, under TTC threshold 3s
TIT.1	Vehicle TIT for a specific scope, under TTC threshold 2s
TIT.2	Vehicle TIT for a specific scope, under TTC threshold 3s
TIT.3	Vehicle TIT for a specific scope, under TTC threshold 4s
DRAC.Max	Maximum value of vehicle DRAC(t) for the defined scope
CPI.1	Vehicle CPI under MADR1 measure
CPI.2	Vehicle CPI under MADR2 measure
PSD.Mean	Mean value of vehicle PSD(t)
PSD.Min	Minimum value of vehicle PSD(t)
RSR	The ratio of time under risk exposures, measured by TIT.2>0
HRR	The proportion of time under high risk conditions, measured by CPI.1>0

These features are interpretable, and represent the risk exposures of a vehicle in driving over space-time domain and in terms of temporal, kinematical and spatial aspects. Of course, some of the features are either redundant or less relevant for risk

grading. However, classical methods of feature selection are generally problematic in an unsupervised scenario, especially with imbalanced data. Herein, the feature selection is performed based on domain knowledge.

For imbalanced clustering, clipping thresholds are introduced to define a measurable scale for similarity comparison. A clipping threshold is described as a boundary level of sufficient safety/risk, and the portion of out-of-range values are truncated. By clipping, the weight of the non-risk class is diluted to a limited range so that the learning emphasises on the risk classes. Moderate clipping thresholds are set to avoid underestimation of the non-risk class, for example, the clipping threshold value for TTC is set as 5s, the threshold of 10g (m/s^2) for DRAC, and threshold of 1.0 for PSD. The clipping thresholds generally describe a larger boundary than the indicator thresholds used to distinguish between risk and non-risk conditions.

Some studies have used behaviour-based features to measure risks, such as speeding (e.g. maximum velocity value), excessive braking (e.g. maximum deceleration value), gap maintenance (e.g. minimum gap, variance), lane keeping (e.g. lateral position deviation), etc. However, the indicator-based features are more sensitive and functional in risk identification, as discussed in Chapter 3. Indicator-based features contribute to improve class separability and somehow reduce the degree of class overlapping. Although there are other complex indicators and procedures to build the features for clustering, transparency is valued over sophistication.

4.3.3 Label identification by classifiers

Label identification by classifiers is proposed to evaluate clustering performance, for conditions without available ground truth data. This method makes use of independent classifiers to investigate whether the clustered labels can be correctly identified by other classifiers and the degree of identification. It is performed by using supervised learning of the features and corresponding labels, and checking the cross-validation prediction performance (e.g. measured by misclassification). This method provides data-driven insights about the underlying information and possible structures that are close to the actual partition, and generates a feasible way for

clustering evaluation.

A bucket of classifiers is employed, therein, weak classifiers serve to figure out instances that are likely to be misclassified, and strong classifiers serve to find the best clustering outcome that produces labels with the highest cross-validation accuracy. A better clustering produces a partition with less overlapping and better distinction of different risk levels, even though using a weaker classifier can offer more precise discrimination. Therefore, a comparison of various clustering outcomes can be made.

Label identification by classifiers is to solve an unsupervised learning problem using supervised concepts. In supervised learning, some metrics are designed to understand the performance of a classifier given a data set (i.e. the features and corresponding labels). It is reasonable to assume that better labels would be easier to fit by classifiers in the training stage, and hence achieve a better generalisation performance (e.g. a lower out-of-sample misclassification rate) on the testing dataset.

For classification, the confusion matrix records the results of correctly and incorrectly identified instances of each class independently (Lever et al., 2016). Herein, a confusion matrix is adapted to focus on two-class clustering problem, as shown in Table 4.2. From the adapted confusion matrix, some metrics can be deduced to perform the evaluation. The metrics and equations are depicted in Table 4.3.

Table 4.2 Confusion matrix adapted for binary clustering

	Identified as positive class by classifiers (minority class) [+]	Identified as negative class (majority class) [-]
Clustered as positive class (minority class) [+]	True Positive (TP)	False Negative (FN) [Type II error]
Clustered as negative class (majority class) [-]	False Positive (FP) [Type I error]	True Negative (TN)

Table 4.3 Metrics built from the confusion matrix

Metrics	Function
TPrate, Sensitivity, Recall	TP/(TP+FN)
FNrate, Miss rate	FN/(TP+FN)
TNrate, Specificity	TN/(FP+TN)
FPrate, 1-specificity, false alarm rate	FP/(FP+TN)
Precision, positive predictive value (PPV)	TP/(TP+FP)
False discovery rate (FDR)	FP/(TP+FP)
False omission rate	FN/(FN+TN)
Negative predictive value	TN/(FN+TN)
Accuracy	(TP+TN)/(TP+FP+TN+FN)
F1 score	2TP/(2TP+FP+FN)

Besides, in classification evaluation, some integrated metrics are commonly used, such as F-measure, Geometric Mean (of sensitivity and specificity) (Díez-Pastor et al., 2015). F-measure uses the parameter β to control the balance of recall and precision. As β decreases, precision is given greater weight (Lever et al., 2016). F-measure is defined as follows:

$$F_\beta = \frac{(1+\beta^2)(Precision \times Recall)}{\beta^2 \times Precision + Recall}$$

With β = 1, the F_1 score is defined, which assigns equal importance to recall and precision.

In addition, the receiver operating characteristic (ROC) curve visualises the trade-off between TPrate (benefits) and FPrate (costs), and the Area Under the ROC Curve (AUC) is computed. Most studies in the literature widely employ AUC as the main criterion for evaluation (López et al., 2013; Díez-Pastor et al., 2015). A greater AUC typically indicates a better performance. AUC is calculated as follows:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

For clustering-based risk grading, TP means the clustered risk instances that are also identified as such by independent classifiers. Recall describes the accuracy of risk identification over the total amount of clustered risk instances, while precision describes the accuracy of risk discovery among the classified risk instances. F-measure represents the ability to detect risk potentials.

For the imbalanced scenarios, some metrics might be misleading, such as accuracy. The higher-risk groups might be under-represented since they are the minority classes. However, the use of recall and precision for the minority and majority classes is recommended respectively, which can lead to the generation of more precise rules for the positive class (Díez-Pastor et al., 2015). Similar to ROC, an alternative visualisation is the precision–recall (PR) curve, which shows the best trade-off between the recall and precision. PR curve is more favourable than the ROC curve to visualise the classifier performance for imbalanced datasets (Lever et al., 2016). As a summarised metric of the PR curve, the area under the PR curve (AUPRC) is recommended for data sets with large class imbalances (Lever et al., 2016). Unlike the ROC AUC, it is less straightforward to calculate AUPRC. Additionally, the class imbalance is a confounding factor that can distort various metrics. Instead of using one specific metric, a basket of metrics is also included as supplementary criteria.

4.4 Clustering-based Risk Assessment

4.4.1 Data description (NGSIM)

Data availability and quality are essential for accurate data mining and risk pattern discovery. The models developed in this chapter utilise the vehicle trajectory data provided in the Next Generation Simulation (NGSIM) Program.

The NGSIM Program was initiated by FHWA which collected high-quality real-world traffic flow and vehicle trajectory data to support the research on microscopic modelling and algorithm testing. The vehicle trajectory data were collected from US Route 101 (Hollywood Freeway) on June 15th, 2005. More description of NGSIM data is provided in Appendix B.

The main variables used in the present models include: (1) vehicle identification number, i; (2) longitudinal coordinate of the front centre of the vehicle with respect to the entry edge of the road segment in the direction of travel, $x_i(t)$; (3) lateral coordinate with respect to the left-most edge of road segment, $y_i(t)$; (4) length of vehicle, L_i; (5) instantaneous velocity, $v_i(t)$; (6) instantaneous acceleration, $a_i(t)$; etc. The data resolution is 0.1 second in relative units.

After data preprocessing, a total of 5,084 instances (i.e. vehicles) is used for modelling, involving 3,203,867 records. Then the feature extraction is conducted, and the dimensionality is vastly reduced, downsized to 5084×12. The basic description of the 12 features is listed in Table 4.4.

Table 4.4 Descriptive information on values of vehicle risk features

	Mean	Std	Min	Q1	Median	Q3	Max
CPI.1	0.002	0.018	0	0	0	0	0.82
CPI.2	0.023	0.273	0	0	0	0	10.89
TIT.1	0.193	0.914	0	0	0	0.006	26.55
TIT.2	0.816	2.25	0	0	0	0.738	54.14
TIT.3	2.209	4.38	0	0	0.351	2.868	85.77
PSD.Mean	0.996	0.015	0.645	0.997	1	1	1
PSD.Min	0.983	0.056	0.186	1	1	1	1
DRAC.Max	0.401	0.569	0	0.181	0.298	0.477	9.8
TET	0.964	1.862	0	0	0	1.3	30.8
TTC.Min	4.593	0.762	0.124	4.474	5	5	5
RSR	0.021	0.376	0	0	0	0.019	2.962
HRR	0.003	0.067	0	0	0	0	0.831

The indicator features are calculated using the full data set available, which was collected from a 640-metre road segment for about 45 minutes. In order to involve more risk conditions in the similarity comparison, large scope of sampling is suggested. Hence, the features are first calculated based on the full length of the lane segments. Besides, some unvalued records are removed because of missing values.

The NGSIM data is representative of high-quality trajectory data, which is also flexible to collect. Relevant techniques for trajectory data collection include traffic surveillance system, high-resolution in-vehicle positioning system, combined use of on-board units (OBU) and front-mounted radar, dashboard camera recording with video processing, to name just a few (e.g. Guido et al., 2010; Sivaraman and Trivedi, 2013b; Castignani et al., 2015; Wong and Wong, 2016). Most importantly, interpretable features are possible to be extracted from trajectory time-series data externally, hence are non-interference in nature.

4.4.2 Hierarchical partitioning

Based on risk pattern similarity, vehicles are clustered into different groups with graded risk ratings. Three clustering algorithms are integrated into the ensemble, which are FCM, K means++, SOM. These clustering algorithms capture the position (the risk level) of an instance in the entire dataset based on multidimensional views of data similarity; a label about the belonged position is assigned to the instance.

The problem of imbalanced data is tackled by conducting the clustering in a progressive manner, and a hierarchy is formed with increasing numbers of clusters. At different hierarchy levels, the group partitioning and boundary values are different, hence, an instance might be marked with different group labels. The hierarchical partitioning is presented based on different numbers of clusters (N). In this way, minority classes (i.e. the highest risk level) are gradually distinguished.

The hierarchical partitioning describes various clustering solutions with different numbers of clusters, as shown in Figure 4.1. Based on some preliminary experiments, a range from 3 to 9 is adopted as the number of clusters. In the ensemble, the clustering run has been repeated 30 times (i.e. 10 times per algorithm) independently, and the resulting label of an instance is determined by the majority voting, to smooth out the slight variability.

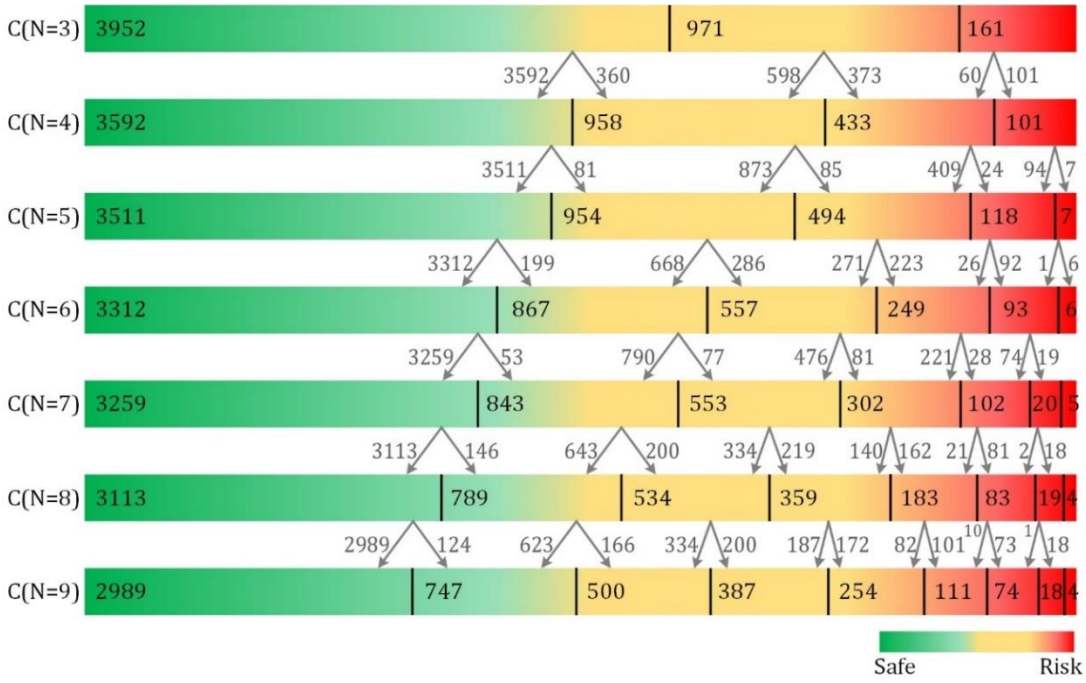


Figure 4.1 Hierarchical risk partitioning by clustering

In Figure 4.1, due to the high class-imbalance, the length of each partition bar stands for the square root of the number of instances assigned into each group. With the increase of N, the resolution is enhanced, and the risk gradings with more levels are revealed. The selection of an optimal value of N depends on both data structure and the requirement of potential applications. Besides, imbalance ratio (IR) is commonly used to measure the class imbalance, which is defined as the ratio of the number of instances in the majority class to the number of the minority class instances (García et al., 2012). A dataset is referred as highly imbalanced if the IR is greater than 9. Herein, even for N=3, the IR is 24.5.

Ensemble learning has shown to be more reliable than an individual algorithm, in both reliability and robustness. Nevertheless, these benefits come at a price, such as, longer computation time, complicated outputs interpretation for end users, since comprehensive information is ensembled. In addition, the ensemble system also illustrates a process of model selection. For the three algorithms in the ensemble, FCM presents a better overall performance, especially in the high stability; K-means generally produces two different outcomes due to random initial conditions; the hyper-parameter tuning and cluster reassignment in SOM are challenging. Herein, for

simplicity, FCM is best-suited for the clustering-based risk grading.

4.4.3 Clustering evaluation

The quality of clustering solutions is evaluated by other independent classifiers. The evaluation is done by investigating whether these clustered labels are correctly identified by other classifiers; and comparing the degree of identification. Since a convincing validation is problematic without ground truth labels, this strategy is useful to generate a data-driven comparison and evaluation.

Prospective classifiers include random forest (RF), classification and regression tree (CART), support vector classification (SVC), k-nearest neighbours (kNN), naive Bayes (NB), and logistic regression (LR), etc. Both weak classifiers (e.g. LR, for figuring out more misclassified cases) and strong classifiers (e.g. RF, for testing performance) are included, to examine the misclassifications under different learning abilities.

Firstly, internal measures (as discussed in Section 4.2.4) are used to provide intrinsic properties about clustering, as shown in Figure 4.2. The plot indicates that, according to the silhouette index (SI) and Calinski-Harabasz score (CH) (as described in Section 4.2.4), solutions with more than five clusters are better in terms of compactness and separation.

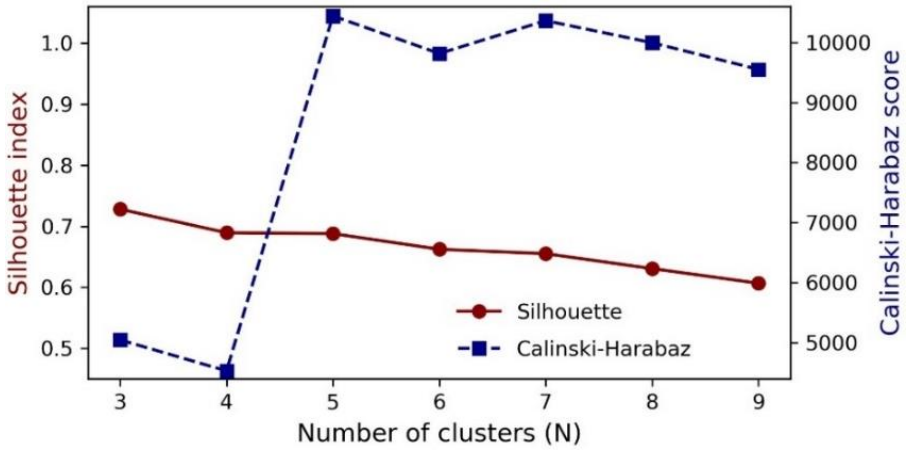


Figure 4.2 Metrics for clustering internal evaluation

Further external evaluation is conducted using the label identification by classifiers.

Given the data size and the extreme imbalance ratio, the stratified 3-fold (i.e. 2/3 data for training, and 1/3 for testing) is set in the cross-validation, since this splitting ratio has a better trade-off of the number of instances for training and testing, and using stratified k-fold ensures that each set contains approximately the same percentage of samples of each target class as the complete set. Necessary classifier hyper-parameter tuning is conducted in each fold based on the training dataset.

For the evaluation criteria, AUPRC is considered as the main metric, while others are listed for supplementary supports, including precision, recall, F1, (ROC) AUC. However, the comparison among clustering solutions with different numbers of groups is not straightforward. Besides, correct identification of the higher risk levels is of greater interests, but for the highest levels, the number of instances is limited. The metrics are calculated for each group, and two types of mean values are produced. One is macro-averaged over all classes, which is unweighted mean. Another is the average weighted by the number of true instances for each group.

The AUPRC and (ROC) AUC for different risk levels of each clustering solution are given in Figures 4.3 and 4.4, respectively. The x-axis denotes graded risk levels, and the y-axis denotes the average values obtained by four strong classifiers, which are RF, SVC, kNN and CART.

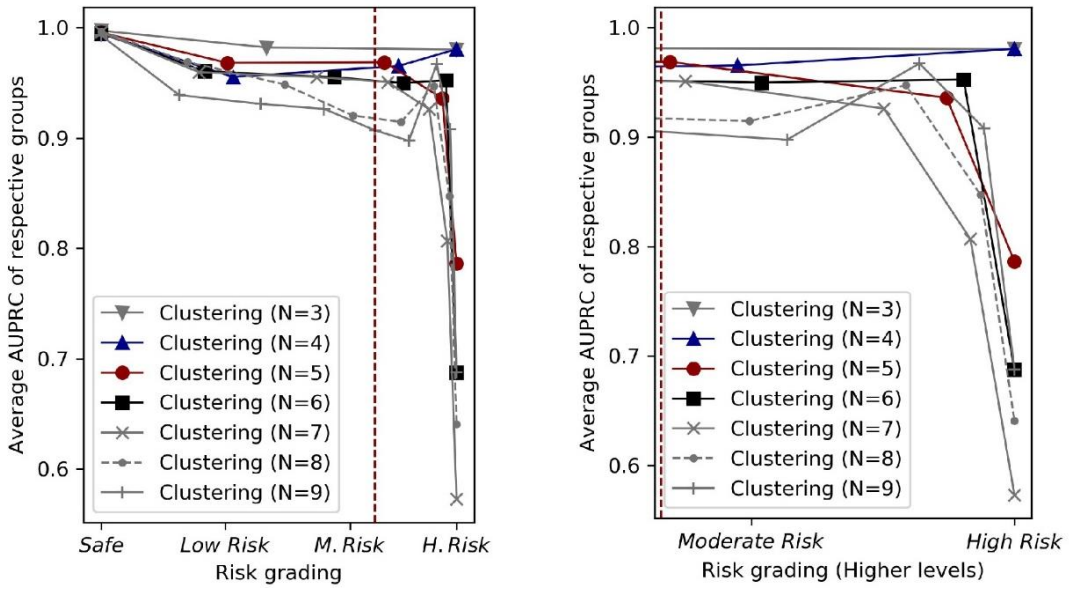


Figure 4.3 Average AUPRC of respective clustering groups

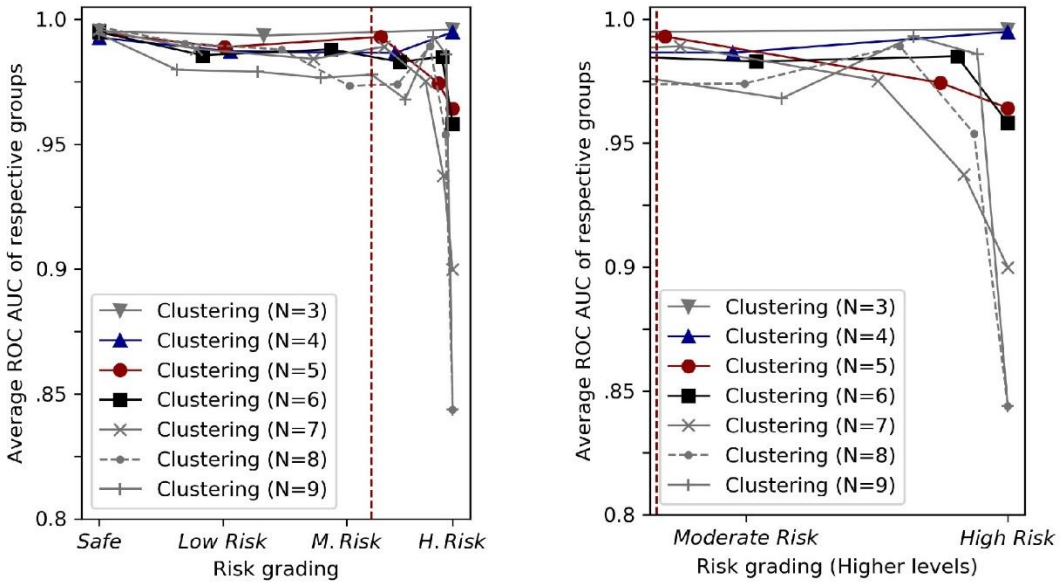


Figure 4.4 Average ROC AUC of respective clustering groups

From the comparison, the clustering solutions with 5-, and 6-groups show better trade-offs in terms of accuracy (measured by AUPRC and ROC AUC) and resolution. For the performance at higher risk levels, the clustering with six groups has a greater recall, and also maintains a higher precision. Herein, N=6 is selected as the preferred clustering, which is indicated by the line defined by the square points.

The clustering performance (for N=6) of the seven classifiers are enumerated, as given in Table 4.5.

The identification of clustered labels by independent classifiers is a feasible way to evaluate the performance of clustering and risk grading. It is reasonable to assume that the clustering with fewer misclassifications has better performance. In addition, the misclassifications are likely to happen on the edge (i.e. the borderline area surrounding a group) of two adjacent groups, or groups with overlapping, which implies that better separation would have fewer instances located on the borderlines and overlapping areas. Whereas for those correctly identified instances, they are likely to be placed in relatively homogeneous positions with respect to the risk labels. Hence, a better clustering would generate well-separated homogeneous areas (i.e. dense clusters). Of course, a small number of clusters would have a lower opportunity to make misclassifications (i.e. fewer borderline and overlapping instances),

however, the resolution (i.e. the ability to identify minority classes) is also low. Therefore, a reasonable clustering can be selected based on misclassification in conjunction with resolution.

As shown in Table 4.5, four classifiers (i.e. SVC, RF, kNN and CART) demonstrate a high accuracy, which implies that an underlying structure close to ground truth is promising to be discovered by the clustering.

Table 4.5 Evaluation by classifiers and clustering performance (for N=6)

	SVC	RF	kNN	CART	NB	LR	AdaBoost
Misclassified	34	53	61	67	390	691	1466
CA	0.993	0.990	0.988	0.987	0.926	0.872	0.712
Macro-average over all classes							
Precision	0.957	0.952	0.981	0.953	0.865	0.762	0.534
Recall	0.987	0.976	0.927	0.978	0.901	0.676	0.657
F1	0.965	0.959	0.944	0.955	0.880	0.690	0.572
ROC AUC	0.992	0.988	0.962	0.987	0.943	0.824	0.800
AUPRC	0.923	0.913	0.910	0.922	0.787	0.577	0.515
Weighted-average over all classes							
Precision	0.994	0.989	0.988	0.988	0.933	0.865	0.893
Recall	0.993	0.988	0.988	0.987	0.926	0.872	0.712
F1	0.993	0.988	0.988	0.988	0.928	0.861	0.656
ROC AUC	0.996	0.994	0.993	0.992	0.956	0.923	0.827
AUPRC	0.988	0.982	0.978	0.977	0.892	0.804	0.725

4.4.4 Risk pattern identification

The risk patterns of each clustered group are identified based on risk indicator features, and the risk levels are recognised. To facilitate the risk pattern decoding, the value distribution of the 12 features are summarised. For N=6, six groups with graded risk levels are identified, as illustrated in Figure 4.5.

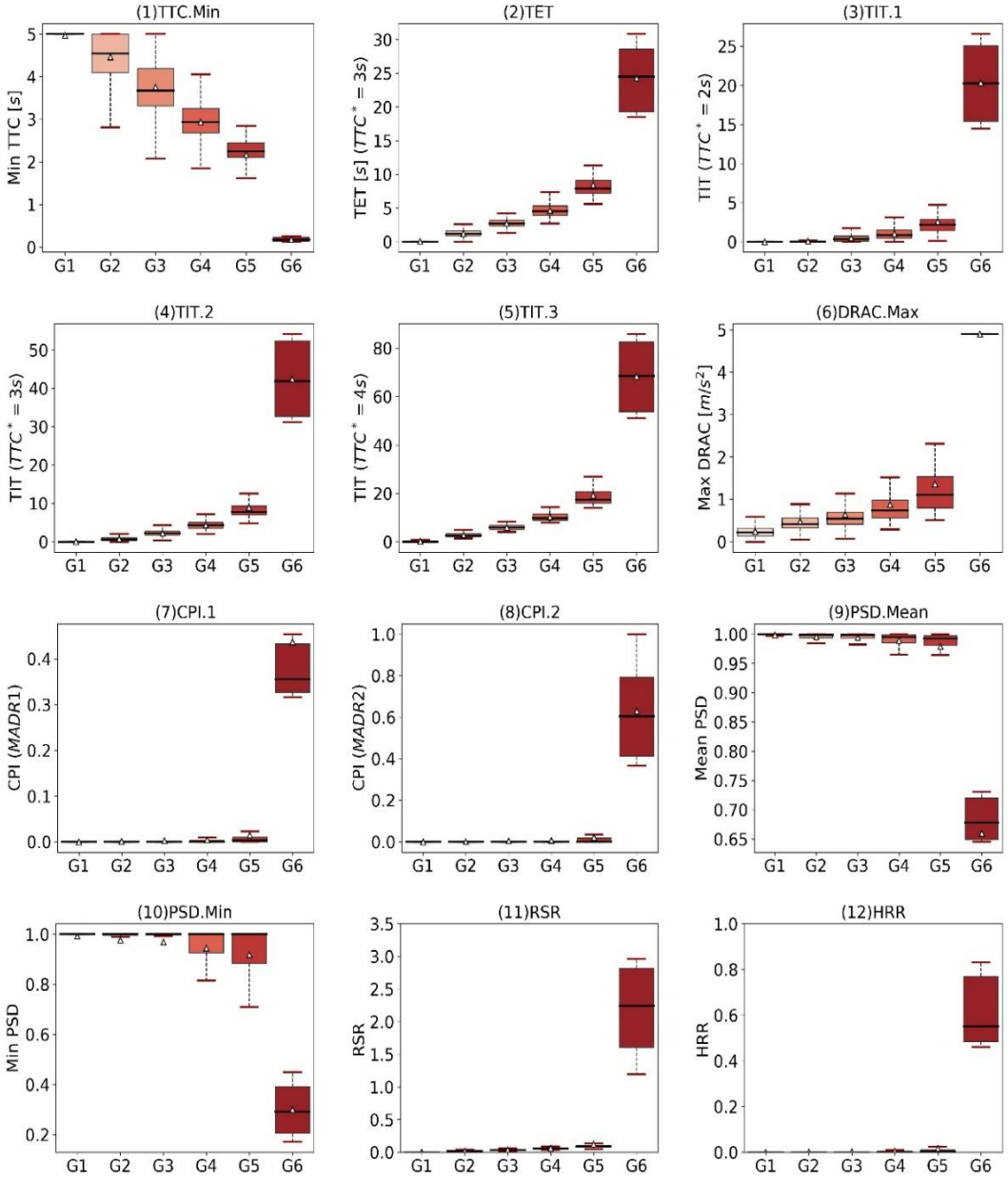


Figure 4.5 Feature value distribution for risk pattern decoding

In Figure 4.5, graded risk levels are represented using a grayscale colour map, where dark represents higher risk levels. According to a comprehensive comparison of the feature values, an annotation of the risk levels is safe level (Group 1), low-risk levels (Group 2: LR-, Group 3: LR+), moderate-risk levels (Group 4: MR-, Group 5: MR+), high-risk levels (Group 6: HR). This annotation is only to illustrate the procedure of unsupervised risk grading and risk level labelling, and users are free to decide other annotations, which are more suitable for their respective applications. Based on the

risk pattern of a group, the driving risks of the vehicles belonging to this group can be inferred. Note that risk levels are calculated with respect to the entire vehicle stream for a specific time and road space, and the definition of feature calculation is discussed in Section 4.5.2.

From Figure 4.5, some of the features tend to perform better in separating certain risk levels, such as TIT.3 and CPI.1. The feature TIT.3 has a less overlapping area (i.e. a region of data space that contains a similar quantity of data from two adjacent groups). CPI-based features serve to isolate the higher risk instances. These well-separated features have less overlapping between adjacent risk levels, hence the values of boundaries are used to set the thresholds for risk estimation. Besides, indicator-based features are interpretable, therefore, they contribute to risk pattern decoding and are easy to adopt by end-users. Furthermore, such feature-based method is helpful to handle classification for imbalanced data set, without the support of cost functions, algorithmic or data level processing (see Section 2.5).

4.5 Benchmark for Risk Estimation

4.5.1 Feature ranking and selection

Before clustering, preliminary feature selection has been conducted subjectively, since in an unsupervised scenario with imbalanced data, feature selection should be conservative, to avoid loss of key information. Herein, the clustering has produced the data labels of risk levels, which makes it possible to select the most important features in a supervised manner. Supervised feature selection is generally guided by the feature importance in learning. An in-depth study on learning-based feature selection is analysed in Chapters 2 and 5, including the methodology of feature selection (in Section 2.6) and feature selection modelling (in Section 5.2).

In this section, feature importance ranking is conducted using mean decrease impurity (also known as Gini importance) in random forest. The methodology can be found in Breiman (2001). Firstly, feature importance ranking is performed to obtain filtered features, and then permutes to find one feature subset with the best performance with respect to accuracy and interpretability. The procedure is based on data subsets

formed by different risk groups. As a post-clustering process, the final feature subset should maintain an equivalent clustering outcome. The feature importance for each data subset (formed by different risk groups) is investigated, as illustrated in Table 4.6.

Table 4.6 Feature importance ranking on different data subsets

	Data subset 1	Data subset 2	Data subset 3	Data subset 4	Full dataset
Groups	5 and 6	From 4 to 6	From 3 to 6	From 2 to 6	All groups
Features					
CPI.1	0.198(1)	0.091	0.035	0.036	0.039
TIT.2	0.161(2)	0.165(2)	0.162(3)	0.160(3)	0.163(3)
TIT.3	0.141(3)	0.233(1)	0.299(1)	0.305(1)	0.284(1)
TIT.1	0.140(4)	0.110(4)	0.074	0.069	0.057
TET	0.101(5)	0.129(3)	0.212(2)	0.189(2)	0.208(2)
RSR	0.061	0.087	0.087	0.100(4)	0.121(4)
HRR	0.060	0.038	0.013	0.015	0.006
CPI.2	0.039	0.041	0.021	0.025	0.012
PSD.Mean	0.039	0.002	0.004	0.023	0.018
PSD.Min	0.020	0.002	0.008	0.004	0.007
TTC.Min	0.020	0.040	0.054	0.062	0.057
DRAC.Max	0.019	0.061	0.030	0.014	0.028
Selected	CPI.1, TIT.2, TIT.3, TIT.1, TET	TIT.3, TIT.2, TET, TIT.1	TIT.3, TET, TIT.2	TIT.3, TET, TIT.2, RSR	TIT.3, TET, TIT.2, RSR

The importance rankings obtained by random forest are in favour of the features with more levels. The importance threshold is set as 0.1. The filtered features mainly include TIT.3, CPI.1, TET, etc. TIT considers both severity and time duration, which offers superior quality compared to TET and TTC. Besides, TIT.3 (i.e. TIT with 4s TTC threshold) is numerically better than TIT.2 and TIT.1, since TIT.3 shows

appreciable distinction with less overlapping, which is also shown in Figure 4.5. For CPI, since CPI.1 and CPI.2 essentially comprise two aspects of CPI evaluation, and their abilities in risk identification are found to be similar, hence one CPI value or an averaged value of the two could be used, instead of using two. In addition, RSR and HRR are measured based on TIT and CPI respectively, but they do not outperform.

The clustering results using different feature subsets are listed in Table 4.7, as well as the similarity to the original clustering outcome. To guide the decision of feature inclusion or rejection, the clustering performance of different feature subsets on each data subset (as described in Table 4.6) are investigated, as demonstrated in Figure 4.6.

Table 4.7 Clustering results using different feature subsets

	Original	Subset 1	Subset 2	Subset 3
Features	12 features	TIT.3; CPI.1	TIT.3; CPI.1; TET	TIT.3; CPI.1; TET; TIT.2
Results and similarity				
Group 1	3312	3234	3264	3295
		97.6%	98.6%	99.5%
Group 2	867	907	883	859
		95.4%	98.2%	99.1%
Group 3	557	576	584	581
		96.6%	95.2%	95.7%
Group 4	249	269	258	250
		92.0%	96.4%	99.6%
Group 5	93	91	88	93
		97.8%	94.6%	100%
Group 6	6	7	7	6
		83.3%	83.3%	100%

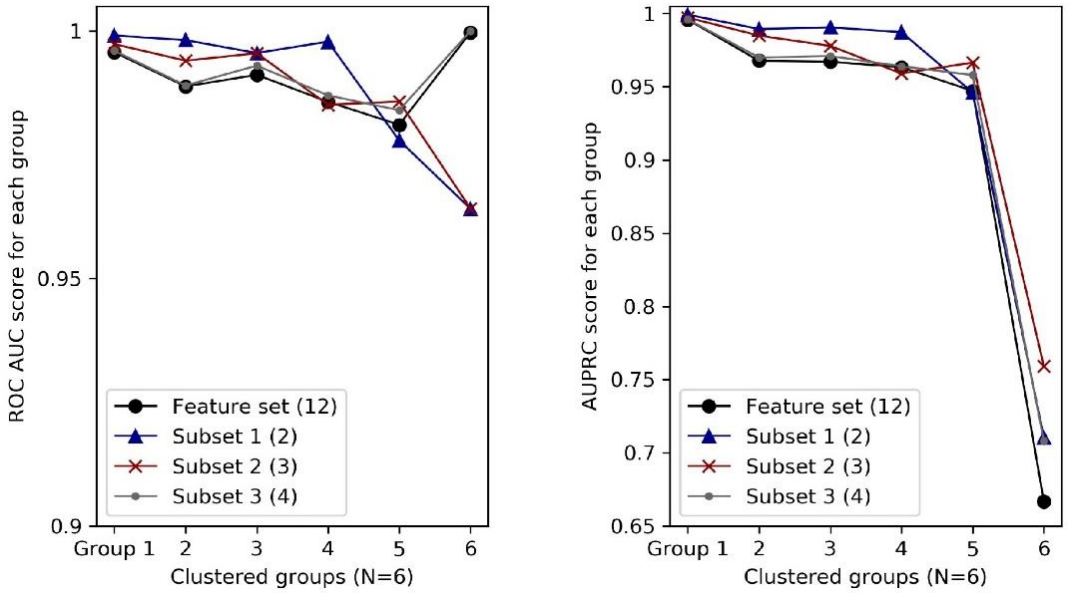


Figure 4.6 Clustering performance on different feature subsets

From Table 4.7 and Figure 4.6, the combination of TIT.3 and CPI.1 can maintain similar clustering outcome, which uses a minimal number of features, and offers appropriate performance. Herein, an eligible feature subset for risk estimation is determined.

For visualisation of feature-label relationships, the TIT.3 and CPI.1 values of instances in each risk level are plotted, as shown in Figure 4.7.

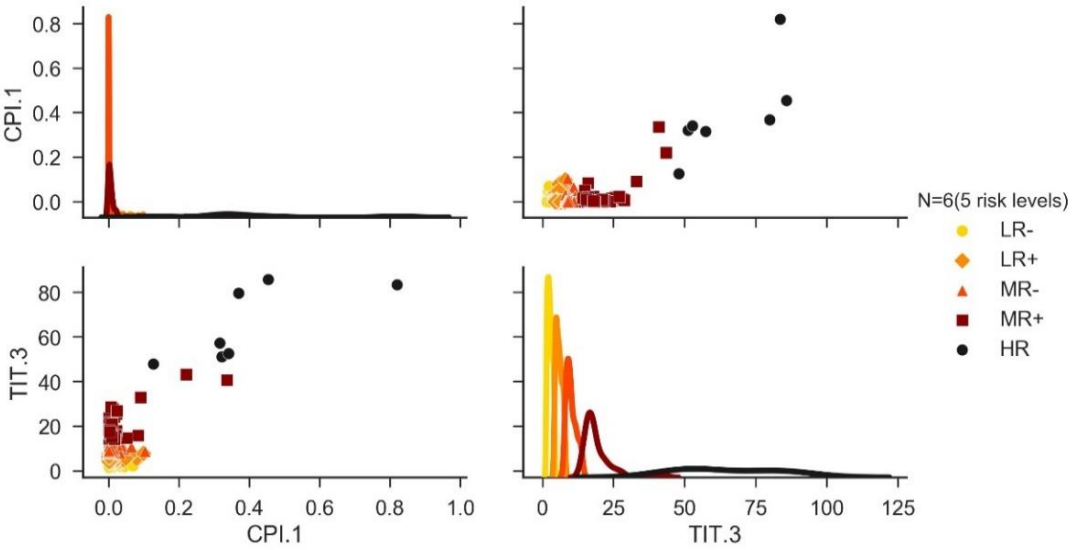


Figure 4.7 TIT-CPI diagrams of risk levels

The subfigures of pairwise relationships and univariate distributions share the x-axis and y-axis across a single row or column. Herein, TIT.3 contributes to a distinct range of risk levels, and CPI.1 exhibits the identification for higher risk levels. CPI and TIT are therefore suitable to be used as key risk signals to estimate risk potentials.

4.5.2 Scope for feature calculation

The risk partitioning of driving behaviour is structurally based on the complete set of available data. There are several considerations to do so, such as, to include more information for a stable risk hierarchy; to reduce the impact of small disjuncts and lack of density, etc. However, since some indicators measure the risk accumulation, hence, the scope for indicator calculation is important to estimate measurable risk levels and calibrate threshold values. The parameters to define such scope are mainly the travelled distance and/or time duration. In addition, since the features are extracted from trajectory time-series data, the information of the data stream may change with time.

To investigate the parameter settings, the data stream is broken into multiple continuous sequences under the same road coverage, which is moving observation windows defined by the travelled distance. Given a data sequence defined by travelled distance ΔD beginning at position $d(n)$, the risk indicator features are calculated based on the data sequence.

$$f_i = mean \sum_{d(n)}^{\Delta D + d(n)} f_i(x)$$

The comparison of different distances under the same road coverage is investigated, as shown in Table 4.8.

Numerically, shorter distances (e.g. 100 m, 200 m) still exhibit equivalent clustering structure. Herein, the suggested value is 100 m distance (or equivalent 5-6s time interval).

Table 4.8 Clustering results and accuracy under different distances

ΔD	640 m	100 m	200 m	300 m	400 m	500 m
Group 1	3312	3292	3422	3489	3489	3423
		99.4%	96.7%	94.7%	94.7%	96.6%
Group 2	867	857	859	844	845	862
		98.8%	99.1%	97.3%	97.5%	99.4%
Group 3	557	575	488	477	476	489
		96.8%	87.6%	85.6%	85.5%	87.8%
Group 4	249	269	230	198	198	225
		92.0%	92.4%	79.5%	79.5%	90.4%
Group 5	93	83	80	71	71	80
		89.2%	86.0%	76.3%	76.3%	86.0%
Group 6	6	8	5	5	5	5
		66.7%	83.3%	83.3%	83.3%	83.3%
Average accuracy		90.5%	90.8%	86.1%	86.1%	90.6%

The consistent feature parameters serve to generate a measurable benchmark for risk grading, which means only using the indicators and thresholds to measure risk levels, instead of clustering every time, since the clustering results may vary if the sample size is insufficient. As a demonstration using 100-metre distance, the hybrid indicators and simplified thresholds for risk estimation are shown in Table 4.9.

Generally, approximate risk ranges (e.g. MR, LR) are primarily determined by indicators, and threshold values are employed to identify the risk levels which the vehicle belongs to. The combined use of hybrid indicators and simplified thresholds is a helpful way to distinguish multiple levels, instead of relying on multiple threshold values of one indicator.

On the other hand, since features are recalculated by the defined time interval or distance, a large amount of data subsets are obtained based on the moving observation windows. The diversity of data subsets enhances the ensemble clustering, which can be considered close to the bagging technique in supervised ensemble learning.

Additionally, a defined scope of data sampling also facilitates the design of strategies to improve clustering performance, for example, diversity-increasing strategies such as random subspace sampling (Díez-Pastor et al., 2015).

Table 4.9 Hybrid indicators and thresholds for risk estimation

	Level	Mean	Std.	(Min, Max)	Threshold
TIT					
	LR-	0.43	0.16	(0.2, 0.7)	0
	LR+	0.95	0.23	(0.7, 1.3)	0.5
	MR-	1.64	0.34	(1.2, 2.2)	1
	MR+	3.13	1.02	(2.1, 4.7)	2
	HR	10.61	2.31	(8.2, 13)	-
CPI					
	LR-	0	0.01	(0, 0)	-
	LR+	0	0.01	(0, 0)	-
	MR-	0	0.02	(0, 0)	-
	MR+	0.01	0.05	(0, 0.1)	0
	HR	0.47	0.2	(0.3, 0.8)	0.3

4.5.3 Risk mapping and positioning

For the identification and visualisation of risk conditions and vehicles with unsafe behaviours and involved in risk conditions, the risk levels of each vehicle are plotted, as shown in Figure 4.8.

In the plots, the x-axis denotes the timestamps, in sub-second interval (0.1s), and the y-axis denotes the travelled distance or locations on each lane, in metre-scale. The high-risk points are enhanced for better visualisation. The vehicle trajectory information is integrated with risk grading, hence, target vehicles with risk potentials can be figured out, and the locations and times with higher risk potentials can also be identified.

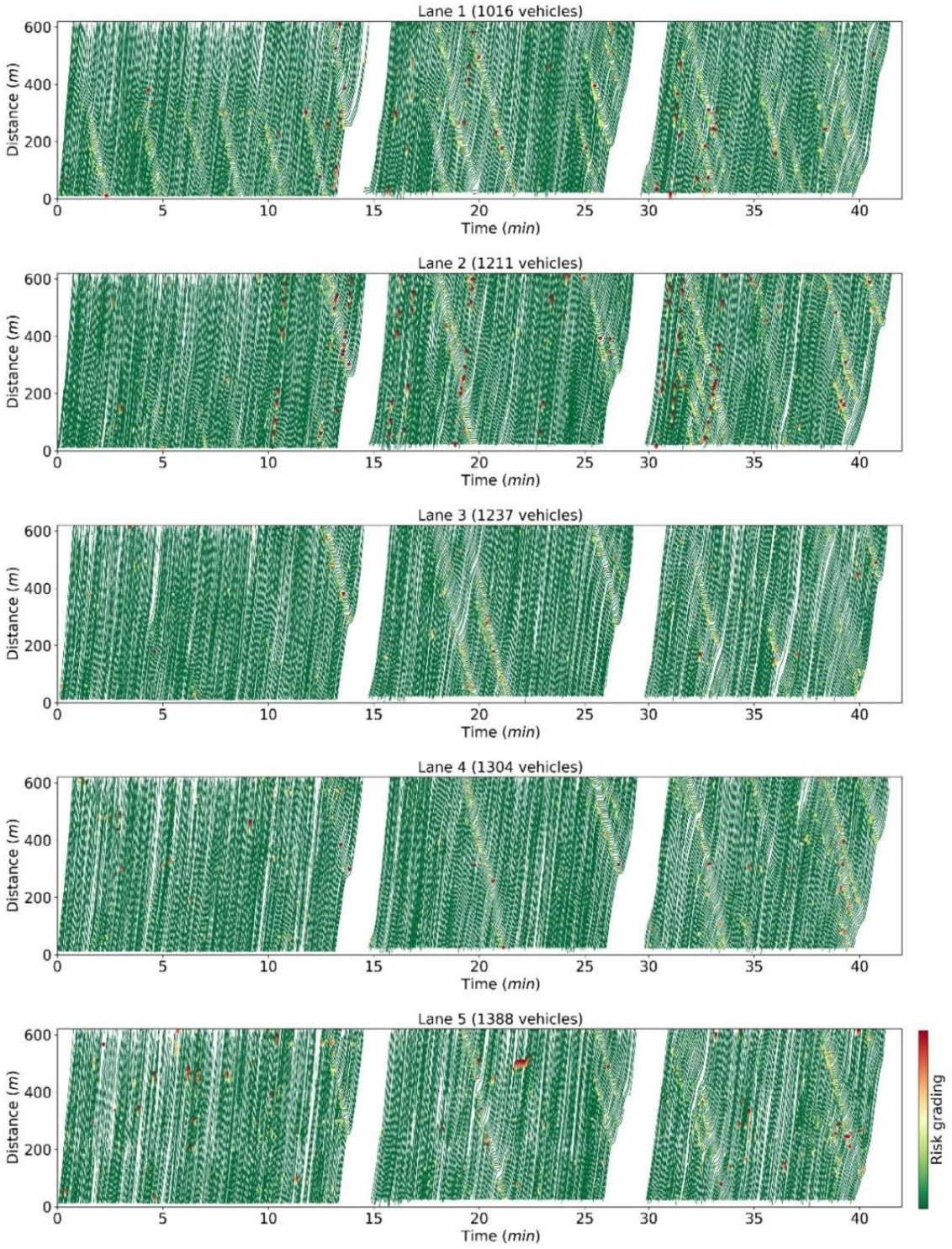


Figure 4.8 Risk mapping and positioning

4.6 Discussion

4.6.1 Application potentials

The development in this chapter results in an unsupervised learning system for risk grading of vehicle driving behaviour, and also provides a data-driven benchmark for risk assessment, with the aim of positioning risk potentials in terms of targeted vehicles, locations in metre-scale, timestamps in sub-second interval. This chapter contributes towards a range of applications, for example:

(1) Predictive crash risk mitigation. A high-resolution risk mapping and positioning is demonstrated in Figure 4.8. The risk patterns (e.g. severity, frequency, trends) offer predictive insights about crash potentials, hence they could be employed as advance signals, and constant monitoring contributes to predictive crash mitigation. The solutions are twofold. One is vehicle-oriented. Since the targeted vehicles with risk exposures are possible to be monitored and identified in advance, proactive intervention could be taken before any prospectively impending crash, such as risk warning by in-vehicle devices or advance driver-assistance systems (ADAS). Another is road-oriented. The locations and time with higher crash likelihoods are possible to be real-time inferred or historically recorded, therefore, a protective strategy can be applied pre-emptively, such as patrol despatch, road enhancement (e.g. remediation of layouts or locations that may lead drivers to more likely indulge in risky driving behaviours). Such dynamic risk grading extends the scope of traditional accident-based solutions, since they reveal additional information on potential crash and risk conditions.

(2) “Pay how you drive”. Measurable risk grading is obtained based on a large group of vehicles within a road segment. The indicator-based risk patterns show the driving behaviours of both safe and risky groups. Hence, such risk grading provides a basis for the design of behaviour-based insurance (also called “pay how you drive”, PHYD), which might be more reasonable than usage-based insurance (UBI, also known as “pay as you drive”, PAYD). Through PHYD, insurance-based incentives can be applied, such as insurance discount for better driving behaviour, which helps to encourage drivers to drive in a safer manner.

Besides, technologically, in view of a large amount of real-world applications which suffer from the challenges of class imbalance problem and unsupervised data labelling (Lin et al., 2017), the findings demonstrate a solution based on feature extraction and clustering. Moreover, the labelling of instances is usually expensive, and some labels are difficult to obtain beforehand. Thus, unsupervised data labelling with interpretable features holds great potential in real-world applications.

4.6.2 Limitations

For risk grading, the major challenge is that the lack of crash instances makes it hard to verify the linkages between the highest risk level and actual crash occurrence. Risk levels are obtained based on the comparison within the samples, and there are no crash cases in the samples. On this aspect, a potential way of validation is to examine the accident records of the drivers/vehicles which are grouped as higher risk levels. Besides, the results are based on data collected from freeways (expressways). Hence, to get a stronger evaluation and validation, an in-depth study using a large data set is recommended, which should cover a wider range of risk conditions, including various types of accidents and near-miss cases collected from expressways, junctions, arterial roads, etc. In addition, the benchmark of risk measures and threshold values can be enhanced using data set collected over a longer period of time. Such work needs a new way of data collection. Furthermore, based on current unsupervised procedure, semi-supervised data fusion can be built, which uses certain labelled instances to refine modelling.

Currently, to grade risk levels, a set of features is considered comprehensively, which have covered risk characteristics in temporal, kinematical and spatial aspects. Of course, these features are far from being exhaustive. Future work should continue to construct new features, such as interpretable features on risk perception, decision-making, time-series characteristics.

Machine learning from highly imbalanced data is very challenging. The ensemble clustering in this chapter is performed based on a general concept of ensemble learning, whereas, there are some ensemble techniques designed for supervised learning, which might be useful, such as bootstrap aggregating, and random subspace

sampling. This chapter has contributed towards the development of unsupervised risk grading, while in-depth research on ensemble clustering for imbalanced data is beyond the scope of the present research study.

4.7. Chapter Summary

In this chapter, a feature-oriented clustering method is proposed to achieve risk grading of a large group of vehicles within a road segment. This method not only contributes to overcoming some intrinsic challenges from imbalanced data, but also produces a trial benchmark for risk mapping and positioning. In summary, this chapter has contributed to the domain knowledge in three areas.

First, for risk grading and imbalanced clustering, indicator-based features are designed. Based on surrogate measures of vehicle conflicts, interpretable features are extracted from a general vehicle driving trajectory, which represents vehicle risk exposures in terms of temporal, kinematical and spatial aspects. Most of the features rely on threshold values to differentiate between risk and safety, which meanwhile contributes towards overcoming the problems of imbalanced data, for example, to reduce the degree of overlapping and lack of density.

Second, clustering-based risk grading is proposed to achieve unsupervised data labelling of risk levels for a large group of vehicles. Based on risk pattern similarity, vehicles are clustered into distinct groups with various risk levels. To obtain reliable and stable outputs, ensemble clustering is built by majority voting of various algorithms, including FCM, K-Means and SOM. Clustering is conducted in a progressive manner to obtain the hierarchical partitioning, which facilitates to identify the highest risk level, and meanwhile addresses the small disjuncts of imbalanced data. Furthermore, label identification by classifiers is proposed to evaluate clustering performance, and guide the selection of the best clustering solution. Besides, key features are identified using feature importance ranking by random forest.

Last but not least, an experimental benchmark for data-driven risk grading is investigated. A high-resolution risk mapping and positioning is demonstrated based

on NGSIM data. The underlying quest is to figure out the risk potentials in terms of targeted vehicles, locations in metre-scale, timestamps in sub-second interval. In addition, the results also generate a better understanding of risk patterns (e.g. severity, frequency, trends), which may offer predictive insights about crash potentials. The findings contribute towards a range of applications, such as predictive crash risk mitigation, “pay how you drive” insurance, among others.

CHAPTER 5
FEATURE LEARNING AND BEHAVIOUR-BASED RISK PREDICTION USING XGBOOST

5.1 Chapter Introduction

This chapter aims to build an approach which is effective and reliable to identify important features for driving assessment, and achieve accurate prediction of risk levels based on driving behaviours. Feature learning is an important and precursory step for machine learning. In this chapter, an integrated feature learning framework is designed, which combines learning-based feature selection, unsupervised risk rating, and imbalanced data resampling. The methodology is introduced in Section 5.2. Section 5.3 elaborates on feature extraction. Massive driving behaviour features are extracted from vehicle movement trajectory, which produces in-depth and multi-view measures on behaviours. Corresponding risk indicator features are also constructed as well. In Section 5.4, the linkages between behaviour features and corresponding risk levels are built using XGBoost (eXtreme Gradient Boosting), and key features are identified according to feature importance ranking and recursive elimination. In addition, data resampling is performed to reduce the risk-safe class imbalance, which contributes to improving learning. Afterwards, risk levels of vehicles in driving are predicted using XGBoost based on the selected key features, which is described in Section 5.5. Model performance evaluation and optimisation directions are also analysed. The final two sections cover the discussion and summary. The modelling is demonstrated using NGSIM dataset.

This chapter includes part of contents in the following papers:

Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C. and Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis and Prevention, 129, 170-179. DOI: 10.1016/j.aap.2019.05.005.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2018). Accident risk prediction based on driving behavior feature learning using CART and XGBoost. Transportation

Research Board 97th Annual Meeting, No. 18-06270.

5.2 Methodology

5.2.1 Feature learning framework

A machine learning framework is designed to select key behaviour features and predict risk levels, which integrates learning-based feature selection, unsupervised risk rating, and imbalanced data resampling. The framework is depicted in Figure 5.1, and the detailed process is illustrated in Algorithm 5.1.

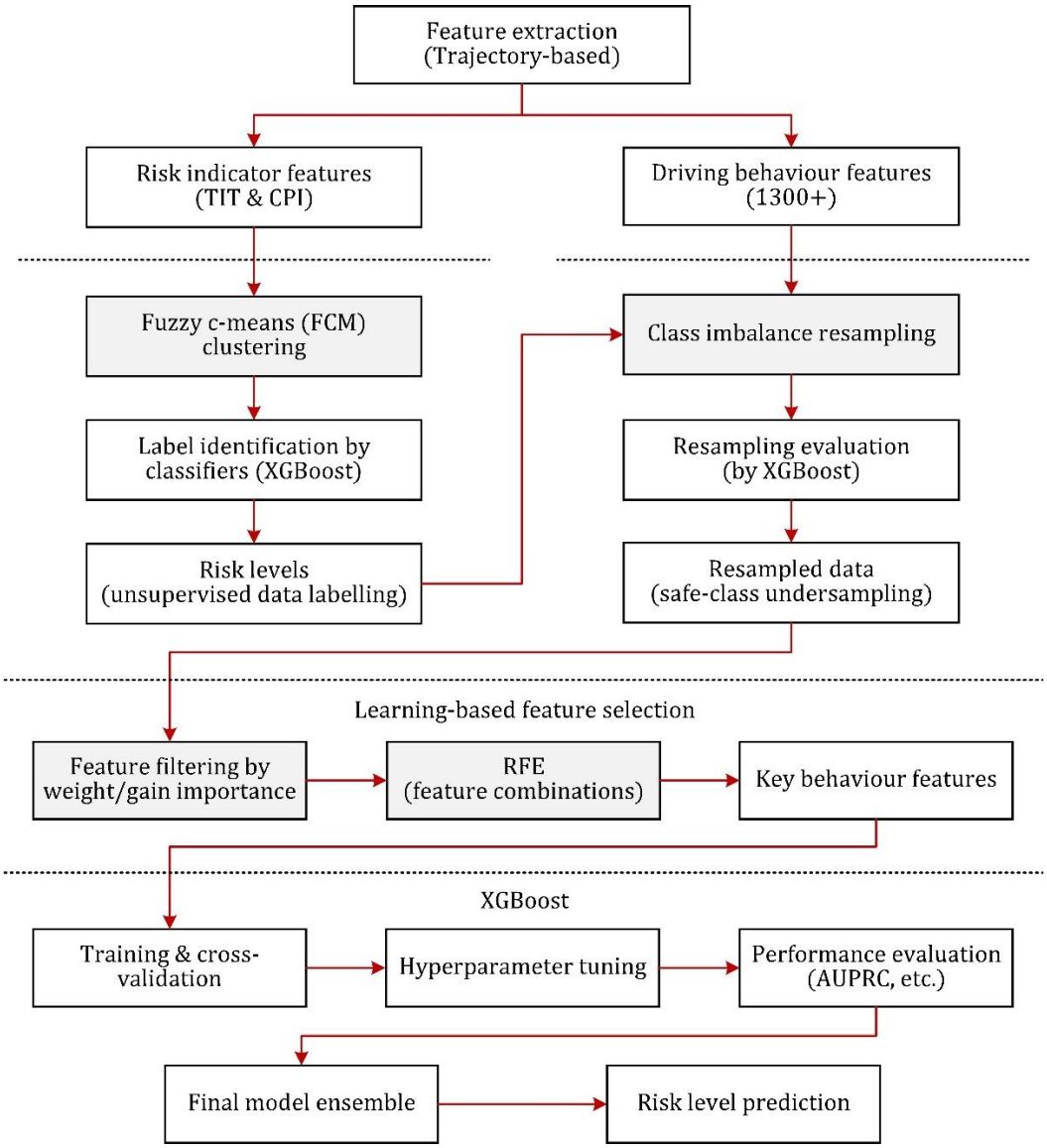


Figure 5.1 Feature learning framework

Algorithm 5.1 Feature learning

1. Feature extraction:
    a. Driving behaviour features of instance i, $x_i^{(b)} = \{f_n(i)\}$;
    b. Risk indicator features, $x_i^{(r)}$, for risk labelling.
2. Risk labelling by clustering:
    a. Select clustering algorithms (e.g. FCM);
    b. Define potential numbers of clusters, K;
    c. Run clustering on $X^r = \{x_i^{(r)}\}$, with various K independently:
        1. Split X^r into K groups;
        2. Produce partition matrix, $i \in C(k), k \in [1, K]$;
        3. Assign the label, $y_i \leftarrow \text{label } C(k)$;
    d. Find the best-suited K based on evaluation by a classifier (i.e. XGBoost);
    e. Risk levels, $Y = \{y_i\}$.
3. Imbalanced data resampling:
    a. Shortlist under-sampling and/or over-sampling strategies;
    b. Apply each strategy on (X^b, Y), where $X^b = \{x_i^{(b)}\}$:
        1. Obtain resampled data $(X^b, Y)'$;
        2. Model (XGBoost) training on $(X^b, Y)'$;
        3. Evaluate model performance;
    c. Select a best-suited resampling strategy.
4. Learning-based feature selection:
    a. Configure hyper-parameters;
    b. Train model (XGBoost) based on $(X^b, Y)'$;
    c. Rank feature relative importance, and remove less important ones;
    d. Recursive feature elimination, and find the best feature subset;
    e. Data set with key features, $(X^b, Y)''$.
5. Risk prediction:
    a. Model training and hyper-parameter tuning based on $(X^b, Y)''$;
    b. Prediction performance evaluation.

5.2.2 Learning-based feature selection

A two-step hybrid method is developed to rank and select key features by machine learning. This procedure firstly filters a set of relative important features based on XGBoost, and then permutes to find an optimal subset from the filtered features using Recursive Feature Elimination (RFE), as illustrated in Algorithm 5.2.

Algorithm 5.2 Feature filtering and RFE

1. Train/tune model (XGBoost) using all features $\{f_n\}$;
2. Feature filtering based on importance ranking:
    a. Calculate relative importance scores, $\{r_n\}$;
    b. Feature filtering by thresholds:
        1. Define thresholds $\{\tau_i\}$;
        2. For each τ_i, do:
            a. Remove features f_n for all $r_n < \tau_i$;
            b. Obtain subset S_i with remaining features;
            c. Re-training with S_i;
            d. Obtain performance A_i;
    c. Select the subset S' with the max $\{A_i\}$.
3. RFE:
    a. N is the feature size of S';
    b. For $n = N, \ldots, 2$, do:
        1. Permute n time, for $k = 1, \ldots, n$:
            a. Remove feature $f_k^{(n)}$, obtain subset $S'^{(n-1)}_k$;
            b. Re-training with $S'^{(n-1)}_k$;
            c. Obtain performance A_k;
        2. For the max $\{A_k\}$, eliminate the feature $f_k^{(n)}$;
        3. Keep $n-1$ important features, obtain subset $S'^{(n-1)}$ and $A^{(n)}$.
    c. Select the feature subset S'' with the max $\{A^{(n)}\}$.

This method combines the advantages of both feature ranking procedures. The permutation importance (mean decrease in accuracy, MDA) is used in RFE, which can find the optimal feature combinations, but the search procedure is computationally intensive, especially for high-dimension feature vectors (Guyon et al., 2002). Tree-based ensemble learning algorithms (e.g. random forest, XGBoost) generate rankings of individual features based on Gini importance (mean decrease in impurity, MDI), which can be integrated to reduce the RFE search space quickly, by shortlisting a set of relative important features. The detailed description is provided in Section 5.2.3.

The optimal feature subset can be selected based on the trade-off between learning performance and model simplicity (i.e. fewer features). Permutation importance is measured by the mean difference of accuracy before and after randomly permuting a feature (Fahad et al., 2014). For MDA, a considerable decrease in accuracy indicates that the feature is highly relevant and useful, contributing to learning improvement, and vice-versa. Whereas an irrelevant feature only carries minimal impact, and a redundant feature also has limited contribution due to the high correlation with other more important ones. Therefore, redundant and irrelevant features could be removed, without loss of accuracy. Benefits of feature selection include better interpretability, simplified modelling, shorter learning time, and enhanced generalisation, among others (Guyon and Elisseeff, 2003; García et al., 2016).

5.2.3 XGBoost

XGBoost is short for eXtreme Gradient Boosting, proposed by Chen and Guestrin (2016). XGBoost is an optimised gradient tree boosting system. The system is designed with parallel learning, cache-aware access, blocks for out-of-core computation, which help to improve efficiency. For algorithmic innovations, an approximate greedy search algorithm for split finding and weighted quantile sketch are introduced, which are faster than gradient boosting. Besides, various hyper-parameters are used to improve learning and control over-fitting.

As an ensemble learning based on boosted trees, XGBoost also provides tree-based feature importance ranking. The feature relative importance can be measured by

several metrics, such as split weight, average gain, etc. Weight is the number of times that a feature is used to split the data across all boosted trees. More important features are used more frequently in building the boosted trees, and the rests are used to improve on the residuals. Instead of counting splits, gain measures the actual decrease in node impurity, which is the total decrease in node impurity (Gini index) weighted by the number of samples it splits, and averaged over all trees, namely, the average gain across all splits the feature is used in. After model training, the linkages between behaviour features and corresponding risk levels are built, and the results of weight-based and gain-based importance are considered.

The detailed description is provided in Chen and Guestrin (2016). As a powerful classifier, in this chapter, the XGBoost serves as the key algorithm in the processes of clustering evaluation, resampling evaluation, feature selection, and prediction. More information about XGBoost is described in Appendix C.

5.2.4 Unsupervised risk rating

Since the ground-truth labels about risk levels are generally not available, clustering-based risk grading is used to generate the labels of risk levels. Crash risk potentials of vehicles in driving are estimated by clustering of a massive number of vehicles based on risk indicator features. The present modelling will select a best-suited clustering algorithm or combination based on clustering performance, to provide a reliable labelling of risk levels. Label identification by classifiers is used for evaluation. Detailed information about unsupervised risk rating and label identification by classifiers are analysed in Chapter 4.

In this chapter, to evaluate and determine clustering results, label identification by classifiers is conducted using XGBoost. As an independent classifier, XGBoost is used to investigate the degree of correct identification of the clustered labels, and determine the main hyper-parameter, namely, the number of clusters (K). The basic assumption is that a better clustering solution could produce labels with higher cross-validation accuracy. In addition, a reasonable clustering can be selected based on misclassification in conjunction with the resolution, as analysed in Chapter 4. Besides, a similar method and modelling process are applied in this chapter, but the

data is pre-processed in a slightly different way, and the clustering algorithm and risk indicator features are further improved, which may result in a difference in clustering outcomes.

5.2.5 Imbalanced data resampling

After data labelling, adaptive data resampling is integrated, to reduce class imbalance problems. A more balanced class distribution could be produced by undersampling of the majority class and/or oversampling of the minority class. Since useful data might be eliminated, the undersampling is only performed on the safe class. Minority oversampling creates artificial instances by interpolation, which needs careful performance evaluation. Herein, the effects on learning performance of various resampling strategies are compared in order to find the best solution.

The technique adopted for undersampling is Edited Nearest Neighbours (ENN). ENN makes the decision to keep or remove a given sample based on a nearest-neighbour algorithm (Wilson, 1972). ENN can keep the main data structure, and generally is better than random undersampling. Based on different execution of the ENN algorithm, there are Repeated ENN (RENN) and All K ENN (AKNN). RENN executes ENN multiple times to reach the desired number of samples, and AKNN runs ENN with a range of values of neighbour number, flags the misclassified instances and removes them at the end (Tomek, 1976).

For oversampling, SMOTE (synthetic minority oversampling technique) is adopted. SMOTE generates new samples by interpolating new points between two instances (Chawla et al., 2002). To reduce the possibility of interpolation with noisy data, hybrid methods are proposed. SMOTE with ENN (SMOTE+ENN) firstly uses ENN to reduce noisy samples next to class boundaries, then applies SMOTE on the well-separated space (Batista et al., 2004). Similarly, SMOTE with support vector machine (SVM-SMOTE) firstly classifies the data by SVM, and makes interpolation according to class belongingness (Nguyen et al., 2009).

5.2.6 Performance evaluation metrics

The prediction performance of a classification algorithm (or classifier) can be described based on a confusion matrix, which reports the numbers of True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN), as depicted in Table 5.1. Based on the confusion matrix, a set of metrics is derived to evaluate the classification results and prediction performance, including some integrated metrics (e.g. AUPRC, AUC), as depicted in Table 5.2.

Table 5.1 Classification confusion matrix

		Predicted value	
		Positive [P]	Negative [N]
Actual value	Positive [P]	True Positive (TP)	False Negative (FN)
	Negative [N]	False Positive (FP)	True Negative (TN)

Table 5.2 Metrics for classification evaluation

Metrics	Description
Recall; TPrate; Sensitivity	TP/(TP+FN)
Precision; positive predictive value (PPV)	TP/(TP+FP)
Misclassification (MC)	FP+FN
Accuracy	(TP+TN)/(TP+FP+TN+FN)
FPrate; False alarm rate	FP/(FP+TN)
(ROC) AUC	$\frac{1 + TP_{rate} - FP_{rate}}{2}$
AUPRC	Area Under the precision–recall (PR) Curve
F1 score	2TP/(2TP+FP+FN)

For imbalanced classification, this chapter uses AUPRC value as the main metric, while others are also included as supplementary criteria, including misclassification, precision, recall, and (ROC) AUC. As multi-class supervised learning, the metrics are calculated for each class, and two types of mean values are produced. One is

macro-averaged over all classes, which is unweighted mean. Another is the average weighted by the number of true instances for each class (Lever et al., 2016). Model performance is evaluated by various metrics via stratified cross-validation. More introduction about these metrics are described in Section 4.3.3.

5.3 Feature Extraction

5.3.1 Feature extraction from trajectory

Features are derived values from raw data, and used as input to a machine learning algorithm. High-quality features (e.g. being informative, relevant, non-redundant, interpretable, etc.) improve the understanding of algorithms, which is important for problem-solving (Guyon and Elisseeff, 2003). Since learning processes rely on the exact information delivered into the algorithms, features are the keys to generate reliable and convincing results.

Two kinds of features are discussed in this chapter, namely, driving behaviour features and risk indicator features. Driving behaviour features are extracted to produce in-depth and multi-view measures on behaviours (e.g. movement characteristics), which not only focus on individual vehicles and vehicle pairs, but also consider relative performance with respect to vehicle platoon and traffic conditions. Risk indicator features are used in labelling risk levels, which are expected to distinguish between risk and safety. In a supervised learning framework, driving behaviour features act as the input, and risk indicator features serve for the target.

Vehicle movement trajectory data are used for feature extraction. The trajectory is an overall reflection of the driver behaviour, including risk perception and response, driving performance and style (Chai and Wong, 2015b). Plenty of spatial-temporal information can be mined from vehicle trajectory. Moreover, trajectory data is non-interference in nature and is flexible to collect externally, as compared with in-vehicle recording data.

The movement trajectory $M_i(t_{a\to b})$ of a forward-moving vehicle i from time t_a

to t_b is defined as a series of position $P_i(t)$ with coordinates $(x_i(t), y_i(t))$.

$$M_i(t_{a\to b}) = [P_i(t_a), P_i(t_{a+1}), \cdots, P_i(t_{b-1}), P_i(t_b)]$$

The raw data R_i and feature vector F_i of the vehicle i with observable attribute statement S_i (e.g. vehicle type, traffic flow conditions, weather, traffic signals) are represented as follows:

$$R_i = [T \quad M_i \quad S_i] = \begin{bmatrix} t_a & x_i(t_a) & y_i(t_a) & s_i(t_a) \\ t_{a+1} & x_i(t_{a+1}) & y_i(t_{a+1}) & s_i(t_{a+1}) \\ \vdots & \vdots & \vdots & \vdots \\ t_b & x_i(t_b) & y_i(t_b) & s_i(t_b) \end{bmatrix} \ldots$$

$$F_i = \left(f_i^{\alpha}(\cdot), f_i^{\beta}(\cdot), f_i^{\gamma}(\cdot), \ldots\right)$$

where, $f(\cdot)$ is an extraction process that measures the information and characteristics in certain aspects. Instead of the large raw data, the feature vector F_i is obtained to represent the information used for machine learning.

The process of features extraction is depicted in Figure 5.2. The processes of deriving and constructing features consist of four steps, elaborated as follows.

Step 1. Extraction of relevant variables from raw data.

A series of variables involving driving behaviour are derived from raw trajectory data, including velocity, acceleration, lateral position, the preceding and following vehicles, etc. Variables related to vehicle pairs, vehicle platoon and traffic conditions are also computed, such as, front gap, average velocity of vehicle stream within a lane segment for a given time window.

Step 2. Feature construction by pre-defined functions.

Some functions are designed to mine in-depth information and build multi-variable relationships. Surrogate measures of traffic conflicts are used to build risk indicator features, such as Time-to-Collision (TTC). Variables are explored from multi-view (e.g. relative change) and multi-scale (i.e. by small time windows). Relationships of variables are built, such as, comparison with proceeding vehicles or vehicle platoon,

similarity match, correlation coefficient, etc.

Step 3. Summarise and select the key information.

Some operations are defined to summarise the key information and profile the data series, including statistical descriptions, threshold-based filtering, aggregated or accumulated values, etc. Descriptive statistics are widely used to profile the data series, including values of minimum, maximum, mean, standard deviation, and percentiles (e.g. the 25th percentile, or the Q1 values).

Finally, some informative and interpretable features are preliminarily shortlisted, and learning-based feature selection is developed to select the most important ones (as described in Section 5.4).

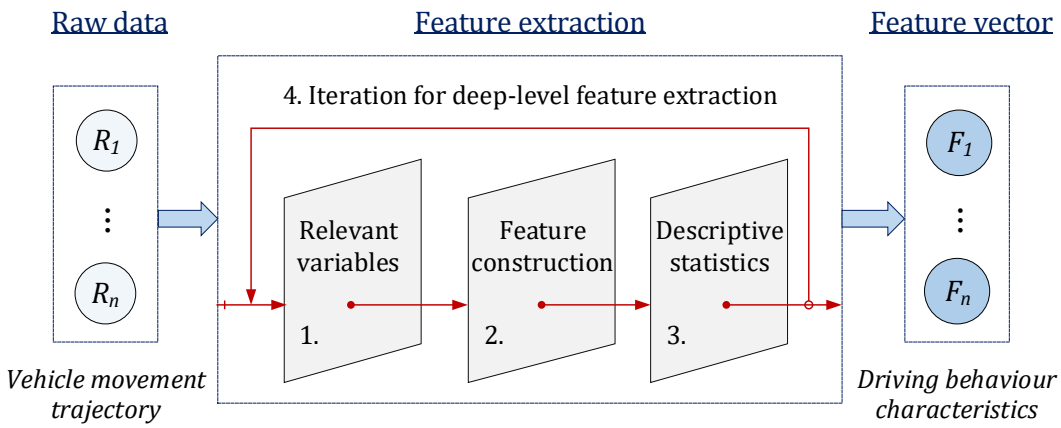


Figure 5.2 Feature extraction process

5.3.2 Driving behaviour features

Based on vehicle movement trajectory data, a total of 1,328 features are preliminarily extracted to describe various aspects of driving behaviour characteristics. The variables, functions, operations to extract and construct driving behaviour features are listed in Tables 5.3, 5.4 and 5.5, respectively. The extracted driving behaviour features are listed in Table 5.6.

Table 5.3 Variables for feature extraction

	Code	Description	No.
Collected data			1-7
Vehicle trajectory	x; y	Time series data, for vehicle i, $x_i(t)$ is longitudinal position defined by vehicle front centre; $y_i(t)$ is the lateral position	1; 2
Trajectory of the preceding vehicles	pv.x; pv.y	$x_{i-1}(t)$; $y_{i-1}(t)$, for preceding vehicles (pv) $i-1$	3; 4
Lane	lane	Lane number of a vehicle travelling on	5
Vehicle type	class	0-motocycle; 1-car; 2-truck	6
Vehicle length	L	L_{i-1}, vehicle length of preceding vehicle	7
Derived variables			8-14
Velocity	vel	$v_i(t) = d[x_i(t)]/dt$	8
Acceleration	acc	$a_i(t) = d[v_i(t)]/dt$, negative value indicates deceleration	9
Jerking	jerk	$j_i(t) = d[a_i(t)]/dt$	10
Front gap	gap	Calculated by $x_{i-1}(t) - x_i(t) - L_{i-1}$	11
Variables of preceding vehicles	pv.vel; pv.acc; pv.jerk	$v_{i-1}(t)$; $a_{i-1}(t)$; $j_{i-1}(t)$	12-14

Table 5.4 Functions for feature extraction

	Code	Description	No.
Functions			15-30
Moving time windows	w1; w2; w3	Moving windows defined by time intervals of 1.0 s (w1), 5.0 s (w2), and 10 s (w3), return time series data	15-17
Difference	dif	$s_{i-1}(t) - s_i(t)$, measure difference of variable s between subject vehicle and preceding vehicles	18

Table 5.4 (continued) Functions for feature extraction

	Code	Description	No.
Percentage change	pct	$\frac{s_i(t)-s_i(t-1)}{s_i(t-1)} * 100$, percentage change of a variable (per 1.0 second)	19
Log ratio	logr	Measure relative change on a logarithmic scale, per 1.0 second, calculated by $\log \frac{s_i(t)}{s_i(t-1)}$	20
Vehicle to flow ratio	vfr	Comparison between a vehicle and the average performance of vehicle platoon in the same lane segment	21
Range	rng	Calculated by $\max(s_i(t)) - \min(s_i(t))$	22
Coefficient of range	crng	Calculated by $\frac{\max(s_i(t))-\min(s_i(t))}{\max(s_i(t))+\min(s_i(t))}$	23
Simple moving average	sma	$MA_i(t)$, mean value of the time series data within a moving window defined by $(t-w,\ t)$	24
Moving standard deviation	msd	Standard deviation of the time series data within a moving window	25
Relative standard deviation	rsd	Relative variability and unitised measure, defined as the ratio of std to mean	26
Bias ratio	emar	Calculated by $\frac{s_i(t)}{s_i(t)-EMA_i(t)}$, where $EMA_i(t)$ is the exponential moving average of the data within a moving window defined by $(t-w,\ t)$	27
Dynamic time warping	dtw	Using DTW algorithm to measure similarity between two temporal sequences	28
Correlation coefficient	scor; pcor	Compute pairwise correlation of two variables, by Spearman correlation (scor), and Pearson correlation (pcor). Besides, TTC-vel relationship (tv), and TTC-pv.vel relationship (tpv) are calculated to refer the responsibility in a conflict condition	29;30

Table 5.5 Operations for feature extraction

	Code	Description	No.
Operations			31-45
Basic centre and dispersion	mean; std	Values of mean and standard deviation	31;32
Extreme values	min; max; p01; p99	Values of minimum, maximum; also consider using 1^{th} and 99^{th} percentiles values (p01, p99) to deal with outliers and noise	33-36
Percentile values	p05; q1; q2; q3; p95	The 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles to represent data profile and distribution pattern	37-41
Mean absolute deviation	mad	Measure variability or dispersion	42
Profile shape	krt; skw	Unbiased kurtosis (krt) over data using Fisher's definition; unbiased skew (skw), normalised by n-1	43;44
Absolute mean	absm	Mean of absolute value	45

Table 5.6 Driving behaviour features

Feature	Code and counts	Total
Basic characteristics of subject vehicle	{vel; acc; gap}.{kurt; mad; max; mean; min; p01; p05; p95; p99; q1; q2; q3; skew; std}(42); {jerk}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(11) {clas}(1)	54
Relative comparison with respect to proceeding vehicles or vehicle platoon	{acc; vel}.{dif}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(22); {acc; gap; vel; jerk}.{vfr}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(44); {acc; vel}.{dif}.{vfr}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(22)	88

Table 5.6 (continued) Driving behaviour features

Feature	Code and counts	Total
Lane keeping and changing	{y}.{std}(1); {lane}.{mean; std; rng}(3)	4
Basic characteristics of preceding vehicles	pv.{vel; acc}.{kurt; mad; max; mean; min; p01; p05; p95; p99; q1; q2; q3; skew; std}(28); pv.jerk.{kurt; mad; max; mean; min; p01; p05; p95; p99; q1; q2; q3; skew; std}(11); pv.{clas}.{mean; truck; motorcycle}(3)	42
Relative change measured by percentage ratio and log ratio	{vel; acc; gap; vel.dif; acc.dif; pv.vel; pv.acc; y; pv.y}.{pct}.{absm; max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(108); {y; vel; gap; pv.y; pv.vel}.{logr}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(55)	163
Microscale behaviour defined by small moving windows	{vel; pv.vel; acc; pv.acc; gap; vel.dif; acc.dif}.{w1; w2; w3}.{rng; crng; sma; msd; rsd; emar}.{mean; std; max; min}(504); {acc; vel}.{dif}.{vfr}.{w1; w2; w3}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(66); {acc; gap; vel; jerk}.{vfr}.{w1; w2; w3}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(132)	702
Behaviour similarity match by DTW	{vel; acc; x}.{dtw}(3); {vel; acc; x}.{w1; w2; w3}.{dtw}.{mean; std; p05; q1; q3; p95; max; min}(72)	75
Behaviour correlations	{vel; acc; tv; tpv}.{pcor; scor}(8); {vel; acc; tv; tpv}.{w1; w2; w3}.{pcor; scor}.{mean; std; p05; q1; q2; p95; max; min}(192)	200
∑		1,328

Feature fusion and in-depth mining can enhance reliability and predictability. To capture safety-critical conditions, various behaviours are assessed, such as braking response, speed matching, gap maintenance. For specific behaviours, multiple measures and functions are designed. For example, the instantaneous changes of movement are measured by several ratio-based functions, including percentage ratio, log ratio and bias ratio, since different forms of ratio have different sensitivities to changes. The intensity of abrupt braking can be reflected from high magnitude jerk, the high percentage change of velocity, when decelerating, etc. The lane keeping features (e.g. y.std) are measured based on vehicle lateral trajectory data that do not involve lane-changing. From predictable perspectives, compared with extreme values, percentile values provide a kind of early signals, which are more helpful in early diagnostics of risk conditions. Besides, microscale behaviours can be measured based on features defined by small moving time windows. The proposed feature extraction procedure is scalable for such kind of trajectory time series data.

5.3.3 Risk indicator features

Chapter 3 has examined the feasibility of using risk indicators to assess pre-accident risk conditions, and hybrid indicators based on TIT and CPI are suggested to measure risks. In Chapter 4, risk indicator features are built for clustering, and important features for risk assessment are identified. According to previous chapters, TIT and CPI are more suitable to build risk indicator features, and label risk levels. The definition and calculation of TIT and CPI are described in Sections 3.3.2 and 3.2.3, respectively. Extended information about the extraction and selection of risk indicator features are described in Sections 4.3.2 and 4.5.1, respectively.

In this chapter, five risk indicator features are built based on TIT and CPI, with different settings of threshold values and threshold measurement. Three TIT-based features are built, namely, TIT.t1 (TTC threshold = 2s), TIT.t2 (TTC threshold = 3s) and TIT.t3 (TTC threshold = 4s). Two CPI-based features are developed, namely, CPI.m1 (MADR1) and CPI.m2 (MADR2), based on two MADR measures adopted in CPI.

5.4 Learning-based Feature Selection

5.4.1 Data preprocessing

Same as Chapter 4, the NGSIM vehicle trajectory data is also used in this chapter as a case study. The description of NSGIM data is located in Section 4.4.1. More description and analysis of NGSIM data are illustrated in Appendix B. The features are calculated using the complete data set available, which was collected from a 640-metre road segment for about 45 minutes. Besides, some unvalued records are removed because of missing values. After data cleaning, a total of 5,084 instances (vehicles) is used for feature modelling, involving 3,203,867 records.

In addition, Savitzky-Golay filter is used in data preprocessing to smooth out potential noise (e.g. unphysical fluctuation) and errors in data acquisition. Savitzky-Golay filter approximates a given signal using a sliding window and a low degree polynomial to model data within that window, and also incorporates the introduced error in the approximation process using linear least squares. This filter reduces the disadvantage of cutting off peaks. For the vehicle driving data, velocity and acceleration are the 1^{st} derivative and 2^{nd} derivative of trajectory, and Savitzky filter is performed at each derivative operation based on 1.0s filter window, using 1^{st} and 2^{nd} order of the polynomial to fit the samples, respectively.

5.4.2 Labelling of risk levels using FCM

As described in Chapter 4, the clustering performance of the shortlisted clustering algorithms (i.e. FCM, K-means, SOM) is tested on preliminary experiments. K-means typically produces different outcomes due to random initial conditions. SOM is hard to converge, and the hyper-parameter tuning and cluster reassignment involve subjective judgement. In comparison, FCM is best-suited in this problem, which presents a better overall performance, especially providing stable and reasonable clustering results. Besides, Fuzzy theories are interpretable, and widely used in behaviour analysis for traffic safety (Chai and Wong, 2015b). Herein, FCM is best-suited in this problem.

Given the imbalanced problems and lack of ground truth labels, several values of the number of clusters (K) are considered in the FCM clustering, ranging from 4 to 6. The clustering results and evaluation are presented in Figure 5.3 and Table 5.7. Label identification by XGBoost provides an evaluation of the clustering results, using models built with various numbers of boosted trees to represent both weak and strong classifiers, as shown in Figure 5.3(a).

Table 5.7 Clustering results and label identification by XGBoost

K	Clustered groups	Accuracy	F1-score[#]	AUC[#]	AUPRC[#]
4	(3921; 943; 212; 8)*	0.994	0.962	0.993	0.942
5	(3653; 900; 425; 98; 8)	0.993	0.970	0.999	0.974
6	(3445; 812; 525; 224; 70; 8)	0.992	0.966	0.977	0.936

* Number of instances in each clustered group, from safe to higher risk levels.

[#] Macro-averaged over all classes.

From Table 5.7, the clustering with 5 groups shows a better trade-off between performance (i.e. higher AUPRC) and resolution (i.e. more groups). A detailed comparison is presented in Table 5.7, based on an XGBoost ensembled with 20 boosted trees (a moderate-level classifier). Label identification by XGBoost demonstrates an accuracy of 99.3%, which can imply that an underlying structure close to ground truth labels is promising to be discovered by both clustering and classifier.

Four risk levels are obtained from the clustering of 5 groups. The reason is that, for the highest risk level, the number of instances is limited (only 8 vehicles), hence they are combined with the sub-highest class. One annotation of the risk labelling is the safe level (group 1; with 3,653 instances), low risk level (group 2; LR; with 900 instances), moderate risk level (group 3; MR; with 425 instances), high risk level (combination of groups 4 and 5; HR; with 106 instances). The safe level indicates a near-zero risk, which has the lowest likelihood to involve in traffic conflicts. A scatter plot of the clustered risk levels is illustrated in Figure 5.3(b). TIT contributes to distinguishing the range of each risk level, and CPI further figures out the instances with a high likelihood to involve an accident.

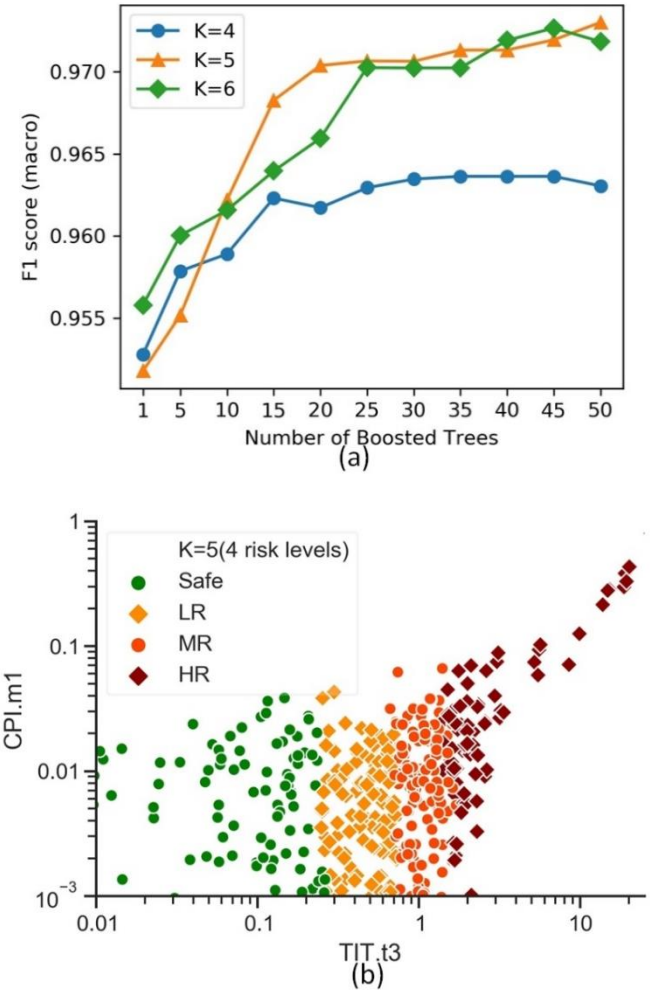


Figure 5.3 Clustering evaluation and risk grouping

5.4.3 Data resampling

To reduce the impact of class imbalance, six resampling strategies are investigated, and the resampled datasets are also generated, as listed in Table 5.8. Under-samplings by Repeated ENN (RENN) and All K ENN (AKNN) reduce the safe-class instances to an amount close to the LR group. Four strategies of combining both under-sampling and over-sampling (i.e. from R3 to R6) are also examined. Over-samplings create instances in the HR group, reaching an amount close to the MR group. Two SMOTE-based hybrid oversampling techniques are applied, which are SMOTE with ENN cleaning (SMOTE+ENN) and SMOTE with classification by support vector machine (SVM-SMOTE).

Table 5.8 Resampling strategies and resampled datasets

No.	Under-sampling	Over-sampling	Counts of resampled data*
R1	ERNN	-	2642 (1211; 900; 425; 106)
R2	AKNN	-	2587 (1156; 900; 425; 106)
R3	AKNN	SMOTE+ENN	2783 (1033; 900; 425; 425)
R4	AKNN	SVM-SMOTE	2906 (1156; 900; 425; 425)
R5	RENN	SMOTE+ENN	2817 (1067; 900; 425; 425)
R6	RENN	SVM-SMOTE	2961 (1211; 900; 425; 425)

* Counts of resampled data for each risk level (from low to high).

The effects of resampling strategies on learning performance are compared in Figure 5.4. From the evaluations for each risk level, under-sampling can improve the learning performance (i.e. AUPRC, precision and recall) of the LR class, and RENN is slightly better than AKNN in this experiment. There is no evidence to show that oversampling the HR class could improve the learning performance of classes of MR and/or LR. In Figure 5.4 (a), the initial dataset and four datasets processed by over-sampling have higher accuracy values, the reasons being that more instances are in the safe class and oversampled HR class, which are also more homogeneous and less overlapping, hence easier to classify. Besides, the obvious improvements of the HR class are based on interpolated data, to which careful attention should be paid. Herein, safe-class under-sampling by RENN is selected to reduce the class imbalance, for better modelling.

The following modelling is based on the under-sampled dataset. The safe class data is under-sampled using RENN, in which 1,211 instances are selected from initial 3,653 safe-class instances, and the total data size drops from 5,084 to 2,642.

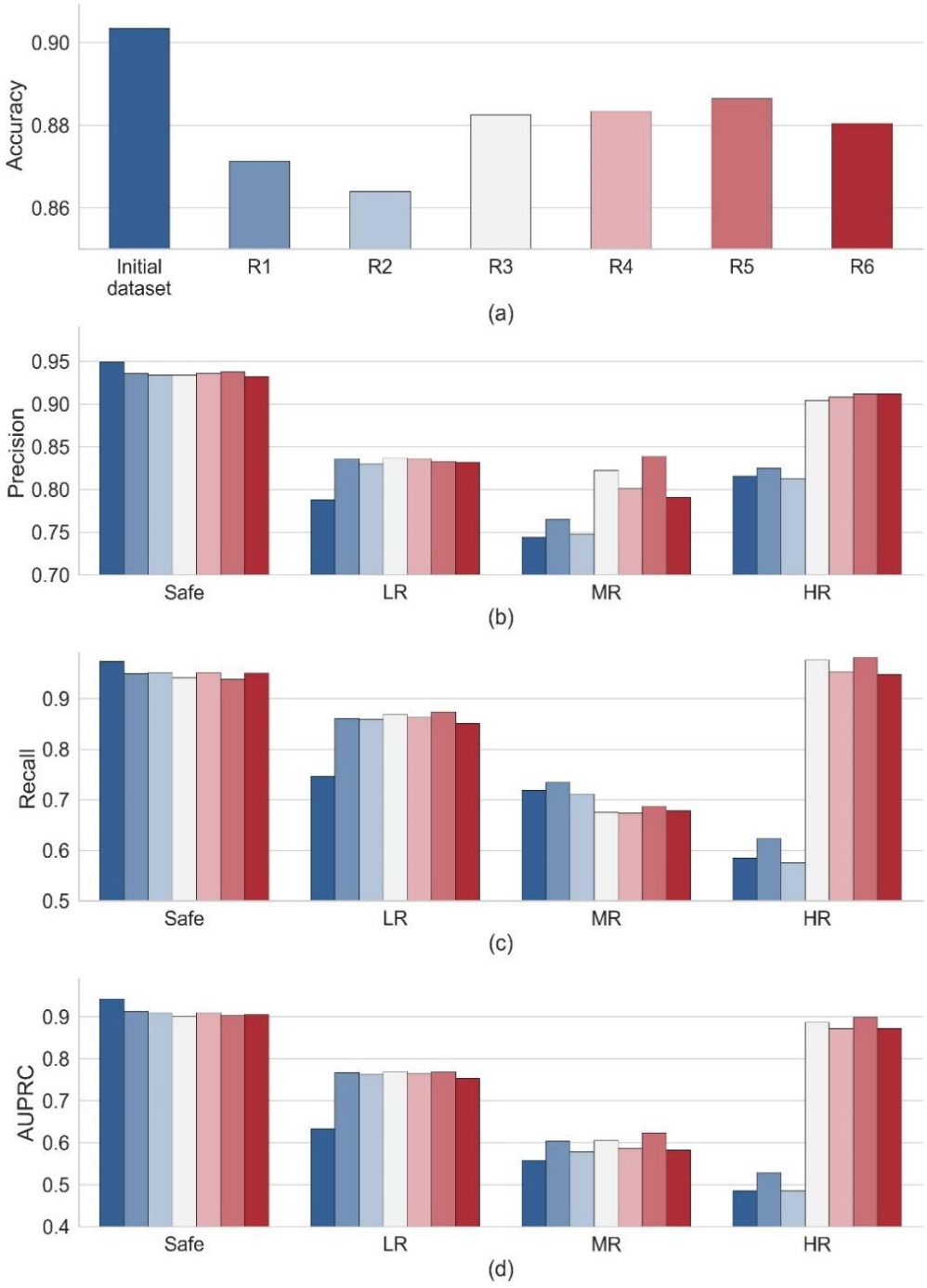


Figure 5.4 Performance comparison of resampling strategies

5.4.4 Feature importance ranking

The linkages between behaviour features and corresponding risk levels are built using XGBoost. The training of an XGBoost model is based on the under-sampled dataset, and tested via 10-fold stratified cross-validation. Hyper-parameters are configured to

build an appropriate XGBoost model for feature ranking. A similar process of hyper-parameter tuning is described in Section 5.5.2.

After model training, an XGBoost model with high accuracy is obtained, and the split weight and average gain for each feature are generated, which are normalised to calculate the weight-based and gain-based relative importance scores, respectively. The scores measure the usefulness of a feature in building the boosted trees in XGBoost. Higher scores indicate greater relative importance. The feature filtering by weight-based and gain-based importance ranking are demonstrated in Figures 5.5(a) and 5.5(b), respectively.

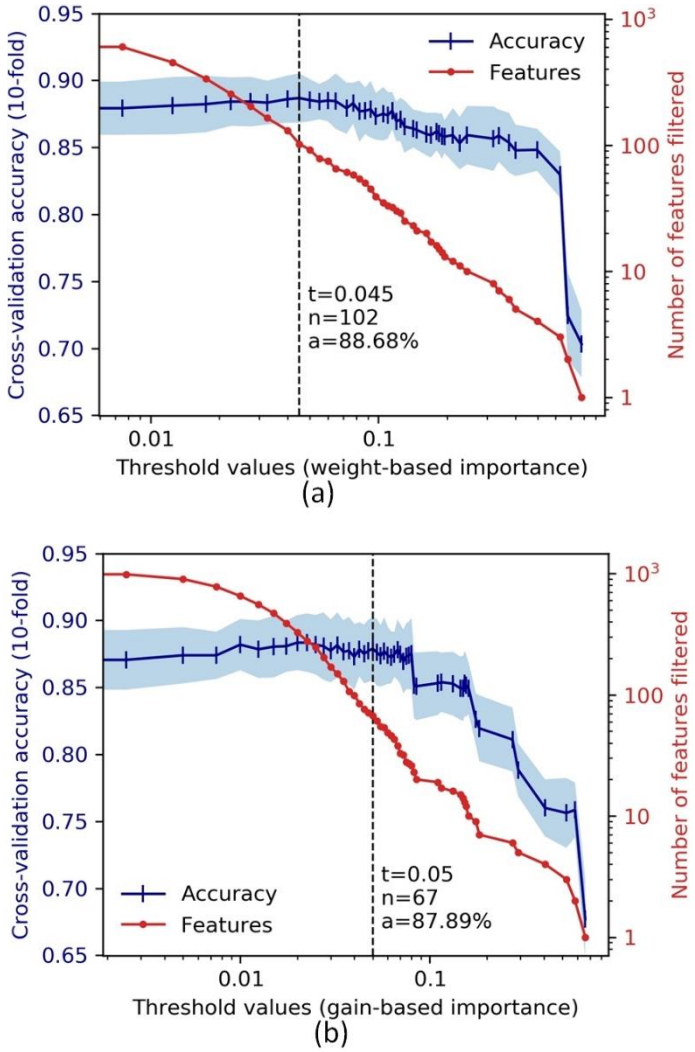


Figure 5.5 Feature filtering by relative importance ranking

A range of score thresholds is defined for a quick feature filtering. In each iteration, the features with the importance scores greater than the threshold value are kept, and the learning performance is estimated via 10-fold stratified cross-validation.

From Figure 5.5(a), a total of 148 promising features are filtered, including 102 features selected according to weight-based importance ranking, and 67 features with higher Gini importance, while noting that some features are duplicated in the two filtered feature subsets. Weight-based selection generally favours the features with more classes, and gain-based selection is biased towards the ones with stronger signals (e.g. impurity).

5.4.5 Feature recursive elimination

In the RFE process, an optimal feature combination is selected from the filtered features, by model re-training and recursively pruning the feature with the least permutation importance from the current set. The learning performance of each iteration is shown in Figure 5.6. RFE starts with all filtered features and ends with the most important one. The top 5 iterations ranked by mean accuracy are listed in Table 5.9.

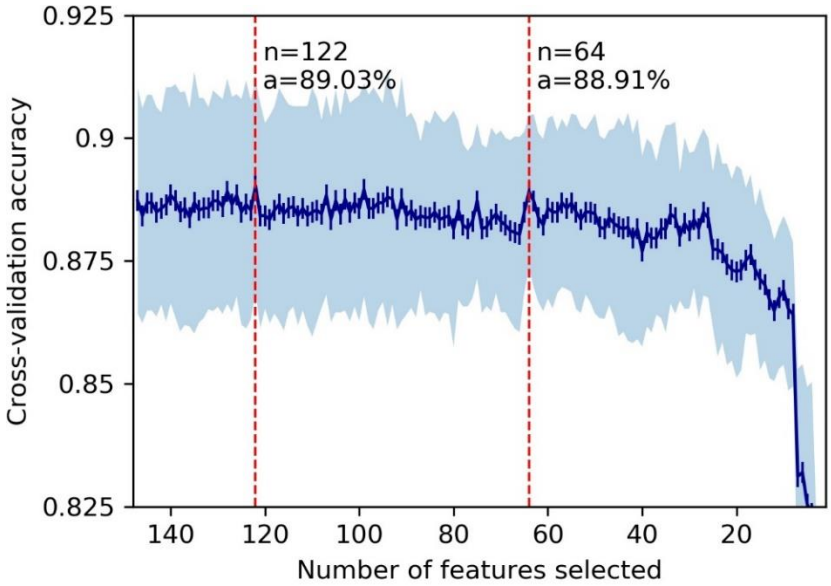


Figure 5.6 Feature selection by RFE

Table 5.9 Top 5 iterations ranked by mean accuracy

	Iteration	Number of features	Mean accuracy	Std. accuracy
1	25	122	0.8903	0.0215
2	48	99	0.8895	0.0216
3	**83**	**64**	**0.8891**	**0.0155**
4	19	128	0.8891	0.0229
5	7	140	0.8884	0.0219

The combination with the best mean accuracy has 122 features. However, the performance at the subset with 64 features reaches an accuracy of 88.91%, but lower variance. Considering the trade-off of complexity and performance, the combination with 64 features is suggested, which has less complexity and a modest decrease in estimated accuracy from 89.03% down to 88.91%. The selected key features are listed in Table 5.10.

From a practical perspective, the identification of key features guides the procedures of data collection, mining and understanding. From Table 5.10, the gap, velocity and acceleration are the most informative variables involving risk assessment. Data capture and processing at shorter time intervals are also important, given that 38 key features are defined based on small moving windows. For data mining, 17 ratio-based features are selected (i.e. 7 log ratio, 6 percentage change, and 4 bias ratio), which can measure the abnormal changes of movement, and capture potential risk signals. More than half of the features involve behaviour comparison and relative performance, which indicates that dynamic benchmarking of the overall performance of vehicle stream is helpful to assess the behaviour of individual vehicles pertinently. Moreover, the fusion of multi-view and in-depth features makes the system more robust and fault-tolerant, and also has high transparency.

Table 5.10 Key features selected

Variable	Features	Counts
Gap	gap.vfr.w3.min; gap.pct.min; gap.pct.q1; gap.p01; gap.logr.p05; gap.vfr.w3.p05; gap.pct.p01; gap.logr.std; gap.min; gap.w1.msd.max; gap.vfr.w3.p01; gap.pct.p05; gap.w1.crng.std; gap.logr.absm; gap.vfr.min; gap.logr.p01; gap.vfr.w2.p95; gap.w1.emar.min; gap.vfr.w2.min; gap.w1.sma.min; gap.w1.crng.max; gap.w2.emar.min	22
Acceleration	acc.dif.w2.sma.mean; acc.dif.vfr.w3.p01; acc.w3.crng.std; acc.dif.w2.emar.std; acc.dif.vfr.w2.q2; acc.dif.w1.sma.max; acc.dif.vfr.w2.p01; acc.vfr.w3.mean; acc.w1.emar.max; acc.dif.vfr.w1.max; acc.dif.w2.rsd.min; acc.dif.mean; acc.w2.sma.mean; acc.dif.vfr.w2.q3	14
Velocity	vel.dif.p99; vel.dif.w3.sma.min; vel.dif.p95; vel.dif.max; vel.dif.w1.sma.mean; vel.w2.scor.std; vel.logr.p05; vel.w1.msd.std; vel.dif.w1.sma.max; vel.dif.p01; vel.dif.w1.rng.mean; vel.dif.mean; vel.dif.w1.msd.max	13
Preceding vehicles	pv.vel.logr.p05; tpv.pcor; pv.vel.logr.p99; tpv.w1.pcor.q3; pv.acc.p05; tpv.w2.pcor.max; pv.acc.w2.rsd.max; pv.vel.pct.std; tpv.w1.pcor.p95; tpv.w2.scor.p95; tpv.w2.pcor.p95	11
Jerk	jerk.vfr.w3.std; jerk.vfr.min	2
Lateral position	y.std; y.pct.p99	2
Total		64

Besides, in the RFE stages, the feature elimination is based on MDA importance. Another feature importance measure is MDI, as shown in Figure 5.7. The results of MDA and MDI are compared, including the selected features, the difference in accuracy, feature correlations, etc.

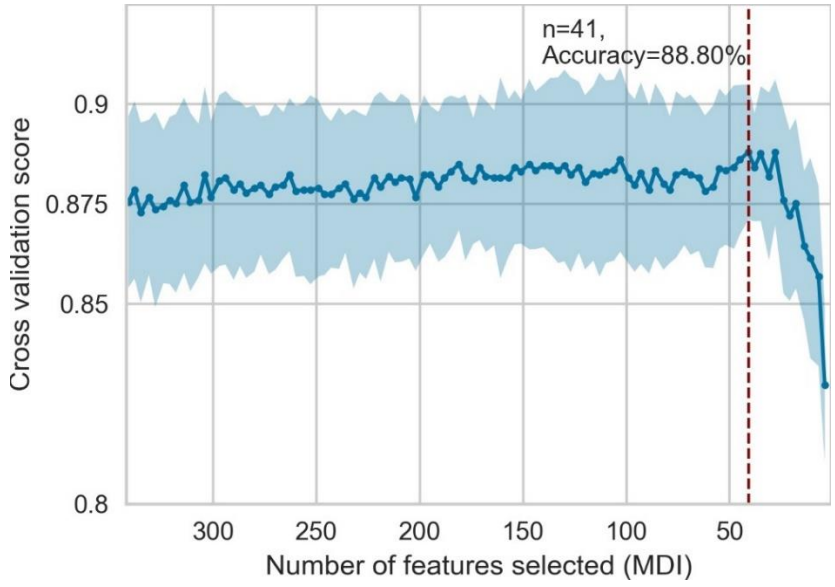


Figure 5.7 Feature selection by RFE based on MDI

As representations of driving behaviour information, key features are able to be used to perform tasks such as driving assessment and risk identification. Combinations of key features can produce insights about driving performance (short-term assessment) and driving style (long-term assessment), such as concentrated and skilled driving (e.g. appropriate gap maintenance, smooth velocity change and well-controlled lane keeping), aggressive and/or distracted driving (e.g. huge gap change, over speeding, excessive braking, great variable difference of the subject vehicle and its surrounding vehicles), etc. In addition, risk-averse or risk-taking behaviours are also able to be inferred from monitoring the driving behaviour and long-term feature analysis. The combinations of key features can be used for driving assessment.

5.5 Crash Risk Prediction based on Key Behaviours

5.5.1 Risk prediction using XGBoost

In order to infer the likelihood of a crash in advance, the antecedent risk conditions and corresponding risk levels should be predicted firstly. In this way, there is a leading time and early signal to achieve crash prediction, and prevention strategies could thereby be conducted before the risk conditions turning into an actual accident event. The risk-based crash inference and prediction is described in Chapter 6. In this

chapter, the focus is more about the prediction of risk levels based on behaviour features. The risk levels of vehicles in driving are predicted based on the supervised learning of key behaviour features and corresponding risk levels. The selected key behaviour features are applied as early signals for risk identification and prediction. The flowchart of the final model fitting and prediction performance estimation is illustrated in Figure 5.8.

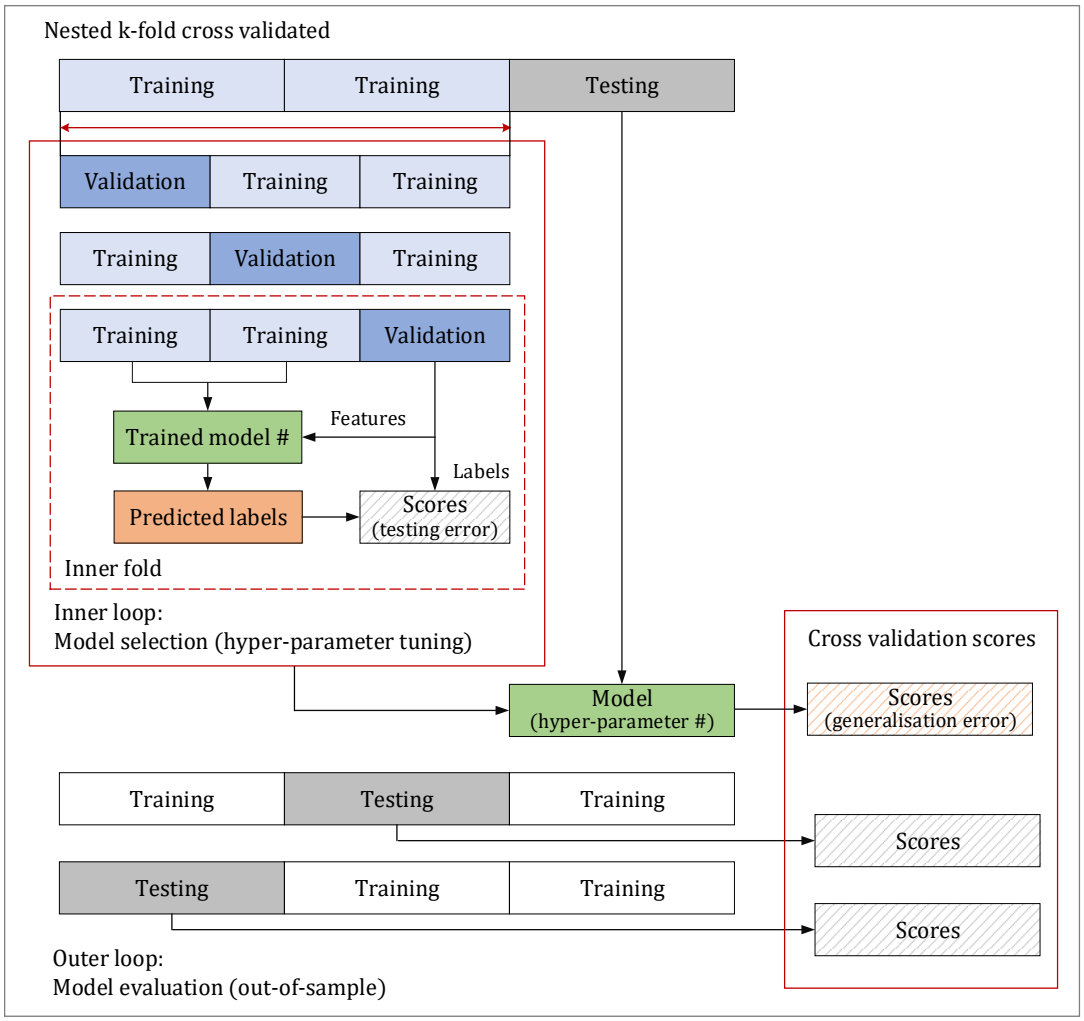


Figure 5.8 Final model fitting and performance estimation

This process is essentially a classification task. The entire dataset is split into separate training datasets and testing datasets, and nested k-fold cross-validation is applied to produce an unbiased evaluation (out-of-sample). The inner loop is used for model selection and hyper-parameter configuration. A trained XGBoost model is obtained based on training input (i.e. key behaviour features) and training target (i.e.

corresponding risk levels). The predicted risk levels of vehicles in the validation datasets are obtained using the trained model, and compared with original labelled risk levels. The model with the best performance is selected. In the outer loop, the selected model is evaluated based on data, which are not used in model building and selection. The predictive power is measured based on the predefined metrics, such as misclassification, AUPRC, etc. If the predictive power is accurate and reliable, the final prediction model is built using the complete data, based on the hyper-parameters selected in the trained XGBoost model. The final model is able to be used to predict the risk levels with new vehicle feature data.

5.5.2 Hyper-parameter tuning

Hyper-parameters tuning is to get a trade-off between bias and variance. XGBoost is a kind of ensemble learning configured with boosted trees. In XGBoost, boosted trees are constructed sequentially, where each new tree is created to correct for the prediction errors made by the sequence of existing trees. The effect is that the model can quickly fit on the training dataset (i.e. less bias), but overfitting occurs (e.g. greater variance in prediction). The learning rate helps to shrink the boosting process by weighting, which makes fitting more conservative. This in turn results in more trees and iterations are added to achieve similar results. For individual boosted trees, tree hyper-parameters can directly control model complexity, such as maximum tree depth, splitting weight, etc. Besides, random subsampling of instances and features also help to decorrelate and improve the model robustness against noise, hence reducing the variance.

A range of hyper-parameters is tuned using Grid Search for model optimisation, as shown in Table 5.11 and Figure 5.9. The cross-entropy loss (log loss) is used for the performance evaluation, which provides a more nuanced view of the model performance (e.g. as compared with accuracy). Logistic loss is defined by the negative log-likelihood of the true labels given a probabilistic prediction, which takes into account the uncertainty of predicted results.

Figure 5.9 shows the model performance of each Grid Search, measured by log loss via 10-fold stratified cross-validation, with the standard deviation values shown as

error bars.

Table 5.11 Key hyper-parameters and tuned values

Hyper-parameters	Description	Tuned value	Accuracy (%)
1. Ensemble hyper-parameters			
Learning rate	Shrink the feature weights of each boosting step	0.1	88.61
Number of estimators	Number of boosted trees added in model	150	88.61
2. Boosted tree hyper-parameters			
Tree depth	Maximum depth of a tree	5	88.72
Splitting weight	Further partitioning of a leaf node in tree building process	1	88.72
3. Subsampling hyper-parameters			
Instance subsample ratio	Random sample of the training data prior to growing trees in every boosting iteration	0.7	88.99
Feature subsample ratio	Random sample of features for each split in tree level	0.8	88.99
4. Regularisation hyper-parameters			
Gamma	Minimum split loss reduction required to make a further partition on a leaf node	0	88.99
Alpha	L1 regularisation term on weights	0	88.99
Lambda	L2 regularisation term on weights	1	88.99
5. Tuning update			89.01

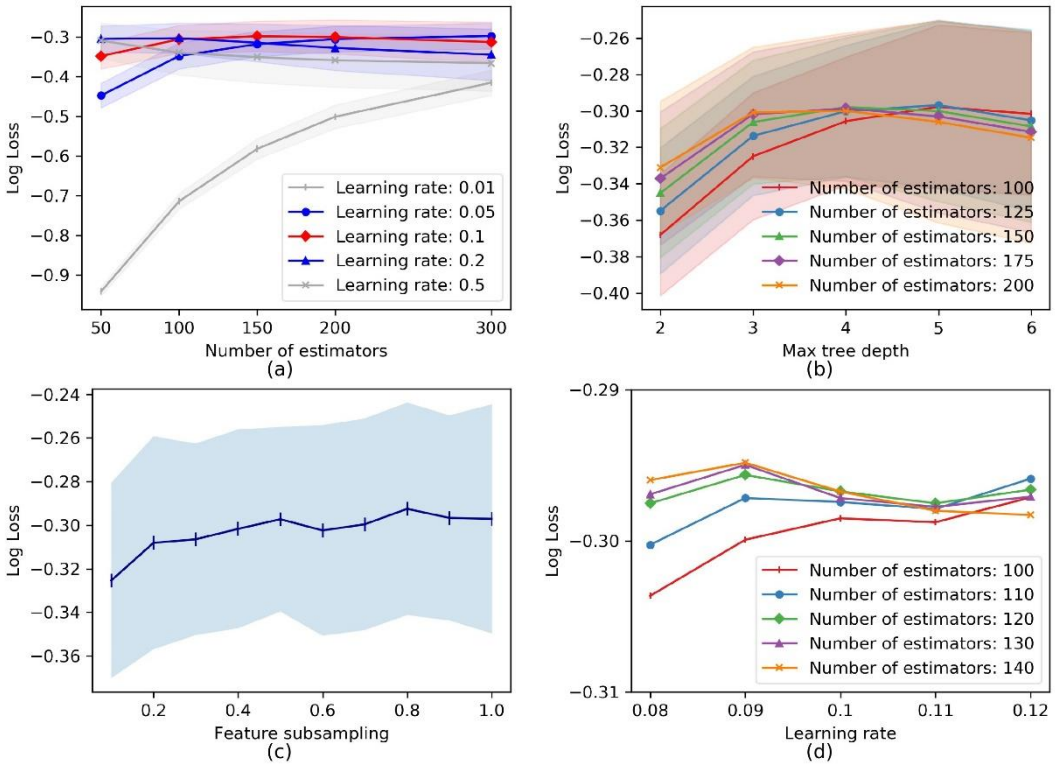


Figure 5.9 Grid Search for hyperparameter tuning

Generally, hyper-parameters values with the best mean and relatively low standard deviation are selected; if several values produce similar performance, the tuned value is selected when a plateau in performance or a point of diminishing returns is observed. The tuned values and accuracy are listed in Table 5.11.

5.5.3 Early stopping

Early stopping of the training procedure at an optimal epoch can control overfitting. The optimal epochs could be reached once the testing performance has not improved after a fixed number of training iterations, for example, over a window of 10% of the total epochs. An inflexion point, where the testing performance starts to decrease while the training performance continues to improve, may indicate overfitting being encountered. The learning curves on separate training and testing datasets over epochs are plotted in Figure 5.10. The inflexion points are observed at around 120 epochs, and the epoch for early stopping is selected.

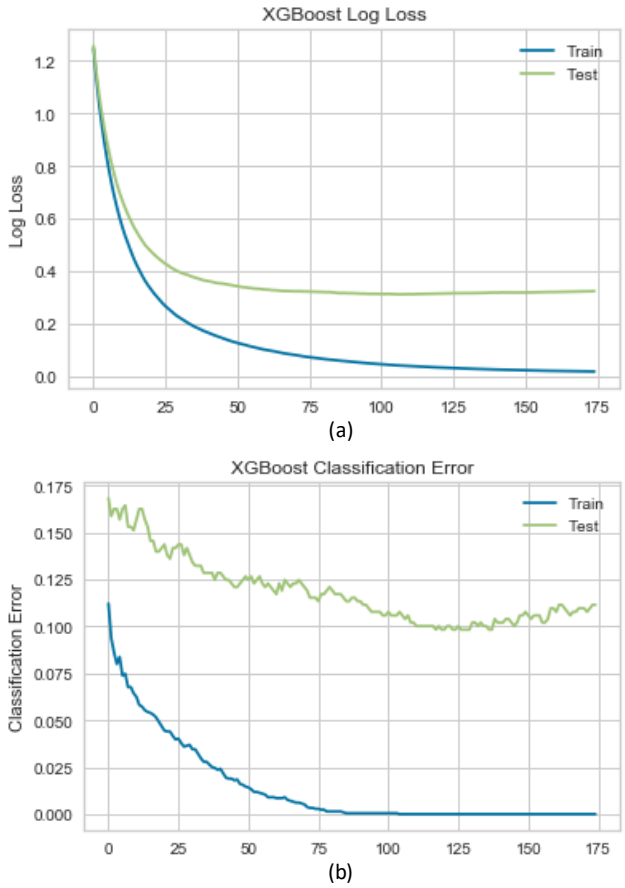


Figure 5.10 Learning curves

5.5.4 Results and evaluation

The risk levels of vehicles in driving are predicted based on key behaviour features. The resampled dataset only keeps the selected key features. The dataset is split into two complementary partitions of training and validation, in a stratified fashion. The splitting ratio is set according to the learning performance, measured by cross-validation accuracy, as illustrated in Table 5.12.

Table 5.12 Splitting ratio and cross-validation accuracy

	Training ratio	Testing ratio	Mean accuracy	Std. accuracy
10-Fold	90%	10%	89.17	01.95
5-Fold	80%	20%	88.08	01.01
3-Fold	67%	33%	87.24	00.76

In Table 5.12, the cross-validation accuracy under different splitting ratio is compared, and this study uses a proportion of 80% for training, and remaining 20% for validation, based on balancing between bias (i.e. higher mean accuracy) and variance (i.e. lower standard deviation of accuracy).

With the tuned hyper-parameters, the final prediction model is obtained by re-training using the complete data. The predictive power (generalisation ability) of the final model can be estimated using the cross-validation performance. The results are the predicted probabilities of a vehicle belonging to each risk level, and the risk level with the highest probability is determined as the predicted label. Herein, a satisfactory prediction with an accuracy of 88.91% is achieved based on XGBoost and key behaviour features. The converted overall accuracy is about 91.66%, which is converted based on the raw data distribution before safe-class undersampling. Detailed estimation of the prediction performance is shown in Table 5.13, where the support is the number of instances in each class.

Table 5.13 Prediction evaluation

Group	Precision	Recall	AUC	AUPRC
Safe	0.952	0.946	0.990	0.990
LR	0.855	0.890	0.966	0.931
MR	0.798	0.781	0.966	0.835
HR	0.824	0.660	0.989	0.844
Macro	0.857	0.819	0.978	0.900
Weighted	0.889	0.889	0.983	0.939

Predicting detailed risk levels is more challenging but valuable. Great performance by isolated validation indicates an accurate and reliable predictive power, whilst allowing the model to work well on unseen data. The prediction of risk levels can be used as early signals for crash potentials and likelihood, which allows measures to be taken to reduce crash proneness, as well as make a crash prediction ahead of time (i.e. in the risk stages). Behaviour-based risk prediction is complementary to indicator-based risk assessment, since the fusion of more feature information helps to make the

system more robust and reliable.

5.5.5 Performance comparison

In a previous study (Shi et al., 2018b), about 50 features were selected using CART, and risk levels were predicted using XGBoost based on 306 vehicles (NGSIM data, from timestamp 7:49:40 AM to timestamp7:52:25 AM), with an accuracy rate of 74%.

The improvement of prediction accuracy relies on advanced feature extraction and selection, and optimisation of hyper-parameter tuning of XGBoost. In this study, XGBoost-based risk prediction achieves an accuracy rate of 90%, which is relatively better than existing similar studies.

5.6 Discussion

To mine more information about driving behaviours, more than a thousand features have been considered comprehensively. The feature extraction is far from being exhaustive. In-depth feature extraction is recommended to further improve modelling, which should cover a broader range of driving behaviours and risk conditions, such as lane-changing, conflicts between motorcycles and vehicles. The interests of feature extraction are mainly twofold, namely, making risk assessment more reliable, and providing early signals for risk-based crash prediction.

Due to the extreme requirement in risk analysis, unsupervised feature learning could be used to obtain deep-level features, which may bring in more information than existing hand-crafted features. The advantage of unsupervised learning is that plentiful information can be extracted without manual intervention, such as clustered factors, deep-level structure of features, complex spatial and temporal relationship, etc. However, the decoding of the learned features is a challenging work, some of which may be difficult to be represented by prior knowledge (Wang et al., 2017). Future works about feature learning are concentrated on using unsupervised and semi-supervised learning algorithms, which learn and decode the representations from time-series data, instead of being hand-crafted. Extended discussion on

clustering-based risk grading is provided in Section 4.6.2.

5.7 Chapter Summary

In this chapter, an integrated machine learning approach is designed to assess driving behaviours and predict risk levels, which combines supervised feature selection and unsupervised risk labelling. This chapter contributes to the safety domain and associated literature in four areas.

First, the extraction of massive features from vehicle trajectory. To mine information about driving behaviours, for individual vehicles, more than a thousand features are extracted comprehensively, which produce in-depth and multi-view measures on behaviours. The feature extraction procedure is scalable for trajectory time series data.

Second, an integrated framework for feature learning and risk prediction. The framework combines learning-based feature selection, clustering-based risk rating and data labelling, and imbalanced data resampling. Four risk levels are obtained by FCM clustering to assess risk potentials of vehicles in driving. Besides, under-sampling of the safe-class data is conducted to amend the biased results derived from class imbalance.

Third, key feature selection by importance ranking and recursive elimination. The linkages between behaviour features and corresponding risk levels are built using XGBoost. The weigh-based and gain-based relative importance are ranked, which produces a filtered feature space. Next, RFE is performed to select an optimal feature subset, and 64 key behaviour features are identified.

Fourth, satisfactory results of behaviour-based risk prediction by XGBoost. The risk levels of vehicles in driving are predicted based on the key features, and predictive power with an overall accuracy of 91.66% is achieved. The approach is effective and reliable to identify important features for risk assessment, and contributes to guiding the direction of feature engineering, which is the key to improve modelling.

CHAPTER 6
AUTOMATED MACHINE LEARNING FOR RISK PREDICTION AND POTENTIAL APPLICATIONS

6.1 Chapter Introduction

This chapter leverages on the findings and techniques developed in the preceding chapters, and develops end-to-end AutoML methods of driving risk prediction. The systems and proof-of-principle prototype demonstration for downstream potential applications are discussed, including risk decision-making for autonomous vehicles (AVs), pay-how-you-drive (PHYD) insurance, driving safety system under the connected vehicle (CV) environment, and short-term crash prediction. Section 6.2 elaborates the main structure and key components of the AutoML system. In Section 6.3, the AutoML for risk decision-making and motion trajectory planning in AVs is discussed, from feature extraction under various sensor configurations, to the prediction performance and data-driven insights for AV designs. Section 6.4 illustrates the risk exposure centric insurance premium (i.e. PHYD), based on the driving risk evaluation using the proposed AutoML. In Section 6.5, a driving safety and behaviour assessment system under the CV environment based on the proposed AutoML is designed, and the roadmaps for the implementation of different system integrations are also sketched. As an extension of the behaviour-based risk prediction, the framework of crash prediction based on detailed risk levels is described in Section 6.6. Ending section is the chapter summary.

This chapter includes part of the contents in the following manuscripts:

Shi X., Wong Y.D., Li M.Z.F., and Chai C. (2019). An automated machine learning (AutoML) method of risk prediction for the decision-making of autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems.

Shi X., Wong Y.D., Li M.Z.F., and Chai C. (2019). Driving risk prediction and data-driven design for pay-how-you-drive (PHYD) insurance: An automated machine learning (AutoML) approach. Transportation Research Part A: Policy and Practice.

6.2 Automated Machine Learning (AutoML)

6.2.1 Domain-specific AutoML for risk prediction

To build an interpretable and successful machine learning system for risk prediction is inherently challenging. Achieving a state-of-the-art performance depends not only on the fundamental power of the core algorithms, but also on careful data processing, right feature engineering, optimised hyperparameter tuning, and well-configured end-to-end pipeline (Feurer et al., 2015; Eggensperger et al., 2019). Decision-tree based ensemble learning algorithms have demonstrated the advantages in interpretability and performance, such as XGBoost, LightGBM.

Moreover, there are greater technical and practical challenges to build machine learning systems used for risk prediction of vehicles in driving. Firstly, it is not straightforward to obtain the ground-truth data labels about risk levels, which is the basic to guide behaviour-risk learning. Feature engineering is an important aspect to improve interpretability and predictive ability, which is the process of using domain knowledge to extract and select implicit information from sensing data, as the input of learning algorithms (García et al., 2016). The performance of a given model also leverages on hyper-parameter tuning, which usually requires manually trying out the best one from all possible values (Feurer et al., 2015).

Risk prediction for real-world applications faces complex scenes and changeable factors, which vary in time and space, thereby efficient machine learning pipelines must incorporate auto-tuning and self-learning procedures (Shahriari et al., 2015). Besides, considering the time and computation constraints, it is more practical to have a lean and fast-response machine learning system that can achieve high performance using small sample data. Certainly, such domain-specific automated machine learning (AutoML) for driving risk prediction enables a lot of benefits but relevant research is still much lacking.

6.2.2 AutoML framework

An AutoML framework is designed to achieve self-optimised modelling of driving

risk prediction and behaviour assessment. The AutoML assembles necessary modelling steps as an end-to-end machine learning pipeline and automates the pipeline to get the features, algorithms, and hyperparameters that return the best performance as measured on validation datasets, which is a process of learning to learn. There are three main components of the AutoML pipeline, including clustering-based risk identification (as developed in Chapter 4), risk-behaviour feature learning (as developed in Chapter 5), model selection and hyperparameter tuning by Bayesian optimisation. Bayesian optimisation guides the self-learning process to find the best-suited algorithms and hyperparameter values within specific computation capacity and time constraints. Besides, the pipeline also incorporates pre-processing components such as imbalanced data resampling, noise filtering, among others. These components are sequentially implemented and cross-validated to build an optimised model. The framework enables an end-to-end AutoML that maps vehicle driving data to risk levels.

The AutoML process is scalable for other classification algorithms, such as random forests, LightGBM, etc. Among the pipeline steps, hyperparameter tuning is the toughest process, which usually requires trying out all the possible values of each hyperparameter, and the optimal values of some hyperparameters depend on settings of others, hence the process entails huge combinations.

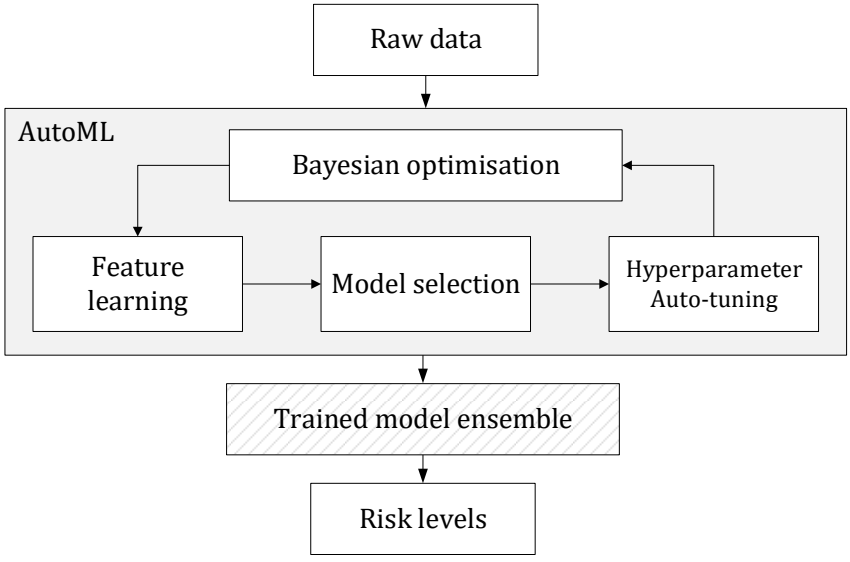


Figure 6.1 AutoML for risk prediction and behaviour assessment

6.2.3 Hyperparameter auto-tuning

Machine learning is optimised by finding the best algorithm configuration and hyperparameter setting that maximise the model performance on validation sets. The process of mapping from the hyperparameters (and algorithm selection) to the performance is known as a black-box function, which is tedious and expensive to optimise. Bayesian optimisation is more efficient in such context, which constructs a probabilistic (surrogate) model of the objective function that maps input values to a probability of a loss, making it easier to optimise than the actual objective function (Shahriari et al., 2015). Besides, by reasoning from the past search results, the next trials can concentrate on more promising values, which reduces the number of trials while finding a good optimum.

There are several methods to form the surrogate function, such as Gaussian Processes, Tree-structured Parzen Estimator (TPE) (Snoek et al., 2012). This chapter adopts the TPE to build the surrogate model (i.e. response surface) of the objective function and guide the exploration of the domain space, according to the demonstrated efficiency and pre-tested performance (Bergstra et al., 2011). The Bayesian hyperparameter optimisation is represented as:

$$x^* = \arg\min_{x \in \chi} f(x)$$

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x) dy$$

where $f(x)$ represents the loss of the objective function to minimise, χ is the domain space of hyperparameter values to search over. y^* is a threshold value of the objective function, y is the actual value of the objective function using hyperparameters x. A probability surrogate model $p(y|x)$ is built based on TPE, which produces a predictive posterior distribution over the performance of the past evaluations. The hyperparameter values to evaluate in the next trials are selected based on the expected improvement (EI). The values with the highest EI are expected to potentially maximise the increase in the performance.

6.3 AutoML for AV Risk Decision-Making

6.3.1 AV decision-making mechanisms

Autonomous vehicles (AVs) have the potentials to dramatically reduce vehicle collisions associated with driver errors and negligence. However, driving scenarios and risk conditions are hugely diverse and complex, and the biggest challenge facing AVs is how to operationalise the required high level of safety and reliability in driving. With maturity in the development of sensor perception and vehicle control, safe decision-making and risk assessment mechanisms are becoming increasingly important towards promoting AV development.

The decision-making mechanisms and technical solutions of self-driving are complex. Briefly, typical models include rule-based systems, end-to-end deep learning, and their hybrids. Rule-based systems (e.g., finite state machine) pre-define thousands of behaviour rules based on expert knowledge, which have certain interpretability and transparency on behaviour-risk reasoning (Paden et al., 2016). To consider all possible driving scenes, the rules are incredibly complex, which essentially limits the prospects of further improvement. End-to-end deep learning is data-driven and scalable, in which the actions (e.g. steering, speed) are directly learned from sensed scenes (e.g. lidar point cloud, camera video) (Hecker et al., 2018). Based on well-tuned models and high-quality big data, deep learning is expected to achieve high performance and be adopted widely, such as LSTM (long short-term memory RNN), and CNN (convolutional neural networks) (Xu et al., 2017). But such scene-to-action learning is more like the intuitions of human drivers, and is less straightforward to inspect and control quantitively (e.g., when assessing risk levels). Besides, motion planning is also related to the future behaviour intentions and trajectories of surrounding vehicles (e.g., within next 8s in Apollo), which encounters high behaviour uncertainty (Altché and de La Fortelle, 2017). Herein, interpretable learning from behaviours to risk levels can benefit existing systems, which adds decisive advantages, such as causal reasoning of behaviour-to-risk, modular division of sensing and decision-making, etc. It is important that behaviour-risk decision-making can be inspected, such as the responsibilities and faults-

checking of individual components (McAllister et al., 2017). Reliable risk prediction can enhance the confidence level of decision-making in uncertainty, and early identification of potential risk conditions can empower AVs with advanced intelligence.

6.3.2 Behaviour-to-risk AutoML for AVs

Safety is one of the core objectives of AV decision-making, and risk prediction helps to determine an optimal path or trajectory to be executed in local motion planning (Katrakazas et al., 2015). Figure 6.2 demonstrates the roles of AutoML in the AV decision-making hierarchy. Given a sequence of route segments planned, the behavioural decision-making layer is responsible for generating dynamically feasible sets of movement based on perceived driving scenes (e.g., relative position, assessment of nearby vehicles). Due to uncertainty, making an optimal choice of next-step movements is not easy, which needs to satisfy all the constraints and criteria, such as being collision-free, efficient and comfortable. Herein, the domain-specific AutoML can offer behaviour assessment and predict the risk levels of instantaneous next-step movements, which helps to decide a better local motion to be executed.

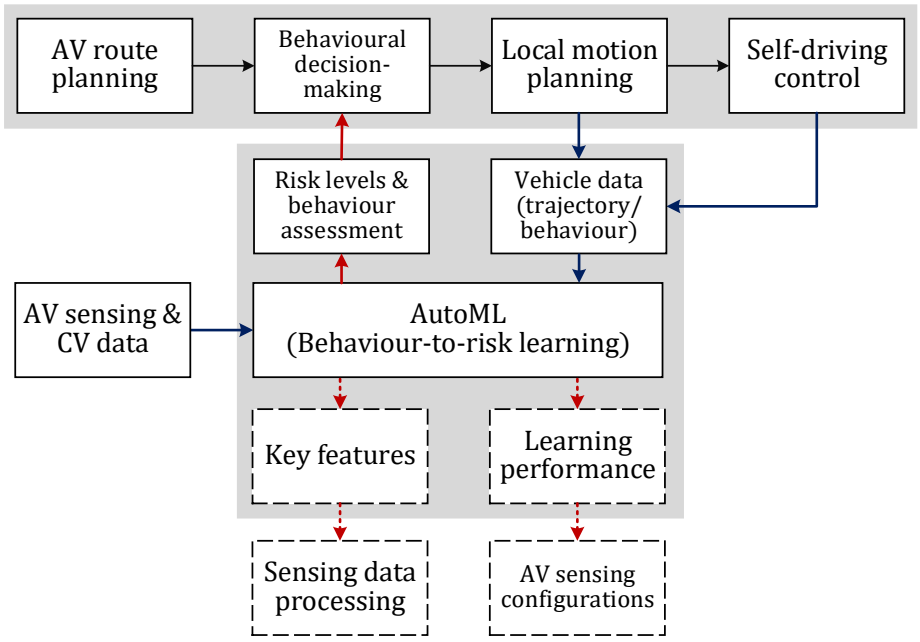


Figure 6.2 AV decision-making hierarchy

6.3.3 AV configurations from risk assessment viewpoints

The proposed AutoML aims to predict risk levels of vehicles in driving and route planning based on driving behaviour features. However, due to different data availability under different sensing quality, the extracted feature sets are different. Herein, four AV configurations are considered to represent various sensing information acquisition conditions, and the model performances under each condition are benchmarked. Meanwhile, data-driven insights about AV configurations from risk prediction viewpoints is another purpose herein.

The data availability about feature extraction involving four AV sensing configurations are listed in Table 6.1. First, the AutoML using data involving individual standalone vehicle is tested as a basis for comparison, which is to predict risk levels from behaviour characteristics of ego vehicle only, hence the add values of AVs in safety enhancement can be quantified from a risk decision-making perspective. Second, a basic AV sensing configuration can detect and measure the position of preceding vehicles, in which extra features are extracted from distance-based variables (e.g., relative position, or front gap). Third, with high-quality sensor perception and advanced fusion (e.g., AV Level 4 and above), more variables and finer features can be derived for risk assessment and driving scene understanding, such as using the relative relationships with the ego vehicle to measure the behaviours of surrounding vehicles. Furthermore, if the AVs also configure CV (connected vehicles) functions, additional information is available from V2V (vehicle-to-vehicle) and V2I (vehicle-to-infrastructure) communication. The AutoML performance and scalability in these four risk prediction scenarios are demonstrated.

The vehicle trajectory data provided in the NGSIM programme is used as the surrogate data, to emulate the safe and risk conditions that AVs could face in naturalistic flow (e.g., presuppose some safe behaviour vehicles as AVs). Herein, surrogate sensing data (e.g., front gap) is calculated using the trajectory data of vehicle pairs, and surrogate connected vehicles data is simulated using the data of all vehicles in the traffic flow. For instance, the front gap data is calculated based on the trajectory and length of the preceding and following vehicles.

Table 6.1 AV configurations and feature sets

Configuration	Sensors	Variables for feature extraction	Feature set
Ego vehicle standalone (movement behaviours only)	OBD (on-board diagnostics); position system + IMU (inertial measurement unit), cameras, etc.	Velocity, acceleration, lateral position, lane tracking, steering angular speed, vibration, jerk, etc.	I
Basic object detection and distance measurement	Basic sensor fusion of: vision odometry; LiDAR point cloud; Radar; semantic map, etc.	Front gap, relative distance, etc.	II
Advanced surrounding sensing and high-quality measurement	Advanced sensor fusion of: vision odometry; LiDAR point cloud; Radar; semantic map, etc.	Derived variables related to vehicle pairs, interactions and driving scenes, etc.	III
Addon connected vehicle (CV) communication	V2V; V2I; etc.	Information related to vehicle platoon and traffic conditions (e.g., average velocity or gap of vehicle stream)	IV

6.3.4 AutoML pipeline architecture

This section elaborates the main steps of the AutoML pipeline, in particular, data labelling and feature engineering adapted for various AV configurations, and auto-tuning by Bayesian optimisation.

Step 1: Feature extraction

Table 6.2 lists the feature extraction adapted for various AV configurations. The detailed processing of massive feature extraction is illustrated in Chapter 5.

Table 6.2 Feature extraction for various AV configurations

	Features	Counts
RISK INDICATOR FEATURES		5
TIT-based	TIT.t1; TIT.t2; TIT.3 (TIT measured by TTC threshold = 3s, 4s, 5s, respectively) (3).	3
CPI-based	CPI.m1 (CPI measured by MADR1 with threshold 3.4 m/s2); and CPI.m2 (CPI measured by MADR2) (2).	2
DRIVING BEHAVIOUR FEATURES		1,357
Ego vehicle standalone (feature set I)	{vel; acc; jerk}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(33); {vel; acc}.{kurt; mad; skew} (6); {vel; acc; y; y'}.{pct}.{ max; mean; min; p01; p05; p95; p99; q1; q2; q3; std; absm}(48); {y'; vel}.{logr}.{ max; mean; min; p01; p05; p95; p99; q1; q2; q3; std; absm}(24); {vel; acc}.{w1; w2; w3}.{rng; crng; sma; msd; rsd; emar}.{mean; std; max; min}(144); y.{std}(1); lane.{mean; std; rng}(3).	259
Gap related features	gap.{kurt; mad; max; mean; min; p01; p05; p95; p99; q1; q2; q3; skew; std}(14); gap.{pct; logr}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std; absm }(24); gap.{w1; w2; w3}.{rng; crng; sma; msd; rsd; emar}.{mean; std; max; min}(72).	110
Vehicle pair related (preceding vehicles)	pv.[feature set I](259); {vel; acc; x}.{dtw}(3); {vel; acc; x}.{w1; w2; w3}.{dtw}.{mean; std; p05; q1; q3; p95; max; min}(72); dif.{acc; vel}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(22); dif.{vel; acc}.pct.{absm; max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(24); dif.{vel; acc}.{w1; w2; w3}.{rng; crng; sma; msd; rsd; emar}.{mean; std; max; min}(144); {vel; acc; tv; tpv}.{pcor; scor}(8); {vel; acc; tv; tpv}.{w1; w2; w3}.{pcor; scor}.{mean; std; p05; q1; q2; p95; max; min}(192).	724
Vehicle stream related (within a lane segment for a time window	{acc; gap; vel; jerk}.vfr.{w0; w1; w2; w3}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(176); dif.{acc; vel}.vfr.{w0; w1; w2; w3}.{max; mean; min; p01; p05; p95; p99; q1; q2; q3; std}(88).	264

Step 2: Unsupervised data labelling of risk levels

The AutoML can conduct unsupervised data labelling of risk levels by the clustering based on risk patterns. This AutoML has a clustering evaluation mechanism (i.e. label identification by XGBoost) to determine an optimised solution, which can select the best-suited clustering algorithms or ensembles based on performance, and also suggest the best-suited clustering result. The detailed processing of unsupervised risk grading is illustrated in Chapter 4.

Step 3: Learning-based feature selection

The feature importance rankings for various sensor configurations are illustrated in Figure 6.3. The number of selected features and corresponding learning accuracy are listed in Table 6.3. The detailed processing of learning-based feature selection is illustrated in Chapter 5.

Table 6.3 Number of features selected by tree-based filtering and RFE

Feature set	Initial size	Weight-based	Gain-based	Total filtered	RFE suggested
I	259	90 (73.31%*)	102 (73.85%)	160	58 (75.06%)
II	369 (I+110)	54 (87.44%)	122 (87.17%)	149	74 (88.15%)
III	1093 (II+724)	132 (89.06%)	203 (88.99%)	281	83 (89.40%)
IV	1357 (III+264)	98 (89.14%)	152 (88.95%)	229	80 (89.33%)

* Corresponding learning accuracy.

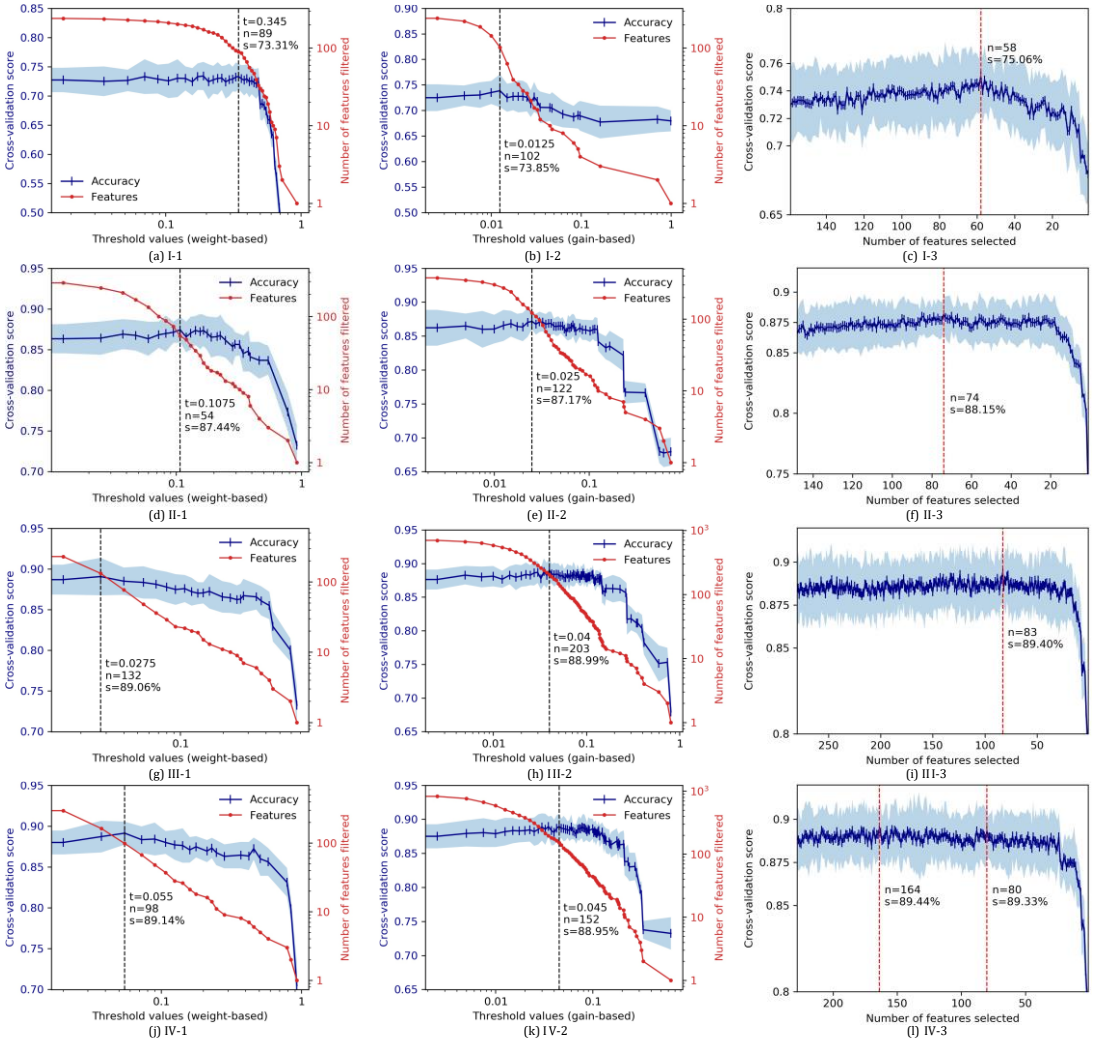


Figure 6.3 Feature self-selection and learning performance (of various AV configurations)

Step 4: Auto-tuning by Bayesian optimisation

The risk levels of vehicles in driving are predicted based on the selected key behaviour features. The modelling of behaviour-based risk prediction is improved in two parts, namely, model training using key features, and hyperparameter tuning by Bayesian optimisation. With the optimised features and hyperparameters, the final prediction model is obtained by re-training using the complete data. The predictive power (generalisation ability) of the final model can be estimated using the nested cross-validation performance.

Based on Bayesian optimisation, the best-suited model configurations and

hyperparameter values are obtained within specific computation capacity and time constraints. The hyperparameters in XGBoost is illustrated in Chapter 5, and the domain distributions of each hyperparameter are listed in Table 6.4.

Table 6.4 Hyper-parameter domain space

Hyper-parameters	Distribution	Domain
Learning rate	Continuous log uniform	(0.05, 0.3)
Number of estimators	-	300
Tree depth	Discrete uniform (integers spaced evenly)	[3, 4, 5, 6]
Splitting weight	Continuous uniform	(0.6, 1.0)
Instance subsample ratio	Continuous uniform	(0.4, 1.0)
Feature subsample ratio	Continuous uniform	(0.4, 1.0)
Gamma	Continuous uniform	(0, 1.0)
Alpha	Continuous uniform	(0, 1.0)
Lambda	Continuous uniform	(0, 1.0)

The domain space of hyperparameters over which to search is created as centred around the pre-tested values and then refine it in subsequent searches (Shi et al., 2019). The logarithmic uniform distribution is used for the learning rate because it varies across several orders of magnitude. The sampling of values in the domain is equally likely (uniform). The number of estimators is set to 300, but this number will not always be reached because early stopping is used to stop the training when validation scores have not improved for 30 iterations (i.e., 10% of total estimators).

Herein, the objective function is to minimise the log loss of the XGBoost model using specific hyperparameters, via 10-fold stratified cross-validation. The surrogate model of the objective function and the selection function for evaluating which hyperparameter values to choose next are based on TPE, and the setting of 500 iteration trials is used. The optimised hyperparameter values and corresponding performance scores are listed in Table 6.5.

Table 6.5 Optimised hyper-parameters and performance scores

	Feature set I	Feature set II	Feature set III	Feature set V
HYPER-PARAMETERS				
Learning rate	0.056	0.086	0.064	0.161
Number of estimators (early stop iteration)	116	153	264	90
Tree depth	6	5	4	5
Splitting weight	0.698	0.628	0.829	0.949
Instance subsample ratio	0.936	0.829	0.677	0.948
Feature subsample ratio (by tree)	0.668	0.833	0.854	0.656
Feature subsample ratio (by level)	0.842	0.645	0.945	0.673
Gamma	0.530	0.517	0.472	0.116
Alpha	0.520	0.934	0.432	0.999
Lambda	0.038	0.999	0.743	0.410
PERFORMANCE				
Misclassification	848	387	338	334
Accuracy	0.734	0.870	0.880	0.883
Accuracy (converted)	0.808	0.907	0.915	0.917
AUPRC (macro)	0.646	0.879	0.894	0.901
AUPRC (weighted)	0.777	0.931	0.939	0.943
AUC (macro)	0.870	0.972	0.977	0.978
AUC (weighted)	0.918	0.979	0.983	0.984

The performance and hyperparameters versus the iterations are plotted to inspect the auto-tuning process, as shown in Figure 6.4. The dark triangles indicate top-five optimal values. The average validation scores increase over time (conversely the loss decreases) as expected, indicating the method is trying better hyperparameter values. As the search progresses, the auto-tuning switches from exploration (e.g., trying new values) to exploitation (e.g., selecting values with better past results), which is more efficient compared to uninformed random or grid search methods.

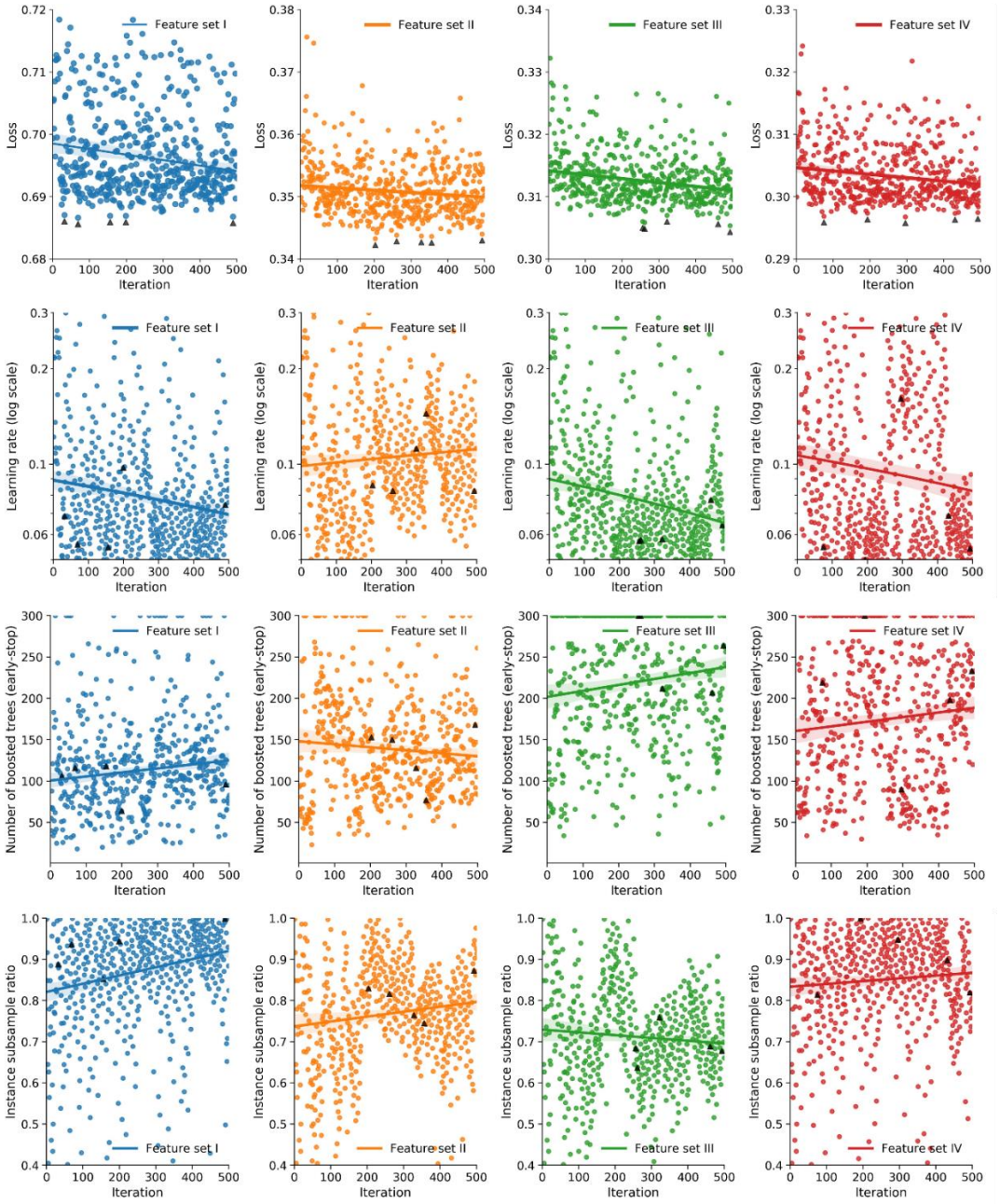


Figure 6.4 Hyper-parameters auto-tuning and corresponding performance

The dependence of loss with each hyperparameter is measured by mutual information (MI, also known as information gain), as shown in Figure 6.5. Key hyperparameters with more information are identified, especially the learning rate, instance subsample, estimators and tree depth. The loss-hyperparameter distribution relationships are plotted in Figure 6.6. There are noticeable trends that the Bayesian optimisation tends to concentrate (i.e., place more probability) the search on evaluating more promising

values. The identification of key hyperparameters and value distribution contributes to the better design of domain-specific AutoML.

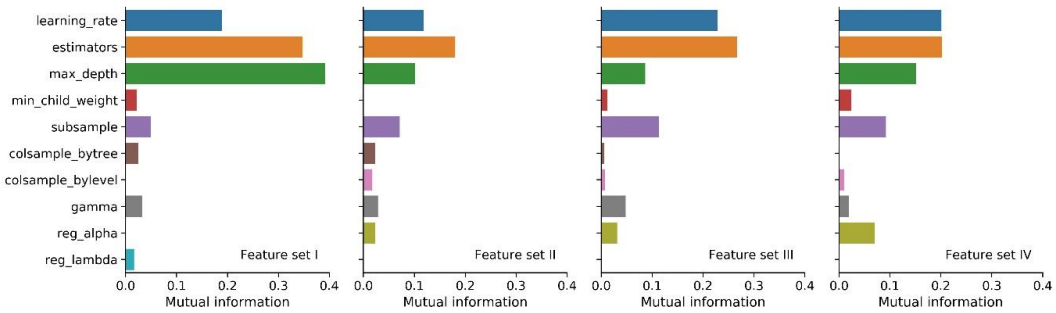


Figure 6.5 Hyperparameter comparison by mutual information of loss

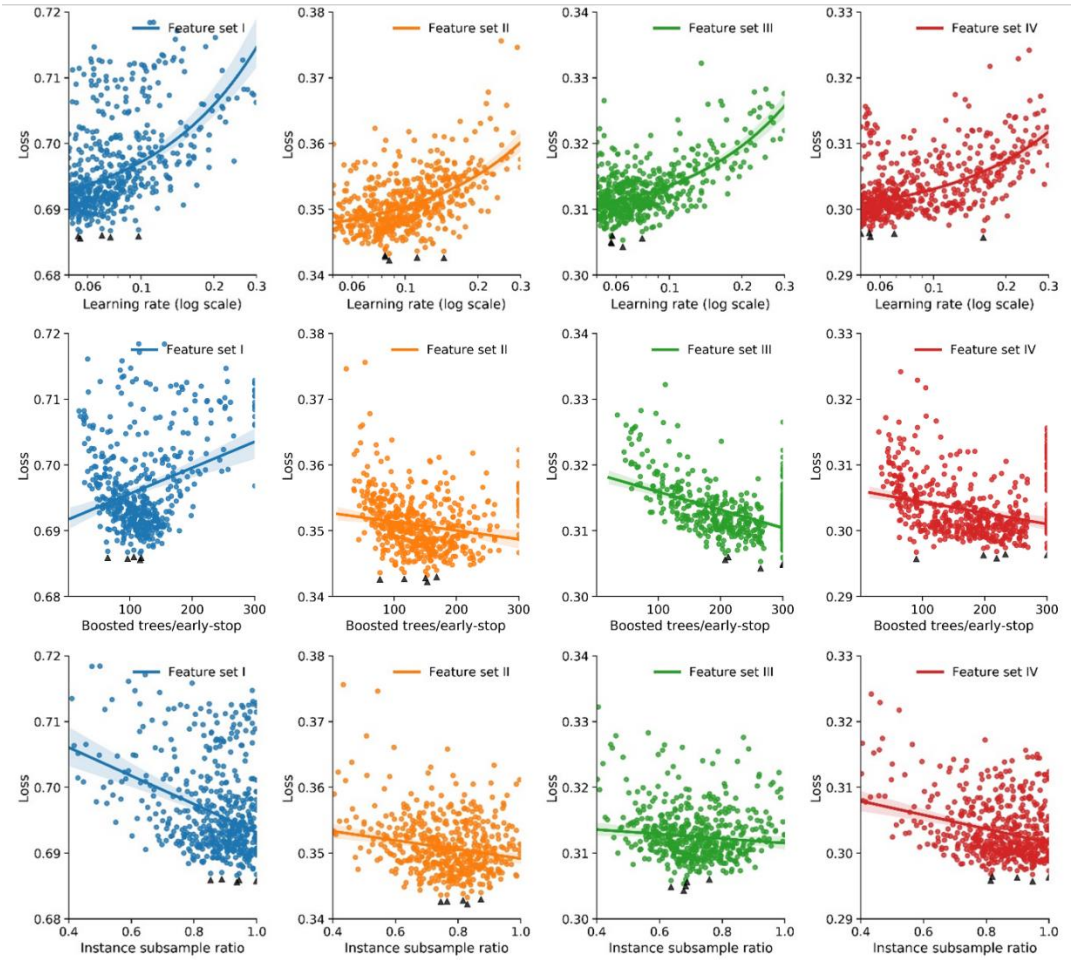


Figure 6.6 Loss versus key hyperparameters

After the Bayesian optimisation phase, an ensemble of several versatile models is

constructed, by iteratively adding the model that maximises ensemble validation performance. The automatic ensemble construction is more robust and less prone to over-fitting, compared in favour of the best setting. With the optimised features and hyperparameters, the final model is generated by re-training using the complete data set.

6.3.5 Prediction performance and key features

The potential capacity of risk prediction under the four AV configurations (corresponding to four feature sets) is compared. AutoML provides the optimised models for each AV configuration. From Table 6.5, AutoML achieves satisfactory results of behaviour-based risk prediction. The converted accuracy is about 91%, which is converted based on the raw data distribution before safe-class undersampling. Detailed estimation of the prediction performance for each risk level and AV configurations (various feature sets) is compared in Table 6.6. The detailed partition of risk levels helps to uncover potential risk conditions from the data without real accident cases, and the predictive power of greater than 95% accuracy is expected for safe-risk classification.

The prediction performance of various feature sets provides data-driven insights about AV sensing configurations from the perspective of the information needed for risk decision-making. Compared with basic features (feature set I), the gap-related features (feature set II) can vastly improve the performance of the higher risk levels, especially the recall of MR and HR. Hence, accurate front gap measurement is critical for risk decision-making. For high-quality measurement about vehicle pair relationship (feature set III), there is about 12% improvement on the precision of the HR level, and about 2% improvement on the overall performance. Additionally, the information about CV (feature set IV) helps to improve about 10% on the recall of HR prediction, but for other metrics, the addon values are not obvious. Thus, the surrounding risk evaluation (e.g., behaviour comparison, risk force field, relative driving performance of preceding vehicles) is helpful to assess the behaviour of individual vehicles pertinently, and contributes to risk decision-making, such as, identify a trajectory with fewer likelihoods of conflicts caused by surrounding

vehicles, or motion planning with defensive driving, etc.

Table 6.6 Prediction evaluation for detailed risk levels

	Risk levels	Precision	Recall	F1-score	AUROC	AUPRC
Feature set I	Safe	0.839	0.888	0.862	0.939	0.927
	LR	0.639	0.797	0.709	0.857	0.733
	MR	0.618	0.296	0.401	0.83	0.541
	HR	0.647	0.208	0.314	0.854	0.385
Feature set II	Safe	0.926	0.947	0.937	0.987	0.985
	LR	0.844	0.852	0.848	0.958	0.926
	MR	0.787	0.739	0.762	0.96	0.824
	HR	0.729	0.66	0.693	0.983	0.779
Feature set III	Safe	0.938	0.952	0.945	0.99	0.989
	LR	0.856	0.866	0.861	0.966	0.938
	MR	0.774	0.767	0.771	0.967	0.831
	HR	0.817	0.632	0.713	0.986	0.817
Feature set IV	Safe	0.94	0.955	0.947	0.99	0.99
	LR	0.856	0.867	0.861	0.967	0.938
	MR	0.788	0.76	0.774	0.967	0.846
	HR	0.813	0.698	0.751	0.987	0.831

The results of feature self-selection provide data-driven insights about the optimal processing and information mining of sensing data. The main variables and functions involved in the key features selected are listed in Table 6.7. For variables acquisition from sensing data, the gap, velocity and acceleration are the most informative variables to assess risk levels. In terms of the functions for sensing data processing, this study designs a series of ratio functions for behaviour data mining, and most of them have appeared in the selected key features, such as log-ratio, percentage change, and bias ratio. These ratio functions help to measure the abnormal changes of movement, and capture potential risk signals. Moreover, different forms of ratio have different sensitivities to changes, hence multiple functions complement the feature

mining. Besides, data processing at shorter time intervals is also important, given that a lot of key features are defined based on small moving windows. In-depth data mining can enhance the reliability and predictability of AutoML.

Table 6.7 Feature auto-selection analysis

	Feature set I	Feature set II	Feature set III	Feature set V
Total	58	74	83	80
NUMBER OF EXTRACTORS INVOLVED IN THE KEY FEATURES SELECTED				
Variables	vel(30); acc(19); y(6); jerk(3)	gap(34); vel(27); acc(8); y(4); jerk(1)	vel(39); gap(24); acc(13); y(6); jerk(1); pv.(19)	gap(32); vel(31); acc(11); y(2); jerk(2); pv.(13)
Functions	rng(12); w3(12); w1(11); pct(10); crng(8); logr(7); sma(5); w2(5); rsd(4); emar(4); msd(3)	w1(21); rng(18); pct(14); w2(11); crng(10); logr(9); w3(9); emar(7); sma(5); msd(5); rsd(3); pcor(3); scor(1)	w1(20); dif(18); w3(15); pct(13); rng(13); sma(12); w2(12); logr(8); crng(7); pcor(7); emar(5); dtw(2); scor(5); msd(4); rsd(2)	w1(21); vfr(18); dif(16); w2(14); w3(12); logr(8); rng(8); pct(7); sma(7); pcor(6); crng(5); emar(5); scor(4); rsd(3); dtw(2); msd(2)
Operations	mean(14); std(9); min(8); max(8); q1(4); p01(3); p99(3); p05(3); absm(2); q2(1); q3(1); p95(1)	std(18); min(10); max(12); mean(9); p01(4); p05(4); p95(4); q1(3); q2(3); p99(2); absm(2); q3(1)	mean(15); min(13); max(12); std(11); p95(6); p01(6); p05(5); p99(4); absm(3); q1(3); q3(2)	min(15); std(12); mean(9); p01(9); p05(9); max(8); p95(6); q3(4); absm(2); q1(1); q2(1); p99(1)

The tree-based AutoML with domain-specific features has the advantages of high transparency, being robust and fault-tolerant, etc. The AutoML can be integrated into the self-driving system to predict the risk levels of instantaneous motion trajectory,

as well as evaluating AV behavioural decision-making based on benchmarks of clustered safe driving.

6.3.6 Data-driven insights for AVs and AutoML

AutoML for predicting detailed risk levels is more challenging but valuable. Accurate assessment of risk levels can improve the confidence level of AV decision-making, which complements intuition-like decision-making in end-to-end deep learning (i.e. scene-to-action). Enhanced performance from the isolated validation indicates an accurate and reliable predictive power, whilst allowing the model to work well on data that are not used in the modelling. The proposed AutoML is helpful in the planning of non-collisional and optimal motion trajectory by adding a clear assessment of risk levels (e.g., accurate safe boundary and optimal risk buffer). Moreover, the risk levels provide early signals for crash potentials and likelihood, which can be used to plan trajectory and/or motion based on pre-emptive identification of unsafe conditions, such as driving safety field theory and defensive driving strategies (Wang et al., 2016; Liu and Khattak, 2016). Besides, the AutoML can enhance the AVs with the ability to quantify and predict the risk levels, and modularise the sensing, decision-making and control, to operationalise the required transparency and interpretability for AV optimisation and regulation.

AutoML reveals the most important behaviour features used for risk assessment, which provides useful insights for information processing and sensor configuration, as well as interpretable risk decision-making mechanisms. Feature fusion and hybrid measures show good performances (e.g. low false alarms, high accuracy rates), and redundancy risk assessment is desirable to promote system reliability. For AV and ADAS (advanced driver-assistance systems), the AutoML for behaviour-based risk prediction can be deployed as a co-pilot system, which can be used to offer reliable and appropriate risk warning and behaviour assessment. The estimation of the risk levels of surrounding vehicles is helpful to improve risk assessment. Conversely, for human-AV interaction, if the detected responses of drivers to hazards are inconsistent with the risk conditions, the system would take over control of the vehicle to reduce the crash risk. Reliable risk prediction and behaviour assessment are important for

improving driver trust in the AV system.

For practical application with computation constraints, the AutoML also supports offline warm start using an ensemble of pre-trained models. The meta-characteristics of new datasets are first captured to identify and match similar datasets in the known data space (Feurer et al., 2014). Thus, the ensembles of similar pre-trained models can be used directly, or the AutoML training can start with similar configurations to seed the Bayesian optimisation. This can boost the overall modelling within a reasonable amount of time or computation limits. Besides, in cognisance of resource limitations, the trade-off between optimisation time versus the settings of pipelines can also be tested.

Since the data about real-world traffic flow involving AVs are not massively collectable currently, this study used surrogate data to emulate the scenes and data available under AV and CV environments. The NGSIM data are about human driver vehicles, of which the behaviours are used to train AVs. Moreover, the model performance is highly dependent on data quality and quantity. Since the size of the dataset is relatively very small, the benefits of hyperparameter optimisation are not well presented. The improvement may have diminishing returns, because of noisy data, small size, and hidden variables that are not measured (Mannering et al., 2016).

Bayesian optimisation is effective, but not guaranteed to find the best hyperparameters, since it has the risk of getting stuck in a local minimum of the objective function. This can be checked by starting an entirely different search. If the subsequent searches focus on similar values, such values are expected to be optimal. Although the random search does not suffer from this issue, given the high dimensionality and complex interactions between hyperparameters, the random search is much more computation expensive.

6.4 Pay How You Drive Insurance

6.4.1 Behaviour-based insurance

Reliable driving assessment can be used to determine a measure for discounts or

penalties for annual car insurance premiums, which provides a basis for the design of behaviour-based insurance (BBI, such as “pay how you drive”, PHYD). Behaviour-based insurance is more reasonable than usage-based insurance (UBI, also known as “pay as you drive”, PAYD) and traditional fixed insurance premiums. Different types of vehicle insurance premium schemes are compared in Figure 6.7.

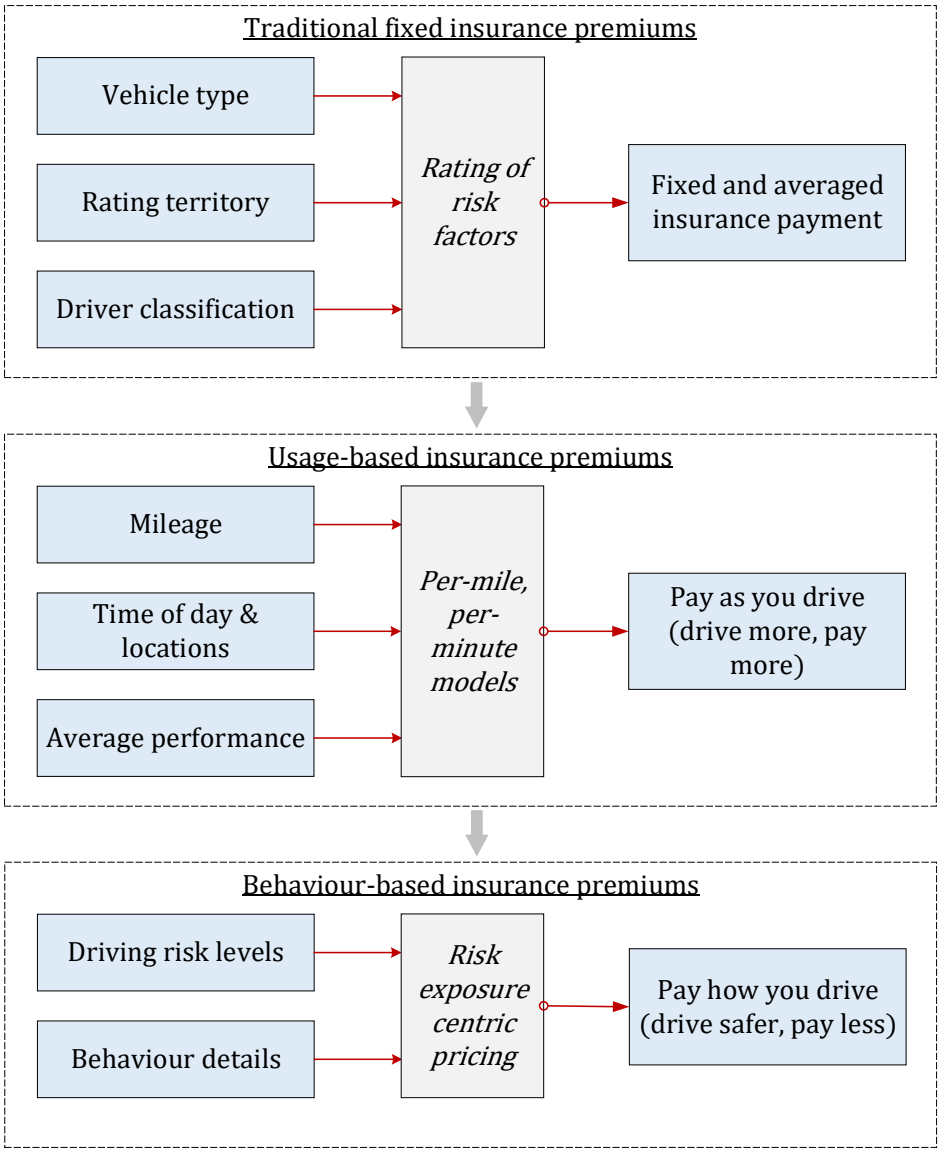


Figure 6.7 Types of vehicle insurance premiums

The traditional insurance premiums are exclusively through a fixed rating of risk factors, such as, vehicle type (e.g. vehicle status, initial purchase price), rating territory (i.e. accident rates and claim records in the area where the driver lives),

driver classification (e.g. age, gender, credit history, driving record, driving habits). PAYD further considers travel and usage information, such as, mileage, time of day, road type, average velocity (Li et al., 2018). Examples of PAYD include per-mile premium, per-minute premium (Bian et al., 2018).

The most recent insurance premium is PHYD, which considers the impact of various driving risk levels, which calculates a personalised risk pricing scheme based on detailed behaviours (Li et al., 2018). PHYD has been recognised as the most promising differentiated commercial vehicle insurance premium strategy (Nai et al., 2016). PHYD insurance is a fair and straightforward policy whereby the safer you drive, the less you pay. PHYD is recognised as a common agreement of next stage insurance premium. However, the key issue is risk assessment and prediction of the risk exposures of vehicles in normal driving, as well as related system design and hardware configurations to conduct the task as a product. The maturity in the development of sensor techniques and AVs enables high-quality data acquisitions for in-vehicle behaviour assessment and risk prediction.

As a data-driven business service, the key of PHYD is the behaviour-based vehicle-level risk grading. The proposed AutoML of behaviour assessment and risk prediction using feature extraction and machine learning fills the research gap. Measurable risk grading is obtained based on a large group of vehicles within a road segment (as described in Chapter 4), and the linkages between driving behaviour features and corresponding risk levels are built (as described in Chapter 5). Another key research issue is how to develop a personalised risk pricing model (Bian et al., 2018). The vehicles with different risk levels in driving can be classified based on the monitoring of driving behaviours, and corresponding insurance premiums can be defined based on driving risk levels.

6.4.2 PHYD prototype

Based on the AutoML, a personalised risk pricing model is proposed by examining the driving risk exposures, and calculating a personalised vehicle insurance premium, as well as providing guidance to enhance driving. The system prototype for the risk exposure centric PHYD is designed, as shown in Figure 6.8.

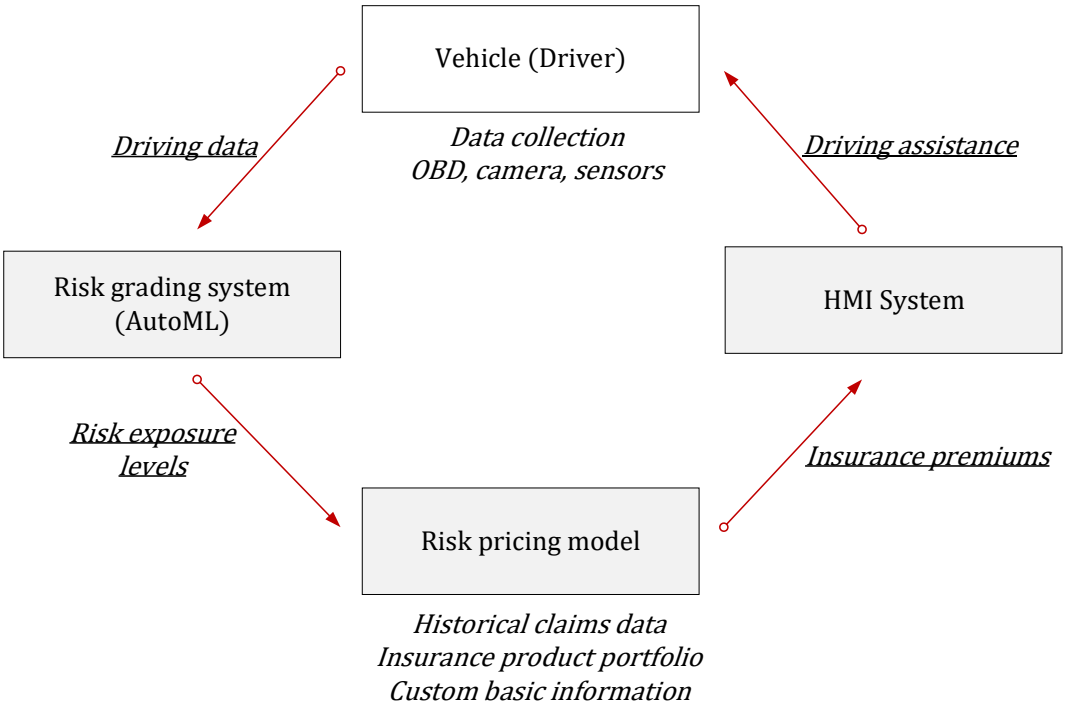


Figure 6.8 Risk exposure centric PHYD prototype

Firstly, data collection and processing module is an essential part, which extracts vehicle driving data from built-in and/or post-equipped sensors and telematics devices, and the AutoML is used to guide the system design and configurations. Secondly, a risk grading module using machine learning is employed to evaluate the driving risk levels in real time. Thirdly, a risk pricing model is established to formulate the vehicle premiums according to the classified driving risk levels.

In this chapter, an appropriate path for personalised PHYD is by offering logical premium discounts as the incentives for safe driving. The calculation of risk pricing is formulated, as follows:

$$P = C_0 - \sum_{r=0}^{R} \sum_{t}^{T} d_t^{(r)} \times p_t^{(r)}$$

where C_0 denotes the basic premium cost per policy term T, r indicates the risk level estimated by the risk grading module, $d_t^{(r)}$ is the discount rate for a corresponding risk level, $p_t^{(r)}$ is the average unit insurance payment of the risk level.

The penalties of unsafe driving will reduce the rewards and discount parts only.

Moreover, through an in-depth assessment of driving behaviours, the system also provides real-time driving assistance and post-driving analytics feedback. The system interface will present instantaneous warning and notification when a risk condition is detected. Only the necessary driving assistance information is displayed to reduce the distraction in driving, such as, unsafe front gap maintenance, defensive driving based on the prediction of risk conditions (e.g. to avoid abrupt braking). The driving scoring and potential insurance premium are generated and updated after each driving, and displayed as post-driving feedback on the smartphone apps and system interface. Besides, detailed driving assessment and suggestions are also presented, as motivation for improving driving skills and becoming eligible for cheaper insurance and discount awards. For risk evaluation and pricing, the estimation of the risk levels of surrounding vehicles is necessary to make a fair and reasonable assessment of the responsibility in risk conditions, since the conflicts involve the interactions of vehicle pairs. The surrounding risk assessment serves to identify some no-fault risk conditions that can be waivable.

PHYD offers benefits to both customers and insurers, as well as society. The business model is simple, but changes the existing insurance business such that lower risk drivers pay less and hence encourages safe driving. The customers will not only have a greater awareness of their driving performance, but also gets advice on how to improve their driving and hence rewards. Besides, instead of paying an averaged and fixed insurance, PHYD aims to price fairly based on the driving skills and risk levels, regardless of the age, gender, or whether the driver has had a previous accident but performs well nowadays. A low-risk insurant with a high mileage could also incur low expenses, which is different from PAYD (long distance of driving incurs high expenses). The PHYD could be preferred especially by younger/older drivers, and drivers with a history of accident records. Moreover, to reduce additional costs, the monitoring devices can be rented by paying a deposit.

Through PHYD, insurance-based incentives encourage the insurant to drive in a safer manner, which in return reduces the total claim costs for insurers. The potential

saving of claim costs would be much more than the reduction of revenues from offering discounts. Moreover, advanced driving assistance is provided by predictive modelling and machine-learning technologies, which may change the business model of insurers, and leads to the transition from being a pure insurance product provider to becoming insurance-service hybrids. Besides, PHYD helps in tackling major road safety risks, and can yield additional societal benefits through reduced accident risks.

6.4.3 System design

For an accurate and reliable risk rating, it is important to obtain high-quality driving data with devices that are easy-to-implement. The developed AutoML system contains a process of learning-based feature selection, which shows that the velocity and front gap are the main variables to evaluate the risk levels of drivers. Conditioned on different hardware configurations and development phases, several system designs are proposed, to seek the most appropriate PHYD deployment, as illustrated in Figure 6.9.

The basic configuration generally includes a dashboard camera, a risk grading system, and a port to access the data from vehicle OBD and/or smartphone sensors, as well as telematics connection, as shown in Figure 6.9 (a). The driving behaviour data is usually available as records stored in OBD (Bian et al., 2018), and vehicle movement data measured by smartphone sensors is an alternative way to obtain the OBD driving data. The dashboard camera is necessary for recording external conditions and video-based front gap measurement. The risk grading system calculates risk levels by a built-in processor or through a cloud-based computation platform. Besides, a user-interaction interface is used for real-time driving assistance. The system is configured in a low intervention manner. For post-driving feedbacks, smartphone apps are designed to display detailed analytics information and for visualisations.

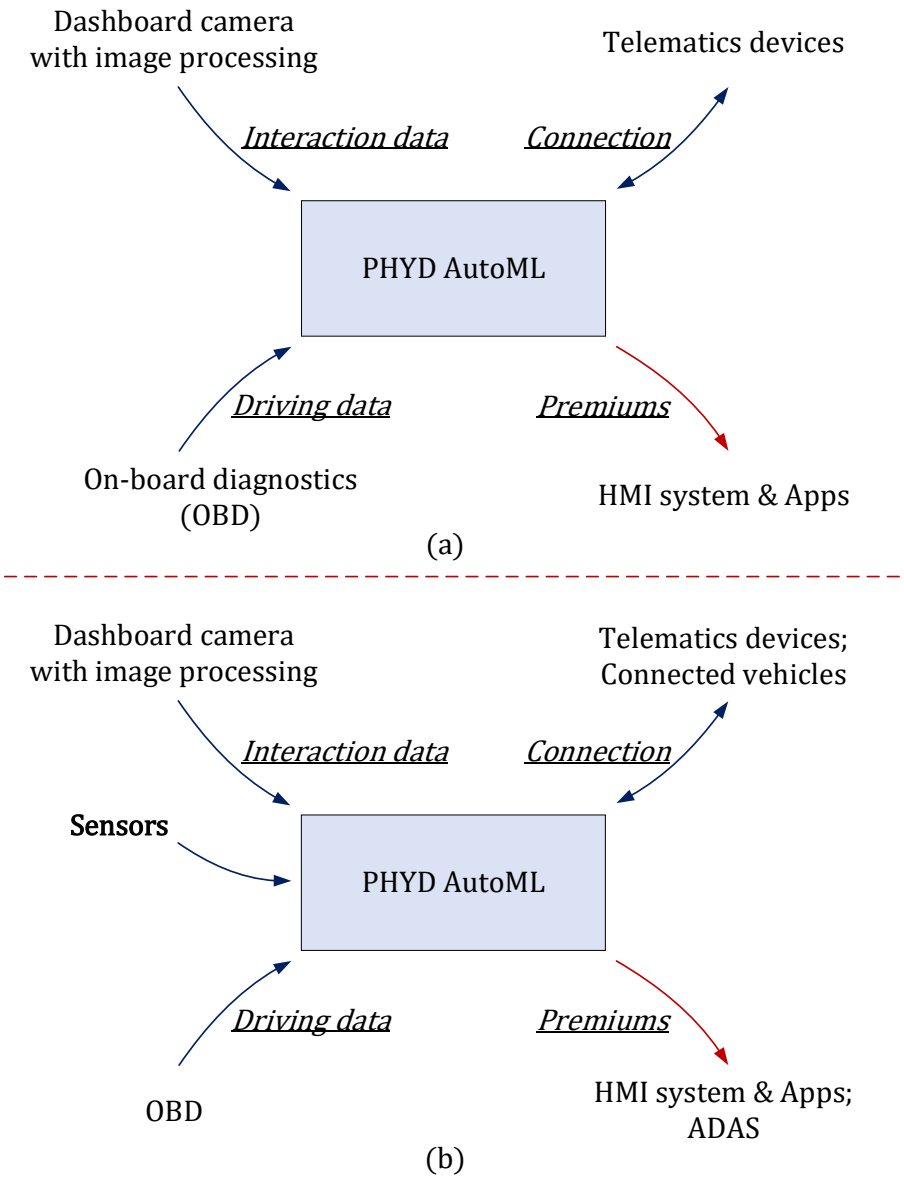


Figure 6.9 PHYD system designs

In Figure 6.9 (b), an improved configuration further integrates in-vehicle sensors (e.g. radar, lidar), which enables more accurate and robust collection of driving-related variables, such as, front gap, longitudinal and lateral acceleration, steering angle, etc. With the increasing availability and commercial usage of in-vehicle sensors, more real-world data is accessible, which provides sufficient instances to model training and refine the risk grading system. Detailed analysis of sensing and system integration is discussed in Section 6.5.

The AutoML is scalable for various device configurations. The end-user device will be a cost-efficient solution for in-vehicle applications with sensors, and the alternative solutions of sensors are also provided, including smart-phone based sensors and dash camera-based data collection. Generally, the costs for customers (end-users) is the rental/purchase costs of the in-vehicle devices.

6.4.4 Risk-based incentives

Risk-based financial incentives can be used with other technical solutions to promote better deployment and engagement. Besides PHYD, potential application scenarios include insurance and regulations for AVs, risk pricing on toll roads, among others.

Insurance and regulations for AVs. The behaviour-based risk assessment can be a basis of the insurance premiums offered to AV providers. Based on interpretable risk-behaviour learning, the responsibility of each party can be differentiated, such as the responsibility of drivers and AV machines, the responsibility of sensing perception techniques and AV decision-making algorithms, etc. A clear risk assessment and responsibility differentiation can help to enhance the regulations of AVs, and encourage the acceptance of end-users. Accurate assessment of risk levels can be also used to benchmark the AV driving performance and conduct behaviour-based regulations on AVs.

Risk pricing on toll roads. Similarly, such behaviour-centric risk pricing policy can be deployed on toll roads, as incentives to encourage safe driving by offering a discount of tolling fees.

6.5 Driving Safety System under CV Environment

6.5.1 Vehicle-to-road collaboration

Connected vehicles (CV) and AVs are two important directions to improve safety, compared with AVs which are already on-road tested, the safety roadmaps under CV environment are still not very clear. The domain-specific AutoML fits well for the driving safety analysis under the CV environment.

CV techniques enable massive acquisition of high-quality driving and traffic data, such as, vehicle to vehicle (V2V) and vehicle-to-infrastructure (V2X), which allow previously unrealisable services to be deployed in a scalable, real-time, and reliable manner (Wan et al., 2014; Guo et al., 2015).

Under the CV platform, the driving safety analysis can be deployed in a vehicle-sensing-cloud structure, the system architecture and integration of the vehicle-to-road collaboration for driving safety are depicted in Figure 6.10 and Figure 6.11, respectively.

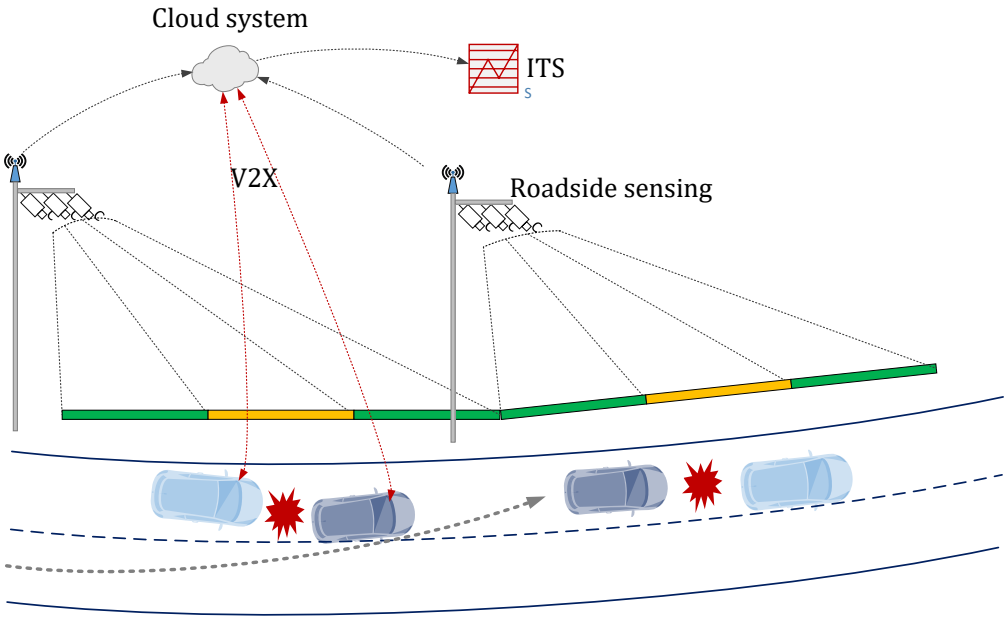


Figure 6.10 Vehicle-to-road collaboration for driving safety

The system includes three parts, namely, road-oriented data acquisition (e.g. roadside sensing for vehicle-level trajectory coordinates), vehicle-based safety systems (e.g. ADAS with risk warning and driving recommendations), and cloud-based AutoML for risk prediction and behaviour assessment. Two types of data acquisition are considered preliminarily, namely, 5G-V2X connection and roadside sensing.

Sensor-equipped vehicles and roadside infrastructure generate big traffic data, including vehicular spatial-temporal trajectories, vehicle stream driving status, etc. Within a cloud platform, the big data is delivered into the cloud-based AutoML

system for data mining, and smart information (e.g. risk exposures, risk-avoidance warning, surrounding sensing, short-term crash prediction, etc) is generated which in turn, is sent back to various receivers, including targeted drivers, authorities' system, roadside infrastructures, etc. The cloud-based structure is efficient to perform the data mining and discovery of smart information, and allow the realisation of a broad range of applications. Technological innovation in this field is accelerating, especially 5G connectivity, which will provide efficient and low-cost data communication and cloud-based computation for sensor perception.

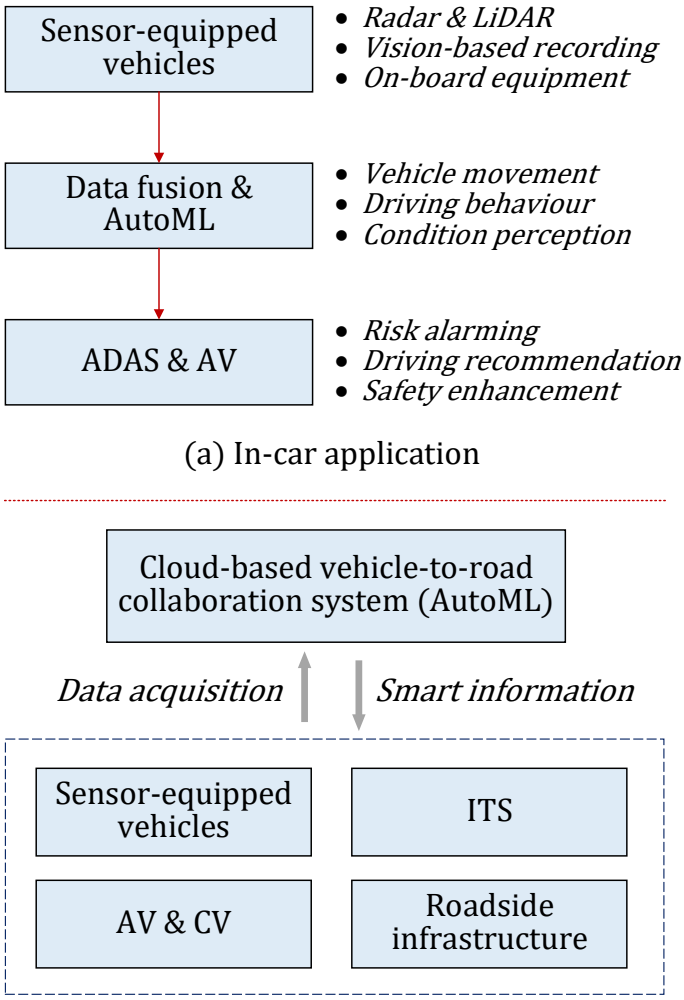


Figure 6.11 System integration under CV platform

The system can provide a high-resolution vehicle-scope risk mapping and positioning, which reveals the risk patterns (e.g. severity, frequency, trends) and early

signals about crash potentials. A high-resolution risk mapping and positioning has been demonstrated in Section 4.5.3. The information can be provided to the drivers and intelligent transport systems (ITS) centre, to proactively safety manage from the vehicle-level and pre-crash risk scope, which extends the scope of traditional accident management (passive and post-accident). Drivers can receive the risk information and driving recommendations from the in-vehicle devices. For the authorities, detailed risk levels can offer extended information on safety, therefore, protective strategies could be applied pre-emptively, such as patrol despatch, flow control, road enhancement (e.g. remediation of layouts or locations that may lead drivers to more likely indulge in unsafe driving).

For traffic safety management, detailed assessment of risk conditions is helpful for authorities in policy-making and targeted countermeasure design to enhance traffic safety. Potential outcomes from the system include, predictive management of traffic accidents on the expressway, driving enhancement systems for commercial vehicles and public buses, the identification and remediation of crash-prone locations, etc. The solutions would bring in great benefits, such as, the reduction of certain types of expressway accidents, the efficient management of road safety, the identification of accident-related congestion, etc.

For vehicle driving assistance, apart from the risk decision-making for AVs as discussed in Section 6.3, this system can be used to assess driving performance (short-term) and driving style (long-term). The vehicle-side services could provide risk-based driving assistance, such as risk alarm and driving recommendation (e.g. re-routing to alternative routes to avoid road segments with risk potentials), etc. Since the vehicles with risk exposures are possible to be identified and tracked in advance, proactive intervention could be taken before any impending crashes, such as risk warning by in-vehicle devices and other ADAS.

6.5.2 Roadside sensing

High-quality data acquisition and processing are important to build a reliable system for applications. Based on the feature learning in AutoML, important variables for risk-behaviour analysis are suggested, which mainly include vehicle movement

variables (e.g. velocity, trajectory, and acceleration) and vehicle interaction variables (e.g. front gap). For the vehicle movement variables, internally they can be accessed from OBD, and externally, on-board sensors can be used to measure these data, such as high-resolution positioning system, gyroscope, etc. For the vehicle interaction variables such as front gap, a direct way is using built-in-vehicle sensors, such as radar and LiDAR (LIght Detection and Ranging).

Alternatively, a cost-efficient way is video-based roadside sensing. Vehicle trajectory data can be collected using vision-based vehicle detection by roadside multiple cameras that cover the entire road segments, and using computer vision algorithms to measure the trajectory of each vehicle and evaluate the distance such as front gap values. Generally, the vehicle trajectory data is obtained by firstly vehicle detection and then trajectory measurement. The grid remapping algorithm for trajectory measurement and coordinate transformation is developed in Chapter 3.

Recent advances in smart vehicles provide prominent and mature techniques for sensor fusion and data acquisition, which can also be deployed on the roadside infrastructures. Primary sensors for data acquisition are shown in Table 6.8. Sensor fusion of roadside cameras and roadside Lidars helps to provide accurate measurements, and promote the system reliability.

Vehicle-mounted lidar sensing provides clean measurements with a wide field-of-view, allowing for object tracking across multiple lanes. Lidar has been extensively used for object detection in autonomous vehicles (Levinson et al., 2011). However, lidar sensing is sensitive to precipitation. Radar sensing works consistently in different weather and illumination conditions, but has a narrow angular field of view, and suited for tracking preceding vehicles in the ego lane (or host lane, where the vehicle is positioned) (Paul et al., 2016; Berriel et al., 2017). A millimetre wave radar sensor continuously measures the speed and position of the preceding vehicles on the road ahead. Video-based data collection still has limitations in certain scenarios, such as, overlapping in dense areas, shadowing, illumination, etc. The data fusion from cameras and sensors helps to provide accurate measurements. Generally, vehicles are detected and tracked by video cameras and sensors, and then the detected objects

discerned to be vehicles or other obstacles based on features such as size constraint, motion characteristics (Mao et al., 2012).

Table 6.8 Sensors for data acquisition

Sensor	Function	Processing	Application
Camera	Visible light (ambient)	Perspective projection, distance calculation by stereo vision, computer vision and video processing	Appearance motion, classification and detection, vehicles position, lane tracking
LiDAR	600-1000 nanometre-wave laser signals at 10-15hz	Distance calculation by backscattered energy in an imaging receiver	Spatial segmenting, scene scanning, vehicle detection, occupancy grid
Radar	Emitted millimetre-wave radio signals	Distance of the object by frequency shift in the received signal; demodulation	On-road tracking, distance measure, object detection

6.5.3 5G-V2X connection

5G information exchange can facilitate the data acquisition and processing of vehicle movement trajectory. The vehicle-level data can be uploaded to cloud computation, which produces accurate behaviour data based on raw data, such as, OBD data, dashboard camera, high-resolution positioning data, etc. Moreover, 5G also simplify the end user side deployment, within a fully connected vehicle road segment, a smartphone app can support the deployment of the safety functions in which the information is from cloud computation.

6.5.4 In-vehicle devices and service

The AutoML is scalable for various device configurations. Based on risk prediction and behaviour assessment, a series of safety enhancement services can be provided to the driver via ADAS, in-vehicle devices and smartphone apps, such as, risk warnings, driving recommendations and behaviour detection (e.g. inattention,

fatigue), which can enhance the risk awareness and optimise decision-making and driving-suitability for drivers.

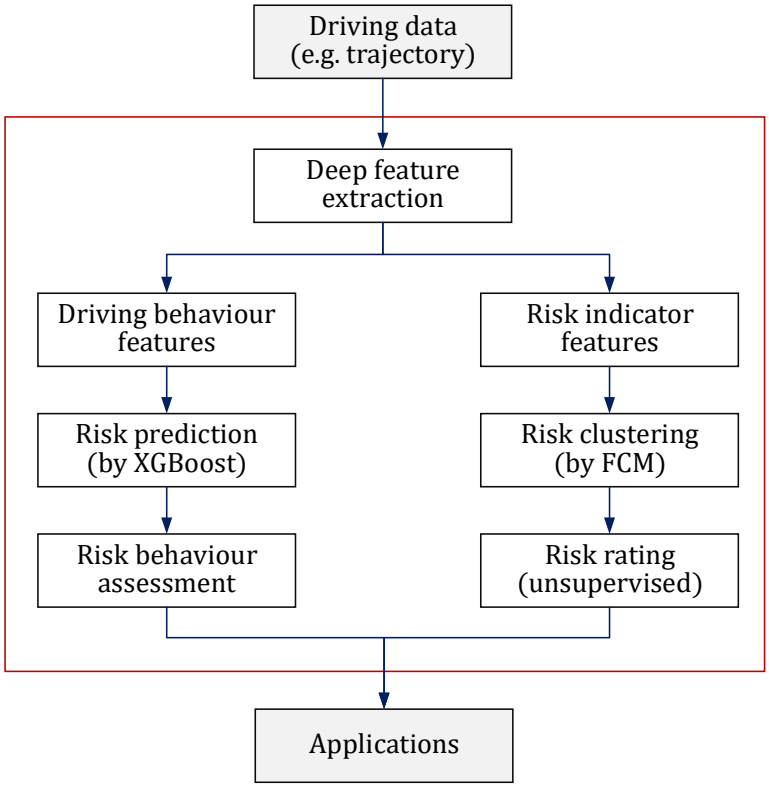


Figure 6.12 Module design and components

For ADAS, the risk prediction and behaviour assessment by AutoML could empower the system with a more reliable and appropriate risk warning and avoidance. Collision avoidance is an important function of ADAS. When the pre-defined indicators and rules exceed certain safety thresholds, the system would provide a collision warning, and autonomous braking assistance would be activated in higher-level AV systems, to reduce the crash risk (Li et al., 2018). Generally, the risk conditions are too complex to be covered by simple predefined rules and thresholds (e.g. using TTC as the indicator). Powered by AutoML, adaptive warning or evasive manoeuvres are initiated based on predicted risk levels of driver behaviours and traffic conditions, as well as the responses of drivers.

Besides, to improve acceptance and helpfulness, the interactions of the ADAS co-pilot service should be personalised based on the driver's driving performance and

risk awareness, especially the risk warnings and driving recommendations. An important characteristic is to foresee the risk conditions and possibility, and provide the co-pilot service that can meet the driver's real needs, otherwise could disturb and distract the driving. Since the driving performance (e.g. driver hazard perception, response to external hazards) is assessed, personalised and matched assistance could be provided, which is important for improving driver acceptance and trust on the system (e.g. false alarms if mismatching with drivers' hazard perception). Existing ADAS provides predefined and undistinguished services to all the drivers, such as HMW (headway monitoring and warning), FCW (forward collision warning), FVM (forward vehicle moving), LDW (lane departure warning).

For traditional vehicles, in-vehicle devices and smartphone apps can be equipped to provide near equivalent functionality. For example, the human-vehicle interaction in the forms of Augmented Reality (AR) and HUD (head-up display) can be designed, which displays trajectory tracking marked with risk levels, risk levels of surrounding vehicles for defensive driving, enhanced risk warnings and avoidance suggestions, among others.

The information mining about risk-behaviour is clearly an area with huge add-values. The proposed AutoML is helpful to provide a platform to analyse the driving log data, and compute the behaviour performance. Accurate assessment of risk levels can also be used to benchmark the driving performance of ADAS and AVs, and the ranking and quantitively standards can be conducted. The risk-behaviour information from vehicle probes can be used by road authorities for all kinds of traffic management measures at the system level. The information can thus be a revenue stream (for the providers such as insurance companies) when sold to agencies.

6.6 Risk-based Short-term Crash Prediction

6.6.1 Crash prediction based on vehicle movements

Future accident events are generally unpredictable, and guaranteed accurate information about accident occurrence is impossible. Herein, crash prediction is decomposed into components that are possible to predict, especially for crashes

involving vehicle conflict. As the precursors of crashes, risk exposures are used to make inference about possible developments and the likelihood of impending crashes, and forecast what would happen under specific conditions.

Crash is predicted as the probability of outcomes of risk exposures, rather than a specific outcome. The outcomes of risk exposures depend not only on the possible risk levels, but also consider the likelihood of the risk level (R). Hence, the risk exposure is an integrated value of the risk level (r) and the likelihood of each risk level over time (t).

$$Risk\ exposure\ = \sum_{r=0}^{R} \sum_{t=0}^{T} Risk\ level_t^{(r)} \times Likelihood_t^{(r)}$$

The risk levels can be predicted based on the assessment of driving behaviours within a time window. The feasibility of risk assessment and prediction are evaluated in the preceding chapters, including indicator-based crash assessment (in Chapters 3 and 4) and behaviour-based risk prediction (in Chapter 5). Furthermore, according to the learning-based feature selection, the key driving behaviour features are identified, which provides guidance for data collection and processing. The framework of risk prediction and crash inference is depicted in Figure 6.13.

The risk likelihood describes the possibility of a risk process occurring in the future. To make crash inference ahead of time, possible vehicle movements and subsequent likelihood should be predicted. Since vehicle movements follow certain physical mechanism, the next-seconds vehicle trajectory can be generated based on past behaviours and interactions with surrounding vehicles, for example, using the method integrating LSTM (long short-term memory recurrent neural networks) and MCMC (Markov chain Monte Carlo) to predict the motion trajectory of individual vehicles in the next seconds. LSTM addresses the $t+n$ (e.g. n<10s) prediction using the historical trajectory time-series and interactions with surrounding vehicles, and MCMC solves the likelihood of prediction based on the impacts of possible actions and status. Based on LSTM+MCMC, different likelihoods of future vehicle movements and subsequent risk levels under each condition can be predicted,

thereby, the distribution of predicted risk exposures is evaluated.

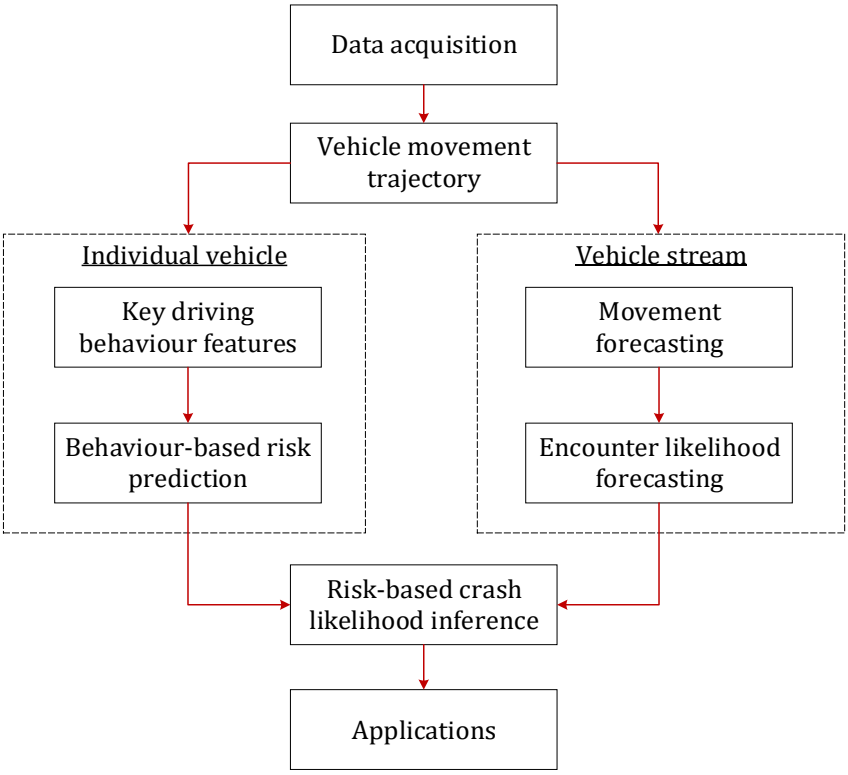


Figure 6.13 Risk prediction and crash inference framework

The risk likelihood describes the possibility of a risk process occurring in the future. To make crash inference ahead of time, possible vehicle movements and subsequent likelihood should be predicted. Since vehicle movements follow certain physical mechanism, the next-seconds vehicle trajectory can be generated based on past behaviours and interactions with surrounding vehicles, for example, using the method integrating LSTM (long short-term memory recurrent neural networks) and MCMC (Markov chain Monte Carlo) to predict the motion trajectory of individual vehicles in the next seconds. LSTM addresses the $t+n$ (e.g. n<10s) prediction using the historical trajectory time-series and interactions with surrounding vehicles, and MCMC solves the likelihood of prediction based on the impacts of possible actions and status. Based on LSTM+MCMC, different likelihoods of future vehicle movements and subsequent risk levels under each condition can be predicted, thereby, the distribution of predicted risk exposures is evaluated.

The risk exposures (i.e. severity and likelihood) are predicted from vehicle

movements and subsequent vehicle conflicts, and the likelihood of an impending accident can be inferred dynamically, as well as the involved vehicles, potential locations and timestamps, etc.

6.6.2 Crash prediction based on risk trends

The trends of risk exposures are important signals for crash prediction. The risk exposure is a complex and dynamic process, which includes not only the risk accumulation for a single vehicle or risk factor, but also the risk aggregation of multiple vehicles and factors in the traffic flow. Risk accumulation would increase the severity of a single factor, which may lead to an accident directly. Risk aggregation is the encounter of multiple risk factors, which would also increase the accident likelihood. Encounter is a traffic situation in which two road users approach each other in time and space and may influence each other's behaviour. The risk accumulation and aggregation interactions are depicted in Figure 6.14.

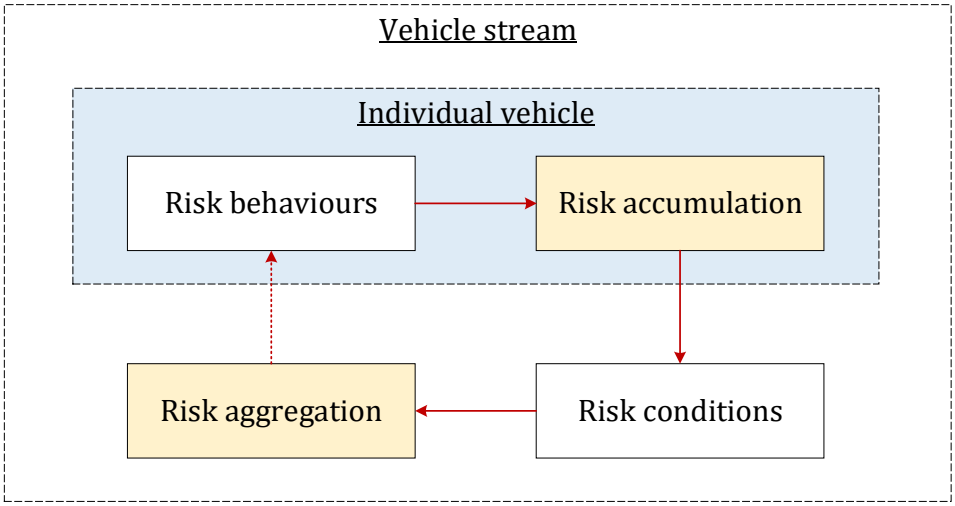


Figure 6.14 Risk accumulation and aggregation interactions

The risk trends are helpful to understand the detailed process of an impending crash, in which the risk factors are able to be highlighted and monitored. The crash predictors, risk aggregation (traffic flow) and risk accumulation (vehicle instance) can be evaluated and monitored by indicators and behaviour-based risk predictions, which are analysed in previous chapters. Based on the risk trends, the targeted vehicle with unsafe behaviour and/or involved in a risk condition is able to be figured out,

and real-time safety countermeasures can be given.

6.6.3 Predictive crash mitigation

Reliable prediction of risk exposures and the development of risk trends are more meaningful to design targeted countermeasures for crash mitigation, since a reliable system should provide appropriate countermeasures matched with the risk conditions accordingly, with higher accuracy and fewer false alarms, which are important to enhance system efficiency and trust. A demonstration of countermeasures for different risk conditions is illustrated in Figure 6.15.

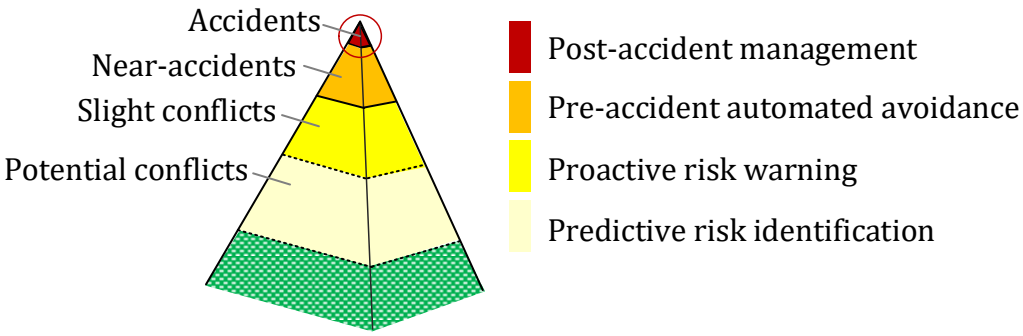


Figure 6.15 Predictive crash mitigation countermeasure framework

The two proposed frameworks (i.e. movement-based crash prediction and trend-based crash prediction) are feasible ways to achieve predictive crash mitigation, both methods can provide a lead time for crash prediction, hence prevention strategies could thereby be applied before the risk conditions turn into an actual accident event. Besides, the targeted vehicles with a higher likelihood of crash are able to be predicted based on the risk exposures and risk trends assessed from driving behaviour features. To mitigate crashes in advance, countermeasures should be taken at the pre-accident risk stages, by correcting the failure mechanism causing the crash, and hopefully contributes towards achieving the near zero-fatality goal.

6.7 Chapter Summary

This chapter designs the AutoML methods for risk prediction and the systems for potential applications, including risk decision-making for AVs, PHYD insurance,

driving safety system under CV environment, and short-term crash prediction, which generates contributions to associated domains in the following three aspects.

(1) A domain-specific AutoML is built based on XGBoost, which enables end-to-end learning from the driving behaviour data to detailed risk levels and corresponding key features. The AutoML integrates the main components of risk prediction modelling into an auto-optimisable pipeline, including unsupervised risk identification by FCM clustering and risk indicators, risk-behaviour feature learning by the hybrid of XGBoost-based filtering and RFE, imbalanced data resampling, model selection and hyperparameter tuning by Bayesian optimisation, among others. Two main mechanisms are designed to improve the modelling performance, one is the massive feature extraction and learning-based selection, another is the hyperparameter auto-tuning. Massive in-depth behaviour features are extracted to capture more useful information on vehicle driving and risk potentials, and the key feature sets with the best modelling performance are self-learned by the AutoML. XGBoost is the key classifier of the AutoML, and interpretable risk decision rules can thus be generated using the tree-based AutoML.

(2) The AutoML is scalable and transparent to achieve accurate prediction of detailed risk levels of vehicles in driving, and assess the unsafe behaviours. AutoML achieves satisfactory results of behaviour-based risk prediction, which has a predictive power of 91% overall accuracy for detailed risk levels (i.e. 4 levels), and greater than 95% accuracy for safe-risk classification. The AutoML platform has a self-learning and auto-optimisation mechanism, which can be easily updated by introducing the most advanced algorithms. Bayesian optimisation guides the self-learning of AutoML by effectively auto-tuning the hyperparameters and exploring the pipeline space for better performance. The identification of key features not only helps to produce better results with fewer computation costs, but also provides data-driven insights about modelling and system design.

(3) Furthermore, application potentials are discussed, and the AutoML can be used in the risk decision-making and motion trajectory planning of AVs and ADAS, PHYD insurance, driving safety system under the connected vehicle environment,

among others. Under the same AutoML framework, the prediction performances of various AV sensing configurations have been compared, which provides data-driven insights about AV safety from the perspective of the information needed for risk decision-making. Moreover, the AutoML reveals the most important behaviour features used for risk assessment, which uncovers useful insights about sensing data processing. These studies contribute to traffic safety by providing a portfolio of techniques, ranging from vehicle-level crash risk prediction to personalised driving behaviour enhancement, which enables the development of effective measures and systems to reduce the likelihood of crashes. The frameworks of risk-based accident prediction and inference model are sketched. Besides, pertinent data acquisition and system integration are also discussed. The main benefits are from the countermeasures of vehicle crash risks. Herein, the innovative techniques developed in this research are thus discussed in an application-centric context, and a brief summary of the main techniques and application potentials is depicted in Figure 6.16.

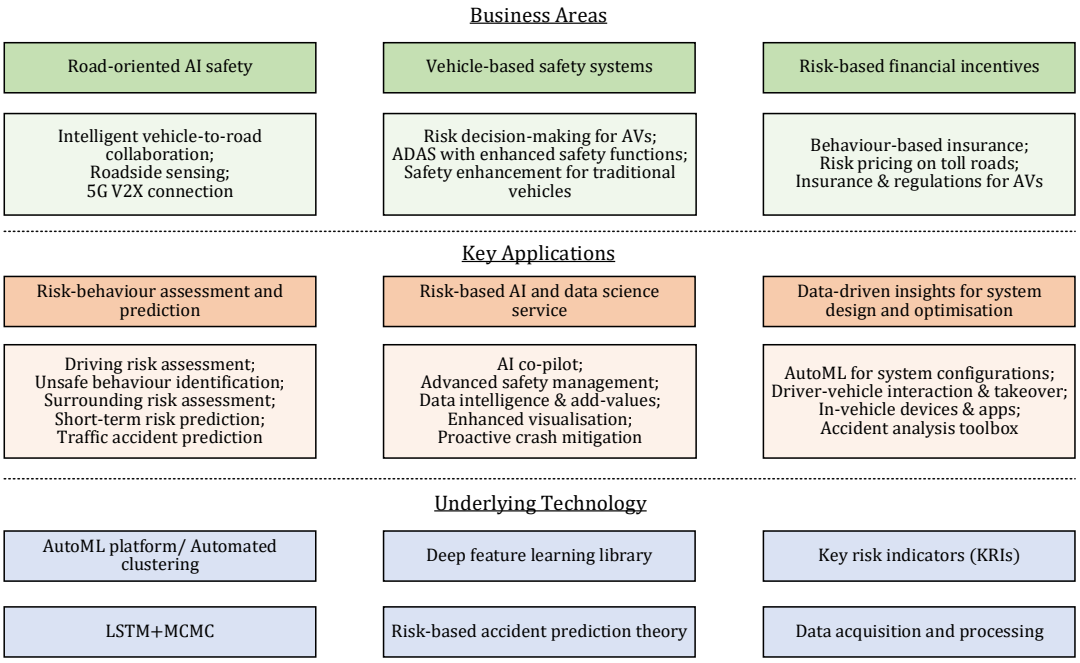


Figure 6.16 Main techniques and application potentials summary

CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS

This concluding chapter includes two main parts, namely, a summary about techniques developed in this study and key contributions, and the future works.

7.1 Summary of Techniques

7.1.1 Surrogate key risk indicators

Accident events are generally unexpected and occur rarely. Pre-accident risk assessment using hybrid and hierarchical indicators (the KRIs) offers new insights about risk exposures, which is an effective way to identify risk levels and thus boost accident prediction. The feasibility of using KRIs to measure pre-accident risk exposures conditioned on real-world accident data has been assessed. Three main findings and contributions are stated as followings:

First, the concept of KRI is formulated to assess risk exposures using hybrid indicators. Seven individual indicators are shortlisted as the basic indicators of KRIs, with evaluation in terms of risk behaviour, risk avoidance, and risk margin. A temporal-spatial case-control study is designed to investigate the feasibility of each indicator with the key findings that TIT can identify many pre-accident risk conditions and CPI can pick out the most severe conditions (the near-accident) from among these risk conditions. In addition, it is found that the impact of different threshold values is not very critical in determining risk levels.

Second, KRI uses hybrid indicators to hierarchically distinguish various risk levels. The expressions of KRIs have been developed mainly based on TIT and CPI, to measure risk severity with three levels, as well as the likelihood. It is also flexible in defining straightforward threshold values for classification of risk levels. The KRIs and their threshold measurements are then further validated by another independent accident case. The KRI-based risk assessment also reveals predictive insights about a potential accident, including at-risk vehicles, locations and time.

Third, a grid remapping method has been developed to obtain vehicle trajectory data

from surveillance video system, which can be applied for coordinates transformation from image pixels to real-world units with high quality. In addition, two real-world (chain-collision) accident events and their antecedent (pre-crash) road traffic movements are retrieved from surveillance video footage, which unveils some insights of pre-accident risks. Hybrid indicators such as KRIs offer new insights about pre-accident risk exposures, which is helpful for accident risk assessment and prediction.

7.1.2 Risk assessment by clustering

Risk grading plays a key role in traffic safety, as it is important in a range of areas, such as, proactive crash prevention, identification of risk determinants, driving behaviour evaluation, among others. There is a perennial quest to develop enhanced solutions to control traffic safety. The risk grading pertains to a distinctly imbalanced problem. In this part, a feature-oriented clustering method is proposed to achieve unsupervised risk grading of a large group of vehicles within a road segment. This method not only contributes to overcoming some intrinsic challenges from imbalanced data, but also produces a trial benchmark for risk mapping and positioning. In summary, this method has contributed to the domain knowledge in three areas, which are stated in the followings.

First, for risk grading and imbalanced clustering, indicator-based features are designed. Based on surrogate measures of vehicle conflicts, interpretable features are extracted from a general vehicle driving trajectory, which represents vehicle risk exposures in terms of temporal, kinematical and spatial aspects. Most of the features rely on threshold values to differentiate between risk and safety, which meanwhile contributes towards overcoming the problems of imbalanced data, for example, to reduce the degree of overlapping and lack of density.

Second, clustering-based risk grading is proposed to achieve unsupervised data labelling of risk levels for a large group of vehicles. To obtain reliable and robust partitioning, ensemble clustering is built by majority voting of the outcomes produced by multiple clustering. Based on risk pattern similarity, vehicles are clustered into distinct groups with graded risk ratings. Clustering is conducted in a progressive

manner to obtain hierarchical partitioning, which facilitates to identify the highest risk level, and meanwhile addresses the small disjuncts of imbalanced data. Furthermore, label identification by classifiers is proposed to evaluate clustering performance, and guide the selection of the best clustering solution. Besides, key features are identified using feature importance ranking by random forest. Herein, trajectory data from the NGSIM Program is used as a case study, and risk grading with six hierarchical levels is established.

Third, an experimental benchmark for data-driven risk grading is investigated. To delineate the risk potentials, a high-resolution risk mapping and positioning is demonstrated based on NGSIM data. The underlying quest is to figure out the risk potentials in terms of targeted vehicles, locations in metre-scale, timestamps in sub-second interval. In addition, the results also generate a better understanding of risk patterns (e.g. severity, frequency, trends), which may offer predictive insights about crash potentials. The proposed method is effective to assess and identify detailed risk potentials from driving behaviour as exhibited by the general vehicle trajectory.

7.1.3 Risk-behaviour feature learning

Accident occurrence is a complex mechanism, with many contributing factors. Generally, driver-centric factors could be found in most crash accidents, and driving behaviour assessment is an important aspect to enhance safety and reduce crashes. Data mining on driving behaviour is key to driving assessment and risk prediction. In this part, a machine learning based approach is designed to select driving behaviour features and predict risk levels, which combines supervised feature selection and unsupervised risk labelling. This method contributes to the safety domain and associated literature in three areas, which are stated as follows:

First, massive features are extracted from vehicle trajectory. To mine information about driving behaviours, for individual vehicles, more than a thousand driving behaviour features are extracted from trajectory data comprehensively, which produce in-depth and multi-view measures on behaviours. The feature extraction procedure is scalable for trajectory time series data. Besides, risk indicator features are also extracted based on surrogate safety measures.

Second, an integrated feature learning framework is developed. The framework combines learning-based feature selection, clustering-based risk rating and imbalanced data labelling, and imbalanced data resampling. To estimate risk potentials of vehicles in driving, unsupervised risk rating is conducted on a large group of vehicles based on extracted risk indicator features. Vehicles are grouped into clusters using fuzzy C means (FCM), and four risk levels are obtained for data labelling. Besides, data under-sampling of the safe class is performed to amend the biased results derived from class imbalance.

Third, key features are identified by importance ranking and recursive elimination. The linkages between behaviour features and corresponding risk levels are built using XGBoost. The weigh-based and gain-based relative importance are ranked using XGBoost, which produces a filtered feature space. Next, RFE is performed to select an optimal feature subset, and 32 key behaviour features are identified.

7.1.4 Behaviour-based risk prediction

Reliable accident prediction and proactive prevention are undoubtedly of great benefit and necessity. Using the integrated feature learning framework, crash risk levels of vehicles in driving can be predicted based on the selected key behaviour features. In this part, satisfactory results of behaviour-based risk prediction by XGBoost, and the predictive power with an accuracy of 91.66% is achieved. The approach is effective and reliable to identify important features for driving assessment, and contributes to guiding the direction of feature engineering, which is the key to improve modelling.

Behaviour-based risk prediction is a creative and feasible way to identify the targeted vehicles with risk exposures, which is the precursor for crash prediction. The risk exposures and conflict severity are predicted dynamically from vehicle stream movement, and the likelihood of an impending accident can be inferred, as well as the involved vehicles, potential locations and timestamps, etc.

Moreover, the entire methods, from surrogate indicators to feature learning, can contribute to a range of applications, and be designed as the prototypes of several

novel solutions for traffic safety, such as, predictive crash risk mitigation, ‘pay how you drive’ insurance, driving behaviour analytics modules, among others. In addition, data acquisition and collection techniques are also discussed.

7.1.5 AutoML and application potentials

A domain-specific automated machine learning (AutoML) is built based on XGBoost, which enables end-to-end learning from the driving behaviour data to detailed risk levels and corresponding key features. The AutoML assembles all necessary machine learning steps as an end-to-end pipeline and automates the pipeline to get the features, models, and hyperparameters that return the best performance as measured on validation sets. The AutoML integrates the main components of risk prediction modelling into an auto-optimisable pipeline, including unsupervised risk identification by FCM clustering and risk indicators, risk-behaviour feature learning by the hybrid of XGBoost-based filtering and RFE, imbalanced data resampling, model selection and hyperparameter tuning by Bayesian optimisation, among others. Besides, the pipeline also incorporates pre-processing components such as imbalanced data resampling, noise filtering, among others.

The AutoML is scalable and transparent to achieve accurate prediction of detailed risk levels of vehicles in driving, and assess the unsafe behaviours. AutoML achieves satisfactory results of behaviour-based risk prediction, which has a predictive power of 91% overall accuracy for detailed risk levels (i.e. 4 levels), and greater than 95% accuracy for safe-risk classification. The AutoML platform has a self-learning and auto-optimisation mechanism, which can be easily updated by introducing the most advanced algorithms. Bayesian optimisation guides the self-learning of AutoML by effectively auto-tuning the hyperparameters and exploring the pipeline space for better performance. The identification of key features not only helps to produce better results with fewer computation costs, but also provides data-driven insights about modelling and system design.

Furthermore, application potentials are discussed, and the AutoML can be used in the risk decision-making and motion trajectory planning of autonomous vehicles (AVs) and ADAS (advanced driver assistance systems), pay-how-you-drive (PHYD)

insurance, driving safety system under the connected vehicle environment, among others. These studies contribute to traffic safety by providing a portfolio of techniques, ranging from vehicle-level crash risk prediction to personalised driving behaviour enhancement, which enables the development of effective measures and systems to reduce the likelihood of crashes.

7.2 Future Works

The directions for further studies include:

(1) Indicator construction and risk data mining

This study provides a proof of concept of the KRI from an experimental standpoint based on real-world accident cases. As a next step, an in-depth study using more accident cases is recommended, to get a stronger validation and build more powerful indicators. The scope of the technique can be extended by applying the method to additional real-world accidents cases, including various types of accidents and near-miss cases collected from expressways, intersections, arterial roads, etc. More accident cases can be analysed to allow examination of the sensitivity and specificity, the false positive rates and false negative rates, and d-primed value, etc. In addition, appropriate threshold values can be defined to further subdivide risk levels.

Naturalistic driving studies (NDS) can be used to improve indicator construction and data mining. In NDS, crash, near-crash, and safety-critical events are identified based on kinematic and video records, and various characteristics and variables about both driving and drivers' behaviours are collected. Besides, some NDS data sources are available to access, such as SHRP2 (the second Strategic Highway Research Program) data. In addition, the benchmark of risk measures and threshold values could be enhanced using data set collected over a longer period of time. Moreover, there is a lack of leading indicators for accident prediction. Leading indicators should continue to be developed with the aim of proactive prevention with a lead time.

For clustering-based risk grading, the major challenge is the lack of crash instances, which makes it hard to verify the linkages between the highest risk level and actual

crash occurrence. Risk levels are obtained based on the comparison within the sample, but there are no crash cases in the sample. On this aspect, a potential way is to perform clustering on NDS data and examine the risk records of the drivers/vehicles which are grouped as higher risk levels.

(2) Deep-level feature extraction and selection

High-quality features are the key to improve modelling. In-depth feature extraction is recommended to further improve the results, which should cover a wider range of driving behaviours and risk conditions, such as lane-changing, conflicts between motorcycles and vehicles, risk response, and other time-series characteristics. Currently, to grade risk levels and mine more information about driving behaviours, more than a thousand features have been considered comprehensively, and important features are identified based on feature learning. These features are far from being exhaustive. The interests of feature extraction are mainly twofold, namely, making risk assessment more robust, and providing early signals for risk-based crash prediction.

Due to the extreme requirement in risk analysis, unsupervised and semi-supervised feature learning could be used to obtain deep-level features, which may bring in more information than existing hand-crafted features. The advantage of unsupervised learning is that plentiful information can be extracted without manual intervention, such as clustered factors, deep-level structure of features, complex spatial and temporal relationship, etc. Furthermore, based on current unsupervised procedure, semi-supervised data fusion can be built, which uses certain labelled instances to refine modelling. Future works about feature learning are concentrated on using unsupervised and semi-supervised learning algorithms, which learn and decode the representations from time-series data, instead of being hand-crafted.

(3) Prototypes development and test-bedding

The key techniques on risk assessment and prediction can be applied in several areas, and some proof-of-principle prototypes and key functional aspects are designed preliminarily. As a next step, the focus is on developing workable prototypes with

more mature machine learning techniques.

Key aspects include: (1) the hybrid modelling of AutoML and deep learning, namely, using the key features selected in AutoML as the input for deep learning; (2) in-depth research on ensemble clustering and resampling methods for imbalanced data; (3) next n-seconds short-time vehicle trajectory and risk prediction using LSTM and MCMC; (4) update the AutoML system with the state-of-the-art algorithms newly developed; adapt the AutoML models for various application needs and driving conditions; enhance in-depth feature extraction; extend to cover a broader range of driving behaviours and risk conditions, such as lane-changing, conflicts between motorcycles and vehicles; (5) refine the techniques and systems to achieve the required performance in real-world applications and massive deployment, such as technologies test-bedding with on-road data collection and sensor fusion, feasibility demonstration of an end-to-end solution, as well as product commercialisation planning; and (6) continue to introduce new concepts and innovative practices on safety and accident research, such as, build strong risk indicators, conduct crash experiments to calibrate indicator threshold values using driving simulators, verify the linkages between the highest-risk level and actual crash occurrence, etc.

REFERENCES

Abdel-Aty, M. and Haleem, K. (2011), "Analyzing Angle Crashes at Unsignalized Intersections Using Machine Learning Techniques", Accident Analysis and Prevention, Vol. 43, No. 1, pp. 461-470.

Abdel-Aty, M. and Pande, A. (2005), "Identifying Crash Propensity Using Specific Traffic Speed Conditions", Journal of Safety Research, Vol. 36, No. 1, pp. 97-108.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F. and Hsia, L. (2004), "Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression", Transportation Research Record: Journal of the Transportation Research Board, No. 1897, pp. 88-95.

Ahmed, M. and Abdel-Aty, M. (2013), "A Data Fusion Framework for Real-Time Risk Assessment on Freeways", Transportation Research Part C: Emerging Technologies, Vol. 26, pp. 203-213.

Ahmed, M., Abdel-Aty, M. and Yu, R. (2012), "Assessment of Interaction of Crash Occurrence, Mountainous Freeway Geometry, Real-Time Weather, and Traffic Data", Transportation Research Record: Journal of the Transportation Research Board, No. 2280, pp. 51-59.

Ahmed, M. and Abdel-Aty, M. (2012), "The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction", IEEE Transactions on Intelligent Transportation Systems, Vol. 13, No. 2, pp. 459-468.

Ahrens, W. and Pigeot, I. (2014), Handbook of Epidemiology, Springer, New York, N.Y.

Allen, B. L., Shin, B. T. and Cooper, P. J. (1978), "Analysis of Traffic Conflicts and Collisions", Transportation Research Record: Journal of the Transportation Research Board, No. 667, pp. 67–74.

Almqvist, S., Hydén, C. and Risser, R. (1991), "Use of Speed Limiters in Cars for

Increased Safety and A Better Environment", Transportation Research Record: Journal of the Transportation Research Board, No. 1318.

Altché, F., and de La Fortelle, A. (2017), "An LSTM Network for Highway Trajectory Prediction", IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 353-359.

American Association of State Highway and Transportation Officials (AASHTO), (2004), A Policy on Geometric Design of Highways and Streets, Washington D.C. pp. 263–266.

Amundsen, F. H. and Hydén, C. (1977), Proceedings of First Workshop on Traffic Conflicts, Oslo, Institute of Transport Economics and Lunds Tekniska Högskola, Sweden.

Archer, J. (2005), "Indicators for Traffic Safety Assessment and Prediction and Their Application in Micro-Simulation Modelling: A Study of Urban and Suburban Intersections", Ph.D. Thesis, KTH Royal Institute of Technology in Stockholm.

Bagdadi, O. (2013), "Assessing Safety Critical Braking Events in Naturalistic Driving Studies", Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 16, pp. 117-126.

Bailly, K. and Milgram, M. (2009), "Boosting Feature Selection for Neural Network Based Regression", Neural Networks, Vol. 22, No. 5, pp. 748-756.

Barraclough, P., Af Wåhlberg, A., Freeman, J., Watson, B. and Watson, A. (2016), "Predicting Crashes Using Traffic Offences. A Meta-Analysis That Examines Potential Bias Between Self-Report and Archival Data", Plos One, Vol. 11, No. 4, e0153390.

Batista, G. E., Prati, R. C. and Monard, M. C. (2004), "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data", ACM SIGKDD Explorations Newsletter, Vol. 6, No. 1, pp. 20-29.

Bergel-Hayat, R., Debbarh, M., Antoniou, C. and Yannis, G. (2013), "Explaining the

Road Accident Risk: Weather Effects", Accident Analysis and Prevention, Vol. 60, pp. 456-465.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011), "Algorithms for Hyper-Parameter Optimization", Advances in Neural Information Processing Systems, pp. 2546-2554.

Berriel, R. F., de Aguiar, E., de Souza, A. F. and Oliveira-Santos, T. (2017), "Ego-Lane Analysis System (ELAS): Dataset and Algorithms", Image and Vision Computing, Vol. 68, pp. 64-75.

Beyan, C. and Fisher, R. (2015), "Classifying Imbalanced Data Sets Using Similarity Based Hierarchical Decomposition", Pattern Recognition, Vol. 48, No. 5, pp. 1653-1672.

Bian, Y., Yang, C., Zhao, J. L. and Liang, L. (2018), "Good Drivers Pay Less: A Study of Usage-Based Vehicle Insurance Models", Transportation Research Part A: Policy and Practice, Vol. 107, pp. 20-34.

Bierlaire, M. (2015), "Simulation and Optimization: A Short Review", Transportation Research Part C: Emerging Technologies, Vol. 55, pp. 4-13.

Birrell, S. A. and Fowkes, M. (2014), "Glance Behaviours When Using an In-Vehicle Smart Driving Aid: A Real-World, On-Road Driving Study", Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 22, pp. 113-125.

Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, New York, NY.

Bloedorn, E. and Michalsi, R. S. (1998), "Data-Driven Constructive Induction", IEEE Intelligent Systems and Their Applications, Vol. 13, No. 2, pp. 30-37.

Breiman, L. (2001), "Random Forests", Machine Learning, Vol. 45, No. 1, pp: 5–32.

Castignani, G., Derrmann, T., Frank, R. and Engel, T. (2015), "Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring", IEEE

Intelligent Transportation Systems Magazine, Vol. 7, No. 1, pp. 91-102.

Chai, C. and Wong, Y. D. (2013), "Automatic Vehicle Classification and Tracking Method for Vehicle Movements at Signalized Intersections", IEEE Intelligent Vehicles Symposium IV, pp. 624-629.

Chai, C. and Wong, Y. D. (2014), "Micro-simulation of Vehicle Conflicts Involving Right-Turn Vehicles at Signalized Intersections Based on Cellular Automata", Accident Analysis and Prevention, Vol. 63, pp. 94-103.

Chai, C. and Wong, Y. D. (2015a), "Comparison of Two Simulation Approaches to Safety Assessment: Cellular Automata and SSAM", Journal of Transportation Engineering, Vol. 141, No. 6, 05015002.

Chai, C. and Wong, Y. D. (2015b), "Fuzzy Cellular Automata Model for Signalized Intersections", Computer-Aided Civil and Infrastructure Engineering, Vol. 30, No. 12, pp. 951-964.

Chai, C., Wong, Y. D. and Wang, X. (2017), "Safety Evaluation of Driver Cognitive Failures and Driving Errors on Right-Turn Filtering Movement at Signalized Road Intersections Based on Fuzzy Cellular Automata (FCA) Model", Accident Analysis and Prevention, Vol. 104, pp. 156-164.

Chan, C. Y. (2006), "Defining Safety Performance Measures of Driver-Assistance Systems for Intersection Left-Turn Conflicts", IEEE Intelligent Vehicles Symposium, pp. 25-30.

Chang, L. Y. (2005), "Analysis of Freeway Accident Frequencies: Negative Binomial Regression Versus Artificial Neural Network", Safety Science, Vol. 43, No. 8, pp. 541-557.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002), "SMOTE: Synthetic Minority Over-Sampling Technique", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357.

Chen, T. and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System",

Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

Chin, H. C. and Quek, S. T. (1997), "Measurement of Traffic Conflicts", Safety Science, Vol. 26, No. 3, pp. 169-185.

Cooper, D. F. and Ferguson, N. (1976), "Traffic Studies at T-Junctions. 2. A Conflict Simulation Record", Traffic Engineering and Control, Vol. 17, pp. 306–309.

Cunto, F. and Saccomanno, F. F. (2008), "Calibration and Validation of Simulated Vehicle Safety Performance at Signalized Intersections", Accident Analysis and Prevention, Vol. 40, No. 3, pp. 1171-1179.

Cunto, F. J. C. and Saccomanno, F. F. (2007), "Microlevel Traffic Simulation Method for Assessing Crash Potential at Intersections", Transportation Research Board 86th Annual Meeting, No. 07-2180.

Dallat, C., Salmon, P. M. and Goode, N. (2017), "Risky Systems Versus Risky People: to What Extent Do Risk Assessment Methods Consider the Systems Approach to Accident Causation? A Review of the Literature", Safety Science, http://dx.doi.org/10.1016/j.ssci.2017.03.012.

Dezman, Z., De Andrade, L., Vissoci, J. R., El-Gabri, D., Johnson, A., Hirshon, J. M. and Staton, C. A. (2016), "Hotspots and Causes of Motor Vehicle Crashes in Baltimore, Maryland: A Geospatial Analysis of Five Years of Police Crash and Census Data", Injury, Vol. 47, No.11, pp. 2450-2458.

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. I. and Kuncheva, L. I. (2015), "Diversity Techniques Improve the Performance of the Best Imbalance Learning Ensembles", Information Sciences, Vol. 325, pp. 98-117.

Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S. (2013), "Similarity-Based Machine Learning Methods for Predicting Drug–Target Interactions: A Brief Review", Briefings in Bioinformatics, Vol. 15, No. 5, pp. 734-747.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M. and

Hankey, J. (2016), "Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data", Proceedings of the National Academy of Sciences, Vol. 113, No. 10, pp. 2636-2641.

Dong, N., Huang, H. and Zheng, L. (2015), "Support Vector Machine in Crash Prediction at the Level of Traffic Analysis Zones: Assessing the Spatial Proximity Effects", Accident Analysis and Prevention, Vol. 82, pp. 192-198.

Dong, Y., Hu, Z., Uchimura, K. and Murayama, N. (2011), "Driver Inattention Monitoring System for Intelligent Vehicles: A Review", IEEE Transactions on Intelligent Transportation Systems, Vol. 12, No. 2, pp. 596-614.

Eboli, L., Mazzulla, G. and Pungillo, G. (2016), "Combining Speed and Acceleration to Define Car Users' Safe or Unsafe Driving Behaviour", Transportation Research Part C: Emerging Technologies, Vol. 68, pp. 113-125.

Eftekhari, H. R. and Ghatee, M. (2018), "Hybrid of Discrete Wavelet Transform and Adaptive Neuro Fuzzy Inference System for Overall Driving Behavior Recognition", Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 58, pp. 782-796.

Eggensperger, K., Lindauer, M., and Hutter, F. (2019), "Pitfalls and Best Practices in Algorithm Configuration", Journal of Artificial Intelligence Research, Vol. 64, pp. 861-893.

El-Basyouny, K. and Sayed, T. (2013), "Safety Performance Functions Using Traffic Conflicts", Safety Science, Vol. 51, No. 1, pp. 160-164.

Elkan, C. (2001), "The Foundations of Cost-Sensitive Learning", International Joint Conference on Artificial Intelligence, Vol. 17, No. 1, pp. 973-978.

Essa, M. and Sayed, T. (2015), "Transferability of Calibrated Microsimulation Model Parameters for Safety Assessment Using Simulated Conflicts", Accident Analysis and Prevention, Vol. 84, pp. 41-53.

Evans, L. (1991), Traffic Safety and the Driver, Van Nostrand Reinhold, New York.

Fahad, A., Tari, Z., Khalil, I., Almalawi, A. and Zomaya, A. Y. (2014), "An Optimal and Stable Feature Selection Approach for Traffic Classification Based on Multi-Criterion Fusion", Future Generation Computer Systems, Vol. 36, pp. 156-169.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015), "Efficient and Robust Automated Machine Learning", Advances in Neural Information Processing Systems, pp. 2962-2970.

Feurer, M., Springenberg, J. T., and Hutter, F. (2014), "Using Meta-Learning to Initialize Bayesian Optimization of Hyperparameters", Proceedings of the 2014 International Conference on Meta-learning and Algorithm Selection, Vol. 1201, pp. 3-10.

Garcia, F., Cerri, P., Broggi, A., De La Escalera, A. and Armingol, J. M. (2012), "Data Fusion for Overtaking Vehicle Detection Based on Radar and Optical Flow", IEEE Intelligent Vehicles Symposium (IV), pp. 494-499.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. and Herrera, F. (2016), "Big Data Preprocessing: Methods and Prospects", Big Data Analytics, Vol. 1, No. 1, pp. 9.

García, V., Sánchez, J. S. and Mollineda, R. A. (2012), "On the Effectiveness of Preprocessing Methods When Dealing with Different Levels of Class Imbalance", Knowledge-Based Systems, Vol. 25, No. 1, pp. 13-21.

Guido, G., Saccomanno, F., Vitale, A., Astarita, V. and Festa, D. (2010), "Comparing Safety Performance Measures Obtained from Video Capture Data", Journal of Transportation Engineering, Vol. 137, No. 7, pp. 481-491.

Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N. Y., Huang, R. and Zhou, X. (2015), "Mobile Crowd Sensing and Computing: The Review of An Emerging Human-Powered Sensing Paradigm", ACM Computing Surveys, Vol. 48, No. 1, pp. 7.

Guo, F. and Fang, Y. (2013), "Individual Driver Risk Assessment Using Naturalistic Driving Data", Accident Analysis and Prevention, Vol. 61, pp. 3-9.

Guo, F., Klauer, S., Hankey, J. and Dingus, T. (2010), "Near Crashes as Crash Surrogate for Naturalistic Driving Studies", Transportation Research Record: Journal of the Transportation Research Board, No. 2147, pp. 66-74.

Guo, J., Song, B., He, Y., Yu, F. R. and Sookhak, M. (2017), "A Survey on Compressed Sensing in Vehicular Infotainment Systems", IEEE Communications Surveys and Tutorials, Vol. 19, Iss. 4, pp. 2662-2680.

Guyon, I. and Elisseeff, A. (2003), "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, Vol. 3, pp. 1157-1182.

Guyon, I. and Elisseeff, A. (2006), "An Introduction to Feature Extraction", Feature Extraction, Springer, Berlin, Heidelberg, pp. 1-25.

Hakkert, A. S. and Gitelman, V. (2014), "Thinking About the History of Road Safety Research: Past Achievements and Future Challenges", Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 25, pp. 137-149.

Hankey, J. M., Perez, M. A. and McClafferty, J. A. (2016), "Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets", Virginia Tech Transportation Institute.

Hapfelmeier, A. and Ulm, K. (2013), "A New Variable Selection Approach Using Random Forests", Computational Statistics and Data Analysis, Vol. 60, pp. 50-69.

Hecker, S., Dai, D., and Van Gool, L. (2018), "End-to-end Learning of Driving Models with Surround-View Cameras and Route Planners", Proceedings of the European Conference on Computer Vision (ECCV), pp. 435-453.

Hong, J. H., Margines, B. and Dey, A. K. (2014), "A Smartphone-Based Sensing Platform to Model Aggressive Driving Behaviours", Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 4047-4056.

Hormann, K. and Agathos, A. (2001), "The Point in Polygon Problem for Arbitrary Polygons", Computational Geometry, Vol. 20, No. 3, pp. 131-144.

Huang, H., Zeng, Q., Pei, X., Wong, S. C. and Xu, P. (2016), "Predicting Crash Frequency Using an Optimised Radial Basis Function Neural Network Model", Transportmetrica A: Transport Science, Vol. 12, No. 4, pp. 330-345.

Hwang, S., Fraser, J. and Simkins, B. J. (2010), "Identifying and Communicating Key Risk Indicators", Enterprise Risk Management, pp. 125-140.

Hydén, C. (1996), "Traffic Conflicts Technique: State-of-the-Art", Traffic Safety Work with Video-Processing, Vol. 37, 3-14.

Iida, Y., Uno, N., Itsubo, S. and Suganuma, M. (2001), "Traffic Conflict Analysis and Modeling of Lane-Changing Behavior at Weaving Section", Proceedings of Infrastructure Planning, Vol. 24, No. 1, pp. 305-308.

Imprialou, M. and Quddus, M. (2017), "Crash Data Quality for Road Safety Research: Current State and Future Directions", Accident Analysis and Prevention, http://dx.doi.org/10.1016/j.aap.2017.02.022.

Izakian, H., Pedrycz, W. and Jamal, I. (2015), "Fuzzy Clustering of Time Series Data Using Dynamic Time Warping Distance", Engineering Applications of Artificial Intelligence, Vol. 39, pp. 235-244.

Jonasson, J. K. and Rootzén, H. (2014), "Internal Validation of Near-Crashes in Naturalistic Driving Studies: A Continuous and Multivariate Approach", Accident Analysis and Prevention, Vol. 62, pp. 102-109.

Kang, H. B. (2013), "Various Approaches for Driver and Driving Behavior Monitoring: A Review", Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 616-623.

Kasper, D., Weidl, G., Dang, T., Breuel, G., Tamke, A., Wedel, A. and Rosenstiel, W. (2012), "Object-Oriented Bayesian Networks for Detection of Lane Change Maneuvers", IEEE Intelligent Transportation Systems Magazine, Vol. 4, No. 3, pp. 19-31.

Katrakazas, C., Quddus, M., Chen, W. H., and Deka, L. (2015), "Real-Time Motion

Planning Methods for Autonomous On-road Driving: State-of-the-art and Future Research Directions", Transportation Research Part C: Emerging Technologies, Vol. 60, pp. 416-442.

Kecman, V. (2005), "Support Vector Machines–An Introduction", Support Vector Machines: Theory and Applications, pp. 605-605.

Kira, K. and Rendell, L. A. (1992), "A Practical Approach to Feature Selection", Proceedings of the 9th International Workshop on Machine Learning, pp. 249-256.

Kockelman, K. K. and Ma, J. (2010), "Freeway Speeds and Speed Variations Preceding Crashes, Within and Across Lanes", Journal of the Transportation Research Forum, Vol. 46, No. 1.

Koetse, M. J. and Rietveld, P. (2009), "The Impact of Climate Change and Weather on Transport: An Overview of Empirical Findings", Transportation Research Part D: Transport and Environment, Vol. 14, No. 3, pp. 205-221.

Kohonen, T. (2013), "Essentials of the Self-Organizing Map", Neural Networks, Vol. 37, pp. 52-65.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015), "Machine Learning Applications in Cancer Prognosis and Prediction", Computational and Structural Biotechnology Journal, Vol. 13, pp. 8-17.

Kunt, M. M., Aghayan, I. and Noii, N. (2011), "Prediction for Traffic Accident Severity: Comparing the Artificial Neural Network, Genetic Algorithm, Combined Genetic Algorithm and Pattern Search Methods", Transport, Vol. 26, No. 4, pp. 353-366.

Lai, J. C., Huang, S. S. and Tseng, C. C. (2010), "Image-Based Vehicle Tracking and Classification on the Highway", IEEE 2010 International Conference on Green Circuits and Systems (ICGCS), pp. 666-670.

Längkvist, M., Karlsson, L. and Loutfi, A. (2014), "A Review of Unsupervised Feature Learning and Deep Learning for Time-Series Modeling", Pattern

Recognition Letters, Vol. 42, pp. 11-24.

Laureshyn, A., Svensson, Å. and Hydén, C. (2010), "Evaluation of Traffic Safety, Based on Micro-Level Behavioural Data: Theoretical Framework and First Implementation", Accident Analysis and Prevention, Vol. 42, No. 6, pp. 1637-1646.

Laval, J. A. and Leclercq, L. (2010), "A Mechanism to Describe the Formation and Propagation of Stop-And-Go Waves in Congested Freeway Traffic", Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 368, No. 1928, pp. 4519-4541.

Lever, J., Krzywinski, M. and Altman, N. (2016), "Classification Evaluation", Nature Methods, Vol. 13, No. 8, pp. 541–542.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S. and Sokolsky, M. (2011), "Towards Fully Autonomous Driving: Systems and Algorithms", IEEE Intelligent Vehicles Symposium (IV), pp. 163-168.

Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008), "Predicting Motor Vehicle Crashes Using Support Vector Machine Models", Accident Analysis and Prevention, Vol. 40, No. 4, pp. 1611-1618.

Li, Y., Zheng, Y., Wang, J., Kodaka, K. and Li, K. (2018), "Crash Probability Estimation Via Quantifying Driver Hazard Perception", Accident Analysis and Prevention, Vol. 116, pp. 116-125.

Lin, L., Wang, Q. and Sadek, A. W. (2015), "A Novel Variable Selection Method Based on Frequent Pattern Tree for Real-Time Traffic Accident Risk Prediction", Transportation Research Part C: Emerging Technologies, Vol. 55, pp. 444-459.

Lin, W. C., Tsai, C. F., Hu, Y. H. and Jhang, J. S. (2017), "Clustering-Based Undersampling In Class-Imbalanced Data", Information Sciences, Vol. 409, pp. 17-26.

Liu, H. and Yu, L. (2005), "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data

Engineering, Vol. 17, No. 4, pp. 491-502.

Liu, H. and Motoda, H. (2012), "Feature Selection for Knowledge Discovery and Data Mining", Springer Science and Business Media, Vol. 454.

Liu, J., and Khattak, A. J. (2016), "Delivering Improved Alerts, Warnings, and Control Assistance Using Basic Safety Messages Transmitted Between Connected Vehicles", Transportation Research Part C: Emerging Technologies, Vol. 68, pp. 83-100.

Llanos, J., Morales, R., Núñez, A., Sáez, D., Lacalle, M., Marín, L. G., Hernándezc, R. and Lanas, F. (2017), "Load Estimation for Microgrid Planning Based on A Self-Organizing Map Methodology", Applied Soft Computing, Vol. 53, pp. 323-335.

Loh, W. Y. (2011), "Classification and Regression Trees", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1, No. 1, pp.14-23.

López, V., Fernández, A., García, S., Palade, V. and Herrera, F. (2013), "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics", Information Sciences, Vol. 250, pp. 113-141.

Lord, D. and Mannering, F. (2010), "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives", Transportation Research Part A: Policy and Practice, Vol. 44, No. 5, pp. 291-305.

Lu, N., Cheng, N., Zhang, N., Shen, X. and Mark, J. W. (2014), "Connected Vehicles: Solutions and Challenges", IEEE Internet of Things Journal, Vol. 1, No. 4, pp. 289-299.

Lum, H. and Reagan, J. A. (1995), "Interactive Highway Safety Design Model: Accident Predictive Module", Public Roads, Vol. 58, No. 3.

Ma, P. E. (2009), "Bayesian Analysis of Underreporting Poisson Regression Model with An Application to Traffic Crashes on Two-Lane Highways", Transportation Research Board 88th Annual Meeting, No. 09-3192.

Mahmud, S. S., Ferreira, L., Hoque, M. S. and Tavassoli, A. (2017), "Application of Proximal Surrogate Indicators for Safety Evaluation: A Review of Recent Developments and Research Needs", IATSS Research, http://dx.doi.org/10.1016/j.iatssr.2017.02.001.

Mannering, F. L., Shankar, V. and Bhat, C. R. (2016), "Unobserved Heterogeneity and the Statistical Analysis of Highway Accident Data", Analytic Methods in Accident Research, Vol. 11, pp. 1-16.

Mao, X., Inoue, D., Kato, S. and Kagami, M. (2012), "Amplitude-Modulated Laser Radar for Range and Speed Measurement in Car Applications", IEEE Transactions on Intelligent Transportation Systems, Vol. 13, No. 1, pp. 408-413.

Meng, Q. and Qu, X. (2012), "Estimation of Rear-End Vehicle Crash Frequencies in Urban Road Tunnels", Accident Analysis and Prevention, Vol. 48, pp. 254-263.

McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., and Weller, A. V. (2017), "Concrete Problems for Autonomous Vehicle Safety: Advantages of Bayesian Deep Learning", International Joint Conferences on Artificial Intelligence (IJCAI-17), pp. 4745-4753.

Minderhoud, M. M. and Bovy, P. H. (2001), "Extended Time-To-Collision Measures for Road Traffic Safety Assessment", Accident Analysis and Prevention, Vol. 33, No. 1, pp. 89-97.

Morris, B. T. and Trivedi, M. M. (2008), "Learning, Modeling and Classification of Vehicle Track Patterns from Live Video", IEEE Transactions on Intelligent Transportation Systems, Vol. 9, No. 3, pp. 425-437.

Nai, W., Chen, Y., Yu, Y., Zhang, F., Dong, D. and Zheng, W. (2016), "Fuzzy Risk Mode and Effect Analysis Based on Raw Driving Data for Pay-How-You-Drive Vehicle Insurance", IEEE International Conference on Big Data Analysis (ICBDA), pp. 1-5.

Nguyen, H. M., Cooper, E. W. and Kamei, K. (2009), "Borderline Over-Sampling

for Imbalanced Data Classification", Proceedings of Fifth International Workshop on Computational Intelligence and Applications, Vol. 1, pp. 24-29.

Nitsche, P., Thomas, P., Stuetz, R. and Welsh, R. (2017), "Pre-Crash Scenarios at Road Junctions: A Clustering Method for Car Crash Data", Accident Analysis and Prevention, Vol. 107, pp. 137-151.

Ozbay, K., Yang, H., Bartin, B. and Mudigonda, S. (2008), "Derivation and Validation of New Simulation-Based Surrogate Safety Measure", Transportation Research Record: Journal of the Transportation Research Board, No. 2083, pp. 105-113.

Paden, B., Čáp, M., Yong, S. Z., Yershov, D., and Frazzoli, E. (2016), "A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles", IEEE Transactions on Intelligent Vehicles, Vol. 1, No. 1, pp. 33-55.

Paefgen, J., Staake, T. and Thiesse, F. (2013), "Evaluation and Aggregation of Pay-As-You-Drive Insurance Rate Factors: A Classification Analysis Approach", Decision Support Systems, Vol. 56, pp. 192-201.

Patel, M., Lal, S. K. L., Kavanagh, D. and Rossiter, P. (2011), "Applying Neural Network Analysis on Heart Rate Variability Data to Assess Driver Fatigue", Expert Systems with Applications, Vol. 38, No, 6, pp. 7235-7242.

Paul, A., Chilamkurti, N., Daniel, A. and Rho, S. (2016), Intelligent Vehicular Networks and Communications: Fundamentals, Architectures and Solutions, Elsevier, New York, N.Y.

Pei, X., Wong, S.C. and Sze, N.N. (2011), "A Joint-Probability Approach to Crash Prediction Models", Accident Analysis and Prevention, Vol. 43, pp. 1160-1166.

Perez, M. A., Sudweeks, J. D., Sears, E., Antin, J., Lee, S., Hankey, J. M. and Dingus, T. A. (2017), "Performance of Basic Kinematic Thresholds in the Identification of Crash and Near-Crash Events Within Naturalistic Driving Data", Accident Analysis and Prevention, Vol. 103, pp. 10-19.

Piramuthu, S. and Sikora, R. T. (2009), "Iterative Feature Construction for Improving Inductive Learning Algorithms", Expert Systems with Applications, Vol. 36, No. 2, pp. 3401-3406.

Punzo, V., Borzacchiello, M. T. and Ciuffo, B. (2011), "On the Assessment of Vehicle Trajectory Data Accuracy and Application to the Next Generation Simulation (NGSIM), Program Data", Transportation Research Part C: Emerging Technologies, Vol. 19, No. 6, pp. 1243-1262.

Qin, J., Fu, W., Gao, H. and Zheng, W. X. (2017), "Distributed K-Means Algorithm and Fuzzy C-Means Algorithm for Sensor Networks Based on Multiagent Consensus Theory", IEEE Transactions on Cybernetics, Vol. 47, No. 3, pp. 772-783.

Quddus, M. (2013), "Exploring the Relationship Between Average Speed, Speed Variation and Accident Rates Using Spatial Statistical Models and GIS", Journal of Transportation Safety and Security, Vol. 5, No. 1, pp. 27-45.

Rodriguez, A. and Laio, A. (2014), "Clustering by Fast Search and Find of Density Peaks", Science, Vol. 344, No. 6191, pp. 1492-1496.

Roshandel, S., Zheng, Z. and Washington, S. (2015), "Impact of Real-Time Traffic Characteristics on Freeway Crash Occurrence: Systematic Review and Meta-Analysis", Accident Analysis and Prevention, Vol. 79, pp. 198-211.

Sahayadhas, A., Sundaraj, K. and Murugappan, M. (2012), "Detecting Driver Drowsiness Based on Sensors: A Review", Sensors, Vol. 12, No. 12, pp. 16937-16953.

Saifuzzaman, M. and Zheng, Z. (2014), "Incorporating Human-Factors in Car-Following Models: A Review of Recent Developments and Research Needs", Transportation Research Part C: Emerging Technologies, Vol. 48, pp. 379-403.

Salmon, P. M., Walker, G. H., M. Read, G. J., Goode, N. and Stanton, N. A. (2017), "Fitting Methods to Paradigms: Are Ergonomics Methods Fit for Systems Thinking?", Ergonomics, Vol. 60, No. 2, pp. 194-205.

Savolainen, P. T., Mannering, F. L., Lord, D. and Quddus, M. A. (2011), "The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives", Accident Analysis and Prevention, Vol. 43, No. 5, pp. 1666-1676.

Sayed, T., Brown, G. and Navin, F. (1994), "Simulation of Traffic Conflicts at Unsignalized Intersections With TSC-Sim", Accident Analysis and Prevention, Vol. 26, No. 5, pp. 593-607.

Scandizzo, S. (2005), "Risk Mapping and Key Risk Indicators in Operational Risk Management", Economic Notes, Vol. 34, No. 2, pp. 231-256.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2015), "Taking the Human Out of the Loop: A Review of Bayesian Optimization", Proceedings of the IEEE, Vol. 104, No. 1, pp. 148-175.

Shi, X., Wong, Y. D., Li, M. Z. F. and Chai, C. (2018a), "Key Risk Indicators for Accident Assessment Conditioned on Pre-Crash Vehicle Trajectory", Accident Analysis and Prevention, Vol. 117, pp. 346-356.

Shi X., Wong Y.D., Li M.Z.F. and Chai C. (2018b), "Accident Risk Prediction Based on Driving Behavior Feature Learning Using CART and XGBoost", Transportation Research Board 97th Annual Meeting, No. 18-06270.

Singh, S. (2015). "Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey", Traffic Safety Facts Crash Stats Report, No. DOT HS 812 115.

Sikora, R. and Piramuthu, S. (2007), "Framework for Efficient Feature Selection in Genetic Algorithm Based Data Mining", European Journal of Operational Research, Vol. 180, No. 2, pp. 723-737.

Sivaraman, S. and Trivedi, M. M. (2013a), "Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis", IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 4, pp. 1773-1795.

Sivaraman, S. and Trivedi, M. M. (2013b), "Integrated Lane and Vehicle Detection, Localization, and Tracking: A Synergistic Approach", IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 2, pp. 906-917.

Snoek, J., Larochelle, H., and Adams, R. P. (2012), "Practical Bayesian Optimization of Machine Learning Algorithms", Advances in Neural Information Processing Systems, pp. 2951-2959.

So, J. J., Park, B. B. and Yun, I. (2015), "Classification Modeling Approach for Vehicle Dynamics Model-Integrated Traffic Simulation Assessing Surrogate Safety", Journal of Advanced Transportation, Vol. 49, No. 3, pp. 416-433.

Sobhani, A., Young, W. and Sarvi, M. (2013), "A Simulation Based Approach to Assess the Safety Performance of Road Locations", Transportation Research Part C: Emerging Technologies, Vol. 32, pp. 144-158.

Stanton, N. A. and Salmon, P. M. (2009), "Human Error Taxonomies Applied to Driving: A Generic Driver Error Taxonomy and Its Implications for Intelligent Transport Systems", Safety Science, Vol. 47, No. 2, pp. 227-237.

Strobl, C., Boulesteix, A. L., Zeileis, A. and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and A Solution", BMC Bioinformatics, Vol. 8, No. 1, pp. 25.

Sun, J. and Sun, J. (2015), "A Dynamic Bayesian Network Model for Real-Time Crash Prediction Using Traffic Speed Conditions Data", Transportation Research Part C: Emerging Technologies, Vol. 54, pp. 176-186.

Sun, Q. C., Odolinski, R., Xia, J. C., Foster, J., Falkmer, T. and Lee, H. (2017), "Validating the Efficacy of GPS Tracking Vehicle Movement for Driving Behaviour Assessment", Travel Behaviour and Society, Vol. 6, pp. 32-43.

Svensson, A. (1998), "A Method for Analysing the Traffic Process in A Safety Perspective", Dissertation, Lund Institute of Technology.

Sze, N. N., Wong, S. C. and Lee, C. Y. (2014), "The Likelihood of Achieving

Quantified Road Safety Targets: A Binary Logistic Regression Model for Possible Factors", Accident Analysis and Prevention, Vol. 73, pp. 242-251.

Talebpour, A., Mahmassani, H. S. and Hamdar, S. H. (2015), "Modeling Lane-Changing Behavior in A Connected Environment: A Game Theory Approach", Transportation Research Procedia, Vol. 7, pp. 420-440.

Tascikaraoglu, A. and Uzunoglu, M. (2014), "A Review of Combined Approaches for Prediction of Short-Term Wind Speed and Power", Renewable and Sustainable Energy Reviews, Vol. 34, pp. 243-254.

Theofilatos, A. and Yannis, G. (2014), "A Review of the Effect of Traffic and Weather Characteristics on Road Safety", Accident Analysis and Prevention, Vol. 72, pp. 244-256.

Thiemann, C., Treiber, M. and Kesting, A. (2008), "Estimating Acceleration and Lane-Changing Dynamics from Next Generation Simulation Trajectory Data", Transportation Research Record: Journal of the Transportation Research Board, No. 2088, pp. 90-101.

Tomek, I. (1976), "An Experiment with the Edited Nearest-Neighbor Rule", IEEE Transactions on Systems, Man, and Cybernetics, No.6, pp. 448-452.

Uno, N., Iida, Y., Itsubo, S. and Yasuhara, S. (2002), "A Microscopic Analysis of Traffic Conflict Caused by Lane-Changing Vehicle at Weaving Section", Proceedings of the 13th Mini-Euro Conference Handling Uncertainty in Transportation Analysis of Traffic and Transportation Systems, pp. 143-148.

van Craenendonck, T. and Blockeel, H. (2015), "Using Internal Validity Measures to Compare Clustering Algorithms", Benelearn, pp. 1-8.

van der Horst, A. R. A. (1990), "A Time-Based Analysis of Road User Behaviour in Normal and Critical Encounters", Doctoral Dissertation, Delft University of Technology, No. HS-041 255.

Vogel, K. (2003), "A Comparison of Headway and Time to Collision as Safety

Indicators", Accident Analysis and Prevention, Vol. 35, No. 3, pp. 427-433.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F. and Fouilloy, A. (2017), "Machine Learning Methods for Solar Radiation Forecasting: A Review", Renewable Energy, Vol. 105, pp. 569-582.

Wahlström, J., Skog, I. and Händel, P. (2017), "Smartphone-Based Vehicle Telematics: A Ten-Year Anniversary", IEEE Transactions on Intelligent Transportation Systems, Vol. 18, No. 10, pp. 2802-2825.

Wan, J., Liu, J., Shao, Z., Vasilakos, A. V., Imran, M. and Zhou, K. (2016), "Mobile Crowd Sensing for Traffic Prediction in Internet of Vehicles", Sensors, Vol. 16, No.1, pp. 88.

Wang, C., Quddus, M. A. and Ison, S. G. (2013), "The Effect of Traffic and Road Characteristics on Road Safety: A Review and Future Research Direction", Safety Science, Vol. 57, pp. 264-275.

Wang, J., Zhang, L., Zhang, D. and Li, K. (2013), "An Adaptive Longitudinal Driving Assistance System Based on Driver Characteristics", IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 1, pp. 1-12.

Wang, L., Abdel-Aty, M., Shi, Q. and Park, J. (2015), "Real-Time Crash Prediction for Expressway Weaving Segments", Transportation Research Part C: Emerging Technologies, Vol. 61, pp. 1-10.

Wang, J., Wu, J., Zheng, X., Ni, D., and Li, K. (2016), "Driving Safety Field Theory Modeling and Its Application in Pre-Collision Warning System", Transportation Research Part C: Emerging Technologies, Vol. 72, pp. 306-324.

Wang, W., Zhang, W., Guo, H., Bubb, H. and Ikeuchi, K. (2011), "A Safety-Based Approaching Behavioural Model with Various Driving Characteristics", Transportation Research Part C: Emerging Technologies, Vol. 19, No. 6, pp. 1202-1214.

Watson, A., Watson, B. and Vallmuur, K. (2015), "Estimating Under-Reporting of

Road Crash Injuries to Police Using Multiple Linked Data Collections", Accident Analysis and Prevention, Vol. 83, pp. 18-25.

Wilson, D. L. (1972), "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data", IEEE Transactions on Systems, Man, and Cybernetics, No. 3, pp. 408-421.

Wong, W. and Wong, S.C. (2016), "Evaluation of The Impact of Traffic Incidents Using GPS Data", Proceedings of the Institution of Civil Engineers - Transport, Vol. 169, pp. 148-162.

Wu, K. F. and Jovanis, P. P. (2012), "Crashes and Crash-Surrogate Events: Exploratory Modeling with Naturalistic Driving Data", Accident Analysis and Prevention, Vol. 45, pp. 507-516.

Wu, K. F., Aguero-Valverde, J. and Jovanis, P. P. (2014), "Using Naturalistic Driving Data to Explore the Association Between Traffic Safety-Related Events and Crash Risk at Driver Level", Accident Analysis and Prevention, Vol. 72, pp. 210-218.

Wu, K. F. and Jovanis, P. P. (2013), "Defining and Screening Crash Surrogate Events Using Naturalistic Driving Data", Accident Analysis and Prevention, Vol. 61, pp. 10-22.

Xia, Y., Liu, C., Li, Y. and Liu, N. (2017), "A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring", Expert Systems with Applications, Vol. 78, pp. 225-241.

Xie, Y., Lord, D. and Zhang, Y. (2007), "Predicting Motor Vehicle Collisions Using Bayesian Neural Network Models: An Empirical Analysis", Accident Analysis and Prevention, Vol. 39, No. 5, pp. 922-933.

Xu, C., Tarko, A. P., Wang, W. and Liu, P. (2013), "Predicting Crash Likelihood and Severity on Freeways with Real-Time Loop Detector Data", Accident Analysis and Prevention, Vol. 57, pp. 30-39.

Xu, H., Gao, Y., Yu, F., and Darrell, T. (2017), "End-to-end Learning of Driving

Models from Large-Scale Video Datasets", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2174-2182.

Xue, S., Lu, J., Wu, J., Zhang, G. and Xiong, L. (2016), "Multi-Instance Graphical Transfer Clustering for Traffic Data Learning", IEEE International Joint Conference on Neural Networks (IJCNN), pp. 4390-4395.

Yang, G., Lin, Y. and Bhattacharya, P. (2010), "A Driver Fatigue Recognition Model Based on Information Fusion and Dynamic Bayesian Network", Information Sciences, Vol. 180, No. 10, pp. 1942-1954.

Yang, H., Ozbay, K. and Bartin, B. (2010), "Application of Simulation-Based Traffic Conflict Analysis for Highway Safety Evaluation", Proceedings of the 12th WCTR, Lisbon, Portugal.

Young, M. S., Birrell, S. A. and Stanton, N. A. (2011), "Safe Driving in A Green World: A Review of Driver Performance Benchmarks and Technologies to Support Smart Driving", Applied Ergonomics, Vol. 42, No. 4, pp. 533-539.

Young, R. A. (2017), "Talking on A Wireless Cellular Device While Driving: Improving the Validity of Crash Odds Ratio Estimates in the SHRP 2 Naturalistic Driving Study", Safety, Vol. 3, No. 4, pp. 28.

Young, W., Sobhani, A., Lenné, M. G. and Sarvi, M. (2014), "Simulation of Safety: A Review of the State of the Art in Road Safety Simulation Modelling", Accident Analysis and Prevention, Vol. 66, pp. 89-103.

Yu, R. and Abdel-Aty, M. (2014), "Using Hierarchical Bayesian Binary Probit Models to Analyze Crash Injury Severity on High Speed Facilities with Real-Time Traffic Data", Accident Analysis and Prevention, Vol. 62, pp. 161-167.

Yu, R., Abdel-Aty, M. and Ahmed, M. (2013), "Bayesian Random Effect Models Incorporating Real-Time Weather and Traffic Data to Investigate Mountainous Freeway Hazardous Factors", Accident Analysis and Prevention, Vol. 50, pp. 371-376.

Yu, Z., Wang, D., You, J., Wong, H. S., Wu, S., Zhang, J. and Han, G. (2016), "Progressive Subspace Ensemble Learning", Pattern Recognition, Vol. 60, pp. 692-705.

Zarshenas, A. and Suzuki, K. (2016), "Binary Coordinate Ascent: An Efficient Optimization Technique for Feature Subset Selection for Machine Learning", Knowledge-Based Systems, Vol. 110, pp. 191-201.

Zeng, D., Liu, K., Lai, S., Zhou, G. and Zhao, J. (2014), "Relation Classification Via Convolutional Deep Neural Network", Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Dublin, Ireland, pp. 2335-2344.

Zhang, L. and Suganthan, P. N. (2014), "Random Forests with Ensemble of Feature Spaces", Pattern Recognition, Vol. 47, No. 10, pp. 3429-3437.

Zhang, Z. (2000), "A Flexible New Technique for Camera Calibration", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, pp. 1330-1334.

Zhang, Z. (2004), "Camera Calibration with One-Dimensional Objects", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 7, pp. 892-899.

Zheng, L., Ismail, K. and Meng, X. (2014), "Traffic Conflict Techniques for Road Safety Analysis: Open Questions and Some Insights", Canadian Journal of Civil Engineering, Vol. 41, No. 7, pp. 633-641.

Zheng, Z. (2012), "Empirical Analysis on Relationship Between Traffic Conditions and Crash Occurrences", Procedia-Social and Behavioral Sciences, Vol. 43, pp. 302-312.

Zheng, Z., Ahn, S. and Monsere, C. M. (2010), "Impact of Traffic Oscillations on Freeway Crash Occurrences", Accident Analysis and Prevention, Vol. 42, No. 2, pp. 626-636.

APPENDICES

Appendix A. Terminologies

Table A.1 Terminologies of surrogate risk indicators

Acronym	Definition	Explanation	References
CPI	Crash Potential Index	The probability that a given vehicle DRAC exceeds its maximum available deceleration rate (MADR) or braking capability during a given time interval	Cunto and Saccomanno, 2008
DRAC	Deceleration Rate to Avoid Crash	Differential speed between a following vehicle and corresponding lead vehicle divided by closing time	Almqvist et al., 1991
MADR	Maximum Available Deceleration Rate	Braking capacity defined by several factors	Cunto and Saccomanno, 2008
PICUD	Potential Index for Collision with Urgent Deceleration	The distance between two consecutive vehicles when they abruptly break and stop completely	Uno et al., 2003
PSD	Proportion of Stopping Distance	The ratio of the remaining distance to the potential collision point and the minimum acceptable stopping distance	Allen et al., 1978
TET	Time Exposed TTC	The length of time a TTC-event remains within a designated TTC-threshold	Minderhoud and Bovy, 2001
TIT	Time Integrated TTC	The integral of the TTC-profile during the time within the threshold	Minderhoud and Bovy, 2001
TTC	Time to Collision	The time until a collision between two vehicles would have occurred if the collision course and speed difference are maintained	van der Horst, 1990

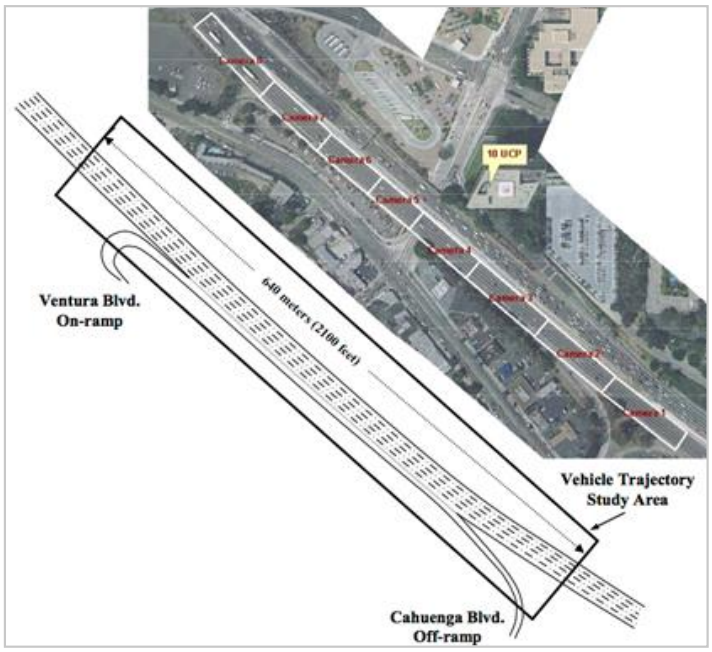
Appendix B. Description and Analysis of NGSIM Data

The Next Generation Simulation (NGSIM) US Route 101 Dataset is used for prediction model development and algorithm test in this study. NGSIM data is widely used as the main dataset in a range of studies, including simulation (Punzo et al., 2011), safety analysis (Cunto and Saccomanno, 2008), behaviour analysis (Wang et al., 2011). The NGSIM program was initiated by US Federal Highway Administration (FHWA) in the early 2000's, and collected high-quality primary traffic and trajectory data from US 101 to support the research on microscopic traffic modelling and testing of the new algorithms.

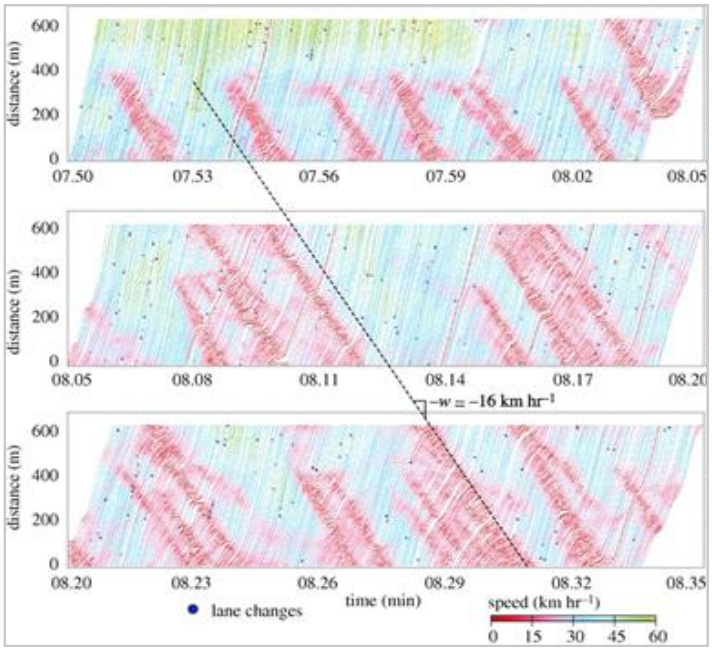
The vehicle trajectory data was collected on the south-bound direction of US 101 (Hollywood Freeway) in Los Angeles, California on June 15th, 2005. The dataset contains transcribed data from 7:50 a.m. to 8:05 a.m., 8:05 a.m. to 8:20 a.m., and 8:20 a.m. to 8:35 a.m. The data was collected using eight video cameras mounted on a 36-storey building located adjacent to the freeway study area. Vehicle trajectory data was transcribed from the video data using a customised software application (NG-VIDEO) developed for NGSIM. The data provides trajectory coordinates of each vehicle, at every 0.1 seconds in relative space and coordinate system.

Main data includes: (1) vehicle identification number (ascending by time of entry into section); (2) lateral (X) coordinate of the front centre of the vehicle with respect to the left-most edge of the road segment in the direction of travel; (3) longitudinal (Y) coordinate of the front centre of the vehicle with respect to the entry edge of the road segment in the direction of travel; (4) vehicle size (length and width) and type (motorcycle, auto, truck); (5) instantaneous variables such as velocity, acceleration, lane position, space headway, time headway in seconds, preceding and following vehicles, etc.

The aerial photograph and schematic drawing about the extent of NGSIM US Route 101 studied area is shown in Figure B.1(a), and the vehicle trajectory from NGSIM dataset is demonstrated in Figure B.1(b) (Laval and Leclercq, 2010).



(a)



(b)

Sources: (a) https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm; and (b) Laval and Leclercq (2010).

Figure B.1 Aerial photograph and vehicle trajectory about NGSIM dataset

Appendix C. XGBoost

XGBoost is a fast, efficient and scalable machine learning system and follows the ideas of Gradient Boosting (GB). GB has been empirically proven to be a highly effective approach for supervised machine learning. Compared with general GB methods, XGBoost optimises the estimation of the objective function by adding a regularisation term based on the loss function. Loss function measures the differences between the real target and prediction value, and the regularisation term penalises the complexity of the model. In XGBoost, a Taylor expansion is employed based on gradient descent to optimise the objective function. Moreover, an approximate greedy search algorithm is introduced to find an optimal tree structure for tree boosting learning, which is faster than exact greedy search used in GB. Besides, hyper-parameters are proposed to improve learning speed and avoid over-fitting (Xia et al., 2017). XGBoost has gained popularity and is employed by numerous winning solutions in machine learning competitions. In Kaggle competition, XGBoost has shown remarkable winning results for a vast array of problems, including prediction, classification and regression.

Appendix D. Vehicle Stream Data by Onsite Recording

Onsite video recording is a direct observation of vehicle stream and accident situations, and has been successfully and widely applied (van der Horst et al., 2014; Roshandel, Zheng and Washington, 2015). Various information is able to be derived from recorded vehicle stream data (Zheng, Ismail and Meng, 2014), such as movement trajectory, driving behaviour, traffic conflict, etc. The movement of vehicle stream along a road segment is able to be recorded by continuous cameras, or from a relatively high position. The cameras should be positioned at vantage places in the vicinity of the road sections in order to obtain high-quality vehicle data, such as no vibration, clear images, and little shadow from surrounding objects. After recording, the detailed data extraction and processing method are provided in Section 3.3.

Onsite recording conducted in Singapore

Onsite video recording was conducted at an AYE road segment on 3 typical weekdays (Tuesday, Wednesday, and Thursday) during the morning peak period from 8:30 pm to 10:30 pm. This AYE road segment is about 800 metres in length, from Clementi Ave 6 (Point 1) to Clementi Flyover (Point 2), with one pair of on-ramp and off-ramp. Six video cameras on a fixed setting were positioned on 3 continuous overhead bridges (A, B, and C) for traffic recording. The road segment layout and recording coverage are shown in Figure D.1.

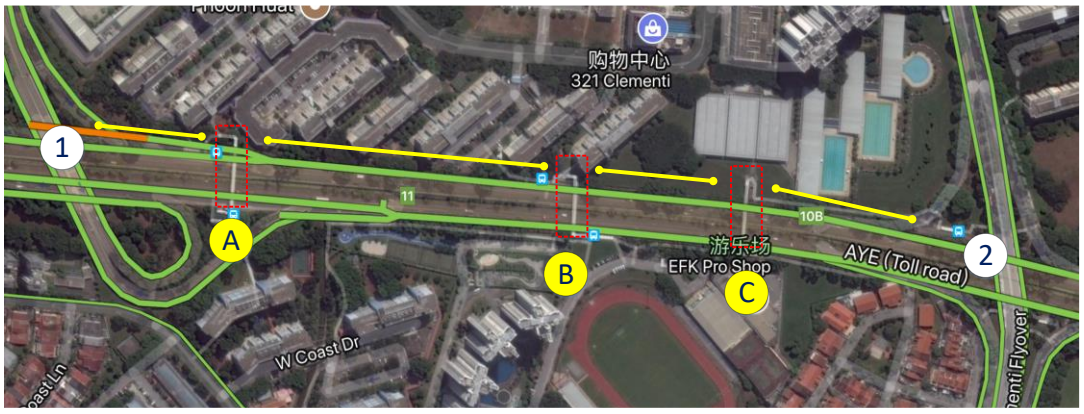


Figure D.1 Road segment layout and recording coverage