# Advanced AI techniques for comprehensive traffic incident analysis: enhancing incident duration prediction and accident risk forecasting

**by Artur Grigorev**

Thesis submitted in fulfilment of the requirements for the degree Doctor of Philosophy.

under the supervision of Dr. Adriana-Simona Mihaita, Dr. Marian-Andrei Rizoiu and Dist. Prof. Fang Chen

University of Technology Sydney
Faculty of Engineering and Information Technology

December 2023

# Certificate of Original Authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Artur Grigorev, declare that this thesis is submitted in fulfilment of the requirements for the award of PhD degree, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: **19/12/2023**

# Abstract

The progress of global urbanization and growth of vehicular traffic have led to an increase in traffic incidents, increasing demand for efficient modeling and prediction methodologies essential for traffic management. To meet this challenge, the thesis proposes the application of advanced machine learning and deep learning techniques to enhance traffic incident modeling. The research contributions are outlined as follows:

Development of a Universal Framework: A novel framework is introduced for predicting traffic incident duration across various road network layouts, handling classification and regression problems, addressing outliers and imbalanced data classes. This framework is designed to handle both classification and regression problems effectively, while also addressing the challenges of outliers and imbalanced data classes. It utilizes feature importance estimation methods, such as SHAP, to identify critical variables, and incorporates anomaly detection techniques like One-Class SVM and Isolation Forest into the prediction models.

Exploration of Data Fusion Techniques: A key contribution of this research is the exploration of data fusion. By integrating different data types, including traffic flow information, textual incident descriptions, and historical traffic flow data, the thesis proposes a methodology to enhance incident duration prediction accuracy. This is achieved through the use of deep learning methodologies like LSTM and ANN encoders, demonstrating the power of combining varied data sources.

Introduction of Visual Transformers: The thesis introduces innovative applications of visual transformers in traffic modeling. The use of the Contextual Vision Transformer network (C-ViT) is a significant advancement, enabling spatial-temporal forecasting of traffic accident risks with higher precision and accuracy, achieving state-of-the-art results and improving upon it through various modifications. This novel application of visual transformers represents a major step forward in traffic incident analysis.

Segmentation and Analysis of Traffic Disruptions: Another major contribution is the development of new methods for segmenting traffic disruptions and associating these disruptions with specific accident reports. This includes an in-depth analysis of the impacts of such disruptions on traffic flow and speed, providing valuable insights into the dynamics of traffic incidents.

Overall, the thesis presents a series of interconnected methodologies that collectively enhance our understanding of traffic incident dynamics. By offering a universal framework and innovative approaches, this research not only contributes to the field of traffic incident modeling and prediction but also opens new avenues for future research and development in traffic management strategies.

# Contents

# List of Figures

# List of Tables

# Acknowledments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Adriana-Simona Mihaita for support, guidance, and valuable mentorship throughout this journey, whose expertise have been instrumental in shaping this research.

I am thankful to Dr. Khaled Saleh and Prof. Massimo Piccardi for their constructive feedback, valuable time, and insightful suggestions that significantly improved the quality of this work.

This work has been done as part of the ARC Linkage Project LP180100114 and funded by iMOVE CRC and supported by the Cooperative Research Centres program, an Australian Government initiative." I'm highly grateful for the support of Transport for NSW, Australia.

Special thanks are extended to my colleagues and friends in the Faculty of Engineering and IT, for their collaboration and constructive feedback.

Also, I am grateful for the stimulating scholarly environment that the Faculty of Engineering and Information Technology and staff. Their collective commitment to academic excellence and quick response has been very supportive throughout my research.

# List of publications

1. Grigorev, A., Mihaita, A. S., Lee, S., & Chen, F. (2022). Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation. Transportation research part C: emerging technologies, 141, 103721. **Published**. Pre-print: https://arxiv.org/abs/2205.05197.
   Code: https://github.com/Future-Mobility-Lab/bi-level-framework.
   CRediT authorship contribution statement. Artur Grigorev: Writing - Original Draft, Methodology, Software, Investigation, Conceptualisation. Adriana-Simona Mihaita: Supervision, Project administration, Writing - Review & Editing, Resources, Conceptualization. Seunghyeon Lee: Writing - Review & Editing, Supervision. Fang Chen: Resources, Supervision.

2. Grigorev, A., Mihăiţă, A. S., Saleh, K., & Piccardi, M. (2022, October). Traffic incident duration prediction via a deep learning framework for text description encoding. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) (pp. 1770-1777). IEEE. **Published**. © 2022 IEEE. Reprinted, with permission. Pre-print: https://arxiv.org/abs/2209.08735
   Code: https://github.com/Future-Mobility-Lab/TIDP_2022.
   CRediT authorship contribution statement. Artur Grigorev: Writing - Original Draft, Methodology, Software, Investigation, Conceptualisation. Adriana-Simona Mihaita: Supervision, Project administration, Writing - Review & Editing, Resources, Conceptualization. Khaled Saleh: Conceptualisation, Supervision. Massimo Piccardi: Supervision, Conceptualisation.

3. Saleh, K., Grigorev, A., & Mihaita, A. S. (2022, October). Traffic Accident Risk Forecasting using Contextual Vision Transformers. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) (pp. 2086-2092). IEEE. **Published**. Pre-print: https://arxiv.org/abs/2209.11180. © 2022 IEEE. Reprinted, with permission.
   Code: https://github.com/Future-Mobility-Lab/ViT-traffic-accident-risk.
   CRediT authorship contribution statement. Khaled Saleh: Writing - Original Draft, Methodology, Investigation, Conceptualisation. Artur Grigorev: Writing - Original Draft, Methodology, Investigation, Conceptualisation. Adriana-Simona Mihaita: Supervision, Project administration, Writing - Review & Editing, Resources, Conceptualization.

4. Grigorev, A., Mihăiţă, A. S., & Saleh, K. (2023, October). Traffic Accident Risk Forecasting Using Contextual Vision Transformers with Static Map Generation and Coarse-Fine-Coarse Transformers. In 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). **Accepted for publication**. © 2023 IEEE. Reprinted, with permission.
   Pre-print: https://www.researchgate.net/publication/372404491_Traffic_Accident_Risk_Forecasting_using_Contextual_Vision_Transformers_with_Static_Map_Generation_

and_Coarse-Fine-Coarse_Transformers.
Code: https://github.com/Future-Mobility-Lab/ViT-traffic-accident-risk.
CRediT authorship contribution statement. Artur Grigorev: Writing - Original Draft, Methodology, Software, Investigation, Conceptualisation. Adriana-Simona Mihaita: Supervision, Project administration, Writing - Review & Editing. Khaled Saleh: Supervision, Resources.

5. Grigorev, A., Mihaita, A. S., & Chen, F. (2023). Predicting Traffic Accident Duration: A Comprehensive Review of Artificial Intelligence Approaches. **Under review at the Engineering Applications of Artificial Intelligence journal**.
CRediT authorship contribution statement. Artur Grigorev: Writing - Original Draft, Methodology, Software, Investigation, Conceptualisation. Adriana-Simona Mihaita: Supervision, Project administration, Writing - Review & Editing. Fang Chen: Supervision.

6. Grigorev, A., Mihăiţă, A. S., Saleh, K., & Chen, F. (2023). Automatic Accident Detection, Segmentation and Duration Prediction Using Machine Learning. IEEE Transactions on Intelligent Transportation Systems (T-ITS) (pp. 1-22). **Published**. © 2023 IEEE. Reprinted, with permission.
Code: https://github.com/Future-Mobility-Lab/AAA-toolkit.
CRediT authorship contribution statement. Artur Grigorev: Writing - Original Draft, Methodology, Software, Investigation, Conceptualisation. Adriana-Simona Mihaita: Writing - Review & Editing, Resources, Supervision. Khaled Saleh: Conceptualisation, Supervision. Fang Chen: Supervision, Resources.

# Chapter 1

# Introduction

Today, Intelligent Transportation Systems (ITS) are an essential component of transport networks in modern cities. These systems monitor and control transport systems, ensuring safety, increasing efficiency, reducing travel time, and lowering air emissions, which significantly impact the economy and health of city populations.

Traffic congestion is a pressing issue faced by numerous cities worldwide. Several factors contribute to this problem, including the growth in population, concentration of the workforce in central areas, and a lack of efficient public transportation modes. Congestion can take two primary forms: recurrent and non-recurrent. Recurrent congestion typically occurs during peak hours when the demand for traffic exceeds the capacity of the roads. The non-recurrent congestion arises from unplanned events such as accidents, breakdowns, adverse weather conditions, etc. Surprisingly, it has been found that almost 60% of traffic congestion is caused by non-recurrent incidents [202] characterized by stochastic behavior, which makes them less predictable. In Australia, the number of road deaths per year has substantially decreased by 70% since the 1970s, indicative of improved road safety measures. Nevertheless, the economic cost of road crashes in Australia remains substantial, with estimates reaching AUD 27 billion in 2017 [73].

Accidents continue to pose a challenge in traffic management despite the implementation of advanced Intelligent Transportation Systems (ITS). The resulting traffic disruptions can cause congestion on multiple roads, leading to significant delays for commuters. Within Australia, the existing incident management and response plan solutions are currently limited, heavily reliant on the expertise of staff members rather than data-driven approaches. However, the adoption of advanced solutions such as machine learning and deep learning techniques holds great potential for improving incident management and response. The Traffic Incident Management System (TIMS) efficiently collects data on incidents and factors influencing their duration. The accurate prediction of incident duration can greatly reduce operational costs and save valuable time for both transport agencies and end-users. It is important to recognize that incident duration is influenced by ongoing traffic congestion and external factors. By estimating the significance of different incident factors and incorporating relevant data into prediction models, the accuracy of predictions can be improved. Therefore, the integration of data-driven approaches and advanced methodologies into transport modeling can revolutionize incident management, ensuring a more efficient and reliable traffic flow.

Most prior studies related to this topic concentrated on testing different machine learning models on specific road types like freeways or highways and focuses primarily on different phases of the incident duration such as clearance time, recovery time, and total incident duration [129]. There is currently a lack of an unified approach that can be applied on all road types, for all accident types, and across various countries with different driving behavior utilizing large amounts of openly available data.

Advances that have the highest impact and which can improve our ability to analyze traffic incidents include:

1. The popularization and wide-scale application of Machine Learning and Deep Learning techniques, which allow for advanced data analysis and more accurate predictions.

2. The availability of high-performance computing devices, which can handle large amounts of data and complex calculations quickly and efficiently.

3. The availability of large datasets related to traffic incidents, including both data from vehicle detectors and data from traffic incident reports.

4. The current lack of complex and sophisticated methodologies for analyzing traffic incidents. This gap in the field presents an opportunity for the development of new, advanced methods that can greatly improve our understanding of traffic incidents and our ability to predict and respond to them.

Important topics in the field of traffic incident research include a range of key areas: 1) the prediction of traffic incident duration, where Machine Learning methods and various approaches to data processing, 2) the detection of traffic incidents, where the focus lies on analyzing traffic flow data and estimating the impact of the incident. Additionally, the spatial and temporal analysis of incidents, which offers insights into the location-specific and time-bound aspects of traffic incidents and their effects.

Current studies in the field of traffic incident analysis predominantly focus on classification, regression and clustering methods. However, research specifically targeting anomaly detection methods is still limited. Anomaly detection methods, such as one-class SVM and isolation forests, possess the potential to effectively detect rare and unusual non-recurring traffic incidents or anomalies in reporting. By incorporating these methods into incident modelling systems, anomaly detection can support the removal of erroneous reports which can adversely affect the model performance. Also, detected anomalies in accident reporting and road situations offer valuable information for further investigation.

The duration of traffic incidents has been analyzed using statistical models such as the log-normal [212] and log-logistics distributions [46], [211], both of which are known for their skewness. These models enable approximate estimation of the incident duration distribution, providing valuable information for understanding the duration of traffic incidents. It has been observed that the incident duration distribution is significantly influenced by various incident case characteristics, such as day/night conditions [252]. Moreover, the chosen method of incident clearance plays a crucial role in predicting the duration of an incident [128]. Furthermore, numerous factors, including weather conditions, traffic density, time period, incident location, and road factors, can exert distinct effects on the duration of different types of incidents [155]. Also, road factors found to be affecting each one of 4 incident types (rear-end, side wipe, collision with fixtures and rollover) in a different way. Therefore, a comprehensive understanding of relevant factors contributing to the accident duration is crucial for accurate modeling and enhancement of our ability to manage and mitigate traffic incident impact.

Incident duration can be modelled in terms of spatial relations (geometric placement of adjacent lanes, angle of adjacency, different parameters of lanes, including speed limits). Recent studies rely on reported incident characteristics [216], [80], but road topology can also play significant role in estimation of the incident probability (e.g. poorly designed junction, wrongly imposed speed limits). According to [49], about 5% of the road junctions are the site of 50% of the accidents in the city of London. Thus, it seems reasonable to analyse incident duration and probability with consideration of the road topology. The task of predicting the duration of an incident usually solved by using Machine Learning methods. Among these methods – tree based methods [178], fuzzy logic [238], Bayesian networks [179], artificial neural networks [18], [12]. And recently [150] studied GBDT as a better performing method for incident duration prediction. Gaussian process regression and artificial neural networks were found to outperform tree methods and SVM in incident duration prediction [80].

Also, estimation of incident duration can be reduced to the classification method [216]. To do this, a specific threshold for the duration is set and a prediction is made whether the incident will last longer than a specified time. Artificial neural networks show high average accuracy for prediction of 4 types of incident severity relying on data on the state of the road (lane, condition of the roadway, weather, light, etc.), time and date. In general, accuracy of classification of traffic accident severity between death, severe, moderate and minor severity accidents is found to be around 60-80% [12].

Recent studies in machine learning involve interpretable models and approaches for model interpretation. Bayesian networks can produce interpret-able models for incident injury severity prediction [174]. Bayesian networks also outperform regression models in incident severity prediction (involving three severity indicators: number of fatalities, number of injuries and property damage) [271]. Interpret-ability is not specific property of the tree models only and by using knowledge distillation one can extract tree rules from different prediction models (e.g. Bayesian network [182]). It allows to represent the model as an interpret-able decision tree and estimate feature importance.

In conclusion, research in the field of traffic incidents encompasses several important topics including the prediction and detection of incidents, spatial and temporal analysis, and the modeling of incident duration. While many studies focus on classification and clustering methods for incident detection, there is a need for more research on anomaly detection techniques to capture rare and unusual incidents. Machine learning methods, such as tree-based methods and artificial neural networks, have shown promising results in incident duration prediction. Additionally, interpretable models, such as Bayesian networks, provide valuable insights and feature importance estimation. Future research should continue to explore the integration of interpretability and anomaly detection techniques to enhance incident detection and analysis, ultimately leading to more effective and efficient traffic management strategies. Traffic congestion remains a major challenge, with both recurrent and non-recurrent congestion causing significant disruptions. Non-recurrent incidents account for a significant portion of congestion, emphasizing the importance of developing effective incident modelling techniques.

## 1.1 Research Project Summary

The research project seeks to develop a universal framework for traffic incident duration prediction, explore data fusion techniques for incident modeling improvement, explore visual transformers for accident risk prediction, and propose novel methods for traffic disruption segmentation and analysis.

This represents a multi-dimensional approach that addresses the challenges of traffic incident detection, duration prediction, and impact analysis.

**Motivation:** with recent advances in the fields of Machine Learning and Deep Learning we can develop a methodology for traffic incident analysis, which will improve the prediction accuracy of traffic incident duration and incident impact estimation.

The main aim of this research study is to improve the traffic incident modelling performance by utilizing advanced machine learning and deep learning methodologies.

Specific objectives are:

- To develop a universal framework for traffic incident duration prediction, which can handle different road network layouts and complexities, account for outliers and propose a methodology for incident duration classification using data-driven approach, optimize both classification and regression problems. In detail, to build a system, which incorporates multiple tasks from the theory of traffic incident analysis: traffic incident detection, incident duration prediction and incident impact analysis.

- To explore the use of data fusion techniques for traffic incident duration prediction, incorporating traffic flow and textual incident description features through deep learning encoding methods into incident duration prediction models.

- To introduce visual transformers for traffic accident risk prediction, incorporating static accident risk maps and various modifications to Visual Transformer architectures for improved prediction performance.

- To develop novel methods for traffic disruption segmentation and association between vehicle detector stations and accident reports, and propose a fusion methodology for combining large datasets to analyze the relationship between traffic accidents and their effects on traffic flow and speed. Successful delivery of this aim consists of the development of a system, which allows to detect traffic incident from traffic flow/speed data, estimate and predict their duration and develop a measure to represent their impact.

### 1.1.1 Research Questions

The rapid development of machine learning and deep learning methodologies presents opportunities to enhance the capacity of existing models in accurately predicting traffic incident durations. In recent years, these advanced techniques have shown considerable promise in a multiple of traffic-related studies, highlighting their potential role in redefining traffic incident management. However, the application of these methodologies for the task of incident modelling remains relatively under-researched. Also, existing studies systematically focusing on methods potential and applicability across specific road network types and incident scenarios. This research aims to address this gap and contribute to the field by constructing a unified modeling framework for incident duration prediction, explore the application and integration of machine learning and deep learning methodologies.

Research questions include:

1. What advanced machine learning and deep learning methodologies can be applied to construct a comprehensive modeling framework for incident management, focusing on the specifics of traffic incident prediction, duration estimation, and incident impact estimation?

2. How can feature importance estimation methods (e.g. SHapley Additive exPlanations) be employed to select the most influential variables from traffic incident reports and determine their contribution to the prediction accuracy of incident duration across various datasets? Also, which accident report characteristic should be prioritized by traffic agencies to improve the traffic incident duration prediction?

3. Can data fusion techniques, specifically incorporating traffic flow and textual incident description data, enhance the accuracy of traffic incident duration prediction? Which encoding methods and deep learning architectures are the most effective for this task?

4. How can the optimal split threshold for modeling short-term and long-term incidents be determined, and what are the best approaches for the regression task of incident duration prediction?

5. Can a universal framework be designed to handle traffic incident data sets from various types traffic networks with different data collection methodologies, and what are the key factors affecting the applicability and accuracy of such a framework?

6. How can visual transformers and their variations, including the Contextual Vision Transformer network (C-ViT), be utilized for traffic accident risk forecasting from spatial and temporal perspectives? Which changes to the network architecture can be integrated into these models for improved prediction performance?

7. What approaches can be developed for accurate segmentation of traffic disruptions and association with specific accident reports and vehicle detector stations? How can this segmentation inform early traffic accident disruption detection and traffic disruption speed impact analysis?

8. How can anomaly detection methods, such as One-Class SVM and Isolation Forest, be incorporated into traffic incident duration prediction models to identify outliers and what will be their impact on the incident duration prediction performance?

9. What is the relationship between traffic accidents and their impacts on traffic flow and speed, and how can this be accurately represented and analyzed using data fusion methodologies, combining large datasets such as CTADS and PeMS?

10. What are the most influential features in accident reports for traffic accident duration prediction using data-driven approaches that have the highest contribution to the incident duration prediction accuracy and how do these features vary across different datasets and incident scenarios?

In conclusion, the findings from this research highlight the potential of advanced machine learning and deep learning methodologies in enhancing the accuracy of traffic incident modelling. Despite limitations, the study's methodologies and findings significantly contribute to the field, laying the groundwork for future research and application of advanced technologies in traffic incident modelling. The study also offers valuable insights into feature importance across various traffic networks, data fusion techniques, application of Visual Transformers to the task of accident risk prediction and methods for traffic disruption segmentation and analysis.

### 1.1.2 Thesis structure

Current thesis type is thesis by compilation, which incorporates published and accepted as well as papers under review. Core chapters correspond to specific papers and mentioned as follows.

The thesis consists of the following chapters:

- **Introduction**. This chapter provides a summary of the research project, outlining its objectives and motivations.
- **Chapter 2. Literature Review**. This chapter presents a comprehensive review of incident modeling studies in the field of traffic incident analysis. The chapter explores the necessary data sets required for thorough incident modeling, including accident logs, traffic states, and external information. Additionally, it provides insights into public data sets used for modeling. The subsequent sections discuss the utilization of machine learning techniques, and the application of advanced deep learning methods in traffic accident analysis. It concludes by summarizing the identified challenges and future research directions in the field. This chapter corresponds to the paper under review 'Grigorev, A., Mihaita, A. S., & Chen, F. (2023). Predicting Traffic Accident Duration: A Comprehensive Review of Artificial Intelligence Approaches'.
- **Chapter 3. Incident duration prediction using a bi-level machine learning framework with outlier removal and intra-extra joint optimisation** . This chapter addresses several challenges related to the prediction of traffic incident duration. We highlight the need for a universal framework applicable to different types of incident data sets on various road network layouts, including arterial roads. Additionally, we discuss the importance of handling outliers and imbalanced data classes in incident duration prediction. Furthermore, we explore the regression problem and investigate the extrapolation performance of machine learning models on different subsets of incident durations. Our objective is to improve the accuracy of incident duration prediction models while identifying important incident report factors for accurate predictions. This chapter corresponds to the published paper 'Grigorev, A., Mihaita, A. S., Lee, S., & Chen, F. (2022). Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation'.
- **Chapter 4. Traffic incident duration prediction via a deep learning framework for text description encoding**. This chapter, focuses on improving the prediction accuracy of traffic incident duration by utilizing additional textual incident description variable and historical traffic flow and speed records. The methodology section presents the details of our modelling framework, including the ML models, LSTM sentiment encoder for textual incident descriptions, and ANN encoder for traffic flow speed. This chapter corresponds to the paper 'Grigorev, A., Mihăiţă, A. S., & Saleh, K. (2023, October). Traffic incident duration prediction via a deep learning framework for text description encoding'.
- **Chapter 5. Spatial-Temporal Traffic Accident Risk Forecasting using Contextual Vision Transformers with Static Map Generation and Coarse-Fine-Coarse Transformers**. This

chapter addresses the problem of traffic accident risk prediction, which has significant implications for city planning, resource allocation, and traffic management strategies. The use of spatial-temporal modeling methods and the availability of publicly accessible data sets have enhanced the automated analysis of traffic data. However, there is still a need for improved prediction accuracy and the integration of contextual information to better understand accident risk, a novel approach inspired by vision transformers, which combines spatial-temporal modeling with contextual information to address the traffic accident risk forecasting problem. This chapter corresponds to two papers Ӧrigorev, A., Mihăiţă, A. S., & Saleh, K. (2023, October). Traffic Accident Risk Forecasting Using Contextual Vision Transformers with Static Map Generation and Coarse-Fine-Coarse Transformers" and 'Saleh, K., Grigorev, A., & Mihaita, A. S. (2022, October). Traffic Accident Risk Forecasting using Contextual Vision Transformers' written in co-authorship. The first paper introduces the use of Visual Transformers for the task of traffic accident risk prediction and the second paper proposes two further enhancements of this approach.

- **Chapter 6. Automatic Accident Detection, Segmentation and Duration Prediction using Machine Learning**. This chapter addresses several challenges in traffic accident analysis, including user-input errors in accident reports, as well as the lack of association between vehicle detector station readings and accident reports. Multiple methods are proposed for traffic disruption segmentation and the association of vehicle detector stations with accident reports. These methods aim to improve the accuracy of accident duration prediction and enable a detailed analysis of accident impact on traffic speed. This chapter corresponds to the paper 'Grigorev, A., Mihaita, A. S., Saleh, K., & Chen, F. (2023). Automatic Accident Detection, Segmentation and Duration Prediction using Machine Learning'.

- **Chapter 7. Discussion, Synthesis and Conclusions**. This chapter includes a discussion, synthesis, and conclusion based on the main findings of the research.

In conclusion, this research aims to address various challenges in traffic incident analysis and modeling. By leveraging advanced machine learning and deep learning techniques, the project strives to improve incident duration prediction, explore data fusion techniques, enhance accident risk prediction using visual transformers, and propose novel methods for traffic disruption segmentation and analysis.

# Chapter 2

# Literature review

## 2.1 Introduction to Traffic accident analysis

Today, Artificial Intelligence (AI) is being used to enhance the performance of different industries and businesses, especially the transport industry. AI technologies such as Machine Learning (ML) and Deep Learning (DL) models can be used to address transportation problems such as traffic management, urban mobility and traffic safety. AI models are used to solve traffic prediction, traffic control, road safety planning and traffic flow optimisation problems [1], [153].

Traffic congestion, which arises in 60% of instances due to unforeseen events [202], poses a significant challenge for numerous urban centres globally. Various factors contribute to congestion, including population growth, the concentration of the workforce in central areas, and the absence of effective public transportation options. Generally, there are two primary types of traffic congestion: a) recurrent congestion, which occurs during peak hours when traffic demand surpasses road capacity, and b) non-recurrent congestion, which results from unpredictable events such as car accidents, vehicle breakdowns, weather-related incidents, and more.

In Australia, the number of road deaths per year was reduced by 70% since the 1970s. However, the annual economic cost of road crashes was estimated at $27 billion per annum in 2017 [73]. In Melbourne, Australia more than 640 km of arterial roads are congested during peak hours with 2.9 tons of CO2 emissions during the years 2014-2015 [15].

Intelligent Transportation Systems (ITS) have become an essential component of urban transport networks in contemporary metropolises. By facilitating the observation and management of the transportation infrastructure, these systems not only increase travel safety but also augment overall transport network efficiency. Consequently, travel durations are minimized, air pollutant emissions are curtailed, and the city's economy and public health are notably improved.

The incorporation of AI techniques into the ITS system has the potential to greatly reduce traffic congestion and its effects on the environment. The main data sources used by Intelligent transportation systems (ITS) are: vehicle detectors (magnetic, infrared, ultrasonic, and microwave), traffic cameras, Global Positioning Systems and Automatic Vehicle Identifiers (e.g. electronic toll collection, access control and speed control) [6]. AI techniques were applied to these kinds of data previously [151], [21]. Multiple various measures can be taken by ITS to reduce the impact of incidents (e.g. variable message signs, toll roads, adaptive cruise control, adaptive traffic light control, transport group priority management) [6].

In Australia, there is a notable deficiency in incident management and response plan solutions, with the majority of transportation management centres relying on the operational experience of their staff members instead of data-driven insights. This highlights a significant gap in the adoption of more sophisticated solutions that leverage existing information sources and contemporary data-driven techniques, such as machine learning and deep learning. Furthermore, the literature reveals a scarcity of research exploring the full range of modelling capabilities that integrate transport modelling with data-driven approaches.

Traffic Incident Management Systems (TIMS) collect data on traffic incidents, encompassing a multitude of factors that affect incident duration. Accurate prediction of the total duration shortly after an incident occurs can potentially decrease operational costs and save end-users time by influencing route planning decisions. Furthermore, the clearance time of accidents is inherently linked to persistent traffic congestion and a variety of external factors, each bearing differing levels of importance. As a result, it is essential to evaluate the significance of these incident-related factors to improve prediction accuracy. The majority of existing research in this domain has primarily concentrated on examining diverse machine learning models for specific road types, such as freeways or highways, and on distinct stages of incident duration, including clearance time, recovery time, and total incident duration [129]. However, there remains a notable absence of a comprehensive approach that can be universally applied to all road types, account for all accident categories, and remain relevant across various countries with unique driving behaviours.

Deep Learning and Machine Learning have become increasingly important tools to improve traffic incident management systems (TIMS). Accurately predicting the total duration of a traffic incident shortly after it occurs is essential to saving operational costs and end-user time, as well as reducing traffic congestion. Understanding the importance of incident factor importance is key to improving the accuracy of predictions. In this chapter, we review the literature related to traffic incident duration prediction and spatial-temporal accident modelling. Specifically, we discuss the challenges associated with each modelling step, the complexity of the task, and the most recent advances in this field, with a focus on the potential of deep learning and machine learning for incident duration prediction. Our goal is to provide a comprehensive overview of the most recent advances in this field, and demonstrate the potential of deep learning and machine learning for incident duration prediction and traffic simulation.

### 2.1.1 Chapter structure

The chapter organisation is detailed as follows:

Section 2.1.2 presents the PRISMA methodology that we have followed for our study, which has revised overall a total of almost 1200 papers on the topic of incident modelling, which have been further filtered and selected down to 75 final papers to provide a comprehensive structured analysis into current gaps and future research directions.

Section 2.2 gives an overview of all the required data sets that one needs to conduct a thorough incident modelling which ranges from accident logs, but also to traffic states such as flow, speed, occupancy, and external related information (weather, events, etc.). We also provide insights into public data sets that have been used for modelling, as many countries restrict access to such data sets due to privacy concerns.

Section 2.3 describes methods of statistical analysis for traffic accident modelling. Section 4 is devoted to the use of Machine Learning in traffic accident analysis including classification and regression tasks, feature selection, imbalanced data set management techniques, anomaly detection, dimensionality reduction, novel machine learning methods and frameworks. Section 5 gives insights into the use of advanced Deep Learning techniques for textual accident report description analysis, accident detection and segmentation from the traffic flow. Finally, in Conlusions we provide a sumamry of the challenges we have dteected as well as future research gaps to be filled.

### 2.1.2 Literature review material and the PRISMA method

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) is the review method that has been applied when organising and analysing literature for this chapter. The process of

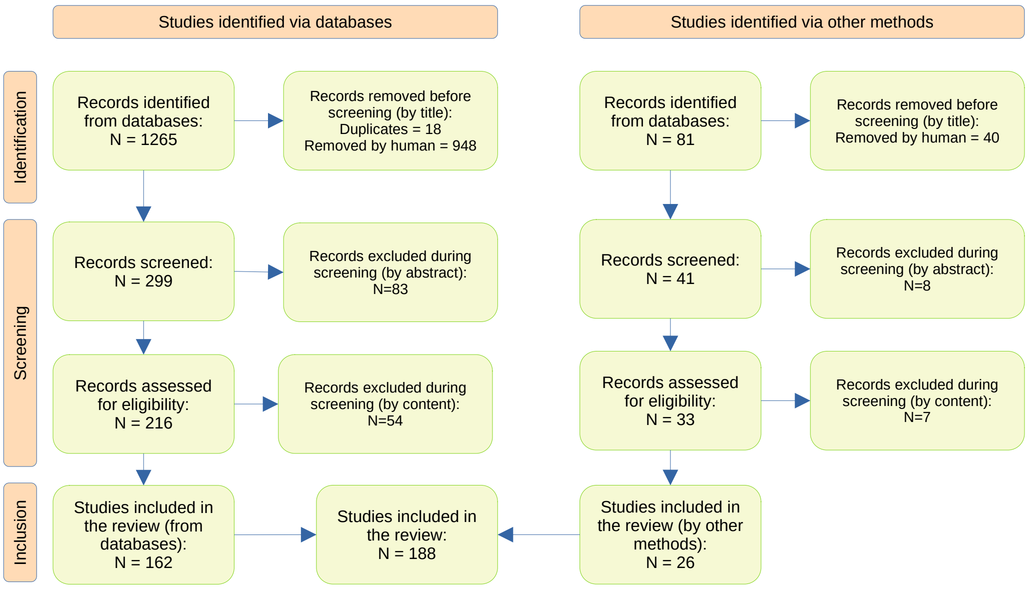



FIGURE 2.1: Flow diagram for systematic review based on the PRISMA approach

reviewing the literature is shown in Figure 2.1. In the first stage, relevant literature has been identified by using publication databases based on keyword search. The list of used keywords:

| No. | Search Query | Technique Used |
|---|---|---|
| 1 | Traffic Incident Duration Prediction | Base query |
| 2 | Traffic Incident Clearance Time Prediction using Machine Learning | Specification |
| 3 | Traffic Incident Clearance Time Prediction using AI | Specification, Generalization |
| 4 | Traffic Accident Duration Prediction | Synonym |
| 5 | Deep Learning models for traffic incident analysis | Specification, Focus shift |
| 6 | Random Forest for Traffic Incident Duration Prediction | Deeper Specification of the first term |
| 7 | Traffic Incident description analysis in congestion mitigation | Double Focus shift |
| 8 | PRISMA method in traffic incident duration prediction research | Contextualization |
| 9 | Literature review on traffic incident duration prediction | Contextualization, Generalization |
| 10 | AI in traffic congestion reduction | Generalization, Focus-shift |
| 11 | Traffic Accident Duration prediction using methods of AI | Specification |

TABLE 2.1: Search queries and the corresponding transformation techniques

Since incidents can be studied across different research areas, it is necessary to always specify the 'traffic' to narrow the search result to the relevant area. The word 'traffic' can also be referred to as a 'molecular transfer process' from Biology. Keywords can include 'clearance time' since this term is very specific to the task of incident duration modelling. The keyword 'machine learning' is the main method used in incident duration prediction. The use of 'traffic incident prediction using random forest' is clearly related to the tasks of classification and regression related to the traffic incident duration modelling. By using very specific terms and specifying area it is possible to locate relevant literature quickly.

The process of creating search queries using linguistic and logical perspectives involves various transformation techniques. Here's a concise explanation of each type of variation and their technical aspects:

- **Synonym**: Replaces words with similar meanings, e.g., "Traffic Incident" to "Traffic Accident," capturing articles with alternative terms.
- **Specification**: Refines queries by adding specifics, e.g., "Machine Learning for traffic incident duration estimation" instead of "Machine Learning," targeting more relevant articles. Another example is "Machine Learning for Incident Duration Prediction", "XGBoost for Accident Clearance Time Prediction".
- **Generalization**: Broadens queries by using general terms or removing constraints, e.g., "AI techniques in traffic incident management" instead of "Traffic Incident prediction using AI," covering related topics.
- **Contextualization**: Queries regarding a specific context, e.g., "PRISMA method in traffic incident duration prediction research," allows finding articles with similar research approaches or contexts.
- **Focus shift**: Alters queries by shifting focus to a different aspect, e.g., "Traffic incident impact estimation" instead of "Traffic Incident duration prediction," exploring sub-topics or aspects of the research area.

Making search queries using these variations maximizes the chances of finding relevant literature. Combining different techniques uncovers a wider range of articles, ensuring a comprehensive understanding of the research area.

The following databases were used for the literature identification:

- ScienceDirect
- Google Scholar
- Research Gate

The alternative source of information on the relevant literature is Connected Papers, which builds a graph of studies based on their semantic similarity. This requires a sample paper to search for similar ones. The search using this approach was performed after the identification of relevant literature using common databases.

The databases were accessed through the University of Technology Sydney, and the publications were limited between 1980 to 2022. In total, 1346 sources were collected, 1265 were found using conventional databases and 81 using Connected Papers.

The PRISMA process for this literature review is detailed in the following:

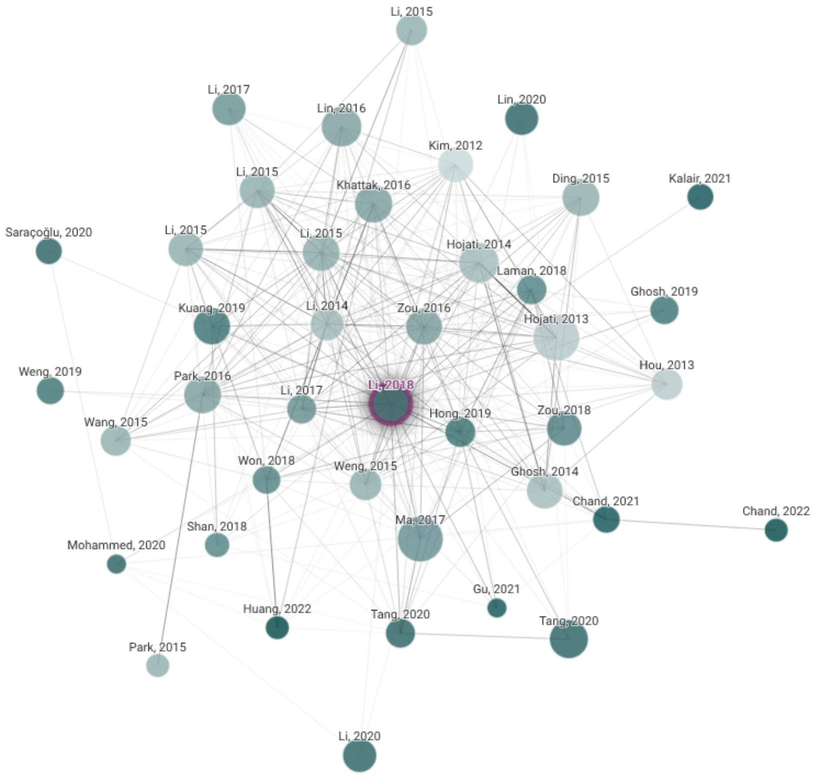



FIGURE 2.2: Connected papers graph example

1. A database of records has been collected using previously described databases and keywords which resulted in 1,265 resources.

2. Resources were screened for duplicates which resulted in 18 resources being removed. Then, records were filtered by paper title, which resulted in the removal of 948 entries mainly due to interference with topics of "internet traffic" (which also relies on the use of machine learning methods and network incident analysis) and "molecular traffic", presence of many studies related to injury statistics and safety analysis related to traffic accidents.

3. Filtered resources (299 in total) were screened by abstract, which resulted in the removal of 83 records due to the mention of unrelated methodologies and findings.

4. Eligible records (216 in total) were then screened by content (reading of methodology and conclusion sections), which resulted in the removal of 54 records.

5. In total, we obtained 162 records from the database search.

6. The most relevant review (Ruimin Li, F. Pereira, M. Ben-Akiva, Overview of traffic incident duration analysis and prediction) from the previous search have been selected and graph of related papers has been built (see Figure 2.2) and a similar process has been performed for semantic similarity search using the Connected Papers service, which resulted in 26 additional papers selected.

7. In total, we obtain 188 relevant resources for the literature review. The highest amount of papers is dated between 2010 and 2022 with peaks in 2002, 2013, 2016, 2018-2021 (see Figure 2.3).

Peaks during this years can be attributed to the introduction of novel Machine Learning methods (e.g. RandomForest in 2002 [134], XGBoost in 2016 [36], onset of the use of Deep Learning methodology in 2015 [117])

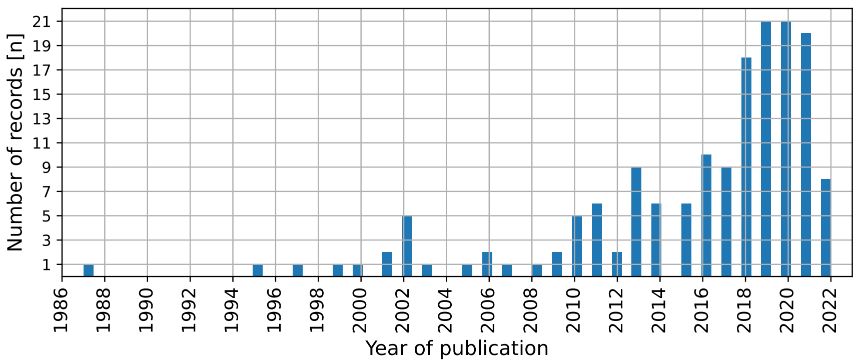



FIGURE 2.3: Reviewed publications grouped by year.

### 2.1.3 Incident duration definitions

Traffic congestion can be recurrent and non-recurrent [3]. Non-recurrent traffic congestion is unexpected congestion, caused by random events affecting traffic flow such as traffic incidents, weather phenomena, vehicle breakdowns, hazards, etc. Recurrent traffic congestion is a predictable regularly occurring congestion, which observed in places where traffic flow regularly exceeds road capacity.[110]

Traffic incident duration, as defined in The Highway Capacity Manual[11], comprises four distinct phases:

- **Incident Detection**: the time interval between the occurrence of an incident and its subsequent reporting,
- **Incident Response**: the time interval between the reporting of an incident and the arrival of the initial investigator at the accident site,
- **Incident Clearance**: the time interval between the arrival of the first investigator and the complete clearance of the incident,
- **Incident Recovery**: the time interval between the clearance of the incident and the restoration of traffic flow to normal conditions.

Determining the precise point in time when a traffic incident occurs can be challenging due to the availability and accuracy of the data. Two potential options to consider are: 1) the time when the car crash happened, or 2) the time when traffic state decline began being registered. Each of these options has its drawbacks:

- Time when the car crash happened: This option refers to the exact moment when the vehicles involved in the accident collided. Choosing this point in time provides a clear starting point for measuring the incident's impact and the time it takes for the entire process, from detection to recovery. However, determining the exact time of the crash can be difficult, as it often relies on witnesses or participants reporting the time accurately. Additionally, in some cases, there might be a delay between the time of the crash and the time when traffic accident starts to be registered (some authorities record the time of accident notification as an accident start time).

- Time when traffic speed decline started being registered: This option focuses on the time when the incident's impact on traffic flow becomes apparent through a noticeable decline in traffic speed. This approach can be more easily determined using traffic sensors or cameras, which continuously monitor traffic conditions. Furthermore, it may better represent the actual impact of the incident on traffic flow. However, one potential drawback is that this approach may not account for the time it takes to detect the incident and initiate the response process.

By considering phases such as clearance time and recovery time separately, a more accurate representation of the overall incident duration can be achieved. However, the collection of data on traffic incidents poses significant challenges due to its complexity. Real-life datasets, in particular, contain a small number of recorded traffic incidents, further complicating the analysis [11].

Duration of detection, response and clearance phases were modelled separately in the literature so far by using Hazard-based duration modelling [169]. Recent study [130] focused on the use of multiple types of distributions (Log-normal, Gamma, etc.) for the four-time intervals within the incident duration structure corresponding to: response team preparation time, response team travel time, incident clearance time, total incident duration time. They found the importance of different distributions to approximate different incident duration stages.

Response time (RT) is defined as a time interval comprised of both response team preparation time and travel time to the incident site. RT was modelled in [94]. In another study, recovery time was analysed on freeway segments in the Southeast Queensland (Australia) [91] and was derived using historical loop-detector-data and traffic incident characteristics at the time and location of the incident. The event of non-recurrent traffic congestion was detected based on the allowable percentage of speed decrease. The time interval of the incident was determined by forward and backward search in time for time intervals, when traffic speed was unaffected, which appear to be bounding for the traffic incident.

Research published in [261] includes the calculation of Recovery Time from the time required for the restoration of travel time during the affected traffic state to travel time during the normal traffic state. That research points to the possibility to use traffic flow data to uncover more precise traffic incident duration instead of relying on the definition given by the response team. As pointed out in Figure 2.4, we can derive the duration of selected phases of traffic incidents from traffic speed data (or possibly traffic flow data). We also want to highlight that guidelines of incident timeline reporting should be taken into consideration when modelling traffic incidents. The reported incident start and end time can mismatch the actual impact of the accident on the traffic speed. Data set providers may not take into consideration time intervals of the incident detection and recovery simply providing the time interval between the report received and the incident cleared message as the total incident duration. This difference in reporting may affect the accuracy of traffic accident impact prediction models. Some incidents may be reported much later than they occured resulting in very short (0-5min) reported accident durations.

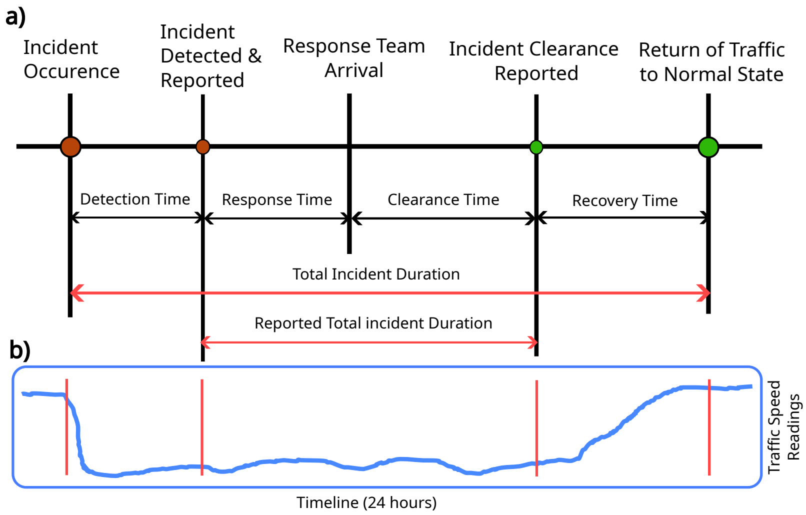



FIGURE 2.4: Traffic Incident Duration representation: a) Timeline segmentation b) Example of speed readings on the day of traffic accident

Understanding these distinct phases is critical for effectively managing and mitigating the impact of traffic incidents. Each phase may be influenced by various factors, and accurately predicting the duration of each phase can help reduce the overall negative impact of incidents on traffic flow, road users, and the environment. By thoroughly examining these phases and incorporating them into prediction models, transportation authorities and incident management systems can better allocate resources, prioritize responses, and minimize the consequences of traffic incidents.

In conclusion, traffic incident duration consists of multiple phases. Nevertheless, data availability for the duration of such phases is rare to find. Duration of phases and total traffic incident duration (assumable, even more correct than recorded by response teams) can be extracted using traffic flow data assuming the data streams are reliable and free of anomalies or outliers that might affect precision and analysis.

## 2.2 Data sets and data availability

The availability of data related to traffic accidents in recent years has enabled a deeper understanding of the factors that lead to these incidents and their outcomes. This data, composed of information such as the location and time of an accident, the type of vehicle involved, the severity of the crash, casualty statistics and economic cost can provide valuable insights into the causes and outcomes of traffic accidents. By employing machine learning techniques, it is possible to make predictions about the risk of future crashes, classify accidents by severity and predict accident duration to develop strategies for mitigating the risk and severity of accidents. This kind of data analysis can help to inform policymakers and road safety management organisations on how to create safer roads and highways, as well as to make driving safer for everyone.

There are multiple publicly available datasets:

- National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS): This dataset contains detailed information on fatal motor vehicle traffic crashes in the United States occurred since 1975 [173].

- National Transportation Atlas Database (NTAD) from United States Department of Transportation's Bureau of Transportation Statistics (BTS) [222]: contains detailed information on non-fatal motor vehicle traffic crashes in the United States since 1994.

- European Commission's Road Safety Atlas [47] provides accident statistics for each European country using interactive maps and satellite images.

- UK Road Safety Statistics [74]: This dataset contains detailed information on fatal and non-fatal road traffic accidents in the UK since 1979.

- California Highway Patrol (CHP) Statewide Integrated Traffic Records System (SWITRS) [41]: a California-wide data set containing detailed information on motor vehicle collisions reported to California Highway Patrol. Accident report details contain data on the location, severity, road condition and victim data including age and degree of injury. Due to the extensive timeline and precision of reporting, this data set was previously used to analyse the effect of country-scale events on crash severity [231].

- World Health Organization's Global Health Estimates [177]: This dataset contains detailed global estimates on road traffic injuries, deaths, and disability-adjusted life years from 1990 to present.

- Australian Road Deaths Database (ARDD) [99] provides basic details of road traffic crash fatalities in Australia as reported by the police each month to the State and Territory road safety authorities. The data set includes information on fatal crashes: year, month, day of the week, time, location, crash type and vehicle type involved.

- Countrywide Traffic Accident Dataset (CTADS) - one of the biggest data sets on traffic accidents is , recently released in 2021 [165] [166], which contains 1.5 million accident reports collected for almost 4.5 years since March 2016, each report containing 49 features obtained from MapQuest and Bing services. This data set was used previously to predict the accident duration [265], [75].

- Caltrans Traffic Performance Measurement System (PeMS) [39] data set contains incident reports with a timeline of events and description of the incident (as a sequence of abbreviations) as it becomes available and status updates from a dispatch unit. Also, the dataset contains 5-minute aggregated traffic speed, traffic flow and traffic occupancy records as well as vehicle detector status. The availability of this data allows analyzing traffic incidents in conjunction to vehicle detector data. A brief description of the incident includes location, area, start time, duration and freeway ID.

- CompassIoT [100] is a data set and API of connected vehicle road trips across Australia which aggregates data from 64 different manufacturers across a number of global car brands. It started from 200,000 connected vehicles in 2018 to over 2.2 million trips and billions of data points in 2023. The data set has low data latency, where real-time API can yield data every 5 seconds. Data on braking acceleration and steering can allow identification of dangerous road conditions and risky behaviour characteristic to road accidents.

- TomTom [220] - historical traffic database with information on road speeds, travel times and traffic density. Allows customised queries for route and area analysis providing statistics on travel time, speed.

### 2.2.1 Characteristics of traffic incidents

Features used by research studies on traffic incident duration are very diverse. Some researchers didn't perform possible feature manipulations despite data availability (e.g. AM/PM peak hour binary feature or night time). Traffic incident research has the possibility to benefit from comprehensive feature derivation based on time, weather and road network data. By tracking the use of features and assessment of their significance, we can make decisions on the concentration of future efforts in data processing. Also, we can develop strategies for feature extraction considering techniques used by other researchers.

For example, in [130] authors use the feature "season" which is represented as a set of four discrete values (summer, winter, spring, autumn). The season can highly affect safety on the road due to weather effects: winter storms, ice on the road, rain showers and other environmental effects (which affect visibility, control and safety) linked to the time of the year. But in some papers [159] authors don't use these features. Rainfall intensity was found highly relevant for the prediction of accident severity in Seoul, South Korea [118].

Roadway geometry is one of the critical factors which can affect the capability of a road system to withstand the incident impact [6]. Various features are used to define the road structure and use machine learning models: road segment length, road segment centroids, gradient, curvature and general road density surrounding the event area [270]. A comprehensive review of spatial network (i.e. road network) theory applications provides an in-depth analysis of graph theory indices including betweenness centrality, ringness, route factor, detour index, and alpha index [19]. All of these indices can be calculated for road networks and incorporated into machine learning pipeline. Accident risk is found to be increasing when traffic speed slows down while traffic density goes up for Yingtian Expressway [141] which highlights the importance of speed-density diagrams for incident-related traffic state visualisation. In that research, no correlation was observed between traffic flow and crash risk. Various Road alignment factors can include: Curve, Tangent, Vertical grade, Super-elevation, Horizontal alignment and Curve length [118].

Identification of road geometries is important for the analysis of incident occurrence. The impact of variables associated with crash frequency was found to be varying across parts of Tennessee, USA [163]. The spatial analysis demonstrated that segment length and median segment width had the highest impact on crash frequency in eastern regions, while commercial land use had the highest connection to crash frequency in southern regions. Multiple studies have indicated that patterns and dependencies in the spatial and temporal dimensions are likely to exist, often represented as clusters or hot spots [7].

The following tables 2.2 and 2.3 include incident characteristics used among different research studies, with citing specific ones, unobserved or rarely observed in papers. For the general set of characteristics, readers can refer to [159], [129].

| Feature | Values | Reference |
|---|---|---|
| Peak hour | $\{0,1\}$ | - |
| Weekday | $\{0,1\}$ | - |
| Weekend | $\{0,1\}$ | [102] |
| Season | $\{winter, autumn, summer, spring\}$ | [130] |
| Time of day | $\{0..23\}$ | [159] |
| Peak hours | $\{Offpeak/AMpeak(6-9AM)/PMpeak(3-6PM)\}$ | [169] |
| Daytime | $\{Evening, night-time\}$ | [169] |

TABLE 2.2: Table of temporal features used to describe traffic incident

| Feature | Values | Reference |
|---|---|---|
| Incident type | $\{Vehiclefire, outofgas, breakdown, etc.\}$ | [110] |
| Number of vehicles involved | $\{1..N\}$ | - |
| Multiple vehicles involved | $\{0,1\}$ | [91] |
| Type of vehicle involved #1 | $\{Motorcycle, Van, Pickup\}$ | [110] |
| Type of vehicle involved #2 | $\{Largevehicle\}$ | [94] |
| Type of vehicle involved #3 | $\{Truck\}$ | [43] |
| Location of incident on the road | $\{Forfreeways: ramp, left/rightshoulder\}$ | [110] |
| Number of lanes | $\{1..N\}$ | [91] |
| Link capacity | $\{N\}$ | [91] |
| Average speed at the time of incident | $\{\}$ | [91] |
| Number of affected Lanes | $\{1..N\}$ | [159] |
| All lines affected | $\{0,1\}$ | [102], [94] |
| Incident Severity | $\{1..N\}$ | [159], [85] |
| Lighting condition | $\{day, night\}$ | [85] |
| Secondary crash | $\{0,1\}$ | [85] |
| Fire, Injury | $\{0,1\}$ | [94] |
| Fatality | $\{0,1\}$ | [91] |
| Traffic disrupted | $\{0,1\}$ | [91] |
| Traffic flow on ajacent lanes | $\{N\}$ | [159] |
| Medical required | $\{0,1\}$ | [91] |
| Rollover | $\{0,1\}$ | [43] |
| Weather #1 | $\{Windy, Clear, Rain\}$ | [11] |
| Weather #2 | $\{Sunny, Cloudy, Storm\}$ | [43] |
| Weather #3 | $\{Rain, Snow, Wind, Fog\}$ | [169] |
| Position within road | $\{Inner, Outer, Middlelane\}$ | [43] |
| Lane number | $\{1..N\}$ | [159] |
| **Other Features** | **Values** | **Reference** |
| Distance from the city center | $\{Rkm\}$ | [159], [91] |
| Traffic condition | $\{congested, uncongested\}$ | - |

TABLE 2.3: Table of features used to describe traffic incident across different studies

## 2.3 Incident duration modelling

Current research predominantly focuses on classification and clustering techniques for incident detection in traffic conditions. However, there is a noticeable scarcity of studies investigating traffic incidents utilizing anomaly detection methods, such as one-class SVM, isolation forests, etc. Non-recurring traffic incidents, characterized by their rarity and unusual nature, can be considered anomalies within the context of traffic patterns. Consequently, deploying anomaly detection methods for incident detection systems can offer adaptability to previously unencountered situations.

In this context, a comprehensive analysis of pertinent anomaly detection techniques, in comparison to established classification and regression methods needs to be addressed. Moreover, identifying anomalous road situations can prove invaluable for subsequent investigations, especially during the early stages of a newly reported accident.

Incident duration can be effectively modelled through the analysis of spatial relations, taking into account the geometric positioning of adjacent lanes, angles of adjacency, and various lane parameters such as speed limits. While recent studies have largely focused on reported incident characteristics as noted in [159] and [81], road topology is an essential factor in estimating incident probability, with poorly designed junctions and inappropriate speed limits contributing to increased risk. In fact, according to [49], approximately 5% of road junctions account for 50% of accidents within the city of London. Therefore, it is crucial to analyze incident duration and probability by incorporating road topology into consideration.

To predict incident duration, various Machine Learning methods have been employed. These techniques encompass tree-based methods as discussed in [178], fuzzy logic systems explored in [238], Bayesian networks analyzed in [179], and artificial neural networks (ANNs) as presented in [18] and [12]. More recently, [150] investigated the Gradient Boosting Decision Tree (GBDT) as an improved method for incident duration prediction. Additionally, it has been found that Gaussian process regression and artificial neural networks surpass tree methods and Support Vector Machines (SVM) in predicting incident duration, as demonstrated by [81]. By utilizing recent advancements in the field of accident analysis, the examination of road topology, spatial relations, and advanced machine learning techniques can lead to better insights into incident duration prediction and risk management.

Overall, understanding the factors that influence incident duration and contribute to the probability of the accident, can help improve traffic management systems and enhance road safety. Therefore, exploring different machine learning methods to predict incident duration is crucial to reduce traffic congestion and improve overall traffic flow.

The majority of previous research has primarily focused on predicting incident duration for specific types of roads, such as freeways or motorways, where data accuracy is generally higher than on arterial roads. As of 2018, only a limited number of studies have applied prediction strategies to normal arterial roads due to the increased modeling complexity and issues related to location mismatching. Most traffic incident duration analysis research concentrates solely on one type of road network, like freeways or highways, as evidenced by [257], [45], [90], and [262]. This observation is further supported by a recent state-of-the-art review published in [129], which highlights the challenges in addressing this problem for arterial roads and the scarcity of studies in this particular domain.

Model transfer between different road types is an important aspect to consider in incident duration prediction research. By studying this transferability, researchers can determine whether a model developed for one type of road network can be effectively applied to another, potentially saving time and resources in the development of new prediction models.

Transfer learning, a technique that leverages knowledge gained from one domain to improve performance in another, can be particularly useful in studying model transfer between road types. By examining the similarities and differences in road features, traffic patterns, and incident characteristics, researchers can gain insights into the factors that contribute to the transferability of prediction models across different road networks.

Exploring model transfer between road types offers several benefits. First, it allows researchers to identify generalizable features and prediction patterns that can be applied across various road networks. This can lead to the development of more robust and versatile prediction models. Second, understanding model transferability can help identify potential limitations and areas of improvement for existing models, ultimately enhancing their accuracy and performance when applied to new road types.

As a result, investigating model transfer between different road types, and comparing different models for generalization across road types can play a crucial role in advancing incident duration prediction research.

The estimation of incident duration can also be approached through classification methods, as suggested by [159]. To implement this, a specific threshold for duration is established, and predictions are made to determine whether an incident will exceed the specified time.

Artificial neural networks (ANNs) have demonstrated high average accuracy in predicting four types of incident severity based on road condition data, such as lane, roadway condition, weather, lighting, and temporal factors like time and date. The overall accuracy for predicting fatal, severe, moderate, and minor severity accidents was found to be between 69-72

Bayesian networks, on the other hand, can generate interpretable models for predicting incident injury severity, as illustrated in [174]. Moreover, they outperform regression models in incident severity prediction, which involves three severity indicators: the number of fatalities, the number of injuries, and property damage, as shown in [271].

It is worth noting that interpretability is not exclusive to tree models. By employing knowledge distillation, tree rules can be extracted from various prediction models, such as Bayesian networks, as detailed in [182]. This allows for the representation of the model as an interpretable decision tree while also estimating feature importance.

It is possible to gain a deeper understanding of incident duration estimation through classification methods, artificial neural networks, and Bayesian networks, and their potential for providing interpretable models and feature importance insights.

### 2.3.1 Traditional accident modelling

The distribution of incident duration has been modeled using various approaches. Earlier studies, such as [212], employed a log-normal distribution, while more recent research has shifted towards utilizing log-logistic distribution, as seen in [46], [210], and [211]. The log-logistic model has been more extensively used and found to offer better goodness-of-fit compared to the log-normal distribution.

Recent studies have also explored multi-component log-logistic models. For instance, [272] introduces a g-component log-logistic model, while [128] presents a competing risks mixture model incorporating a multinomial log-logistic model.

However, estimating the actual distribution of incident duration can only provide approximate information. Moreover, the distribution of incident duration has been found to be heavily influenced by various incident case parameters, such as day or night conditions, as noted in [252].

Various **hazard-based models** have been employed to analyze traffic incident duration, as demonstrated in [169] and [91]. These models utilize a hazard function to describe the conditional probability that an incident will conclude during a specific time interval, given that it has already persisted until the beginning of that interval. Recent studies have also explored multi-component log-logistic models, with the authors in [272] introducing a g-component log-logistic model and [128] presenting a competing risk mixture model that incorporates a multinomial log-logistic model.

In papers attributed the to early 2000s, authors primarily used one or two distributions to fit traffic incident duration data. But starting from the 2010s, we can observe the use of different distributions within one research for the approximation of different phases of incident duration with comparing them according to the BIC score [130], [11] - the Bayesian information criterion (BIC).

In the recent study, clearance duration and impact duration were modelled using Weibul, Log-normal, Log-logistic distribution [85]:

- Log-logistic model outperformed Weibul and Log-normal models based on the comparison by AIC criteria.
- Incident impact duration was based not on incident durations reported by response teams, but estimated from historical speed data from BlueTOAD device pairs located on road segments.
- Hazard-based modelling approach allowed to estimate the impact of incident characteristics on impact and clearance durations. Most of the characteristics (night time, severity, EMS involvement, etc) were found to be affecting both durations in the same way but with a different degree of impact. But some characteristics demonstrated the opposite effect on duration (percentage of the lane closure, peak hour, summer/fall season, involvement of towing vehicles).

Incident impact modelled based on traffic flow data has the potential to be more accurate than reported by incident response teams (due to possible reporting errors). Because of the significant observed difference in modelling accuracy for impact and clearance duration [11], the modelling was proposed to be made separately for each incident type (crashes, hazard, etc) to see how one model can perform better than another in each case.

The fact that different distributions can be used to approximate different phases of traffic incident [130], implies that we can fit different distributions not only to phases but also split the dataset by specific variables (e.g. peak hour), which can also lead to different estimation in feature importance among resulting datasets (e.g. what is important for peak hour incidents, can be less important than for non-peak hour incidents).

### 2.3.2 Log-transformation of the Incident Duration variable

In predictive machine learning tasks, it's crucial to examine the statistical properties of the target variable. Particularly, the target variable's distributional characteristics can significantly impact the accuracy and performance of the predictive models. Often, the incident duration distribution is not normally distributed; instead, it follows a skewed distribution such as a log-normal or log-logistic (see Figure 2.5). This skewness can be due to numerous extreme values, possibly caused by a few incidents lasting extraordinarily longer than the majority. The logarithmic transformation is a monotonic transformation known for its effectiveness in stabilizing variance, reducing skewness, and making the target predicted variable (incident duration) resemble a Normal distribution more closely. This is especially critical for linear models and other algorithms that have an implicit or explicit assumption of normality or symmetry in the target variable distribution. The log transformation can help reducing the skewness and decreasing the influence of extreme values. For log-normally distributed data, a log transformation will result in a normal distribution. For log-logistically distributed data, a log transformation may not lead to a normal distribution but can still significantly reduce skewness. Therefore, for AI-based traffic incident duration prediction tasks, if preliminary data analysis indicates that incident duration follows a skewed distribution (log-normal or log-logistic), a log transformation is a beneficial preprocessing step.

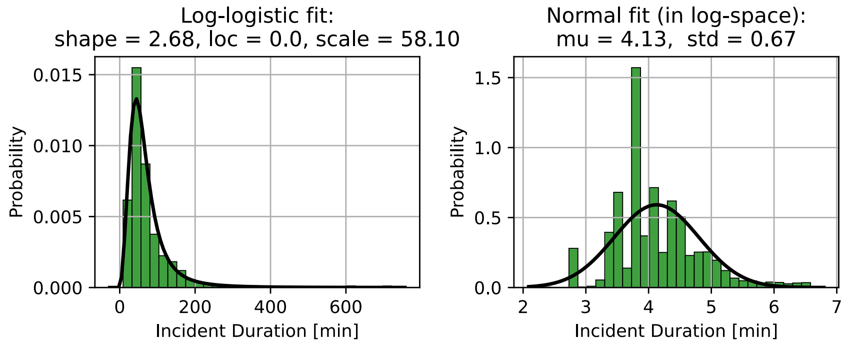



FIGURE 2.5: Example of distribution fit for the sample data of San-Francisco data set[75]

### 2.3.3 Detection and estimation of traffic accident duration from traffic flow data

Main factors contributing to traffic jams are high traffic load (e.g. during peak hours), a bottleneck (a spatial aspect of road geometry) and local disturbances in the flow (e.g. actual traffic incidents which act as a trigger of traffic jam) [223]. Accident detection systems can rely on offline or real-time data. The emerging approach is to use computer vision methods to detect traffic accidents using CCTV cameras. The spatial-temporal near-accident condition detection system has been recently proposed leveraging object detection, segmentation and tracking [98]. Advanced Driver Assistance Systems are intended to mitigate or prevent crashes by providing vehicle drivers with the necessary information to avoid collisions. To accomplish this task the truck driver behaviour (speed reduction) encountering vulnerable road users (e.g. cyclists at intersections) have been studied [201]. The study of vehicle behaviour using GPS data can provide valuable insight into driver behaviour to further assist general drivers to avoid accidents. Geographically Weighted Poisson Regression (GWPR) models were used to model frequencies of harsh driving behaviour events, which were found to be positively correlated with segment length and presence of traffic lights and negatively with neighbourhood complexity (which is a density area in proximity of the event) [270]. Road features can be used to predict the traffic accident risk since they affect the driver's behaviour. Traffic state identification systems related to traffic jams can rely on traffic speed and density data. The previously used algorithmic approach can allow to identification of interruptions and moving jams [141].

## 2.4 Machine Learning in traffic accident analysis

The figure 2.6 illustrates the machine learning pipeline used for predicting the duration of traffic accidents. The pipeline consists of 1) data preprocessing (cleaning, data imputation, label encoding, and outlier detection), 2) data manipulation which may include the feature transformation (Principal Component Analysis and Latent Dirichlet Allocation, Log-transformation of target variable) and feature selection (e.g. using correlation-based feature selection, univariate feature selection, recursive feature elimination), model training (e.g. using linear regression, support vector machines, k-nearest neighbors, decision trees, random forests, and neural networks), feature importance estimation (using Gini importance, permutation importance, or SHAP), 3) model training and evaluation (including cross-validation, confusion matrix, and ROC curve), and model deployment (accuracy, precision, recall, F1-score, RMSE and MAPE). The model is then validated to ensure its accuracy and reliability. The pipeline shows a general way of predicting the duration of stochastic events (like accidents) using machine learning methods and can be used for other similar tasks (e.g. prediction of accident severity).

A machine learning (ML) workflow for traffic accident duration prediction includes various stages regarding preparing the data, engineering and selecting features, training and evaluating models, and ultimately deploying the model:

- Data Preparation: The first step in the process involves preparing the data for analysis. This includes imputing missing values, standardizing the data, and converting various data types. These preprocessing steps are crucial to ensure that the ML algorithms can effectively learn from the data.

- Feature Engineering: The next step involves transforming the raw data into a set of features that can be used as input for the ML algorithms. This includes type conversion, encoding categorical variables, and functional conversion, such as the vectorization of text data.

- Feature Transformation: This stage involves applying techniques like Principal Component Analysis (PCA) and Latent Dirichlet Allocation (LDA) to reduce the dimensionality of the feature set. This can help improve the performance of the ML algorithms by reducing noise and computational complexity.

- Feature Selection: In this step, various techniques like Correlation-based Feature Selection, Univariate Feature Selection, and Recursive Feature Elimination are used to select the most relevant features for the prediction task. This helps to further reduce the complexity of the model and improve its generalization capability.

- Model Training: At this stage, several ML algorithms such as Linear Regression, Support Vector Machines, k-Nearest Neighbors, Decision Trees, Random Forests, and Neural Networks are trained on the prepared data.

- Model Evaluation: The trained models are evaluated using Cross-Validation, with metrics like accuracy, precision, recall, F1 score, RMSE, and MAPE. The evaluation process helps in selecting the best performing model for the prediction task.

- Feature Importance Estimation: Techniques like Permutation Importance and SHAP are used to estimate the importance of different features in the model. This can provide insights into which features have the most significant impact on the prediction of traffic accident duration.

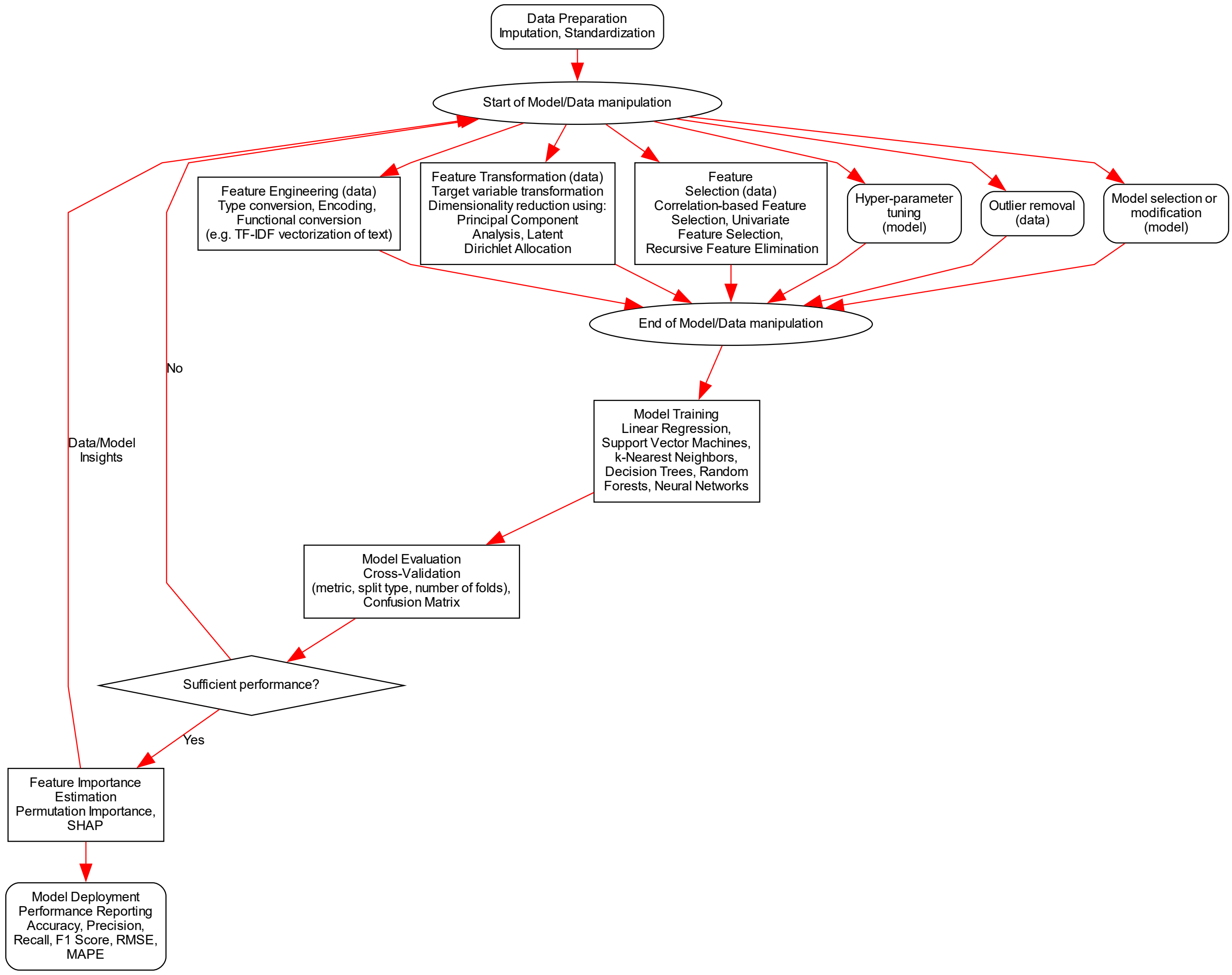



FIGURE 2.6: Machine Learning pipeline for traffic accident duration prediction using table data

- Iterative Decision Point: A decision is made based on the model's performance. If the performance is sufficient, the process moves to the next step of model deployment. Otherwise, the model and data are further refined using techniques like hyperparameter variation, outlier removal, model selection or model modification.

- Model Deployment: Once a satisfactory model and data organization are obtained, the resulting system is deployed for traffic accident duration prediction. Performance metrics like accuracy, precision, recall, F1 score, RMSE, and MAPE are reported as model performance estimations.

As a typical ML workflow for traffic accident duration prediction, it includes decision points and iterative steps for refining the model and data, it highlights the importance of data preprocessing, feature engineering, model training and evaluation, and iterative refinement to achieve the best possible model for predicting traffic accident durations.

The task of predicting the duration of an incident can be solved by using Machine Learning methods. Among these methods are: tree-based classification methods [210], [178], fuzzy logic [238], Bayesian networks [179], linear regression analysis (LR) [108], artificial neural networks (ANN)

[239], [12], [18], support-vector regression (SVR)[233]. Recently GBDT(gradient-boosted decision trees) have been revealed to be a better performing method for incident duration prediction [150]. Gaussian process regression and artificial neural networks were found to outperform tree methods and SVRs in incident duration prediction [81].

Extreme Learning Machines (ELM) [95] - is a machine learning method, which incorporates a feed-forward neural network initialised with random weights and consequent training step based on produced random feature mapping, designed to avoid overfitting of neural network.

Several classic machine learning methods are widely utilized for traffic incident duration modeling represented in Table 2.5.

| **Method** | **Description** | **Traffic Accident Duration Prediction** | **Advantages** | **Limitations** |
|---|---|---|---|---|
| k-Nearest Neighbors (kNN) [59] | Predicts based on the majority vote or average of the k closest neighbouring data points using a distance metric. | kNN predicts traffic accident duration using data on similar accidents in the training data. By considering the k closest neighbors, the algorithm accounts for the local structure of the data. | Simple to implement<br>Can handle non-linear patterns | Sensitive to the choice of k<br>Poor scalability: computationally expensive for large datasets |
| Linear Regression (LR) | Models the relationship between features and the dependent variable in regression tasks using linear equations. | LR predicts traffic accident duration by modelling the linear relationship between accident features and duration. However, it may underperform if the relationship is not linear. | Easy to interpret<br>Computationally efficient | Assumes linear relationship<br>Simplicity: May not capture complex patterns |
| Random Forests (RF) [26] | Combines the average or majority votes from multiple decision trees trained on randomly selected data subsets using bootstrap aggregation. | RF can predict traffic accident duration by using multiple decision trees which reduces the sensitivity of the model to noise in the data, providing a more robust prediction. | Robust to noise<br>Can handle missing values<br>Less prone to overfitting<br>Allows feature importance estimation | May require a large number of trees for robustness against outliers<br>Computationally expensive |
| Support Vector Machines (SVM) [78] | Effective method for both classification and regression tasks. | SVMs can be used to classify accident severity or predict the exact duration of the incident. Support Vector Regression (SVR) can predict continuous values, while SVM can categorize accidents based on certain criteria (e.g. short/long duration, severity). | Can handle non-linear patterns<br>Robust to noise<br>Effective with high-dimensional data | Computationally expensive for large datasets and high-dimensional data<br>May require tuning of hyperparameters |
| Bayesian Models [23] | Bayesian techniques encompass a wide range of methods suitable for regression and classification tasks. | For traffic accident duration prediction, Gaussian Process Regression (GPR) [72] and Bayesian networks [170] can be utilized. Bayesian models account for uncertainty in their predictions and are particularly beneficial when dealing with limited data, measurement error and/or missing observations. GPR is a powerful nonlinear method, can be used to interpret nonlinear systems without prior knowledge | Handles uncertainty<br>Resilient to noise<br>Efficient for low amounts of data<br>Flexible modeling of relationships | Computationally demanding for large datasets<br>May need expert knowledge to define model structure |
| Neural Networks (NN) [241] | Neural networks, particularly deep learning models, can be used for regression and classification tasks. | Feed-forward neural networks (FFNN) and recurrent neural networks (RNN), including Long Short-Term Memory (LSTM) [89], can predict traffic accident durations, considering the temporal aspect of the data | Can capture complex patterns<br>Effective with large datasets | Requires large datasets<br>Computationally expensive<br>May require tuning of hyperparameters |

TABLE 2.4: Classic Methods

More advanced methods are represented in Table 2.5.

These models can be applied to both classification and regression problems, with the exception of logistic regression (used only for classification tasks) and linear regression (employed solely for regression tasks).

Recently, a two-step approach for predicting the duration of traffic incidents has been proposed [115]. The first step was binary classification, which used a cost-sensitive Bayesian network to predict whether an incident would last for more or less than 30 minutes. The second step was regression, which used the k-nearest neighbours approach to predict the duration of the incident. Although the approach was effective, it had one significant limitation: manually selecting the threshold for binary classification may not yield the optimal performance when using data-driven approaches or machine learning methods (e.g. it could result in class imbalance). Therefore, a fixed and varying incident duration threshold needs to be compared to find the best class balance for classification models; even more, a comparison with a multi-class classification approach may also be employed. Also, more advanced regression models together with outlier removal procedures can be used to model groups of incidents separately providing a better and more precise prediction of the incident duration [75]. Alternative to the incident duration threshold split, clustering methods can also be used to detect groups of traffic incident reports to model them separately. Overall, we want to highlight that within the accident report dataset, there can exist multiple report groups and when using separate modelling may provide better accident duration prediction performance [265].

In one of the recent researches about the classification of driving state, multiple hyper-optimised ML models were tested, and the entire feature space was visualised using t-SNE[255]. RandomForest provided the highest accuracy of prediction, but more advanced tree-based models exist that utilise gradient boosting, which we will be using in our research (e.g. gradient-boosted decision trees).

To assess the performance of advanced tree-based models (e.g. LightGBM), supplementary conventional ML models can be employed [38]. This comparison includes both well-known tree-based models and non-tree-based models, such as k-nearest neighbors and Logistic Regression.

A significant portion of previous research has focused on predicting incident duration for specific types of roads, such as freeways or motorways, where data accuracy is generally higher compared to arterial roads. As of 2018, only a limited number of studies have attempted to apply prediction strategies to arterial roads due to the increased modeling complexity and challenges associated with location matching [257]-[45]-[90]-[262]. This observation is supported by a recent state-of-the-art review published in [129], which highlights the difficulty of addressing this problem for arterial roads and the scarcity of research in this area.

The majority of traffic incident duration analyses tend to concentrate on a single type of road network, such as freeways or highways, rather than exploring diverse road types. This narrow focus may limit the applicability and generalizability of the developed models, as they may not account for the unique characteristics and challenges associated with different road networks.

### 2.4.1 Classification and regression tasks in the incident duration prediction

The machine learning pipeline follows the general structure including the use of regression/classification models, feature correlation tests, data split according to train-test-validation schema, model calibration and validation, and prediction performance evaluation [66]. One of the studies on accident hot-spot clustering [7] highlights that methods applied for vehicle accident analysis can be applied to various kinds of point-based events like crime, natural hazards, etc. A supervised machine learning model built upon extreme gradient boosting models has been used to predict rail-road accidents (derailments) and to rank, the importance of contributing factors using ANOVA and Gini criteria [29]. Crash prediction models can be historical and real-time. The comprehensive review on real-time crash prediction approaches has been performed in [93].

### Classification and regression definitions

In order to analyze traffic incident data and estimate the duration of incidents, we first define the matrix of traffic incident features as follows:

$$X = [x_{ij}]_{i=1..N_i}^{j=1..N_f} \tag{2.1}$$

Here, $N_i$ represents the total number of traffic incident records used in our modelling, and $N_f$ is the total number of features characterizing the incident (such as severity, number of lanes, type, neighbourhood, etc.), according to each specific dataset.

Estimating the duration of traffic incidents can be approached as a classification task [159]. In this case, a specific threshold for the duration is set, and the model predicts whether the incident will last longer than the specified time. Artificial neural networks have been shown to achieve high average accuracy (around 69-72

For the incident duration classification problem, we define the incident duration classification vector as:

$$\begin{cases} Y_c = [y_i^c]i \in 1..N & y_i^c \in 0, 1 \\ Ymc = [y_i^{mc}]_{i \in 1..N} & y_i^{mc} \in 0, 1, ..., M \end{cases} \tag{2.2}$$

In this notation, $N$ is the duration of the traffic incident (in minutes), $Y_c$ is the vector of binary values taking values in 0, 1, and $Y_{mc}$ is the vector of integer values for the multi-class classification problem definition, taking $M$ discrete values. More specifically, a binary classification model aims to distinguish between short-term and long-term incident durations, split by the incident clearance threshold $T_c$. Thus, the incident duration classification task is to predict $y_i^c$, where $Y_c$ takes one of the binary values:

$$\begin{cases} y_i^c = 0 & \text{if } y_i \leq T_c, \quad \text{short-term incidents} \\ y_i^c = 1 & \text{if } y_i > T_c, \quad \text{long-term incidents} \end{cases} \tag{2.3}$$

Here, $T_c$ denotes the incident duration threshold. By predicting whether an incident is short-term or long-term based on its features, the model provides valuable insights for traffic management and resource allocation during traffic incidents.

### Evaluation of prediction accuracy

To evaluate the regression or classification model performance the most commonly used metrics are represented in Table 2.6

The variables in the formulas are defined as follows:

- $n$: The number of samples or instances in the dataset.
- $A_t$: The actual value or true duration of the traffic accident at time $t$.

- $F_t$: The predicted value or forecasted duration of the traffic accident at time $t$.
- $tp$: The number of true positives, which are cases where the model correctly predicts the positive class (e.g., accidents with long duration).
- $tn$: The number of true negatives, which are cases where the model correctly predicts the negative class (e.g., accidents with short duration).
- $fp$: The number of false positives, which are cases where the model incorrectly predicts the positive class (e.g., predicting an accident as long-duration when it is actually short-duration).
- $fn$: The number of false negatives, which are cases where the model incorrectly predicts the negative class (e.g., predicting an accident as short-duration when it is actually long-duration).

Various metrics used for the regression tasks, RMSE and MAPE being the most common for the regression task (see Table 2.7).

### 2.4.2 Feature importance and feature selection

There are models of different complexity used to approximate traffic incident duration and duration of its phases. Khattak [108] used simple linear model to approximate clearance time, which can be defined as a function of incident parameters and coefficients:

$$clearancetime = A_1 * feature_1 + A_2 * feature_2 + ... + A_2 * feature_N \tag{2.4}$$

By using this simple model, we can easily determine the most important features by consequently removing them and estimating the resulting error (e.g. estimate change in mean squared error in relation to the removed feature). This method for determining feature importance is called Recursive Feature Elimination [79] and is a general method which is applicable to different approximation models.

The Shapley Additive explanation (SHAP) [148] offers an advanced methodology for estimating feature importance, as it combines estimations derived from multiple models trained on various subsets of the dataset. These subsets are selected based on both feature-scale and index-scale considerations.

A typical SHAP plot (Figure 3.14) visually represents the SHAP values and their impact on model predictions. In this plot, each point corresponds to a specific feature and is positioned along the vertical axis (Oy-axis) according to its SHAP value score. The points are color-coded based on their respective values, ranging from low to high. The horizontal axis (Ox-axis) represents the influence of each feature on the overall prediction output.

The SHAP approach offers several benefits when estimating feature importance. By aggregating information from multiple models and dataset subsets, it provides a more robust and comprehensive representation of the importance of each feature. Additionally, the SHAP plot allows for easy interpretation of the relationships between features and their impact on model predictions, making it an invaluable tool for understanding complex models and guiding feature selection processes.

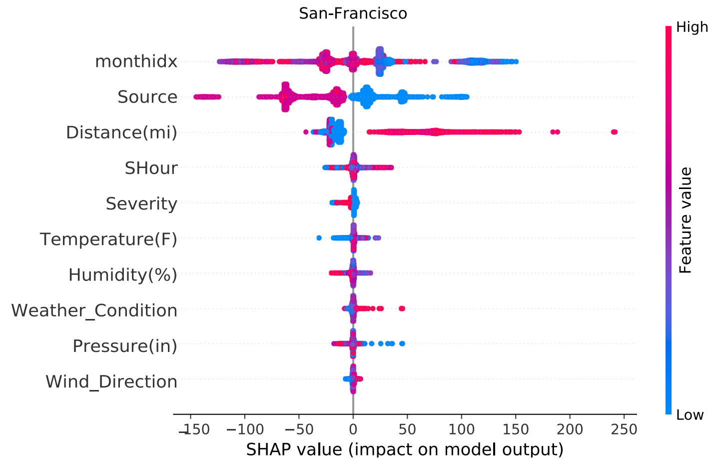



FIGURE 2.7: Feature importance for All-to-All regression using XGBoost for San-Francisco, USA [75]

### Features relevant to traffic accident modelling

In traffic incident duration regression analysis, achieving a balance between the number of features employed and the dataset size is crucial. Employing an excessive amount of features with a limited dataset size can lead to over-fitting. Some features might be beneficial, irrelevant, or have varying degrees of criticality, while others may not impact prediction results at all. By conducting a feature importance analysis, we can advise traffic management agencies to record the most crucial data and exclude less relevant information related to traffic incidents. Moreover, we can refine the precision of specific observations, such as weather conditions, which have been demonstrated to play a substantial role in certain research studies.

In traffic incident duration regression analysis, achieving a balance between the number of features and dataset size is crucial. The risk of overfitting arises when excessive amount of features is employed with a limited dataset size. To mitigate this risk, conducting a feature importance analysis is of high importance. This type of data-driven analysis helps to determine the relevance and significance of each feature in predicting traffic incident duration. Based on the results of such analysis, it is advisable for traffic management authorities to focus on recording and detailing crucial data while excluding irrelevant information. Furthermore, refining the precision of specific observations, such as weather conditions, is essential. Various studies show that weather conditions may have an impact on traffic incident behaviour [94].

For example, during the summer and autumn seasons in 2009, response team preparation time on freeways in Washington, USA, was notably higher [94]. Despite a higher preparation time, no significant effects were observed on clearance time and response team travel time. However, it was revealed that peak hours played a crucial role in response team preparation procedures, as the primary objective was to expedite incident resolution during peak hours. This finding underscores the importance of efficiently handling incidents during periods of high traffic flow. Moreover, a separate study examining Beijing traffic incident data from 2008 [130] also recognized the significance of "peak hour" in relation to response team travel time and clearance time, but not in terms of response team preparation time. These insights contribute to a better understanding of response procedures for incidents on freeways, highlighting the importance of quick and efficient response during peak hours.

By exploring the effects of various features and their significance on incident duration, researchers can develop more accurate models that take into account critical factors while discarding less relevant ones. This can lead to improved prediction performance and better decision-making in traffic management and incident response strategies.

A promising approach for analyzing feature importance is using decision trees from tree-ensemble models, as proposed in [38]. A data-driven technique can be employed to fuse information from multiple sources, as exemplified in [2], where the Gini-index, extracted from Random Forests, was utilized to estimate feature importance. However, it is crucial to recognize that a random model might exhibit significant variability in data mapping when the relationship between features and the target variable is weak. This can result in the feature importance value being heavily influenced by the randomness employed in the model evaluation.

In order to address this challenge, it is essential to consider multiple random seeds, model ensembles or use methods that provide a more stable and robust estimation of feature importance. Techniques such as the Shapley Additive explanation (SHAP) [148] offer more advanced approaches for estimating feature importance, as they fuse estimations from multiple models trained on various subsets of the dataset. This approach can help to mitigate the influence of random seeds and provide more reliable insights into the importance of different features.

By employing a combination of data-driven methods and ensemble techniques, researchers can more accurately assess the importance of various features in the context of traffic incident duration prediction. This can lead to the development of more effective models that take into account critical factors while discarding less relevant ones, ultimately improving prediction performance and enhancing decision-making in traffic management and incident response strategies. Furthermore, adopting a data-driven approach tailored to specific objectives can emphasize the types of data that need to be collected for particular models and prediction tasks, such as accident risk, duration, or severity prediction.

By focusing on the unique requirements of each task, researchers can effectively identify the most relevant features and data sources, which in turn can enhance the performance of the predictive models. This targeted approach can also help to streamline data collection efforts, ensuring that resources are allocated efficiently and that redundancy in data collection is minimized.

A research study using Beijing traffic incidents data from 2008 [130] found the importance of "peak hour" value for the response team travel time and clearance time, but not for response team preparation time. Surprisingly, during summer and autumn, response team preparation time was higher, with no noticeable effect on clearance and response team travel time. The type of incident "overturned vehicle" found to have significant effect on response team preparation and incident clearance time with no effect on travel time. The most important factors for the total incident duration time were: 1) bike involvement, which can imply human injury and 2) night shift (10pm-6am), which was linked to higher incident severity and consequently higher clearance time.

One important conclusion is that rather than using the number of affected lanes[159] (which is only part of information related to road segment), we can also add a resulting value called “all lines affected”[102] or even more informative - ratio of affected lanes (e.g. 50%) indirectly incorporating the number of lanes of a section, where incidents occurred into the model. Also, we can include the number of lanes for the each involved section as a feature.

Modelling of freeway incident response time in Washington State, USA in 2009 [94] found morning peak hours as the most influencing feature on response team preparation delay, which was found to be linked to response procedures (goal of response team was to resolve incidents during peak hours as soon as possible). Response time was also lower during summer and winter. Because of the response strategy of the transport agency we can observe a clear difference in incident duration due to response priority. Priorities in the elimination of traffic incidents (which consider, for example, number of blocked lanes) can be defined as a part of traffic incident response guidelines. Priority of traffic incident elimination (and corresponding response team actions, like involvement of towing vehicle) can be used as a feature estimated from the traffic incident characteristics (and act as an additional information to the model).

Weekend and nightly incidents were also associated with significantly longer clearance and impact (including recovery) duration of traffic incidents on freeway [85]. This observation was attributed to a lower number of staff on duty during weekends and nights [85], [94].

The average traffic speed during a 60-min time interval was found to be statistically significant for the traffic incident duration modelling [91]. One can use aggregated traffic flow data (e.g. represented as a speed of traffic or possibly traffic flow count) within specific time intervals as a feature.

Lighting conditions can be calculated much more precisely than just using binary day/night values [85]. Using longitude, latitude and time, we can determine the angle between the sun and traffic direction at the time and place of the incident. Also, we can calculate precise lighting conditions based on the elevation of the Sun above the horizon during the time of the incident.

Effects of lighting conditions on driver behaviour were assessed using an interactive driving simulator [92]. According to the research, driver perception was found to be limited during night-time; drivers were also found to be limiting travel speed due to impaired visibility (which includes an incorrect and untimely perception of the road). Installation of road markers (to improve the perception of the road curve) and rumble strips were proposed. The impaired visibility during night-time and the proposed measures point that it is possible to derive visibility of the road structure from the point of traffic incidents and corresponding lighting conditions (distance of driver's eyesight in relation to road segments).

### 2.4.3 Interpretable modelling and data visualization

Interpretable machine learning, as defined in [168], involves extracting relevant knowledge from machine learning models to provide insights into the problem domain, guiding further actions and discovery. This knowledge can include feature importance, dataset visualization, or learned relationships within the data.

Various methods and techniques can be employed to create interpretable models:

- **Decision-tree-based methods:** These methods represent the model as an interpretable decision tree, illustrating the classification decisions and feature importance estimates. Examples include Classification and Regression Trees (CART) and C4.5 [120].

- **Knowledge distillation:** This technique is utilized to extract decision rules from different prediction models, such as Bayesian networks [182], and can be used to create interpretable tree or rule-set models.

- **SHAP (SHapley Additive exPlanations):** SHAP values [148] provide a unified measure of feature importance that can be used with any machine learning model, allowing for interpretability and comparison across models.

- **UMAP (Uniform Manifold Approximation and Projection):** UMAP [158] is a dimensionality reduction technique that creates interpretable visualizations of high-dimensional datasets while preserving local and global structures. Can be fine-tuned either for machine learning tasks (classification, clustering, etc) or visualization.

- **t-SNE (t-Distributed Stochastic Neighbor Embedding):** t-SNE [152] is another dimensionality reduction technique that excels at visualizing high-dimensional data by placing similar data points close together in a lower-dimensional space, making complex structures more interpretable.

- **Bayesian networks:** These models can produce interpretable representations for predicting incident injury severity [174]. They have been shown to outperform regression models in incident severity prediction, involving indicators such as the number of fatalities, injuries, and property damage [271].

In summary, interpretable machine learning offers valuable insights by extracting relevant knowledge from various models, which can then be utilized by decision-makers. Decision tree-based methods, knowledge distillation, SHAP, UMAP, t-SNE, and Bayesian networks are particularly useful for creating interpretable models and estimating feature importance. SHAP's flexibility, being applicable to any machine learning model, makes it a particularly valuable tool for model interpretability.

### 2.4.4 Imbalanced dataset classification

In traffic incident duration prediction, we may encounter rare incidents with durations exceeding 60 minutes, representing extreme values. Traffic incident duration is often modeled using Log-normal, Gamma, and Weibull distributions, as they exhibit asymmetric and long-tailed distributions. Thus, classification methods for imbalanced datasets can be beneficial for classifying these rare incidents.

Three primary approaches for handling imbalanced classes include:

- Under-sampling - Reducing the size of the majority class to achieve class balance.

- Over-sampling - Increasing the size of the minority class (e.g., by generating synthetic samples) to achieve class balance.

- Combined approach - Utilizing both under- and over-sampling techniques to balance class proportions [188].

Situations in which these approaches can be used can be attributed to data set sizes and computational resource availability (see Table 2.8). For example, in the case of limited computational resources, the data may be sub-sampled (e.g. 20% of data points can be extracted from the total amount of data) to perform experiments and establish and test the prediction pipeline before deploying the model and using it on complete data. A combined approach may be used to perform a under-sampling (to attribute to computational resource limitations) and then over-sampling of minority class to achieve the class balance. But it is necessary to state that these approaches are used to enhance the model performance during model training and should not be used on the validation data set. It's important to maintain the original distribution of the validation set, so it accurately represents the real-world data that the model will encounter. It is important to split your data into training, validation, and test sets before applying these techniques. If you were to perform over-sampling or use a combined approach before splitting, information from the validation/test sets could inadvertently leak into the training set, leading to overly optimistic performance estimates.

Various methods implement each approach for addressing imbalanced datasets:

- Random under-sampling [214] - This method involves randomly selecting samples from the majority class to achieve better balance with the minority class.
- Tomek Links [260] - This technique identifies pairs of instances in which one is from the majority class and the other is from the minority class, and they are nearest neighbors. The majority class instance in each pair is removed to balance the dataset.
- SMOTE - Synthetic Minority Over-sampling Technique [32] - This method generates synthetic examples for the minority class by selecting random points on line segments between k neighbors in a multidimensional space.
- Borderline-SMOTE [84] - An improvement over SMOTE that focuses on generating synthetic samples near the decision boundary, where the minority class is at a higher risk of being misclassified.
- ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning [87] - This technique uses the distance distribution among k neighbors to determine the number of synthetic samples to generate. The sample generation process is similar to SMOTE, but the number of samples generated is adaptive based on the frequency of the distances among k neighbors.
- Granular Repetitive SVM Under-sampling [217] - This method leverages the Support Vector Machines (SVM) algorithm for under-sampling since only a few support vectors are crucial for classification in SVM. By selecting instances near the decision boundary, this technique reduces the majority class while maintaining informative samples.

These techniques can be combined or modified to create tailored solutions for specific imbalanced dataset classification problems (see Table 2.8). The choice of method should be guided by the dataset's characteristics, the problem domain, and the desired classification performance.

### 2.4.5 Boosted models and ensembles

Different machine learning techniques exist in the literature and each demonstrates different extrapolation performance; AdaBoostSVM and Extreme, for example, are machine learning approaches which has to property to make predictions with reduced overfitting.

AdaBoost is a meta-estimator, a machine learning method which implies a training of a set of weak classifiers with adaptive change to weights samples depending on the correctness of classification (after each boosting operation, weights of samples are adapted so next trained weak estimator gives higher priority to miss-classified samples) [200]. Classifiers are then combined into weighted voting (linear) combinations.

AdaBoost and SVM were combined to solve the classification problem of imbalanced datasets [233] using spatial-temporal traffic data for the case of automatic incident classification.

Extreme Learning Machines (ELM) [124] - is a machine learning method, which incorporates a feed-forward neural network initialised with random weights and consequent training step based on produced random feature mapping, designed to avoid overfitting of neural network. The method is two-step: 1) Neural network initialised using random weights (this way we perform feature mapping into ELM feature space) 2) Then Moore-Penrose generalised inverse performed on the hidden layer output matrix and solving feature classification problem using Gaussian estimation. ELM is very fast to train and provides better generalisation than Artificial Neural Network, also demonstrated remarkable efficiency in comparison to SVM, Naive Bayess and ANN on the traffic incident detection task on I-880 freeway in California [124].

### 2.4.6 Anomaly/outlier detection

A majority of studies in the literature have primarily focused on applying state-of-the-art machine learning models for classifying incident severity, as seen in [171], or predicting their duration, as demonstrated in [129]. However, very few studies have addressed the issues of outliers or imbalanced data classes, which can significantly impact the performance of these models.

Numerous studies, such as [233], [18], and [124], have employed classification or clustering methods for detecting traffic incidents based on the analysis of traffic conditions. However, there is a noticeable lack of research on traffic incidents that utilize anomaly detection techniques, such as one-class SVM, isolation forests, and others. Non-recurring traffic incidents are rare and atypical by nature, and thus detecting these incidents can be framed as an anomaly detection task within traffic data. By leveraging anomaly detection methods, incident detection systems can be adapted to recognize previously unseen situations, enhancing their overall effectiveness.

In this context, it is crucial to compare and evaluate applicable anomaly detection methods alongside well-established classification and regression techniques. This comparison can help identify the most suitable approaches for detecting traffic incidents, particularly those that are non-recurring and difficult to predict. Furthermore, identifying road situations detected as anomalous can be invaluable for conducting in-depth investigations during the early stages of a newly reported accident.

Different anomaly detection methods will be used, including those which can produce a measure of an anomaly for each data point (One-Class SVM, Isolation Forest). This will be used to compare anomaly detection with regression methods (GBDT, ANN) for the task of incident probability estimation. Similarly, anomaly detection can be used in comparison with recently used classification (GBDT, ANN) and regression (e.g. Gaussian process regression) methods, for the task of incident duration estimation. As stated in [150], tree models perform badly at modelling incident duration with the long-tail distribution. Thus, anomaly detection methods can be used to model such kind of rare duration (which is placed in the long tail) as anomalies of duration.

The Isolation Forest (IF) method, as proposed in [140], is an innovative outlier removal technique that employs forests of random split trees for isolating anomalies within the data. For each tree in the forest, the algorithm randomly selects a feature and a corresponding feature value. The dataset is then recursively partitioned into two subsets at each step until each data point becomes "isolated" (separated from the remaining data points). Outliers tend to have a small tree depth, meaning they are easily separable after only a few splits across the selected features. The tree depth is then averaged across all the "isolation" trees and used as an anomaly score for each data point. For example, if a data point has an average tree depth of 1.3, it indicates that the point can be quickly isolated through a limited number of splits.

On the other hand, the Local Outlier Factor (LOF) method, introduced in [28], estimates the anomaly score based on the local deviation of data density within the k-nearest neighborhood. LOF relies on calculating the local reachability density (LRD) for each data point, which represents the inverse of the average reachability distance (RD) from the neighboring data points to the selected point. The RD, in turn, signifies the distance to the most distant neighbor within a k-sized neighborhood (where k is a hyperparameter). The LOF score for a data point is then calculated as the ratio between the LRDs of its neighbors and its own LRD. LOF scores can be greater than 1 (indicating higher LRD than neighbors), less than 1 (lower LRD than neighbors), or equal to 1 (similar data density as neighbors). By sorting the data points based on their LOF scores, we can select a specific percentage of data points with the highest LOF values to be removed as outliers. The LOF method operates on the assumption that outliers are situated in areas with low data point density, whereas non-outliers reside in high-density regions.

In summary, both Isolation Forest and Local Outlier Factor methods offer unique approaches to identifying and removing outliers within a dataset. While IF leverages the concept of tree depth in random split trees, LOF exploits the local deviation of data density within a predefined neighborhood.

The IF method can be applied to identify unusual traffic accidents based on their anomaly scores. These scores will help in isolating incidents that deviate significantly from the majority of the data, potentially revealing underlying patterns or factors that contribute to these outliers. For example, a cluster of high-scoring accidents (based on normality score) in a specific location might indicate a problematic road design that needs attention.

Similarly, the LOF method can be utilized to identify traffic accidents that deviate from their local neighborhood in terms of data density. By analyzing the LOF scores, one can pinpoint accidents that are unusual compared to their surrounding incidents. This information can be valuable for identifying specific situations where accidents tend to be more uncommon, helping authorities prioritize interventions and improvements.

The Isolation Forest (IF) and Local Outlier Factor (LOF) methods can be effectively applied to the analysis of traffic accident reports, providing valuable insights and helping identify unusual patterns or incidents.

In conclusion, applying the Isolation Forest and Local Outlier Factor methods to the analysis of traffic accident reports can help identify unusual incidents and patterns that may otherwise be overlooked. By recognizing these outliers, authorities can make informed decisions to improve road safety and ultimately reduce the number and severity of traffic accidents.

Outlier detection can be described as an activity of data analysis. Data analysis can be combined with clustering methods resulting in cluster-level analysis of accident reports. It recently resulted in traffic incident duration modelling for specific clusters with the use of model ensemble. Similar to this approach, clustering methods, such as K-Means, DBSCAN, or Hierarchical Clustering, can group similar incidents together, while IF and LOF can identify unusual or outlier events within these clusters.

One-class SVM, Covariance estimator, Local outlier factor method and Isolation Forest has been applied for outlier detection in machine learning pipeline to predict rail-road accidents [29]. Anomalies detected in historical traffic state data (traffic flow, traffic speed, occupancy) and vehicle trajectory [53], which can be associated with disruptions produced by traffic accidents. Three main approaches for anomaly detection include the use of statistical models, similarity-based models (which rely on data difference measures and neighbourhood estimation methods to find outliers) and pattern-mining methods (which resolve the correlation problem of similarity-based models but are very time-consuming).

Multiple models in many research studies failed to predict extreme values for the traffic incident duration [130], [207]. Machine learning method GBDT which demonstrates superior performance on a wide variety of tasks also known as failing at predicting very long incident duration, which is represented as extreme values within part of long-tailed distribution [150].

In conclusion, anomaly detection can be an effective tool for improving the estimation of incident probability and duration, as it can identify and isolate rare events (or anomalous accident reports) and also can eliminate data records which can contribute to the bias of the prediction model (which can negatively affect model's performance [246]). As such, anomaly detection can be a valuable addition to existing methods when attempting to model complex data sets.

### 2.4.7 Dimensionality reduction methods

Usually, traffic flow data is represented as traffic state readings across multiple (up to hundreds) vehicle detectors in the transport network. Options to use such a high-dimensional data set include: 1) use the closest to the point of interest readings (e.g. traffic flow on 5 closest road segments) [159] 2) use all the data available on traffic flow in the network [159] 3) Perform dimensionality reduction before modelling [233].

Principal Component Analysis (PCA) is a statistical method that uses an orthogonal transformation to transform multidimensional data into a sequence of linearly uncorrelated variables (components). Each resulting component is a linear combination of input features. The very first component has the greatest variance. The method was proposed in 1901 by Pearson [245]. PCA was used to reduce 23 dimensions of spatial-temporal signals to 2 dimensions for the task of Automatic Incident Classification proposed in [233].

Uniform Manifold Approximation and Projection (UMAP) [158] is a new dimension reduction method (2018), based on the use of Riemann geometry and algebraic topology. UMAP optimises the placement of data in a small dimension space to minimize cross-entropy between two topological representations. The disadvantage of the method is its non-interpretability and non-invertibility (one cannot inverse mapping to feature data). UMAP can be used both for dimensional reduction and data visualisation. To perform clustering and modelling of data points in reduced dimensions and to perform the data visualisation, it is necessary to use different sets of options (e.g. minimal distance between points in reduced dimension needs to be set to zero when preparing data for clustering).

Dimensionality reduction methods can help when we have multidimensional data with a lot of dimensions. A lesser number of dimensions can provide simpler and faster models or allow to process of large data sets within computational resource constraints.

### 2.4.8 Summary on the use of Machine Learning models

The majority of studies rely on Support Vector Machines or Naive Bayes for traffic incident duration prediction (see Table 2.9). The use of more advanced methods like XGBoost or GBDT is rare which is surprising given their effectiveness. This can be explained by a generally slow attribution of both sophisticated models and data-driven approaches to the traffic accident research which we observe in the literature. Also, the complexity of machine learning pipelines has increased in recent years due to the need for the incorporation of more data science knowledge to merge with traditional transport modelling techniques.

Another finding is that various Machine Learning pipeline elements (like dimensionality reduction or feature selection) are rarely used across incident duration prediction studies (see Table 2.10). This reflects a lack of advanced machine learning techniques that can be explored for incident modelling. The lack of popularity among these pipeline approaches leaves room for more innovative ideas for data filtering and feature ranking from the beginning of the incident duration prediction modelling. The use of SHAP and feature selection is found to be lacking in studies, while it may provide a list of entries in the incident reporting form with the highest contribution to the accuracy of the traffic incident duration prediction. An additional description of the most important features may provide a further increase in prediction accuracy. Bi-level frameworks allow the separation of the task of the incident duration prediction into incident duration classification and regression tasks. Allocating different kinds of models to each task may improve the overall model performance. The use of bi-level frameworks as a technique for prediction performance improvement is also found to be rare. Dimensionality reduction had a low relevance in multiple prior studies due to the small size and low dimensionality of incident reports (see the table of data set sizes used in years prior to 2018 [129]), but with current advancements in traffic data collection (see Section 2) and significant consequent increase in amount and variety of data collected, we see a rise in the relevance of data pre-processing methods.

## 2.5 Deep Learning in traffic accident analysis

Another example of classification task in accident modelling includes accident detection and accident risk prediction, which can rely on high-resolution traffic data (speed, occupancy, volume) and use Deep Learning methods like Convolutional Neural networks [97]. Variational Long Short-Term Memory Encoder has been proposed in [58] to perform a short-term traffic flow prediction. It demonstrated better performance than LSTM and Stacked Autoencoder models. Short-term (5 to 30 minutes), mid-term (30 minutes to multiple hours) and long-term prediction (day to multiple days) capabilities of the method were explored. The methodology is in general intended to predict the traffic state using historical data.

For future research, it is essential to consider the diversity and complexity of traffic accident data. ML methods may be more suitable for tasks with simpler patterns, smaller datasets, and limited computational resources. They also offer the advantage of being more interpretable, which could be beneficial for traffic management authorities aiming to understand the underlying relationships between accident factors. On the other hand, DL methods excel in tasks involving complex patterns, spatial-temporal dependencies, and high-dimensional data, while requiring a larger amount of data than ML models. These methods can automatically learn features, potentially improving modeling efficiency and reducing the need for manual feature engineering or domain knowledge involved in ML methods.

To address the challenges and limitations associated with both ML and DL methods, researchers should explore hybrid models that combine the strengths of both approaches. For instance, ensemble techniques that integrate ML and DL models could be utilized to leverage the interpretability of ML methods while benefiting from the advanced pattern recognition capabilities of DL methods.

Moreover, traffic management authorities should prioritize data collection, as well as the establishment of standardized protocols for data integration. This will enable researchers to develop more accurate and robust traffic accident prediction models, which can ultimately lead to more effective traffic management strategies and accident prevention measures. Data standardization is crucial since it allows for the exchange and rapid deployment of existing models to new traffic networks, also it allows data sets to grow larger because of the combination of multiple data sets from different traffic networks and even countries.

Data standardization is a fundamental aspect in the development of machine learning and deep learning models for traffic incident analysis, as it allows for the transfer and integration of existing models to new traffic networks, thus improving the adaptability of approaches to the extensions or modifications of traffic networks. Furthermore, standardization facilitates the combination of multiple data sets from various traffic networks and even countries, which enhances the robustness and specificity of the models. By having standardized data, Deep Learning models can be trained on the data from a broader range of traffic networks. Therefore, in the near future, traffic management authorities should prioritize data standardization efforts to facilitate the development and deployment of more unified protocols and systems on data collection regarding traffic speed, flow and accident report data.

Deep learning is becoming increasingly important in the field of traffic accident modelling. By leveraging the power of artificial intelligence and machine learning, deep learning algorithms can be used to detect patterns in historical accident report data that may be indicative of future accidents [172]. This methodology can help to develop more effective strategies for preventing accidents, as well as allow traffic management agencies to analyze existing road structures for the probability of accident risk. Deep learning has been used to develop predictive models that can identify accident-prone areas [191] and predict accident risk based on driver behaviour [208], as well as to predict the impact of various events on traffic flow [258].

Overall,the majority of studies implementing Deep Learning techniques for traffic incident duration prediction rely on classic ANN/MLP and rarely use recurrent or convolution networks (see the summarising Table 2.11 of the most popular deep learning techniques used so far). The use of recurrent networks is mostly attributed to the analysis of textual incident reports or messages over social networks. This short summarising of Deep learning shows that there is a significant gap in the current incident modelling to leverage such powerful techniques and this is mostly related to the data availability - traffic flow counts, speed, details traffic signal control, etc. There are different approaches that can be used to improve the incident duraiton prediciton and in the following subsections we detail the most common ones as follows.

### 2.5.1 Machine Learning vs Deep Learning methods

The following table (Table 2.12) presents a detailed comparison between Machine Learning (ML) and Deep Learning (DL) methods for traffic accident modeling. The comparison encompasses various aspects, including model complexity, feature engineering, data requirements, computational requirements, interpretability, and applicability. By examining the pros and cons of each approach in the context of traffic accident modelling, this table aims to provide insights for researchers and practitioners to make informed decisions when selecting the most suitable approach.

The choice between Machine Learning (ML) and Deep Learning (DL) methods for traffic accident modeling largely depends on the specific problem, data availability, and computational resources. Also, both ML and DL methods have their advantages and limitations when applied to traffic accident modelling tasks.

### 2.5.2 Spatial-temporal models for traffic incident modelling

In recent years, the field of traffic accident research has experienced a growing reliance on data-driven methodologies. Various problem areas have been addressed through these methods, including:

- Traffic accident duration prediction: A range of deep learning techniques has been developed to predict the duration of traffic events [234], [269], [172]. These methods leverage historical time series data combined with recurrent models (e.g. Long-short Term Memory networks (LSTM)) and their derivatives or hybrid models (e.g. Convolutional LSTM, Seq2Seq models).

- Accident detection: Various Deep Learning techniques have been employed to automatically identify traffic accidents [123], [190]. These methods utilize sensor data, traffic flow information, and other relevant factors to detect incidents in real-time, with the aim of facilitating prompt response and minimizing the impact on traffic.

- Estimation of severity: Research in this area, focuses on assessing the severity of traffic accidents, taking into account factors like vehicle speed, weather conditions, and road infrastructure [189], [209]. In particular, traffic accident features were represented as an image [189] to be processed by Computer Vision methods of Deep Learning approach. Also, a customized loss function was used to optimize the model for the most relevant metrics for the task (precision and recall). Accurate severity estimation can help prioritize emergency response efforts and inform preventative measures.

- Spatial-temporal modeling methods: Spatial-temporal models mainly using two-dimensional grid representation and used to predict the traffic accident risk. More recently, advanced modelling techniques (e.g. double-path ConvGRU network [240], or using Visual Transformers

[196]) have emerged, enabling the traffic accident risk prediction using high-dimensional spatial, semantic (Points-of-Interest proximity), and temporal data. These approaches consider the complex patterns between various factors to forecast accident-prone locations based on historical accident frequency data and predict the accident occurrence within the future time-frame.

The application of these data-driven methods has significantly enhanced the automated analysis of traffic data, particularly with the increasing availability of public datasets. Traffic accident risk prediction offers several key benefits, including:

- Identifying high-risk areas within a traffic network: This allows traffic management authorities to make informed decisions regarding infrastructure improvements and resource allocation.
- Assessing road design and resource allocation: By pinpointing areas with a high likelihood of accidents, authorities can evaluate existing road designs and make necessary adjustments to mitigate future incidents.
- Timely prediction of high-risk situations: Predicting high-risk situations on the road enables authorities to implement appropriate traffic management strategies, reducing the likelihood of accidents and minimizing their impact on traffic flow.
- Implementing risk-reducing traffic management strategies: Accurate risk prediction facilitates the development and implementation of targeted, effective strategies to improve overall road safety.

In summary, the increasing use of data-driven methodologies in traffic accident research has greatly enhanced our ability to predict, detect, and assess incidents. These approaches have far-reaching implications for traffic management, enabling authorities to make informed decisions and implement effective strategies to promote road safety.

In recent years, deep learning techniques have been increasingly employed for traffic accident risk prediction, offering more advanced and accurate methods to forecast the likelihood of incidents. One of the pioneering studies in this area utilized a Stack Denoise Autoencoder (SDAE) for risk prediction based on human mobility data within the Japanese traffic network [35]. However, this research did not take into account traffic flow or time-related factors, such as periodicity.

Subsequently, [192] explored the use of Long Short-Term Memory (LSTM) networks to enhance risk prediction compared to SDAE. This approach incorporated additional data on air quality, traffic flow, and weather, represented as short-term and periodic components, offering a more comprehensive analysis of contributing factors. In [267], the authors proposed a Coarse and Fine-grained prediction methodology applied to the target accident risk map, further refining risk estimation.

RiskOracle [268] employed a Graph-Convolution network and utilized hierarchical coarse-to-fine modeling to provide minute-level predictions, as opposed to day-level [259] or hour-level [35] predictions. In [259], the authors addressed the spatial heterogeneity problem by developing an ensemble of region-specific ConvLSTM models, known as Hetero-ConvLSTM. This approach considered weather, environmental, and road conditions in Iowa, US, over an 8-year period, but did not account for points of interest (POIs).

Incorporating semantic features, coarse and fine-grained risk maps, [237] utilized Graph-convolution neural networks and attention-based LSTMs for improved prediction. More recently, [240] has emerged as the state-of-the-art (SoTA) in accident risk prediction. The authors proposed a weighted loss function to address the zero-inflated issue, which arises from an increase in zero-risk grid cells due to higher granularity in predictions. They also developed an ensemble of models that process semantic and geo features for more accurate risk assessment.

In summary, deep learning methods have significantly advanced the field of traffic accident risk prediction, incorporating a wide range of factors and offering more accurate and granular forecasts. These techniques have evolved over time to address various challenges, such as periodicity, spatial heterogeneity, and zero-inflation, providing valuable insights for traffic management and road safety efforts.

Deep learning algorithms can provide a better understanding of accident risks and impacts in traffic networks in which drivers are operating, allowing traffic management agencies to develop more effective strategies for preventing accidents or mitigating their impact. Additionally, deep learning algorithms have the potential to improve the accuracy and speed of emergency response. However, the use of deep learning in traffic accident prediction is not without drawbacks, such as the need for large amounts of data and the potential for bias in the results due to the algorithm's reliance on existing data [219] (these algorithms may need to be tested against extrapolation between time intervals, different areas and reporting source). Nevertheless, deep learning methodology allows to develop important tools for traffic authorities, as it can provide valuable insights into the causes of accidents, and data patterns which can point to potential risks and impacts and in total, help traffic management authorities to develop better strategies for preventing, responding and reducing the impact of traffic accidents.

### 2.5.3 Textual Accident Description analysis

Traffic accident reports usually contain a textual description of the accident [165] [166]. In recent years, multiple systems were presented to detect traffic accidents using text analysis of social networks content [228], [195], [9]. Various methods were also proposed for traffic-related sentiment analysis of social networks: sentiment classification using ontology and latent Dirichlet allocation [10], the use of gated recurrent unit (GRU) model and generative adversarial networks to estimate the traffic information sentiment [30]. Overall, sentiment analysis has been performed for various traffic rules including 'yellow light rule' using social network analysis [30], [145].

Also, accident reports can contain a category and subcategory definition of the accident (e.g. types of vehicles involved, multiple or single-vehicle crash, etc). The unique property of text description of an accident is that it can contain information regarding event categories not predefined in the accident reporting form [228].

A typical pipeline for textual description preprocessing includes [9]: 1) Tokenization - text being split into a list of words called tokens, 2) Stop-word removal - the removal of pronouns, prepositions, symbols and articles not providing any valuable information for accident description, 3) Lemmatization and stemming - words are reduced to their base form (e.g. involved -> involve, injuries -> injury, reported -> report) or to their root form (e.g. injuries, injury -> injur) 4) Case-conversion - text is converted into lower case, where the difference between uppercase and lowercase words is not relevant 5) Part-of-speech (POS) tagging - each word gets its type associated to it (e.g. traffic ->

noun, stop -> verb), 6) text representation conversion, which relies on a Bag-Of-Words (BOW) representation (each word is represented as a one-hot encoded vector) or on a neural-network-based word embedding method like Word2Vec or FastText, which capture semantic similarities between words. The Word2Vec approach has a significant limitation - the inability to represent a new word which was absent in the training data set with a vector. The FastText resolves this issue by representing each word as a sum of related n-gram vectors.

After the data preparation and representation conversion, various recurrent models are then used to perform tasks related to text analysis.

Incident description features were used in topical text modelling [50]. Previously, the LSTM architecture has been used for the task of detection of incidents from social media data [264]. LSTM was also successfully used for stock price time-series prediction [203], making it applicable for the modelling of traffic flow/speed time-series data.

Text analysis of accident reports is vital for understanding the underlying causes of traffic accidents. By providing insight into the information provided to describe accidents, text analysis can help to identify dangerous conditions that lead to accidents. Traffic accident descriptions can contain inaccuracies due to human factor (e.g. inaccurate accident timeline), which highlights the importance of automated accident detection and timeline estimation from traffic state data [213]. Ultimately, text analysis is a powerful tool for gaining even deeper insight into the causes of traffic accidents unconstrained by accident reporting forms and developing strategies to reduce them.

### 2.5.4 Implications of Artificial Intelligence development on traffic accident analysis

The accelerated progress in the area of Artificial Intelligence (AI) profoundly impacts various domains, including the analysis of traffic accidents. AI's capacity to handle large datasets, forecast hazardous situations, and augment real-time responses can markedly transform our approach towards understanding, preventing, and addressing traffic accidents in the near future. There are multiple ways AI is reshaping the arena of traffic accident analysis and prevention:

- Enhanced Data Collection and Analysis, growing amounts of data: The development of computer hardware and software in recent years has greatly improved the collection, processing, and analysis of traffic data. As a consequence, the amount and dimensionality of data is rapidly growing. Artificial Intelligent systems can benefit from large volumes of data. These methods can also process data from various sources such as videos from traffic cameras, GPS devices, high-dimensional geo-spatial data and messages on social media platforms. This may help to identify accident patterns, contributing factors, and high-risk areas more efficiently, enabling targeted interventions to reduce the frequency and severity of accidents in identified areas.

- Real-time Monitoring and Response: traffic surveillance systems (including video data from traffic cameras) can utilize AI mdethods for monitor traffic conditions in real-time, detecting incidents as they occur. This enables quicker response times from emergency services, potentially saving lives and reducing the severity of injuries.

- Enhanced Accident Analysis: AI can assist in reconstructing traffic accidents more accurately by analyzing data from multiple various sources simultaneoulsly, such as vehicle sensors, traffic cameras, Points-of-Interest in proximity, accident reports, traffic speed and flow readings from nearby traffic detectors. This can help identify the highly-complex patterns of related factors leading to the accident affecting the design of future safety measures.

- Pre-accident Analysis: traffic authorities can leverage machine learning algorithms to anticipate potential accidents before they occur. By analyzing data from diverse sources such as weather reports, historical accident data, and real-time traffic conditions, AI can predict high-risk areas and the likelihood of an accident in a specific time frame. These predictive insights could help authorities implement preventive measures and drivers adjust their routes or driving behaviour.
- Accident-related Medical Assistance Prediction: AI can also play a crucial role in predicting necessary medical assistance. By analyzing accident report data (on the moment of report), AI can assist in predicting the likelihood of certain injuries based on the nature of the accident, traffic state readings in proximity. This can help emergency responders and hospitals prepare more effectively for incoming patients, potentially improving outcomes.
- Enhanced Emergency Services Dispatch: AI systems can aid in the swift and efficient dispatch of emergency services in the event of an accident. By analyzing data from the accident, like the severity of the crash, the number of vehicles involved, and the potential number of injuries, AI can make informed recommendations about the resources required at the scene.
- Enhanced Road Design and Planning: Data-driven insights from AI can help urban planners and engineers design safer roads (e.g. by manipulating the road design in software and estimating possible accident occurrence rate and predicted accident severity outcomes using AI model). By analyzing accident data, AI can identify problematic road designs that lead to higher accident rates.
- Economic Implications: Reducing the number of traffic accidents through AI could have significant economic implications. Fewer accidents mean less money spent on healthcare, vehicle repairs, and insurance claims.

In conclusion, AI has the potential to enhance accident-related data collection and analysis, real-time response (of both clearance and medical assistance teams), accident pattern analysis, and road design assessment, and result in economic savings.

| Method | Description | Traffic Accident Duration Prediction | Advantages | Limitations |
|---|---|---|---|---|
| Light Gradient Boosted Machines (LightGBM) [107] | Enhances gradient boosting in tree-based models, employing Gradient-based One-Side Sampling (GOSS) to exclude data points with small residuals. | LightGBM predicts traffic accident duration by using a more efficient and scalable approach to gradient boosting. By focusing on data points with large residuals through GOSS, it achieves faster training while maintaining accuracy. | Faster training than GBDT and XGBoost<br>Scalable to large datasets<br>Can handle missing values and categorical features | May require tuning of hyperparameters<br>Sensitive to noisy data<br>Could perform worse than XGBoost due to non-greedy tree split search |
| CatBoost [54] | A powerful gradient boosting library developed for efficiently handling categorical features without extensive preprocessing, using Ordered Boosting and Oblivious Trees. | CatBoost is fit for traffic accident duration prediction as it efficiently processes typical categorical features of accident reports like road types, weather conditions, and vehicle types, and can be applied to both regression and classification tasks. | Designed for handling categorical features<br>Robust to overfitting<br>Applicable to regression and classification tasks<br>Can handle missing values | Computationally expensive for large datasets<br>Hyperparameter tuning may be necessary |
| Extreme Learning Machines (ELM) [96] | A single-hidden-layer feedforward neural network (SLFN) initialized with random parameters, where optimization performed only for the output layer's weights. | Extreme Learning Machines are suitable for traffic accident duration prediction on large data sets, as they provide faster training compared to traditional feed-forward neural networks while maintaining good accuracy. The method randomly assigns input weights and biases, focusing solely on the optimization of weights of the output layer. | Faster training than traditional neural networks<br>Can handle non-linear patterns | May struggle with complex patterns or categorical data (may require one-hot encoding)<br>Less interpretable than other methods |
| Adaptive Boosting (AdaBoost) [156] | This algorithm combines multiple weak learners into a strong learner, adjusting training instance weights adaptively. | AdaBoost can be employed for predicting traffic accident duration, forming a strong predictive model by combining multiple weak learners. By adaptively adjusting the weights of training instances, the model focuses on harder-to-predict instances, enhancing overall accuracy. | Robust to overfitting<br>Resistant to noise<br>Meta-model: can use various weak learners (e.g. decision trees) | May be computationally expensive based on the weak learner type and depth of boosting<br>May necessitate weak learner hyperparameter tuning |
| Gradient Boosting Decision Trees (GBDT) [60] | A technique that trains a sequence of models, where each subsequent model is added to minimize the residuals of the previous models. | GBDT is suitable for predicting traffic accident durations by progressively enhancing prediction performance based on the residuals of prior models. The final prediction is an ensemble of all weak learners. | Resistant to overfitting<br>Handles missing values effectively<br>Accommodates mixed feature types | Sensitive to noisy data<br>Longer training time (typical for boosting methods) |
| eXtreme Gradient Boosting (XGBoost) [37] | Determines split values by performing an exhaustive search over all possible splits for each feature, incorporating a regularization parameter. | XGBoost can predict traffic accident durations with high accuracy due to its ability to exhaustively search for optimal split values and control model complexity with regularization. | High prediction accuracy<br>Parallelizable for faster training<br>Regularization prevents overfitting | Longer training time (common for boosting methods)<br>May require tuning of hyperparameters |

TABLE 2.5: Advanced Methods used in traffic accident duration prediction

| **Metric** | **(R) Regression or (C) classification** | **Formula** | **Description** | **Application to Traffic Accident Duration Prediction** |
|---|---|---|---|---|
| MAPE | R | $\frac{1}{n}\sum_{t=1}^{n}\left\|\frac{A_t-F_t}{A_t}\right\|$ | Regression metric. Measures the average percentage error of regression predictions | Evaluates the average percentage difference between the predicted and actual accident durations. Lower values indicating better performance. Previous studies indicate MAPE values around 20-65% [121]. $A_t$ should not be zero value. |
| RMSE | R | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t-F_t)^2}$ | Regression metric. Quantifies the average difference between regression predictions and actual values, sensitive to large errors | Assesses the average difference between the predicted and actual accident durations. Lower values indicating better performance |
| Accuracy | C | $\frac{tp+tn}{tp+tn+fp+fn}$ | Classification metric. Indicates the proportion of correct classification predictions | Measures the overall proportion of correct predictions for all accident duration categories. Higher values indicating better performance |
| Precision | C | $\frac{tp}{tp+fp}$ | Classification metric. Reflects the accuracy of positive predictions made by the model | Indicates the fraction of accurately predicted accident durations out of all predictions for a specific category. Higher values indicating better performance |
| Recall | C | $\frac{tp}{tp+fn}$ | Classification metric. Measures the proportion of true positives out of all actual positive instances | Represents the model's ability to identify all the relevant cases for a specific accident duration category. Higher values indicating better performance |
| F1-score | C | $2*\frac{precision*recall}{precision+recall}$ | Classification metric. Provides a balanced measure of classification performance considering precision and recall | Gives a balanced performance assessment for the model, considering both precision and recall for accident duration categories. Higher values indicating better performance |
| AUC | C | Area under the Receiver operating characteristic (ROC) curve | Classification metric. Measures the overall classification performance across all classification thresholds | Assesses the model's ability to distinguish between different accident duration categories. Higher values indicating better performance |

TABLE 2.6: Summary of metrics for evaluating traffic accident duration prediction models.

| **Metric** | **Studies** |
|---|---|
| MAE | [226][256][71][242][205][215][235][187][119][82][183][130][164][8][81][273] |
| RMSE | [226][256][71][242][205][215][235][126][82][128][130][164][8][81] [132][76][181][139][273][68][13][75] |
| MAPE | [226][256][69][127][242][274][103][86] [205][215][235][126][193] [187][135] [119][159][128][130][164][129][216] [76][181][139][273][67][68][13][75][44][272][115] |
| MSE | [8][76] |
| AUC | [266][167][269] |
| Recall | [198][159][75] |
| Precision | [198][159] |
| F1 | [198][159][75] |

TABLE 2.7: Metrics used across reviewed papers.

| Criteria | Under-sampling | Over-sampling | Combined Approach |
|---|---|---|---|
| Large dataset | Suitable | - | May be used |
| Small dataset | - | Suitable | - |
| Extreme imbalance | - | May be used | Suitable |
| Limited computational resources | Suitable | - | May be used |

TABLE 2.8: Considerations for using over/under sampling methods

| Method | Studies |
|---|---|
| Random Forest | [206][167][164] |
| XGBoost | [159][216][184][76][75] |
| Support-vector machines (SVM) | [256][266][206][98][215][167][183][164][129][81][139][93][71][205][164] [216][248][67][68] |
| Linear Regression | [65][187][135][181][137][115] |
| Naive Bayes | [146][180][266][206][215][167][183][129][216][111][273][68][136][236][93][115] |
| Decision Tree | [198][180][235][81] |
| Gradient-boosted Decision Trees | [159][76][75] |
| K-Nearest Neighbours | [146][136] |

TABLE 2.9: Most popular Machine Learning methods used across reviewed papers

| Method | Studies |
|---|---|
| Principal component Analysis (PCA) | [29][233] |
| Linear discriminant analysis (LDA) | [127][205][187] |
| SHapley Additive exPlanations (SHAP) | [103][184][75] |
| Feature Selection | [230][206][119] |
| Clustering | [71][103][67][136] |
| Ensemble | [167][81] |
| Bi-level frameworks | [115][75] |

TABLE 2.10: Most popular Machine Learning pipeline elements used across reviewed papers.

| Network type | Studies |
|---|---|
| Artificial Neural Network (ANN) | [226][256][242][206][167][187] [119][82][164][129][81][31][269][76][139][67] |
| Multilayer Perceptron (MLP) | [71][205][164][269][67] |
| Recurrent Neural Network (RNN) | [269][236] |
| Long Short-Term Memory (LSTM) | [205][269][76][67][236] |
| Convolutional Neural Network (CNN) | [103] |

TABLE 2.11: Most popular Deep Learning methods used across reviewed papers

| Aspect | Machine Learning (ML) | Deep Learning (DL) | Note | Used Methods |
|---|---|---|---|---|
| Model Complexity | Relatively simple algorithms | Complex architectures with multiple layers and nonlinear functions | DL models may capture intricate patterns and relationships between accident factors, while ML models may be sufficient for simpler patterns. | ML: Linear Regression, Random Forest, kNN, GBDT, XGBoost, LightGBM, CatBoost<br>DL: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Graph Neural Networks (GNN) |
| Feature Engineering | Requires manual selection, transformation, and combination of features | Capable of automatic feature learning | DL models can reduce the need for manual feature engineering and domain knowledge, potentially improving the modeling efficiency in traffic accident prediction. | ML: manual feature engineering and preprocessing procedures<br>DL: Capable of automatic feature learning (e.g. CNN network) |
| Data Requirements | Suitable for smaller datasets | Requires larger amounts of data for optimal performance | ML models may be more suitable for situations where data collection is challenging, while DL models may excel with larger datasets related to traffic accidents. | ML: Used for Small to moderate datasets<br>DL: Used for Large datasets |
| Computational Requirements | Lower, can be trained on standard CPUs | Higher, often requires specialized hardware - Graphics Processing Units (GPU) | The choice between ML and DL models for traffic accident prediction may depend on available computational resources. | ML: Standard CPUs (some methods can be implemented on GPU)<br>DL: Specifically GPUs due to exceeding computational demands |
| Interpretability | Generally more interpretable (easier to understand model predictions) | Often considered "black boxes" due to complexity | ML models may provide better insights into the relationships between accident factors, while DL models may sacrifice interpretability for increased performance. | ML: Mostly interpretable<br>DL: Requires aditional methods for interpretability |
| Applicability | Better suited for simpler patterns or smaller datasets | Excels in tasks with complex patterns, spatial-temporal dependencies, or high-dimensional data | In traffic accident modeling, DL methods may be more appropriate for tasks involving spatial-temporal data or complex relationships, while ML methods may suffice for simpler tasks. | ML: Simpler patterns or smaller datasets<br>DL: Complex patterns, spatial-temporal data, high-dimensional data |

TABLE 2.12: Comparison of Machine Learning and Deep Learning Methods for Traffic Accident Modeling

## Chapter 3

# Incident duration prediction using a bi-level machine learning framework with outlier removal and intra-extra joint optimisation

## 3.1 Introduction

### 3.1.1 Context

Traffic congestion is a significant concern for many cities around the world. Congestion arises due to various factors, including increased population, workforce concentration in central areas, or the lack of efficient public transport modes. Two forms of congestion are typically predominant: a) recurrent traffic congestion during peak hours when traffic demand exceeds the road capacity, and b) non-recurrent traffic congestion caused by unplanned events such as car accidents, breakdowns, weather, public manifestations etc. Previous studies have shown that almost 60% of traffic congestion is due to non-recurrent incidents with a stochastic behaviour in space and time [202]. In Australia, the number of road deaths per year has been reduced by 70% since the 1970s. However, the annual economic cost of road crashes was estimated at $27 billion per annum in 2017 [73]. Traffic Incident Management Systems (TIMS) collect data on traffic incidents, including information on different incident duration factors. Accurately predicting the total duration shortly after an incident took place could save operational costs and end-user time (through affecting the route planning). Moreover, the clearance time of accidents is highly related to the ongoing traffic congestion and several external factors with different weights of importance. Therefore, it is essential to estimate the incident factor importance to improve the accuracy of predictions. Most prior studies related to this topic concentrated on testing different machine learning models on specific road types like freeways or highways and focused primarily on different phases of the incident duration such as clearance time, recovery time, and the total incident duration [129]. There is currently a lack of an advanced approach that can be applied on all road types, for all accident types and across various countries with different driving behaviour.

### 3.1.2 Challenges and contribution

The accuracy of predicting the incident duration is often determined more by the modelling methodology, the feature construction, and the result interpretation rather than by the model in use. In this work, we address several open questions or challenges concerning the prediction of the traffic incident duration.

**The first** challenge is to develop a universal bi-level framework applicable to different incident data sets reported on various road network layouts. The majority of prior works had studied the prediction of incident duration on specific types of roads (freeways or motorways) [257]-[45]-[90]-[262], where the data accuracy is higher than on arterial roads; as of 2018, very few applied the prediction strategies on normal arterial roads due to the high modelling complexity and a location mismatching; the majority of traffic incident duration analysis studies focus only on one type of road network (freeways, highways, etc.); this is revealed by a recent state-of-the-art paper published in [129] which emphasises the difficulty of solving this problem for arterial roads and the lack of studies in this field. Our study proposes a framework capable of predicting the incident duration regardless of the road network or its complexity.

**Secondly**, the majority of studies in the literature have concentrated on applying state of the art machine learning models mostly for classifying the incident severity [171] or their duration[129]. However, very few have treated the problem of outliers or imbalanced data classes. Our study addresses both of these issues by proposing a varying threshold procedure that can facilitate binary duration classification threshold selection by considering both class balance and model performance. We also test multi-class classification on data sets split into three equally-sized parts according to incident duration: short, medium or long term. Previous research studies were selecting incident duration thresholds by simple reasoning (e.g. choosing mean, median, percentiles, etc) [116]-[274]-[128]-[125]. We, on the contrary, test multiple different thresholds for three different data sets. Furthermore, we propose our own optimisation approach which we denote intra-extra joint optimisation (IEO) together with an outlier removal procedure (ORM) and advanced machine learning modelling.

**Thirdly**, we further solve the incident duration regression problem and also perform different regression scenarios to test the extrapolation performance of ML models on various incident data sets. We utilise thresholds selected during the classification threshold evaluation procedure to analyse the extrapolation performance by training ML models and making predictions on several duration subsets. It allows us to find the best ML model and the best extrapolation approach for the regression problem on each duration subset (e.g. short-term incidents) of each data set. For the regression problem, we also detect the most influential factors that affect the incident duration that traffic centres need to prioritise in order to predict incident duration with higher accuracy. Our end goal is to improve the extrapolation ability of machine learning models on the task of incident duration prediction and find the best modelling approach for short-term and long-term incidents.

**Lastly**, the majority of studies are primarily focusing on choosing a single winning algorithm or approach that works for a specific case study. Unfortunately, we show that the performance of ML models is highly affected by the data set and the chosen methodology: data quality, the available features, and the additional parameter tuning and optimisation techniques applied in this work. We try to develop the universal framework for traffic incident duration prediction applicable to different traffic incident data sets. We choose and adapt the best modelling approaches to each data set and show how this can affect the accuracy and performance of the models. This method allows high flexibility that can be applied for classification and regression predictions on various network types and different data sets.

The most similar research to the current work was published in [116] and relied only on one data set, one method for classification (Bayesian network), one method for regression (K-nearest neighbours), and authors selected static threshold (30 min) to alleviate the class-imbalance problem. This current study provides a significant contribution by advancing on multiple aspects from a large pallet of machine learning models to multiple data sets with unique features, up to outlier removal and joint optimisation.

**Overall, main contributions are the following**:

- to the best of our knowledge, this is the first research study proposing a bi-level prediction framework using a large pallet of several machine learning models applied for both incident duration classification and regression, with the scope of predicting the incident duration on different road types across two different countries (Australia and U.S.A.). Overall, our methodology is agnostic of the location, the network, or the size of the network and can be adapted to any new incident log data set that can be made available.

- we propose a binary versus multi-class classification approach in order to find the best optimal threshold to identify short versus long-term incidents via both quantile analysis and varying threshold data split.
- we propose a novel intra/extra joint optimisation algorithm that integrates baseline ML models with outlier removal and hyper-parameter optimisation techniques across the validation cycle.
- we propose several extrapolation scenarios of analysing the impact of missing logs in the precision of the prediction model and reflect on what type of logs should be best used for tailoring to the prediction problem needs.
- we conduct a feature importance selection sing the SHAP method, which allows graphical interpretation of variables impact on the model output, before we conclude on the most important factors affecting traffic incidents.

Overall, this research lays the foundation stone of bi-level predictive methodologies regarding the traffic incident duration and can provide accurate information for both the end-user route choice modelling as well as for the operational centres which need to optimise their operations under non-recurrent traffic congestion. Moreover, this work contributes to our ongoing objective to build a real-time platform for predicting traffic congestion and to evaluate the incident impact during peak hours (see our previous works published in [160]-[204]-[154]).

The chapter is organised as follows: Section 1 discusses related works, Section 2 presents the data sources available for this study, Section 3 showcases the methodology, Section 4 presents the numerical results for binary and multi-class classification tasks, Section 5 presents the numerical results of the regression part of the framework, Section 6 details on the feature importance evaluation and Section 7 is reserved for conclusions and future perspectives.

### 3.1.3 Related works

**Incident data interpretation:** The definition of traffic incident duration phases is provided in the Highway Capacity Manual [11], and it consists of the following time-intervals: 1) **incident detection time** which is the time interval between the incident occurrence and its reporting, 2) **incident response time** standing for the time interval between the incident reporting and the arrival of the first investigator at the location of the accident, 3) **incident clearance time** representing the time interval between the arrival of the first investigator and the clearance of the incident, 4) **incident recovery time** indicating the time interval between the clearance of the incident and the return of traffic flows to normal conditions.

The **total incident duration** is the time interval between the first incident log, and the returning of traffic flows to normal conditions. In our work, we use the term **incident duration** for the time lapse between the detection of an incident and the clearance of the incident, as officially reported in traffic logs provided by local traffic authorities. Therefore we do not include the incident recovery time as this information is not recorded in the three data sets provided. However, different phases of traffic incident duration (e.g. clearance, recovery time) can be modelled individually upon availability; this type of research is rare because of the complexity of data collection for traffic incidents and small amounts of recorded traffic incidents in real-life datasets [129], [11].

When it comes to the data interpretation in the literature, the incident duration distribution has been modelled as log-normal [212] and more recently as log-logistics distribution [46], [210]. In a recent study, [85], incident clearance time and the total impact duration were modelled using Weibull, log-normal, log-logistic distributions and compared using the Akaike information criterion (AIC) criteria; findings have revealed that log-logistic distribution was outperforming other distributions. As distribution utilisation is highly related to the specificity of each data set, for this study, in which we use three different data sets, we further apply a comparison among several distribution modelling choices by using the AIC criteria.

According to [232], different statistical methods were applied to model traffic incidents: 1) fixed parameter regression 2) random parameter regression 3) quantile regression. In this study, log-transformation of the target variable (incident duration) also applied. Random parameter regression found to give better statistical fit for incident duration models than fixed parameter regression, and therefore provide more accurate predictions of incident durations. It also highlighted that fixed-parameter regression model may give non-accurate incident duration predictions due to over/under- estimation of dependency between variables and incident durations. Also, there were no substantial difference found between fixed parameter regression and quantile regression in the case of 2015 Virginia incident data set. The benefit of quantiled regression is the ability to model the relationship of any quantile (rather than only average incident duration) of the incident duration vector with a set of explanatory variables [109]. Ordinary Least Squares model can provide the predicted mean of the incident duration. On the contrary, quantile regression provides estimates for every quantile, which represented as a conditional distribution of incident durations, without providing single value as the incident duration prediction. Quantile regression coefficients represent the change in the incident duration in a given quantile category in relation to independent variables. Similar to this approach, variable importance can be estimated within each traffic incident duration group.

**Machine Learning for incident duration prediction:** While several statistical modelling techniques have been applied previously, more recently, new approaches in machine learning (ML) modelling have emerged as a more advanced way of predicting the incident duration due to their capacity to easily account for new data sources, as well as for removing the linearity assumptions between features and the predicted class [91]. Examples of such approaches are: artificial neural networks (ANNs) [144], genetic algorithms [119], support vector machines (SVMs) [227], k-Nearest-Neighbours (kNNs) [244] and decision-trees (DTs) [88]. The recently proposed Gradient-Boosted Decision Trees (GBDTs) have been shown to provide superior prediction performance when compared to Random Forests, SVMs and ANNs [150]. However, it is known that GBDT can easily over-fit when the prediction target has a long-tail distribution, as is the case of the traffic incident duration distribution [150]. XGBoost [36] is another decision-tree enhancement method that has gained popularity recently in the machine learning community due to its tree boosting capability, loss function regularisation and adaptive learning rate. It was employed in several international competitions, winning 17 out of the 29 Kaggle competitions singled out on the 2015 Kaggle blog; it was also employed by every team in the top-10 in the 2015 KDDCup [20] for solving various problems such as store sales prediction, web text classification, hazard risk prediction, and product categorisation. XGBoost's popularity is also due to its scalability (it can run on a single machine, as well as on distributed and paralleled clusters), its capacity to handle sparse data and its ability to handle instance weights in approximate tree learning (see the recent paper published by [36] where authors proposed an end-to-end tree boosting system with cache-aware and sparsity learning features). While each of these methods has its advantages and disadvantages, building a fast and reliable prediction framework that could be applied for real-time operations represents a true challenge.

One of the recent research studies [116] presented a two-step approach for traffic incident duration prediction. A cost-sensitive Bayesian network was used to perform binary classification of traffic incidents by choosing a threshold of 30 minutes and then performing regression for each class using kNN. While the approach is functional, one major drawback for the classification problem is to manually choose the class split threshold, as it can lead to severe class imbalance; to overcome this issue, in our study, we perform both a fixed and a varying threshold set-up to find the best class balance for our classification models; even-more, we propose as well a comparison with a multi-class classification approach and debate on the benefits and drawbacks of using classifiers for such problems; we also enhanced more advanced regression models together with outlier removal procedures that would provide a better and more precise prediction of the incident duration precondition in minutes. Overall, the cost sensitivity of incorrect classification can be further extended to the cost-based regression metrics. We propose our enhanced ML models with a proposed intra and extra joint optimisation technique and outlier removal procedure to have even more precise predictions.

In one of the recent research studies on applying machine learning, which was related to the classification of driving state, multiple hyper-optimised ML models were tested, and entire feature space was visualised using t-SNE for entire feature space visualisation [255]. RandomForest provided the highest prediction accuracy, but more advanced tree-based models exist that utilise gradient boosting, which we will be using in our research (e.g. gradient boosted decision trees).

To verify the performance of advanced tree-based methods (as LGBM - Light Gradient Boosted Model), additional conventional ML models can be used [38]. We decided to also include LGBM and compare it to conventional ML models with non-tree based models (KNN and Logistic Regression).

**On the feature selection:** It is generally not enough to use all the possible features for the regression analysis of traffic incident durations. Using a high amount of features combined with a small data set size can lead to over-fitting. Some features can be helpful or useless, more or less critical, while others do not impact the prediction results significantly. By performing a feature importance analysis, we can recommend to traffic management facilities to record the most critical data and omit redundant data related to traffic incidents. For example, one can increase the prediction accuracy by using as additional features the weather conditions, which were found to play a significant role in some research studies (e.g. during the summer and autumn seasons in Washington – USA in 2009, the preparation time of the rescue team was higher on freeways [94]). In some countries with cold weather, the response times can be much higher, while in regions with sunny weather most of the year, the weather impact on the intervention team can be neglected. Overall, the weather impact on the traffic incident duration prediction needs needs customised via a data-driven feature important analysis. Peak hours were the most influencing feature on response team preparation delay, which was found to be linked to response procedures (the goal of the response team was to resolve incidents during peak hours as soon as possible). A research study using Beijing traffic incidents data from 2008 [130] found the importance of "peak hour" value for the response team travel time and clearance time, but not for the intervention team preparation time. Our study conducts a feature importance ranking based on the best performing ML models we have proposed and provides a detailed overview of their impact. Different approaches to feature importance estimation use tree-based models (e.g. Random Forest, Light Gradient Boosted Machines - LGBM, extreme gradient bosoting models - XGBoost). For example, one can use produced decision trees from the tree-ensemble model [38]. A data-driven approach was used to perform information fusion from different sources [2], which involved the use of Gini-index extracted from Random Forests as a method to estimate feature importance. Nevertheless, the single random model can have a noticeable variance in data mapping when there is a weak connection between features and the target variable by making the feature importance value dependent on the random seed for the model. The Shapley Additive explanation (SHAP) [148] provides a more advanced approach for feature importance estimation because it fuses estimation from multiple models trained across many different subsets (which selected both feature-scale and index-scale) of the dataset. These studies motivated the utilisation of the Shap Values for our feature importance ranking across three different data sets, all with different features and incident information.

**On the future application of our research:** In comparison with other work, the research proposed in our paper comes not only with a significant prediction capability for all types of incident data sets with various features, but it can be further extended for solving the route scheduling problem within traffic simulation modelling, which will incorporate the adaptation of agents to occurring traffic incidents. Apart from analysing the effects of traffic control measures [112], it is possible to analyse the effect of additional information such as the predicted incident duration, which can be performed both for scheduling and online rescheduling of dynamic agent re-routing. Furthermore, simulation can be performed with and without such information to estimate the possible benefits of the incident duration prediction modelling within the traffic system. Also, using an online rescheduling procedure requires the simulation to be performed at the level of dynamic agents within a micro-simulation model, which could benefit from new re-routing schemes when traffic disruptions occur along the route.

## 3.2 Case study

In order to test the efficiency of the proposed bi-level framework, we have used three different data sets from two different countries: Australia and U.S.A. The three data sets represent incident logs from an arterial road suburb in Sydney, a motorway in Sydney, Australia, and a road area from San Francisco, U.S.A. The data sets are all recorded by different means and allow us to explore the impact of the prediction framework across various types of road networks. The three data sets are represented in Fig. 3.1 and are detailed as follows.

**Victoria Rd - arterial network, Sydney:** The first data set (dataset AR) contains one-year incident logs from the Victoria arterial road from Sydney, Australia (in 2017) (see Table 6.4 for a summary of features, in which the + symbol under each data set column and for each line indicates whether that variable is present or not in the data set - for example, the TZName variable is present in the Arterial Roads data set but not in the M motorway data set). It contains information on 5,134 traffic

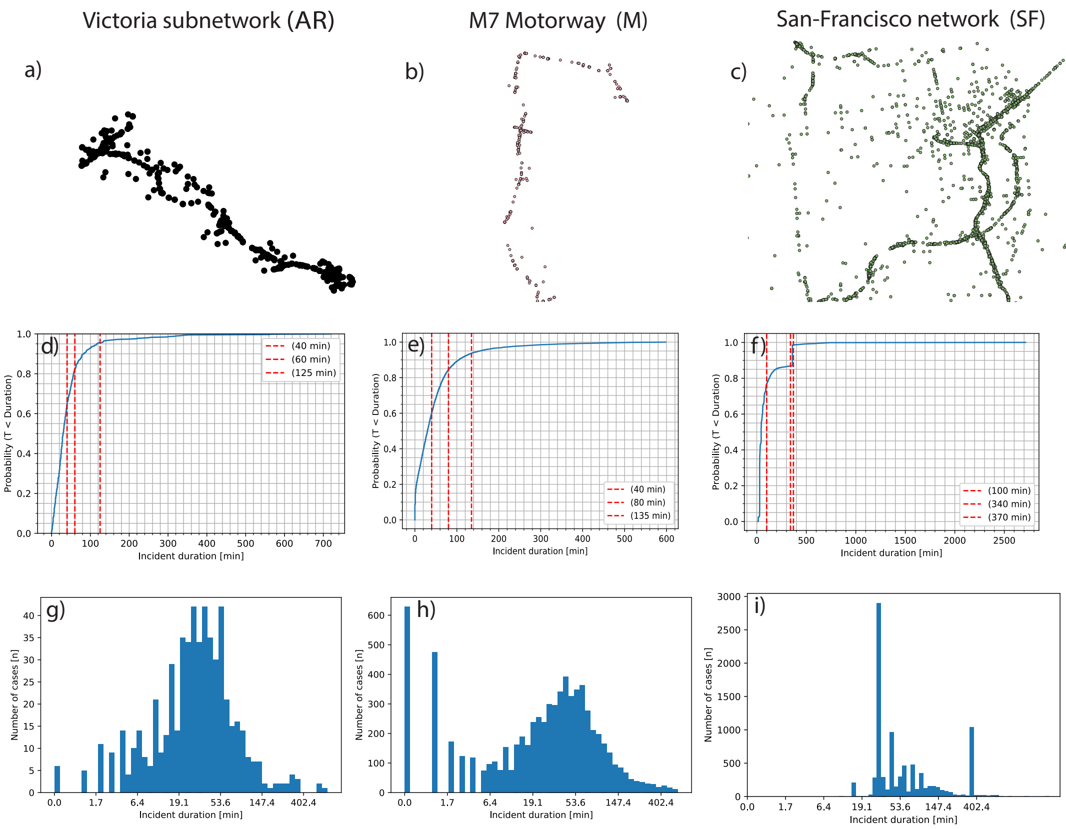



FIGURE 3.1: Data profiling for all data sets in our study: Victoria Rd (A) - a) network mapping, d) ecdf - empirical cumulative distribution function g) distribution plot; M7 motorway (M) - b) network mapping, e) ecdf h) distribution plot; San Francisco (SF) - c) network mapping, f) ecdf i) distribution plot.

| Variable | AR | M | Values | Description |
|---|---|---|---|---|
| Location | + | + | $\mathbb{N}, \mathbb{N}$ | $X, Y$ in GDA Lambert coordinates |
| Hour of day | + | + | $\{0, 1, \ldots, 23\}$ | - |
| Peak Hour | + | + | $\{1, 0\}$ | Value is 1 if hour belongs to $\{7 \ldots 9\}$ or $\{16 \ldots 18\}$ hour interval |
| Day of the week | + | + | $\{1 \ldots 5\}$ | Weekday numbers from Monday to Friday |
| Weekend | + | + | $\{0, 1\}$ | Value is 1 for Saturday and 0 for Sunday |
| Month of the Year | + | + | $\{1, 2, \ldots, 12\}$ | - |
| Incident Subtype | + | + | $\{Bus, car, bicycle, animals, etc.\}$ | Field indicating cause of incident |
| Affected lanes | + | + | $\{1, 2, 3, 4, Alllanes, breakdown, nodata\}$ | Number of lanes affected by the accident |
| Direction | + | + | E, W, N, S, E-W, N-S, One/Both | Affected traffic direction |
| Incident Source | + | + | $\{ICEMS/ISENTRY, OPERATOR, etc\}$ | Source of the incident report |
| Unplanned | + | | $\{0, 1\}$ | Value is 1 if incident is planned, 0 otherwise |
| Average Temperature | + | + | $\{11.13C - 32.4C\}$ | Average temperature for the time of the incident |
| Rainfall | + | + | $\{0 - 85mm\}$ | Rainfall for the time of the incident |
| Public holidays | + | + | $\{0, 1\}$ | Value is 1 if days is a public holiday |
| Sector ID | + | + | R+ | Defined by TMC |
| TZName | + | | R+ | Traffic zone name as Defined by the Bureau of Transport Statistics |
| Section ID | + | | R+ | Road section on which the incident occurred |
| Section Speed | + | | $R + [Km/h]$ | Section speed limit |
| Section Lanes | + | | $\{1, 2, 3, 4, 5, 6\}$ | Number of section lanes |
| Section class | + | | $R+$ | As defined by TMC |
| Street ID | + | | $R+$ | As defined by TMC |
| Intersection ID | + | | $R+$ | As defined by TMC |
| Distance from CBD | + | | $R+$ | distance between the traffic incident and the city CBD |
| Section Capacity | + | | $\{0 \ldots 3100\ vehicles/hour\}$ | Maximum flow capacity of the section |

TABLE 3.1: Traffic incident features for Sydney Arterial roads (AR) and M7 motorway (M).

incidents with different incident types (e.g. hazards, breakdowns, accidents) and subtypes (e.g. work zone, accident with truck). Our current study focuses on 574 "Accidents" since these induce the longest clearance time in the current subnetwork according to the traffic management centre (TMC). Traffic 'Accidents' have a mean duration of 44.59 minutes and a maximum of 719 minutes. Weather data represented as average daily temperature (in Celsius) and precipitation rate (in millimetres) are obtained from the Observatory Hill station in Northern Sydney, which is the closest station to the analysis area. Public holiday data represented as boolean values for public and regional holidays in 2017 in New South Wales, Australia. The area geometry features contain the sector ID as defined by TMC, the code of the official area where the accident occurred (as defined by the Bureau of Transport and Statistics), and supplementary information such as section capacity, section speed limit, and the number of lanes. These features are available for all road sections in the Victoria sub-network, and they were extracted from the official traffic simulation model of the Victoria network, developed in Aimsun and previously used by the authors for conducting an incident impact analysis and traffic prediction [243].

**M7 motorway, Sydney:** The second data set is a motorway data set (data set M), consisting of 7,194 traffic accidents along the M7 motorway in Sydney, Australia, during the same year 2017. The mean duration of motorway accidents is 47.2 minutes, with a maximum duration of 598 minutes (9.96 hours). This data set also includes weather data (average daily temperature and precipitation). This set of features is similar to the arterial roads data set AR without the geometric features of the lanes (section lanes, section class), intersection ID, distance from the central business district (CBD); this is due to the complexity of mapping of a traffic incident to a correct location along the motorway. We make the observation that for both Data set AR and M, the traffic flow information of the affected road sections was omitted for this study since we found previously no significant improvement to the prediction accuracy [160].

**San-Francisco road network:** The last data set is from San-Francisco, U.S.A. (data set SF) and includes information on accidents from all types of roads in the city. It is part of a more considerable initiative entitled "A Countrywide Traffic Accident Dataset", recently released in 2021, which contains 1.5 million accident reports collected for almost 4.5 years since March 2016 [166]. The SF data set contains 49 features describing the accidents as detailed in [166] (due to a large table of feature, we refer the reader to the cited paper and not duplicate this feature information). This study focuses on the "accident" type duration prediction as being the most severe one. We extract and use 8,754 accident records related to the San-Francisco area. As observed from Fig. 3.1 c), a significant part of the accidents occurred along the “US-101” highway and “John F. Foran” Freeway. Accidents have a mean duration of 100 minutes and a max duration of 2,715 minutes.

**Data sets profiling:** Each data set undergoes a profiling procedure by investigating the empirical cumulative distribution functions (ECDF) - as plotted in Fig. 3.1 d), e), f), and their equivalent log-space distribution plots (as represented in Fig. 3.1 g), h), i). The ECDF function presents thresholds of data behaviour (marked in red) across each data set which reveal indicative thresholds of a different behaviour around specific incident duration (see for example Fig. 3.1d) versus Fig. 3.1f) where the first inflection point is around 40min for data set AR versus 100min for data set SF. Findings reveal significant anomalies representative of each data set. For example, data set AR contains a reduced amount of traffic accidents with small incident duration (zero or less than 4 min), data set M contains an increased number of accidents with zero or one-minute duration, while the data set SF despite not presenting any short term incident duration below 17 minutes, it contains a large number of incidents of 29 and 360 minutes which raises the question of either these are outliers in the data set or simply reveal a road network behaviour in terms of incident management in the area; this also might indicate that it will present unique behaviour under the prediction framework and that different processing techniques needs to be applied for this data set. We also observe that the incident duration is long-tail distributed, which is likely to pose difficulties for prediction algorithms due to the presence of extreme values (either small or large).

## 3.3 Methodology

Clearing accidents in a short time represents a high priority task for traffic management centres (TMC) worldwide. For example, in New South Wales, Australia, the target clearance time for traffic incidents is 45 minutes, but this limit might differ in other countries. Therefore, in the rest of this chapter, we will refer to this threshold as “incident clearance threshold ($T_c$)” and any incidents cleared before this threshold (e.g. $< 45$ min) as "short-term"; incidents which lasted more than the clearance threshold (e.g. $>= 45$ min) will be referred to as “long-term” traffic incidents. A unique threshold will be derived for each dataset and will be discussed further. The methodology of this study has its origins in our previous work applied only for arterial roads [160], which we further extend and improve via the joint optimisation and outlier detection enhancements of the prediction framework. The methodology we propose for modelling the incident duration prediction problem is using a bi-level prediction

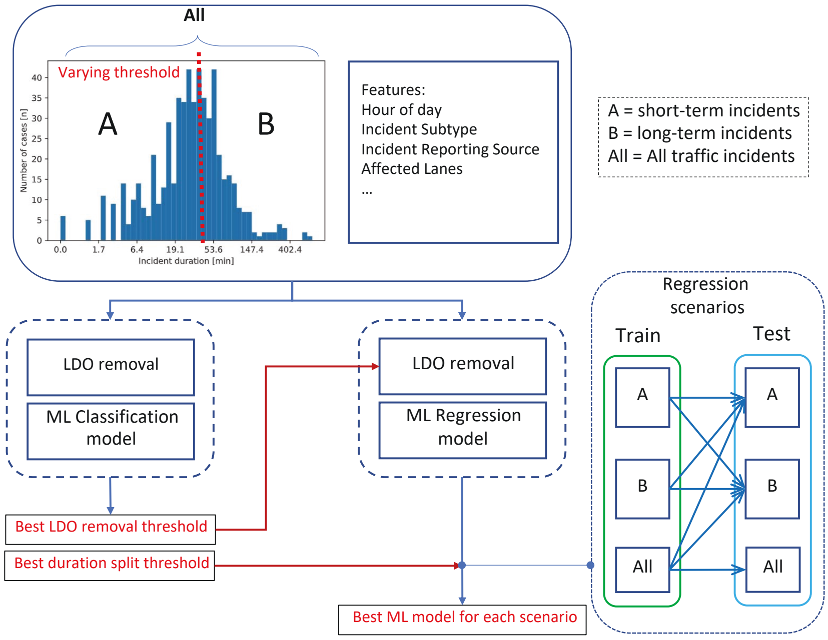



FIGURE 3.2: The proposed bi-level modelling framework for traffic incident duration prediction.

framework combining a classification and regression modelling, as represented in Fig. 3.2. This approach has been constructed by considering the real-time operational goals of TMC and providing short duration prediction into the life-cycle of the incident management.

Based on the initial traffic incident information, the first step is the deployment of a fast classification method which would only predict whether the accident will be either short-term (subset A) or long-term (subset B) - see incoming data set from Fig. 3.2 where the data is split in two parts based on $T_c$). Next, we test various duration thresholds and select the optimal $T_c^o$, which provides a good class balance and classification performance for each dataset. Once the optimal $T_c^o$ has been found, a further regression modelling is applied for predicting a more precise duration of future incidents down to the minute level.

Due to the main challenge of this task, we further propose an outlier removal approach (ORM) detailed in Section 3.3.7 and our innovative Intra/Extra Joint Optimisation modelling coupled with several machine learning models trained via a hyper parameter tuning (we denote this approach as IEO-ML and is further detailed in Section 3.3.9).

The boosted regression framework is finally applied under several regression scenarios (see section Section 3.3.6), which are constructed to evaluate the framework capability to predict under all possible situations. For example, when we only have a subset A available (short-term incidents) but the TMC would like to predict long term incident (subset B) we denote this as a Scenario A-to-B (training the models on subset A and making predictions on subset B); all scenarios are constructed based on the assumptions that the framework needs to be robust in order to predict any type of incident durations, under all possible data shortage or lack of information availability. In the following subsection, we further provide the mathematical and theoretical modelling of each of the steps described above.

### 3.3.1 Classification and regression definitions

Using all available data sets and the incident information, we first denote the matrix of traffic incident features as:

$$X = [x_{ij}]_{i=1..N_i}^{j=1..N_f} \tag{3.1}$$

where $N_i$ is the total number of traffic incident records used in our modelling and $N_f$ is the total number of features characterising the incident (severity, number of lanes, type, neighbourhood, etc.) according to each specific data set (see examples provided in Table 6.4). For the incident duration classification problem, we denote the incident duration classification vector as:

$$\begin{cases} Y_c = [y_i^c]_{i\in 1..N} & y_i^c \in \{0,1\} \\ Y_{mc} = [y_i^{mc}]_{i\in 1..N} & y_i^{mc} \in \{0,1,2\} \end{cases} \tag{3.2}$$

where N is the duration of the traffic incident (in minutes), $Y_c$ is the vector of binary values taking values in $\{0,1\}$, and $Y_{mc}$ is the vector of integer values for the multi-class classification problem definition, taking values in $\{0,1,2\}$. More specifically, in the first stage we create a binary classification modelling with the purpose of identifying short versus long-term incident duration, split by the incident clearance threshold $T_c$. Thus our task is to predict $y_i^c$, where $Y_c$ takes one of the binary values:

$$\begin{cases} y_i^c = 0 & \text{if } y_i \leq T_c, \quad \text{short-term incidents} \\ y_i^c = 1 & \text{if } y_i > T_c, \quad \text{long-term incidents} \end{cases} \tag{3.3}$$

where the threshold is varied every 5min between $T_c \in \{20, 25, ..., 70\}$. Subsequently, the multi-class method identifies the best two thresholds to separate between short, mid and long-term incident duration. The main purpose of this approach is to test the limits of the class balance which would maintain good model performance, and is expressed as follows:

$$\begin{cases} y_i^{mc} = 0 & \text{if } y_i \leq T_c^1, \quad \text{short-term incidents} \\ y_i^{mc} = 1 & \text{if } y_i \in [T_c^1, T_c^2], \quad \text{mid-term incidents} \\ y_i^{mc} = 2 & \text{if } y_i \geq T_c^2, \quad \text{long-term incidents} \end{cases} \tag{3.4}$$

where $T_c^1$ and $T_c^2$ take several values as further detailed in Section 3.4.3. The binary classification approach implemented with a computation time constraint for operational purposes (more details on computation time comparison can be found in **??**).

The regression problem is further structured with a more fine-grained incident duration prediction in mind. The main objective motivating the regression modelling consists in more precise information regarding the duration of incidents which can fall into a wide class which contains mostly incident logs with a reported duration between and 0 minutes and 30 minutes (for these cases, the traffic centres require more detailed precision to the minute level as a 5-min accident has different handling procedures than more severe accidents of 30min for example). The incident duration regression vector ($Y_r$) is represented as:

$$Y_r = [y_i^r]_{i\in 1..N}, \; y_i^r \in \mathbb{N} \tag{3.5}$$

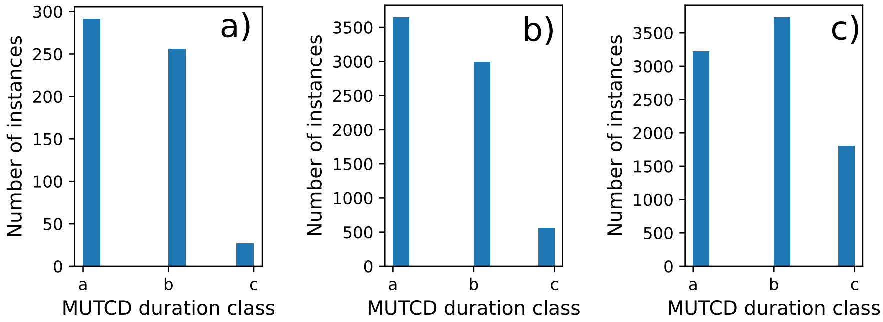



FIGURE 3.3: Distribution of incident durations according to MUTCD duration classes: a) Arterial roads, Sydney, Australia b) M7 Motorway, Sydney, Australia c) San-Francisco, USA

and the regression task is to predict the traffic incident duration $y_i^r$ based on the traffic incident features $x_{i,j}$. The regression models go via an extensive cross-validation procedure with hyper-parameter tuning, with the test of outlier removal using a joint optimisation approach as further detailed in the Section 3.3.4-Section 3.3.7-Section 3.3.9.

### 3.3.2 Applicability of knowledge-based incident duration classification guidelines

According to the The Manual on Uniform Traffic Control Devices (MUTCD) official guidelines [221] Section 6I, traffic incidents divided into three classes: a) Major - with expected duration more than 2 hours b) Intermediate — expected duration of 30 minutes to 2 hours c) Minor — expected duration under 30 minutes.

First, the MUTCD classification seems to be general knowledge-based system and does not consider specifics of each data set / country regulations / specifics of applied incident response guidelines. We approach the classification task from data analysis point of view in relation to application of ML models and try to infer these thresholds from the actual data sets. Shifting to MUTCD classification approach will also make incident duration classes imbalanced (see Figure 3.3). Second, this classification may not be applicable due to road networks heterogeneity [131] and consequent differences in incident duration distribution. As can be seen from Figure 3.1, all three data sets have different distribution of incident durations and therefore such classification may be biased in each case. Overall, in our study, we aim to provide insights from data analysis point of view.

### 3.3.3 Selection of baseline machine learning models

We have tested and deployed several ML models for both the classification and regression problems for this current work, which have served as baseline models to compare our proposed optimisation approach. These are listed as follows: a) gradient boosting decision trees - GBDT [60] which rely on training a sequence of models, where each model is added consequently to reduce the residuals of prior models; b) extreme gradient decision trees - XGBoost [37] which finds the split values by enumerating over all the possible splits on all the features (exhaustive search) and contains a regularisation parameter in the objective function; c) random forests - RF [26] which applies a bootstrap aggregation (bagging, which consists of training models on randomly selected subsets of data) and uses the average (or majority of votes) of multiple decision trees in order to reduce the sensitivity of a single tree model to noise in the data; d) k-nearest neighbours - kNN [59] which uses for the prediction on data points the majority of votes or the average from k closest neighbouring data points from the training set (based on a distance metric); e) linear Regressions - LR - a standard predictor using linear equations to model the relation between the features and the regression variable; f) light gradient boosted machines - LGBM [107] which applies gradient boosting to tree-based models; it also uses a Gradient-based One-Side Sampling (GOSS) and excludes data points with small residuals for finding split value. The models have been used for both classification and regression problems (except logistic regression applied to classification only and linear regression to regression problem only). They are the main base on which we further enhance and develop our outlier and joint optimisation prediction algorithm used in the current bi-level incident duration prediction framework.

### 3.3.4 Hyper-parameter tuning through randomised search

Most machine learning algorithms have a set of hyper-parameters related to the internal design of the algorithm that cannot be fitted from the training data. Both GBDT and XGBoost present dozens of hyper-parameters, out of which the most important ones are max_ depth, learning_rate, min_ child_weight, gamma, subsample, colsample_ bytree and scale_ pos_ weight [24]. The hyper-parameters are usually tuned through randomised search and cross-validation. The most extensive search technique is the grid-search, in which several equally spaced points are chosen in the most credible interval for each parameter, and for each point combination, a model is fitted and tested through cross-validation. The grid-search parameter tuning is straightforward; however, the grid-search scales poorly as the number of hyper-parameters increases. In this work, we employ a Randomised-Search [22] which selects a (small) number of hyper-parameter configurations randomly to use through cross-validation.

To determine the optimal number of iterations for models and data sets, we perform iterative testing. The number of random-search iterations is from 25 to 250 with step 25. For example, on Fig. 3.4, (Arterial roads, Sydney), we see that XGBoost is the best performing model starting from 120 iterations, and it is already close to optimum starting from 175 iterations. The second-best performing model is LGBM, but increasing the number of iterations does not seem to have a significant effect on the model performance which seems to be quite stable without many fluctuations across all evaluation metrics. Other methods perform significantly worse (more than 110% MAPE). For San-Francisco, we see the superior performance of LGBM. The second best is XGBoost. Since there are no metric improvements across iterations for most models, the number of iterations is essential only for XGBoost. According to the results, we decide to search for hyper-parameters for 250 random parameter combinations for each model. We evaluate each combination using a 5-fold cross-validation and then providing results using a 10-fold cross-validation using best combination.

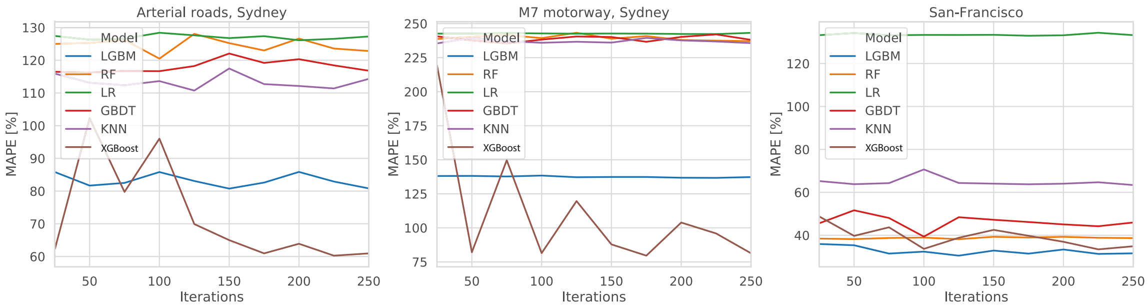



FIGURE 3.4: Performance testing of ML models across three different data sets

### 3.3.5 Model Performance Evaluation

The performance of classification model is evaluated using the Precision, Recall, Accuracy and F1-score and defined as:

$$Precision = \frac{tp}{tp+fp}, \tag{3.6}$$

$$Recall = \frac{tp}{tp+fn}, \tag{3.7}$$

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}, \tag{3.8}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}. \tag{3.9}$$

where $tn$ represents true negatives, $fn$ - false negatives, $tp$ - true positives, $fp$ - false positives. For example, We refer to true positives the incidents which have been predicted to be in a specific class (say short-term) and indeed they were short-term upon validation, false positives the incidents which were predicted to be short term but were not, etc.

We use F1-score as a target metric for classification experiments as F1 represents the balance between Precision and Recall, and is in general a better performance metric to use when we are facing an uneven class distribution rather than interpreting the Accuracy results which take into consideration the total number of both false positive, false negative together with the true positives and true negatives; therefore for uneven class balances (especially the ones with fewer incident records), one should rely less on Precision and Accuracy metrics. To evaluate the regression models we use the mean absolute percentage error defined as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{3.10}$$

where $A_t$ are the actual values and $F_t$ - the predicted values, $n$ - number of samples. Other metrics have been calculated but we will keep them concise due to large amount of experiments to show.

### 3.3.6 Regression scenarios definition

The main objective of the bi-level framework is that the regression accuracy can benefit from different setups for different data subsets. For an even better accuracy compared to the classification problems, we are further developing more complex regression models that can provide incident duration prediction at minute-level accuracy. This is the second step of the bi-level prediction framework to be applied when more precision is needed at the minute level regarding the incident duration length. When training such regression models, a crucial step is the size of the data set and the distribution of the target variable (incident duration). Due to the long tail distribution of incident duration and the class imbalance problem previously identified, we need to design and construct various regression models capable of learning from various types of data sets to make accurate predictions. However, with limited information (small data set size), the prediction results can be skewed. This is the primary motivation that led to the construction of several scenarios of model training, validation and prediction that can be applied under both complete or incomplete data sets from traffic centres. By using the classification thresholds identified previously, we split the traffic incident data set into two subsets: subset A (with duration below threshold $T_c$) and subset B (with duration above threshold $T_c$) as previously defined at the beginning of Section 3.3. We further contract several scenarios of subset combinations for training-validation-testing detailed with the aim of extrapolating the model performance:

***Scenario All-All:*** we use the entire data set and apply several regression models using a 10-fold cross-validation approach and different hyper-parameter search methods. This approach will show us the general performance across various methods.

***Scenario A-to-B:*** we use subset A (short-term incidents) for training the regression models and evaluate the prediction on subset B (long-term incidents). In this scenario, we will analyse methods to extrapolate to higher values of the target variable.

***Scenario A-to-A:*** we use subset A for training the regression models and predict on subset A. In this scenario, we will analyse the prediction ability of methods with long-term incidents excluded (which includes values from the tail of the incident duration distribution).

***Scenario B-to-A:*** we use subset B for training the regression models and predict on subset A. In this scenario, we will analyse methods to extrapolate to lower values of the target variable.

***Scenario B-to-B:*** we use subset B for training the regression models and predict on subset B. In this scenario, we will analyse the prediction ability on long-term incidents.

***Scenario All-to-A:*** we use all the data for training the regression models and predict on each fold within subset A. In this scenario, we will analyse the effect of having access to all types of incident logs in the training phases, both long-term and short-term incidents and how their presence might affect or not, the prediction of short-term incidents duration only. This is to evaluate if using all types of records, including rare events, will help or not to predict better short incidents.

***Scenario All-to-B:*** we use all the data for training the regression models and predict on each fold within subset B. In this scenario, we will analyse the effect of having access to all types of incident logs in the training phases, both long-term and short-term incidents and how their presence might affect or not, the prediction of long-term incidents duration only. This is to evaluate if using all types of records, including short-term events, will help or not to predict better long incidents.

Indeed, from an operational perspective the scenario All-to-All is the ideal situation when traffic management centres would have in their data base both long term and short-term incidents. However, from an operational perspective, several records of short incidents for example and not being kept all the time, while long incidents are often being transferred to various other division if they last more than one day, and they become more of a road infrastructure problem rather than an operational problem which requires constant intervention. Therefore, various incident logs can be imbalanced – some containing more short-term incidents, and others more long-term incidents. The main idea is to provide a good deep dive into the effects of data availability on the model training. For example, training any model only on short term incidents as these are the only ones available will most likely not provide good prediction results in case of long-term incidents and vice versa.

### 3.3.7 Outlier removal methods (ORM)

As previously discussed in Section 3.2-Fig. 3.2 during the data profiling, we observed that the traffic incident logs contain outliers appearing as either minor incidents, rare traffic incidents with highly long duration and/or as errors in incident reports. Therefore, to reduce the side-effect of outliers on all models, we deploy two commonly used outlier removal methods. The IsolationForest (IF) [140] is an outlier removal method, which uses forests of randomly split trees. For each tree, the method randomly selects a feature and a random feature value. The data set is divided into two parts in each step until each data point becomes "isolated" (split from the rest of the data). If the data point is an outlier, it will have a small tree depth (e.g. data point gets quickly separated from the rest by selecting values in just a few features). Tree depth is then averaged between all the "isolation" trees and considered an anomaly score (e.g. if the average tree depth for a point is 1.3, the point is easily separable after a small number of splits). LocalOutlierFactor (LOF) [28] is another outlier removal method, which estimates the anomaly score from local deviation of density within k-nearest neighbourhood. LOF relies on the calculation of a local reachability density (LRD), which represents the inverse of the average reachability distance (RD) of neighbouring data points from the selected data point. Reachability distance (RD) represents the distance to the most distant neighbour within a k-sized neighbourhood (k is also hyper-parameter). LOF of data point then represents the relation between LRDs of neighbours and its LRD and can take values: a) above 1 (higher LRD than its neighbours), b) below 1 (lower LRD than neighbours) and c) equal to 1 (data point has the same density as neighbours). According to the LOF score, we can sort data points and select specific per cent of data points, which have the highest LOF to be eliminated. LOF method relies on the fact that outliers belong to the area where the density of data points is low, while regular data points belong to the high-density area. To summarise, the above outlier removal procedures are applied in conjunction with the proposed optimisation framework and regression models and show a significant improvement in prediction accuracy as further detailed in Section 3.5.3.

### 3.3.8 Outliers from ORM point of view

We would like to make the observation that all the incidents have scalar degree of anomaly when applying outlier removal method. herefore, there are no discrete categories of outliers and normal data points from an outlier method point of view. We simply remove a per cent of data points (e.g. 2%) with the highest degree of anomaly. These points are either easily separable using IF method (tree-based) or remain on a low local density area for the LOF method (distance-based).

So does our outlier removal method actually remove long-term incidents failing to distinguish them from outliers? ML methods in our case, find outliers not only by the value of duration but by including all reported variables (e.g. 25 in the case of Arterial roads). Our aim in this work is to remove incident reports which have very rare characteristics overall, which are also known to negatively affect the ML method performance [147].

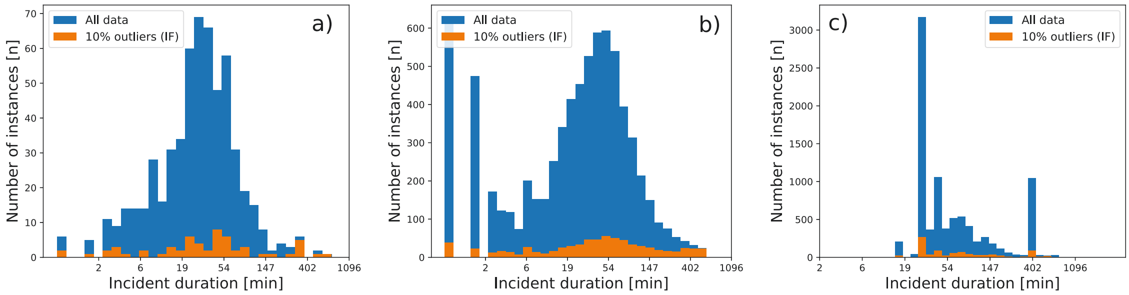



FIGURE 3.5: Data sets with 10% of points with the highest anomaly score removed using IsolationForest: a) Arterial roads, Sydney, Australia b) M7 Motorway, Sydney, Australia c) San-Francisco, USA

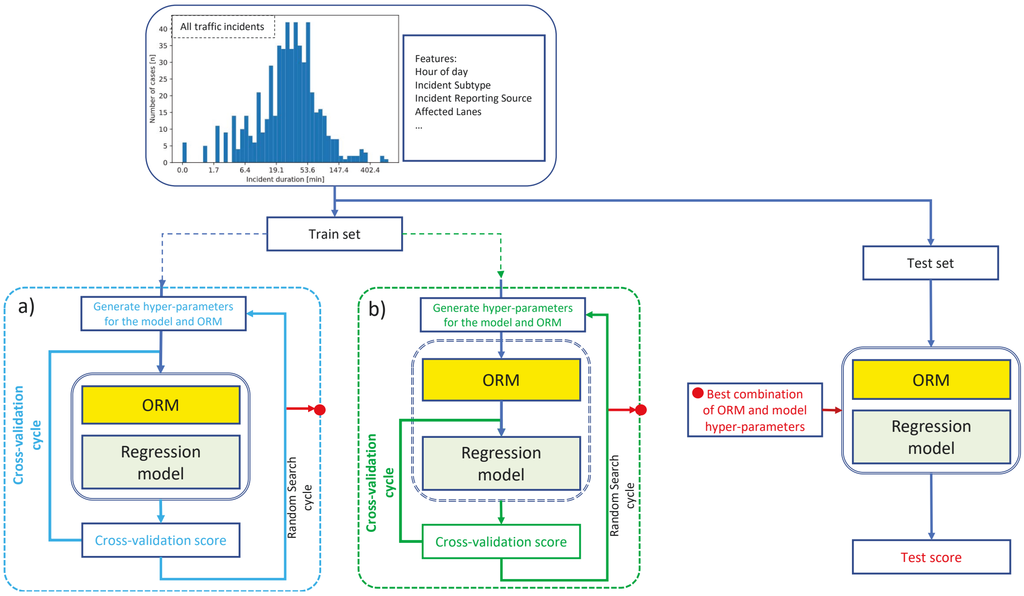



FIGURE 3.6: IEO-ML algorithm with a) Intra joint optimisation schema for the EO-ML algorithm, b) Extra joint optimisation schema for the IO-ML algorithm. Red dot on schema blocks represents output in the form of the best combination of ORM and model hyper-parameters

Fig. 3.5 Showcases Data sets with 10% of points with the highest anomaly score removed using IsolationForest: a) Arterial roads, Sydney, Australia b) M7 Motorway, Sydney, Australia c) San-Francisco, USA. By performing experiments with an outlier removal (isolation forest, 10% of point with the highest anomaly rate removed), we see how many incidents were removed according to each duration interval. An important finding is that outliers do not reside in the area of long-term incidents but rather scattered among the general population of incidents.

### 3.3.9 Intra/Extra Joint Optimisation for ML regression prediction (IEO-ML)

This section presents our novel enhancements of ML regression models by constructing an **intra/extra optimisation technique** to jointly optimise the hyper-parameters of the regression models together with previous outlier optimisation methods. In the rest of the chapter, we denote this approach as **IEO-ML**, where ML is one of the regression models previously described (GBDT, XGBoost, RF, kNN, LR, LGBM). We introduce this approach for multiple reasons: 1) the traffic incident data is prone to errors during the data collection, which is attributed to human factors (e.g. presence of incidents with 0 and 1-minute durations, for example), 2) an outlier removal performance cannot be assessed on the new dataset with no marking for outliers; thus, we can assess outlier removal performance by looking at model performance with outlier removal applied, use joint outlier removal and modelling to assess the outlier removal performance metrics, 3) both the outlier removal method and models have hyper-parameters forming a single hyper-parameters space, 4) we assume that the outlier removal can be performed either inside (Intra - see Fig. 3.6a)) or outside (Extra - see Fig. 3.6b)) of the cross-validation cycle, and we evaluate the effect of such an approach on the model performance, 5) Intra joint optimisation can provide a more effective outlier removal since common hyper-parameters will be found for different data subsets, which allows ORM to be adapted to different possible combinations of incidents in case of the model deployment and prediction on the newly acquired incident log. Overall we want to compare and observe the impact of each technique on the accuracy of regression models and detect the best combination of Intra/Extra joint optimisation and various ML regression models.

Further, we present our proposed IEO-ML algorithm in conjunction with the two outlier removal methods IF and LOF, and several regressions models. Our approach explores the following combinations of ML models in selected working base (decimal or logarithm) with outlier removal and intra/extra joint optimisation; for example, we denote as *iLOF-LT-MLmodel* a "joint optimisation of any available baseline ML model with LOF in a log-transform base within a cross-validation cycle (an intra optimisation)". As an observation, ORM has specific hyper-parameters but one parameter in common - the percentage of removed samples, which we assume to be outliers (ORperc). Thus, to solve the ORM problem, we assume that the amount of outliers in each data set (ORperc) can take values up to 5%. EJO is performed only once and before the cross-validation cycle, but IJO is performed within each fold in a number of times which is equal to the number of folds. Thus, ORperc has values in $\{0, 1 \ldots 5\%\}$ for EJO, in $\{0, 1/5, \ldots, 5/5\}$ for IJO to ensure a comparable amount of removed samples from both approaches. Results for all combinations of the proposed approach inside the incident duration prediction framework are further provided in Section 3.5.3 for eLOF-ML models, iLOF-ML, iIF-ML, eIF-ML (e.g. eIF-ML is a "joint ML optimisation using IF optimised outside (e) of the cross-validation cycle").

**Data:** Traffic incident reports (feature vector $X$, duration vector $Y_r$)
**Input:** HPSm (Hyper-Parameter Space for Model),
ORM: Outlier Removal Method,
HPSor: Hyper Parameter Space for ORM,
Model: ML regression model $\in \{GBDT, XGBoost, RF, kNN, LR, LGBM\}$,
Iters: Number Of Iterations (number of random search steps for hyper-parameter optimisation),
Folds: number of folds for cross-validation,
sample: function for random sampling from the hyper-parameter space,
FoldIndexes: function to get sample indexes for training folds and test fold,
extra: boolean variable stating the use of extra joint optimisation,
intra: boolean variable stating the use of intra joint optimisation,
split: function to split data set into two parts - train/test and validation parts
**Output:** Predicted duration vector $Y_r$

```
x_tr, y_tr, x_te, y_te = split(x, y);
P = [] ;                                   // temporary cross-validation prediction vector
results = []
for it ← 1..Iters do
    HYPm ← sample(HPSm)
    HYPor ← sample(HYPor)
    idx_train = [] ;                       // indexes of train samples
    idx_valid = [] ;                       // indexes of validation samples
    res = 0 ;                              // scoring results
    if extra then
        x = ORM(x, HYPor) ;                // if EO then filter the outliers from the feature vector
    for k ← 1..Folds do
        idx_train, idx_valid = FoldIndexes(x,k);
        x_train ← x_tr[idx_0^train], ..., x_tr[idx_N^train] ;   // array of feature vector samples for training
        y_train ← y_tr[idx_0^train], ..., y_tr[idx_N^train] ;   // array of duration vector samples for training
        x_valid ← x_tr[idx_0^valid], ..., x_tr[idx_N^valid]
        y_valid ← y_tr[idx_0^valid], ..., y_tr[idx_N^valid]
        if intra then
            x_filtered^train = ORM(x_train, HYPor) ;   // if IO then filter outliers
        initialize_model(Model, HYPm) ;                // random hyper-parameter initialisation
        m ← fit_model(Model, x_filtered^train, y_filtered^train) ;   // fitting the model to the filtered train set
        y_pred ← predict(m, x_valid) ;                 // performing predictions
        P = [P; y_pred]
    end
    res ← Metric(y_tr, P) ;                // scoring the accuracy of predictions using performance metric
    r = [] ;                               // Initializing hash-array
    r['metric'] = res ;                    // populating hash-array with resulting metric
    r['HYPm'] = HYPm
    r['HYPor'] = HYPor
    results = [results; r] ;               // collecting results for sampled hyper-parameters into array
end
best = sort(results, by =' metric')[0] ;   // selecting the best combination of hyper-parameters
initialize_model(Model, best['HYPm'])
x_filtered^tr, y_filtered^tr = ORM(x_tr, best['HYPor']) ;   // applying ORM to the training set
m ← fit_model(Model, x_filtered^tr, y_filtered^tr)
Y_r ← predict(m, x_te) ;                   // performing predictions
```

**Algorithm 1:** Intra and extra joint optimisation algorithm with outlier removal and ML regression modelling.

The algorithm represents the modified cross-validation cycle within the randomised hyper-parameter tuning procedure. We use multiple iterations (in fact, attempts) to find optimal parameters both for the selected model (HYPm) and the outlier removal method (HYPor). On every iteration, we sample hyper-parameter sets from hyper-parameter spaces. Then, if extra joint optimisation selected, an outlier removal procedure performed using all the data before the fold-rotation cycle. Then we perform an n-fold cross-validation procedure, where we split data set into training and testing parts (by preserving ratio between them at F-1:1, where F is the number of folds) according to sequentially generated indexes (e.g. in case of 500 data points, fold 0 will represent indexes from 0 to 100 for the testing set, rest of the folds - indexes from 100 to 500 for the training set, fold 1 - 100-200 for the testing set, rest - 0-100 and 200-500 for the training set, etc). Then, if intra joint optimisation is selected within the cross-validation cycle, we perform outlier removal with sampled hyper-parameters using only the train subset within each train-test split. Hyper-parameters for ORM include the percentage of samples to be removed. After removing outliers, we train a model using a train set and make predictions on the test set.

All arrays with actual and predicted samples collected to be used after the fold-rotation cycle for the model accuracy estimation using specified metric. Since we are selecting test folds in order and making predictions on them, the predicted duration vector will be composed of prediction results composed of these folds. So, first, we collect the resulting metric together with hyper-parameters, actual and predicted labels. To collect data we use hash-array, which is represented as an array, where each element can be addressed by name and not by index as for conventional array. Then we perform the sorting procedure, which will order solutions according to the resulting metric, where we select the best combination of hyper-parameters. Furthermore, finally, we obtain the predicted duration vector by filtering data using the ORM method, training model on the train/test part and making predictions on the validation part.

## 3.4 Incident classification results

This section details the results of the first layer of the bi-level prediction framework related to the classification prediction findings, either via a standard binary classification with varying threshold analysis or via a multi-class classification enhanced by outlier removal procedures.

### 3.4.1 Binary incident classification results using varying split thresholds

The first classification problem that we address is to predict whether an incident duration will be lower or greater than a selected threshold (we classify short-term versus long-term traffic incidents), which can then be used to supply the initial assessment needs of the traffic management centre (TMC) under fast decision times. For example, an operational clearance threshold for the Sydney TMC has been currently established at 45min based on previous operational field experience; however, choosing a fixed threshold for classification can have a significant impact on the results of any prediction algorithm and is highly dependent on the incident duration distribution chart (as represented in Fig. Fig. 3.1-g, h, i). Fig. 3.2 showcases the data split for the binary classification problem where the threshold $T_c$ (dashed red line) is varying according to the two set-ups mentioned above: every 5 minutes ($T_c \in \{20, 25, \dots, 70\}$). We name as Subset A all incident duration records which are lower or equal to $T_c$, (if $y_i \leq T_c$), and as Subset B all the incident duration records which are higher than $T_c$ (if $y_i >$

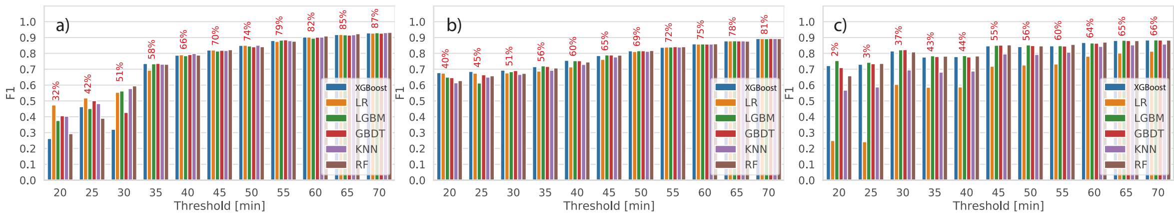



FIGURE 3.7: Incident duration classification using varying thresholds for a) data set AR b) data set M c) data set SF. The red percentage above each set of ML results indicate the percentage split of Subset A and B for that particular $T_c$.

$T_c$). Based on the variation of $T_c$, the size of Subsets A and B will have an impact on the prediction algorithms and this impact is further quantified.

The results of the binary classification approach of incident durations using a varying split threshold are detailed in Fig. 3.7 (for a 5-minutes frequency split) across all data sets. More specifically, Fig. 3.7 presents the F1 results obtained for each ML model that we have developed (XGBoost, LR, LGBM, GBDT, kNN, RF); we observe that other performance metrics have been calculated such as Accuracy, Precision and Recall and these are provided in the **??**). For example, Fig. 3.7a) showcases the classification results for data set AR in which the blue bar represents the F1-result of the XGBoost classifier (F1=0.28) when the data set has been split in Subset A containing incidents with a duration less than 20min (32% of all incident records fall in this subset) and Subset B containing incidents with duration higher than 20min (the rest of 68% of incident records). Therefore, the percentage numbers written in red above each ML result represent the percentage of records lower than the $T_c$ threshold chosen for this experiment. The split around $T_c = 20min$ is not ideal given the data imbalance (32% versus 68%) and the low F1 score; therefore further variations have been undertaken which have reported an increased $F1 = 0.8$ for $T_c = 45min$. According to these results, if we use the best performing binary classifier, we need to select a threshold between 35 and 50 minutes because: a) it will reduce the imbalance between classes (and thus reduce the effects of imbalanced classification, which is vital for modelling when using a small data set); b) there is only a tiny improvement in F1-score after $T_c > 40$min; c) it will be a reasonable split for short incidents lower in terms of field operation management. An exciting finding is revealed for $T_c \in \{20, 25\}$min: we record an overall lousy performance across all ML models in all data sets (F1-score less than 0.5) while some did not even take effect, such as GBDT; for this reason, we exclude from consideration any thresholds which provide an F1-score of less than 0.5. Furthermore, we set our minimum acceptable F1-score to 0.75, and any model performing lower than this threshold will not be considered for further optimisation. By analysing all sub-figures in Fig. 3.7 which provide both a good F1 score and class balance, we conclude that the optimal thresholds for the binary classification problem are the following: a) $T_c = 40$min for the arterial road network in Sydney (Fig. 3.7a: $F1 = 0.79$ and a class balance of 66% for small incident duration), b) $T_c = 45$min for the motorway network in Sydney, (Fig. 3.7b: $F1 = 0.75$, class balance = 65%) and c) $T_c = 45$min for the San Francisco network (Fig. 3.7c: $F1 = 0.83$, class balance=55%).

The other important finding is the cases when $T_c > 45$min which present a significant improvement across all models on all performance metrics, with the best result being the one when Subset A incorporates all incidents lower than 70min (which represents the majority of incidents); this is easily explained by the fact that we use almost all the entire data set for training of the models. However,

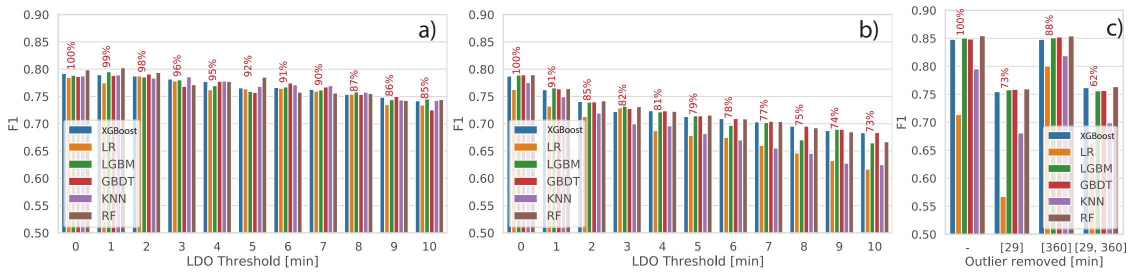



FIGURE 3.8: Outlier removal for a) data set AR b) data set M c) data set SF

the binary classification can be a rough estimate. If TMCs need a higher prediction precision instead of incidents less than 45min or higher (which can last up to several days), then several regression and multi-class classification models are needed to provide more precise predictions. These will be further detailed in Sections 6 and 7. We will further use the detected optimal thresholds for each data set to perform the split between subset A and B in various scenarios of the incident duration regression problem.

Tree-based models yield similar results. However, in multiple cases (e.g. 35, 45, 50, and 60-minute thresholds for data set AR, 25, 30, 40, 60-minute thresholds for data set M), XGBoost produces a slightly better result than other tree-based models. Thus, we are selecting XGBoost as the best model for the incident duration classification.

### 3.4.2 Classification with outlier removal

After selecting the optimal thresholds for binary classification, we further assess the effect of: a) **low-duration outliers** (LDO) (which we define as reports of incidents with zero or less than a few minutes duration) and b) **high-duration outliers (HDO)** as in the San-Francisco data set, by trying different outlier removal procedures, as depicted in Fig Fig. 3.8.

For example, an LDO Threshold of 1min represents removing outliers below 1 minute (e.g. 0min) and the percentage above each removal test. For example, 99% indicates the number of samples remaining after such removal. Removing these outliers is essential since it represents errors in the incident reporting and may affect the accuracy of prediction. For example, Fig. 3.8a represents the LDO removal from the data set AR, up until 10min reported incident durations; by removing these outliers, we observe that the F1-score does not fall below the acceptable threshold of 0.75 until 5min (this indicates that removing all accidents reported with a duration of 0 or lower than 5min does not reduce the model performance. Therefore, we applied an LDO removal for all traffic incidents for this data set with a duration below 5min. For the data set M, the effect of LDO outlier removal is more significant, as depicted in Fig. 3.8b. This data set contains a lot of incidents with duration of 0 and 1 minute (which represents almost 15% of the entire data set); by removing these, we observe that the highest F1-score drops down to 0.74 across all ML models, which falls below the acceptable threshold for a good prediction accuracy). Therefore, we decide to remove only incidents with duration of 0min or 1min from this data set. Lastly, in the case of the San-Francisco data set, we have a completely different range of outliers since there are no incidents reported with a duration of fewer than 17 minutes (see Fig. 3.8c). There are multiple incidents cleared off at around 29min and 360min (as represented as well in **??**, which can be identified as HDO. However, by removing these HDO data points from the ML model training (representing almost 38% of all incident records), we observe a depreciation of the F1 score from 0.85 to 0.76 for XGBoost, while some models dropped to lower values below 0.7). Therefore, the removal of HDO for the San Francisco data set can not be adopted due to several reasons: 1) we cannot separate "rounded" duration from actually reported duration, 2) the amount of these values is almost half of the data, which becomes property of the data set, 3) these outliers still convey information related to the separation between short-term and long-term traffic accidents and 4) all models perform better when using the entire data set than with outlier removal, which makes the ORM procedure in this case non-necessary. Finally, we observe that the outlier procedure is highly related to the specificity of the data set and the incident area location, not by making default assumptions on either LDO or HDO.

### 3.4.3 Multi-class classification

While binary classification can provide fast insights in the overall incident duration, traffic incidents can have more precise duration definition and can be split (based on the histogram profiling) into short-term, mid-term, long-term. In this case one needs to solve a multi-class classification problem. We have split this problem in two subsection in which we analyse the impact of choosing three equally sized classes, versus quantile varying split thresholds and analyse the best approach.

#### Equally split multi-class classification

Firstly, we analyse the impact of using equally-sized classes (based on duration percentiles of almost 33% from each data set). We use F1-macro to assess the performance of a multi-class classification, defined as the unweighted average of class-wise F1-scores:

$$\text{F1-macro} = \frac{1}{N} \sum_{i=0}^{N} \text{F1-score}_i \tag{3.11}$$

where i is the class index and N is the number of classes. Table 3.2 contains the F1-macro scores across all three data sets for a 3-class prediction problem which can be calculated across each data set independently. For example, $C1$ for data set AR in Sydney contains incidents between $0-24$min, while $C1$ for the SF data set contains incidents between $0-30$min; similarly, the $C3$ class for the SF data set contains substantial incidents which can reach up to 2,715min (45h) (this is consistently larger than 710min or 595min in Australia). The F1-macro score is aggregated across all classes, and a low value (below 0.5) indicates that we cannot use a 3-class split for the data set AR (F1-macro=0.35) and M (F1-macro=0.46), but we can do so for the data set SF (F1-macro=0.72). The significant difference between these data sets is the number of records (584 incident records for the data set AR versus 8,754 records for the data set SF), which may affect model performance. The precision of predictions on the data set indicates how many classes we can have to distinguish traffic incidents by duration. However, each data set's specificity seems to dictate the best classification approach to be done and further justifies the need for a more refined regression prediction approach.

| Dataset | $[0-33\%]_{C_1}$ | $[33-66\%]_{C_2}$ | $[66-100\%]_{C_3}$ | F1-macro(3-class) | F1 (2-class) |
|---|---|---|---|---|---|
| Data set AR | 0-24 min | 25-44 min | 44-710 min | 0.35 | 0.79 |
| Data set M | 0-24 min | 25-54 min | 54-598 min | 0.46 | 0.74 |
| Data set SF | 0-30 min | 31-71 min | 72-2,715 min | 0.72 | 0.85 |

TABLE 3.2: Multi-class classification results for equally-sized 3-class split

**Varying multi-class classification via quantile split**

To analyse the effect of splitting data into more varying groups we performed a multi-class classification procedure using quantiles and the F1 results are provided in Figure 3.9 for three data sets: Figure 3.9a) when using the Arterial Roads in Sydney, Australia and b) when using M7 Motorway data set and c) when using the San Francisco data set. The result metric represents an average of F1-scores across classes, where multi-class classification performed as 3 one-vs-all classifications.

The low/high threshold matrix represented in Figure 3.9 indicates a 3-class split performance and allows for the modelling of different size groups separated by quantile thresholds. As an example, the Ox axis in Figure 3.9a) represents the first threshold split ranging from [10% to 80%], while Oy represents the second threshold split percentage ranging from 20% to 90%. The coloured dots represent the F1 scores obtained when splitting the data according to the two thresholds; for example, the combination quantile pair of [10%;20%] gives an F1 score of 0.34, meaning a multi-class split of data logs between the following three classes $\{C1 = [0-10\%]$, $C2 = [10\%-20\%]$ and $C3 = [20\%-100\%]\}$ does not provide good accuracy. Instead, when using the first quantile threshold of 0.3 and the second quantile threshold of 0.6 (meaning $\{C1 = [0-30\%], C2 = [30\%-60\%]\ and\ C3 = [60\%-100\%]\}$), we obtain the highest F1-macro score, F1 = 0.44.

In the case of M7 Motorway (see Figure 3.9b), we obtain the best performance for 20% and 60% quantile thresholds (meaning $\{C1 = [0-20\%], C2 = [20\%-60\%] and C3 = [60\%-100\%]\}$; $20\%, 40\%, 40\%$ size grouping. Other options include $\{20\%, 70\%\}$ and $\{10\%, 60\%\}$ duration thresholds.

In the case of San-Francisco (see Figure 3.9c), we obtain the best performance for 10% and 90% quantile thresholds (meaning $\{C1 = [0-10\%], C2 = [10\%-90\%]$ and $C3 = [90\%-100\%]\}$; this means that the best data split when using quantile thresholds for San Francisco is a $\{10\%, 80\%, 10\%\}$ size grouping. This is highly explained by the incident distribution plots for the San Francisco area which is different than the rest of data sets.

To further see the impact on error by various incident duration groups we introduce the Quantiled Time-Folding and present the results in Figure 3.10. Incident reports are separated into equally-sized duration groups to perform the procedure of cross-validation (each 9 folds evaluated against 1 excluded fold, repeated 10 times). For all three data sets, incidents with the longest duration have the highest contribution to error, even though they represent only 10% of the data set. Considering this error, we may choose to use the hybrid classification-regression framework, where we perform regression only for intervals with acceptable prediction error. Qunatiled Time-Folding can also be useful to see the contribution to error of every duration group and possible extrapolation error towards incidents with unobserved duration groups. Also, the RMSE metric showcased in Figure 3.10 is related to the scale of duration observed in the fold (e.g. high durations can easily translate in high errors), whereas if we

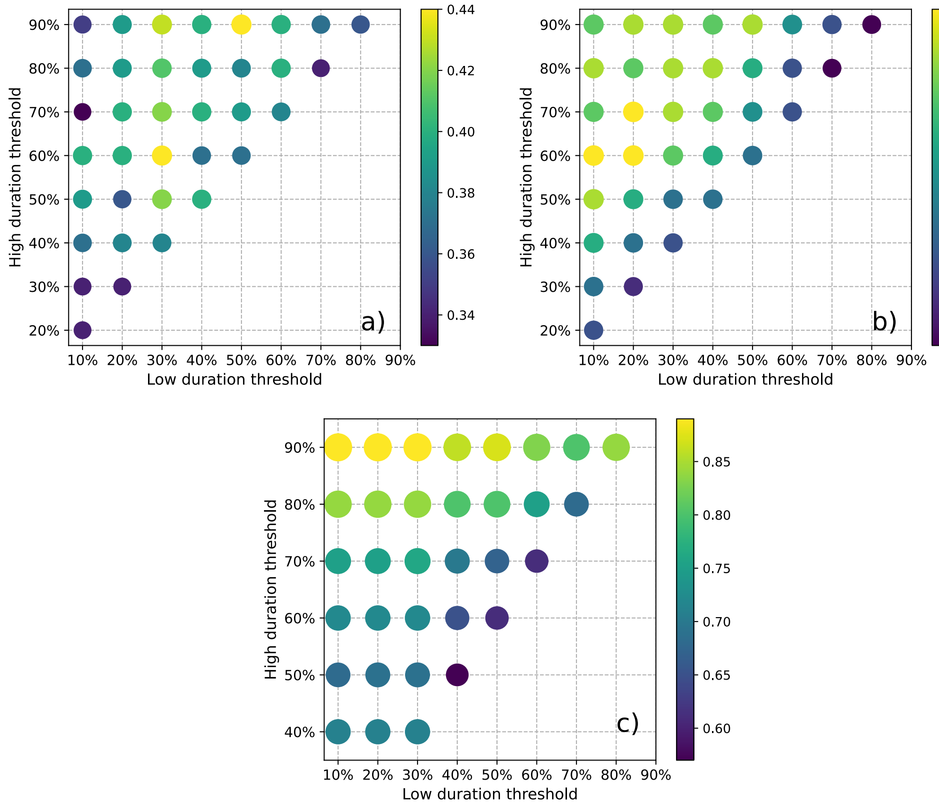



FIGURE 3.9: Multi-class (3-class) classification using quantile splits for a) data set AR b) data set M c) data set SF

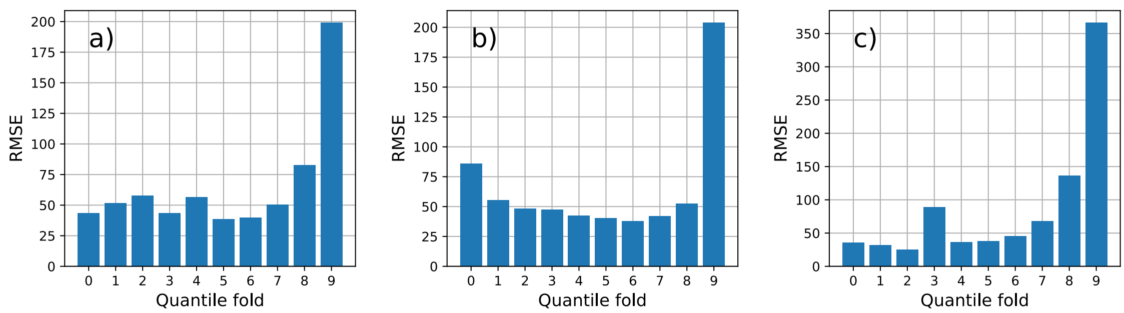



FIGURE 3.10: Regression using Quantiled Time Folding for a) data set AR b) data set M c) data set SF

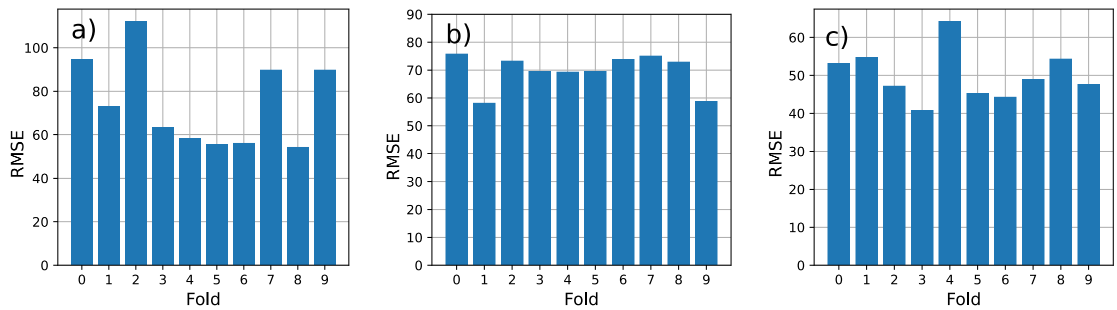



FIGURE 3.11: Regression using randomised 10-folds for a) data set AR b) data set M c) data set SF

adopt a regular 10-fold cross validation (see Figure 3.11), the RMSE error remains below 125.0 for most of the folds.

## 3.5 Incident duration prediction using regression: results

The final objective of the bi-level framework is to predict with an accuracy at the minute level the length of a freshly reported incident, regardless of its previous classification as either short, medium or large. Therefore, the second step of the bi-level prediction framework is to develop more advanced regression models that can adjust to each data set independently and over-perform baseline ML models previously used to solve classification problems. When training such regression models, a significant step is the size of the data set and the distribution of the target variable (incident duration). Due to the long tail distribution of incident duration and the class imbalance problem previously identified, we need to design and construct various regression models capable of learning from various types of data sets to make accurate predictions. However, with limited information (small data set size), the prediction results can be skewed (this effect of prediction skewing will be further discussed). This section first presents the regression results obtained across several scenarios of model training, validation and testing, followed by results of our proposed Intra-Extra Optimisation algorithm applied over all baseline ML models.

### 3.5.1 Regression scenarios results and comparison

In order to find the best set-up that works for traffic incident prediction in TMCs, we test various regression scenarios (detailed previously in Section 3.3.6), which show the extrapolation performance for different ML methods. The outlier removal procedures (LDO, HDO) together with the classification thresholds (which separate short-term and long-term duration of incidents) are selected as described in Section 3.4.1-Section 3.4.2. The primary purpose of this section is to recommend the best scenario set-up for model training and validation when parts of the data set might be hidden. Table 3.3, Table 3.4 and Table 3.5 present the MAPE results for all 7 scenarios (All-to-All, AtoA, AtoB, BtoB, BtoA, AlltoA AlltoB) using all the Baseline ML models across all three data sets (and a dedicated winning regression model across each scenario - last column). Overall, XGBoost seems to be the best regression model in a majority of scenarios across data set AR and M (Table 3.3,Table 3.4): 1) the improvement from using XGBoost shows the lowest MAPE for scenario AtoA of 49.11 and 67.92 correspondingly (predicting short term incidents only using only short term training information), 2) XGBoost also the best performing model for All-to-All regression (59.36% and 85.98% MAPE correspondingly). The main difference between LGBM and XGBoost results is that LGBM struggles with extrapolation to lower values as seen in scenario B-to-A for all data sets: 292.68% vs 77.66% MAPE for data set A, 663.12% vs 180.77% MAPE for data set M, 166.06% vs 32.62% MAPE for data set SF for LGBM and XGBoost correspondingly.

In the SF data set, the LGBM is the best performer reaching a MAPE of 9.34% for the AtoA scenario (which is almost 10 times better than the same scenario for the M data set) and 33.16% MAPE for All-to-All scenario. This is a significant improvement that reveals what model is adapting to what data set, but most importantly, that each data set reacts differently to the seven scenarios.

| Model | LGBM | RF | LR | GBDT | KNN | XGBoost | Best model |
|---|---|---|---|---|---|---|---|
| AlltoAll | 82.76 | 117.28 | 110.99 | 113.41 | 107.79 | **59.36** | XGBoost |
| AtoA | 60.17 | 59.49 | 59.92 | 62.08 | 58.35 | **49.11** | XGBoost |
| AtoB | 64.46 | 64.39 | 64.34 | **63.82** | 64.68 | 64.39 | GBDT |
| BtoA | 292.68 | 381.61 | 367.16 | 348.09 | 349.62 | **77.66** | XGBoost |
| BtoB | 29.52 | **25.03** | 45.14 | 46.26 | 43.82 | 27.55 | RF |
| AlltoA | 117.78 | 121.82 | 175.48 | 176.71 | 120 | **51.18** | XGBoost |
| AlltoB | 34.39 | 37.47 | 32.11 | **31.67** | 35.57 | 37.46 | GBDT |

TABLE 3.3: MAPE results for all 7 scenarios on data set AR

| Model | LGBM | RF | LR | GBDT | KNN | XGB | Best model |
|---|---|---|---|---|---|---|---|
| AlltoAll | 135.59 | 226.6 | 229.53 | 229.46 | 229.82 | **85.98** | XGBoost |
| AtoA | 95.89 | 95.38 | 107.29 | 104.87 | 105.26 | **67.92** | XGBoost |
| AtoB | 68.78 | 69.01 | 69.49 | **68.62** | 69.79 | 68.69 | GBDT |
| BtoA | 663.12 | 939.59 | 818.08 | 878.47 | 854.81 | **180.77** | XGBoost |
| BtoB | 34.14 | 51.02 | 52.33 | 50.99 | 48.68 | **31.18** | XGBoost |
| AlltoA | 233.48 | 406.43 | 387.25 | 398.13 | 402.02 | **76.71** | XGBoost |
| AlltoB | 34.38 | 34.34 | **34.21** | 34.48 | 36.89 | 34.98 | LR |

TABLE 3.4: MAPE results for all 7 scenarios on data set M

In the following, we provide a summarised comparison across a selection of few scenarios and their performance.

**Scenario AtoA** uses short-term traffic accidents (below $T_c$) for both training and the prediction. XGBoost shows a significant performance for AR and M data sets compared with other scenarios; more specifically, they outperform by 10% all models in data set AR (MAPE=51.2) and 30% all models in dataset M (MAPE=68.4). For the SF data set, the improvement is even larger (MAPE=12.7), but XGboost loses ground over LGBM, which reaches a MAPE=11.0. The comparison of scenarios AtoA and AlltoA shows that adding incidents with a longer duration can severely affect the prediction performance across all data sets, regardless of the size or location of the incident logs. For the best prediction performance on data sets AR, M and SF, we need to split the data and use separate models for the short-term incidents as predictions become skewed towards longer incident duration. Thus, if we predict short-term incidents using only short-term incidents data logs, we obtain a higher accuracy across all data sets.

| Model | LGBM | RF | LR | GBDT | KNN | XGBoost | Best model |
|---|---|---|---|---|---|---|---|
| AlltoAll | **33.16** | 36.88 | 128.42 | 41.85 | 64.24 | 37.03 | LGBM |
| AtoA | **9.34** | 11.91 | 16.07 | 12.56 | 14.05 | 11.44 | LGBM |
| AtoB | 68.08 | 65.77 | 67.21 | **65.53** | 66.26 | 65.84 | GBDT |
| BtoA | 166.06 | 191.55 | 389.07 | 211.61 | 302.46 | **32.62** | XGBoost |
| BtoB | **23.69** | 28.76 | 70.18 | 31.08 | 37.6 | 27.61 | LGBM |
| AlltoA | 45.35 | 50.74 | 218.49 | 60.03 | 99.06 | **35.49** | XGBoost |
| AlltoB | 24.28 | **23.97** | 45.08 | 25.49 | 30.82 | 24.78 | RF |

TABLE 3.5: MAPE results for all 7 scenarios on data set SF

**Scenario AtoB** is unique because regression models are trained on Subset A, which contains short-term incident duration logs while they are trying to predict long-term incidents; therefore, the performance is much worse than for AtoA scenario since incidents with long duration are much scarcer and have unique traffic conditions. BtoB scenario shows lower error than AtoB across all three data sets (e.g. BtoB provides 23.69% MAPE and AtoB provides 65.53% MAPE for best models for data set SF). Vice-versa, **Scenario BtoA** shows very high extrapolation errors across all methods to lower values. Adding short-term incidents into the training set of long-term incidents (when we move from BtoA to AlltoA scenario) significantly reduces the error (76.71% MAPE for scenario AlltoA, data set M using XGBoost), but it is still significantly higher than for AtoA scenario (67.92% MAPE for M data set using XGBoost). **Scenario BtoB** shows better performance (e.g. MAPE=31.18% for data set M using XGBoost) than using data addition (such as the case of AlltoB, where MAPE=34.21% using best model) or any extrapolation (as in the case of AtoB, where MAPE=68.62% using best model). By comparing scenarios AtoB and AlltoB we observe a significant performance improvement when adding data for long-term incidents and predicting subset B (from 63.82% to 31.67% MAPE for dataset AR using best model), where error is still higher than for BtoB (25.03%, AR, best model). **Scenario BtoA** shows high prediction errors across all scenarios highlighting a bad extrapolation accuracy when predicting short-term incidents duration using long-term traffic incident data. It means that prediction of the duration of short-term incidents should be performed separately from long-term incidents. Thus, we can't use long-term incidents to predict the duration of short-term incidents and vice versa if we are looking at maximising model performance with limited data set; the second reason lies mainly in different traffic behaviour along with severe accidents that can last for several hours which are harder to clear off - these require similar previous events in order to be predicted for their duration.

#### Fusion framework for the incident duration prediction

In comparison to the above proposed framework, we also present a fusion framework approach, which can be applied when the incident duration category is unknown. When an incident occurs, the incident duration category is not known, but we have a historical data on traffic incidents which allows us to predict the incident duration category and apply specialised regression models (oriented towards the prediction on subsets of short-term and long-term incidents). We further propose two possible approaches to this problem:

- **the pipeline approach** (see Fig. 3.12a ): we train a classification model using a historical data available to predict the incident duration category. Then we predict the incident duration category using the available incident reports. This prediction decides which model we need to use for a further regression (either specialised on short-term or long-term incident duration prediction). The prediction result of the specialised model is then considered to be the final prediction. Specialised regression models are trained on their corresponding subsets. In this case, the decision about the incident duration class is made by the classification model only, which becomes the most important part of the model that is highlighted by significantly improved results (see Tables 3.3 to 3.5).
- **the fusion approach** (see Fig. 3.12b ): instead of relying on the classification model to decide on the incident duration subset, we place a decision-making function on the additional "fusion

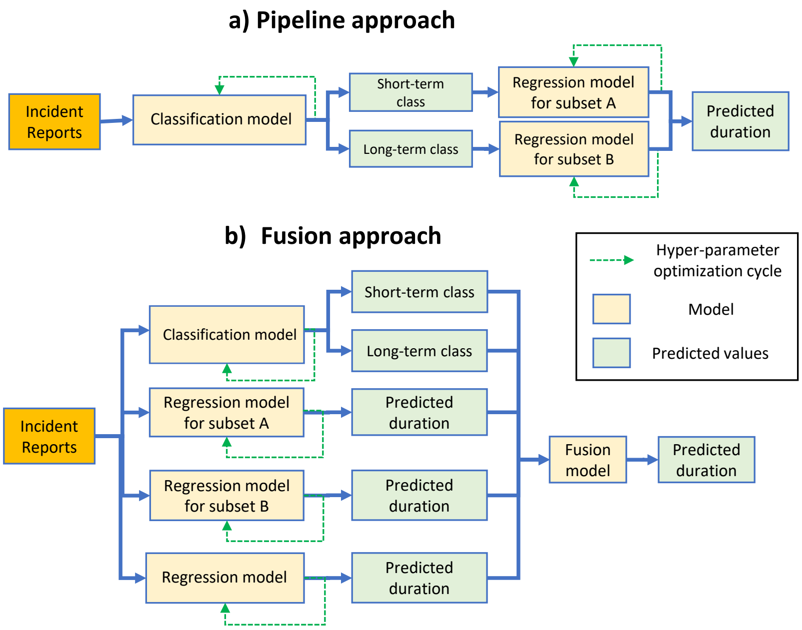



FIGURE 3.12: Pipeline (a) and fusion (b) approaches for the bi-level framework structure

model", which is the global regression model; it now receives the prediction results from the classification model, the regression models specialised on short-term and long-term incidents (subsets A and B) and from the regression model trained on historical data of traffic incidents regardless of the incident duration group. After training all these models on historical data, we perform the incident duration prediction on this historical data. We then use these predictions (such as the predicted incident class, the incident duration predicted by short-term incident duration regression model, the incident duration predicted by the short-term incident duration regression model, and the incident duration predicted by the regression model) in order to train the global fusion model to make a final prediction of the incident duration; we call this the global fusion model and the predicted duration is a result of multiple models fused in a centralised architecture.

The fusion approach can be perceived as the ensemble model, which allows to solve the computational problem of model training. Ensemble models may perform better than single models due to three main reasons: [51]: a) statistical: without sufficient data, a model can find multiple hypothesis about the data approximation which has the same accuracy. Each of these hypotheses can lean towards its local optima. By averaging hypotheses, we may find a better approximation of the data; b) computational: many machine learning models may get stuck in a local optima (e.g. stochastic gradient descent in the case of neural networks or the greedy split finding in the case of decision trees). An ensemble constructed by models performing local search from many different starting points may provide a better prediction performance than the individual models [51, 16], c) representational: each model forms an approximation (representation) of the data, which forms a local representation hypothesis. By combining models it is possible to extend the space of representable functions.

In the case of a bi-level framework we have statistical, computational and representational reasons to expect a better performance from using an ensemble model rather than a single model, since we use different kinds of models on different subsets (in our case a simple regression model, a classification model, a regression model for subset A, a regression model for subset B). In other words, by splitting the data and by using multiple models we obtain models that are having different local optima (subset A, subset B models) and a different representation of the data (classification and regression models); in this way we can obtain a better prediction performance using model ensemble than using individual models.

Finally, we compare the fusion model, single regression model (e.g. for the data set SF it is the model with the best performance for the task of All-to-All regression) and the pipeline model (where the choice between the regression models depends on the predictions from the classification model) performance on all three data sets in Fig. 3.13. We evaluate all model performance on each fold in a randomised 10 fold cross-validation. We observe that the fusion model performs at least not worse than a single model on all three data sets. We use XGBoost as a fusion model. We also use the corresponding best models for each subset of each data set (see Tables 3.3 to 3.5) with hyper-parameter optimisation (e.g. LightGBM as a single model, performing All-to-All regression task for the data set SF, RandomForest as a best classification model for data set SF according to Fig. 3.7). There is a subtle difference in the average RMSE score among the folds for data set A ( see Fig. 3.13a) where the average RMSE for the fusion model is 59, for the single regression model is 59.8, for the pipeline model is 62.2). The same is for the data set M (see Fig. 3.13b) where 68.9, 68.4 and 70.8 are the average RMSEs for the fusion, single and pipeline models correspondingly). There is a significant improvement in the average RMSE score for data set SF (see Fig. 3.13c) with an improvement from 73.6 to 58 of the average RMSE when using of the fusion model instead of the single model); the pipeline model didn't show any improvement in the model performance. Overall, results show that data availability (the amount of information available about the incidents, which is high for the data set SF) can significantly affect the performance of the fusion model.

Given the observed performance (from a subtle difference to significant improvement) we recommend to use the fusion approach within bi-level framework for the task of the incident duration prediction.

### 3.5.2 Outcomes and recommendations

From an operational perspective the scenario All-to-All is the ideal situation when traffic management centres would have in their data base both long term and short-term incidents. However, from an operational perspective, several records of short incidents for example and not being kept all the time, while long incidents are often being transferred to various other division if they last more than one day, and they become more of a road infrastructure problem rather than an operational problem which requires constant intervention.

Scenario modelling shows that the baseline ML models are not improving when facing incident duration extrapolation or data addition (e.g. AlltoA versus AlltoA, BtoB versus AlltoB); these two training set-ups badly affect the model performance extrapolating in any direction.

By evaluating regression scenarios, we highlight the importance that incidents from different duration groups need to be modelled separately in order to significantly improve the accuracy of duration predictions (see (see Table 3.5) for SF data set: if we use all available data, to predict the incident

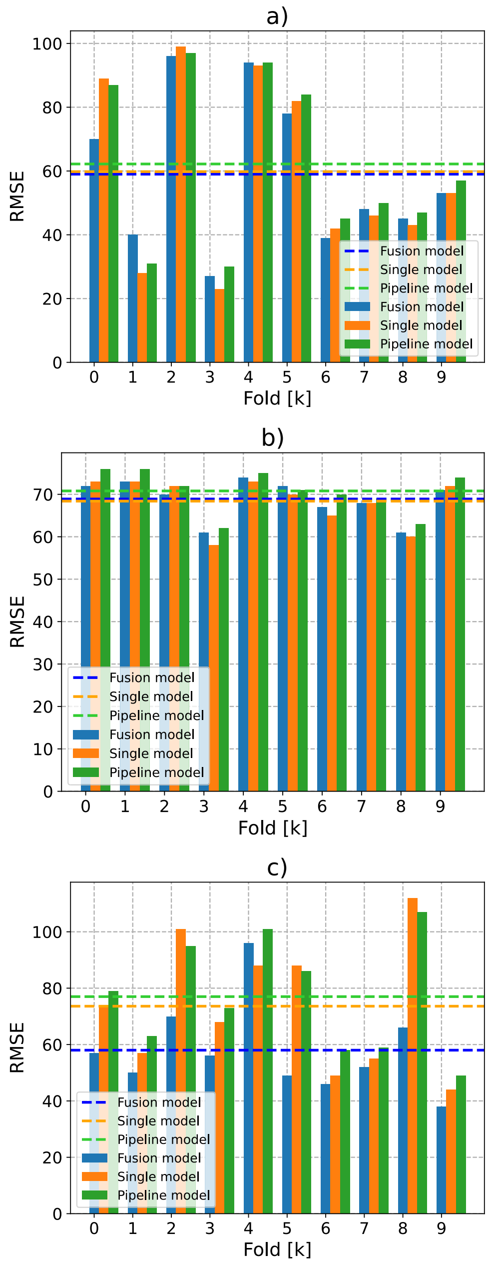



FIGURE 3.13: Comparison of the fusion and single model performance for a) data set AR b) data set M c) data set SF. Dashed lines represent the average RMSE score across all folds for each corresponding model

| $ML_j$ | Log | Unprocessed | iIF-Log | eIF-Log | eLOF-Log | iLOF-Log | Best approach |
|---|---|---|---|---|---|---|---|
| LGBM | 80.4 | 81.1 | 79.9 | 82 | **78.4** | 80.8 | **eLOF-Log-LGBM** |
| RF | 80.3 | 121.9 | 79.5 | 80.7 | **78.5** | 79.1 | **eLOF-Log-RF** |
| LR | **80.0** | 128.4 | 80.4 | 81.6 | 80.5 | 80.5 | Log-LR |
| GBDT | **79.4** | 128.2 | 82.0 | 81.3 | 81.4 | 83.4 | **Log-GBDT** |
| KNN | 82.9 | 127.4 | 82.3 | 86.2 | 81.7 | **81.3** | iLOF-Log-kNN |
| XGBoost | **59.4** | 61.1 | 60.8 | 59.8 | 60.9 | 59.9 | **Log-XGboost** |
| Best $ML_j$ | XGBoost | XGBoost | XGBoost | XGBoost | XGBoost | XGBoost | |

TABLE 3.6: MAPE results for All-to-All scenario of data set A, using different ORM approaches and incident duration transformation, via the proposed IEO-ML approach.

duration, we will have $MAPE = 33.16\%$ (lower is better), but if we managed to categorise incidents into the short-term group, we could model these incidents with only 9.34% error, which is a significant improvement. Also, the classification may point us to which data we need to include in the modelling because if we use all data to predict the duration of the short-term incident (scenario AlltoA), we will have a much higher error 45.35% MAPE than just using the short-term incident for modelling. From the comparison regression of extrapolation scenarios (e.g. scenarios AlltoA versus AtoA), we see how significant can be the impact of having incidents with long duration in the training set when we need to predict the duration of short term incidents, and therefore ML methods become biased towards long-term incidents, which significantly reduces their performance. If we can perform incident duration regression, then we are able to perform incident duration classification as well. We can do this before performing the regression in each group. In other words, our scenario modelling shows modelling advantages of classifying incidents into duration groups.

Therefore, it is essential for the bi-level framework and traffic incident duration prediction to use separate models for short-term and long-term traffic incidents. Moreover, tree-based methods significantly outperforming LR demonstrates that traffic incident regression is a complex non-linear problem that requires more advanced investigations. This aspect was the one that motivated our research to further improve and build a better ML framework for any type of incoming data set, and the results of this novel IEO-ML framework are further detailed in the following section.

### 3.5.3 Regression results for proposed IEO-ML model

In this section, we employ our proposed Intra-extra joint optimisation approach previously presented in Section 3.3.9 and we further present the results of the All-to-All regression scenario, with a log-transformation of incident duration and several outlier removal techniques such as the LocalOutlierFactor (LOF) and the IsolationForest (IF), previously described in Section 3.3.7. All results across the three data sets are presented in Table 3.6-Table 3.7-Table 3.8.

For the data set A (Table 3.6), we observe a significant impact of using the log-transformation of the incident duration vector via the resulting MAPE (see Unprocessed versus Log columns). Since the log-transformation provides a significant improvement among majority of ML models, we decide to use it in our outlier removal scenarios. When comparing results across all models, both regular and re-enforced by our IEO approach (column comparison - see Best $ML_j$ results), we observe that XGBoost is the best performing baseline model for this data set reaching a 59.4 MAPE. Furthermore, when comparing results across regular ML models versus our proposed IEO-ML enhancements (row

| $ML_j$ | Log | Unprocessed | iIF-Log | eIF-Log | eLOF-Log | iLOF-Log | Best approach |
|---|---|---|---|---|---|---|---|
| LGBM | 124.6 | 138.0 | **123.6** | 126.8 | 125.1 | 124.1 | **iIF-Log-LGBM** |
| RF | 126.3 | 238.6 | 126.6 | **125.7** | 127.1 | 126.6 | **eIF-Log-RF** |
| LR | 130.7 | 245.9 | **129.8** | 129.9 | 131.1 | 131 | **iIF-Log-LR** |
| GBDT | **126.7** | 240.1 | 126.9 | 126.7 | 127.2 | 126.9 | **Log-GBDT** |
| KNN | 139 | 248.2 | **135.1** | 137 | 139.4 | 138.2 | **iIF-Log-KNN** |
| XGBoost | 78.6 | 113.2 | **77.5** | 80.6 | 78.3 | 79.6 | **iIF-Log-XGBoost** |
| Best $ML_j$ | XGBoost | XGBoost | XGBoost | XGBoost | XGBoost | XGBoost | |

TABLE 3.7: MAPE results for All-to-All scenario of data set M, using different ORM approaches and incident duration transformation, via the proposed IEO-ML approach.

| $ML_j$ | Log | Unprocessed | iIF-Log | eIF-Log | eLOF-Log | iLOF-Log | Best approach |
|---|---|---|---|---|---|---|---|
| LGBM | 29.9 | 32.6 | 29.7 | **29.5** | 30.2 | 29.9 | **eIF-Log-LGBM** |
| RF | 28.9 | 38.7 | **28.7** | 28.9 | 28.8 | 28.9 | **iIF-Log-RF** |
| LR | 72.6 | 140.5 | 72.8 | 73.1 | 73.3 | **72.4** | **iLOF-Log-LR** |
| GBDT | **31.2** | 46.3 | 31.5 | 31.4 | 32.4 | 32.2 | **Log-GBDT** |
| KNN | **61.5** | 108.6 | 61.7 | 62.5 | 62.2 | 61.8 | **Log-KNN** |
| XGBoost | 31.7 | 35.1 | 31.9 | 31.6 | 32.7 | **31.0** | **iLOF-Log-XGBoost** |
| Best $ML_j$ | RF | LGBM | RF | RF | RF | RF | |

TABLE 3.8: MAPE results for All-to-All scenario of data set SF, using different approaches for ORM and incident duration transformation, via the proposed IEO-ML approach.

comparison), then the extra optimisation approaches seem to outperform the intra optimisation approaches (see iIF-Log versus eIF-Log and eLOF-Log versus iLOF-Log columns). The last column indicates the best approach that won across all proposed IEO approaches where for example, eLOF-Log-RF model is read as the extra optimisation method applied together with the Local Outlier Factor and Random Forest over the log scale data transformation; for this data set A results indicate a similar performance between using baseline ML models with log transformation versus enhanced IEO-ML - for example the joint optimisation provides an improvement (eLOF-log-LightBGM, eLOF-log-RF) versus the cases cases when only the baseline ML with the log-transformation was used (e.g. Log-LR, Log-BDT). However, the A data set is very small and has a special behaviour when compared to the others as further results revealed.

For the data set M (Table 3.7), when we use Log-transformation, we observe very high MAPE scores (100% and higher), except for XGBoost, which provides a MAPE of 78.6%. When comparing the models with each other against the IEO enhancements as well (column comparison), using XGboost as a baseline seems to over-perform all the other approaches, with the best results being a MAPE=77.5 for iIF-Log-XGBoost. When comparing against the proposed approaches (row comparison), the Intra joint optimisation using Isolation Forest in log-transform shows the best performance on this data set for four models (iIF-Log-LGBM, iIF-Log-LR, iIF-Log-kNN, iIF-Log-XGBoost), which can be attributed to data set data structure - outliers can be better analysed using tree-based outlier removal methods rather than distance-based LOF. For the majority of models (4 out of 6), our proposed joint optimisation algorithm obtains the best results for this data set.

For the data set SF (Table 3.8), we observe two competing models - LGBM and Random Forests with a prevalence for Random Forests (column comparison - see Best $ML_j$ results). Also, we observe a considerably lower MAPE score for the best performing models which reached the lowest threshold of 28.7 across all the data sets used in this study. This reveals the power of more complete and larger data sets which can significantly improve the model performance. When comparing the IEO approaches (row comparison), the intra joint optimisation shows improvement across three models and more specifically for the best performing model on this data set, RF. One consistent finding across all results is the fact that the log-transformation of the incident duration vector should be used at all times for incident duration prediction since it significantly improves predictions accuracy; this is mostly related to the long tail distribution and extreme outliers which can affect the final errors in the model performance evaluation. Overall, the best performing models are considered to be XGBoost and Random Forests.

**To summarise**, every data set has its specifics in the data structure, which make some models and outlier removal methods performing better than others. Thus, it is necessary to deploy different models and outlier removal approaches on every data set. Conventional models (KNN and Linear Regressions) show the highest error which is almsot twice in comparison to tree-based models. Thus, tree-based models are preferred options for solving the incident duration prediction together with adapted optimisation and outlier techniques. Overall, we proved that our proposed intra joint optimisation is improving the regression results across multiple data sets (especially data sets M and SF in 7 out of 12 cases). The joint optimisation of the model together with the outlier removal method shows a significant improvement in majority of cases (12 out of 18) across all three data sets.

### 3.5.4 Bi-level framework implementation

The code for the bi-level framework exploring previously described scenarios can be found by the link:
`https://github.com/Future-Mobility-Lab/bi-level-framework`

## 3.6 Feature importance impact and evaluation

Finally, we evaluate the feature importance using a Shapley value calculation in order to estimate the contribution of each feature to the final prediction score. Each point related to a feature is shown in Fig. 3.14 and represents the SHAP value score (Oy-axis), coloured by its value (from low to high),while the Ox-axis shows the impact of that feature information on the entire prediction output. The used models for this feature importance analysis are the winning models of each data set (A, M, or SF) as previously discussed.

The hour-of-the-day when the incident started is among the top 5 features sorted by importance (ranked on the $1^{st}$ place for data set A, $3^{rd}$ for M and $4^{th}$ for SF). For example, Fig. 3.14a) showcases that as the hour of the day increases (getting closer to midnight) the traffic durations are lower as the congestion is lower and rescue teams arrive faster to the accident location; this is the opposite on the motorways as Fig. 3.14b) reflects that rescue teams havea a harder time reaching the incident location in the evening, which is mostly explained by the high distance of the motorway from the local incident management centre. The incident reporting source also has a high significance ( ranked as $7^{th}$ most important for A, $2^{nd}$ for M, $2^{nd}$ for SF). The Ox-axis on SHAP plots represents the impact on model output (e.g. the effect on the predicted duration value). Even though the average temperature is considered significant, its effect on the regression model output is very small $[-5min; +5min]$ for data set

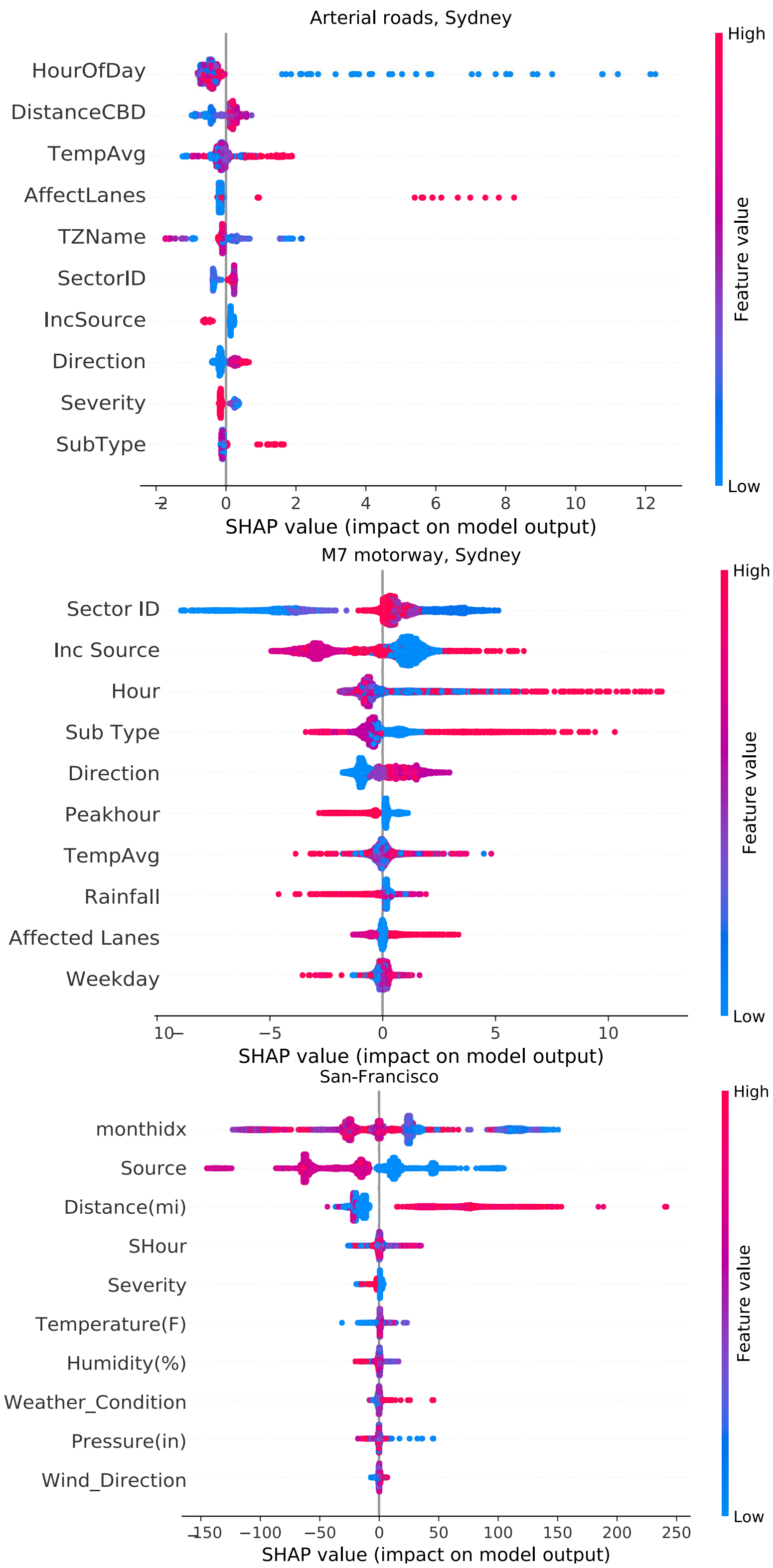



FIGURE 3.14: Feature importance for All-to-All regression using XGBoost for a) Arterial roads, Sydney, Australia b) M7 motorway, Sydney, Australia c) San-Francisco, USA

AR, $[-5min;+5min]$ for data set M, $[-25min;+25min]$ for data set SF. The distance from CBD (DistanceCBD) is important in the data set A, as it can point at some problematic areas, therefore causing a higher incident duration. The number of affected lanes is also an important feature for incident duration prediction on arterial roads in Sydney. The model outputs for the M7 motorway revealed that is highly dependent on the sector ID (similar to the traffic zones in the data set A), which may be linked to the nature of the location or to the distance from incident management agencies. The average daily temperature also affects predictions ($3^{rd}$ place in A, $7^{th}$ in M and $6^{th}$ in SF). Weather factors (rainfall) are found to play a significant role in the M and SF data sets (humidity and barometric pressure may be predictors of rainfall). Different incident sub-types in the M data set (e.g. car, motorcycle, truck, multi-vehicle) contribute to the difference in the accident duration. Severity is weakly connected to the incident duration in the A and SF data sets. It is important to note that the SF data set contains 49 features, but 39 are of very low importance for the incident duration prediction. The length of the affected road segment (Distance in SF) may also be an essential feature which is not found in Sydney data sets. Overall, the specificity of each data set is reflected once again not only in the models that may be more successful than others but also in the way that the same model can provide various feature importance due to each country, their unique landscape and different way of dealing with the disruptions.

### 3.6.1 Short-term vs long-term incident duration prediction feature importance

We further perform a comparison of feature importance for the duration prediction of short-term vs long-term traffic incidents, across all data sets.

**Arterial Roads Feature Importance, Sydney Australia.**

Fig. 3.15 showcases the Feature importance for All-to-All regression using XGBoost for a) short-term incidents b) long-term incidents of Arterial roads, Sydney, Australia. When analysing long-term incidents, one important observation is the direct influence of the number of affected lanes on the severity and duration of disruptions. However, this feature is found to have low importance for short-term incidents. The farther short-term incidents happen from the CBD, the longer it takes to clear them off. The location of the incident is extremely important for both long and short term incidents, most likely due to the easiness to reach the affected location by the intervention teams. Another important factor affecting the short term incidents in Sydney seems to be the travel patterns for commuting [month of the year, day of week, sectionID, section capacity]. Also, the DayOfWeek (value ranges from 0 to 6), we see that the higher the value (closer to the end of the week), the longer it takes for the incident to clear. Also, some sectors reflected by the SectionID feature demonstrate a lower incident duration, which may highlight that some specific areas of the city are less affected by traffic incidents.

**Motorway Feature Importance, Sydney Australia.**

Fig. 3.16 showcases the Feature importance for All-to-All regression using XGBoost for a) short-term incidents b) long-term incidents of M7 Motorway, Sydney, Australia. One immediate observation is the fact that the data has 3 sources of reporting, and this can be seen as three different distributions in the top 1 most important feature ranked in Figure 4a). The source reporting the incidents seems to be

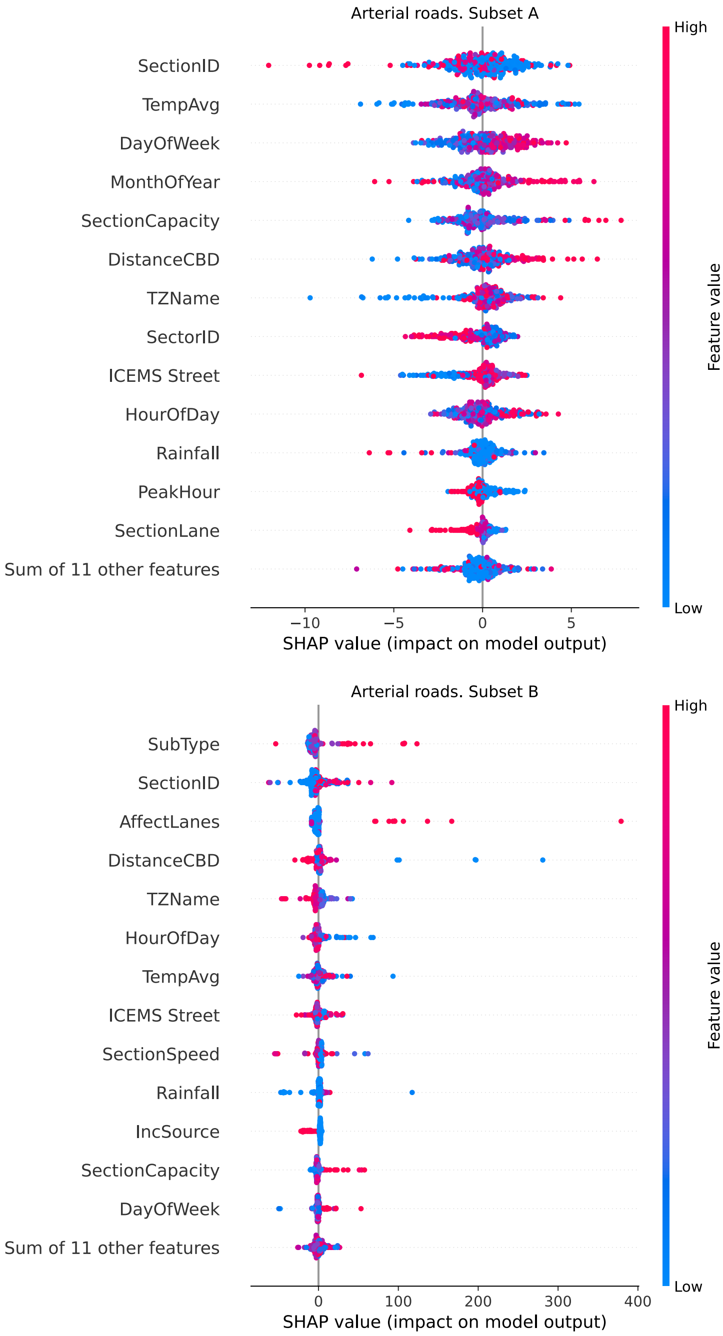



FIGURE 3.15: Feature importance for All-to-All regression using XGBoost for a) short-term incidents b) long-term incidents of Arterial roads, Sydney, Australia

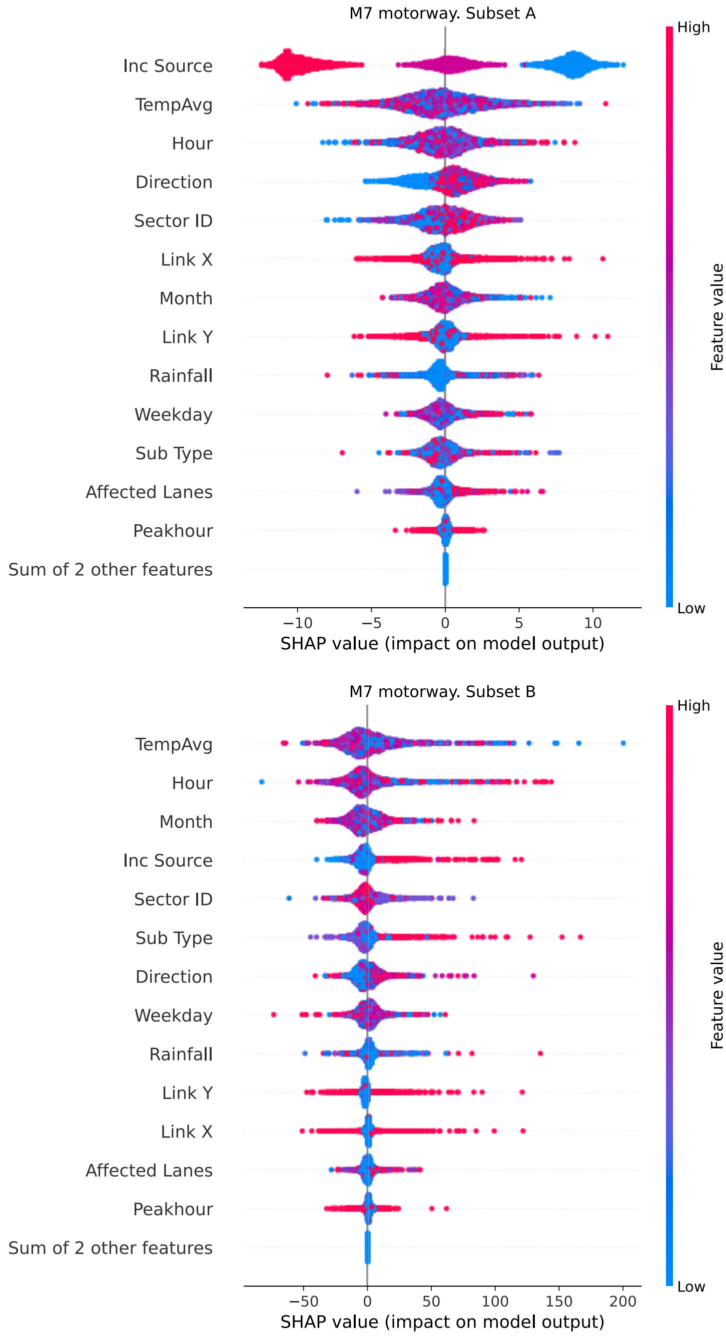



FIGURE 3.16: Feature importance for All-to-All regression using XGBoost for a) short-term incidents b) long-term incidents of M7 Motorway, Sydney, Australia

the one factor which influence the most the incident duration. When comparing the top features for both short versus long term incidents, these are almost the same in both subsets: average temperature, the hour when the incident happened, the Sector ID, the direction of travel and the source of information that reported the incidents. Overall, for this data set, same features can be collected for both types of incidents.

#### San Francisco Feature Importance, U.S.A.

Finally, Fig. 3.17 showcases the feature importance for All-to-All regression using XGBoost for a) short-term incidents b) long-term incidents of San-Francisco, USA. This data set is very different than the rest, but as in the case of M7 motorway, the source reporting the incident seems to be most important factor affecting the duration – this is mostly related to the way the information is received to the centre (from road users, from local traffic agents, from video camera surveillance, etc.). We observe that short-term and long-term incident are very different in their nature and incident characteristics found to have different importance in the prediction of the incident duration. For the SF data set, the most important features are Source, monthidx, Shour, regardless of the incident duration. In terms of large accidents however, the distance from the CBD is very important while for small accident the humidity plays an important factor ranking 4th (which might indicate that weather in San Francisco can cause small traffic accidents to happen often). Overall, despite all data sets being different, their specificity and feature important is highly related to their setup, the location of the network and the way the management centre received and handle the disruption. In order to help improve the prevention techniques more effort should be invested in understand which source of incident reporting causes the most errors overall and why.

#### Additional results for the threshold variation along all data sets such as (Accuracy, Precision and Recall)

The prediction of traffic accident duration is a critical aspect of traffic management and emergency response planning. Accurate forecast enables timely and accurate response, impacting urban mobility and safety. The challenge is to make a predictive model that balances various evaluation metrics, such as Accuracy, Precision, and Recall (see Figure 3.20). These metrics are influenced by the choice of incident duration threshold within the model.

#### Additional information with regards to the computational time of various baseline ML models across the three data sets

The findings indicate the RF and kNN seem to be the slowest models to train versus LGBM and XGBoost and LR which are faster from a computational time point of view (see Figure 3.21).

## 3.7 Conclusion

In this chapter, we present a novel bi-level framework designed for the prediction of incident durations. This framework combines baseline machine learning models, dedicated to both classification and regression tasks. Moreover, an outlier removal procedure is incorporated into the framework to

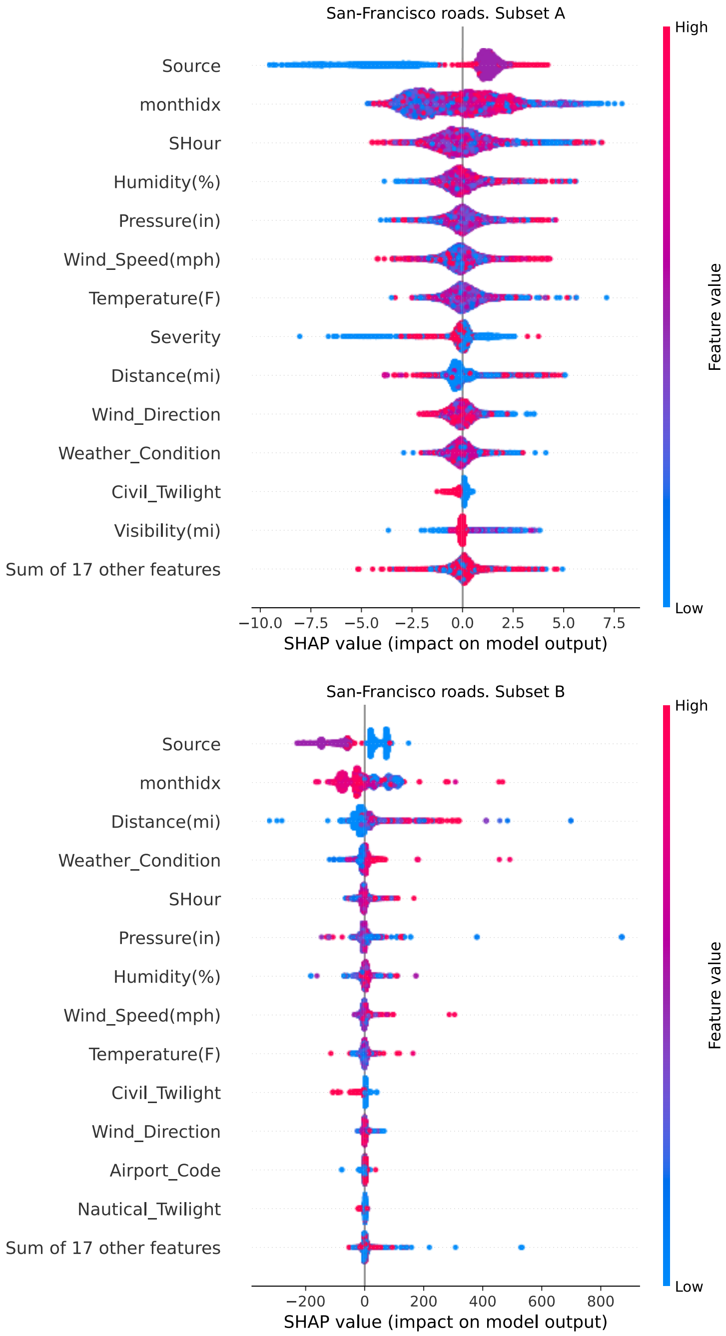



FIGURE 3.17: Feature importance for All-to-All regression using XGBoost for a) short-term incidents b) long-term incidents of San-Francisco, USA

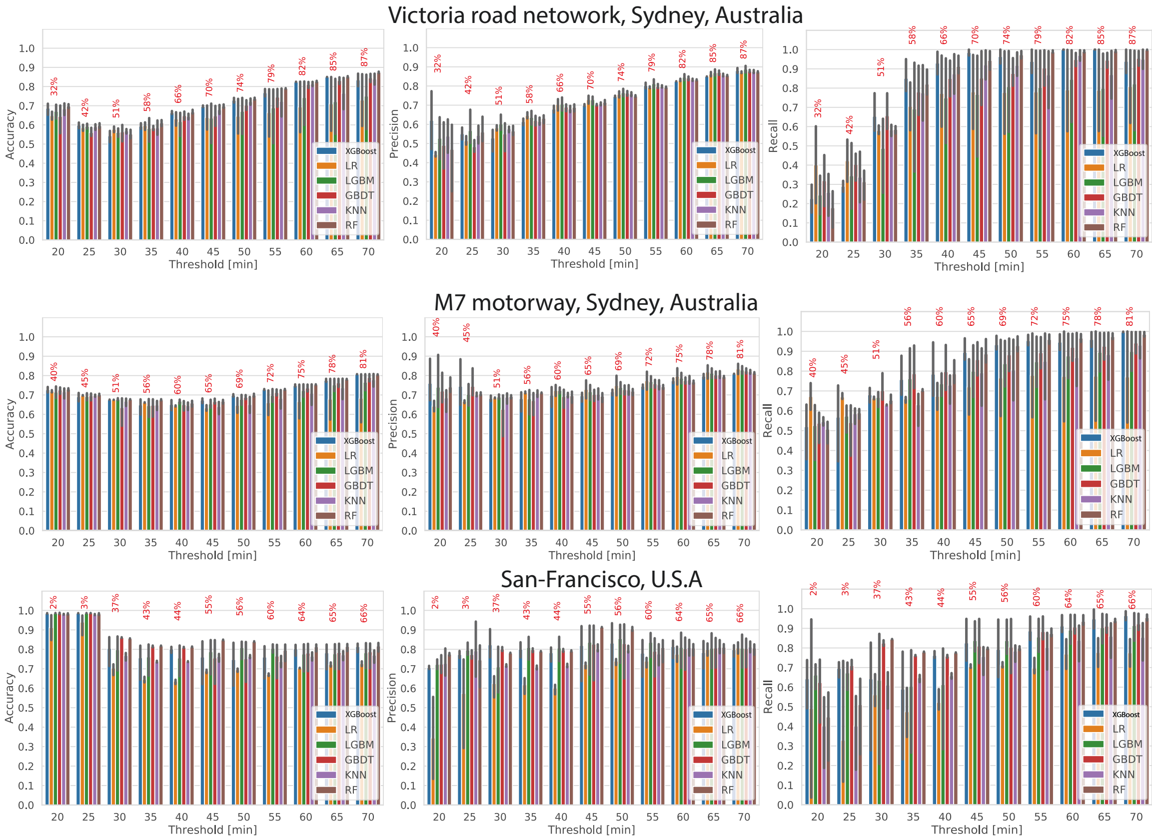



FIGURE 3.18: Binary classification performance using varying incident duration threshold

enhance the accuracy of the predictions. Additionally, a novel intra-extra joint optimization technique is introduced, further improving the efficacy and performance of the framework. To validate the proposed approach, three distinct datasets, from Australia and the United States of America, were utilized across various testing and validation scenarios.

**Major contributions:** The study has several contributions to the field. Firstly, it was found that the optimal duration classification thresholds for incident prediction, estimated using data-driven approach, were consistent across different data sets, with durations of 40 minutes for data set AR, 45 minutes for M, and 45 minutes for SF. This alignment with the threshold used by Sydney TIMS, based on their expertise in on-the-field operations, reaffirmed the coherence of our threshold split with realistic operational times.

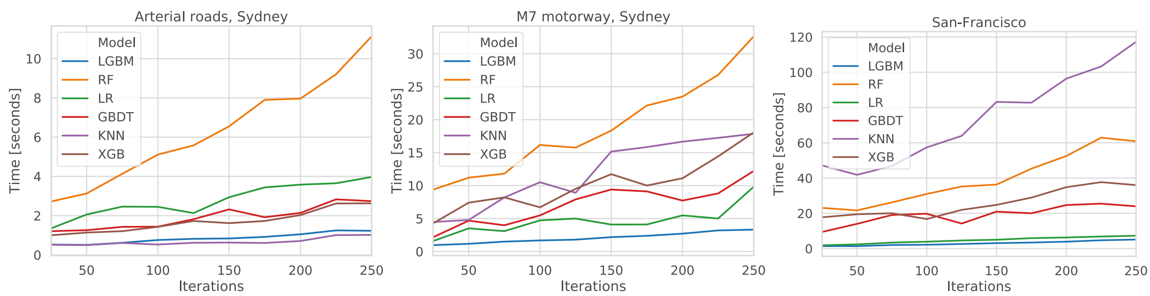



FIGURE 3.19: Performance testing of ML models across three different data sets

Additionally, the tree-based models, such as XGBoost and RandomForest, emerged as the best performing and most robust models in both classification and regression experiments. Moreover, the research highlighted the necessity of modeling short-term and long-term traffic accidents separately, as their distinct characteristics had an adverse effect on model performance when combined.

The proposed IEO-ML approach, which purpose is to combine hyper-pamater space of both model and outlier removal method to enhance prediction accuracy, showcased superior performance compared to baseline ML models in a majority of cases. Furthermore, the analysis of feature importance revealed that various factors, including time, location, type of accident, reporting source, and weather, consistently ranked among the top 10 critical features across all three data sets. By enhancing the precision of these important features and eliminating non-important ones from incident reports, TIMS can significantly enhance the quality of data acquisition.

**Limitations of this study:** One of the biggest challenges when studying the problem of incident duration prediction represents data availability. In most cases, the privacy around traffic incidents represents the main reason why data sets are not released publicly. For example, the two data sets from Australia are private and have only been released for the purpose of this study, whereas only the San Francisco data set is made open publicly. Many other countries around the world have not yet fully released their incident logs, and this represents a challenge for this topic. However, if more incident data logs become available, they can represent a good test best for our approach.

Regarding the model performance, we make the observation that the performance of ML methods is highly affected by the data sets and the used methodology. Our approach shows a better performance for 4 of 6 methods in the case of San-Francisco, but if looked more precisely into details, KNN (where there is no improvement) produces an error that is twice as large as the best performing model (GBDT). The same is for data set A when using the LR method. And, with only GBDT left with no improvement may point to the fact that GBDT is robust to outliers and does not need outlier removal (as observed on all three data sets). As can be seen for the data set A, where MAPE is high (80%), there is a very weak connection between features and the labelled data, and thus the performance for all methods is poor. Therefore, there is not much effect from the outlier removal approach on poor data sets or for methods that are weaker by design.

**Future research** can explore the integration of predicted traffic incident duration into the route planning process using traffic simulation. This implies vehicles considering predicted short-term traffic incident durations and assuming they will be resolved prior to reaching the incident location, thus reducing travel time. The cost of prediction error (e.g. time loss due to encountering the incident in the case of inaccurate duration prediction) and the benefits of estimating traffic accident duration can be evaluated by leveraging simulation models. Additionally, this approach (informing the end-users with predicted incident duration) can be employed not only for offline route planning but also for online route planning, considering departure time. The future study may aim to optimize route planning by incorporating predicted traffic incident duration, providing a more accurate and efficient decision-making process for road users.

Providing additional results for the threshold variation along all data sets such as (Accuracy, Precision and Recall).

Providing additional information with regards to the computational time of various baseline ML models across the three data sets. The findings indicate the RF and kNN seem to be the slowest models

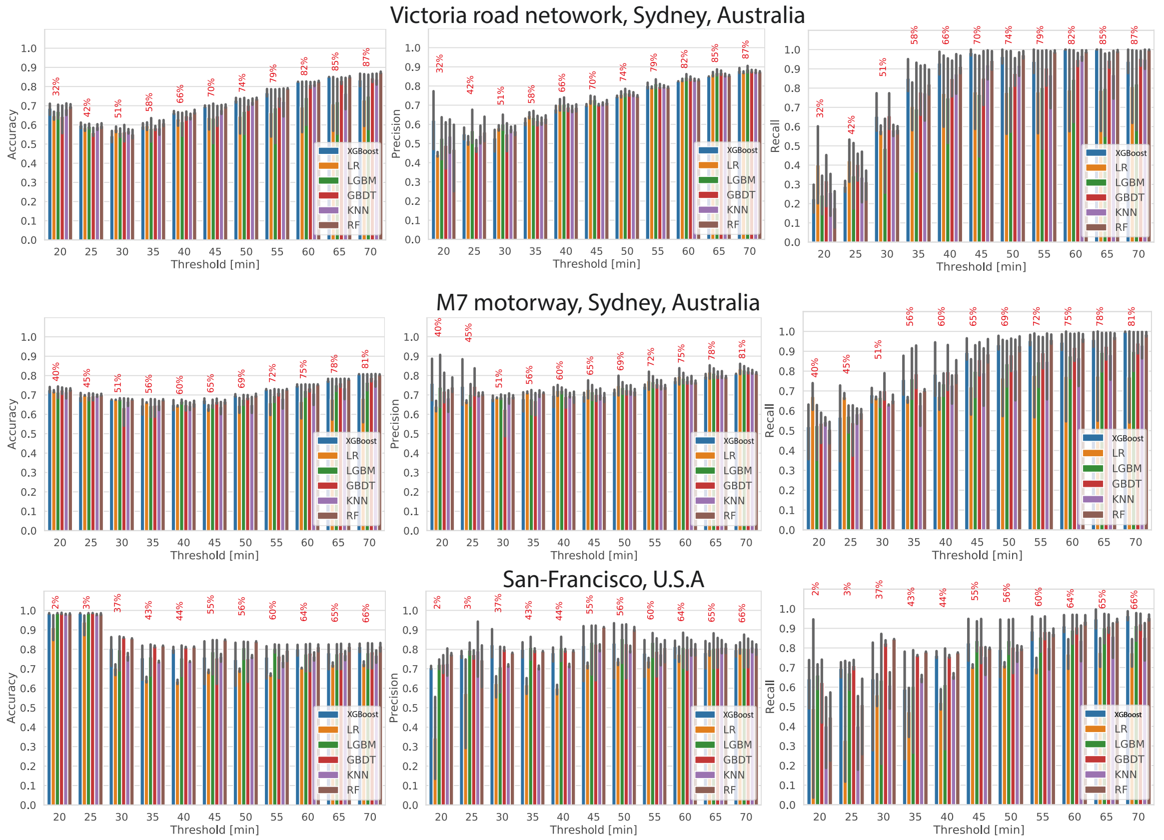



FIGURE 3.20: Binary classification performance using varying incident duration threshold

to train versus LGBM and XGBoost and LR which are faster from a computational time point of view.

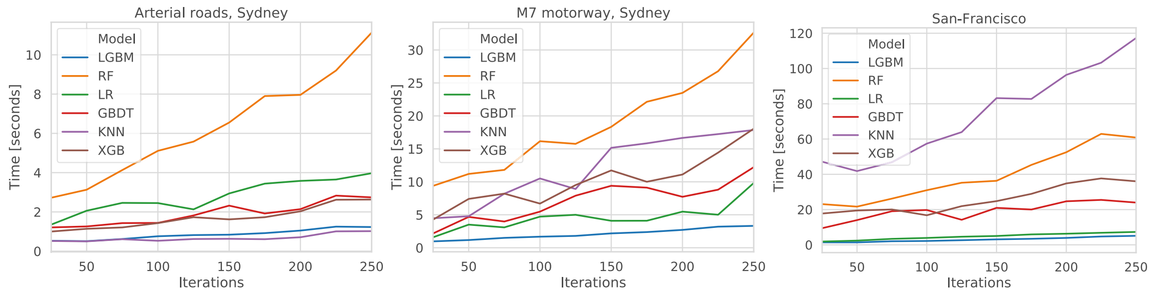



FIGURE 3.21: Performance testing of ML models across three different data sets

# Chapter 4

# Traffic incident duration prediction via a deep learning framework for text description encoding

## 4.1 Introduction

Traffic management centres (TMCs) play a crucial role in managing traffic incidents by storing pertinent information such as textual descriptions and GPS coordinates. However, they often encounter uncertainty in predicting the duration and injury severity and impact of disruptions at the onset of incidents. To address this, historical traffic flows and accident descriptions can be employed to enhance incident duration predictions. In this chapter, we propose an advanced incident duration prediction framework that combines incident reports with additional features related to the incident. This framework utilizes a hybrid machine learning (ML) modeling approach, incorporating deep learning encoding of additional features like textual incident description and historical traffic flow. The utilization of feature encoding is justified due to the high dimensionality of incident descriptions and traffic flow/speed measurements, which can lead to overfitting in traditional ML models. Additionally, the small size of typical incident report datasets may exacerbate overfitting issues. By implementing the proposed framework, TMCs can enhance their incident duration predictions, thereby improving traffic management strategies.

This chapter is organised as follows: Section 4.1 presents the challenges and reviews the related works; Section 4.2 introduces the data sources we have used as well as our traffic flow mapping algorithm for feature construction; Section 4.3 proposes our modelling framework and explains the ML models we have used, the LSTM sentiment encoder for textual incident descriptions, and the ANN encoder for traffic flow speed; Section 4.4 introduces the results before summarising all findings in the Conclusions section.

### 4.1.1 Related Work

There are multiple research papers which use baseline incident reports from TMC with different machine learning models to predict the traffic incident duration [121]. The use of traffic flow and incident description features is found to be rare and mostly specific - topical text modelling [50] for the task of the incident duration detection, modelling or incident impact prediction by using traffic flow readings [61]. And its scarcity is highlighted since it requires the involvement of additional specific models with a feature fusion approach. In other words, traffic flow data is rarely combined with textual incident description and an actual incident reports since it requires a higher system complexity.

But feature combination can be observed in some specific research studies related to the traffic incident impact prediction, which rely heavily on the historical traffic flow data with and without consideration of features that are describing the incident [61]; other works have addressed a similar approach [243], [159]. Also, these works don't focus on the incident duration prediction.

Sometimes, researchers try to apply uniform ML approaches or specific models for all the subtasks. Separate RBM models were applied to different kinds of features and feature fusion representing a uniform application of ML method to different data sets [122]. Also, kNN and Bayesian cost-sensitive networks were combined for the task of the incident duration prediction [116]. But neither of these research studies investigated a deep dive into their model selection.

Since we have the incident description and incident severity values in our incident reports, we can utilise specific models for the task of sentiment classification. Previously, the LSTM architecture has been compared with Support Vector Machines, Artificial Neural Networks, Deep Belief Networks and Latent Dirichlet Association on the task of detection of incidents from social media data [264]. LSTM was also successfully used for stock price prediction [203], making it applicable for modelling of traffic flow/speed time-series data. Despite its superior performance, we need to uplift and bring significant modifications to this architecture. Since we are planning to use encoded time series with machine learning methods, we need a controllable size of the feature vector to simultaneously avoid overfitting and provide enough information for ML methods. This is why we propose to use LSTM coupled with ANN, where the ANN feature vector size and the activation function are varied.

## 4.2 Case study

In this study we assume that textual incident reports as well as historical traffic flows and speed data (including the ones from the moment when an incident happened) are readily available at the moment the incident was reported and sufficient to make the prediction of its duration.

### 4.2.1 Incident description data set and baseline feature set

A Countrywide Traffic Accident Data set (CTADS) has been recently published [165]-[166], which contains about 1.5 million traffic accident records across 49 states of United States of America from February 2016 to December 2020 (version 4). Each incident report contains 47 features describing the traffic accident. The majority of these traffic accidents were recorded in the state of California. The most notable features include: a) Incident Severity (valued from 1 to 4), b) Start and End Time of the incident (from which the traffic incident duration is derivable), c) The road extent affected by the accident, d) textual Incident Description, d) weather and lighting conditions. For the extended description of features please refer to the original paper describing the data set [166]. This data set allows us to use the textual incident description and, hence, apply a sentiment analysis methodology (based on the incident severity) [12]. We further refer to these features as a baseline feature set, excluding the textual incident description.

### 4.2.2 Traffic flow and speed data

To collect the data on traffic flows and speed we rely on the Caltrans Performance Measurement System (PeMS) [33], which provides aggregated 5-minute precision measurements of traffic movements across California. Although there is a lot more data for the Los Angeles area (which may be considered in our future research), we decided to concentrate on the area of the city of San Francisco. We focus on 83 Vehicle Detection Stations (VDS) placed in that area, and we try to manually associate each incident occurred in that area with a VDS in their 500m proximity. VDS in PeMS may have detector failures and incomplete readings, which is common across the data set and should be taken into account. Even though the PeMS data set contains data on reported incidents, we decided to use the descriptions from the Countrywise Traffic Accident Data Set since it provides a high-quality description of each incident (47 features in each incident report) extracted from Bing and MapQuest services.

In total, from 9,275 incidents in the area (extracted from CTADS) we have obtained 1,932 traffic incident reports in a 500m proximity next to VDS stations, which we were able to associate with the

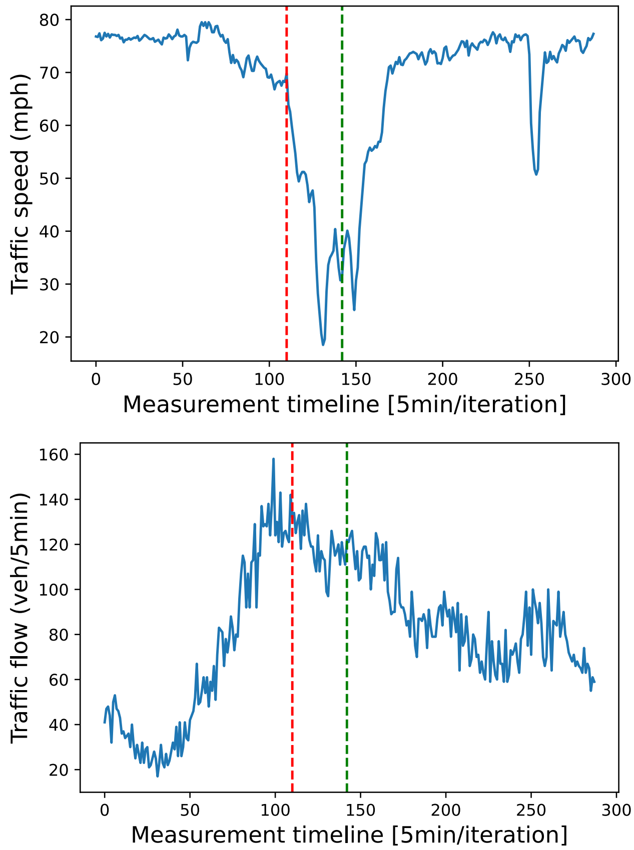



FIGURE 4.1: a) Traffic speed and b) Traffic flow plots for the VDS associated to incident A-4798 (accident on US-101 Southbound with duration of 31 5-minute iterations - actual reported incident clearance time, without considering the incident recovery time). The red line denotes the start of the accident, and the green line the end of the accident. The blue line denotes the speed evolution in the vicinity of the incident location (drops almost to 20km/h) while the flow is still running at high values due to large numbers of vehicles blocked in traffic.

correct (no detector faults) and complete traffic flows and speed readings. Incident to VDS association is necessary since both are represented as points and it is not clear which incident is related to which detector since incidents on different separate roads can be in proximity of one detector (also, since we have a different representation of street names in VDS and the incident data sets). The task of VDS-to-incident assignment can be a topic for additional research, but in this chapter we summarize our extracted mapping strategy as follows. We extract the following speed and flow readings from each VDS station:

1. Speed – Traffic Speed from the 24h leading to the incident occurrence.
2. Flow – Traffic Flow from the 24h leading to the incident occurrence.
3. Speed7 – Traffic Speed on the same weekday, the week before the incident.
4. Flow7 – Traffic Flow on the same weekday, the week before the incident.
5. SD – the vector difference between the traffic speed on the day of the incident and on the same weekday, the week before the incident.
6. FD – the vector difference between the traffic flow on the day of the incident and on the same weekday, the week before the incident.

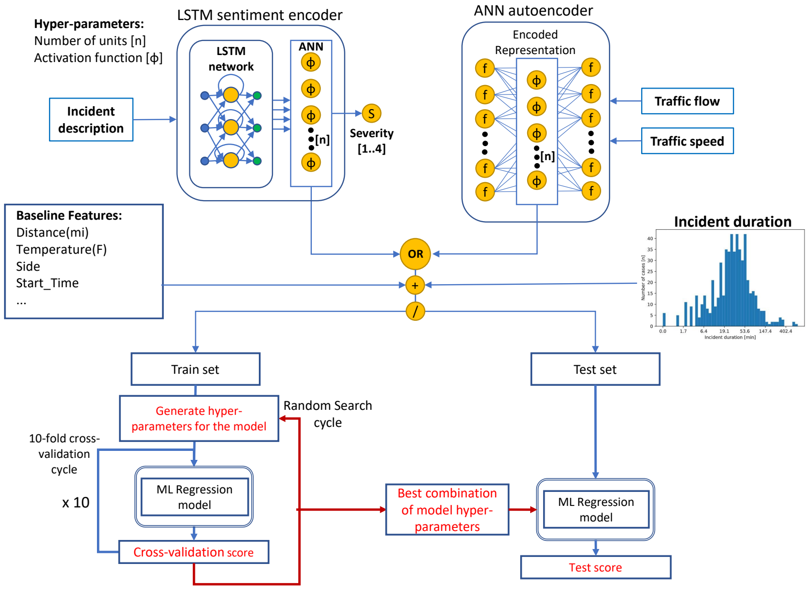



FIGURE 4.2: The structure of the proposed framework

Each of these feature vectors contains 288 values, which correspond to 5-minute readings throughout the day. Since each of these vectors have a high dimensionality, we decide to perform dimensionality reduction via an ANN autoencoder.

The use of dimensionality reduction is justified since a large number of explanatory variables can cause model overfitting [199], [194].

**Regarding the 288 input values on the day of incident:** the traffic data is taken from the time between the incident start and minus 24h before of its occurrence and not during the entire day after the incident has been lodged.

Figure 4.1 shows an example of a traffic speed drop during the incident A-4798. After we have analysed different traffic flows and speed plots we expect that the traffic speed will be the most useful single feature for the task of incident duration prediction as the traffic flow measurement seems to be not affected by the accident (as the majority of vehicles will be waiting for the congestion to clear off the road, and will still be counted as part of the traffic flow). We will also use speed measurements from the weekday, 7 days before the incident in order to obtain the complete picture between what is a regular traffic flow condition versus disrupted traffic condition on the same time and same day of the week. We make the observation that we have also conducted a detailed feature ranking and selection (via SHAP values, forward feature selection, etc.) to several incident data sets which are not presented here due to space limitations.

The point A-4798 point was selected just as an example for a traffic speed drop and its usefulness to the prediction problem; in reality, we have analysed about 100 traffic flow and speed plots before drawing the conclusions (we provide several shapshots of flow and speed reading in the supplementary material). As an observation, by adding severity classification probabilities (from the LSTM-ANN model) to the feature vector for the task of incident duration prediction doesn't seem to be useful since we already included Severity, which is a strong feature.

| |
|---|
| Accident on I-280 Northbound at Exit 57 King St. |
| Right hand shoulder blocked due to accident on I-280 Northbound after Exits 54 54A 54B US-101. |
| Lane blocked due to accident on US-101 Presidio Pkwy Southbound at Exit 438 CA-1. |
| Accident on I-80 Westbound at Exits 1 1C / Bryant St / 8th St. |
| Second lane blocked due to accident on I-80 Eastbound at Exits 2B 2C Harrison St. |
| Lane blocked due to accident on US-101 Golden Gate Brg Southbound at Exit 439 Transit Transfer Facility. |
| Right hand shoulder blocked due to accident on I-280 Northbound at Exit 52 San Jose Ave. |
| Right hand shoulder blocked due to accident on US-101 Southbound at Exits 429B 429C Bay Shore Blvd. |
| Lane blocked on exit ramp due to accident on I-280 Northbound at Exit 55 Cesar Chavez. |
| Right hand shoulder blocked due to accident on I-280 Northbound at Ocean Ave. |

TABLE 4.1: Example of the Incident Description values

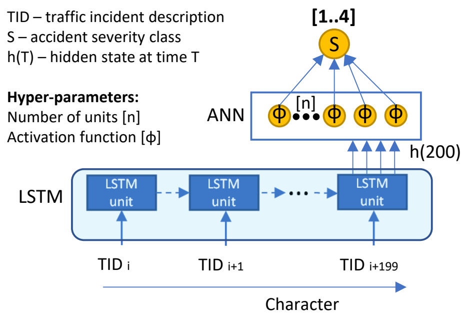



FIGURE 4.3: LSTM sentiment encoder structure.

## 4.3 Methodology

The overall goal of this study is to predict the duration of incidents (see Figure 4.2). In order to accomplish this, the data is utilized as combination of the baseline feature set and either the encoded textual description or the encoded traffic flow/speed values. It is worth noting that the encoder components of both the LSTM-ANN network and the ANN autoencoder have hyper-parameters, such as the number of units and the activation functions, which are fine-tuned to achieve optimal encoding. Once the encoded representations associated with each incident are obtained, an exhaustive search is conducted to identify the optimal hyper-parameters for each ML regression model at each case of the encoded representation. This approach allows for the adaptation of the encoder and ML models to the data, yielding the best cross-validation results.

### 4.3.1 LSTM-ANN for the textual incident description encoding

Textual Incident Description in the CTADS data set describes type of disruption caused by the incident and/or location (Table 4.1).

To perform the encoding of the textual description of the incident we use a combination of character-level LSTM and ANN for the sentiment analysis (Figure 4.3). We use the textual incident description from all the available traffic incident reports for the San Francisco area (9,275 incident records). Firstly, we set the target variable for the LSTM classification model as the incident severity (values 1 to 4).

Secondly, we use the encoded representation of the textual description extracted from the LSTM sentiment classification model to use it as additional features for the task of incident duration prediction.

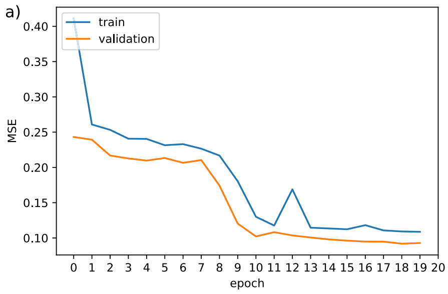



FIGURE 4.4: Example of LSTM network training results using 12 units, a ReLU activation function, 10 epochs, 80 hidden units. a) Train-validation score over 20 epochs

The incident description text is only provided at the beginning of the incident reporting timeline, and no temporal evolution is found across multiple countries for which we analysed the incident logs in our previous work [75].

Each textual description is formed into repeated strings up to 200 characters in length and each character in that string is then encoded by using one-hot encoding.

In order to showcase the importance of the textual incident description for the tasks of incident duration prediction and incident severity classification, we perform a word importance analysis using the LIME method (provided in the supplementary material **appendix**). We further train a an LSTM model with 80-units hidden state vector. We use the encoding of the incident description by using different numbers of neurons and different activation functions. An example of training results for one of the variants is shown on Figure 4.4. Traffic incidents descriptions were used to predict the incident severity. The data set was split into train, validation and test sets by proportion 70:20:10. Training results show that the LSTM sentiment encoder needs at least 15 epochs to converge, so we decided to train each variant of the LSTM sentiment encoder for 15 epochs. We use Root Mean Squared Error (RMSE) as the loss function.

#### The use of MSE versus cross-entropy

MSE is a legitimate metric for the classification when the target feature is represented as an ordered variable [63] in which MSE is preferred instead of the Cross-Entropy (CE) loss in order to reduce the model complexity and the probability of over-fitting. In our research we determined that CE required Nx5 sized matrix for the intermediate feature vector to the target value classification, while the MSE solution requires only Nx1 matrix, where N is size of the intermediate feature vector). MSE loss is also superior to CE loss for class-imbalanced datasets [106] and our incident severity feature distribution poses an imbalanced classification problem.

### 4.3.2 Artificial Neural Network Encoder for the traffic flow/speed encoding

In this study, we explore the use of Artificial Neural Network (ANN) Autoencoder [114]r as a data processing tool for handling incident-related traffic speed and traffic flow features. By performing hyper-optimization of various parameters such as different activation functions and a varying number of neurons in the bottleneck layer, we generate encoded speed/flow reading representations associated with each accident. The flow and speed values are normalized to the maximum observed traffic speed

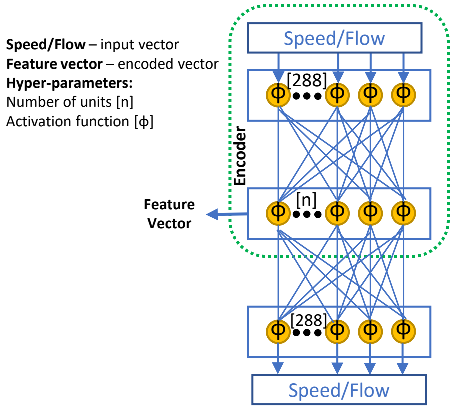



FIGURE 4.5: The structure of the ANN autoencoder

and flow within the data set. To enhance the encoding model's performance, we utilize all available time series data for training the Autoencoder model. We combine the normalized flow and speed data sets, enabling the model to capture the overall time series rather than focusing solely on individual aspects of speed and flow levels at specific station. This way we obtain a general encoding model for traffic speed/flow time-series. It is worth noting that using ANN for auto-encoding provides valuable benefits, such as dimensionality reduction and improvement in model robsutnes, particularly when dealing with extreme outliers. Consequently, the outputs from the ANN's autoencoder bottleneck layer are extracted and utilized as features (encoded representations of traffic flow/speed readings associated with the accident) in the Machine Learning (ML) models employed in this study.

The following activation units were used in the bottleneck layers of the ANN autoencoder and the LSTM sentiment encoder: a) the Rectified Linear Unit (ReLU) [5] which is a piecewise linear function (output values are $[0;+\infty]$ b) the Exponential Linear Unit (ELU) [224], which was developed to reduce bias shift (which leads to weight oscillations) c) the Tanh - a hyperbolic tan function which has the property of equalizing training over layers [104]; its output can take values in the interval $(-1;+1)$ d) the Sigmoid activation function which output can take values in the interval $(0;1)$.

### 4.3.3 Baseline Machine Learning model selection

When all encoding has been finalised, we first use the following ML regression models as a baseline to perform the incident duration prediction:

a) gradient boosting decision trees - GBDT [250] which rely on training a sequence of models, where each model is added consequently to reduce the residuals of prior models; b) extreme gradient decision trees - XGBoost [36] which rely on an exhaustive search of split values by enumerating over all the possible splits on all the features and contains a regularisation parameter in the objective function; c) random forests - RF [27] which applies a bootstrap-aggregation (bagging, which consists of training models on randomly selected subsets of data) and uses the average (or majority of votes) of multiple decision trees in order to reduce the sensitivity of a single tree model to noise in the data d) Support Vector Regression (SVR) machines [56] which are characterized by the use of kernels and symmetrical loss function (equal penalization of high and low errors), e) Decision Trees (DT) regression models [25] which rely on the repetitive process of splitting and generates a set of rules which can be used for the value prediction, f) Linear Regression (for which we use standard Ordinary

Least Squares optimisation) which represents the relation between features and the target variable as a linear equation targeting to minimize the residual sum of squares between the actual and the predicted values of the target variable.

#### Model performance evaluation

To evaluate the regression models on the task of the incident duration prediction we use the mean absolute percentage error and the root mean squared error defined as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2} \tag{4.2}$$

where $A_t$ are the actual values and $F_t$ - the predicted values, $n$ - the number of samples. We do make the observation that other performance metrics have been obtained (MAE, SMAPE), but given the current page limitations, we focus on MAPE, RMSE results only.

#### Hyper-parameter tuning for the proposed regression model

We use 10-fold cross-validation to overcome the over-fitting problem [64] and to assess the generalization performance of the ML models. In each scenario, the data set is partitioned into 10 folds. The ML regression model is trained on 9 folds to make prediction on the remaining fold. The procedure is then repeated 10 times and the accuracy results are averaged across several repetitions.

### 4.3.4 MAPE versus RMSE comparison and their non-linear relationship

There is a non-linear relationship between MAPE and RMSE when performing regression, which can be verified by using different regression data sets. We tested this hypothesis on the Concrete Compressive Strength (CCS) Data Set from UCI Machine Learning Repository by using 1000 evaluations of random 9:1 train-test splits using Random Forest evaluated against MAPE and RMSE. Fig. 4.6a) presents the MAPE versus RMSE plot in which we observe that, the same MAPE result (e.g. 12%) may be attributed to multiple RMSE results (e.g. from 3.5 to 6.5). A similar situation observed for 45% of MAPE on CTADS using Random Forest (see Fig. 4.6)b). Therefore, the occurrence of a higher RMSE error when MAPE becomes lower (as in our paper) and vice-versa is a correct result. MAPE vs RMSE compared between 100-units random vectors with 1-10 value interval using 10,000 evaluations (see Fig. 4.6)c). As can be seen from all three sub-plots, the decrease in MAPE doesn't necessarily mean a decrease in RMSE. For our study we focused on discussing the MAPE metric, which is widely used in the literature on the topic of incident duration prediction since its intuitive meaning (e.g. a 30% MAPE means a 30% deviation of prediction from the actual incident duration) and a less inclination to high errors from outliers such as the case of RMSE. The results are part of the optimal Pareto Front [marked in orange] which showcases that our proposed method can obtain the set of optimal feature combination scenarios rather than only one winning scenario. To conclude, despite an assumption on linear dependence between the RMSE and the MAPE metrics (assumption that both metrics should be

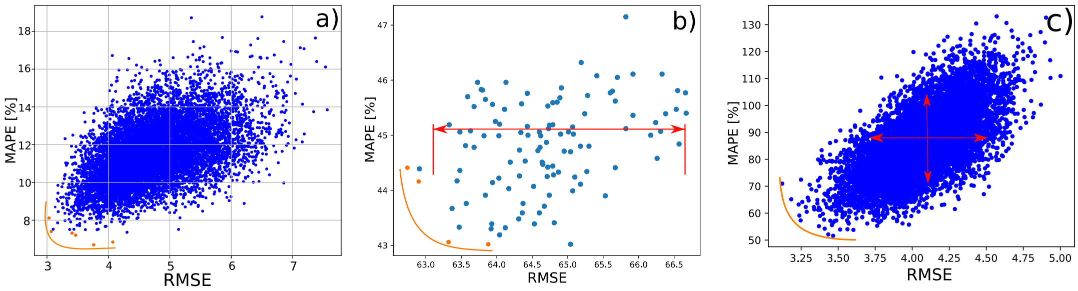



FIGURE 4.6: RMSE vs MAPE results for a) CCS data set, b) CTADS - incident duration c) Random vectors

reduced in an efficient solution), both in our incident duration case and the CCS data set, we observe a Pareto front of efficient solutions (no solution is sufficient in both metrics, making our results stand strong).

### 4.3.5 Comparison to other baselines

It is hard to perform a comparison between different studies on the traffic incident duration prediction since different data sets are used for research purposes [121]. Majority of these data sets are also private and rely on different sets of features. CTADS data set appeared only recently (2019) and there is still no uniform convention on which data subset to use as a baseline, since the data set is big (1.5 million records) and heterogeneous (it includes reports from all kinds of traffic networks around United States). Indeed, in our previous work we have compared various ML-DL approaches against logs from Australia and USA, which can be used as extended results.

## 4.4 RESULTS

### 4.4.1 Best model selection

First, we try to find the three best models which show high performance of the baseline feature set consisting of traffic accident reports for which we have available traffic flow counter data. We do so by performing a cross-validation as described in 4.3.3 and a performance evaluation as detailed in 4.3.3. Figure 4.7 shows the average MAPE score for the 10-fold cross-validation obtained across several ML models such as Random Forests (RF), GBDT, XGBoost, kNN, Decision Trees (DTs), Linear Regression (LR) and Support Vector Regression (SVR). Given that the majority of traffic incident duration prediction methods published previously have reported a MAPE score below 50% [121], we select RandomForest, GBDT and XGBoost as the best performing models as their MAPE score falls below 46%. Next, we evaluate these three models against the baseline feature set when we apply our novel modelling approach as previously explained in sections 4.3.1-4.3.2: traffic flow, speed via ANN autoencoding and textual incident description via LSTM sentiment encoding.

There are in total 140 scenarios describing combinations of additional features [7 speed/flow/text features x 5 unit count x 4 activation functions] for each of the top three ML models. Given the restricted space allocation for this article, in Tables 4.2-4.4 we present only the top 8 best scenario results ranked against MAPE for each ML model.

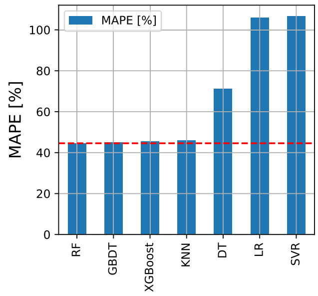



FIGURE 4.7: Regression results for baseline feature set across different ML models.

| AdditionData | units | activation | MAPE | RMSE |
|---|---|---|---|---|
| baseline | | | 44.99 | 58.4 |
| LSTM-sent | 12 | relu | 41.89 | 65.03 |
| Flow7 | 8 | tanh | 41.92 | 64.61 |
| LSTM-sent | 16 | tanh | 42.05 | 65.04 |
| Speed7 | 16 | tanh | 42.28 | 63.97 |
| LSTM-sent | 8 | tanh | 42.43 | 64.13 |
| LSTM-sent | 16 | relu | 42.56 | 65.82 |
| Flow | 16 | sigmoid | 42.57 | 64.53 |
| Speed7 | 2 | sigmoid | 42.59 | 64.76 |

TABLE 4.2: Top 8 best scenario results for GBDT-enabled framework

Findings reveal that the encoded textual description is among the top 3 configurations for every regression model as seen from Tables 4.2-4.4. Models also demonstrate a preference for the way of encoding: 1) the Tanh activation function forms a majority in the top results for GBDT both for encoding the incident description and flow/speed features (Table 4.2), 2) the ReLU activation function forms a majority in the case of XGBoost (Table 4.4). This observation can point on a preference in the way of encoding features when using specific regression models. The best performing model among the top three finalists, when using all additional features seems to be GBDT: the best results are obtained when encoding the traffic incident description and when using the traffic flow 7 days before the incident with 12 units and the ReLU activation function [$MAPE$ = 41.89%, Table 4.2] (therefore including the information on the regular traffic flow profile on the same weekday, together with the incident report proves important for the task of incident duration prediction).

Other models show a higher MAPE or RMSE results for the incident duration prediction (see RF

| AdditionData | units | activation | MAPE | RMSE |
|---|---|---|---|---|
| baseline | | | 44.58 | 57.6 |
| Flow | 4 | tanh | 43.02 | 63.88 |
| LSTM-sent | 4 | elu | 43.02 | 65.04 |
| LSTM-sent | 12 | relu | 43.06 | 63.32 |
| Flow7 | 16 | sigmoid | 43.19 | 64.04 |
| Flow | 16 | elu | 43.30 | 63.92 |
| FD | 4 | elu | 43.32 | 64.12 |
| Flow7 | 16 | tanh | 43.33 | 63.47 |
| FD | 4 | sigmoid | 43.39 | 64.53 |

TABLE 4.3: Top 8 best results for RF

| AdditionData | units | activation | MAPE | RMSE |
|---|---|---|---|---|
| baseline | | | 45.44 | 63.41 |
| Flow | 8 | relu | 43.44 | 69.93 |
| LSTM-sent | 4 | tanh | 43.58 | 71.03 |
| Speed7 | 16 | tanh | 43.63 | 71.62 |
| SD | 4 | relu | 43.73 | 70.58 |
| Speed7 | 16 | relu | 43.80 | 71.92 |
| LSTM-sent | 16 | elu | 43.81 | 70.45 |
| LSTM-sent | 8 | relu | 43.82 | 72.19 |
| Flow7 | 2 | relu | 43.85 | 72.94 |

TABLE 4.4: Top 8 best results for XGBoost

enabled results in Table 4.3 with lowest $MAPE = 43.2\%$ for a combination of baseline, regular traffic flow, 4 layer units and a tanh activation function); similar findings appear for XGBoost-enabled results in Table 4.4 with the lowest $MAPE = 43.44\%$, when using again the regular flow features, 8 layer units and ReLU activation function. This experiment shows that an accurate incident duration prediction immediately after the event has occurred is possible, leveraging the incident description and the measured traffic flow on the day of accident, which may prove very useful for TMCs to incorporate directly in their incident management platforms. Lower MAPE does not necessarily mean lower RMSE as seen from the baseline and additional data scenarios, but the LSTM sentiment encoding seems to be the approach that obtains the best RMSE score (64.13) when combined indeed with other variations of the activation function and number of hidden units (as shown in Table 4.2).

### 4.4.2 Parallel coordinates for scenario setup

To supplement the findings, we also provide a parallel categories representation of all the 140 scenarios for the GBDT model in Figure 4.8, which highlights the best combination of activation functions that seem to be working best alongside the character-level LSTM sentiment encoder of traffic flow incident textual description and speed information - mostly from previous daily speed profiling using historical data. The worst results seem to be the ones obtained when using only the speed or flow difference vector alongside the baseline incident features.

Encoding using Sigmoid and Tanh activation units on average performs best, probably because of the limited value range: Tanh an Sigmoid allow encoded representations to take values in ranges $[-1; +1]$ and $(0; 1)$ correspondingly, ReLU and ELU can take unlimited positive values. These results indicate which value ranges work best for encoded representation.

Comparison of MSE and CE implementations of LSTM severity classification metric for the purpose of obtaining feature vector representation of Incident Description (see Fig. 4.8) shows that a sentiment classifier with Cross-entropy (lstmsentCE) as a target metric with one-hot encoded severity values is more efficient (left column attributed to LSTMsentCE shows more blue rows associated with low metric values than LSTMsentMSE - sentiment encoder which predicts severity as a single value). Comparison between the number of units shows preference for 4 units since the presence of the lowest error and absence of the highest error rows. Among the activation units, the Sigmoid is the best performer showing more low error results than other units. This scenario is to show how feature vector representing incident description may may be efficiently encoded to be used with conventional

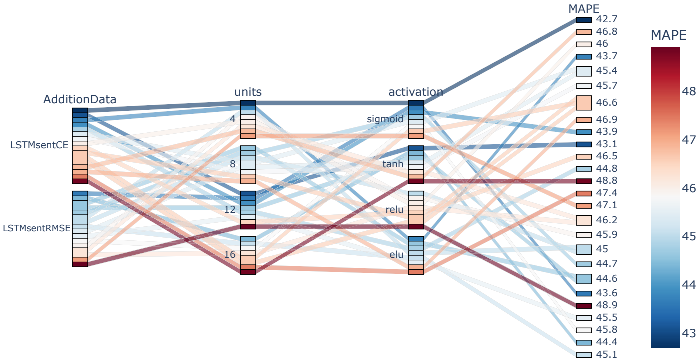



FIGURE 4.8: Parallel categories representation for all regression scenarios with GDBT.

GBDT machine learning method: using cross-entropy for the severity classification, using 4 units and the Sigmoid as activation function.

## 4.5 Conclusion

In this chapter, we propose a novel framework that integrates machine learning techniques with traffic flow and description features encoded through deep learning methods for the purpose of predicting incident duration. Our approach demonstrates a stable improvement across all regression models used. The results highlight the importance of employing specific deep learning encoding approaches in regression models, which enhances model performance by leveraging past historical traffic speed/flow information and textual incident descriptions. Efficiently encoding incident-related features is the initial step in modeling the impact of traffic incidents on traffic flow. Our future work focuses on exploring the spatial and temporal dynamics analysis of incident impact. However, our research has certain limitations. Firstly, we only utilized data from San Francisco as our study area, while data availability for traffic accidents and flow extends to the entire area of United States of America. Secondly, we considered traffic speed and flow only day/week before the incident occurred, but collecting data on traffic counts over longer periods has the potential to construct more accurate traffic speed and flow profiles, further improving predictions. The societal impact of our research lies in the methodology for improved data availability for Traffic Management Centers (TMC) in predicting incident durations. This can assist TMCs in effectively managing incidents and traffic, including announcing predicted incident duration times to response teams and road users and allocating appropriate resources. Consequently, timely and accurate incident response can reduce congestion time for drivers and decrease the overall impact of incidents on traffic flow. The code for the paper can be found: `https://github.com/Future-Mobility-Lab/TIDP_2022`.

### 4.5.1 Word importance for severity classification

To estimate word importance in the Incident Description feature, word count matrix has been transformed to a normalized TF-IDF representation (term frequency–inverse document frequency) [48]. N-gram value range is (1,2). Then linear dimensionality reduction has been performed using truncated singular value decomposition to 50 componenets for 7 iterations. Then we used GBDT classification model to fit incident severity and three quantiled groups (ratio 33%:33%:33% to represent equaly sized

| Severity group=0 | | Severity group=1 | | Severity group=2 | | Severity group=3 | |
|---|---|---|---|---|---|---|---|
| Weight? | Feature | Weight? | Feature | Weight? | Feature | Weight? | Feature |
| +0.644 | lanes blocked | +1.559 | chavez | +2.805 | 280 | +0.982 | lanes blocked |
| +0.345 | two lanes | +1.190 | cesar | +1.909 | 280 northbound | +0.427 | two lanes |
| +0.231 | blocked due | +0.973 | <BIAS> | +0.828 | blocked | +0.365 | blocked due |
| +0.034 | due to | +0.894 | st | +0.740 | to accident | +0.174 | on i |
| +0.007 | to accident | +0.475 | i | +0.736 | accident | +0.110 | due to |
| -0.407 | lanes | +0.467 | to | +0.721 | i 280 | +0.008 | to accident |
| -0.620 | blocked | +0.465 | on | +0.697 | chavez st | -0.076 | cesar chavez |
| -0.689 | i | +0.351 | at | +0.677 | accident on | -0.127 | lanes |
| -0.704 | <BIAS> | +0.309 | northbound | +0.448 | two lanes | -0.546 | blocked |
| -0.748 | to | +0.307 | cesar chavez | +0.375 | lanes blocked | -0.621 | i |
| -0.760 | st | +0.289 | two | +0.336 | at cesar | -0.666 | on |
| -0.769 | accident | +0.153 | due | +0.194 | due to | -0.672 | <BIAS> |
| -0.793 | two | -0.031 | northbound at | +0.187 | blocked due | -0.678 | to |
| -0.797 | due | -0.101 | blocked due | +0.138 | lanes | -0.710 | at |
| -0.800 | on | -0.125 | due to | +0.070 | northbound at | -0.762 | two |
| -0.818 | at | -0.310 | at cesar | +0.022 | on i | -0.773 | due |
| -0.869 | northbound | -0.330 | two lanes | -0.160 | northbound | -0.918 | accident |
| -0.924 | 280 | -0.372 | lanes | -0.208 | due | -0.953 | st |
| -0.979 | chavez | -0.466 | lanes blocked | -0.354 | cesar chavez | -0.961 | cesar |
| -1.149 | cesar | -0.647 | accident on | -0.358 | at | -0.997 | chavez |
| | | -0.684 | i 280 | -0.369 | two | -1.116 | 280 |
| | | -0.692 | chavez st | -0.498 | on | -1.140 | northbound |
| | | -0.711 | accident | -0.509 | to | | |
| | | -0.728 | to accident | -0.534 | i | | |
| | | -0.913 | blocked | -0.891 | st | | |
| | | -1.993 | 280 northbound | -0.994 | <BIAS> | | |
| | | -2.728 | 280 | -1.110 | cesar | | |
| | | | | -1.479 | chavez | | |

FIGURE 4.9: Word importance estimation using LIME method for incident severity groups

groups with duration intervals 0-29min, 30-71min and 72-2750min) of the incident duration. Classifer predictions were then analyzed for feature importance using Local Interpretable Model-agnostic Explanations (LIME) method [148], where every feature represents 1 word or 2 word combination presence in the incident description. The technique has been explored in other studies to analyse word importance in traffic crash reports [14].

One or more combinations of word in the description can contribute to the incident being classifed into one of severity groups (Fig. 4.9) - presence of "lanes blocked" and "two lanes blocked" has the highest contribution to the incident being classifed into highest (3) or lowest (0) severity group. Severity 1 or 2 is more related to the actual location, which represented as word describing Cesar Chavez St and I-280 Interstate Highway. High positive and opposite high negative contribution of words towards severity group observed for severity groups 1 and 2, where "280" and "chavez" have high opposite contributions, making this groups easily separable. When we perform classification towards equaly sized incident duration groups, "lanes blocked" has the highest positive contribution of the incident to be classified into low duration group. If accident happens on Cesar Chavez St, it can be easily classified into low duration group signifying importance of location for the task of incident duration prediction. High negative contribution of "lanes blocked" observed for duration group 1 with the highest contribution of "280" word meaning that incident appears on I-280 Interstate Highway.

### 4.5.2 Traffic flow and traffic speed on the day of the incident

The following plots represent recorded traffic speed and flow on the day of the incident and week before in 500m proximity of the incident along the road (see Fig. 4.11 and 4.12). Reports in CTADS data set indicate that the highest impact of traffic incident is attributed to significant decrease in traffic speed, while traffic flow stays the least affected by disruption.

| duration group=0 | | duration group=1 | | duration group=2 | |
|---|---|---|---|---|---|
| **Weight?** | **Feature** | **Weight?** | **Feature** | **Weight?** | **Feature** |
| +1.307 | lanes blocked | +0.548 | 280 | +0.389 | chavez st |
| +0.653 | two lanes | +0.444 | northbound | +0.256 | 280 northbound |
| +0.461 | blocked due | +0.357 | blocked | +0.149 | blocked due |
| +0.422 | lanes | +0.218 | chavez | +0.132 | northbound at |
| +0.326 | to accident | +0.214 | st | +0.092 | at cesar |
| +0.324 | on i | +0.213 | accident | +0.075 | cesar chavez |
| +0.255 | at cesar | +0.182 | cesar chavez | +0.068 | to accident |
| +0.230 | due to | +0.095 | two lanes | +0.062 | cesar |
| +0.216 | northbound at | +0.091 | cesar | +0.017 | to |
| +0.211 | chavez st | +0.050 | due to | -0.036 | <BIAS> |
| +0.177 | accident on | +0.039 | i 280 | -0.057 | lanes blocked |
| +0.026 | i 280 | +0.034 | lanes | -0.080 | due |
| -0.123 | st | +0.029 | 280 northbound | -0.088 | at |
| -0.153 | cesar chavez | -0.013 | on | -0.133 | two lanes |
| -0.232 | blocked | -0.030 | <BIAS> | -0.232 | accident |
| -0.232 | 280 northbound | -0.037 | two | -0.264 | chavez |
| -0.275 | i | -0.069 | to accident | -0.383 | st |
| -0.290 | at | -0.072 | northbound at | -0.502 | northbound |
| -0.348 | on | -0.077 | i | -0.580 | lanes |
| -0.405 | chavez | -0.129 | blocked due | -0.594 | 280 |
| -0.437 | northbound | -0.204 | chavez st | -0.633 | blocked |
| -0.439 | 280 | -0.655 | lanes blocked | | |
| -0.440 | to | | | | |
| -0.449 | due | | | | |
| -0.485 | accident | | | | |
| -0.544 | two | | | | |
| -0.724 | cesar | | | | |
| -0.918 | <BIAS> | | | | |

FIGURE 4.10: Word importance estimation using LIME method for incident duration groups

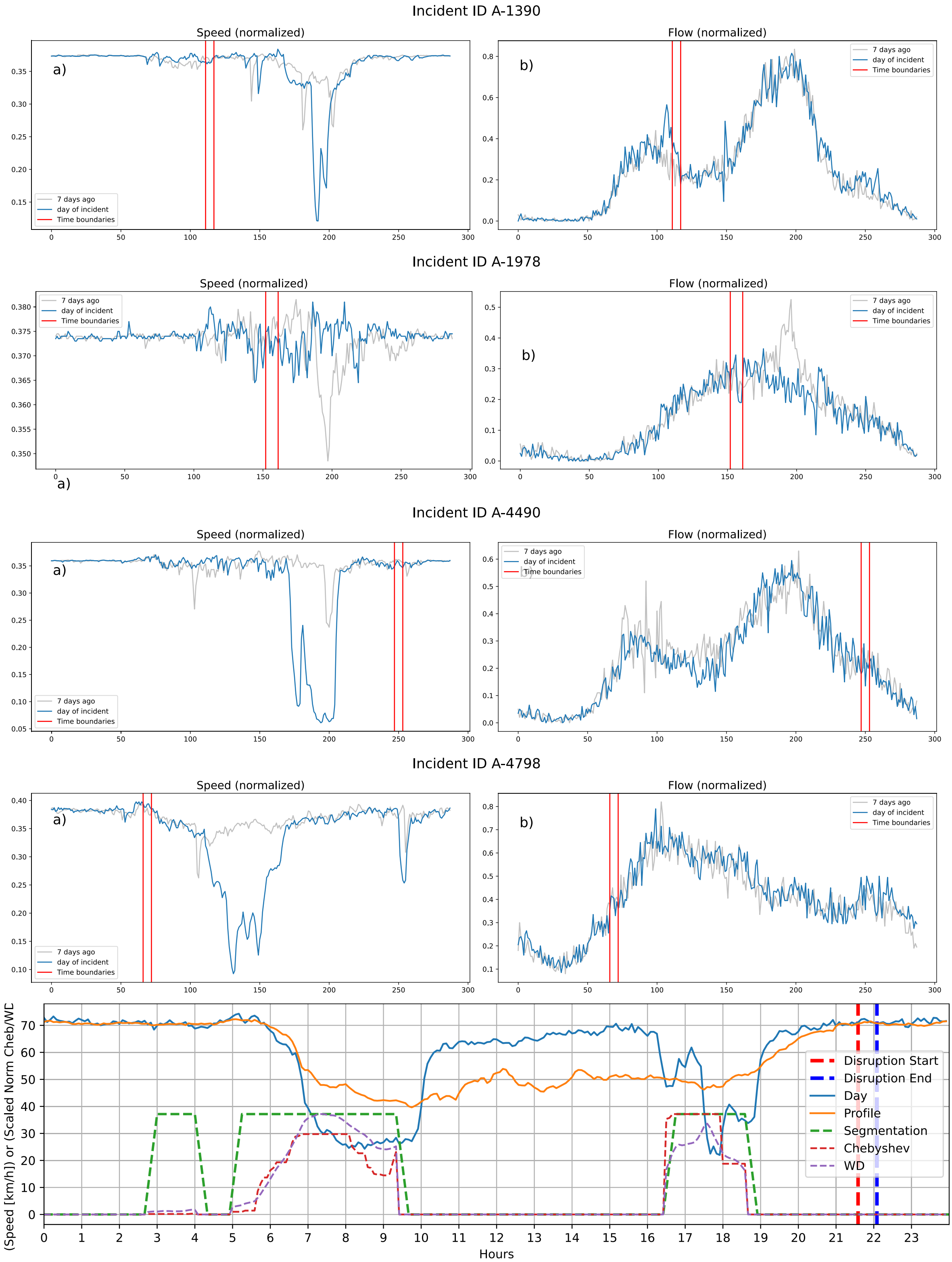



FIGURE 4.11: Traffic speed and flow during the day of the incident. Part #1

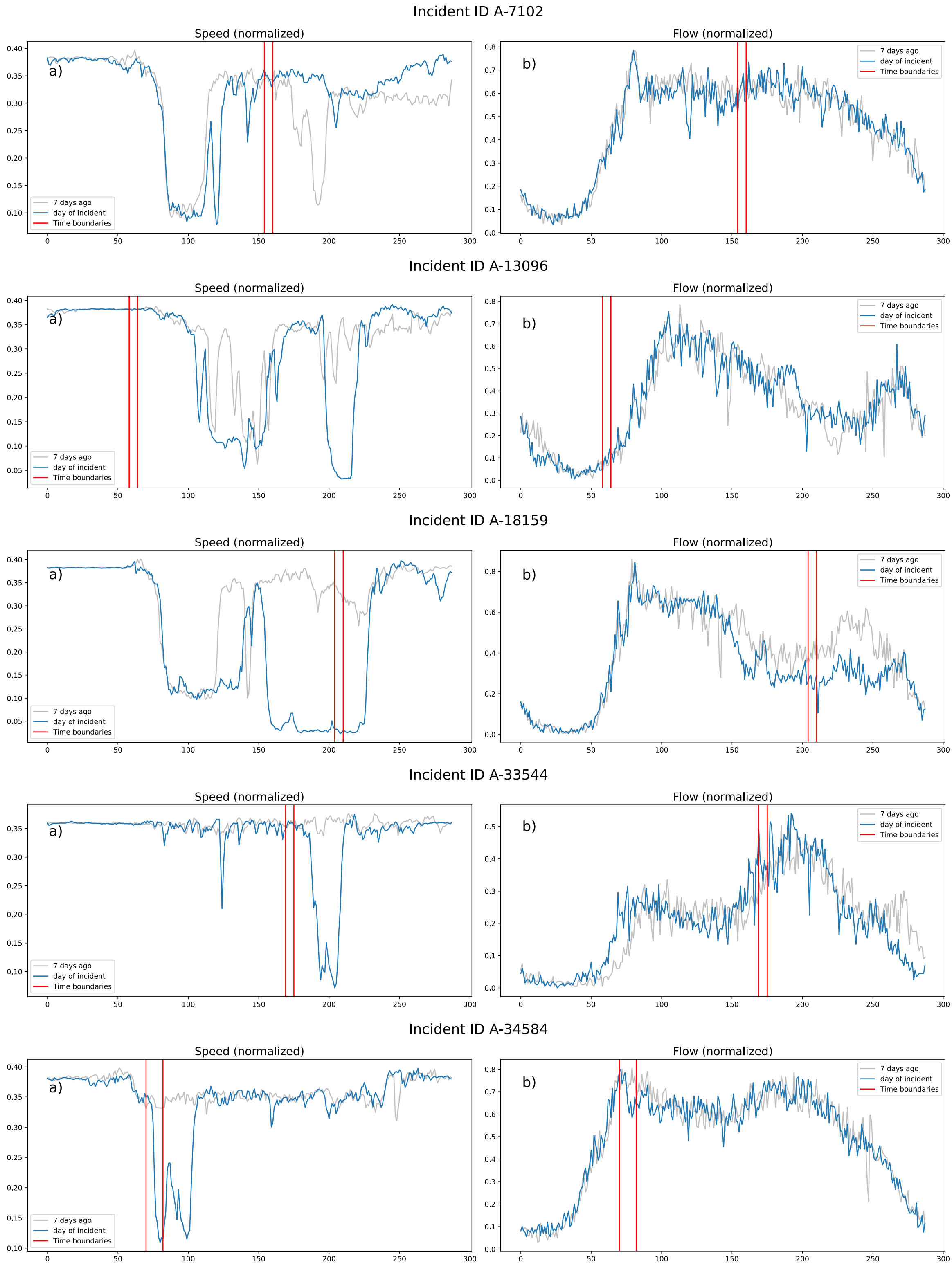



FIGURE 4.12: Traffic speed and flow during the day of the incident. Part #2

## Chapter 5

# Spatial-Temporal Traffic Accident Risk Forecasting using Contextual Vision Transformers with Static Map Generation and Coarse-Fine-Coarse Transformers

## 5.1 Introduction

Traffic accidents pose a significant impact on global health and economics, with an upward trend in incidents particularly notable in developing countries [175]. The issue persists with over 5 million accidents annually in the United States alone [4], and 1.35 million fatalities worldwide in 2016 [176].

Traditionally, traffic accident risk forecasting is viewed as a time-series prediction task, requiring separate models to handle spatial and temporal aspects. Despite initial Deep Learning attempts to predict traffic accident risks, some studies didn't consider traffic flow or time-related factors [35]. Subsequent research [192], [267], [268], [259], [237], [240] offered enhancements by incorporating additional contextual data like air quality, weather, and the condition of roads.

This research on traffic accident risk prediction can offer several benefits to different stakeholders involved in urban planning and traffic management:

**Traffic Management Authorities**: The predictive insights offered by this research can aid traffic management authorities in deploying resources effectively. If certain areas are predicted to have a high risk of accidents at particular times, they can arrange for additional traffic police deployment or emergency medical services in those areas in advance. This can lead to faster response times and potentially save lives in the event of accidents.

**Emergency Services**: Predicting high-risk scenarios and their potential locations can significantly enhance emergency services' readiness. Knowing when and where accidents are likely to happen means ambulances, fire services, and police can be strategically located to respond rapidly when needed.

In recent years, the field of traffic accident research has witnessed a significant rise in the utilization of computational methodologies. Various challenging issues have been explored, including the prediction of traffic accident duration [121], the detection of accidents as they occur [185], and the estimation of severity. More recently, the field has witnessed the development of spatial-temporal modeling techniques, which enable accident risk prediction through the analysis of high-dimensional spatial, semantic, and temporal datasets [240]. The application of these advanced methodologies has greatly enhanced the automated analysis of traffic data, especially considering the increasing availability of publicly-accessible datasets. The prediction of traffic accident risk offers valuable insights for several purposes: firstly, it allows for the identification of high-risk areas within the traffic network, thus aiding decision-making processes within traffic management authorities. Secondly, it facilitates the allocation of resources and the evaluation of road design to minimize the occurrence of future accidents. Thirdly, it enables the timely anticipation of high-risk situations on the road. Lastly, it facilitates the implementation of traffic management strategies aimed at reducing risk.'

This paper introduces a series of enhancements to our previously proposed novel approach that relies on Vision Transformers [55], [247] to forecast traffic accident risk. Our approach leverages the spatio-temporal nature of the problem and the influence of contextual information in a unified end-to-end model. Specifically, we introduce the Coarse-Fine-Coarse Transformer architecture and static map incorporation into ViT architecture.

The code for the paper can be found by the following link: `https://github.com/Future-Mobility-Lab/ViT-traffic-accident-risk`

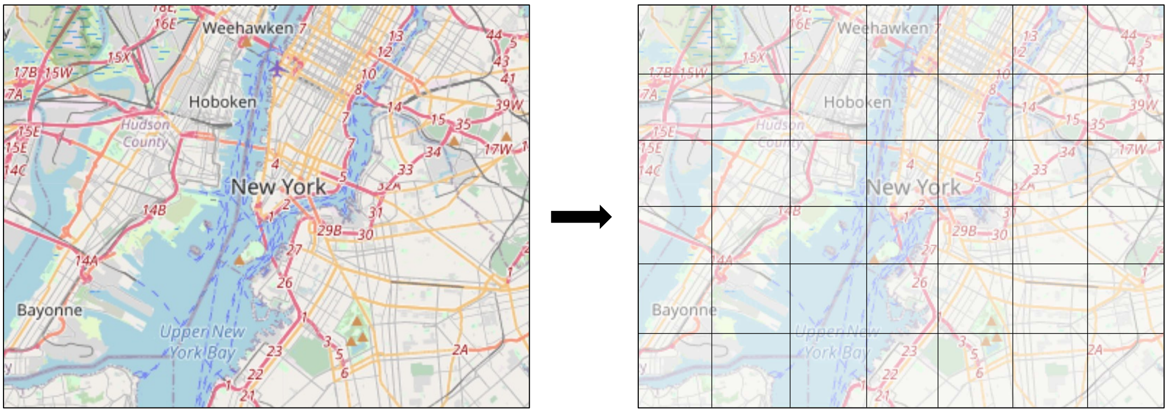


FIGURE 5.1: City grid representation

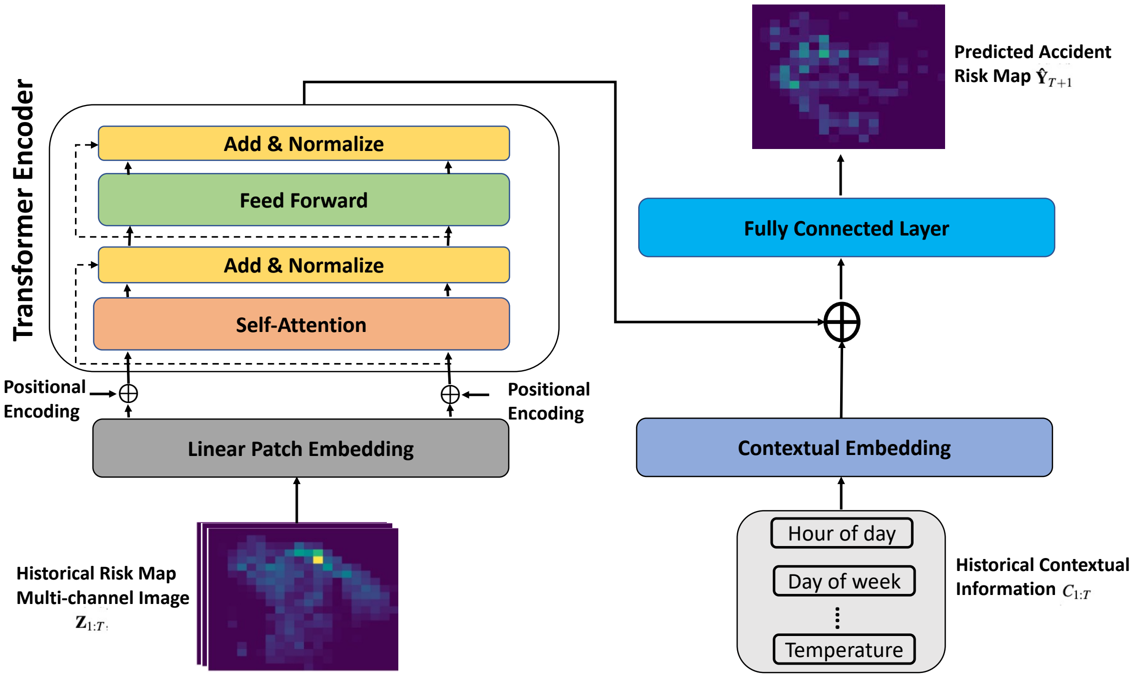



FIGURE 5.2: The building blocks of our proposed C-ViT model.

Across various studies, the problem of forecasting traffic accident risk is commonly addressed as a time-series forecasting task. This involves utilizing historical data on traffic accidents within a specific area, with the potential incorporation of contextual information relating to those accidents. The primary objective is to accurately predict the future risk of traffic accidents for the specific area. Given the nature of the problem, combining both spatial and temporal elements, it is commonly approached using multiple types of model architectures. This implies the necessity of the utilization of a spatial approach, focusing on the affected geographic area, as well as a temporal approach, applied over a defined period of time.

### 5.1.1 Related works

In this study, we compare a set of diverse models to predict traffic accident risk. These models have shown effectiveness in capturing spatial-temporal patterns on the task of traffic accident risk prediction:

RNN-GRU [42]: This model utilizes a variant of deep recurrent neural networks (RNN) known as the gated recurrent unit (GRU). It approaches the traffic accident risk forecasting problem by treating it as a time-series prediction task. The model includes a hidden state that allows it to keep track of long-term dependencies, making it particularly suited for time-series prediction tasks such as traffic accident risk forecasting.

SDCAE [35], [34]: One of the first works on traffic accident risk prediction using Deep Learning has been performed with human mobility data using a Stack Denoise Autoencoder (SDAE) on the Japan traffic network [35], but traffic flow and time-related matters (including periodicity) were not considered. The model is based on the stacked denoised convolutional auto-encoder architecture. This model is able to extract local spatial features from a city grid and, with the autoencoder structure, it can learn a compressed representation that captures the essential spatial patterns related to traffic accident risks.

H-ConvLSTM [259]: The model that combines deep convolution layers with RNN-based LSTM layers. It extracts spatio-temporal features by using a sliding window over the city's grid cells. It relies on a sliding window approach over the grid cells of the city to understand the variations in spatial patterns over time.

GCN [249]: The GCN model is a deep learning approach that leverages graph convolutional neural networks. It represents historical traffic accident data as a graph, allowing it to capture long-term spatio-temporal dependencies. Nodes represent different locations and edges indicate spatial proximity or similarity, the model can uncover long-term connectivity-based patterns.

GSNet [240]: a recent model that incorporates GCN, LSTM, and attention mechanisms to learn complex spatial-temporal correlations in traffic accident risk. It combines the strengths of graph convolution, recurrent modeling, and attention-based mechanisms. Currently, GSNet is considered the state-of-the-art method for the NYC and Chicago datasets.

C-ViT[196]: our previously proposed SoTA model [196], an example of application of Computer Vision model to non-vision task of accident risk prediction. The model utilizes a transformer-based architecture to predict traffic accident risk. It consists of three components: historical risk map encoding, historical contextual information encoding, and a transformer encoder. The historical risk maps are divided into image patches, which are individually passed through a linear embedding layer. The contextual information is encoded using a linear embedding layer. The transformer encoder, with its self-attention mechanism, captures global contextual dependencies across different patches, thereby enhancing the accuracy of future accident risk prediction. Compared to the existing state-of-the-art GSNet model, C-ViT demonstrated competitive performance while offering a more computationally efficient solution.

RiskOracle [268] relied on Graph-Convolution network, utilizing hierarchical coarse-to-fine modelling and proposing minute-level predictions in comparison to day-level [259] and hour-level [35]. In [259] authors have constructed over the ConvLSTM by highlighting the spatial heterogeneity problem and proposing an ensemble of region-specific ConvLSTM models (Hetero-ConvLSTM); they considered weather, the environment and the road condition in Iowa, US for over 8 years of observations, but POIs were not considered. Semantic features, coarse and fine grained risk maps were considered in [237], where also Graph-convolution neural networks and attention-based LSTMs were used. A more recent work in [240] represents the State-of-Art (SoTA) in the field of accident risk prediction, where the authors propose a weighted loss function to address the zero-inflated issue (increase in the number of zero-risk grid cells due to the increase in the granularity of predictions) and making ensemble of models by processing semantic and geo features.

Current study relies on the use of original Visual Transformer [55], [247], which has been widely applied to various tasks in areas of Computer Vision. Transformer models have multiple variations including Convolution Neural Network Enhanced Transformer, Hierarchical Transformer, Transformers with Local Attention, Deep Transformer [142].

So far, risk accident prediction relied mostly upon graph-based methods and spatial-temporal modelling. While this approach worked for limited case study applications, we highly believe that in order to scale it up, this approach can benefit from using visual analysis techniques. Thus, in this work we are re-formulating the problem of traffic accident risk forecasting and we are proposing a novel approach inspired by one of the recent best performing deep learning based architectures for computer vision tasks, the vision transformers [55]. In our proposed model we jointly model and take into account the spatio-temporal nature of the traffic accident risk forecasting problem as well as the influence of contextual information on it using a single unified end-to-end model.

An earlier version of this chapter was presented at the IEEE ITSC 2022 Conference and was published in its Proceedings [196]. The current chapter provides a significant expansion (including two new added sections) of our previous work to further improve accident risk prediction results. The current chapter expansion includes results on Coarse-Fine-Coarse and Static Map Visual transformer architectures.

In Section 5.2, a detailed description about the proposed methodology will be presented. Then, in Section 5.3, we will introduce the datasets we utilised for training and evaluating the performance of our approach, the experiments setup and the baseline approaches from the literature we compared our approach against. Next, we introduce the Coarse-Fine-Coarse Transformer architecture to improve accident risk prediction results in Section 5.4. Then, we propose an incroporation of static maps into ViT architecture in Section 5.5. Finally, in Section 5.6, we conclude our paper.

The code for the chapter can be found by the following link:

`https://github.com/Future-Mobility-Lab/ViT-traffic-accident-risk`

## 5.2 Methodology

**Grid Representation:** We model a specified city region, determined by latitude and longitude bounds, as a uniform grid. This grid comprises $I$ rows and $J$ columns, with each cell being identical in size.

**Traffic Accident Risk:** The risk of traffic accidents at time $t$ for a specific grid cell $i$, designated $Y_t^i$, is quantified as the cumulative weighted sum of various types of traffic incidents that have transpired in that cell. Based on the classification provided in [240], traffic accidents are divided into three types, each assigned a specific weight: minor accidents are given a weight of 1, accidents causing injuries have a weight of 2, and fatal accidents receive a weight of 3. As an example, let's consider a grid cell that has seen two fatal accidents, one accident causing injuries, and four minor accidents. The cumulative traffic accident risk for this grid cell would then be calculated as $(2\times3)+(1\times2)+(4\times1) = 10$.

**Problem Formulation:** We redefine the traffic accident prediction problem from a standard time-series prediction task to an image regression task. We interpret the series of historical traffic accident risk maps, $\mathbf{Z}_{1:T}$ where $\mathbf{Z} \in \mathbb{R}^{I\times J}$ spans the time frame $[1 : T]$, as an image $X$ with a resolution of $I \times J$ and $T$ channels. This image, combined with historical contextual data $C_{1:T}$, is input into our C-ViT model to generate a forecast of the accident risk map for the next hour, $\hat{\mathbf{Y}}_{T+1}$, with $\mathbf{Y} \in \mathbb{R}^{I\times J}$.

### 5.2.1 Contextual Vision Transformer (C-ViT) Model

Given the aforementioned formulation, we compile the traffic accident risk maps $\mathbf{Z}_{1:T}$ as a unified single image with size $T \times I \times J$, where $T$ is the number of channels, $I$ is the image's height and $J$ is the image's width, which we pass as an input to our proposed novel C-ViT model. Our C-ViT model's architecture is inspired by the recently introduced vision transformer network [55] that has been achieving competitive results to the convolutional neural network (ConvNet) architecture for image classification tasks [247], [55].

The C-ViT model is composed of several crucial building blocks, namely the historical traffic accident risk map encoding stage, the historical contextual information encoding stage, and the transformer encoder stage. Each of these components plays a distinct role in the overall functioning of the model. The historical traffic accident risk map encoding stage focuses on encoding the past traffic accidents to generate a comprehensive risk map. Similarly, the historical contextual information encoding stage aims to encode relevant contextual factors to capture their influence on accident risks. Lastly, the transformer encoder stage employs transformer-based architectures to enable effective information processing and feature extraction from the encoded risk map and contextual information.

**Historical Risk Map Encoding:**

Given the historical risk maps as a unified single image $X$ with size $T \times I \times J$, we first encode it into a representation that could be easily digested and learned using our transformer encoder. As it was shown in [229], transformer encoders can work better with input data as a sequence of tokens. Thus, we divide the unified single image into a sequence of equally sub-images $X_p$ which we refer to it as an image patch sequence. We can think of the image patches as a sub-spatial regions of a number of cells within the city's grid representation that we defined in Section **??**. The rationale behind this patching process is derived by the assumption that grid cells that are spatially closer to each others will have some geographical and spatial correlations that could potentially be exploited by our model for conducting a better traffic accident risk forecasting.

Here $X_p$ has a size of $N \times T \times P \times P$, where $P$ is the height/width of the image patch and $N$ is the total number of sequences of image patches, which is defined by $N = IJ/P^2$. The operation of dividing the unified single image into a sequence of image patches $X_p$ can be shown in Fig. 5.3. The image patches sequence are then individually passed through a linear embedding layer which is essentially a learn-able linear projection operation in order to get a sequence of trainable flattened image patches of size $D$, which we refer to as patch embeddings. Additionally, similar to [55], we have an extra learnable embedding token appended before the sequence of patch embeddings to be passed to

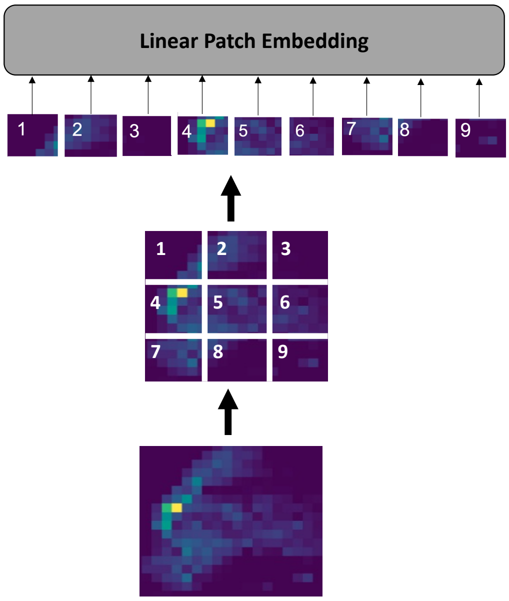



FIGURE 5.3: Description of the first stage of the historical risk Map encoding. Given a unified single image $X$, it is then divided into equally-sized image patches that are passed individually to the linear patch embedding layer.

the transformer encoder and we refer to this embedding as a "regression token". The regression token embedding acts as an image representation which its output is transformed inside the transformer encoder into the predicted accident risk map $\hat{\mathbf{Y}}_{T+1}$.

Since the transformer encoder does not have the notion of order in its input sequence tokens, an additional position embeddings are added to each patch embedding. There are a number of pathways to define position embedding, and in our current model we follow the formulation introduced in [229]. In this formulation, the position encoding $PE$ vector is defined by using a wide spectrum of frequencies of sine/cosine functions as follows:

$$\begin{aligned} PE_{(a,2k)} &= \sin\left(a/10000^{2k/D}\right) \\ PE_{(a,2k+1)} &= \cos\left(a/10000^{2k/D}\right) \end{aligned} \tag{5.1}$$

where $a$ represents the position, and $k$ is the dimension. From the above formulation, once can conclude that for each dimension $k$ of $PE$ vector, it has a corresponding sinusoid that spans a frequency range from $2\pi$ to $10000 \cdot 2\pi$. In other words, this will allow the model to be mindful of the order in the sequential patch embedding by using unique relative positions. The dimension of the $PE$ vector is similar to the linear patch embedding layer's dimension which is $D$.

**Historical Contextual Information Encoding** As discussed in Section **??**, besides the historical accident risk maps, our C-ViT model takes into account also the historical contextual information $C_{1:T}$ for the city grid representation. In our model and similar to [240], we took into account the following contextual features: 1) the time period of the day, 2) the day of the week, 3) whether the day is a holiday or not, 4) the weather condition (clear, cloud,..etc), 5) the weather temperature, and 6) traffic condition (inflow and outflow). Given those contextual features, we encode them via a learnable linear embedding layer of dimension $D$, whose output is fused together with the output from the transformer encoder via a concatenation operation.

**Transformer Encoder** The main building block of our transformer encoder is the multi-head self-attention module [229]. In total we have six layers inside our transformer encoder. Internally, each layer is composed of a both self-attention head and feed-forward fully connected sub-layers. Additionally, each sub-layer is followed by two residual connections and a normalisation operation. The multi-head self-attention, or the multi-scaled dot-product attention, works based on the mapping between the so-called ‘query’ vectors and the pair (key, value) vectors. The dimension of the query and key vectors is $d_k$, where the values vector dimension is $d_v$. The attention operation itself is computed by taking the dot-product between the query and the key vectors divided by the square root of $d_k$ before finally passing them to the softmax function to get their weights by their values. Since the scaled dot-product attention operation is done multiple times, the queries, keys and values vectors are extended into matrices $Q, K, V$ respectively. The following formula is the description of how the scaled dot-product attention operation is calculated:

$$\text{Attention}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5.2}$$

## 5.3 Experiments and Results

In this section, we first present the datasets we utilised for training and evaluating the performance of our proposed approach. Then, we provide the details of the setup for our experiments, the evaluation metrics and the compared baseline approaches from the the literature. Finally, the quantitative and qualitative results of our proposed approach on real-life datasets are evaluated and discussed.

### 5.3.1 Datasets

In our study, we utilize two publicly accessible real-world datasets for forecasting traffic accident risk: NYC[1] and Chicago[2].

The NYC dataset, reported from January 1, 2013, to December 31, 2013, consists of about 147K accidents, 173,179K taxi trips, 15,625 points of interest (POIs), 8,760 weather reports, and data about a road network consisting of 103K segments. An additional feature unique to this dataset is its Point of Interest (POI) data, which provides information about specific locations like residences, schools, cultural facilities, recreation spots, social services, transportation hubs, and commercial centers.

The Chicago dataset, reported from February 1, 2016, to September 30, 2016, contains approximately 44K accidents, 1,744K taxi trips, 5,832 weather reports, and data about a road network comprising 56K segments.

As it can be seen from Table 5.1, both datasets have historical traffic accidents and historical taxi trips. Both datasets include historical traffic accidents and taxi trips data. The traffic accident data provides details about time, date, location (latitude and longitude), the number of causalities, weather condition (clear, cloudy, rainy, snowy, or mist), temperature, and road segment data (i.e., road length,

[1] https://opendata.cityofnewyork.us/
[2] https://data.cityofchicago.org/

| Dataset | Attributes | Range/Count |
|---|---|---|
| NYC | Reporting Duration | 1 Jan 2013 - 31 Dec 2013 |
| | Accidents | 147K |
| | Taxi Trips | 173,179K |
| | POIs | 15,625 |
| | Weathers | 8,760 |
| | Road Network | 103K |
| Chicago | Reporting Duration | 1 Feb 2016 - 30 Sep 2016 |
| | Accidents | 44K |
| | Taxi Trips | 1,744K |
| | Weathers | 5,832 |
| | Road Network | 56K |

TABLE 5.1: Datasets Statistics

width, and type). The taxi trip data, which includes location and times of pick-up and drop-offs, is utilized to compute the inflow/outflow of traffic condition in each area.

### 5.3.2 Experiment Setup

Before we train and evaluate our proposed C-ViT model, we first pre-process the two datasets. The first pre-processing stage was to perform a grid representation by dividing each city map of the two datasets (i.e. NYC and Chicago) into equally-sized grid cells each with a dimension of ($2KM \times 2KM$). Secondly, similar to [240], we group all the accidents that happened in each grid cell based on their location over the reported duration time for each dataset (for each grid cells with no road segments/accidents, we set its traffic accident risk to zero).

In this study, we employed a systematic approach to divide the data into distinct subsets for training, validation, and testing purposes. Following a similar strategy as outlined in a previous study utilizing GSNet architecture [240], the data-sets were partitioned with a ratio of 60% for training, 20% for validation, and 20% for testing. To ensure the reliability of the splits, we strictly prevented any instances of overlapping accidents based on the occurrence times, ensuring that no accident shared the same grid cell and specific time stamp across the three subsets. The periodicity of traffic accidents, as observed in both datasets, was set to 1 hour. Prior to the analysis, each data split went through standardization using mean and standard deviation normalization. This normalization process helps with accelerating the model training and enhancing the efficiency of the model.

Regarding the implementation details of our C-ViT model, the size of the historical traffic risk maps $X$ was set to $7 \times 20 \times 20$ which corresponds to a total 7 historical traffic accident risks across the city grid with $I$ rows $\times$ $J$ columns of size 20. Here we chose 7 historical accident risks specifically to conform with the work done in the literature [34], [240] for a fair comparison provided later. For each grid cell, the 7 historical accident risks comes from the most recent accident risks in past 3 hours in addition to the past accident risks in the last 4 weeks. The prediction horizon pf the traffic accident risk was set to 1 (i.e next hour) similar to [34], [240].

In our study, we have followed a set of comprehensive steps for data pre-processing and have detailed the implementation of our proposed enahncement to C-ViT model. These steps ensure the validity of our research while facilitating comparison with previous studies. The specific steps of data preprocessing, and the parameters of the C-ViT model, are presented in the following table (Table 5.2).

| **Processing Step / Implementation Detail** | **Specification** |
|---|---|
| Datasets | NYC, Chicago |
| Grid Representation | Each city map divided into grid cells of $2KM \times 2KM$ |
| Accident Grouping | All accidents in each grid cell grouped based on location and duration time |
| Data Split | Training: 60%, Validation: 20%, Testing: 20% |
| Overlapping Accident Control | No overlapping accidents based on time |
| Data Standardization | Mean and standard deviation normalization |
| Traffic Accidents Periodicity | 1 hour |
| Historical Traffic Risk Map Size | $7 \times 20 \times 20$ |
| Historical Accident Risks | 7 (most recent accident risks in past 3 hours + past accident risks in the last 4 weeks) |
| Prediction Horizon of Traffic Accident Risk | 1 (next hour) |
| Dimension ($D$) of Linear Patch Embedding | 64 |
| Dimension ($D$) of Position Embedding Layer | 64 |
| Dimension ($D$) of Linear Embedding Layer of the Historical Contextual Encoder | 64 |
| Resolution of Input Patches ($P$) to the Patch Embedding Layer | 5 |
| Number of Self-Attention Heads | 8 |
| Dimension of the Final Output Fully Connected Layer | 128 |
| Optimization Function | Weighted Mean-Squared Error (MSE) |
| Loss Weighting Procedure | Focal loss |
| Risk Value Classes | 0, 1, 2, ≥3 |
| Loss Function Weights | 0.05, 0.2, 0.25, 0.5 |
| Training Epochs | 200 |
| Optimizer | Adam |
| Learning Rate | 0.003 |
| Batch Size | 32 |

TABLE 5.2: Setup for X-ViT Model and Preprocessing Procedures

Since we formulated the traffic accident risk prediction task as an image regression task, we have therefore optimised our C-ViT model during the training phase using a weighted mean-squared error (MSE) loss function. The reason for using the weighted MSE loss function instead of using the standard MSE loss function, is to try to combat the unbalanced nature of the traffic risk prediction problem, also known as the zero-inflated problem [17]. The procedure for weighting our loss function is motivated by the focal loss introduced in [138], where we holistically divided the total training samples into four distinctive classes based on their traffic accident risk values. Those risk values are (0, 1, 2, ≥3). Similar to [240], the loss function weights were set to 0.05, 0.2, 0.25 and 0.5 respectively. In total, we have trained our C-ViT model for 200 epochs using the Adam optimiser with a learning rate of 0.003 and the batch size was set to 32.

### 5.3.3 Evaluation Metrics

In order to evaluate the performance of our trained C-ViT model, we utilised the three commonly used metrics for the traffic accident risk prediction task [149], [240], namely root mean squared error

TABLE 5.3: Performance evaluation of our C-ViT model against a number of baseline approaches from the literature over the NYC and Chicago datasets.

| Dataset | NYC | | | Chicago | | |
|---|---|---|---|---|---|---|
| Model | RMSE ↓ | Recall ↑ | MAP ↑ | RMSE ↓ | Recall ↑ | MAP ↑ |
| RNN-GRU [42] | 8.3375 | 28.09% | 0.1228 | 12.6482 | 17.83% | 0.0664 |
| SDCAE [34] | 7.9774 | 30.81% | 0.1594 | 11.3382 | 18.78% | 0.0753 |
| H-ConvLSTM [259] | 7.9731 | 30.42% | 0.1454 | 11.3033 | 18.43% | 0.0716 |
| GCN [249] | 7.7358 | 31.78% | 0.1623 | 11.0835 | 18.95% | 0.0805 |
| GSNet [240] | 7.6151 | 33.16% | 0.1787 | 11.3726 | 19.92% | 0.0822 |
| C-ViT (ours) | **7.0053** | **33.86%** | **0.1875** | **9.4456** | **20.93%** | **0.0980** |

(RMSE), Recall and mean average precision (MAP). The three evaluation metrics are calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{Y}_n\right)^2}, \tag{5.3}$$

$$\text{Recall} = \frac{1}{N}\sum_{n=1}^{N}\frac{H_n \cap A_n}{|A_t|}, \tag{5.4}$$

$$\text{MAP} = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{j=1}^{|A_t|}\text{PR}(j)\times\text{REC}(j)}{|A_n|}, \tag{5.5}$$

where $N$ is the total number of samples to be evaluated, $Y_n$, $\hat{Y}_n$ are the ground truth and the predicted risk values for all grid cells of sample $n$ respectively. $A_n$ corresponds to the set of grid cells of sample $n$ that have an actual/true traffic accident risk values. $H_n$ corresponds to the set of grid cells within $A_n$ with the highest traffic accident risk values. On the other hand, PR($j$) corresponds to the precision of the grid cells starting at 1 and ending at grid cell $j$. Similarly, REC($j$) corresponds to the recall value for grid cell $j$ which is set to 1 in case there was a traffic accident risk at it and set to 0 otherwise.

Based on the definition of these three evaluation metrics, we can deduce that the lower the score of RMSE is, the better is the quality of prediction coming out of the model. On the other hand, the higher the recall and MAP scores are, the better is the accuracy of the model.

### 5.3.4 Baselines

We have compared the performance of our proposed C-ViT model to 5 different baseline approaches from the literature and in the following we will briefly describe each approach:

- **RNN-GRU [42]:** This model is based on one variant of deep recurrent neural networks (RNN), the gated recurrent unit (GRU) model. This model casts the traffic accident risk forecasting problem as a time-series prediction problem and tries to model the temporal dependency among historical traffic accidents risk.

- **SDCAE [34]:** This model is based on the stacked denoised convolution auto-encoder architecture, which focuses mainly on capturing/modelling the spatial features between different cells within a city grid area for a better prediction of the traffic accident risk.

TABLE 5.4: Performance evaluation of our C-ViT model against a number of baseline approaches from the literature over the high frequency times of accidents in the NYC and Chicago datasets.

| Dataset | NYC | | | Chicago | | |
|---|---|---|---|---|---|---|
| Model | RMSE ↓ | Recall ↑ | MAP ↑ | RMSE ↓ | Recall ↑ | MAP ↑ |
| RNN-GRU [42] | 7.3546 | 30.76% | 0.1301 | 9.0421 | 18.66% | 0.0758 |
| SDCAE [34] | 7.2806 | 31.22% | 0.1536 | 8.7543 | 20.58% | 0.1002 |
| H-ConvLSTM [259] | 7.2750 | 31.43% | 0.1498 | 8.5437 | 18.93% | 0.0770 |
| GCN [249] | 7.0958 | 33.04% | 0.1647 | 8.4484 | 20.42% | 0.0933 |
| GSNet [240] | 6.7758 | 34.15% | 0.1769 | 8.6420 | 21.12% | 0.1052 |
| C-ViT (ours) | **6.2658** | **34.46%** | **0.1802** | **7.0353** | **21.95%** | **0.1247** |

- **H-ConvLSTM [259]:** As the name implies, this model combines both deep convolution layers with RNN-based LSTM layers to extract the spatio-temporal features of the traffic accident risk problem by having a sliding window over the city's grid cells; this allows to have sub-regions that could potentially capture the heterogeneity among the different types of spatial regions.
- **GCN [249]:** This model is a deep learning model that relies on graph convolution neural network to represent the historical traffic accident data as a graph to capture the long-term spatio-temporal dependency among historical traffic accidents risk data.
- **GSNet [240]:** A recent model that learns the complex spatial-temporal correlations of traffic accidents risk by using a combination of GCN, LSTM and attention mechanism. To the best of our knowledge, GSNet is currently the SOTA method on the NYC and Chicago data-sets.

### 5.3.5 Results

In Table 5.3, we report the results of our C-ViT model in comparison to the aforementioned baseline approaches from the literature over the total testing splits for both NYC and Chicago data-sets. As it can be noticed, our model has outperformed all the baseline approaches from the literature in terms of RMSE, recall and MAP scores over the two data-sets. It is worth noting from the results, that those models (our C-ViT, GSNet, GCN and H-ConvLSTM) which account for the spatio-temporal property of the traffic accident risk prediction problem, are the top performing approaches on the two data-sets.

The closest competitor baseline approach to our C-ViT model, was the GSNet, which to the best of our knowledge, was the SOTA on the two data-sets before our proposed approach. As it can be seen, our C-ViT model has improved the RMSE, recall and MAP scores in comparison to GSNet especially across the Chicago dataset by a relatively large margin. Furthermore, our C-ViT has more competitive advantage over GSNet in terms of the efficiency. As it can be shown in Fig. 5.4, the number of parameters required by our C-ViT model for training are far lower than those needed for GSNet (saving more than 23x parameters) which makes our approach more suitable for real-time deployment.

In order to further evaluate the performance of our proposed C-ViT model, in Table 5.4 we report the RMSE, recall and MAP scores of our model when compared to all the other baseline approaches

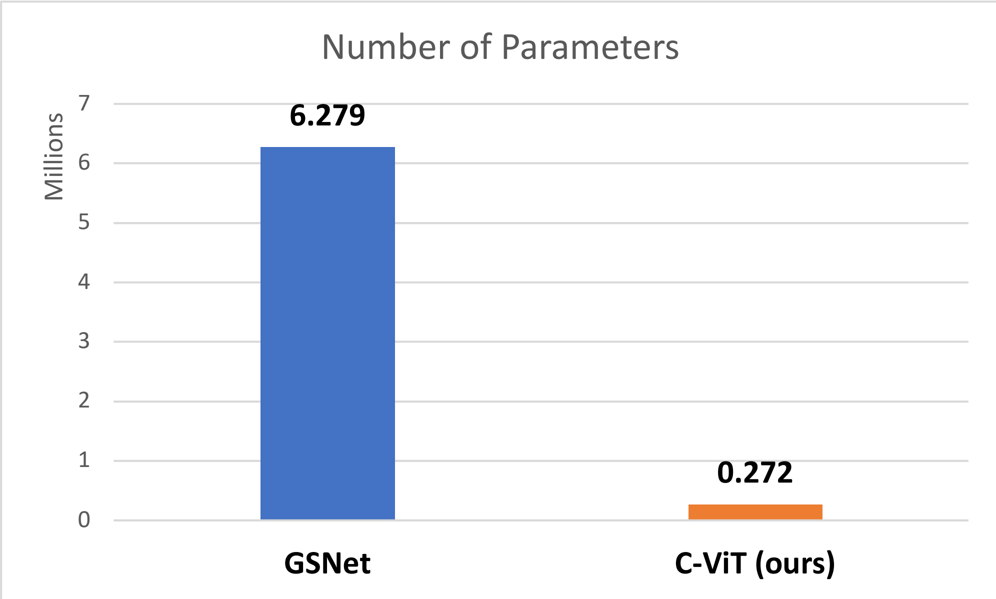



FIGURE 5.4: Comparison between our proposed C-ViT model and GSNet [240], in terms of the number of training parameters.

over peak hours of frequent traffic accidents that resulted from the testing split of both the NYC and Chicago data-sets. Those times of high frequency of traffic accidents are essentially during morning/evening rush hours which are within 7:00-9:00 AM and 04:00-07:00 PM. As it can be seen from the reported results, our C-ViT model continues to achieve more robust results than all other compared baseline approaches. This further prove the utility and quality of our proposed approach that it has a consistent performance across different settings.

## 5.4 Coarse-Fine-Coarse Visual Transformer (CFC-ViT)

One of the issues to solve in the topic of accident risk prediction is the zero-inflated issue - the imbalance between the amount of non-zero and zero accident risk cells. This issue can be resolved by using a comparison mask or variations of focal loss [240]. Another issue, which is usually ingnored is the fine granularity of accident risk map. For example, in the grid representation, cells can be separated and of minimal 1x1 cell size (see Fig. 5.5).

Current computer vision methods applied to the task of accident risk prediction produce 'blurred' results due to intrinisic limitations of convolution network architectures [133], [77]. To resolve this issue we propose an alternative approach which consists of up-scaling the patches before the embedding to allow fine-grained processing by internal layers of transformer, and then down-scaling embedding to match the original output shape 5.6. Up-scaling may be a necessary step in the of use of asymmetric convolutional networks for segmentation to increase the detailization of results [143] since there are no upscaling layers at the final part of the asymmetric network. We upscale patches before the embedding to perform fine-grained processing of these patches. The dimensionality of embedding is also increased proportionally to the patch size; processed embedding then downscaled by the same rate. This allows the network to form intermediate results of higher dimensionality, which when down-scaled, will produce more fine-grained image.

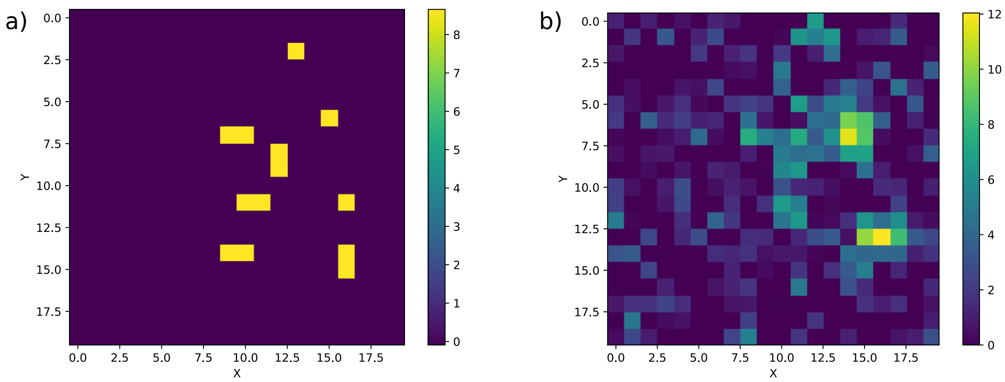


FIGURE 5.5: Example of GSNet predictions (after training for 2 epochs, when the best performance is observed): a) Actual map of the accident occurrence b) Predicted map of the accident occurrence.

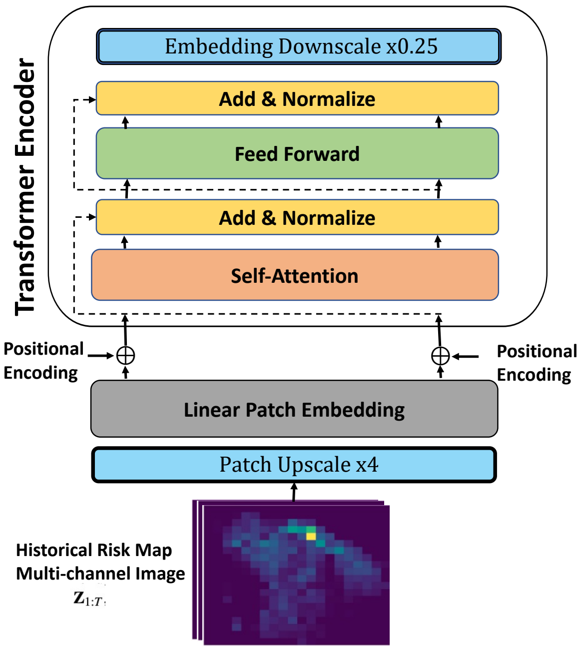



FIGURE 5.6: Coarse-Fine-Coarse Transformer

TABLE 5.5: Performance evaluation of our CFC-ViT model on Chicago data set

| | RMSE | Recall | MAP | HFT-RMSE | HFT-Recall | HFT-MAP | Data set | CFC Scale factors |
|---|---|---|---|---|---|---|---|---|
| 0 | **8.62** | 19.38 | 0.08 | **6.48** | 19.89 | 0.09 | chicago | 4x, 0.25x |
| 1 | 9.24 | 18.84 | 0.06 | 6.85 | 20.85 | 0.07 | chicago | 2x, 0.5x |
| – | – | – | – | – | – | – | – | – |
| - | 9.45 | 20.93 | 0.098 | 7.035 | 21.95 | 0.125 | chicago | 1x,**baseline (multi-epoch)** |

TABLE 5.6: Performance evaluation of our CFC-ViT model on NYC data set

| | RMSE | Recall | MAP | HFT-RMSE | HFT-Recall | HFT-MAP | Data set | CFC Scale factors |
|---|---|---|---|---|---|---|---|---|
| 0 | 7.09 | 33.17 | 0.1808 | 6.45 | 33.90 | 0.1751 | NYC | 4x, 0.25x |
| 1 | **6.81** | 32.15 | 0.1838 | **6.14** | 33.24 | 0.176 | NYC | 2x, 0.5x |
| – | – | – | – | – | – | – | – | – |
| - | 7.0053 | 33.86 | 0.1875 | 6.2658 | 34.46 | 0.1802 | NYC | 1x, **baseline (multi-epoch)** |

Results for the Chicago data set show a significant improvement in the RMSE metric results both for 2x and 4x scale factors (see Table 5.5 where the RMSE is 8.62 as compared to 9.45 translating in a 8.78% improvement). There is an inverse dependence observed between the scale factor and the Recall or MAP metrics: the increase in the scale factor lowers RMSE but MAP and recall also decrease. However, given the robustness of the RMSE metric, the improvement is consistent.

Results for the NYC data set show that the prediction performance can increase at a specific scale factor (2x) and decrease at different scale factor (4x) (see Table 5.6. These results suggest that the optimal scale factor for each data set can exist, which leads to a deciated optimization task of finding the optimal scale factor value. Results both for NYC and Chicago data sets show a non-linear dependency between the RMSE, the MAP (mean average precision) or Recall metrics. These metrics are intended for different purposes (RMSE for the regression, MAP and Recall for the classification results) and therefore can produce different results based on the characteristics of the predicted values.

Overall, our new proposed CFC-ViT approach shows an improvement in the RMSE results, but these results and other metrics depend on the scale factor parameter. The optimal scale factor can vary for each data set and can be found using other optimization techniques.

## 5.5 Application of the Static Map Generation

The use of Attention layers is a computer vision technique which implies an estimation of attention maps from different images. Since each image may have different areas of attention, the attention map is generated for every case of prediction (which we can call the dynamic attention estimation). But in the case of accident risk prediction, we predict on the same area each time. Therefore, we can use the statically generated attention map (static attention estimation). We evaluate multiple scenarios of combining dynamic (DA) and static attention (SA) estimations using varying combination operations. To further utilise the advantage of a non-volatile area, we also try to generate the Static Accident Risk Map (S-ARM) so our network needs to predict the offset of the accident risk (relative accident risk) from the statically generated risk map values instead of predicting the absolute accident risk values. Therefore, another contribution of this work is to further combine the Predicted Offset Accident Risk Map (PO-ARM) with the Static Accident Risk Map (S-ARM) (see Fig. 5.7).

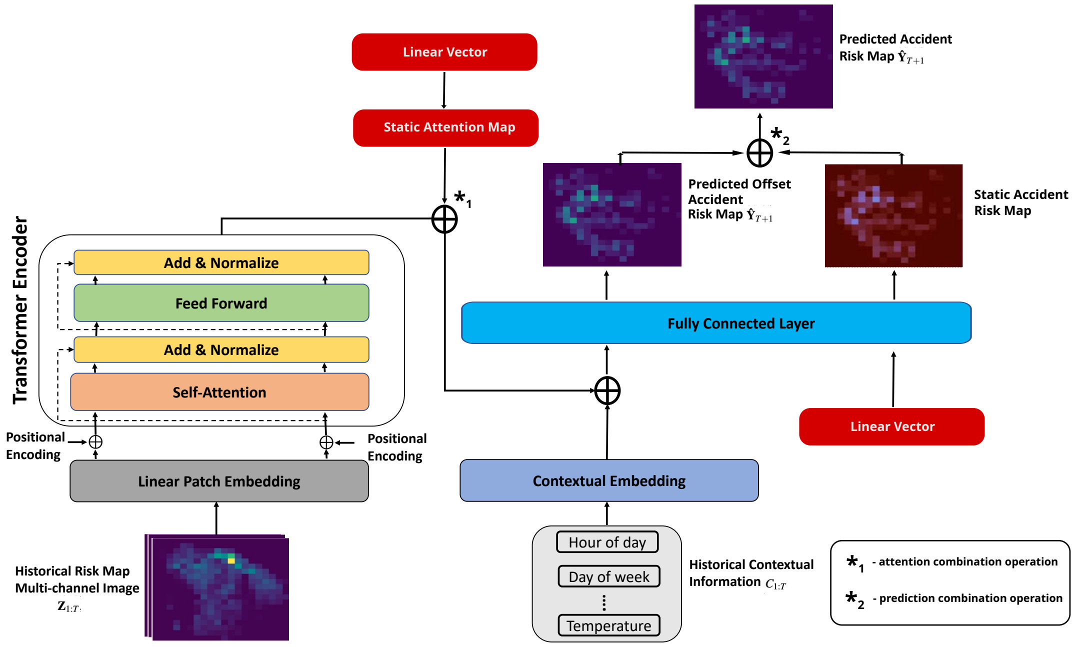



FIGURE 5.7: The building blocks of our proposed XViT model with Static Map Generation

### 5.5.1 Pipeline description

The generalisation performance of the Transformer model can be greatly improved by using one-epoch training [113]. Therefore we use results obtained from one-epoch training in further scenarios. Other parameters of the setup are the same as in Section 5.3.2.

### 5.5.2 Description of combination operations

The use of static map generation at the beginning of the attention layer as well as near the network output can remove the necessity for the network to predict the absolute risk values (static map is assumed to act as a static image and network is required to predict the relative risk from the one in a static map). We test multiple different approaches to achieve the benefit of using the static map generation. Different constraint functions can be used to limit the range of values observed from the static map. Also, the actual static map can be combined differently with the final and the intermediate network values.

Combination operations for the attention layer that we have considered:

1. None - using only the original pipeline structure with no static map,
2. tanh(map)+x - the static map is bounded by the tanh function in order to obtain the static map values distribution between (-1,+1), combined with layer input values,
3. tanh(map)*x - the same as above, but combined using a multiplication operation,
4. map - we use the static map instead of the attention layer inputs,

5. map + x - we combine the static map with the attention layer inputs using the "plus (+)" operation,

6. sigmoid(map)+x - static map values are distributed between (0,+1) and are further combined with the attention layer inputs by using the "plus (+0)" operation (a linear offset combination),

7. sigmoid(map)*x - the same as above, but combined using the "multiplication (*)" operation.

Combination operations for the network output that we have implemented:

1. tanh(map)+x in which the static map values (map) are combined with the intermediate predictions (transformer output - x),

2. sigmoid(map)+x - same as above, but using sigmoid as a static map constraining function,

3. tanh(map)*x - tanh is used as a static map constraining function, and the static map values (map) are combined with the intermediate predictions by using the "multiplication (*)" operation,

4. head(map+non) - the static map values combined with the non-risk features and passed through the feed forward neural network,

5. head(map)+head2(x+non) - the static map is passed through a separate feed forward neural network, while other predictions together with the non-risk features are passed through the second network of the same structure,

6. head(map)*head(x) - the accident features are passed through the same network as the static map values and then combined using the "multiplication (*)" operation,

7. head(non)+head(x) - the non-risk features and the accident risk features are passed through the same network, and then combined using the plus operation,

8. None (head(x+non))- this is the original ViT implementation.

The constraint functions (tanh and sigmoid) are tested with the assumption that values close to the actual normalised accident risk values will be observed right after the network parameter initialisation. Due to the variation, we name this derivative model an XViT model.

### 5.5.3 S-ARM results

The results for the Chicago and NYC data sets are provided in Tables 5.7-5.8, RMSE results conveniently represented on Figures 5.9-5.8. The results are also provided for the high-frequency hours (HFT) - meaning the RMSE errors obtained only when using the HFT hours when more traffic is normally expected in the city. The use of the static map generation didn't show an improvement on the NYC data set. In fact, the results are a bit worse but closely realted to the baseline (7.05 RMSE for the best combination vs 7.00 RMSE using original baseline multi-epoch approach); however we observe that there is an improvement in the recall results for the high-frequency hours (34.84 for the best combination vs 34.25 when using the multi-epoch baseline). But results forthe Chicago data set show a very significant improvement across all the metrics (e.g. from 9.45 to 9.01 in RMSE, from

TABLE 5.7: Performance evaluation of our XViT model for a number of combination operations on NYC data set. Top 20 results.

| | RMSE | Recall | MAP | HFT-RMSE | HFT-Recall | HFT-MAP | S-ARM Combination Operations |
|---|---|---|---|---|---|---|---|
| 0 | 7.05 | 33.72 | 0.19 | 6.46 | 34.84 | 0.18 | **tanh(map)*x, tanh(map)+x** |
| 1 | 7.07 | 33.49 | 0.19 | 6.48 | 34.43 | 0.18 | sigmoid(map)+x, sigmoid(map)+x |
| 2 | 7.10 | 33.21 | 0.19 | 6.50 | 33.87 | 0.18 | sigmoid(map)+x, head(map)+head(x) |
| 3 | 7.10 | 33.57 | 0.19 | 6.50 | 34.15 | 0.18 | sigmoid(map)+x, tanh(map)+x |
| 4 | 7.11 | 33.26 | 0.19 | 6.49 | 33.76 | 0.18 | none, tanh(map)*x |
| 5 | 7.11 | 33.38 | 0.19 | 6.52 | 34.53 | 0.19 | sigmoid(map)*x, tanh(map)+x |
| 6 | 7.11 | 33.39 | 0.19 | 6.52 | 34.71 | 0.18 | tanh(map)*x, sigmoid(map)+x |
| 7 | 7.11 | 33.85 | 0.19 | 6.50 | 34.81 | 0.19 | sigmoid(map)*x, tanh(map)*x |
| 8 | 7.12 | 33.21 | 0.19 | 6.52 | 34.50 | 0.18 | tanh(map)*x, head(map)+head(x) |
| 9 | 7.13 | 33.33 | 0.19 | 6.54 | 34.71 | 0.18 | tanh(map)+x, sigmoid(map)*x |
| 10 | 7.13 | 33.36 | 0.19 | 6.52 | 34.64 | 0.19 | sigmoid(map)+x, sigmoid(map)*x |
| 11 | 7.13 | 33.43 | 0.19 | 6.54 | 34.53 | 0.19 | tanh(map)+x, tanh(map)+x |
| 12 | 7.14 | 32.88 | 0.19 | 6.55 | 33.41 | 0.18 | map+x, head(map)+head(x) |
| 13 | 7.14 | 33.09 | 0.19 | 6.55 | 34.18 | 0.19 | map+x, sigmoid(map)+x |
| 14 | 7.14 | 33.35 | 0.19 | 6.55 | 34.25 | 0.19 | none, tanh(map)+x |
| 15 | 7.14 | 33.36 | 0.18 | 6.52 | 34.67 | 0.18 | tanh(map)+x, tanh(map)*x |
| 16 | 7.14 | 33.55 | 0.19 | 6.54 | 34.81 | 0.19 | sigmoid(map)*x, sigmoid(map)+x |
| 17 | 7.15 | 32.87 | 0.19 | 6.55 | 33.87 | 0.19 | sigmoid(map)*x, head(map)+head(x) |
| 18 | 7.15 | 33.09 | 0.19 | 6.56 | 33.87 | 0.18 | map+x, tanh(map)+x |
| 19 | 7.15 | 33.24 | 0.19 | 6.55 | 34.71 | 0.19 | map+x, sigmoid(map)*x |
| 20 | 7.15 | 33.30 | 0.19 | 6.57 | 34.36 | 0.18 | none, sigmoid(map)+x |
| – | – | – | – | – | – | – | – |
| - | 7.25 | 33.39 | 0.19 | 6.53 | 34.25 | 0.19 | **baseline (1-epoch)** |
| - | 7.0053 | 33.86 | 0.1875 | 6.2658 | 34.46 | 0.1802 | **baseline (multi-epoch)** |

20.93 to 22.24 in Recall, from 21.95 to 23.46 in HFT-Recall). More than that, the 1-epoch training also shows a significant improvement in case of the baseline ViT structure (from 9.45 to 9.25 RMSE, from 20.93 to 21.77 Recall, from 7.035 to 6.93 in HFT-RMSE).

This slight reduction in the model performance in case of the NYC data set and significant improvement in case of the Chicago data set can be interpreted through the concept of local optima and data set size. There may be multiple local optima for the accident risk approximation across historical accident risk records (e.g. multiple average risk maps for different months). This optima can have an ability to show a good approximation of the accident risk, but since the road networks and the city structures change over time, different local optima can appear over time as well. So finding just one static accident map may not be optimal for a large data set, but may show benefit in the case of small data set (Chicago has just 44K accident records in comparison to 147K for NYC attributing to 1 full year of records and these are mostly short-time accidents in Chicago - just 8 months). We conclude that there is evidence that the proposed method and the use of multiple static maps can be a topic of the future research which can bring improvement over large data sets.

Another important observation is that the same set of combination operations gives the best results in the case of the ViT network with a generated static map: "tanh(map)*x" in the attention layer and "tanh(map)+x" near the network output. This not only signifies the use of the constraint function tanh, but also shows where to use each combination operator (addition and multiplication). We also observe that the use of non-risk features is not present among the top 20 results for NYC data set (see Table 5.7, Figure 5.8), while for the Chicago data set it is present in 9 combinations out of 20, which may

TABLE 5.8: Performance evaluation of our XViT model for a number of combination operations on Chicago data set. Top 20 results.

| | RMSE | Recall | MAP | HFT-RMSE | HFT-Recall | HFT-MAP | S-ARM Combination Operations |
|---|---|---|---|---|---|---|---|
| 0 | 9.01 | 22.24 | 0.11 | 6.80 | 23.46 | 0.13 | **tanh(map)*x, tanh(map)+x** |
| 1 | 9.06 | 22.30 | 0.10 | 6.85 | 23.59 | 0.13 | tanh(map)*x, head(map)+head(x) |
| 2 | 9.09 | 22.30 | 0.10 | 6.81 | 23.18 | 0.11 | none, head(non)+head(x) |
| 3 | 9.10 | 21.77 | 0.10 | 6.80 | 22.63 | 0.13 | none, head(map)+head2(x+non) |
| 4 | 9.12 | 21.88 | 0.10 | 6.80 | 23.05 | 0.13 | sigmoid(map)*x, head(map)+head2(x+non) |
| 5 | 9.14 | 22.90 | 0.11 | 6.82 | 24.14 | 0.12 | sigmoid(map)*x, head(non)+head(x) |
| 6 | 9.15 | 22.12 | 0.11 | 6.91 | 22.63 | 0.13 | none, sigmoid(map)+x |
| 7 | 9.15 | 22.18 | 0.11 | 6.94 | 22.91 | 0.13 | tanh(map)*x, sigmoid(map)+x |
| 8 | 9.16 | 21.59 | 0.11 | 6.98 | 21.81 | 0.13 | tanh(map)*x, sigmoid(map)*x |
| 9 | 9.17 | 22.72 | 0.10 | 6.82 | 24.01 | 0.12 | tanh(map)*x, head(non)+head(x) |
| 10 | 9.19 | 21.59 | 0.10 | 6.83 | 22.63 | 0.12 | tanh(map)*x, head(map)+head2(x+non) |
| 11 | 9.22 | 22.00 | 0.11 | 6.88 | 22.63 | 0.13 | tanh(map)*x, head(x+non) |
| 12 | 9.22 | 22.18 | 0.11 | 6.97 | 23.18 | 0.13 | none, tanh(map)+x |
| 13 | 9.22 | 22.24 | 0.11 | 7.00 | 23.18 | 0.13 | sigmoid(map)*x, tanh(map)+x |
| 14 | 9.22 | 22.24 | 0.11 | 7.00 | 23.32 | 0.13 | sigmoid(map)*x, head(map)+head(x) |
| 15 | 9.24 | 22.00 | 0.11 | 6.99 | 22.77 | 0.13 | none, head(map)+head(x) |
| 16 | 9.25 | 21.77 | 0.10 | 6.93 | 22.36 | 0.12 | **baseline (1-epoch)** |
| 17 | 9.29 | 22.00 | 0.11 | 6.96 | 22.63 | 0.13 | sigmoid(map)*x, head(x+non) |
| 18 | 9.29 | 22.12 | 0.11 | 6.97 | 22.77 | 0.13 | tanh(map)+x, head(map)+head2(x+non) |
| 19 | 9.31 | 21.88 | 0.10 | 7.08 | 22.50 | 0.13 | tanh(map)+x, sigmoid(map)*x |
| 20 | 9.34 | 22.60 | 0.10 | 6.97 | 23.87 | 0.12 | sigmoid(map)+x, head(non)+head(x) |
| – | – | – | – | – | – | – | – |
| 16 | 9.25 | 21.77 | 0.10 | 6.93 | 22.36 | 0.12 | **baseline (1-epoch)** |
| - | 9.45 | 20.93 | 0.098 | 7.035 | 21.95 | 0.125 | **baseline (multi-epoch)** |

indicate the difference in quality of these features in both data sets.

## 5.6 Conclusion

In this work, we have presented a novel approach for the task of traffic accident risk forecasting. In our approach we re-formulated the problem as an image regression problem and introduced a unique contextual vision transformer network (C-ViT) that can efficiently model the traffic accident risk forecasting task from both spatial and temporal perspectives. The proposed approach has been evaluated on two publicly available data sets for the traffic accident risk problem. Furthermore, our proposed C-ViT model has been compared against a number of baseline approaches from the literature and it has outperformed them with a large margin while only requiring less than 23 times the number of training parameters.

The combination of static accident risk map with the ViT model (XVit) provides an even more significant improvement over the previous method in case of the New-York data set, thus establishing the new SoTA in the study area. The operation combination method has a potential for improvement (e.g. more different combination methods and constraint functions can be tested). Improvements in results obtained in the current research can also highlight the applicability of vision transformers for non-visual tasks.

The Coarse-Fine-Coarse Visual Transformer (CFC-Vit) architecture allows for fine-grained processing of the accident risk map and introduces an additional scale factor parameter which affects

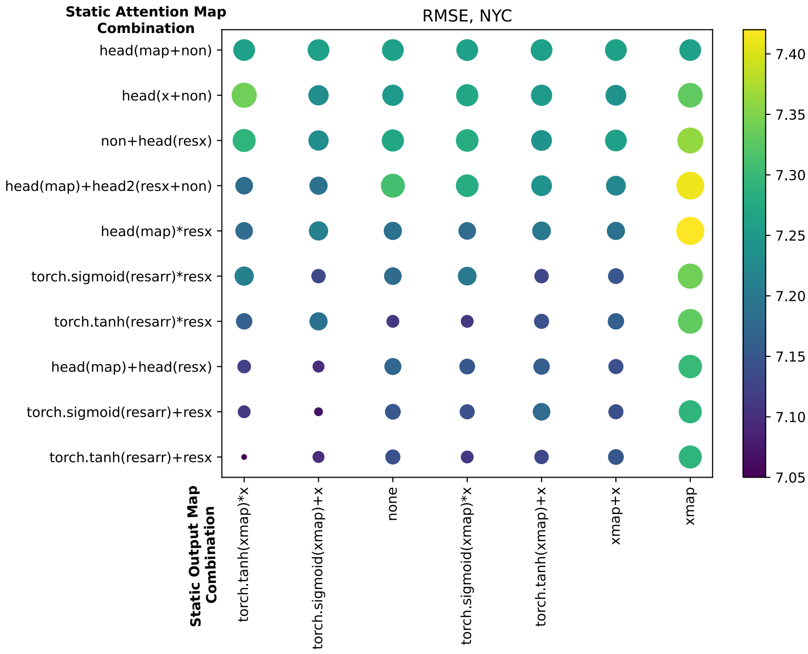



FIGURE 5.8: Root Mean Squared Error from performance evaluation of our XViT model for a number of combination operations on NYC data set

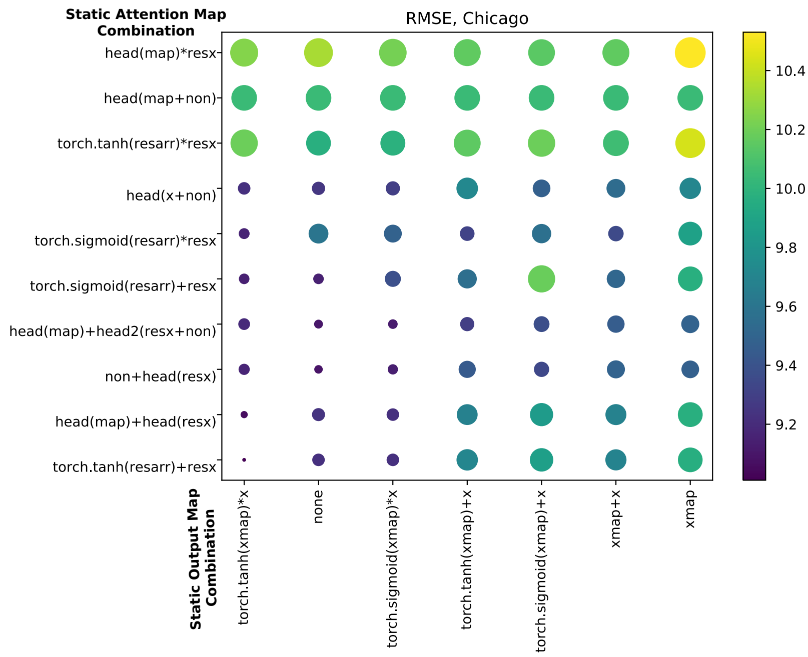



FIGURE 5.9: Root Mean Squared Error from performance evaluation of our XViT model for a number of combination operations on Chicago data set

(and may improve) the prediction performance. There is a non-linear dependence between RMSE and the scale factor observed for both data sets, which may suggest that the optimal scale factor for the accident risk map processing may exist and be different for each data set.

Overall, the use of visual transformers and its variations for traffic accident risk prediction outperforms previously used approaches. Further applications of image and video processing methods may provide further improved results and open alternative approaches for the task of accident risk prediction.

## Chapter 6

# Automatic Accident Detection, Segmentation and Duration Prediction using Machine Learning

## 6.1 Introduction

The number of vehicles has been substantially increasing during the past decades, which currently leads to an increase in the number of traffic accidents [175]. The National Highway Traffic Safety Administration (NHTSA) reported more than 5 million traffic accidents happening in the United States during year 2013 [4]. Traffic Managements Agencies usually rely on Traffic Incident Management Systems (TIMS) to collect data on traffic accidents, including information on various accident, traffic state and environmental conditions. Accurately predicting the total duration of an incident shortly after it is being finished, will help in improving the effectiveness of accident response by providing important information to decide the required resources to be allocated (response team size, equipment, traffic control measures) [111]. Traffic accident is a rare event with stochastic nature. The effect of the accident can be observed as an anomalous state in the time series of traffic flow [218].

**Challenges:** The traffic accident analysis may be a challenging task due to incorrect or incomplete accident reports, including the set and the quality of the accident characteristics that have been reported. Accident reports can contain user-input errors related to the accident duration such as: 1) an approximate reporting of accident's start and end time 2) reporting of the accident start time could have been done after the incident finished in reality 3) a 'placeholder' accident duration reporting (filling report with the approximate duration value due to unavailability of data by the moment of reporting). In our previous research [75], [160] we found that timeline-related errors are present in accident reports across three different data sets from both Australia and the United States of America, which creates the possibility of observing that such errors can occur in other data sets from around the world as well, due to multiple human and technical factors that can arise. To forecast the accident impact it is crucial to have an accurate and correct data regarding the observed disruption timeline. We emphasise that disruptions observed in a recorded traffic state can be automatically segmented and associated with a reported accident at the same time and location as when the accident occurred, which allows to eliminate user-input errors from reports and improve the accident duration prediction performance in many traffic management centres around the world. To help address this issue, in our paper we propose various methods for a correct traffic disruption segmentation, the method for an association between vehicle detector stations and accident reports.

Another important challenge is that many incident data sets around the world are private and not shared for public investigation; for those open data sets, there are several missing information fields, or even worse, incomplete information regarding the traffic conditions in the vicinity of the accidents. Even often publish crash data sets are limited in size as well and contain a very small number of records. This represents a tight constraint when testing one framework over multiple countries with different traffic rules and regulations. For our studies we have oriented our attention towards two big open data sets - CTADS (Countrywise traffic accident data set) which contains 1.5 million accident reports and the Caltrans Performance Measurement System (PeMS) which provides data on traffic flow, traffic occupancy and traffic speed across California. Despite both being extensive data sets, vehicle detector station readings from PeMS are not associated with traffic accident reports from CTADS either by time, location or coverage area. The lack of such association makes it impossible to analyse the relation between accidents and their effects on traffic flow and speed. To address this challenge, in our paper we introduce the following mapping algorithm which will secure several steps such as :

- an association of Vehicle Detection Stations (VDS) with reported accidents in their proximity,
- a segmentation of traffic speed disruptions from detector readings,
- an association of detector stations with reported accidents (we will further show that this step is necessary due to many detected user-input errors in accident reports).

As a result, we obtain traffic disruptions segmented by the traffic speed associated with reported accidents. This association makes it possible to perform various important tasks of the accident analysis: 1) prediction of the traffic accident impact on the traffic speed based on accident reports, 2) prediction of the traffic accident duration derived directly from the effect of disruption on the traffic speed (impact-based duration), 3) analysis of disruption propagation (each detected disruption can be studied for spatial-temporal impact within the traffic network). Through this work, we will focus on the prediction of the impact-based accident duration and lay the foundation for a further research.

Overall, the main contributions (summarised in Figure 6.1) of our paper are as follows:

1. We propose a fusion methodology of two large data sets (CTADS and PeMS) for a detailed traffic accident analysis. To the best of our knowledge, this is the first research study proposing the methodology for merging of these two large data sets, which allows an association between observed disruptions in traffic flow and the reported accidents.

The research of this nature (fusion of traffic flow and accident reports) has been performed before [70], [52], but our methodology has the following advantages: 1) Our disruption segmentation model can be fine-tuned via hyper-parameter search to find optimal disruption detection rate, 2) The method produces difference estimates proportional to the degree of observed disruption, which allows for control of false positives rate via threshold choice, 3) We evaluate multiple comparison metrics for traffic speed difference estimation, 4) The segmentation algorithm is more complicated and includes pre-processing convolution, test of multiple difference metrics, adjustment to selectivity and cyclic shift for difference window, 5) our methodology is modular, where each logical part can be further refined and studied in a separate research.

2. We propose a novel methodology for the disruption mining using a combination of different metrics (which we further find to have properties important for disruption segmentation): a) the Wesserstein metric, which allows us to measure the disruption severity and b) the Chebyshev metric, which provides a higher selectivity for the disruption mining and a rectangular shape of the disrupted segments, allowing an automated disruption segmentation. We detail all unique properties of both metrics utilized together to allow an accurate disruption segmentation.

3. We perform the estimation of traffic accident disruption duration from traffic speed via the above metrics which allows us to alleviate user-input errors in accident reports.

4. We evaluate multiple machine learning models by comparing both the reported and the estimated accident duration predictions extracted from traffic speed disruptions.

5. We introduce a new modelling approach which focuses on the amount and shape of the the disruption associated with an accident, which allows a further analysis and modelling of accident impact.

Overall, this research forms the foundation for a new early traffic accident disruption detection, traffic disruption speed impact analysis and the use of observed traffic accident durations for correcting errors in user reports. Moreover, this work contributes to our ongoing objective to build a real-time

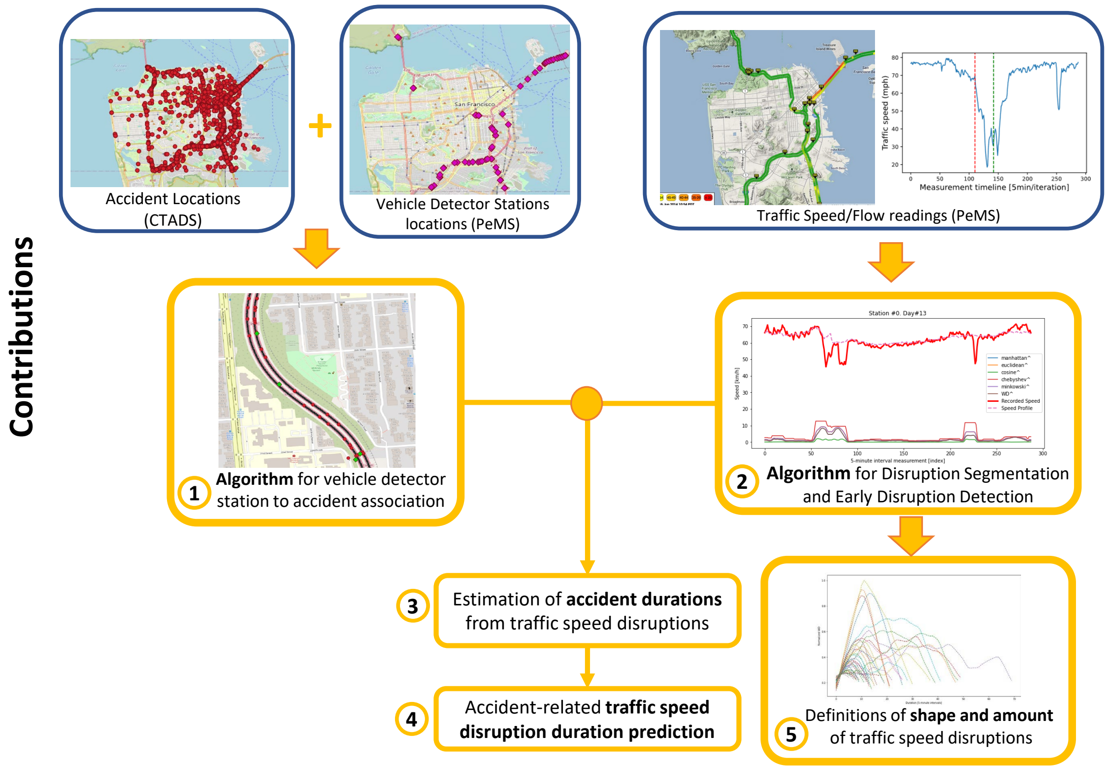



FIGURE 6.1: Contributions and data-flow schema for association of traffic speed readings with accident reports

platform for predicting traffic congestion and to evaluate the incident impact (see our previous works published in [160]-[204]-[154]).

The chapter is further organised as follows: Section 6.2 discusses related works, Section 6.3.1 presents the data sources available for this study, Section 6.3 showcases the methodology, Section 5.3.5 presents the disruption segmentation results, showcases the result of data set fusion and Section 6.7 provides conclusions and future perspectives.

## 6.2 Related Works

Multiple studies rely on user-input-based incident reports from Traffic Management Centers (TMC) with different machine learning models to predict the traffic incident duration [129]. The use of traffic flow features is found to be rare and mostly specific - incident detection and incident impact prediction by using traffic flow [61]. In other words, traffic flow data is rarely combined with actual incident reports since it requires a higher system complexity and extensive data collection.

There were numerous studies related to accident detection from traffic flow using anomaly detection techniques [185]. Various methods used for anomaly detection in time series are applicable for the task of traffic disruption detection. The ability to perform the detection of actual disruption, which should give us actual shapes of disruptions and time intervals allows in-depth analysis of usual accident statistics including the effect of the type of accident on the pattern of disruption in traffic flow. By integrating data on traffic state with accident reports we are able to further connect traffic flow disruption patterns to various accident characteristics (hour of the day, weather conditions, crash type, type of vehicle involved - truck/car [57], the effect of road pavement types [225], road design and road operation [253], etc).

Various machine learning models are used to solve the task of traffic accident duration prediction[129] including k-nearest neighbours (KNN) and Bayesian networks [116], Recursive Boltsman Machines and Support Vector Machines(SVM) [251] and Random Forests (RF)[83].

The definition of traffic incident duration phases is provided in the Highway Capacity Manual [11] and includes the following time-intervals: 1) incident detection - the time interval between the incident occurrence and its reporting, 2) incident response - time between the incident reporting and the arrival of the response team, 3) incident clearance time between the arrival of the response team and the clearance of the incident, 4) incident recovery - the time between the clearance of the incident and the return of traffic state to normal conditions. In this research, we rely on total incident duration - the time between incident occurrence and return of the state to normal conditions. Also, we analyse the subset of traffic incidents - traffic accidents. As we found during the data investigation, traffic accident duration is reported at the time when the incident is cleared by the response team, which doesn't include the duration of the effect that the accident produces on traffic flow. Traffic incident duration prediction studies rely on incident reports without emphasizing on the duration of observed incident effects. In this research we try to solve this issue by proposing the methodology for disruption segmentation from traffic speed.

Analysis of the effect of traffic incidents has been performed previously using Caltrans PeMS data, where the measure of incident impact was represented as a cumulative travel time delay [161], which is an aggregated value. However, traffic state recovery from disruptions is not necessarily following a single pattern - it may be slowly dissipating, we may observe secondary crashes, it may have a high or low impact, etc. Traffic accident duration prediction methodology relies on reported traffic accidents, but actual reports may contain user-input errors and be misaligned with the actual shape of disruption produced by the accident. Therefore, the approach for disruption segmentation may provide the accident duration estimated from the actual shape of disruption in traffic flow.

## 6.3 Methodology

The new proposed framework is represented in Figure 6.1 which we support across some initial definitions for our modeling approach (see next sub-section). First, we associate the road segments with their corresponding Vehicle Detector Stations (VDS) from the Caltrans PeMS data set, as well with the locations of reported accidents (see Algorithms 1 and 2 proposed in sub-section 6.3.4). The main outcome of this algorithm is that traffic accidents will get associated with the traffic flow, speed and occupancy readings from the VDS stations.

Second, we propose a new algorithm for early disruption detection and segmentation, detailed in sub-section 6.3.5. By detecting disruptions that occurred in time-space proximity of reported traffic accidents, we obtain the estimated traffic accident duration. This gives us much more information to include in the model training than just the simple accident duration: 1) the disruption shape in terms of modifications of speed data profiles from the standard patterns 2) the accident duration estimated from the impact on the traffic speed 3) the cumulative accident impact estimation.

### 6.3.1 Case study

Before diving into the methodology, we provide a brief introduction into the data sets in use for showcasing our approach, which helps establishing the modelling base and understanding of the steps taken. We make the observation that the current methodology can be applied on any incident and traffic state data set which can contain a time component, and is not bounded to the chosen data sets for exemplification.

#### CTADS: Accident reports data set

We rely on accident reports from the "Countrywide Traffic Accident Dataset" (CTADS), recently released in 2021 [165], [166], which contains 1.5 million accident reports collected for almost 4.5 years since March 2016, each report containing 49 features obtained from MapQuest and Bing services. We select the area of San-Francisco, U.S.A and extract data for 9,275 accidents (see Figure 6.2).

FIGURE 6.2: CTADS reported accidents for San-Francisco

The Countrywise Traffic Accident Data Set (CTADS) offers an insight into the extent of recorded accidents. Particularly the Bing data subset has start and end locations of accident extents, CTADS includes accident extents calculated using Havesine distance formula. Properties of extents may help to fine-tune our algorithm. We further rely on the Bing data subset to derive our conclusions regarding the accident extent properties. Figure 6.3 represents the distribution of these accidents' extents in a histogram, showing how often certain disruption extents occurred within the data. Additionally, Figure 6.4 depicts the empirical cumulative distribution function (ECDF) for the same data. The ECDF provides a complementary perspective to the histogram, showing the proportion of data points that fall below each value on the x-axis (e.g. around 90% of reported accidents have the road extent below 500m). A summary of important statistics, derived from the data, is provided in Table 6.1. This table includes key measures including the interquartile range (0.25 to 0.75 quantiles). We can safely choose 500m as the maximum accident extent for our association algorithm between vehicle detectors and accident points.

#### PeMS: Traffic speed and flow data set

We rely on Caltrans Performance Measurement System (PeMS) [33] to collect data on traffic flow and speed. This data set provides aggregated 5-minute measurements of traffic flow, speed and occupancy across California. We decided to extract the data for the area of San-Francisco (see Figure 6.5a), which contains 83 Vehicle Detection Stations (VDS) placed in that area (see 6.5b), and we try to associate

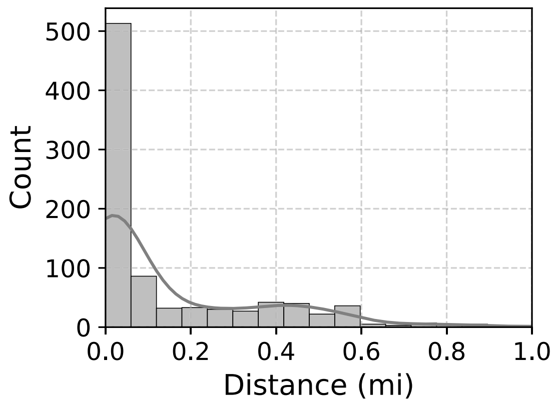



FIGURE 6.3: CTADS: Bing - Histogram for recorded accident extent

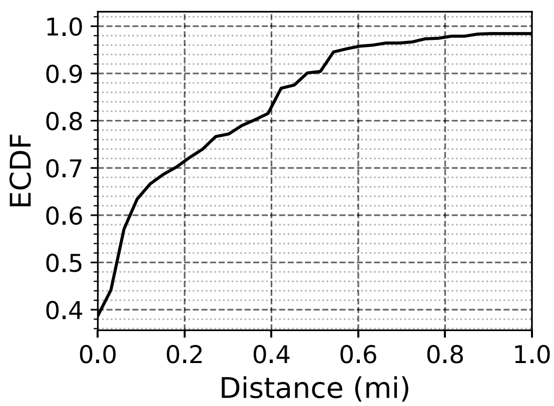



FIGURE 6.4: CTADS: Bing - ECDF for recorded accident extent

TABLE 6.1: Statistics on accident extent from CTADS (Bing part) data set

| Statistic | Value [$km$] |
|---|---|
| Mean | 0.16 |
| Median | 0.04 |
| 0.95 Quantile | 0.57 |
| 0.05 Quantile | 0 |
| Standard Deviation | 0.27 |
| Variance | 0.08 |
| Interquartile Range [0.25, 0.75] | 0.27 |

each traffic accident occurred with each of San-Francisco VDS in their 500m proximity using the algorithm detailed in the following section. In total, from 9,275 accidents in the area (extracted from CTADS) we have obtained 1,932 traffic incident reports which we were able to associate with the correct and complete traffic flow and speed readings from a VDS.

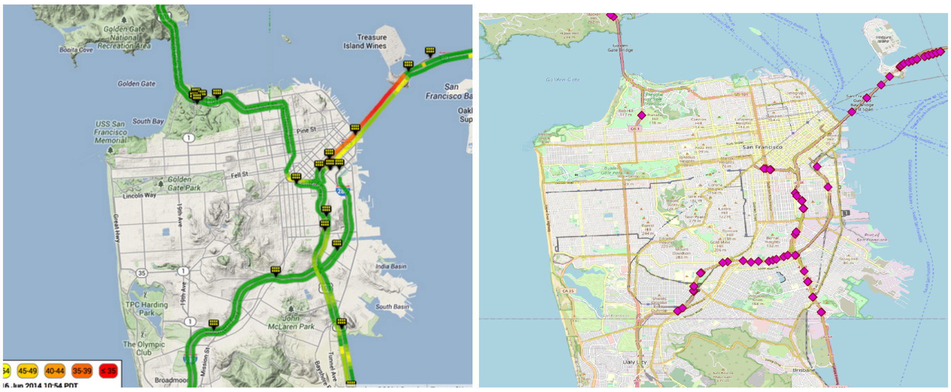


FIGURE 6.5: 1) PeMS data set area coverage for San-Francisco (the map is available at https://pems.dot.ca.gov/) 2) Mapping of the Vehicle Detection Stations from PeMS data set. OpenStreetMap excerpt showing San Francisco. Available at:https://www.openstreetmap.org/#map=12/37.7612/-122.4395

### 6.3.2 Speed difference estimation definitions

In the current study we compare the performance of multiple difference metrics that will help us to correctly estimate the impact of an accident and the deviation from the historical speed patterns. These metrics are defined as follows:

a) The Chebyshev difference is a measure of the maximum difference between corresponding elements of two one-dimensional vectors $u$ and $v$ and is expressed as:

$$D_{\text{Cheb}}(u, v) := \int \max_i(|u_i - v_i|) \tag{6.1}$$

The metric that best captures the concept of similarity in relation to the application at hand should be considered optimal, regardless of the model used for classification or regression. This hypothesis has been tested and validated across nine datasets and five prediction models [40]. This Chebyshev metric, commonly used in data analysis, has been found to outperform other metrics on a variety of tasks.

b) The Wasserstein difference, also known as the earth mover's distance, is a measure of the minimum "work" required to transform one probability distribution $u$ into another $v$. It is expressed as:

$$D_{\mathrm{WD}}(u,v) = \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| \mathrm{d}\pi(x,y) \tag{6.2}$$

This metric was introduced by Leonid Kantorovich in 1942 [105] and has found applications in fields such as computer vision, image processing, and natural language processing.

c) The cosine difference, also known as the cosine similarity, is a measure of the similarity between two one-dimensional vectors $u$ and $v$. It is expressed as:

$$D_{\mathrm{C}}(u,v) = \frac{u \cdot v}{|u|_2 |v|_2}. \tag{6.3}$$

This metric is commonly used in information retrieval and has also found applications in recommender systems and document clustering [197].

d) The Euclidean difference is a measure of the distance between two one-dimensional arrays $u$ and $v$ in a Euclidean space. It is expressed as:

$$D_{\mathrm{E}}(u,v) = \left(\sum (w_i |(u_i - v_i)|^2)\right)^{1/2} \tag{6.4}$$

This metric is commonly used in fields such as machine learning, computer vision, and signal processing.

e) The Minkowski difference is a generalization of the Euclidean difference and is a measure of the distance between two one-dimensional arrays $u$ and $v$ in a Minkowski space. It is expressed as:

$$D_{\mathrm{M}}(u,v) = (\sum |u_i - v_i|^p)^{1/p}. \left(\sum w_i(|(u_i - v_i)|^p)\right)^{1/p}. \tag{6.5}$$

This metric is a generalization of other distance metrics, such as the Manhattan distance (when $p = 1$) and the Euclidean distance (when $p = 2$), and is commonly used in fields such as physics, engineering, and data science [162].

f) The Bray-Curtis difference metric [24] between two vectors $\mathbf{u}$ and $\mathbf{v}$ is given by:

$$D_{\mathrm{BC}}(\mathbf{u},\mathbf{v}) = \frac{\sum_{i=1}^{n} |u_i - v_i|}{\sum_{i=1}^{n} (u_i + v_i)}, \tag{6.6}$$

where $n$ is the number of dimensions in the vectors.

g) The Canberra difference metric [101] between two vectors $\mathbf{u}$ and $\mathbf{v}$ is given by:

$$D_{\text{Can}}(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{n} \frac{|u_i - v_i|}{|u_i| + |v_i|}, \tag{6.7}$$

where $n$ is the number of dimensions in the vectors.

### 6.3.3 Accident duration prediction task definitions

Using all available data sets and the incident information, we first denote the matrix of traffic incident features as:

$$X = [x_{ij}]_{i=1..N_i}^{j=1..N_f} \tag{6.8}$$

where $N_i$ is the total number of traffic incident records used in our modelling and $N_f$ is the total number of features characterising the incident (accident severity, vehicles involved, number of lanes, etc) according to the accident report data set.

Traffic Speed represented as a vector with 5-minute averaged readings from Vehicle Detector Stations:

$$S = [s_i]_{i=1..N} \tag{6.9}$$

where $N$ is the total amount of traffic speed readings.

Within this research we assess the performance of Machine Learning models on tasks of predicting reported and estimated accident duration. We define the task of accident duration prediction as a regression problem.

The incident duration regression vector ($Y_r$) is represented as:

$$Y_r = [y_i^r]_{i \in 1..N},\ y_i^r \in \mathbb{N} \tag{6.10}$$

and the regression task is to predict the traffic accident duration $y_i^r$ based on the traffic incident features $x_{i,j}$. The regression models go via an 10-fold cross-validation procedure with hyper-parameter tuning.

The estimate the accident duration prediction performance we use the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2} \tag{6.11}$$

where $A_t$ - actual value, $F_t$ - predicted value.

### 6.3.4 Algorithm for vehicle detector station to accident association

In order to match correctly what traffic conditions reflect best the effects of each incident, we further define the association procedure between traffic accidents and VDS (Accident-to-VDS), for the San Francisco area. We observe that only a few traffic accidents have VDS stations in their proximity to allow a good traffic speed and flow extraction, as shown in Figures 6.2 and 6.5.

In order to find the traffic incidents for which we can have traffic flow and speed data, we develop a mapping algorithm (Accident-to-VDS) which consists of two parts (see Algorithm 2-3), defined by the following steps:

1. We extract primary and secondary road lines from Open Street Map.
2. Road segments are then transformed into points at 2-meters equal distance.
3. Each VDS station and accident are mapped to the closest road point (up to 10m distance).
4. From this step we use the following algorithm to process the point-based representation of VDS, accidents and road segments (see Algorithm 2). The vdsPoints array contains tuple of form (VDS ID, x and y coordinates), each point in accidentPoints contains an array visitedBy (initialized to be empty) to maintain a list of stations in proximity of the accident and assignedVDS as a resulting nearest VDS station to the accident along the road.

---

**Algorithm 2:** Accident-to-VDS: Accident to VDS mapping algorithm

```
Input: point
Output: None
Access global arrays: roadPoints, accidentPoints, vdsPoints
Function visitNearestPoints(VDSID, point, currentHops)
accidents := findNearestAccidents(point, accidentPoints, 10m)
for i = 0 to length(accidents) do
    a := accidents[i]
    a.visitedBy.append([VDSID, currenthops]) ;        //Recording visits from stations to
     internal accident list
end
if currenthops < 500/2 then
    ;                                                 //Limiting the travel distance from VDS
    roadpoints:=findNearestRoadPoints(point, roadPoints, 3m) for i = 0 to length(roadpoints) do
        rp := roadpoints[i]
        if VDSID not in rp.visitedBy then
            ;                                         //Preventing the infinite recursion
            rp.visitedBy.append(VDSID) visitNearestPoints(point, currentHops + 1)
        end
    end
else
    Return
end
```

---

The algorithm relies on a recursive function to implement the process of visiting road points (see Algorithm 3). The association part of the algorithm works as follows:

1. We select the current VDS station.
2. We move (jump by points) in all possible directions available from the starting and forthcoming points in a 3m radius. This radius allows us to move along the road jumping between road points. Movement in all possible directions allows to grasp the propagation of the traffic congestion associated with the accident. The maximum available distance is set to 500m (250 jumps) and allows to limit the observable impact distance.

3. By moving across points we collect traffic incidents in the 5m proximity of each point and associate them with the current VDS station.

---

**Algorithm 3:** The recursive function for traveling across road points

---

```
Input: roadPoints, accidentPoints, vdsPoints
Output: assignedAccidents
for i := 0 to length(vdsPoints) do
    vds: = vdsPoints[i]
    visitNearestPoints(vds, 0)
end
assignedAccidents = []
for i := 0 to length(accidentPoints) do
    accident := accidentPoints[i]
    if length(accident.visitedBy) > 0 then
        accident.assignedVDS = sort(accidetn.visitedBy, sortvalue = hops)[0] ;    //Choosing
         closes VDS station
        assignedAccidents.append(accident)
    end
end
return assignedAccdents
```

---

The algorithm is recursive and relies on the list of visited points for each VDS. At the end of the algorithm, we have a subset of traffic accidents with their associated VDS which allows us to extract the traffic flow and speed in the vicinity of the accident. Ideally, all traffic accidents should have associated traffic flow but given their unavailability (due to detector coverage), we select accident reports which have associated traffic flow information currently available from the PeMS data set.

### 6.3.5 Algorithm for automated disruption segmentation (ADS)

Once the accidents have been mapped and associated to their VDS stations which allows us to select the flow/speed that match the day of the incident, etc, we are using the extracted traffic state parameters to propose a new automated disruption segmentation (ADS) method. The algorithm for the segmentation of disruptions via traffic speed works as follows:

1. A time series pre-processing step prepares all the data for segmentation (see Alg. 4):
   (a) Calculate the average monthly profile for daily traffic speed measurements;
   (b) Iterate over the traffic speed time series using a moving window of 1-hour time interval (in total there are twelve measurements of 5-minute each)
   (c) On each iteration perform a comparison of a 12-unit window between the monthly profile and the current day of measurements. The resulting single value is added to the resulting time series sequence.
   (d) Calculated the time series differences (TS) choosing the above defined metrics will be then adjusted by selectivity (using the power function, which will keep values closer to one for the least affected by the function and minor values the most suppressed) and normalized to produce nTS and pTS arrays respectively.
2. The time-series segmentation step (see Algorithm 5):

(a) A first order derivative (dTS) is calculated for the resulting time series of the previous stage (nTS), which returns positive peaks when entering the disruption and negative peaks when exiting the disruption state.

(b) We iteration over resulting derivative time series to record the opening and closing of each disruption in each time series. If two consecutive positive peaks (opening times) are observed then we choose the largest one between the two (we will further debate on this aspect in our future work plans). We repeat the same for consecutive negative peaks.

(c) We then associate the detected disruptions with the accident reports: for each accident report, we extract the traffic speed time series on the day of the accident and if both opening and closing times are recorded, we perform an association of the accident with these times and extract the actual time series sequence for further analysis.

**Enhancing selectivity:** We use the convolution with the kernel (1,1,1), which attributes to the morphological dilation operation, to facilitate the work of the segmentation algorithm. By applying this convolution we make multiple consequent differences to be accumulated ; for example, assuming we have a sequence of 0.3, 0.1, 0.1, 0.2 and 0.2 as differences for each 5-minute step, therefore a total of 0.9 change over 4 iterations. The convolution (1, 1, 1) will produce the values of 0.5, 0.4, and 0.5 by making a sequence of high values from the sequence of small changes (see Figure 6.6). The dilation operation is primarily used in computer vision tasks to make connected groups from closely placed scattered points to facilitate a further image analysis.

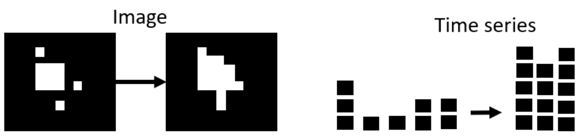



FIGURE 6.6: The application of dilation operation to an image and time series

To obtain the monthly profile, the traffic speed measurement sequence was obtained for a duration of 1 month from the VDS before the accident occurred, and was done separately for each accident. This sequence then gets reshaped into a matrix of the form $[number\_of\_days; 288]$, where columns contain the total number of measurements across an entire day (24*12=288). The monthly average was then calculated across axis 1 (number of days) to obtain a vector with 288 values of measurements. This vector gets recalculated for a number of days of observations from each detector to be comparable with the VDS daily measurements.

As an observation, the constants *pThreshold* and *nThreshold* represent thresholds for change that observed in the time series of the metric derivative; they allow us to define a positive and negative change of the difference metric, the selectivity defines power function coefficient to suppress the non-significant and filter the most significant disruptions.

### 6.3.6 Modification of the algorithm for automated real-time early disruption detection

Since our proposed algorithm doesn't look into the future and calculates different metrics based on the currently observed traffic speed and a few measurements in the past (11 units in the current study, equivalent to 55 minutes from the past), we can perform an early accident detection which will consist

**Algorithm 4:** Algorithm for automated disruption segmentation. Part 1

**Input:** $monthlyProfile, speedReadings, selectivity, shift$
**Output:** $cTS$
; `//Accidents array contains a day number, starting and ending index for segmented traffic disruptions`
$step := 1$
$windowsize := 12$
$i := windowsize$
$lastDiff = 0$
$DS = []$
**while** $i < length(speedReadings)$ **do**
  $A := speedReadings[i - windowsize : i]$ ; `//Look-back window of readings`
  $B := monthlyProfile[i - windowsize : i]$
  $diff := metric(A, B)$
  $DS.append(diff)$
  $lastDiff = diff$
**end**
**for** $i = 0$ *to windowsize* **do**
  ; `//Padding array with the latest observed value to obtain full-day readings`
  $DS.append(lastDiff)$
**end**
$pTS = power(TS, selectivity)$ ; `//The use of power function to improve selectivity of significant disruptions`
$nTS = cyclic_shift(shift)$ ; `//The use of cyclic shift operation`
$nTS = normalize(pTS)$
$dTS = derivative(nTS)$ ; `//First order derivative allows to decompose metric results into positive and negative change to the disruption amount`
$cTS = convolution(dTS, [1, 1, 1])$
**return** $cTS$

in calculating and comparing the first-order differential (FOD) of Chebyshev metric based on the monthly profile. The detection of significant positive peaks (e.g. 0.3-0.5 of normalized difference metric) can identify the amount of disruption in real-time. The end of the disruption can be detected using the same approach in real-time as well by observing a significant negative peak.

## 6.4 Results

### 6.4.1 Data exploration and setup

CTADS data set contains traffic accident reports, which after an initial data mining investigation, we found to contain several user-input errors; for example, a lot of traffic accident durations have been rounded to 30 or 360 minutes (see Fig. 6.7d)); or the incident start time which was reported is unrelated to any disruptions observed by the vehicle detector stations in the proximity - see Figure 6.7 in which we have provided two different examples of speed recorded during two different accidents A-5198 and A-4490; the red lines indicate the official reported start and end time of the accidents, while in reality the accidents have had a long lag in spreading across the network - see Fig. 6.7a) or were reported much later that the official speed drop was recorded - see Fig. 6.7b).

At this step we observed a significant amount of user-input errors in accident reports, which affect the accident duration/impact analysis: 1) accidents can be reported earlier or later than its occurrence (observable disruption misalignment in time) 2) a report can be filled with "placeholder" duration

**Algorithm 5:** Algorithm for automated disruption segmentation. Part 2

```
Input: cTS, pThreshold, nThreshold, selectivity
Output: Accidents
;     //Accidents array contains a day number, starting and ending index for segmented
      traffic disruptions
state := 0
Accidents = []
for i := 0 to length(cTS do
    if cTS[i] > pThreshold then
        ;                //Significant positive peak identifies the start of disruption
        if state <> +1 then
            state = +1
            enteridx = i
        else
            if cTS[i] > cTS[enteridx] then enteridx = i;
            ;                //Choosing the largest change from previously observed
        end
    end
    if cTS[i] < nThreshold then
        ;                //Significant negative peak identifies the end of disruption
        if state <> -1 then
            state = -1 exitidx = i
        else
            if cTS[i] < cTS[enteridx] then exitidx = i;
        end
    end
    if i mod 288 == 0 and i > 0 then
        ;                //Reset segmentation procedure at the end of each day
        state = 0
        Accident.append([i div 288, enteridx, exitidx])
    end
end
return Accidents
```

values not representing the actual accident duration 3) there may be no observable disruption in traffic speed despite the accident report (due to placement and management of the accident) (false positive) 4) there may be accident-related traffic disruptions not grasped by accident reports (false negative). Therefore, incorrect accident start time, duration and end time, unreported presence or abscence of disruption make it necessary to estimate accident duration characteristics from traffic state data instead of relying on user reports. In this chapter our proposed methodology is really meant to solve the user-reporting issues related to traffic accidents and to be applied automatically on any data set, regardless of its nature or geo-location.

The use of PeMS data set allows to estimate the impact of accidents on the traffic states (flow, speed). For our scenarios, we choose the area of San-Francisco with accidents recorded from 2016 to 2020 in the CTADS data set. We then obtain Vehicle Detector Station locations from PeMS, the road network shape from OpenStreetMap and we perform an association of CTADS accident reports with VDS stations along the road within 500m proximity. We then try to segment the disruption time interval occurred on the day of an accident. Further, we associate observed disruptions in the traffic speed series with actual accident reports. The purpose of this step is to reduce user-input errors in accident reports and to enhance the modelling of traffic disruptions with an analysis of traffic speed.

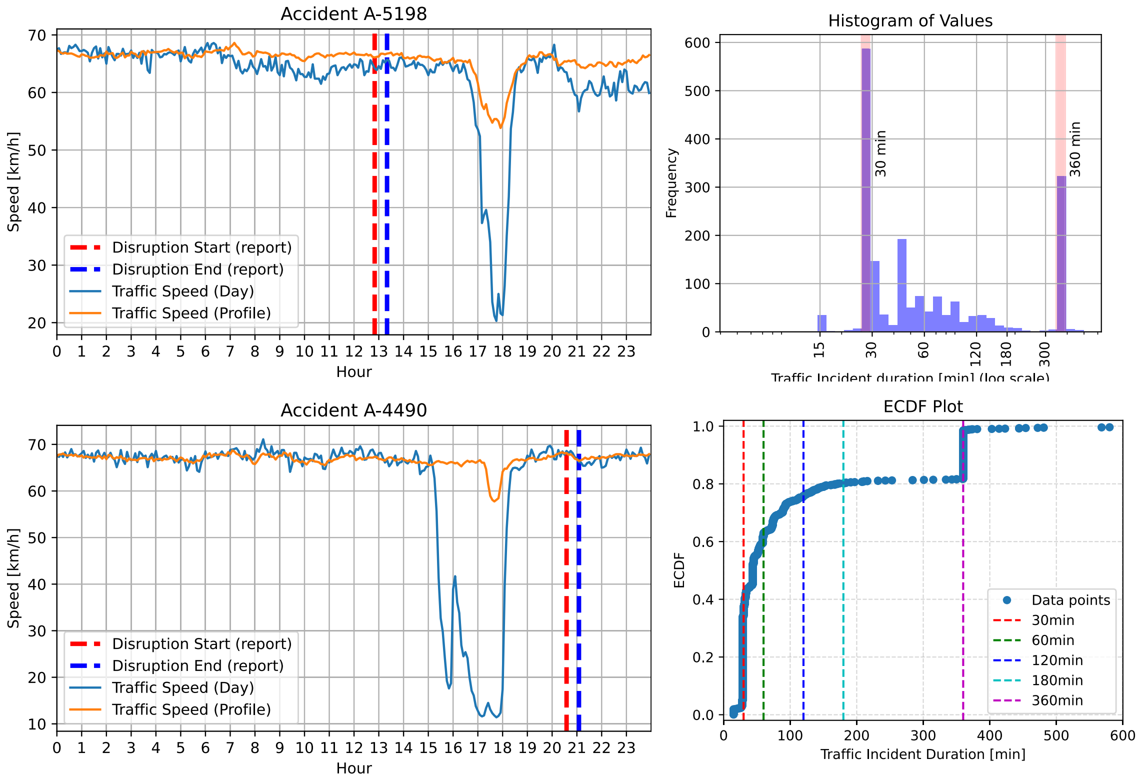



FIGURE 6.7: User input errors located within the CTADS data set

### 6.4.2 Metric performance comparison

We apply the difference metrics detailed earlier in Section 6.3 to a monthly traffic speed/flow profile (monthly readings averaged to one day) and reading on the day of the traffic accident. There are two approaches to applying the difference calculation: 1) a global difference - when we try to find the difference between the monthly profile and traffic flow/speed readings on the day of the accident; the global approach is too broad and will not allow the actual comparison between disruptions localized in time (metric results can be very similar between the very long subtle disruption and abrupt but impactful one). We measure the amount of difference that occurred within a moving time window (we choose twelve 5-minute time intervals equivalent to one hour). Traffic speed/flow readings from the moving window are taken right before the currently observed value to ensure that the difference estimation algorithm is not looking into the future.

To compare the metric performances we provide an example of speed readings from one of the detector stations. Each difference metric demonstrates its specifics as represented in Figure 6.8: 1) the Chebyshev metric, which we define as the maximum difference between the monthly profile and the observed readings, produces a noticeably rectangular shape and demonstrates a higher selectivity towards major disruptions than other metrics; the Chebyshev metric will be further used for the automated accident segmentation; 2) the use of Cosine metric allows to detect the change in the traffic state - speed decrease and increase both represented as positive peak values, 3) the Wasserstein difference allows for smooth representation of the amount of disruption (conceptually, it measures the amount of work necessary to change one shape into another, which we can rephrase as the amount of work produced by an accident to deviate the traffic state from the normal operation), 4) the Minkowsky, Euclidean and Manhattan difference metrics show little to no difference to the Wesserstein distance; we choose to use the Wasserstein difference since its connection to physical interpretation.

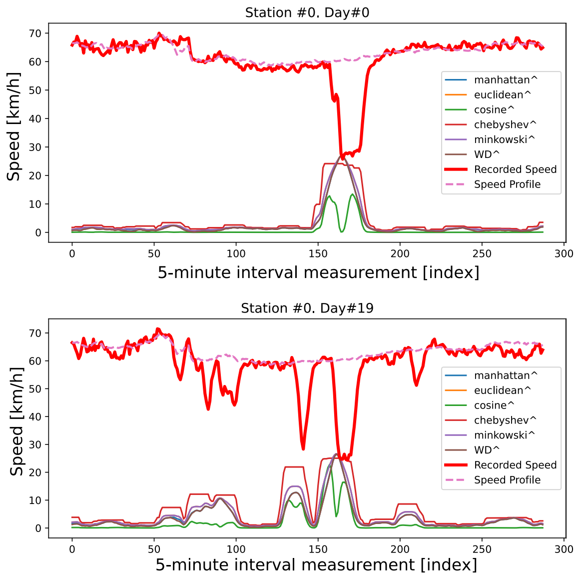



FIGURE 6.8: Various metrics applied to difference between recorded speed and speed profile

Examples of applying our proposed algorithm are presented on Figure 6.9. The 'Disruption Start' and 'Disruption End', which are represented as dashed blue and red vertical lines correspondingly show a reported accident timeline. The 'Day' (blue line) represents the traffic speed on the day of the incident and 'Profile' shows the average speed for every 5-minute interval across 14 days of measurements. Application of the 'Wasserstein distance (WD)' shows a gradual measurement of the observed disruption, while the 'Chebyshev' metric shows the nearly rectangular outline of a time interval where disruption is observed. This 'rectangular' result of the 'Chebyshev' metric was the main consideration for the development of the presented algorithm. One of the main observations from the figures as well as from the procedure of manual markup was that accidents were primarily reported 1-2 hours after the return of traffic state to normal conditions (which we define as the end of disruption). Other observation is that accident timeline is often misreported as a 'rounded' value of either 30 or 360 minutes. Application of both metrics shows a clear outline of disruption shape observed in traffic speed.

### 6.4.3 Combination of our proposed methodology with modern methods for accident scene segmentation

Image segmentation methods can be utilized to output a degree to which an accident is observed in an image [263], ultimately helping to create an accident timeline. By leveraging the power of semantic segmentation, the method can quantify the extent of the accident by assigning scores or probabilities to different elements within the scene. Here's how this can be done: 1) Accident-related object detection (Spatial analysis): Semantic segmentation can identify accident-related objects in the scene, such as damaged vehicles, debris, or injured pedestrians; by calculating the proportion of these objects within the segmented image, it is possible to assign a degree or score that represents the severity or extent of the accident at that specific moment, 2) Temporal analysis: by analyzing the segmented images over

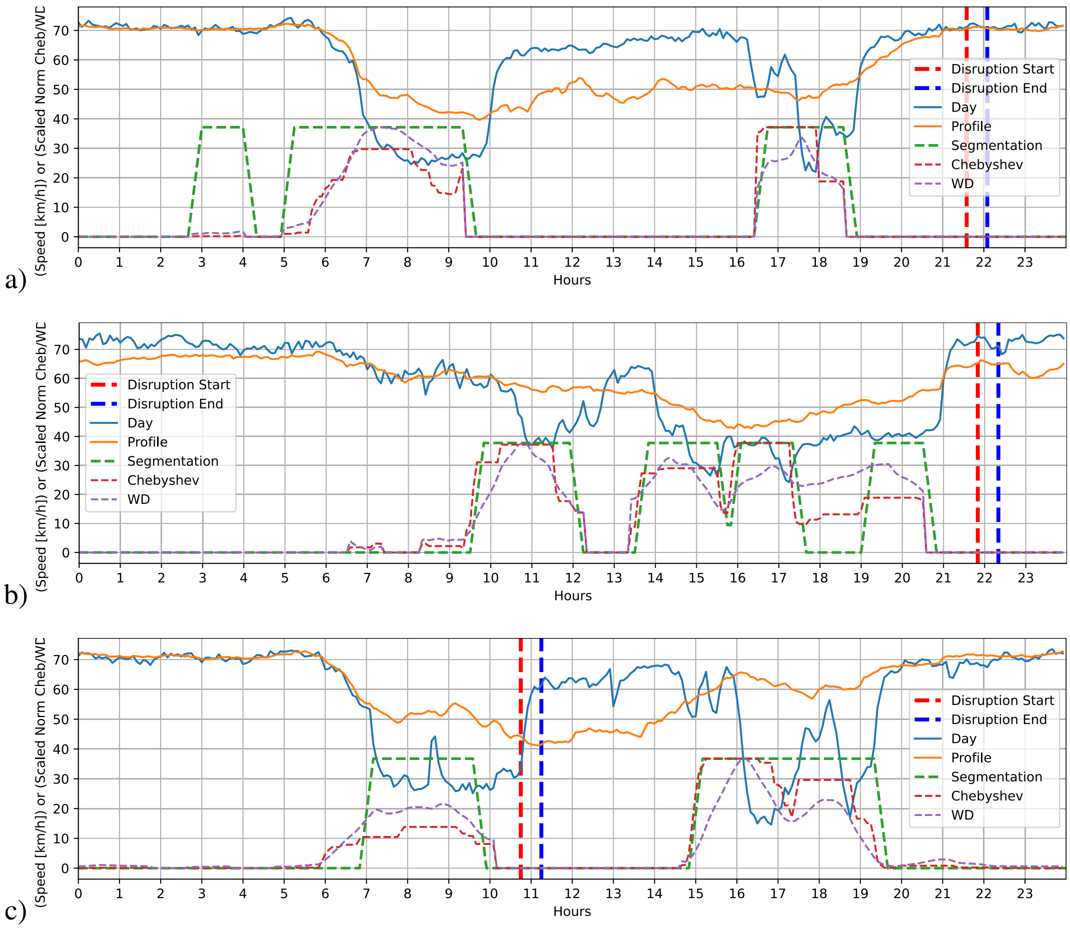



FIGURE 6.9: Results disruption segmentation algorithm application for accidents a) A-5764, b) A-8119, c) A-9931

time, we can track changes in the accident scene, such as the motion of vehicles or the appearance of new accident-related elements. This enables the creation of a timeline that reflects the progression of the accident and the associated changes in the severity or extent of the event, 3) Probability-based analysis: advanced segmentation methods can output probability maps that indicate the likelihood of each pixel belonging to a specific class or label; by analyzing these probability maps, it is possible to compute a score that represents the degree of accident occurrence within the scene over the timeline, 4) Accident phase classification: The degree to which an accident is observed can also be used to classify distinct phases of the accident, such as pre-collision, impact, and post-collision. By evaluating the changes in accident-related object proportions or scores over time, the segmentation method can identify critical moments or transitions between different accident phases. This information can be used to construct a detailed accident timeline that highlights the key events and their corresponding degrees of severity.

There is potential to connect our proposed methodology with accident scene segmentation research approaches to create a more comprehensive and accurate framework for analyzing traffic accidents and predicting disruption durations. Here's how the two research approaches can be integrated: 1) Improved incident duration prediction: The segmentation output from the first research can be used as input for the early detection and disruption segmentation algorithm in the second research. This would allow for a more accurate identification of critical moments in the accident timeline and better prediction of incident durations, 2) Integration of mathematical metrics: the Wasserstein and Chebyshev metrics proposed in the second research can be used to refine the segmentation results obtained from the accident scene segmentation over timeline. This would help to improve the performance of the accident scene segmentation and contribute to a more accurate incident duration prediction, 3) Joint machine learning model: The speed disruption segmentation from our research can be combined with the semantic segmentation methods to create a joint machine learning model. This integrated model could leverage both the event-driven dynamic context and the mathematical metrics for segmentation to improve its predictions for incident duration and severity. By connecting these two research approaches, a more comprehensive framework for analyzing traffic accidents and predicting disruption durations can be developed. This integrated approach would benefit from the strengths of both methods, enabling more accurate and reliable predictions for incident durations. Ultimately, this could lead to improvements in road safety, emergency response, and traffic management.

In conclusion, image segmentation methods can be employed to not only segment the accident scene but also to quantify the degree to which an accident is observed in an image. This information can be used to create an accident timeline that reflects the progression of the accident, the severity of the event, and the critical moments when interventions or safety measures could have been taken. This approach in combination with our proposed disruption segmentation method can potentially contribute to better accident analysis, road safety improvements, and more effective emergency response strategies.

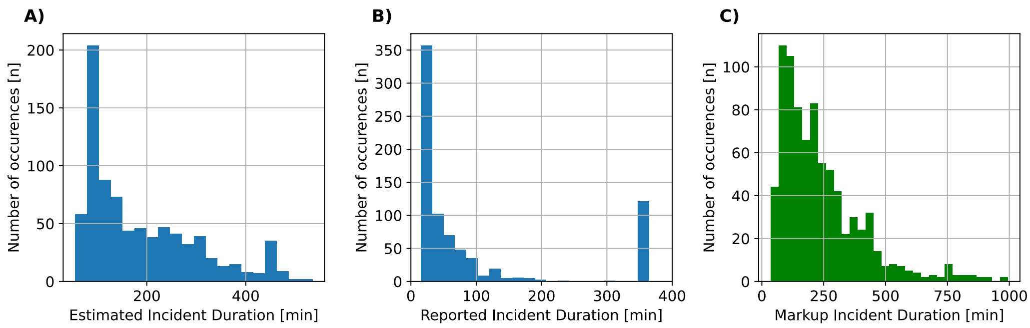
### 6.4.4 Automated disruption segmentation results

Figure 6.9 presents the results obtained from our algorithm for the automated disruption segmentation. The segmentation line (dotted blue) represents the estimated disruption intervals represented as 0 and 1 to perform our visualisation investigation better. Figure 6.9a) shows that there may be multiple observed disruptions in a 300*5=1500 time interval. Due to errors in accident reports regarding the starting time and the duration of the accident, it is non-trivial to determine which disruption is associated with the accident. The situation may be easier in the case when only one disruption is observed during the day. According to our algorithm, we select the largest disruption on the day the accident was reported. Figures 6.9b) and 6.9c) highlight additional specific situations which need to be considered: 1) higher traffic speed at the end of the day than observed from the monthly profile, 2) unstable traffic speed approaching normal traffic conditions with high frequency, 3) slight misalignment of disruption intervals with the visually observed disruption intervals. All these problems can be addressed by using manual segmentation with deployment of Deep Learning models since there are advanced computer vision methods proposed in recent years (e.g. autoencoders for segmentation).

### 6.4.5 Comparison of estimated, reported and manual markup of accident durations

There is a significant difference between the estimated and the reported accident durations that we would like to highlight: 1) the reported accident durations contain a large amount of 30 and 360 minutes duration values (nearly 40% of data - see Figure 6.10a)) while the estimated accident durations using our approach have an average duration of 58 minutes, while the reported is 108 minutes (which is by assumption skewed due to 360 placeholder values), 3) the estimated accident durations are distributed between 90 and 355 minutes (0.10 and 0.90 quantiles correspondingly) (see 6.11b)), while the reported durations are distributed between 29 and 360 minutes (see 6.11a) and manually detected disruptions distributed between 75 and 440 minutes), which highlights that disruptions observed from traffic speed are much shorter than reported in the original data set, 4) There is no noticeable correlation between observed and reported durations with high amount of horizontal anomalies in reported accident durations (see Figure 6.11). Traffic accident duration is most common to follow log-normal or log-logistic distribution [129] and on resulting plots, we see that accident reports are found to represent log-normal distribution to less extent than manual markup or estimated accident duration.

To perform the ablation study, we perform a manual markup of disruptions observed in traffic speed for 800 accidents, which will be discussed in the corresponding section.




a)

FIGURE 6.10: Distribution of accident durations for a) estimated, b) reported accident durations for the area of San Francisco, c) results of manual markup of disruptions observed in traffic speed

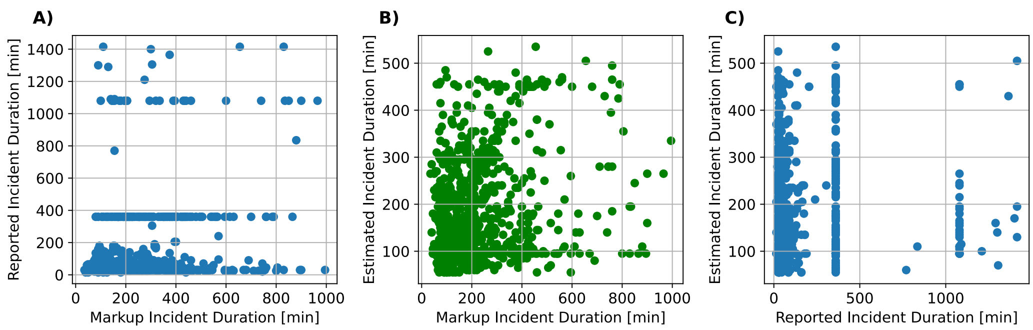



FIGURE 6.11: Scatter plot for a) estimated and b) reported accident durations for the area of San Francisco, c) results of manual markup of disruptions observed in traffic speed

### 6.4.6 Extraction of disruption shapes

In previous subsections we applied a Chebyshev metric to perform segmentation of disruptions. To analyse the disruption impact we apply the Wasserstein difference between monthly speed profile and daily traffic speeds and extract the corresponding disruption intervals. Wasserstein difference, originally named an Earth Mover distance, has an intuitive physical interpretation - the minimum "cost" of altering one pile of earth into the other, which is assumed to be the amount of earth that needs to be moved times the mean distance it has to be moved. In application to traffic state, it is the minimum amount of work necessary to alter the traffic state to disrupted condition, or in other words - the amount of disruption. We compare normalized metric values since every at every vehicle detector station there is a different average traffic speed. As in our proposed algorithm, we use a 12-units moving window (one hour) to estimate the Wasserstein difference between traffic speed measurements and provide the plot for the first 40 segmented disruptions, which allows for shape analysis of traffic disruption amount (see Figure 6.12): 1) We observe the similarity between multiple disruptions - they have a 'hill' shape, 2) there are secondary (double 'hill') and long-lasting disruptions. The observed shapes can be defined through the parametric equation to perform the classification of disruption effects and facilitate the prediction of disruption impact timeline since we observe that high-peak fast-ascending disruptions have a probability to end sooner than slowly ascending ones. The analysis of the speed of ascendance has potential to perform the early classification of disruptions, which is planned for further research.

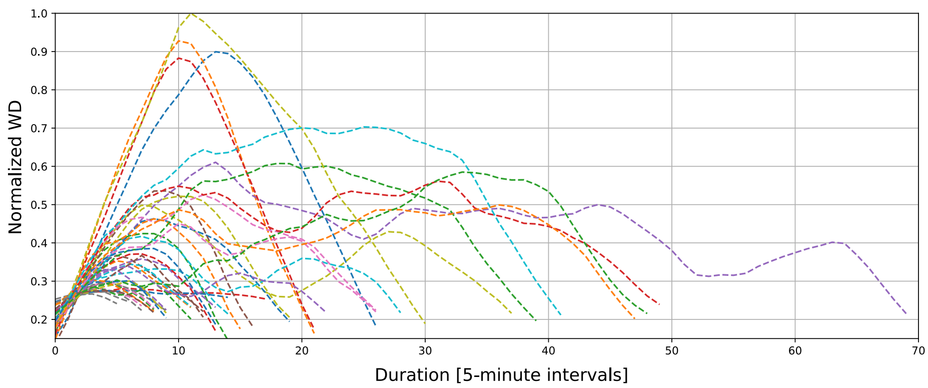



FIGURE 6.12: Normalized Wasserstein distance plot for disruption shapes extracted for segmented intervals

### 6.4.7 Accident duration prediction

We further compare a regression model prediction performance on the CTADS data set by using on the training data set both our estimated versus the reported accident durations. We report results of a 10-fold cross-validation over 820 accident reports for which we performed a Vehicle Detector Station association and manual markup of traffic disruptions from traffic speed for ablation study. Firstly, we need to consider that the performance using reported durations from CTADS can be affected because of the presence of user-input errors in the form of placeholder values. Secondly, the nature of estimated accident durations is different since accident response teams usually report the end of the accident at the moment they finished the accident clearence, without estimating the time for the traffic to return to a normal condition, which would require additional presence, calculations and access to measurements.

We have further extended the current results by adding newtables with several machine learning models on the task of predicting a target variable.

Table 6.2 shows the Mean Absolute Error (MAE) results. The model with the lowest MAE is the CatBoost model, with an estimated MAE of 17.55, followed by the Ridge Regression with an estimated MAE of 17.87. The highest MAE is reported by the Linear Regression model (76.76). The CatBoost model outperforms all the other models by a significant margin, with the next best model (Ridge Regression) having an estimated MAE that is only slightly lower.

Table 6.3 shows the Root Mean Squared Error (RMSE) results. Here, the CatBoost model also has the lowest RMSE, with an estimated value of 22.55. The next best model is the Ridge Regression with an estimated RMSE of 22.21. The highest RMSE is reported by the SVM model, with an estimated value of 208.29. As with the MAE results, the CatBoost model outperforms all the other models by a significant margin. All the methods use default parameters as they are presented in Scikit-learn [186] and corresponding modules.

When we are using accident reports to predict the estimated accident duration, we obtain a better performance using the RMSE metric across all the regression models, which may be connected to the lower amount of long accident durations than reported.

These best-performing models are all complex tree methods, which utilize multiple learners (via ensembles and boosting) to gain better predictive performance. They work well with mixed types of data (numeric and categorical), can capture non-linear relationships, and are less prone to overfitting. On the contrary, Linear Regression assumes a linear relationship between the input variables and the single output variable, KNN assumes that similar instances are near to each other, and SVM assumes that the data is linearly separable by a hyperplane in a feature space. Low performance of these methods shows that these assumptions may not align well with the data in case of traffic accident reports.

The reported duration, as provided directly from the source or via some other form of direct measurement is subject to more variability due to factors such as measurement errors (incorrectly reported duration), reporting biases ("rounded" 30 and 360 minute durations), or other uncontrolled external influences (late accident detection, disruption effects misaligned to reported accident timeline). We expect that correct estimation of the incident duration contributes to reduction in modelling complexity due to reduced effect of outliers, bias and errors on prediction performance.

In contrast, the manual and estimated durations are derived using more controlled processes and algorithms. The manual duration calculated by a consistent procedure, minimizing the room for error. The estimated duration, relies on parametric model, would also tend to have less variation due to the model fine-tuning to minimize prediction error based on the available data.

Overall, the CatBoost model consistently outperforms all the other models across all metrics.

## 6.5 Ablation study

In this chapter, we propose using the F1 score to estimate the quality of time interval segmentation in binary time series (see Figure 6.13) in which we provide two different examples of different stations with both manual markups of the incidents - red markups- and our segmentation algorithms - blue markups- that is more efficient at detecting multiple incidents throughout the 24h time period and not only one single isolated event. The value on Y-axis shows a positive 1.0 value if the interval contains the disruption. Examples are provided for Accidents with ID A-1024015 and A-1034382 from CTADS data set.

TABLE 6.2: Mean Absolute Error (MAE) Results

| Model | Reported | Manual | Estimated |
|---|---|---|---|
| KNN [116] | 44.11 | 26.22 | 19.73 |
| RandomForest [83] | 26.52 | 21.89 | 17.21 |
| XGBoost [36] | 24.22 | 23.06 | 18.29 |
| LinearRegression | 76.76 | 24.12 | 17.82 |
| LightGBM [107] | 36.57 | 22.43 | 18.26 |
| SVM [251] | 84.82 | 23.70 | 17.55 |
| GBDT [254] | 26.50 | 22.37 | 17.46 |
| CatBoost [54] | 23.96 | 21.58 | 17.55 |
| NeuralNetwork [62] | 55.34 | 24.33 | 19.27 |
| RidgeRegression [157] | 84.72 | 24.26 | 17.87 |
| Target | (Reported) | (Manual) | (Estimated) |

TABLE 6.3: Root Mean Squared Error (RMSE) Results

| Model | Reported | Manual | Estimated |
|---|---|---|---|
| KNN [116] | 142.97 | 35.27 | 24.45 |
| Random_Forest [83] | 93.73 | 29.94 | 21.79 |
| XGBoost [36] | 82.67 | 31.75 | 23.61 |
| Linear_Regression | 117.53 | 32.54 | 22.35 |
| LightGBM [107] | 99.77 | 30.55 | 23.58 |
| SVM [251] | 208.29 | 34.23 | 23.66 |
| GBDT [254] | 73.14 | 30.46 | 22.11 |
| CatBoost [54] | 73.05 | 29.64 | 22.55 |
| Neural_Network [62] | 124.21 | 33.38 | 23.18 |
| Ridge_Regression [157] | 134.71 | 32.48 | 22.21 |
| Target | (Reported) | (Manual) | (Estimated) |

Given a ground truth dataset with original reported accident duration, we perform a manual labelling of segments and obtain a set of predicted segments obtained from our automated segmentation algorithm, we compute the precision and the recall of the algorithm, and then combine them into a single F1 score.

F1-score is a popular metric used to evaluate the quality of binary classification models defined as follows:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$
$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

where true positives are the number of correctly classified positive instances, false positives are the number of negative instances classified as positive, and false negatives are the number of positive instances classified as negative.

F1-score is defined as the harmonic mean of precision and recall, given by:

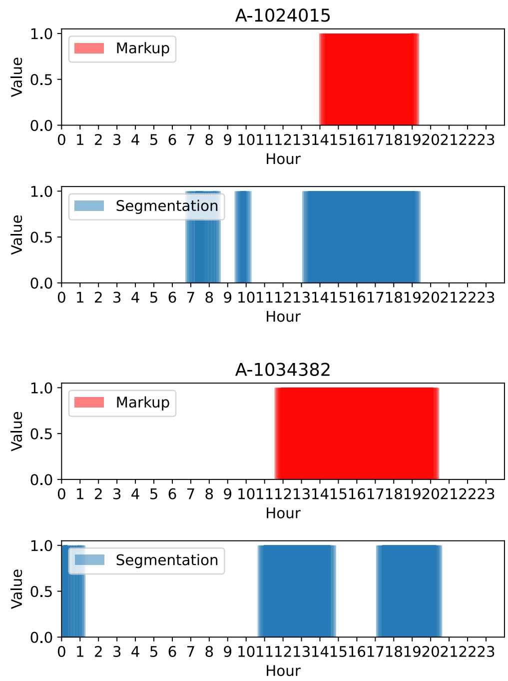



FIGURE 6.13: Manual markup and algorithm segmentation comparison. Time series segments represented as binary values of 0 and 1.

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F1-score ranges from 0 to 1, with higher values indicating a better classification performance.

In the case where a time series is represented as a series of points with values of 1 for segmented intervals and 0 for intervals with no segments, F1-score can be applied to estimate the quality of the time interval segmentation.

To apply the F1-score, we need a ground truth dataset with manually labelled segments (and we obtain this manual markup for 820 accidents), and a set of predicted segments obtained from our automated segmentation algorithm. We can use these two sets to compute the precision and recall of the segmentation algorithm, and then combine them into a single F1-score.

Precision measures the proportion of true positives among all the predicted positives. In the context of time interval segmentation, the precision measures the accuracy of the algorithm in detecting the true segments. The Recall measures the proportion of true positives among all the actual positives. In the context of time interval segmentation, the recall measures the completeness of the algorithm in detecting all the true segments.

To apply the F1-score to estimate the quality of time interval segmentation, we can compute the precision and recall for each segment, and then compute the overall F1-score as the weighted average of precision and recall, weighted by the number of segments. This provides a single metric that reflects the quality of the time interval segmentation.

As a result (see Figure 6.14), the official reported incident segmentation is found to be very off (with a mean F1-score of 0.29 - Figure 6.14a)); next, the segmentation done by the algorithm while selecting only the interval closest to the reported timeline yields the highest average F1-score of 0.51

- Figure 6.14c)) with a peak at 0.3; lastly, when considering multiple segmented incident intervals detected from our algorithm, it produced a slightly lower F1 score of 0.47 - Figure 6.14b)), but more evenly distributed. Overall, the algorithm performance that we propose in this chapter yields a higher precision in detecting disruptions from time series of traffic speeds than from the reported accident timeline. The use of multiple segments produced by the algorithm can highlight multiple disruptions while producing just a slight decrease in the quality of results. The error for multiple intervals segmentation increases because more additional intervals are considered in the evaluation of the metric, which may lay outside of originally marked intervals (see Figures 6.13 and 6.9).

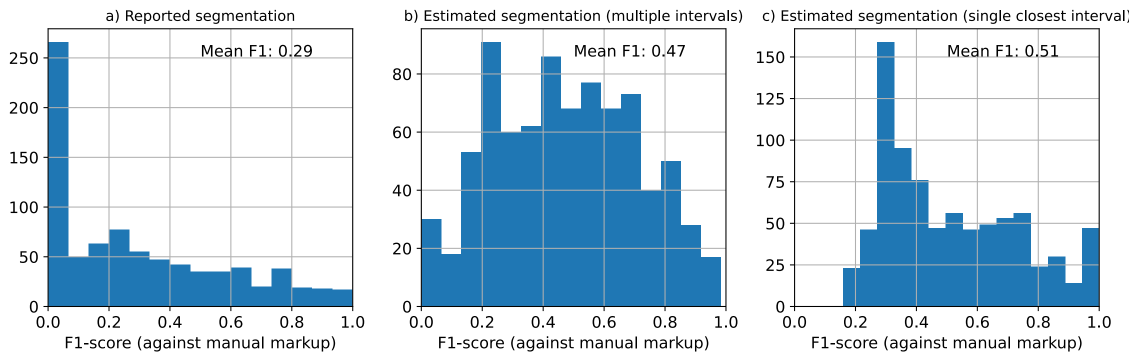



FIGURE 6.14: Histogram of F1-score against manual markup for a) reported accident time interval and b) estimated segmentation when algorithm detecting multiple disruption intervals c) estimated segmentation for the single closest interval to reported incident occurrence time

### 6.5.1 Parameter importance study

For our model, we have the following variables and their intervals of variation:

- **gran**: Granularity, an integer value controlling the level of detail (moving window size) in the metric estimation function. In the provided search space, the range of gran is [2, 40] with a step of 1. Default value is 12.
- **kernel_size**: A list of float values used as weights in the dilation convolution operation. The search space for the kernel is the size of the convolution [1, ... 4], float values primarily intended to implement pre-processing operation for the day time-series. Default value is 3.
- **selectivity**: A float value between 0.01 and 4.0 that determines the power coefficient in post-processing difference estimations Default value is 2.0.
- **shift**: An integer value between -32 and +32 that represents a cyclic shift of the resulting time series to attribute to a shift in convolution operation and facilitate to overall adaptation to the target segmentation. Default value is 0.
- **threshold**: A float value that serves as a threshold in the interval processing function, which is used to perform the binarization of the normalized output array by disruption degree. In the provided search space, the range of the search space for the threshold is [0.01, 0.99]. Default value is 0.15.

At the begininning we perform a hyper-parameter search across all the mentioned parameters but also include a search among metric list (Bray-Curtis, Canberra, Chebyshev, Manhattan, Correlation, Cosine, Euclidean, Minkowski difference metrics) to determine the best performing difference metric for our algorithm. By performing search across 3,000 iterations we then estimate the avrage f1 score obtained when using each metric (see Figure 6.15). The Chebyshev metric yields higher f1 score than other metrics, possibly due to the structure and interpretation of the metric: high difference between maximum and minimum traffic speed measurements within a time window can indicate the presence of the disruption.

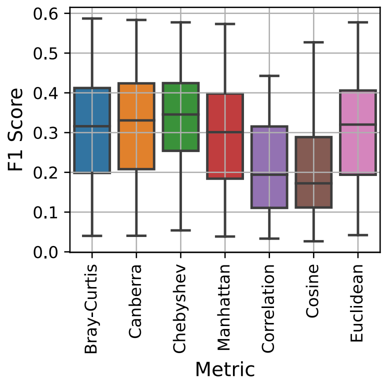



FIGURE 6.15: Connection between metric and F1 score

Our next step is to perform a hyper-parameter search for the Chebyshev difference metric only for 1,000 iterations. We obtained a significant improvement in the average f1 score for multiple interval comparison - 0.62 (a significant improvement from 0.52). As can be seen from scatter plots (see Figure 6.16), there are noticeable positive (kernel size vs f1 score), negative (binarization threshold vs f1 score) and peaking trends (shift vs f1 score) observed in results. Optimal values for the binarization threshold are located at lower values (between 0.01 and 0.4). Overall, the algorithm requires a positive shift in the post-processing function, which contributes to a substantial increase from 0.52 to 0.62 in f1 score when considering the positive shift of the resulting array.

We further provide a Correlation heatmap between algorithm parameters (see Figure 6.17) and the resulting f1 score: 1) The highest Pearson correlation values are with variables threshold (-0.55), shift (0.49), followed by selectivity (-0.32). There are no significant correlations between model parameters themselves.

In conclusion, the hyper-parameter search led to the selection of the Chebyshev metric, which demonstrated the highest average F1 score. Fine-tuning the disruption segmentation algorithm hyper-parameters significantly improved the average F1 score. Trends and optimal parameter values were identified, and the correlation heatmap showed that threshold, shift, and selectivity had the highest Pearson correlation with the F1 score.

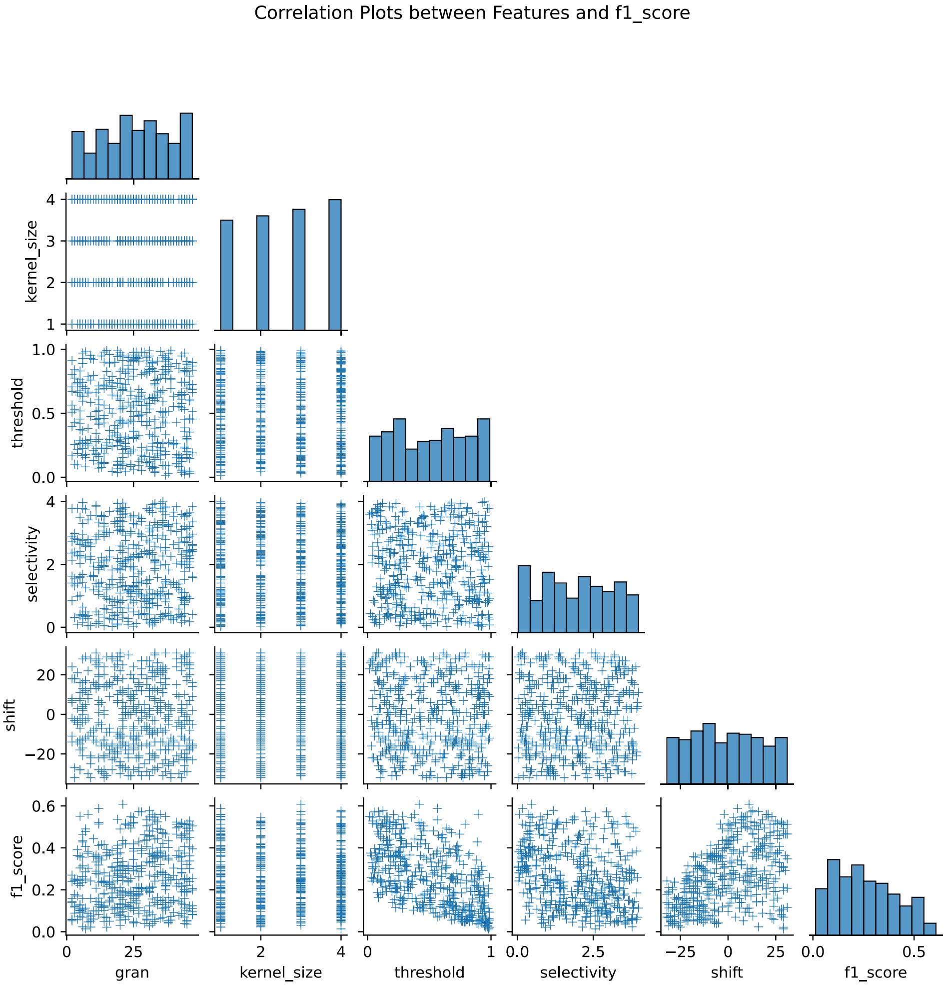



FIGURE 6.16: Scatter plots and histograms between model parameters and F1 score

## 6.6 Application of the methodology

We further publish the code and describe the functionality of the toolkit which can be described as a novel approach for associating geospatial and traffic data to traffic incidents, using data from three different sources such as OpenStreetMap (OSM), the Countrywise Traffic Accident Data Set (CTADS), and the Performance Measurement System (PeMS). Data preprocessing includes various steps, such as filtering data by geographic area, converting road data to point data, and aligning different datasets based on corresponding road points, the toolkit manages to link traffic conditions with incident locations. A specialized algorithm, referred to as CTADS2VDS, is applied to form these associations. Furthermore, the toolkit retrieves traffic speed data for each day of an incident and for the preceding two-week period, offering insights into traffic conditions leading up to the accident.

For authorities, such a toolkit can be incredibly valuable. Not only can it provide comprehensive information about the conditions of traffic accidents, it can also help in identifying patterns or trends related to accident progression and the traffic conditions leading up to these incidents. This can be useful for traffic management, road infrastructure planning, and the design of traffic safety measures. The data-driven approach used by the toolkit can enable authorities to make informed decisions based on observed traffic speed disruption, rather than relying on reported estimates or assumptions. The

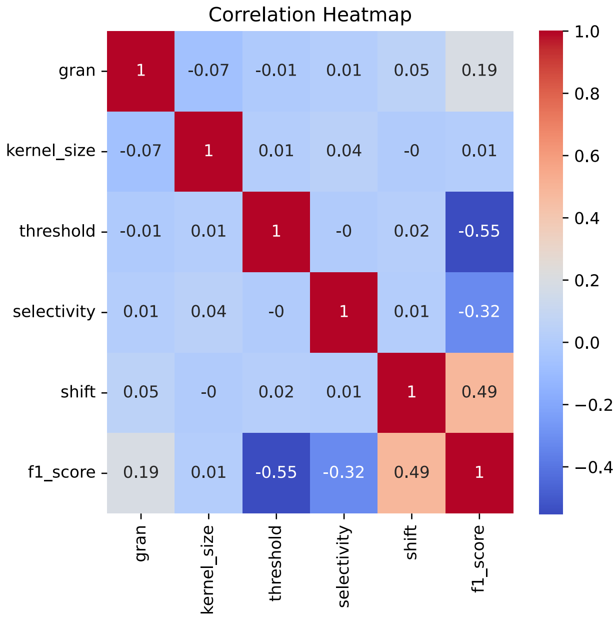



FIGURE 6.17: Correlation matrix for model parameters

methodology followed by this toolkit is modular and versatile. It can be adjusted and optimized based on specific requirements or challenges encountered in different regions. Alternatively, the toolkit can be used to monitor the effectiveness of traffic safety measures by comparing the rate of observed disruptions before and after the implementation of these measures. The capability to analyze accident duration or severity in relation to traffic flow data can also assist authorities in prioritizing their efforts to improve road safety.

There are several potential obstacles related to the code and its adaptation for real-world scenarios, including: 1) Scalability: Handling big data environments, especially in larger urban networks, requires optimized algorithms capable of distributed computing. Two main resource-intensive components are traffic speed/flow data retrieval and CTADS2VDS point mapping, both of which can be improved by implementing parallel versions of algorithms, 2) Interoperability: The original code must be able to parse and handle OSM map and CTADS incident report data formats (e.g., CSV, OSM) and connect to the PeMS database. In the case of alternative data sources and formats, it requires implementing additional data preprocessing steps, 3) Algorithm Precision: performance and limitations of disruption segmentation algorithms explored in the results section. These issues can be mitigated by the following properties of our methodology: 1) Algorithm Fine-Tuning: Fine-tuning of the disruption segmentation algorithm can be performed automatically using hyper-parameter search for best performance on alternative sources of data, 2) Sensitivity-specificity control: maintaining high incident detection rates while minimizing false alarms is a key challenge. The disruption segmentation algorithm allows us to estimate the “degree” of disruption before applying the binarization threshold. This property allows for false-alarm control using fine-tuning of the detection threshold. Balancing sensitivity (identifying real incidents) and specificity (avoiding false alarms) often involves trade-offs and can be fine-tuned to specific data set.

In urban networks with hundreds of measurement locations, the data retrieval is a bottleneck, since each accident report will require a request for daily and fortnight measurements at specific detectors.

Depending on the speed of VDS data retrieval, the amount of data that can be obtained in an acceptable amount of time can be limited.

The code for the paper can be found by the following link:

https://github.com/Future-Mobility-Lab/AAA-toolkit/tree/main

The Table 6.4 presented below provides a summary of the key parts involved in the Accident Analysis& Association (AAA) Toolkit codebase. Each row represents a specific segment of the code, outlining the corresponding inputs required and outputs produced for each segment. The sequence of code parts indicates the flow of data and the transformation processes that occur from acquiring the initial raw data to ultimately applying segmentation algorithms on the compiled information.

| Code Part | Input | Output | Use Case |
|---|---|---|---|
| Download OSM data | Geographic coordinates or region name | Raw OSM data file | - |
| Download CTADS data | Geographic coordinates or region name, CTADS database access | Raw CTADS data file | - |
| Filter CTADS and OSM by area | Raw OSM and CTADS data files | Filtered OSM and CTADS data files | - |
| Convert OSM to points | Filtered OSM data file | Point-based OSM data | Simplifies the geographical data to allow point-based algorithm processing. |
| Get VDS points from PEMS | Geographic coordinates or region name, PEMS database access | VDS data file | - |
| Apply road alignment | Point-based OSM data, VDS data, and CTADS data | Aligned VDS and CTADS data with OSM road points | Initial stage for the association of traffic flow/speed and accident data with specific roads, enabling more accurate accident modelling. Allows algorithms application from road point of view. |
| Apply CTADS2VDS algorithm | Aligned VDS and CTADS data | Associated VDS points with accident points | Links traffic speed/flow detector ID and accident report data. |
| Download traffic speed data(CTADS2TS) | CTADS data with accident dates, PEMS database access | Traffic speed data for the incident day and the two weeks prior | Used to investigate the impact of traffic accidents on traffic speed, leading to more informed speed limit policies. |
| Apply segmentation algorithm | Associated VDS points with accident points, traffic speed data | Time series of disruption degree, disruption intervals associated with incidents | Enables the time series analysis assiciated with the disruption. |

TABLE 6.4: Inputs, Outputs, and Use Cases of AAA Toolkit Code Parts

### 6.6.1 False Positives Rate analysis

The issue of false alarms in the incident detection task can be significant. Traffic authorities may need the control over incident detection specificity. Since our segmentation algorithm provides real values after applying a difference metric, the value of false positives can be controlled by selecting an appropriate threshold of binarization. We provide a receiver operating characteristic curve (ROC) curve for comparison across total merged timeline of incidents and represent manual and estimated segmentation procedures as a binary classification problem. Parameters like granularity and binarization threshold can be fine-tuned according to specific metric (e.g. Area under ROC curve, F1-score or heuristics of metrics) to increase the amount of true positives while reducing the amount of false positives. We utilized F1-score as it able to grasp both of these values in a single formula. As shown on Figure 6.18, our proposed methodology, even without the tuning of hyper-parameters, allows to maintain a high detection rate while keeping the false alarm rate low.

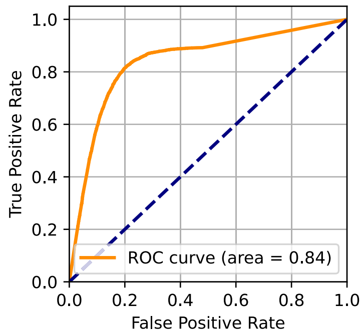



FIGURE 6.18: Receiver operating characteristic for all accidents. Markup vs Estimation

We further look into specifics of disruption detection for various accidents (see Figures 6.19 and 6.20). For some accidents, high detection rate cannot be achieved without increasing the false positives rate. It is important to note that the selection of the binarization threshold plays a crucial role in controlling algorithm performance. A lower threshold might increase the sensitivity, thereby escalating the detection rate, but at the cost of specificity, leading to more false positives. Conversely, a higher threshold might reduce false alarms but may also miss some real incidents, thus lowering the detection rate. Therefore, the end users can fine-tune the parameters according to their specific needs, demonstrating the flexibility and adaptability of our proposed methodology.

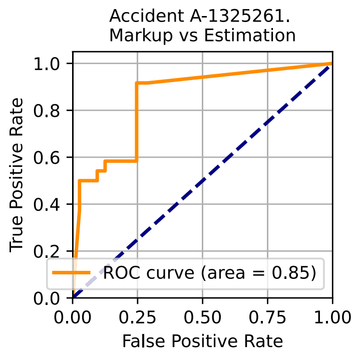



FIGURE 6.19: Receiver operating characteristic for accident A-1325261. Markup vs Estimation

## 6.7 Conclusion

Our methodology aims to automatically detect, segment, and extract traffic disruptions and accidents using distance metrics. This approach improves incident prediction accuracy across multiple machine learning models and provides better fit to manual markup of observed traffic speed disruptions. By obtaining the intervals and shapes of traffic disruptions, we can model the impact of accidents with greater precision, using traffic state measurements rather than just reported parameters (duration, start

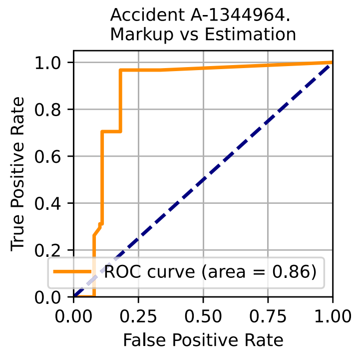



FIGURE 6.20: Receiver operating characteristic for accident A-1344964. Markup vs Estimation

time, etc). This approach provides more data on the accident and allows us to study accident impacts in greater detail.

**Relevance of this work can be summarized in following points:** 1) Enhancement of Traffic Management Systems: Integrate the proposed early detection and disruption segmentation algorithm into existing traffic management systems to improve and automate accident detection and corresponding data collection. This will help to minimize congestion and the overall impact of accidents on traffic flow, 2) highlight of reporting errors to standardize data reporting: Establish standardized guidelines and protocols for reporting traffic accidents, including the accurate reporting of the location, start and end times, number of lanes affected, and other relevant details; this will ensure that data-driven models can accurately predict accident severity and disruption length, 3) highlight the necessity of creating of data standards policies across countries for collecting necessary traffic accident information, 4) development of Incident Response Strategies by utilizing the improved incident prediction models to develop data-driven accident response strategies, including the dynamic traffic rerouting and real-time traffic guidance; this will help to mitigate the impact of traffic accidents on road users and reduce the risk of secondary incidents; 5) Data Fusion for a better traffic accident analysis: due to observed improvement in the quality of prediction arising from data fusion, traffic Authorities can consider integrating data sets from private companies for jointly analysing traffic datasets of various types to improve traffic safety by improving accuracy of traffic accident duration prediction.

**Future research in this area:** 1) Algorithm's complexity can be expanded by incorporating custom kernels, which can be found using hyper-parameter search, 2) Disruption measurements obtained over time can enable the prediction of traffic accident impact propagation with greater accuracy than relying solely on reported values, 3) The proposed methodology can be extended to include disruptions beyond accidents, such as construction or road closures, which can improve the accuracy of impact prediction, 4) Further improvement can also be achieved by performing data fusion and incorporating external data sources, such as weather and events, into the accident impact prediction models. We are currently modelling the cascading effect on traffic disruptions and how these can be automatically identified based on multiple incoming traffic state streams; the main challenge of detecting subsequent incidents lie in the time-span duration of the first accident which is normally stochastic in nature.

**Limitations of this work:** The current modelling approach has been applied to a San Francisco data set due to its public availability and easiness to access. However, we would like to test the approach on multiple other countries and incident databases across the globe; the main challenge is the lack of both traffic states and traffic accidents logs to be released with synchronised timelines.

# Chapter 7

# Discussion, Synthesis and Conclusions

## 7.1 Literature review: discussion, synthesis and conclusions

**Summary of the main findings:** This research identified a significant gap in both the transport management sector and the academic literature with regard to the implementation of advanced methods in incident management and response plan solutions. Current practices largely rely on operational experience rather than data-driven decision-making. The research also highlights the importance of using deep learning and machine learning in traffic incident management systems (TIMS) to accurately predict incident durations and identify the most critical factors influencing incident clearance time. This approach has the potential to save operational costs, reduce end-user time, and decrease traffic congestion. This review also identified several challenges and gaps in the field of traffic accident analysis including imbalanced classification, skewed distribution of accident duration data, reporting errors and anomalies, high-dimensionality of the data, low data availability, incorporation of textual accident descriptions, utilization of historical traffic data, the use of novel machine learning and deep learning models, and incident-related traffic state identification.

**Discussion of the implications:** The findings of this study have several implications for the field of traffic management and incident response. First, they emphasize the need for a shift from experience-based decision-making to data-driven approaches in transport management. This will enable more efficient incident management and the allocation of resources. Second, the application of deep learning and machine learning in TIMS can lead to more accurate predictions of incident durations across different road types, accident types, and countries with varying driving behaviour. This enhanced prediction capability will contribute to existing knowledge and inform the development of future research and traffic management strategies.

**Limitations and future research:** The research has some limitations, such as the focus on Australian traffic management centres and the lack of exploration of all possible modelling capabilities. Future research should investigate a more comprehensive approach to incident duration prediction, encompassing various road types and accident types across different countries. Additionally, further studies should aim to improve the integration of transport modelling and data-driven solutions, utilizing the full potential of deep learning and machine learning in TIMS.

Based on the identified challenges and gaps, several potential future research directions have been proposed:

- Data set integration and fusion models: Combining data sets such as traffic flow, speed, and occupancy with traffic accident reports can enhance incident duration modelling. This may require the use of data fusion models or feature embedding methods.

- Utilization of textual data: Incorporating natural language processing techniques to analyze textual accident reports can provide valuable insights and improve prediction accuracy. In general, traffic accident reports contain unstructured description of accident parameters. Application of natural language processing techniques can facilitate the data formalization (into a tabular form) and summarization.
- Application of advanced machine learning and deep learning models: Exploring more sophisticated ML and DL models can help identify nonlinear relationships and threshold effects in traffic incident duration prediction. The area of traffic accident studies is a recipient for the use of novel AI methods. Consideration that advanced methods underrepresented in this area, implies multiple future studies utilizing these methods.
- Integration of advanced ML pipeline elements into frameworks: Anomaly detection, hyperparameter optimization, dimensionality reduction, and sampling techniques can enhance prediction performance. So far, the integration of ML methods into frameworks, in this study area, is of rare occurrence.
- Real-time incident reporting text analysis: Natural Language processing methods can be utilized for examining textual incident descriptions over timeline to enhance the duration prediction accuracy of incidents occurring in real time.
- Standardization of accident reporting forms based on data-driven approaches utilizing the feature importance estimation techniques: assessing the impact of specific factors, such as weather conditions, on incident duration prediction accuracy, detailing and choosing features that have the highest contribution to the prediction accuracy. Data-driven approach provides highlighting of accident parameters relevant for tasks of prediction and classification.
- Road Type extrapolation tests and model bias considerations: Ensuring the model's performance remains reliable when extrapolating data to other road networks (e.g. cross-network test) or when applied to different time and space contexts (e.g. time-based cross-validation). The bias towards data used for model training can be studied using various extrapolation scenarios.
- Advanced data pre-processing methods: Implementing dimensionality reduction and feature extraction techniques (e.g. using autoencoder) to manage the growing volume and variety of data collected in traffic networks. Both high dimensionality and data availability pose a challenge in data analysis which emphasise the use these methods.

In conclusion, traffic incident duration prediction is a complex and important task that can benefit from further research involving sophisticated artificial intelligence models. By addressing the identified challenges and gaps, future research has the potential to significantly improve traffic incident duration prediction performance, ultimately leading to enhanced traffic flow and reduced impact from traffic incidents. The study highlights the need for more advanced incident management solutions that leverage deep learning and machine learning techniques to accurately predict traffic incident durations and identify the most important factors affecting incident duration. Implementing such data-driven approaches will result in better resource allocation and improved traffic management, ultimately benefiting end-users and society in general.

## 7.2 Bi-level framework: discussion, synthesis and conclusions

**Summary of the main Findings:** The work around a bi-level framework for traffic incident duration prediction presented a universal bi-level framework that addresses several challenges for different road network layouts. The study proposes a framework capable of predicting incident duration regardless of the road network or its complexity. It addresses the issues of outliers and imbalanced data classes by proposing a varying threshold procedure, optimizes both classification and regression problems, and highlights the most influential factors that affect the incident duration. The research demonstrates that the performance of machine learning models is highly affected by the dataset and the chosen methodology, emphasizing the need for a flexible and adaptable approach. This chapter lays the groundwork for bi-level predictive methodologies regarding traffic incident duration, ultimately aiding incident modelling.

**Discussion of the implications:** The proposed bi-level framework for traffic incident duration prediction has significant implications for the field of study. It fills a gap in the literature by providing a universal framework that can be applied to different traffic incident datasets and various road network types. This approach offers a more comprehensive solution to the problem of incident duration prediction, considering varying incident duration threshold analysis, joint hyper-parameter optimization algorithms, and feature importance selection. By identifying the most influential factors that affect incident duration across three different types of road networks, this research can help traffic authorities prioritize their efforts and improve their decision-making processes. The framework's adaptability also allows for more accurate predictions and better-informed decisions.

**Limitations and Future Research:** While the work in this chapter addressed several challenges in predicting traffic incident duration, it acknowledges some limitations. The performance of machine learning models is highly dependent on the quality and size of the dataset and the chosen methodology. Future research could focus on exploring more advanced machine learning techniques to improve model performance further. Additionally, more extensive and diverse datasets could be used to test the framework and further validate its applicability to various road network types and incident scenarios. Incorporating real-time traffic data and dynamic road conditions could also enhance the prediction accuracy and better reflect real-world complexities.

**Conclusions:** Overall the work around the bi-level framework modeling contributes to the ongoing development of a real-time platform for predicting traffic congestion and evaluating the incident impact. The proposed bi-level framework for traffic incident duration prediction is a significant advancement in the field, offering a flexible, adaptable, and comprehensive solution. By addressing the challenges of predicting incident duration on different road network layouts and accounting for various influential factors.

## 7.3 Data fusion for traffic incident duration prediction: discussion, synthesis and conclusions

**Summary of the main findings:** This chapter proposed a novel framework for predicting incident duration by integrating machine learning prediction methods with traffic flow and textual incident description features encoded via several Deep Learning methods. The approach showed stable and significant improvement across all models. Our research also highlighted the importance of using specific deep-learning encoding approaches for regression models to further enhance performance. The study revealed that encoding incident-related features efficiently is crucial for predicting traffic incident impacts. We also investigated the importance of words in incident descriptions using the LIME method, which showed that certain word combinations contribute to the classification of incidents into specific severity or duration groups.

**Discussion of the implications:** This research contributes to the ongoing objective of building a real-time platform for predicting traffic congestion and evaluating incident impacts during peak hours. The proposed framework can provide accurate information for both end-user route choice modeling and operational centers looking to optimize their operations under non-recurrent traffic congestion. Furthermore, the study lays the foundation for bi-level predictive methodologies concerning traffic incident duration, which can be beneficial for Traffic Management Centers (TMCs) in improving incident and traffic management.

**Limitations and future research:** The study has some limitations, including focusing only on the San Francisco area and considering traffic speed and flow only one week before the incident. Future research could incorporate more extensive geographical areas, like California, and longer periods of traffic count data to build traffic speed/flow profiles for more accurate predictions. Also, additional methods of time series encoding may be utilized. Further work could also explore the spatial and temporal dynamic prediction of incident impact using graph-based modelling approaches. The availability of the predicted incident duration data integrated with data on traffic flow and textual incident description can improve the TMC incident and traffic management, reducing the time that people spend in traffic congestion caused by incidents.

## 7.4 Visual transformers for traffic accident risk prediction: discussion, synthesis and conclusions

**Summary of the main findings:** This research introduces a novel approach to traffic accident risk forecasting by reformulating the problem as an image regression task and proposing a unique Contextual Vision Transformer network (C-ViT) that efficiently models traffic accident risk from both spatial and temporal perspectives. The proposed approach outperforms existing methods, requiring significantly fewer training parameters. Additionally, incorporating a static accident risk map with the ViT model (XVit) further improves performance, establishing a new state-of-the-art. The Coarse-Fine-Coarse Visual Transformer (CFC-Vit) architecture allows for fine-grained processing of the accident risk map and introduces an additional scale factor parameter, which can enhance prediction performance.

**Discussion of the implications:** The findings of this study demonstrate the potential of visual transformers and their variations for traffic accident risk prediction, surpassing previous approaches. This research highlights the applicability of vision transformers for non-visual tasks and suggests that further applications of image and video processing methods may yield even better results and open alternative approaches for accident risk prediction. The proposed methods can contribute to more accurate and efficient traffic management, accident prevention, and policy-making in urban environments.

**Limitations and future research:** This study has several limitations, such as the potential for improvement in operation combination methods and constraint functions. There may also be a non-linear dependence between RMSE and the scale factor observed for different datasets, suggesting that an optimal scale factor for accident risk map processing may exist and vary between datasets. Future research can explore alternative combination methods and constraint functions, investigate the optimal scale factor (e.g. the use of weeks, months of historical data) for various datasets, and apply image and video processing methods to further improve accident risk prediction.

**Conclusions:** This research presents a novel approach for traffic accident risk forecasting using visual transformers and their variations, outperforming existing methods. By incorporating static accident risk maps and Coarse-Fine-Coarse Visual Transformer architectures, the proposed methods show significant improvement in prediction performance. These findings can contribute to better traffic management, accident prevention, and policy-making, while also opening new avenues for applying vision transformers to non-visual tasks and exploring image and video processing methods for accident risk prediction.

## 7.5 Accident Segmentation: discussion, synthesis and conclusions

**Summary of the main findings:** This research focused on addressing the challenges in traffic accident analysis due to incorrect or incomplete accident reports and the need for accurate traffic disruption segmentation. We proposed novel methods for traffic disruption segmentation and association between vehicle detector stations and accident reports. We also developed a fusion methodology for combining two large datasets, CTADS and PeMS, to analyze the relationship between traffic accidents and their effects on traffic flow and speed. Through the evaluation of multiple machine learning models, we introduced a new modeling approach that focuses on the amount and shape of the disruption associated with an accident. This research lays the foundation for early traffic accident disruption detection, traffic disruption speed impact analysis, and the use of observed traffic accident durations for correcting errors in user reports.

**Discussion of the implications:** Our findings have significant implications for the field of traffic accident analysis and prediction. By accurately segmenting and analyzing traffic disruptions, we can better understand the impact of accidents on traffic flow and speed. This can lead to improved traffic incident management and more effective allocation of resources in response to accidents. Furthermore, the fusion of two large datasets enables the investigation of traffic accidents across different countries and traffic conditions, contributing to a more comprehensive understanding of the factors influencing accident duration and impact. Our research also provides a foundation for the development of real-time platforms for predicting traffic congestion and evaluating incident impact.

**Limitations and future research:** This research has some limitations, such as the reliance on two large datasets which may not be representative of all traffic conditions worldwide. Additionally, the fusion methodology may not be applicable to all data sources, and the machine learning models tested may not be optimal for all situations. Future research should focus on expanding the analysis to more diverse datasets and investigating other machine learning models for improved prediction performance. Moreover, further research could explore the spatial-temporal impact of disruptions within the traffic network and analyze the influence of various accident characteristics on traffic flow patterns.

**Conclusions:** In conclusion, this research contributes significantly to the field of traffic accident analysis by addressing challenges related to data quality and segmentation, proposing novel methods for traffic disruption analysis, and evaluating the performance of multiple machine learning models. The findings of this research have important implications for traffic incident management, resource allocation, and the development of real-time platforms for predicting traffic congestion and incident impact. Further research is needed to refine the methodologies and expand their applicability to a broader range of traffic conditions and datasets.

## 7.6 Final thesis Conclusion

In conclusion, this thesis demonstrates the potential for leveraging advanced machine learning, deep learning techniques to improve traffic incident duration prediction, accident risk forecasting to support decision-making in traffic management. By utilizing novel approaches such as bi-level frameworks, contextual vision transformers, estimation of observed incident duration via time series segmentation and integrating deep learning methods with traffic flow and description features, these studies contribute to the development of more accurate and efficient traffic management systems.

These advancements have important societal implications, as improved incident prediction and management can lead to reduced congestion, more efficient allocation of resources, and better-informed policy-making. While there are limitations, there are more areas for future research, such as exploring broader geographical areas, optimizing methodologies, and further investigating the applications of vision transformers for non-visual transportation tasks, these studies lay the groundwork for the continued development and application of cutting-edge techniques in the field of traffic management and incident response.

# Bibliography


[1] Rusul Abduljabbar et al. "Applications of artificial intelligence in transport: An overview". In: *Sustainability* 11.1 (2019), p. 189.

[2] Zouhair Elamrani Abou Elassad, Hajar Mousannif, and Hassan Al Moatassime. "A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems". In: *Transportation research part C: emerging technologies* 118 (2020), p. 102708.

[3] Martin W. Adler, Jos van Ommeren, and Piet Rietveld. "Road congestion and incident duration". In: *Economics of Transportation* 2.4 (2013), pp. 109 –118. ISSN: 2212-0122. DOI: https://doi.org/10.1016/j.ecotra.2013.12.003. URL: http://www.sciencedirect.com/science/article/pii/S2212012213000269.

[4] National Highway Traffic Safety Administration. *Traffic safety facts 2013*. U.S. department of transportation, 2013.

[5] Abien Fred Agarap. "Deep learning using rectified linear units (relu)". In: *arXiv preprint arXiv:1803.08375* (2018).

[6] Mohamed Abdullah Mohamed Ahmed Al-Bordiny. "Function, Requirements and Applications of Intelligent Transportation Systems (ITS)". PhD thesis. Faculty of Engineering, Tanta University, 2014.

[7] M Al Hamami and TC Matisziw. "Measuring the spatiotemporal evolution of accident hot spots". In: *Accident Analysis and Prevention* 157 (2021), p. 106133.

[8] Hamzah Al-Najada and Imad Mahgoub. "Real-time incident clearance time prediction using traffic data from internet of mobility sensors". In: *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE. 2017, pp. 728–735.

[9] Farman Ali et al. "Traffic accident detection and condition analysis based on social networking data". In: *Accident Analysis and Prevention* 151 (2021), p. 105973. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2021.105973. URL: https://www.sciencedirect.com/science/article/pii/S000145752100004X.

[10] Farman Ali et al. "Transportation sentiment analysis using word embedding and ontology-based topic modeling". In: *Knowledge-Based Systems* 174 (2019), pp. 27–42. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2019.02.033. URL: https://www.sciencedirect.com/science/article/pii/S0950705119300942.

[11] Abdulla Mohammed Saeed Alkaabi, Dilum Dissanayake, and Roger Bird. "Analyzing clearance time of urban traffic accidents in Abu Dhabi, United Arab Emirates, with hazard-based duration modeling method". In: *Transportation Research Record* 2229.1 (2011), pp. 46–54.

[12] Sharaf Alkheder, Madhar Taamneh, and Salah Taamneh. "Severity prediction of traffic accident using an artificial neural network". In: *Journal of Forecasting* 36.1 (2017), pp. 100–108.

[13] Bahar N Araghi et al. "A comparative study of k-NN and hazard-based models for incident duration prediction". In: *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2014, pp. 1608–1613.

[14] Cristian Arteaga, Alexander Paz, and JeeWoong Park. "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach". In: *Safety Science* 132 (2020), p. 104988. ISSN: 0925-7535. DOI: https://doi.org/10.1016/j.ssci.2020.104988. URL: https://www.sciencedirect.com/science/article/pii/S0925753520303854.

[15] Linking Melbourne Authority. *Department of Economic Development, Jobs, Transport and Resources Annual Report for 2014-2015*. URL: https://www.parliament.vic.gov.au/file_uploads/Linking_Melbourne_Authority_Annual_Report_2014-2015_CwBGv8WN.pdf (visited on 02/02/2021).

[16] Michel Ballings et al. "Evaluating multiple classifiers for stock price direction prediction". In: *Expert Systems with Applications* 42.20 (2015), pp. 7046–7056. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2015.05.013. URL: https://www.sciencedirect.com/science/article/pii/S0957417415003334.

[17] Jie Bao, Pan Liu, and Satish V Ukkusuri. "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data". In: *Accident Analysis & Prevention* 122 (2019), pp. 239–254.

[18] Pablo Barcellos et al. "A novel video based system for detecting and counting vehicles at user-defined virtual loops". In: *Expert Systems with Applications* 42.4 (2015), pp. 1845–1856.

[19] Marc Barthelemy. "Spatial networks". In: *Physics reports* 499.1-3 (2011), pp. 1–101.

[20] R Bekkerman. *The present and the future of the KDD cup competition: an outsider's perspective*. 2015.

[21] Abdelmoudjib Benterki et al. "Artificial intelligence for vehicle behavior anticipation: Hybrid approach based on maneuver classification and trajectory prediction". In: *IEEE Access* 8 (2020), pp. 56992–57002.

[22] James Bergstra and Y. Bengio. "Random Search for Hyper-Parameter Optimization". In: *The Journal of Machine Learning Research* 13 (Mar. 2012), pp. 281–305.

[23] Christopher M Bishop, Michael E Tipping, et al. "Bayesian regression and classification". In: *Nato Science Series sub Series III Computer And Systems Sciences* 190 (2003), pp. 267–288.

[24] J Roger Bray and John T Curtis. "An ordination of the upland forest communities of southern Wisconsin". In: *Ecological monographs* 27.4 (1957), pp. 326–349.

[25] L Breiman et al. "Classification and regression trees (the wadsworth statistics/probability series) chapman and hall". In: *New York, NY* (1984), pp. 1–358.

[26] Leo Breiman. "Random Forests". In: *Mach. Learn.* 45.1 (2001), 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[27] Leo Breiman et al. "Cart". In: *Classification and regression trees* (1984).

[28] Markus Breunig et al. "LOF: Identifying Density-Based Local Outliers." In: vol. 29. June 2000, pp. 93–104. DOI: 10.1145/342009.335388.

[29] Raj Bridgelall and Denver D Tolliver. "Railroad accident analysis using extreme gradient boosting". In: *Accident Analysis and Prevention* 156 (2021), p. 106126.

[30] Donglin Cao, Shuru Wang, and Dazhen Lin. "Chinese Microblog Users' Sentiment-Based Traffic Condition Analysis". In: *Soft Comput.* 22.21 (2018), 7005–7014. ISSN: 1432-7643. DOI: 10.1007/s00500-018-3293-8. URL: https://doi.org/10.1007/s00500-018-3293-8.

[31] Hsin-li Chang and Tse-pin Chang. "Prediction of freeway incident duration based on classification tree analysis". In: *Journal of the Eastern Asia Society for Transportation Studies* 10 (2013), pp. 1964–1977.

[32] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[33] Chao Chen et al. "Freeway performance measurement system: mining loop detector data". In: *Transportation Research Record* 1748.1 (2001), pp. 96–102.

[34] Chao Chen et al. "Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data". In: *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*. IEEE. 2018, pp. 328–333.

[35] Shun-Yuan Chen, Yung-Chuan Chen, and Cheng-Hao Hsieh. "Learning Traffic Accident Risk Prediction using Deep Learning". In: *IEEE Transactions on Intelligent Transportation Systems* 17.12 (2016), pp. 3409–3418.

[36] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[37] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015).

[38] Tianyi Chen et al. "Predicting lane-changing risk level based on vehicles' space-series features: A pre-emptive learning approach". In: *Transportation research part C: emerging technologies* 116 (2020), p. 102646.

[39] Tom Choe, Alexander Skabardonis, and Pravin Varaiya. "Freeway performance measurement system: operational analysis tool". In: *Transportation research record* 1811.1 (2002), pp. 67–75.

[40] "Choosing the metric: a simple model approach". In: *Meta-Learning in Computational Intelligence* (2011), pp. 97–115.

[41] California Highway Patrol (CHP). *Statewide Integrated Traffic Records System (SWITRS)*. URL: https://www.chp.ca.gov/switrs/ (visited on 02/02/2021).

[42] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[43] Yi-Shih Chung, Yu-Chiun Chiou, and Chia-Hua Lin. "Simultaneous equation modeling of freeway accident duration and lanes blocked". In: *Analytic Methods in Accident Research* 7 (2015), pp. 16 –28. ISSN: 2213-6657. DOI: https://doi.org/10.1016/j.amar.2015.04.003. URL: http://www.sciencedirect.com/science/article/pii/S2213665715000275.

[44] Younshik Chung. "Development of an accident duration prediction model on the Korean Freeway Systems". In: *Accident Analysis and Prevention* 42.1 (2010), pp. 282–289.

[45] Younshik Chung, Lubinda Walubita, and Keechoo Choi. "Modeling Accident Duration and Its Mitigation Strategies on South Korean Freeway Systems". In: *Transportation Research Record Journal of the Transportation Research Board* 2178 (Jan. 2011), pp. 49–57. DOI: 10.3141/2178-06.

[46] Younshik Chung, Lubinda F Walubita, and Keechoo Choi. "Modeling accident duration and its mitigation strategies on South Korean freeway systems". In: *Transportation research record* 2178.1 (2010), pp. 49–57.

[47] European Commission. *Road Safety Atlas*. URL: https://ec.europa.eu/transport/road_safety/specialist/library/atlas/index_en.htm (visited on 02/02/2021).

[48] Carolyn J Crouch. "A cluster-based approach to thesaurus construction". In: *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*. 1988, pp. 309–320.

[49] Rafael Prieto Curiel, Humberto Gonzalez Ramirez, and Steven Richard Bishop. "A novel rare event approach to measure the randomness and concentration of road accidents". In: *PloS one* 13.8 (2018).

[50] Soumitra Das, Sachin Mohanty, and Pushpak Bhattacharyya. "Topical text modeling for incident detection on twitter". In: *Expert Systems with Applications* 122 (2019), pp. 182–195.

[51] Thomas G. Dietterich. "Ensemble Methods in Machine Learning". In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6.

[52] Loukas Dimitriou and Eleni I Vlahogianni. "Fuzzy modeling of freeway accident duration with rainfall and traffic flow interactions". In: *Analytic Methods in Accident Research* 5 (2015), pp. 59–71.

[53] Youcef Djenouri et al. "A survey on urban traffic anomalies detection algorithms". In: *IEEE Access* 7 (2019), pp. 12192–12205.

[54] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support". In: *arXiv preprint arXiv:1810.11363* (2018).

[55] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[56] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. "Support vector machines for spam categorization". In: *IEEE Transactions on Neural networks* 10.5 (1999), pp. 1048–1054.

[57] Laura Eboli, Carmen Forciniti, and Gabriella Mazzulla. "Factors influencing accident severity: an analysis by road accident type". In: *Transportation Research Procedia* 47 (2020). 22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18th – 20th September 2019, Barcelona, Spain, pp. 449–456. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2020.03.120. URL: https://www.sciencedirect.com/science/article/pii/S2352146520303197.

[58] Mehrdad Farahani et al. "Short-term traffic flow prediction using variational LSTM networks". In: *arXiv preprint arXiv:2002.07922* (2020).

[59] Evelyn Fix and JL Hodges. "Discriminatory analysis, nonparametric discrimination". In: (1951).

[60] Jerome Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29 (Nov. 2000). DOI: 10.1214/aos/1013203451.

[61] Shota Fukuda et al. "Short-term prediction of traffic flow under incident conditions using graph convolutional recurrent neural network and traffic simulation". In: *IET Intelligent Transport Systems* 14.8 (2020), pp. 936–946.

[62] Stephen I Gallant et al. "Perceptron-based learning algorithms". In: *IEEE Transactions on neural networks* 1.2 (1990), pp. 179–191.

[63] Lisa Gaudette and Nathalie Japkowicz. "Evaluation methods for ordinal classification". In: *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*. Springer. 2009, pp. 207–210.

[64] S Geisser. *An Introduction to Predictive Inference*. 1993.

[65] Milad Ghasri et al. "Hazard-based model for concrete pouring duration using construction site and supply chain parameters". In: *Automation in Construction* 71 (2016), pp. 283–293.

[66] Hamid Gholami et al. "Mapping wind erosion hazard with regression-based machine learning algorithms". In: *Scientific Reports* 10.1 (2020), p. 20494.

[67] Banishree Ghosh. "Predicting the duration and impact of the non-recurring road incidents on the transportation network". PhD thesis. 2019.

[68] Banishree Ghosh and Justin Dauwels. "Comparison of different Bayesian methods for estimating error bars with incident duration prediction". In: *Journal of Intelligent Transportation Systems* 26.4 (2022), pp. 420–431.

[69] Banishree Ghosh et al. "Dynamic prediction of the incident duration using adaptive feature set". In: *IEEE Transactions on Intelligent Transportation Systems* 20.11 (2018), pp. 4019–4031.

[70] Banishree Ghosh et al. "Dynamic prediction of the incident duration using adaptive feature set". In: *IEEE Transactions on Intelligent Transportation Systems* 20.11 (2018), pp. 4019–4031.

[71] Banishree Ghosh et al. "Predicting the duration of non-recurring road incidents by cluster-specific models". In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2016, pp. 1522–1527.

[72] Mark N Gibbs. "Bayesian Gaussian processes for regression and classification". PhD thesis. Citeseer, 1998.

[73] Australian Government. *Road Safety*. 2017. URL: https://infrastructure.gov.au/roads/safety/.

[74] UK Government. *UK Road Safety Statistics*. URL: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data (visited on 02/02/2021).

[75] Artur Grigorev et al. "Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation". In: *Transportation research part C: emerging technologies* 141 (2022), p. 103721.

[76] Artur Grigorev et al. "Traffic incident duration prediction via a deep learning framework for text description encoding". In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2022, pp. 1770–1777.

[77] Jiuxiang Gu et al. "Recent advances in convolutional neural networks". In: *Pattern recognition* 77 (2018), pp. 354–377.

[78] Steve R Gunn et al. "Support vector machines for classification and regression". In: *ISIS technical report* 14.1 (1998), pp. 5–16.

[79] Isabelle Guyon and Andre Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[80] Khaled Hamad, Mohamad Ali Khalil, and Abdul Razak Alozi. "Predicting Freeway Incident Duration Using Machine Learning". In: *International Journal of Intelligent Transportation Systems Research* (2019), pp. 1–14.

[81] Khaled Hamad, Mohamad Ali Khalil, and Abdul Razak Alozi. "Predicting freeway incident duration using machine learning". In: *International Journal of Intelligent Transportation Systems Research* 18 (2020), pp. 367–380.

[82] Khaled Hamad et al. "Predicting incident duration using random forests". In: *Transportmetrica A: transport science* 16.3 (2020), pp. 1269–1293.

[83] Khaled Hamad et al. "Predicting incident duration using random forests". In: *Transportmetrica A: transport science* 16.3 (2020), pp. 1269–1293.

[84] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning". In: *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1*. Springer. 2005, pp. 878–887.

[85] Henrick J. Haule et al. "Evaluating the impact and clearance duration of freeway incidents". In: *International Journal of Transportation Science and Technology* 8.1 (2019), pp. 13 –24. ISSN: 2046-0430. DOI: https://doi.org/10.1016/j.ijtst.2018.06.005. URL: http://www.sciencedirect.com/science/article/pii/S2046043018300522.

[86] Henrick J Haule et al. "Evaluating the impact and clearance duration of freeway incidents". In: *International journal of transportation science and technology* 8.1 (2019), pp. 13–24.

[87] Haibo He et al. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning". In: July 2008, pp. 1322 –1328. DOI: 10.1109/IJCNN.2008.4633969.

[88] Qing He et al. "Incident duration prediction with hybrid tree-based quantile regression". In: *Advances in dynamic network modeling in complex transportation systems*. Springer, 2013, pp. 287–305.

[89] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.

[90] Ahmad Hojati et al. "Analysing freeway traffic incident duration using an Australian data set". In: *Road and Transport Research* 21 (June 2012), pp. 16–28.

[91] Ahmad [Tavassoli Hojati] et al. "Modelling total duration of traffic incidents including incident detection and recovery time". In: *Accident Analysis and Prevention* 71 (2014), pp. 296 –305. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2014.06.006. URL: http://www.sciencedirect.com/science/article/pii/S0001457514001791.

[92] Sungmin Hong et al. "The Effect of Road Environment Factors on Freeway Traffic Crash Frequency during Daylight, Twilight, and Night Conditions". In: (Jan. 2014).

[93] Moinul Hossain et al. "Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements". In: *Accident Analysis and Prevention* 124 (2019), pp. 66–84.

[94] Lin Hou et al. "Modeling freeway incident response time: A mechanism-based approach". In: *Transportation Research Part C: Emerging Technologies* 28 (2013). Euro Transportation: selected paper from the EWGT Meeting, Padova, September 2009, pp. 87 –100. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2012.12.005. URL: http://www.sciencedirect.com/science/article/pii/S0968090X12001519.

[95] Guang-Bin Huang, Dian Hui Wang, and Yuan Lan. "Extreme learning machines: a survey". In: *International journal of machine learning and cybernetics* 2.2 (2011), pp. 107–122.

[96] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications". In: *Neurocomputing* 70.1-3 (2006), pp. 489–501.

[97] Tingting Huang, Shuo Wang, and Anuj Sharma. "Highway crash detection and risk estimation using deep learning". In: *Accident Analysis and Prevention* 135 (2020), p. 105392.

[98] Xiaohui Huang et al. "Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video". In: *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 6.2 (2020), pp. 1–28.

[99] Department of Infrastructure Regional Development and Cities. *Australian Road Deaths Database - ARDD*. URL: https://www.bitre.gov.au/statistics/safety/fatal_road_crash_database (visited on 02/02/2021).

[100] Compass IoT. *Compass IoT*. URL: https://www.compassiot.com.au/ (visited on 02/02/2021).

[101] Davor Ivankovic and Mirjana Kujundzic Tiljak. "Comparison of two or more samples of quantitative data". In: *Acta medica Croatica: Casopis Hravatske akademije medicinskih znanosti* 60 (2006), pp. 37–46.

[102] Roxana J Javid and Ramina Jahanbakhsh Javid. "A framework for travel time variability analysis using urban traffic incident data". In: *IATSS research* 42.1 (2018), pp. 30–38.

[103] Kieran Kalair and Colm Connaughton. "Dynamic and interpretable hazard-based models of traffic incident durations". In: *Frontiers in future transportation* 2 (2021), p. 669015.

[104] Barry L Kalman and Stan C Kwasny. "Why tanh: choosing a sigmoidal function". In: *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. Vol. 4. IEEE. 1992, pp. 578–581.

[105] Leonid Kantorovich. "On the translocation of masses". In: *Management Science* 5.1 (1958), pp. 1–4.

[106] Sota Kato and Kazuhiro Hotta. "Mse loss with outlying label for imbalanced classification". In: *arXiv preprint arXiv:2107.02393* (2021).

[107] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[108] Asad J Khattak, Joseph L Schofer, and Mu-Han Wang. "A simple time sequential procedure for predicting freeway incident duration". In: *Journal of Intelligent Transportation Systems* 2.2 (1995), pp. 113–138.

[109] Asad J Khattak et al. "Modeling traffic incident duration using quantile regression". In: *Transportation Research Record* 2554.1 (2016), pp. 139–148.

[110] Hyung Jin KIM and Hoi-Kyun CHOI. "A Comparative Analysis of Incident Service Time on Urban Freeways". In: *IATSS Research* 25.1 (2001), pp. 62 –72. ISSN: 0386-1112. DOI: https://doi.org/10.1016/S0386-1112(14)60007-8. URL: http://www.sciencedirect.com/science/article/pii/S0386111214600078.

[111] Woon Kim and Gang-Len Chang. "Development of a Hybrid Prediction Model for Freeway Incident Duration: A Case Study in Maryland". In: *International Journal of Intelligent Transportation Systems Research* 10 (Jan. 2011). DOI: 10.1007/s13177-011-0039-8.

[112] Luk Knapen et al. "Within day rescheduling microsimulation combined with macrosimulated traffic". In: *Transportation Research Part C: Emerging Technologies* 45 (2014), pp. 99–118.

[113] Aran Komatsuzaki. "One epoch is all you need". In: *arXiv preprint arXiv:1906.06669* (2019).

[114] Mark A Kramer. "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE journal* 37.2 (1991), pp. 233–243.

[115] Li Kuang et al. "Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor". In: *Journal of Intelligent Transportation Systems* 23.2 (2019), pp. 161–174.

[116] Li Kuang et al. "Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor". In: *Journal of Intelligent Transportation Systems* 23.2 (2019), pp. 161–174.

[117] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[118] Jonghak Lee et al. "Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study". In: *Applied Sciences* 10 (Dec. 2019), p. 129. DOI: 10.3390/app10010129.

[119] Ying Lee and Chien-Hung Wei. "A computerized feature selection method using genetic algorithms to forecast freeway accident duration times". In: *Computer-Aided Civil and Infrastructure Engineering* 25.2 (2010), pp. 132–148.

[120] Roger J Lewis. "An introduction to classification and regression tree (CART) analysis". In: *Annual meeting of the society for academic emergency medicine in San Francisco, California*. Vol. 14. Citeseer. 2000.

[121] Jie Li et al. "An overview of graph embedding: Problems, techniques and applications". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12.5 (2018), pp. 1–36.

[122] Linchao Li et al. "A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction". In: *Engineering Applications of Artificial Intelligence* 93 (2020), p. 103686.

[123] Linchao Li et al. "Real-time traffic incident detection based on a hybrid deep learning model". In: *Transportmetrica A: transport science* 18.1 (2022), pp. 78–98.

[124] Linchao Li et al. "Traffic incident detection based on extreme machine learning". In: *Journal of Applied Science and Engineering* 20.4 (2017), pp. 409–416.

[125] Ruimin li. "Traffic incident duration analysis and prediction models based on the survival analysis approach". In: *IET Intelligent Transport Systems* 9 (Jan. 2014). DOI: 10.1049/iet-its.2014.0036.

[126] Ruimin Li. "Traffic incident duration analysis and prediction models based on the survival analysis approach". In: *IET intelligent transport systems* 9.4 (2015), pp. 351–358.

[127] Ruimin Li, Francisco C Pereira, and Moshe E Ben-Akiva. "Competing risk mixture model and text analysis for sequential incident duration prediction". In: *Transportation Research Part C: Emerging Technologies* 54 (2015), pp. 74–85.

[128] Ruimin Li, Francisco C Pereira, and Moshe E Ben-Akiva. "Competing risks mixture model for traffic incident duration prediction". In: *Accident Analysis and Prevention* 75 (2015), pp. 192–201.

[129] Ruimin Li, Francisco C Pereira, and Moshe E Ben-Akiva. "Overview of traffic incident duration analysis and prediction". In: *European transport research review* 10.2 (2018), pp. 1–13.

[130] Ruimin Li and Pan Shang. "Incident duration modeling using flexible parametric hazard-based models". In: *Computational intelligence and neuroscience* 2014 (2014), pp. 33–33.

[131] Xiaobing Li. "Analysis of Large-Scale Traffic Incidents and En Route Diversions Due to Congestion on Freeways". In: (2018).

[132] Xiaobing Li et al. "Sequential prediction for large-scale traffic incident duration: Application and comparison of survival models". In: *Transportation research record* 2674.1 (2020), pp. 79–93.

[133] Zewen Li et al. "A survey of convolutional neural networks: analysis, applications, and prospects". In: *IEEE transactions on neural networks and learning systems* (2021).

[134] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[135] Lei Lin, Qian Wang, and Adel W Sadek. "A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations". In: *Accident Analysis and Prevention* 91 (2016), pp. 114–126.

[136] Lei Lin, Qian Wang, and Adel W Sadek. "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction". In: *Transportation Research Part C: Emerging Technologies* 55 (2015), pp. 444–459.

[137] Lei Lin, Qian Wang, and Adel W Sadek. "Duration prediction of urban freeway traffic accidents based on the M5P tree and hazard-based duration model". In: ().

[138] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[139] Yunduan Lin and Ruimin Li. "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data". In: *Accident Analysis and Prevention* 145 (2020), p. 105696.

[140] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest". In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[141] Tong Liu et al. "Using empirical traffic trajectory data for crash risk evaluation under three-phase traffic theory framework". In: *Accident Analysis and Prevention* 157 (2021), p. 106191.

[142] Yang Liu et al. "A survey of visual transformers". In: *arXiv preprint arXiv:2111.06091* (2021).

[143] Shao-Yuan Lo et al. "Efficient dense modules of asymmetric convolution for real-time semantic segmentation". In: *Proceedings of the ACM Multimedia Asia*. 2019, pp. 1–6.

[144] Jorge Lopes et al. *Dynamic forecast of incident clearance time using adaptive artificial neural network models*. 2013.

[145] Hao Lu et al. "Social signal-driven knowledge automation: a focus on social transportation". In: *IEEE Transactions on Computational Social Systems* 8.3 (2021), pp. 737–753.

[146] Yi-Chi Lu. "Detecting outliers for improving the quality of incident duration prediction". PhD thesis. University of Maryland, College Park, 2021.

[147] Yi-Chi Lu. "Detecting Outliers for Improving the Quality of Incident Duration Prediction". In: (2021).

[148] Scott Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874* (2017).

[149] Chen Ma et al. "Point-of-interest recommendation: Exploiting self-attentive autoencoders with neighbor-aware influence". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 697–706.

[150] Xiaolei Ma et al. "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method". In: *IEEE Transactions on Intelligent Transportation Systems* 18.9 (2017), pp. 2303–2310.

[151] Yifang Ma et al. "Artificial intelligence applications in the development of autonomous vehicles: A survey". In: *IEEE/CAA Journal of Automatica Sinica* 7.2 (2020), pp. 315–329.

[152] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[153] Mirialys Machin et al. "On the use of artificial intelligence techniques in intelligent transportation systems". In: Apr. 2018, pp. 332–337. DOI: 10.1109/WCNCW.2018.8369029.

[154] Tuo Mao et al. "Boosted Genetic Algorithm using Machine Learning for traffic control optimization". In: *IEEE Transactions on Intelligent Transportation Systems* (2021).

[155] Xinhua Mao et al. "Risk factors affecting traffic accidents at urban weaving sections: Evidence from China". In: *International journal of environmental research and public health* 16.9 (2019), p. 1542.

[156] Dragos D Margineantu and Thomas G Dietterich. "Pruning adaptive boosting". In: *ICML*. Vol. 97. Citeseer. 1997, pp. 211–218.

[157] Gary C McDonald. "Ridge regression". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1 (2009), pp. 93–100.

[158] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[159] Adriana-Simona Mihaita et al. "Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting". In: *arXiv preprint arXiv:1905.12254* (2019).

[160] Adriana Simona Mihaita et al. "Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting". In: *CoRR* abs/1905.12254 (2019). arXiv: 1905.12254. URL: http://arxiv.org/abs/1905.12254.

[161] Mahalia Miller and Chetan Gupta. "Mining traffic incidents to forecast impact". In: *Proceedings of the ACM SIGKDD international workshop on urban computing*. 2012, pp. 33–40.

[162] Hermann Minkowski. *Raum und Zeit*. Vieweg, 1909.

[163] Amin Mohammadnazar et al. "Understanding how relationships between crash frequency and correlates vary for multilane rural highways: Estimating geographically and temporally weighted regression models". In: *Accident Analysis and Prevention* 157 (2021), p. 106146.

[164] Zainab Ali Mohammed, Mohammed Najm Abdullah, and Imad Husain Al Hussaini. "Predicting incident duration based on machine learning methods". In: *Iraqi Journal of Computers, Communications, Control and Systems Engineering* 21.1 (2021), pp. 1–15.

[165] Seyed Mohammad Moosavi et al. "Accident Type Detection Using Text Mining of Traffic Accident Reports". In: *IEEE Access* 7 (2019), pp. 109960–109976.

[166] Sobhan Moosavi et al. "A countrywide traffic accident dataset". In: *arXiv preprint arXiv:1906.05409* (2019).

[167] Moggan Motamed et al. "Developing a real-time freeway incident detection model using machine learning techniques". PhD thesis. 2016.

[168] W. James Murdoch et al. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080. ISSN: 0027-8424. DOI: 10.1073/pnas.1900654116. eprint: https://www.pnas.org/content/116/44/22071.full.pdf. URL: https://www.pnas.org/content/116/44/22071.

[169] Doohee Nam and Fred Mannering. "An exploratory hazard-based analysis of highway incident duration". In: *Transportation Research Part A: Policy and Practice* 34.2 (2000), pp. 85 –102. ISSN: 0965-8564. DOI: https://doi.org/10.1016/S0965-8564(98)00065-2. URL: http://www.sciencedirect.com/science/article/pii/S0965856498000652.

[170] Chris J Needham et al. "A primer on learning in Bayesian networks for computational biology". In: *PLoS computational biology* 3.8 (2007), e129.

[171] Hoang Nguyen, Chen Cai, and Fang Chen. "Automatic classification of traffic incident's severity using machine learning approaches". In: *IET Intelligent Transport Systems* 11.10 (2017), pp. 615–623.

[172] Hoang Nguyen et al. "Deep learning methods in transportation domain: a review". In: *IET Intelligent Transport Systems* 12.9 (2018), pp. 998–1004.

[173] National Highway Traffic Safety Administration (NHTSA). *Fatality Analysis Reporting System (FARS)*. 2020. URL: https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

[174] Juan de Ona, Randa Oqab Mujalli, and Francisco J Calvo. "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks". In: *Accident Analysis and Prevention* 43.1 (2011), pp. 402–411.

[175] World Health Organization. *Global status report on road safety 2015*. World Health Organization, 2015.

[176] World Health Organization. *Global status report on road safety 2018: summary*. World Health Organization, 2018.

[177] World Health Organization. *World Health Organization's Global Health Estimates*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/estimated-number-of-road-traffic-deaths (visited on 02/02/2021).

[178] Kaan Ozbay and Pushkin Kachroo. "Incident management in intelligent transportation systems". In: (1999).

[179] Kaan Ozbay and Nebahat Noyan. "Estimation of incident clearance times using Bayesian Networks approach". In: *Accident Analysis & Prevention* 38.3 (2006), pp. 542–555.

[180] Kaan Ozbay and Nebahat Noyan. "Estimation of incident clearance times using Bayesian Networks approach". In: *Accident Analysis and Prevention* 38.3 (2006), pp. 542–555.

[181] Halit Ozen et al. "Multi-step approach to improving accuracy of incident duration estimation: case study of Istanbul". In: *Tehnički vjesnik* 26.6 (2019), pp. 1777–1783.

[182] Hyoshin Park, Ali Haghani, and Xin Zhang. "Interpretation of Bayesian neural networks for predicting the duration of detected incidents". In: *Journal of Intelligent Transportation Systems* 20.4 (2016), pp. 385–400.

[183] Hyoshin Park, Ali Haghani, and Xin Zhang. "Interpretation of Bayesian neural networks for predicting the duration of detected incidents". In: *Journal of Intelligent Transportation Systems* 20.4 (2016), pp. 385–400.

[184] Amir Bahador Parsa et al. "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis". In: *Accident Analysis and Prevention* 136 (2020), p. 105405.

[185] Siavash Parsa et al. "Real-Time Detection and Classification of Objects in Traffic Scenes". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 1095–1100.

[186] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[187] Francisco C Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. "Text analysis in incident duration prediction". In: *Transportation Research Part C: Emerging Technologies* 37 (2013), pp. 177–192.

[188] Ronaldo Prati, Gustavo Batista, and Maria-Carolina Monard. "Data mining with imbalanced class distributions: Concepts and methods". In: Jan. 2009, pp. 359–376.

[189] Md Adilur Rahim and Hany M Hassan. "A deep learning based traffic crash severity prediction framework". In: *Accident Analysis & Prevention* 154 (2021), p. 106090.

[190] Gokul Rajesh et al. "A deep learning based accident detection system". In: *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2020, pp. 1322–1325.

[191] Honglei Ren et al. "A deep learning approach to the citywide traffic accident risk prediction". In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 3346–3351.

[192] Zhenhua Ren et al. "Deep Spatial-Temporal Residual Networks for Citywide Crowd Flows Prediction". In: *arXiv preprint arXiv:1710.00025* (2017).

[193] RM Zahid Reza and Srinivas S Pulugurtha. “Forecasting short-term relative changes in travel time on a freeway”. In: *Case Studies on Transport Policy* 7.2 (2019), pp. 205–217.

[194] Mohammad Ali Sahraei et al. *An Overview of Traffic Accident Prediction Models*. Feb. 2021. ISBN: 978-625-7677-10-3.

[195] Angelica Salas, Panagiotis Georgakis, and Yannis Petalas. “Incident detection using data from social media”. In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. 2017, pp. 751–755. DOI: 10.1109/ITSC.2017.8317967.

[196] Khaled Saleh, Artur Grigorev, and Adriana-Simona Mihaita. “Traffic Accident Risk Forecasting using Contextual Vision Transformers”. In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2022, pp. 2086–2092.

[197] Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1988.

[198] Abdulsamet Saracoglu and Halit Ozen. “Estimation of traffic incident duration: A comparative study of decision tree models”. In: *Arabian Journal for Science and Engineering* 45.10 (2020), pp. 8099–8110.

[199] Ziad Sawalha and Tarek Sayed. “Transferability of accident prediction models”. In: *Safety science* 44.3 (2006), pp. 209–219.

[200] Robert E Schapire. “Explaining adaboost”. In: *Empirical inference*. Springer, 2013, pp. 37–52.

[201] Ron Schindler and Giulio Bianchi Piccinini. “Truck driversâ€™ behavior in encounters with vulnerable road users at intersections: Results from a test-track experiment”. In: *Accident Analysis and Prevention* 159 (2021), p. 106289.

[202] David Schrank and Tim Lomax. “The 2002 Urban Mobility Report (College Station, TX: Texas Transportation Institute, Texas AandM University, June)”. In: (2002).

[203] Pranab Kumar Sen and Sujit Kumar Dhar. “Stock price prediction using LSTM, RNN and CNN-sliding window model”. In: *International journal of scientific and engineering research* 9.11 (2018), pp. 2046–2054.

[204] S Shafiei et al. “Short-Term Traffic Prediction Under Non-Recurrent Incident Conditions Integrating Data-Driven Models and Traffic Simulation”. In: *Transportation Research Board 99th Annual Meeting*. 2020.

[205] Qiang Shang, Tian Xie, and Yang Yu. “Prediction of duration of traffic incidents by hybrid deep learning based on multi-source incomplete data”. In: *International journal of environmental research and public health* 19.17 (2022), p. 10903.

[206] Qiang Shang et al. “A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis”. In: *Journal of Advanced Transportation* 2019 (2019).

[207] Luou Shen and Min Huang. “Data mining method for incident duration prediction”. In: *International Conference on Applied Informatics and Communication*. Springer. 2011, pp. 484–492.

[208] Xiupeng Shi et al. *Accident risk prediction based on driving behavior feature learning using CART and XGBoost*. Tech. rep. 2018.

[209] Swarnima Singh and Vikash Yadav. "An Improved Particle Swarm Optimization for Prediction of Accident Severity". In: *IJEER* 9.3 (2021), pp. 42–47.

[210] Brian Lee Smith and Kevin W Smith. "An investigation into incident duration forecasting". In: *ITS America 11th Annual Meeting and Exposition, ITS: Connecting the AmericasIntelligent Transportation Society of America (ITS America)*. 2001.

[211] Kevin Smith and Brian L Smith. "Forecasting the clearance time of freeway accidents". In: (2002).

[212] Edward C Sullivan. "New model for predicting freeway incidents and incident delays". In: *Journal of Transportation Engineering* 123.4 (1997), pp. 267–275.

[213] Homa Taghipour et al. "A novel deep ensemble based approach to detect crashes using sequential traffic data". In: *IATSS Research* 46.1 (2022), pp. 122–129. ISSN: 0386-1112. DOI: https://doi.org/10.1016/j.iatssr.2021.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0386111221000455.

[214] Muhammad Atif Tahir, Josef Kittler, and Fei Yan. "Inverse random under sampling for class imbalance problem and its application to multi-label classification". In: *Pattern Recognition* 45.10 (2012), pp. 3738–3750.

[215] Jinjun Tang et al. "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review". In: *Analytic Methods in Accident Research* 27 (2020), p. 100123.

[216] Jinjun Tang et al. "Traffic incident clearance time prediction and influencing factor analysis using extreme gradient boosting model". In: *Journal of Advanced Transportation* 2020 (2020).

[217] Yuchun Tang et al. "SVMs Modeling for Highly Imbalanced Classification". In: *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39 (Mar. 2009), pp. 281 –288. DOI: 10.1109/TSMCB.2008.2002909.

[218] Athanasios Theofilatos et al. "Predicting Road Accidents: A Rare-events Modeling Approach". In: *Transportation Research Procedia* 14 (2016). Transport Research Arena TRA2016, pp. 3399–3405. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2016.05.293. URL: https://www.sciencedirect.com/science/article/pii/S235214651630299X.

[219] Tatiana Tommasi et al. "A deeper look at dataset bias". In: *Domain adaptation in computer vision applications* (2017), pp. 37–55.

[220] TomTom. *TomTom*. URL: https://www.tomtom.com/ (visited on 02/02/2021).

[221] Federal Highway Administration of the United States Department of Transportation. *The Manual on Uniform Traffic Control Devices*. 2017. URL: https://mutcd.fhwa.dot.gov/htm/2009/part6/part6i.htm.

[222] U.S. Department of Transportation (USDOT)/Bureau of Transportation Statistics (BTS). *National Transportation Atlas Database (NTAD)*. 2020. URL: https://www.bts.gov/ntad.

[223] Martin Treiber and Arne Kesting. "Traffic flow dynamics". In: *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg* (2013), pp. 983–1000.

[224] Ludovic Trottier, Philippe Giguere, and Brahim Chaib-Draa. "Parametric exponential linear unit for deep convolutional neural networks". In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2017, pp. 207–214.

[225] Takahiro Tsubota et al. "Effect of Road Pavement Types and Ages on Traffic Accident Risks". In: *Transportation Research Procedia* 34 (2018). International Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)Emerging Transport Technologies for Next Generation Mobility, pp. 211–218. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2018.11.034. URL: https://www.sciencedirect.com/science/article/pii/S2352146518303235.

[226] Gaetano Valenti, Maria Lelli, and Domenico Cucina. "A comparative study of models for the incident duration prediction". In: *European Transport Research Review* 2.2 (2010), pp. 103–111.

[227] Gaetano Valenti, Maria Lelli, and Domenico Cucina. "A comparative study of models for the incident duration prediction". In: *European Transport Research Review* 2.2 (2010), pp. 103–111.

[228] Sebastian Vallejos et al. "Mining Social Networks to Detect Traffic Incidents". In: *Information Systems Frontiers* 23 (Feb. 2021). DOI: 10.1007/s10796-020-09994-3.

[229] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[230] Eleni I Vlahogianni and Matthew G Karlaftis. "Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties". In: *Computer-Aided Civil and Infrastructure Engineering* 28.6 (2013), pp. 420–433.

[231] David Waetjen and Fraser Shilling. "Rapid Reporting of Vehicle Crash Data in California to Understand Impacts from COVID-19 Pandemic on Traffic and Incidents". In: (2021).

[232] Behram Wali, Asad J Khattak, and Jun Liu. "Heterogeneity Assessment in Incident Duration Modelling: Implications for Development of Practical Strategies for Small and Large Scale Incidents". In: ().

[233] Li-Li Wang, Henry Y.T. Ngan, and Nelson H.C. Yung. "Automatic incident classification for large-scale traffic data by adaptive boosting SVM". In: *Information Sciences* 467 (2018), pp. 59 –73. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2018.07.044. URL: http://www.sciencedirect.com/science/article/pii/S0020025518305681.

[234] Senzhang Wang, Jiannong Cao, and Philip Yu. "Deep learning for spatio-temporal data mining: A survey". In: *IEEE transactions on knowledge and data engineering* (2020).

[235] Shi Wang, Ruimin Li, and Min Guo. "Application of nonparametric regression in predicting traffic incident duration". In: *Transport* 33.1 (2018), pp. 22–31.

[236] Shuangcheng Wang et al. "Research on a dynamic full Bayesian classifier for time-series data with insufficient information". In: *Applied Intelligence* (2022), pp. 1–17.

[237] Shujie Wang et al. "Traffic Accident Risk Prediction with Attention-based Multi-Modal Fusion Networks". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2021, pp. 2031–2040.

[238] Wenqun Wang, Haibo Chen, and Margaret Bell. "A study of the characteristics of traffic incident duration on motorways". In: *Traffic And Transportation Studies (2002)*. 2002, pp. 1101–1108.

[239] WenQun Wang, Haibo Chen, and MARGARET C Bell. "Vehicle breakdown duration modeling". In: *Journal of Transportation and Statistics* 8.1 (2005), p. 75.

[240] Yuxiang Wang et al. "GSNET: Graph-Structured Network for Traffic Scene Understanding". In: *IEEE transactions on pattern analysis and machine intelligence* (2021).

[241] Brad Warner and Manavendra Misra. "Understanding neural networks as statistical tools". In: *The american statistician* 50.4 (1996), pp. 284–293.

[242] Chien-Hung Wei and Ying Lee. "Sequential forecast of incident duration using artificial neural network models". In: *Accident Analysis and Prevention* 39.5 (2007), pp. 944–954.

[243] Tao Wen et al. "Integrated incident decision-support using traffic simulation and data-driven models". In: *Transportation research record* 2672.42 (2018), pp. 247–256.

[244] Yuan Wen et al. "Traffic Incident duration prediction based on k-nearest neighbor". In: *Applied Mechanics and Materials*. Vol. 253. Trans Tech Publ. 2013, pp. 1675–1681.

[245] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1 (1987). Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, pp. 37 –52. ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(87)80084-9. URL: http://www.sciencedirect.com/science/article/pii/0169743987800849.

[246] Minsu Won. "Outlier analysis to improve the performance of an incident duration estimation and incident management system". In: *Transportation research record* 2674.5 (2020), pp. 486–497.

[247] Bichen Wu et al. "Visual transformers: Token-based image representation and processing for computer vision". In: *arXiv preprint arXiv:2006.03677* (2020).

[248] Wei-wei Wu, Shu-yan Chen, and Chang-jiang Zheng. "Traffic incident duration prediction based on support vector regression". In: *ICCTP 2011: Towards Sustainable Transportation Systems*. 2011, pp. 2412–2421.

[249] Zonghan Wu et al. "Graph wavenet for deep spatial-temporal graph modeling". In: *arXiv preprint arXiv:1906.00121* (2019).

[250] Ying Xia and Jungang Chen. "Traffic flow forecasting method based on gradient boosting decision tree". In: *2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*. Atlantis Press. 2017, pp. 413–416.

[251] Siyao Xiao. "Traffic accident duration prediction based on natural language processing and a hybrid neural network architecture". In: *2021 International Conference on Neural Networks, Information and Communication Engineering*. Vol. 11933. SPIE. 2021, pp. 194–202.

[252] Hong Yang et al. "Development of an automated approach for quantifying spatio-temporal impact of traffic incidents". In.

[253] George Yannis et al. "Use of Accident Prediction Models in Road Safety Management – An International Inquiry". In: *Transportation Research Procedia* 14 (2016). Transport Research Arena TRA2016, pp. 4257–4266. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2016.05.397. URL: https://www.sciencedirect.com/science/article/pii/S2352146516304033.

[254] Jerry Ye et al. "Stochastic gradient boosted distributed decision trees". In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 2061–2064.

[255] Dewei Yi et al. "A machine learning based personalized system for driving state recognition". In: *Transportation Research Part C: Emerging Technologies* 105 (2019), pp. 241–261.

[256] B Yu et al. "A comparison of the performance of ANN and SVM for the prediction of traffic accident duration". In: *Neural Network World* 26.3 (2016), p. 271.

[257] Bin Yu and Zhengfeng Xia. "A Methodology for Freeway Incident Duration Prediction Using Computerized Historical Database". In: July 2012, pp. 3463–3474. ISBN: 978-0-7844-1244-2. DOI: 10.1061/9780784412442.351.

[258] Rose Yu et al. "Deep learning: A generic approach for extreme condition traffic forecasting". In: *Proceedings of the 2017 SIAM international Conference on Data Mining*. SIAM. 2017, pp. 777–785.

[259] Jing Yuan et al. "Hetero-ConvLSTM: A Hierarchical Spatiotemporal Network for Accident Risk Prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 20.3 (2018), pp. 764–776.

[260] Min Zeng et al. "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data". In: *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*. IEEE. 2016, pp. 225–228.

[261] Xiaosi Zeng and Praprut Songchitruksa. "Empirical Method for Estimating Traffic Incident Recovery Time". In: *Transportation Research Record* 2178.1 (2010), pp. 119–127. DOI: 10.3141/2178-13. eprint: https://doi.org/10.3141/2178-13. URL: https://doi.org/10.3141/2178-13.

[262] Chengjun Zhan, Albert Gan, and Mohammed Hadi. "Prediction of Lane Clearance Time of Freeway Incidents Using the M5P Tree Algorithm". In: *IEEE Transactions on Intelligent Transportation Systems* 12 (Dec. 2011), pp. 1549–1557. DOI: 10.1109/TITS.2011.2161634.

[263] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. "Exploring event-driven dynamic context for accident scene segmentation". In: *IEEE Transactions on Intelligent Transportation Systems* 23.3 (2021), pp. 2606–2622.

[264] Zhiyuan Zhang, Zhiwei Chen, and Xianpei Zhu. "Deep learning based incident detection from social media data". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 1059–1064.

[265] Yuexu Zhao and Wei Deng. “Prediction in traffic accident duration based on heterogeneous ensemble learning”. In: *Applied Artificial Intelligence* 36.1 (2022), p. 2018643.

[266] Qikang Zheng et al. “Investigating the predictability of crashes on different freeway segments using the real-time crash risk models”. In: *Accident Analysis and Prevention* 159 (2021), p. 106213.

[267] Ruiming Zhou et al. “Foresee: A Multi-Scale Multi-Task Network for Accident Risk Prediction”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2020, pp. 2021–2030.

[268] Ruiming Zhou et al. “RiskOracle: A Graph-Convolution Network for Fine-Grained Accident Risk Prediction”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2020, pp. 1437–1446.

[269] Weiwei Zhu et al. “Dynamic prediction of traffic incident duration on urban expressways: A deep learning approach based on LSTM and MLP”. In: *Journal of intelligent and connected vehicles* 4.2 (2021), pp. 80–91.

[270] Apostolos Ziakopoulos. “Spatial analysis of harsh driving behavior events in urban networks using high-resolution smartphone and geometric data”. In: *Accident Analysis and Prevention* 157 (2021), p. 106189.

[271] Fang Zong, Hongguo Xu, and Huiyong Zhang. “Prediction for traffic accident severity: comparing the Bayesian network and regression models”. In: *Mathematical Problems in Engineering* 2013 (2013).

[272] Yajie Zou, Kristian Henrickson, and Lord. “Application of finite mixture models for analysing freeway incident clearance time”. In: *Transportmetrica A: Transport Science* 12.2 (2016), pp. 99–115.

[273] Yajie Zou et al. “Application of the bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management”. In: *Journal of advanced transportation* 2021 (2021), pp. 1–9.

[274] Yajie Zou et al. “Jointly analyzing freeway traffic incident clearance and response time using a copula-based approach”. In: *Transportation research part C: emerging technologies* 86 (2018), pp. 171–182.