Technische Universität München
Master′s Program in Transportation Systems

Master′s Thesis

Real-time Driving Intention Prediction and Crash Risk Estimation from Naturalistic Driving Data using Machine learning

Vishal Mahajan

Supervised by:

Dr. Christos Katrakazas
Univ.-Prof. Dr. Constantinos Antoniou

October 21^{st}, 2019

Disclaimer

I hereby confirm that the presented thesis work has been done independently and using only the sources and resources as are listed. This thesis has not previously been submitted elsewhere for purposes of assessment.

October 21^{st}, 2019 Vishal Mahajan

Acknowledgments

I would like to express my sincere appreciations to Dr. Christos Katrakazas and Prof. Constantinos Antoniou for supervising my thesis. Their guidance and valuable feedback throughout the work was crucial in the successful completion of the work.

I will always be grateful to my parents and my brother for their love and unceasing support.

Abstract

Road crashes are one of the critical issues in the road transportation. The crash studies are done to estimate the likelihood of a crash on a road section, whereas crash impact has received relatively less attention. The focus of the most studies is to establish relationship of crashes with the factors mainly traffic effects. The analysis of microscopic factors such as individual driving intention/ maneuver has not received that attention due to requirement of a large disaggregate data. Identification and prediction of driving intention is fundamental for avoiding collisions as it can provide useful information to drivers and vehicles in their vicinity.

This study utilizes a large real traffic dataset Highway Drone Dataset (HighD Dataset) collected on German Freeways. A driving intention prediction model using clustering for automatic labelling and supervised deep learning for real time prediction is introduced and validated. The model is able to predict the lane change intention on an average 3 seconds before the vehicle crosses the lane. This study introduces a new Surrogate Safety Measures (SSM) to estimate rear-end crash severity and uses it along with likelihood indicator - Modified Time to Collision (MTTC) to determine the total crash risk. This study analyzes the effects of traffic conditions and driving intention on the estimated crash risk. The results of the study indicate that the traffic congestion plays a significant role in increasing the likelihood and severity of the rear-end crashes. As far as driving intention is concerned, lane changing is found to be associated with increase in likelihood and severity of a rear-end crash during free-flow conditions. This study is an attempt to develop a comprehensive risk estimation method and analyze its dependence on macroscopic and microscopic factors of the crash. The utilization of large unlabeled and disaggregate data increases the prospects of transferability of the approach and its practical application for highway safety by researchers and practitioners.

Contents

List of Figures

List of Tables

List of Abbreviations

ACI	Average Crash Impact
ACL	Average Crash Likelihood
CI	Crash Index
CIF	Criticality Index Function
CNN	Convolutional Neural Networks
CRIM	CRash IMpact
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DH	Distance Headway
DRAC	Deceleration Required to Avoid Collision
EDA	Exploratory Data Analysis
FD	Fundamental Diagram
GM	Gaussian Mixtures
HighD Dataset	Highway Drone Dataset
LIDAR	Light Detection and Ranging
LSTM	Long Short Term Memory
MDN	Mixture Density Networks
ML	Machine Learning
MTTC	Modified Time to Collision
NDS	Naturalistic Driving Studies
NGSIM	Next Generation SIMulation
NN	Neural Network
OR	Odds Ratio
PCA	Principal Component Analysis
RF	Random forest
RNN	Recurrent Neural Network
RVM	Relevance Vector Machines
SSM	Surrogate Safety Measures
SVM	Support Vector Machines
TH	Time Headway
TTC	Time to Collision

Part I.

Introduction and Background Theory

1. Introduction

Road crashes are one of the most critical issues in the transportation sector. According to WHO (2018), Road crashes are the number one cause of injury related deaths in the world. This results in loss of millions of lives each year as well as incurs social and economic costs associated with such crashes. Every year more than 1.35 million people die in road accidents whereas between 20 million - 50 million suffer injuries and in some cases life-long disabilities (WHO, 2018). These accidents cause significant social and economic costs to the society. WHO (2018) estimated that for most countries, road crashes related costs can reach upto 3% of the GDP and therefore countries all around the world are trying to bring down the road crash related fatalities and injuries. For example, the European Union has adopted "Vision Zero" approach to eliminate the deaths and injuries in the road accidents and make European roads safer (European Union, 2019).

In case of a vehicle collision, the terms "crash" and "accident" are used interchangeably, but the term accident implies something which was accidental and not preventable. Thus, crash is a preferred term to imply something which can be prevented, unlike accidents. Treat et al. (1979) categorized the crash causing factors under three categories namely, Human, Environment and Vehicular factors. Human factors refer to the crash causes attributable due to driver inattention, wrong maneuver, improper driving, etc. Later, Oh et al. (2005) introduced Traffic dynamics also as one of the contributing factors to the road crashes. Out of these crash factors, Treat et al. (1979) identified driver factors as the most probable cause of crashes. Even today, 94% of the road accidents are related to human errors and bad decisions (National Highway Traffic Safety Administration, 2016).

The driver behaviour and traffic characteristics are highly dynamic factors and thus significant from the point of view of real-time road safety. On the aggregated level, certain traffic patterns or states can also be found to be pre-cursor to crash conditions (Abdel-Aty and Pande, 2007). Traffic state refers to characteristics of the traffic flow, speed and density on a road segment. Oh et al. (2005) argued that instabilities in the traffic state causes road accidents and preventing such traffic conditions can help in reducing the chances of a crash. Detection of such dangerous traffic states such as congestion can help to decide appropriate traffic management strategies and reduce the chances of a crash. This has found applications in Intelligent Transportation Systems (ITS) such as Variable Speed Limits and Ramp Metering which are found to have significant safety benefits by reducing the likelihood of a crash (Abdel-Aty and Pande, 2008).

The availability of the traffic data is crucial factor in crash studies and these studies can be classified based on the data collection methods into three categories, 1) Epidemiological driving studies, 2) Naturalistic Driving studies and 3) Empirical Driving studies (Fitch and Hanowski, 2012). The Epidemiological driving studies use crash databases for assessment of crash risk. The Empirical studies are done in controlled environments with the help of simulators or test track. In empirical studies, safety surrogates are used for crash risk estimation. The Naturalistic Driving Studies (NDS) are based on the in-situ driver behaviour investigation using cameras and sensors and have more validity than empirical studies (Fitch and Hanowski, 2012).

Traditionally, the researchers used data from historical crash records and traffic loop detectors to obtain road section parameters such as average speed, traffic flow, etc. These parameters are aggregated over space and time and that's why the studies using this data are referred to as aggregate studies (Abdel-Aty and Pande, 2007, Xu et al., 2013). In these studies, limited information about the pre-crash driver behaviour is available. The study of driver behaviour has benefited from the relatively new and advanced methods of traffic data collection such as video cameras Alexiadis et al., 2004, Light Detection and Ranging (LIDAR) (Zyner et al., 2018, Herty and Visconti, 2018), drone videography (Krajewski et al., 2018, Chen, Zeng, et al., 2017), smartphones (Guido et al., 2012), on-board devices (McCall and Trivedi, 2007), etc. These new methods help in obtaining detailed naturalistic driving data at disaggregate level. The naturalistic driving refers to the driver performance and behaviour in real world scenario. The disaggregate data has made possible to conduct traffic analysis at the vehicle level by utilizing data such as vehicle motion, steering, braking, speed, driver's alertness, etc. These attributes from this data can be utilized to estimate or predict the vehicle trajectory, maneuver or collision risk (Lefèvre, Vasquez, et al., 2014) in immediate future for tactical and operation driving decisions. The tactical and operational decisions are decision for short-term or instantaneous future such as braking, acceleration, lane changing, etc. Since the the unit of analysis is a vehicle, these study of human/ driver factors can be referred to as individual/ disaggregate studies (Abdel-Aty and Pande, 2007).

The disaggregate studies play an important role in understanding and modelling the driver behaviour under different driving context such as lane changing (Zheng, 2014). This has played a significant role in the design of vehicle safety systems and other advanced road safety applications such as Cooperative and Autonomous driving which relies on information about surrounding vehicle behaviour (V2V: Vehicle-To-Vehicle systems) and infrastructure (V2I: Vehicle-To-Infrastructure systems) (Andrews, 2012, Özgüner et al., 2007). Some of the applications of V2V systems are intersection collision warning, lane change warning, pre-crash sensing. The real time traffic management infrastructure is now more advanced than ever which has applications in V2I systems such as traffic congestion, accident, road closure warning, etc. These systems also called as intelligent safety system, enhance the safety by increasing the driver's situation awareness and predicting their actions (McCall and Trivedi, 2007). **The identification of the driver behaviour in the context of traffic dynamics and its effect on crash risk can help in gaining further insights in the driving behaviour for appli-**

cations in real-time safety.

1.1. Problem Definition

As discussed in the previous section, the dynamic factors such as driver behaviour and traffic conditions are important determinants of crashes. The identification of these factors is crucial for making safe driving decisions in real-time and estimating the prevailing risk. Risk is defined as the product of probability and severity of a collision for the subject vehicle (Lefèvre, Vasquez, et al., 2014). The analysis of impact of driver and traffic factors on crash risk is important to understand the underlying interactions, in view of highly complex driving behaviours and traffic situations.

The disaggregate data has found multiple applications in driver behaviour identification and prediction. The driver behaviour is manifested on the road by the driving intentions. Driving intention is the preference of one driving maneuver over the available set of driving maneuvers. For example, Lane changing and lane keeping are the most common lateral driving maneuvers on a freeway (Gayko, 2012). The study on these maneuvers usually follows these steps: 1) Selection of maneuvers, 2) Collection of naturalistic driving data, 3) Manual Labelling of the data in terms of maneuvers/ intentions, 5) feature extraction and formulation of the classification problem and 4) Model building to predict the driving intentions and model evaluation (Mandalia and Salvucci, 2005, Morris et al., 2011).

On one hand, availability of the large naturalistic driving data-sets has become relatively easier, but the manual labeling of driving maneuvers is a roadblock towards use of large data-sets due to associated labelling costs and time. This slows the applications of large data-sets for driver behaviour analysis. **Therefore, there is a need for unsupervised classification which can help to automate the maneuver labeling process**. Labeling of maneuvers can help to analyze the crash risk with respect to driving intentions using a large data. This can also help to study the interactions effects of driving intention and traffic state on crash risk which has not been investigated much before. Data driven labeling based on unsupervised classification also holds the promise for its application in real time prediction models with much easier validation and transferability of results, which is difficult with manual labeling.

The crash studies use regression or a classification formulation of the risk. In either formulation, the aim of the studies is to identify when a crash is likely to occur based on the past crash data and the conditions prevailing at that time (Abdel-Aty and Pande, 2007). The main emphasis of the crash data based studies is estimation of the risk in terms of the likelihood whereas crash severity has not received as much attention. This could be due to issues related to under reporting of less severe accidents in developed countries and developing countries (Naji and Djebarni, 2000), which makes the fatal/ severe accidents as the main focus the studies. Further, the less frequency of the crashes necessitate long duration of the data collection, which has attracted the use of micro-simulations or naturalistic driving data for safety evaluation via SSM (Mahmud et al., 2017). The SSM are the measure of the temporal and spatial proximity of the two or more vehicles and are used to estimate a crash risk. Similar to data based studies, SSM are also mainly focused on the crash likelihood. By treating all the likely conflicts with equal severity, this could lead to over representation of the high impact conflicts. The SSM formulations used by Ozbay et al. (2008) and Chan (2006) combine the likelihood and severity factor in one indicator, but the SSM to measure the severity separately (independent of likelihood) are not found in the available literature, especially for rear-end conflicts. Marchesini and Weijermars (2010) also emphasized the need to further investigate the traffic effects on the crash severity.**Therefore, there is a need to introduce a new surrogate for rear-end crash severity**.

In view of the introduction of a fresh surrogate for crash severity and automatically labelled driving maneuvers, there is a need to holistically determine the aggregate crash risk and analyze its correlation with driving intention. This is identify if certain driving intentions are associated with higher risk in a particular traffic scenario. **The method to estimate aggregated crash risk in terms of separate likelihood and severity surrogates is needed. This can help to better understand the interaction effects of the traffic state and driving intention, which have not been investigated before using a large dataset.**

1.2. Research Questions, Aim and Objectives of the study

Based on the problems discussed above, there are three research questions of this study:

- Can unsupervised classification methods help to label the raw data accurately for its utilization in predicting real time driving intention?
- How to take into account crash severity using surrogate measures for risk assessment?
- Which driving intentions and traffic conditions are more risky?

To answer these research questions, The primary aim of this study is to extract driving maneuvers from naturalistic driving data to automatically label the driving data and evaluate the labelling by building a prediction model for the driving intention. The secondary aim is to develop a method for crash risk estimation by considering severity and then evaluate the correlation of risk with the driving intention and traffic conditions. The research aim will be achieved through objectives listed in table 1.1

Table 1.1.: Research Objectives and Methods

Objective	Method	Chapter
To review the highway crash risk estimation methods	Literature review	2
To review the driving intention/ maneuver prediction methods	Literature review	2
To label a trajectory data-set in terms of the lateral driving maneuver namely lane changing and lane keeping on highways	Application of clustering algorithms to identify the driving maneuver classes and labeling	3
To Predict real-time driving intention namely lane changing and lane keeping on highways	Application of deep learning sequence prediction model for real-time driving maneuver prediction	3
To estimate aggregated crash risk in terms of likelihood and impact	Development of the methodology for quantifying aggregated crash risk from trajectory data using surrogate safety measures	3
Evaluate the safety impact of traffic state using aggregated crash risk	Application of clustering algorithms to identify traffic state into free-flow and congestion and estimation of crash risk for these states and subsequent evaluation of the risk using significance tests	3
Evaluate the safety impact of the driving maneuver using the aggregated crash risk	Estimation of the crash risk for the lane changing and lane keeping driving intention and subsequent evaluation of the risk using significance tests	3

1.3. Contributions

This study has attempted to make following contributions towards achieving its aims:

- Data-driven approach for unsupervised labeling of lateral driving maneuvers into lane keeping and lane-changing from a large naturalistic driving data. The approach is envisioned to reduce effort of manually labeling driving manoeuvres, in order for them to be classified efficiently.
- An efficient real-time intention prediction using deep learning model to proactively predict lane-changing intentions.
- Development of an aggregated crash risk indicator for rear-end collision in terms of likelihood and severity surrogates and its application on the freeway driving data.
- Analysis of rear-end crash risk for lane changing intention and traffic congestion.

1.4. Research Framework and Thesis Outline

The research framework developed for this study is shown in figure 1.1. Firstly, a preliminary study (Current chapter 1) is conducted for problem identification, framing of the research question and objectives. Afterwards, a detailed literature review (Chapter 2) on driving intention prediction and crash risk estimation methods. The review is aimed to identify research gaps. The is followed by methodology (Chapter 3). The first and second part of the methodology deals with generic machine learning pipeline which is adapted for prediction of date for driving maneuvers and identifying traffic states. The third part presents the method for aggregate risk estimation by selection/ formulation of likelihood and severity surrogates to estimate rear-end crash. The last part of methodology describes the method for hypothesis testing for significance of traffic and intention for increased risk. Thereafter, Data collection and analysis section (Chapter 4,5) describes the data-set used to this study and results of the Exploratory Data Analysis (EDA). The results (Chapter 6) of the driving maneuver labeling and intention prediction model are shown and compared with the literature. Thereafter, results from estimated crash risk and significance of intention and traffic state is presented. Finally, the findings of the study are summarized and its applications, limitations and future research directions are presented (Chapter 7).

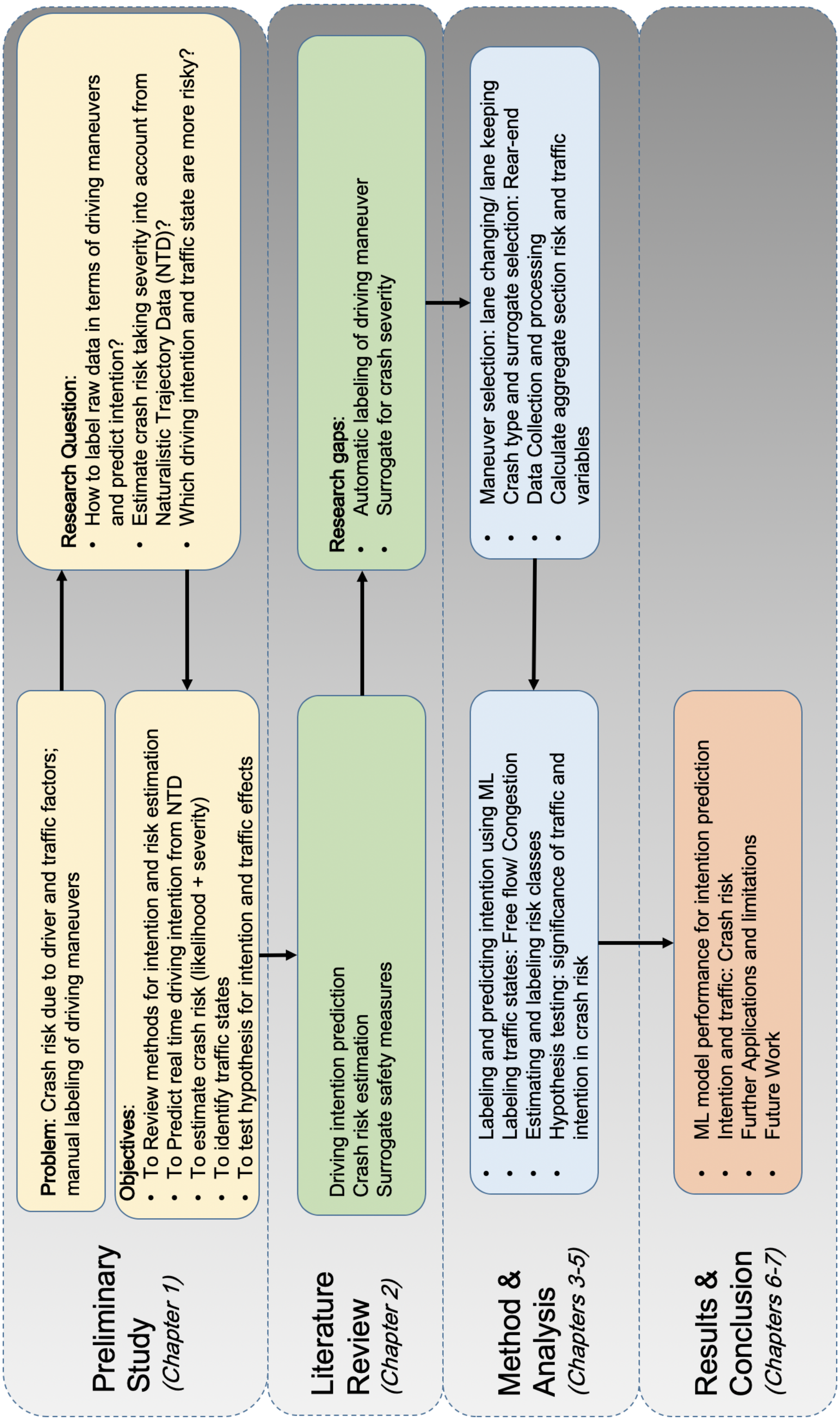
Preliminary Study
(Chapter 1)
Problem: Crash risk due to driver and traffic factors; manual labeling of driving maneuvers
Objectives:
• To Review methods for intention and risk estimation
• To Predict real time driving intention from NTD
• To estimate crash risk (likelihood + severity)
• To identify traffic states
• To test hypothesis for intention and traffic effects
Research Question:
• How to label raw data in terms of driving maneuvers and predict intention?
• Estimate crash risk taking severity into account from Naturalistic Trajectory Data (NTD)?
• Which driving intention and traffic state are more risky?
Literature Review
(Chapter 2)
Driving intention prediction
Crash risk estimation
Surrogate safety measures
Research gaps:
• Automatic labeling of driving maneuver
• Surrogate for crash severity
Method & Analysis
(Chapters 3-5)
• Labeling and predicting intention using ML
• Labeling traffic states: Free flow/ Congestion
• Estimating and labeling risk classes
• Hypothesis testing: significance of traffic and intention in crash risk
• Maneuver selection: lane changing/ lane keeping
• Crash type and surrogate selection: Rear-end
• Data Collection and processing
• Calculate aggregate section risk and traffic variables
Results & Conclusion
(Chapters 6-7)
• ML model performance for intention prediction
• Intention and traffic: Crash risk
• Further Applications and limitations
• Future Work

Figure 1.1.: Research Framework

2. Literature Review

This chapter on literature review is aimed to comprehend the research gaps in the area of driving intention and crash risk estimation. It encompasses definitions, state of art methods on these topics and shortcomings faced by the researchers. The chapter starts with a brief introduction of the concepts related to driving, methods for driving intention estimation and prediction. This is followed by a second section on crash risk estimation wherein methods and the tools have been discussed. Lastly, the relevant surrogates for measuring the crash risk for a rear-end crash are discussed for identification of areas of further improvement.

2.1. Driving Intention

Driving is a skilled task which needs understanding of the environment to make suitable driving decisions. A driver perceives the environment and processes the information in the brain to make a logical decision. FHWA (2004) describes driver behaviour and decision making at five different levels:

1. Pre-trip: Decisions made before starting a trip such as destination choice, mode choice, departure time choice

2. Strategic en-route: Decisions made during a trip at the time scale of minutes such as route choice. These decisions are executed at a time scale of minutes.

3. Tactical route execution: These are small and multi-part decisions executed at a time scale of few seconds. This includes decisions such as lane changing and overtaking.

4. Operational Driving: Decisions made at near-instantaneous scale for immediate actions during the trip such as acceleration.

5. Vehicle Control: The instantaneous decisions to control the vehicle such as steering.

The driver behaviour modelling for microscopic traffic models and real-time safety applications is done at the tactical route execution and operational driving level (Hamdar, 2012). According to Hamdar (2012), the models for freeway traffic at the tactical level correspond to lane changing, merging, overtaking, etc. Driver behaviour prediction at tactical level is significant for real-time motion planning (Katrakazas et al., 2015). The tactical decision making is also relevant for crash risk analysis since these decisions correspond to the driving actions with consequences in immediate future (few seconds).

The vehicle motion at the tactical level consists of driving maneuvers. Lefèvre, Vasquez, et al. (2014) defines driving maneuver as ”the physical movement or series of moves requiring skill and care”. Vehicle manoeuvres can also be understood as characterizations of a vehicles motion on road with regards to its position and speed attributes (Katrakazas et al., 2015). While driving, a human or autonomous driver constantly estimates the probable maneuvers of other drivers in order to take the decision to select the best possible maneuver. For example, when a driver notices the turn indicators of the preceding vehicle, he will become aware of the preceding driver's intention to change lane or turn, and accordingly plan his manoeuvre such as braking or steering around. Therefore, **Driving intention** can be defined as the maneuver preferred by the driver over a set of available manoeuvres based on the perceived surroundings. According to Lefèvre, Vasquez, et al. (2014), maneuver level reasoning extracts high level characteristics which leads to reliable estimation of long-term motion and risk and also easier to generalize.

With regards vehicle lateral control on highways, lane keeping and lane changing are fundamental aspects of highway driving and are mutually exclusive maneuvers (Gayko, 2012). Lane keeping describes the task of driving within the current lane without any intention of leaving it. In other words, if the vehicle stays in the same lane, it will belong to the lane keeping maneuver. On the other hand, lane changing is the manoeuvre associated with leaving the current lane and entering into new lane. A vehicle's manoeuvre is referred to as lane changing if during the recorded trajectory, a full lane change manoeuvre takes place. Figure 2.1 illustrates a lane changing process so as to distinguish between the two classes. A vehicle is moving straight on the highway in a specific lane, and then decides to change. At Point B, the vehicle starts executing a lane change maneuver which is described by its lateral movement. At point C, the lane changing event occurs as the vehicle crosses from one lane to the other, while at point D, the driver completes the lane change process returns to lane keeping. Therefore, the trajectory of the vehicle consists of two lane keeping stages (AB and DE) and one lane changing stage (BD). A lane change is fully executed if all points (i.e. B, C and D) lie within the observed section of the highway.

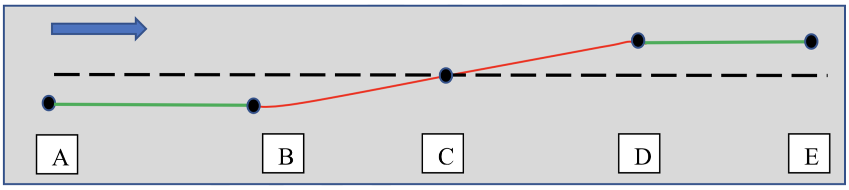

Figure 2.1.: Illustration of lane keeping (AB and DE) and lane changing processes (BD) of a vehicle

Execution of a lane changing requires inter-related decisions among drivers in a definite hierarchy which are affected by the necessity and desirability of changing the lane in which a vehicle is moving (Gipps, 1986). Lane changes can be discretionary or mandatory depending on how the driver perceives conditions on the current and target lanes as well as environmental conditions which influence the decision of changing lanes (Ahmed, 1999). The examples of discretionary lane change can be overtaking or passing whereas the mandatory lane change might be necessitated due to lane closures, merging, etc. During lane changing usually the speed of the subject vehicle increases, but disturbances in the adjacent lane might be observed e.g., if the lane change manoeuvre is aggressive (Ioannou and Stefanovic, 2005). To counteract for these disturbances, and enable safe lane changing, drivers interacting for a lane change should be alert and attentive. However, this is not always a fact as in (Lee, Olsen, et al., 2004) it was indicated that turn indicators are used for 44% of performed lane changes.

The table 2.1 lists the details of the reviewed literature for the data driven intention prediction. The study of the driver intention is studies using the microscopic data at the individual driver or vehicle level. The microscopic data usually consists of the GPS traces (Yang et al., 2018), or vehicle sensor such as radars and cameras are used (Morris et al., 2011; Özgüner et al., 2007). For this purpose, data from traffic videos or real-world driving is used to capture the naturalistic driving. The studies by Hu et al. (2018) and Woo et al. (2017) used real world vehicle trajectories extracted from the video recordings. The data-sets from traffic videos observe a limited section of a road for some time and thus capture the real world behaviour of the vehicles driving on that road section.These data-sets are generally large and capture the behaviour of neighbouring vehicles and traffic interactions also. Some studies use driving data captured by observing certain/ selected drivers over a period of time. Some of the examples of studies using data from selected drivers are Leonhardt and Wanielik (2017), Mandalia and Salvucci (2005), Morris et al. (2011), and Wang, Murphey, et al. (2016) with the advantages that driver behaviour is observed for a long period of time, but it does not include interactions with neighbouring vehicles (Fitch and Hanowski, 2012). The simulated or controlled environments such as fixed driving simulators and traffic simulation are also used to study the intentions. In studies by Dang et al. (2017), Wissing et al. (2017), and Xing and Xiao (2018), the vehicle/ trajectory data extracted from the simulated environment is used for intention analysis.

The driving intention prediction tasks, can be formulated as a regression or a classification problem as mentioned in table 2.1. The classification formulation is mostly common among the intention prediction. It is applied by discretizing vehicle states in terms of manoeuvres, such as lane changing and lane keeping, which enables faster real-time discrimination between maneuver intentions. Sometimes, regression formulation is also combined with the classification set-up to predict indicators such as time to lane change (Dang et al., 2017; Wissing et al., 2017). The trajectory prediction also uses regression formulation to predict position and speeds values, so as to map vehicle motion on the road as in trajectory prediction. But the literature on trajectory prediction is not discussed here, since intention estimation is indirectly estimated from the predicted trajectories as a secondary step.

Machine learning and data-driven approaches are popular for maneuver prediction. It can be seen form the table that Support vector machines (Mandalia and Salvucci, 2005; Woo et al., 2017), Relevant vector machines (Morris et al., 2011), Neural networks (Leonhardt and

Table 2.1.: Literature Review on Data driven Driving Intention prediction on Freeway

Analyzed/ Predicted intention	Formulation		Model	Data used	Manual Labelling	Labelling basis	Prediction (seconds)	Source
	Regression	Classification						
Lane change		✓	SVM	Real-driving	✓	Lane markings	0.3	Mandalia and Salvucci, 2005
Lane change, deceleration		✓	RNNs	Simulation	✓	Lane markings, vehicle motion	-	Xing and Xiao, 2018
Lane-change/ turn		✓	CNNs and LSTMs	vehicle sensor	✓	Lane markings	-	Beglerovic et al., 2018
Lane change			NN and MDN	Traffic	✓	Lane markings	2	Hu et al., 2018
Lane change	✓	✓	LSTM	Driving Simulator	✓	Lane markings	2.7	Dang et al., 2017
Lane change		✓	CNN	Real-driving	✓	Lane markings	0.4 - 1	Wang, Murphey, et al., 2016
Lane change		✓	RVM	Real-driving	✓	Lane markings	3	Morris et al., 2011
Lane change	✓	✓	SVM	Driving Simulator	✓	Controlled experiment	1.6	Wissing et al., 2017
Lane change		✓	SVM	Traffic	✓	Controlled sampling	1.7	Woo et al., 2017
Lane change		✓	NN	Real-driving	✓	Maneuver type	2	Leonhardt and Wanielik, 2017

Support Vector Machines (SVM),Relevance Vector Machines (RVM), Neural Network (NN), Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN)
Long Short Term Memory (LSTM), Mixture Density Networks (MDN)

Wanielik, 2017), Convolutional neural networks (Beglerovic et al., 2018; Wang, Murphey, et al., 2016), Recurrent Neural networks (Beglerovic et al., 2018; Dang et al., 2017; Xing and Xiao, 2018) and probabilistic models (Hu et al., 2018) are used for driving intention detection and prediction. These models have been able to accurately detect the driving intention namely, lane changing with a prediction horizon varying from 0.4 to 3 seconds. The prediction horizon is usually the time between the detection of lane changing intention and actual lane change (vehicle crossing the lane marking and entering a new lane). Mandalia and Salvucci (2005) presents prediction results in terms of the time before a maneuver is even executed and thus cannot be directly compared with other studies.

The performance of machine learning depends on the quality of the labelled data which is used for the model training. Therefore, adequate time and cost is invested to ensure the quality of the labelled data. The studies in table 2.1 usually use manually labelled data. The data is labelled for lane changing intentions by considering the position of vehicle with respect to the road layout or lane markings. The majority of the studies in table 2.1 use the lane markings as a reference to identify the lane change intention and subsequently label the data into lane changing and lane keeping classes. Morris et al. (2011) uses partially automated labelling but it also uses information about the deviation of the vehicle with respect to the lane marking. Wissing et al. (2017) and Woo et al. (2017) annotate the data under controlled experiment settings under manual supervision. The consideration of the road layout for labelling the lane changing intention can slowdown the generalization of the models to new or arbitrary layouts due to additional data processing. Therefore, the use of features corresponding to only vehicle motion to identify maneuver to change the lanes needs further investigation.

Research Gaps: Driving Intention

Based on the above literature review on driving intention, It is observed that the prominent practice is that the data are either manually annotated and labelled into lane keeping or lane changing instances based on the information such as lane position. This limits the development of lane change detection models from large naturalistic traffic data-sets. Consequently, there is a gap in the literature with regards to the utilization of large naturalistic driving datasets and automatic unsupervised labelling of data, for lane change detection for real-time applications. The automatic labelling would also enable use of large data-set for studying the driving intention for other applications such as crash risk, traffic conditions, etc.

2.2. Crash risk

Crash risk signifies the likelihood and the consequences of a crash for the subject vehicle. Therefore, estimation of crash requires estimation of both the likelihood and impact/ severity of a crash. The likelihood of a risk refers to the chances of the occurrence of a particular event. Here, probability, frequency and likelihood are used interchangeably to signify the chances of the happening of an event or a crash. The severity refers to the consequences or the impact of an event. Therefore, an crash of low likelihood and low severity implies a low risk crash. Furthermore, risk in high likelihood and high severity is extreme. The risk due to high likelihood-low severity and low likelihood-high severity crashes can be interpreted between levels of low risk and extreme risk.

2.2.1. Factors of a Crash

Crash risk in a particular situation depends on the four factors of a crash (Chapter 1). In this section, a more detailed study of the crash factors and their impact on safety is presented. According to Oh et al. (2005), the crash factors and their components are listed below:

1. Driver characteristics: Behaviour, driving skill, etc
2. Environment: Geometric design of the road, road safety furniture, weather conditions
3. Traffic dynamics: Traffic state characterized by traffic flow, speed and density
4. Vehicle characteristic: type, shape, size and condition of the vehicle

These four causal factors can also be understood in terms of their scale. The individualistic or microscopic factors such as driver or vehicle characteristics depend on the interactions of the one or two vehicles. In other words, If a crash happens strictly due to these factors, only the related vehicles are affected and the impact is generally localized. On the other hand, environment and traffic factors can impact the crash risk over a road link or even network. Due to this reason, the impact of these factors is global or macroscopic as it has an affect over all the vehicles in that situation.

The crash factors can also be classified based on its relationship with time. The factors such as road geometry, intersection, etc do not change with time on a daily basis and thus road safety analysis involving such static factors is referred to as off-line studies (Hourdos et al., 2006). The traditional crash investigation and road safety analysis such as accident studies, before-after studies and road safety audits also fall in this category. These studies are conducted by associating crash frequency with the geometric design to identify dangerous road sections.

The dynamic factors especially traffic state can vary greatly with time and thus the aim of the studies is the analysis of real time crash risk (Abdel-Aty and Pande, 2007). The traffic characteristics and dynamics such as congestion, speeds and type of the surrounding vehicles, etc are significant in determining the driver behaviour (Hamdar, 2012). Due to this reason, impact of traffic characteristics on crash likelihood has been extensively studied due to its applications in real time traffic control and management. Hourdos et al. (2006) described these studies as as online studies since they use real time measurements of the crash factors i.e., traffic state.

2.2.2. Crash-risk studies

These crash factors through complex interactions determine the crash risk. The risk is studied to understand, estimation or predict the risk under prevailing conditions on road. This understanding helps to implement traffic management measures such as variable speed limits, ramp metering, etc. to ensure that the risk is within the desirable limits and likelihood of an crash is reduced (Abdel-Aty and Pande, 2008).

In these studies, risk is assessed in a time varied manner i,e., identification of traffic conditions which might be collision prone. The table 2.2 summarizes the crash risk studies based on the objective of the study, components of risk, type of crash factors and input data. The detailed about these studies follows in the next paragraph.From the table, it can be seen that the main objective of the crash risk studies is the estimation, prediction or forecasting or classification. The estimation studies are aimed at the analysis and explanatory investigation of the risk of a crash. The crash studies are done at aggregated or disaggregate level of crash risk. The aggregated crash risk is defined for the road section or link, whereas disaggregate risk corresponds to individual vehicle, generally for a specific crash type.

Aggregate Studies

The studies on aggregated crash risk are most common for freeways or highways and are mostly done using crash data. Wang, Quddus, et al. (2009) and Zheng (2012) used the crash data to establish the association of the aggregated crash likelihood with the traffic dynamics. Few studies like the one by Stipancic et al. (2018) have used both the crash data and driving data, which is not found to be common probably due to data availability issues. The prediction of crash risk can be formulated as a regression or a classification problem. In regression formulation, the objective is the prediction of the crash occurrence by assigning a likelihood or probability to the prevailing traffic conditions (Hossain and Muromachi, 2013; Lee, Hellinga, et al., 2003; Oh et al., 2005; Xu et al., 2013; Zheng et al., 2010). In the classification formulation, the traffic conditions are assigned a discrete binary class such as crash vs no-crash conditions (Hourdos et al., 2006; Pande, Abdel-Aty, and Miyake, 2010). A study by Russo et al. (2016) developed off-line severity prediction models with features related to road geometry and average traffic volumes.

The aggregate risk analysis based on crash data is possible only if enough data is available which necessitates long periods of data collection. Then there is also a problem with synchronization of the crash occurrences and traffic flow data from loop detectors at different locations (Hourdos et al., 2006). The crash studies are commonly conducted for the crash

Table 2.2.: Characteristics of the road crash/ risk studies

Main Objective of the study	Components of Crash		Data			Factors of Crash/ Features				Scenario	Source
	Likelihood	Severity	Crash	Loop Detector	Trajectory	Microscopic		Macroscopic			
						Static	Dynamic	Static	Dynamic		
Crash rates forecasting	✓		✓						Traffic	Freeway	Golob, Recker, and Alvarez, 2004
Crash prediction	✓		✓						Traffic	Freeway	Xu et al., 2013
Crash type Classification	✓		✓	✓				Road	Traffic	Freeway	Pande, Abdel-Aty, and Das, 2010
Crash condition classification	✓		✓	✓					Traffic	Freeway	Oh et al., 2005
Crash occurrence prediction	✓		✓	✓					Traffic	Freeway	Zheng et al., 2010
Real-time crash prediction	✓		✓	✓					Traffic	Freeway	Lee, Hellinga, et al., 2003
Real-time crash prediction	✓		✓	✓					Traffic	Urban Freeway	Hossain and Muromachi, 2013
Crash likelihood analysis	✓		✓	✓					Traffic	Freeway	Wang, Quddus, et al., 2009
Crash likelihood analysis	✓		✓	✓					Traffic	Freeway	Zheng, 2012
Crash condition detection	✓		✓	✓				Pavement	Traffic, weather	Freeway	Hourdos et al., 2006
Crash Severity Analysis		✓	✓					Road	Traffic	Rural	Russo et al., 2016
Crash likelihood analysis	✓		✓		✓			Road	Traffic	Intersections	Stipancic et al., 2018
Rear End crash risk estimation	✓				✓				Traffic	Expressway	Kuang et al., 2017
Rear-end crash risk estimation	✓			✓		vehicle type	vehicle motion		Traffic	Urban	Dimitriou et al., 2018
Rear-end crash risk estimation	✓				✓		vehicle motion			Work Zone	Weng and Meng, 2014
Lane-change risk estimation	✓				✓		vehicle motion			Highways	Herty and Visconti, 2018
Lane-change risk estimation	✓				✓		vehicle motion			Freeway	Chen, Shi, et al., 2019
Lane-change risk estimation	✓				✓		vehicle motion			Work Zone	Park et al., 2018
Driver Intention risk	✓				✓		vehicle motion			Intersection	Lefèvre, Laugier, et al., 2013
Driving style classification					✓		vehicle motion			Freeway	Xue et al., 2019
Driver behaviour Classification					✓		vehicle motion			Intersections	Aoude et al., 2011
Driver behaviour Classification					✓		vehicle motion			Freeway	Cheung et al., 2018
Driver Intention Identification	✓				✓		vehicle motion			Intersections	Chen, Liu, et al., 2018

prone sections due to sufficient availability of data which might limit the generalization of the results to sections with less frequent accidents. The crash data also lacks attributes corresponding driver behaviour or intention before the crash. Due to this reason, the consideration of drivers intention is not common in aggregated studies since crash data is aggregated and therefore analysis at individual level is not possible (Abdel-Aty and Pande, 2007). Therefore, disaggregate studies rely on data from individual vehicle for a more detailed analysis.

Disaggregate Studies

At disaggregate level, studies are mostly done for analysis of crash risk rather than actual crash. This is why, the risk is generally estimated using safety surrogates (discussed on section 2.3) at level of individual vehicle. Further the studies focus at a specific crash type, which could possibly due to lack of detailed accident data by crash type. Since disaggregate studies need vehicle motion data to analyze the crash risk, these studies generally use naturalistic driving data or micro-simulations. The two types of crashes have received much attention namely rear-end crash and lane-change crashes. The study by Dimitriou et al. (2018), Kuang et al. (2017), and Weng and Meng (2014) estimated the rear-end crash likelihood using trajectory data or loop-detectors data. Chen, Shi, et al. (2019), Herty and Visconti (2018), and Park et al. (2018) estimated the likelihood of a lane change crash using the trajectory data.

Risk estimation is also done by using the vehicle motion characteristics to associate the driver behaviour or style with certain level of risk. These studies generally use classification methods to identify driver on the scale of safe to dangerous, compliant drivers-violating drivers on freeway or intersections (Aoude et al., 2011; Cheung et al., 2018; Xue et al., 2019). A recent study by Chen, Liu, et al. (2018) has tried to estimate the crash risk for driving intentions at the intersections.

The disaggregate studies generally use micro-simulations or naturalistic driving data for data collection to conduct risk assessment. The SSM are used to identify near crash or crash like events from this data (Johnsson et al., 2018). The detailed discussion on SSM follows in next section 2.3. The naturalistic driving data offers higher validity since the in-situ driver behaviour is captured.

Research Gaps: Crash Risk

From the above discussion on crash studies, it is clear that crash likelihood is the main focus during the estimation and prediction of the crash risk. The other equally important component i.e., severity has not received equal attention. Marchesini and Weijermars (2010) also suggested that the impact of traffic flow interactions on the severity needs further investigation.The other observed trend is that the aggregated and disaggregated studies using the crash data attempt to establish the effects of traffic on the crash occurrences or risk. The studies on crash risk due to a driving intention under a particular traffic conditions are not prevalent.

2.3. Surrogate measures for Rear-end conflicts

SSM are widely used to estimate the conflicts and help to undertake the safety assessment without actual crashes (Johnsson et al., 2018). The indicators are the measure of the temporal and spatial proximity and in some cases the severity or kinetic energy of the conflict. SSM are based on the principle that during/ before the occurrence of a conflict event, one vehicle must take an evasive action to avoid a collision (Gettman and Head, 2003).

SSM are helpful in capturing the likelihood and severity of an accident in an objective way. Since there is no actual collision, certain thresholds of proximity in time and space or evasive action such as braking/ deceleration are defined to indicate the critical events. The hypothesis made in the use of SSM is that the vehicle conflicts will possibly result into crashes if the thresholds are exceeded. The use of SSM done in both online and offline road safety studies. The SSM and validate the crash risk using trajectory data from micro-simulation tools or naturalistic driving. The studies in table 2.2 using trajectory data rely on SSM for estimation of the crash risk. Different types of conflicts such as rear end, lane changing, right angle conflicts, road run off, etc can be captured by SSM (Azevedo et al., 2018; Ozbay et al., 2008).

As the name suggests, Proximity based measures are function of the separation between the vehicles on a conflicting trajectory in terms of distance or time. The smaller is the distance or time separation between them, Higher is the risk of collision. Proximity based measures have been widely used for risk estimation for rear conflicts on lane driving as well as intersection conflicts. Distance Headway (DH) is a simple indicator used to measure the spatial proximity. It is defined as the distance gap between the two vehicles in car following scenario as shown in fig. 2.2. Similarly, Time Headway (TH) is the measure of temporal proximity between two vehicles. It is defined as the time it takes for the following vehicle to reach the position of preceding vehicle at a specific point of time. It can be calculated using equation 2.1. In the context of rear-end crash, it can capture the rear-end crash if the preceding vehicle suddenly applied hard braking and comes to halt immediately. The drawback of Time headway is that it is independent of the velocity of the preceding vehicle.

$$TTC = \frac{D}{(v_f)} \tag{2.1}$$

Time to Collision (TTC) is one of the commonly used surrogate safety measure for rear-end conflicts. TTC is defined as the time which remains to the occurring of a collision between two vehicles, if the collision course and relative velocity between the two vehicles is maintained (Hyden, 1996). The formula for calculation of TTC is given in 2.2. It can be seen that TTC depends on the relative velocities of the approaching vehicles and the distance between them. TTC is meaningful when the following vehicle is faster than preceding vehicle. When the preceding vehicle is faster than the approaching vehicle, TTC is a negative value and not of practical meaning. Therefore, a low non-negative value of TTC indicates higher risk of collision. The TTC-Relative speed plot of two vehicles at different spacings is shown in figure 2.3. From a practical perspective, TTC is an indicator of the available time to collision avoidance manoeuvre such as braking or lane change (Mahmud et al., 2017) and thus can be used to indicate the urgency of such manoeuvre (Lee, Olsen, et al., 2004). Due to this characteristic, TTC is widely used for analysis of rear-end type collision scenarios by analyzing the TTC values during the real or simulated motion of a vehicle. The TTC values are compared with a minimum or desired TTC which is defined on the basis of perception time, reaction time of a driver and the driving conditions, to identify the critical driving events (Mahmud et al., 2017). There is a wide spectrum of the desired values of TTC used by researchers varying from 0.9 to 4 seconds depending on the road configuration such as intersections or highways, frequency of conflict, among others (Meng and Qu, 2012; Sayed et al., 1994). In case small relative velocities between vehicles, the TTC can approach very high values or infinity. Due to this reason, Inverse TTC is used for many practical purposes. Inverse TTC, as the name suggests is defined as the reciprocal of the TTC.

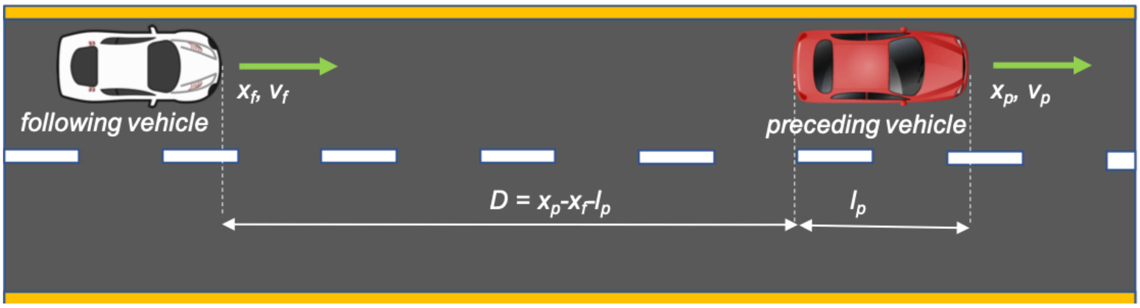

Figure 2.2.: Car following scenario

$$TTC = \frac{D}{(v_f - v_p)},\ if\ v_f > v_p \tag{2.2}$$

Where,
D = $x_p - x_f - l_p$
v_f is the speed of the following vehicle
v_p is the speed of the preceding vehicle as shown in fig. 2.2

TTC is defined at a particular point of time and by this very definition, it ignores the past behaviour. To overcome this, Several other extensions have been proposed by researchers. Time Exposed TTC and Time Integrated TTC take into account the the past driving behaviour by using the accumulated values of TTC during the safety critical time i.e., when TTC is below the threshold over a period of time (Minderhoud and Bovy, 2001).

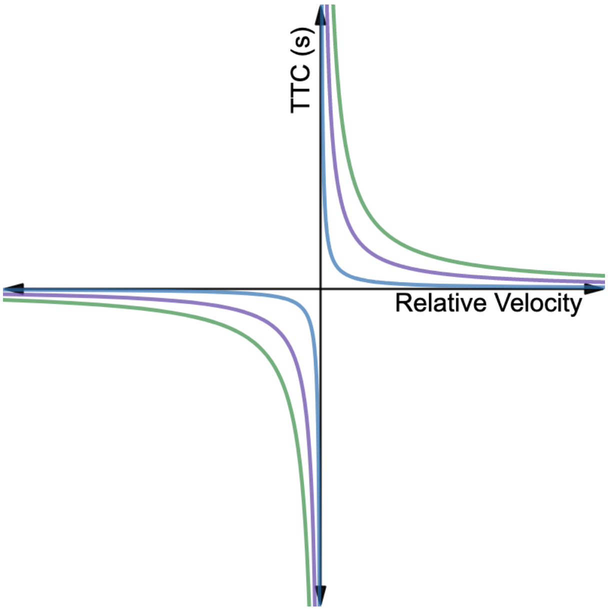

Figure 2.3.: TTC - Relative Velocity $(v_f - v_p)$ Plot (The curves indicate constant spacing between vehicles)

TTC has a drawback when the preceding vehicle is faster than the following vehicles, as in that case TTC is a negative value. Ozbay et al. (2008) proposed MTTC which addresses this drawback. It considers accelerations of the preceding and following vehicle along with their relative velocities (equation 2.3, 2.4) as shown in fig. 2.4. The consideration of vehicle acceleration alongside its velocity helps MTTC to account for the conflict cases where preceding vehicle is faster than the following vehicle. Thus, MTTC can cover more collision scenarios. Similar to TTC, Small values of MTTC suggests high risk of potential conflict. Ozbay et al., 2008 used MTTC threshold of 4 seconds to represent potential conflicts and Yang, 2012 further used MTTC in an exponential decay function to estimate the potential conflict risk. In their study, Yang, 2012 also validated the use of MTTC lane merging/ lane following situation. As in the case of TTC, MTTC indicates the collision proximity at a point of time and it does not gives an idea about the past. But it is also possible to devise integrated MTTC which takes into account the accumulated MTTC using methods similar to Time Exposed TTC or Time Integrated TTC, as discussed above. This aspect has been addressed in this study in the Methodology section.

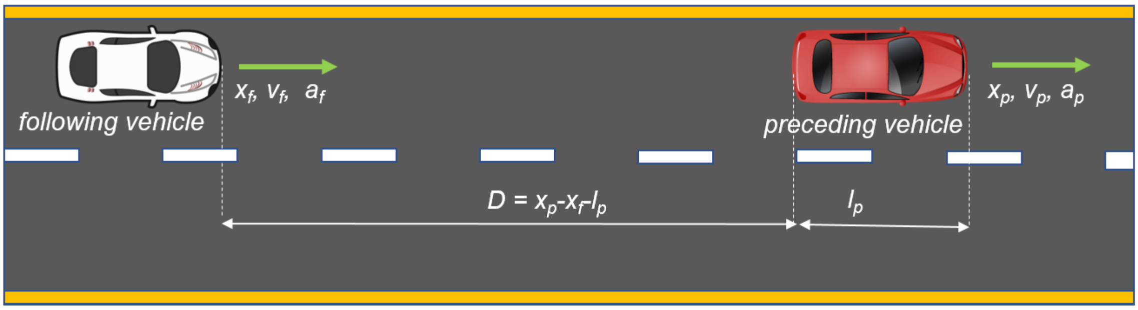

Figure 2.4.: Car following scenario

$$t_1 = \frac{-\Delta v - \sqrt{\Delta v^2 + 2\Delta a D}}{\Delta a} t_2 = \frac{-\Delta v + \sqrt{\Delta v^2 + 2\Delta a D}}{\Delta a}, if\ \Delta a \neq 0 \tag{2.3}$$

$$MTTC = \begin{cases} min(t_1, t_2) & \text{if } t_1 > 0, t_2 > 0, \\ max(t_1, t_2) & \text{if } t_1 * t_2 < 0, \\ D/\Delta v & \text{if } \Delta a = 0 \end{cases} \tag{2.4}$$

Where, $\Delta v = v_f - v_p$
$\Delta a = a_f - a_p$
D = $x_p - x_f - l_p$

From the above discussion on time based surrogate measures, it is found that MTTC is appropriate for analyzing potential rear-end conflict situations as it can take into account wide range of conflict possibilities in case of varying relative velocities and relative accelerations of the following vehicles. But one drawback of TTC or MTTC is that it only gives an indication of the likelihood of an accident and not the impact or severity of the accident because it is possible for a pair of vehicles to have an equal MTTC for different combinations of relative velocities and relative accelerations. But is it obvious that crashes at high speed will have different consequences from the crash at low speed.

There are some non-temporal surrogate safety measures which which can be used to estimate the accident danger. Deceleration Required to Avoid Collision (DRAC) is a commonly used SSM which is used to estimate the danger of a rear end crash by using the formulation in equation 2.5 (Mahmud et al., 2017). This formulation can be seen also in a different way as made of the product of relative velocities of preceding-following vehicle and inverse of the TTC as shown in equation 2.6. Therefore, although it considers differential speeds, but does not give an impact measurement since different combinations of differential speed is possible for same value of DRAC, but the impact at high speeds might be different than those at low speeds as stated above.

$$DRAC = \frac{(v_{following} - v_{preceding})^2}{X_{following} - X_{preceding} - l_{preceding}} \tag{2.5}$$

$$DRAC = \frac{(v_{following} - v_{preceding})}{TTC} \tag{2.6}$$

where, TTC $= D/\Delta v$
D $= X_{following} - Xpreceding - l_{preceding}$
$\Delta v = v_{following} - v_{preceding}$

There are few SSMs which take into account the severity and likelihood of a crash in single formulation. Crash Index (CI) uses MTTC to indicate the combined severity and likelihood of the possible crash. Ozbay et al., 2008 used the equation 2.7 to indicate the severity as well as the likelihood of the the conflict by using adapted formulations from kinetic energy and MTTC.

$$CI = \frac{(v_p + a_p * MTTC)^2 - (v_f + a_f * MTTC)^2}{2 * MTTC} \tag{2.7}$$

Another measure called Criticality Index Function (CIF) introduced by Chan (2006) also takes into the conflict severity into consideration by using follwing two hypothesis:

1. Crash consequences are more severe at high speeds
2. longer the available time to a collision, the crash can be avoided

Accordingly, the formulation given by Chan (2006) is given in equation 2.8. The equation given by him has two components i.e., CIF is directly proportional to the square of the velocity and inversely proportional to the TTC. Thus the main difference between the CI and CIF is that the numerator in CIF depends only on the vehicle speed rather than the relative speed of following and preceding vehicle, which is the case in CI.

$$CIF = \frac{(v)^2}{TTC} \tag{2.8}$$

Although, CI and CIF give a total measure of the crash risk i.e., its likelihood and severity., but it is not possible to study the crash likelihood and crash severity separately using CI and CIF. Therefore, there is a need to introduce a SSM which estimates the severity for rear-end crash separately. A surrogate to quantify severity separately can help to better understand the relationship of crash severity with traffic conditions and driving intentions and improve the risk estimation/ prediction models.

2.4. Conclusions and Research Gaps

From the above literature review, the following research gaps are summarized:

1. The methods for lane changing prediction mainly use manually labelled dataset. The methods to automatically label maneuvers from a large naturalistic driving data are not found in the literature. Thus, there is a need to develop an automatically maneuver labeling method which is efficient for real-time prediction settings.
2. The effects of traffic dynamics and driving intention on crash risk are generally studied separately. Impact of driving intention along with traffic conditions on crash risk has not been fully explored.
3. A greater emphasis is towards crash likelihood in both aggregate and disaggregate stud-

ies. Crash severity is investigated in off-line studies but not enough literature is found for the on-line studies. Lack of SSMs for explicit investigation of crash severity also supports the need to develop severity surrogates and investigate the relation with traffic conditions and intentions.

Part II.

Method and Analysis

3. Methodology

This chapter presents the methodology for achieving the objectives of this study. Contextually, the methodology consists of three modules namely methods for driving maneuver identification and intention prediction, traffic state identification and crash risk estimation. An overview of the complete methodology is given in fig. 3.1.

The methodology is motivated by the data driven analysis and use of statistical and Machine Learning (ML) tools. These methods are used in above mentioned three modules of the study to different extent i.e., driving intention module consist of classification and prediction tasks whereas traffic state module consists of only classification task. This is why the core of each of these module is a machine learning pipeline. A ML pipeline is basically a process consisting of steps, but not limited to, such as data collection and processing, features extraction, modeling and evaluation (Liu et al., 2017). This ML pipeline is adapted to the objectives of each module i.e., estimation and/ or prediction. This chapter contains three sections. The first section describes the generic ML pipeline. The subsequent sections demonstrate the approach for application of ML pipeline for driving intention prediction, traffic state identification and crash risk estimation from the disaggregate data.

3.1. Machine Learning Pipeline

Oxford Dictionary (2019) defines Machine learning as *"a type of artificial intelligence in which computers use huge amounts of data to learn how to do tasks rather than being programmed to do them"*. This is done by tuning a model to learn the data distribution from the available data. The flowchart of the ML pipeline adopted for this study is shown in fig. 3.2. The black arrows represent the normal work-flow while the grey-dotted arrows represent the feedback wherein the project steps are revisited during the ML work-flow due to unsatisfactory results. This characteristic of the ML pipeline makes it an iterative process at different levels. The steps in ML pipeline are discussed in below:

1. **Defining project goal**: This is the first step for the ML project/ task. This step is dedicated to defining the goals of the machine learning project and accordingly allocating resources for the project. The project goals may need to be revisited if they are found to be unrealizable during the project.

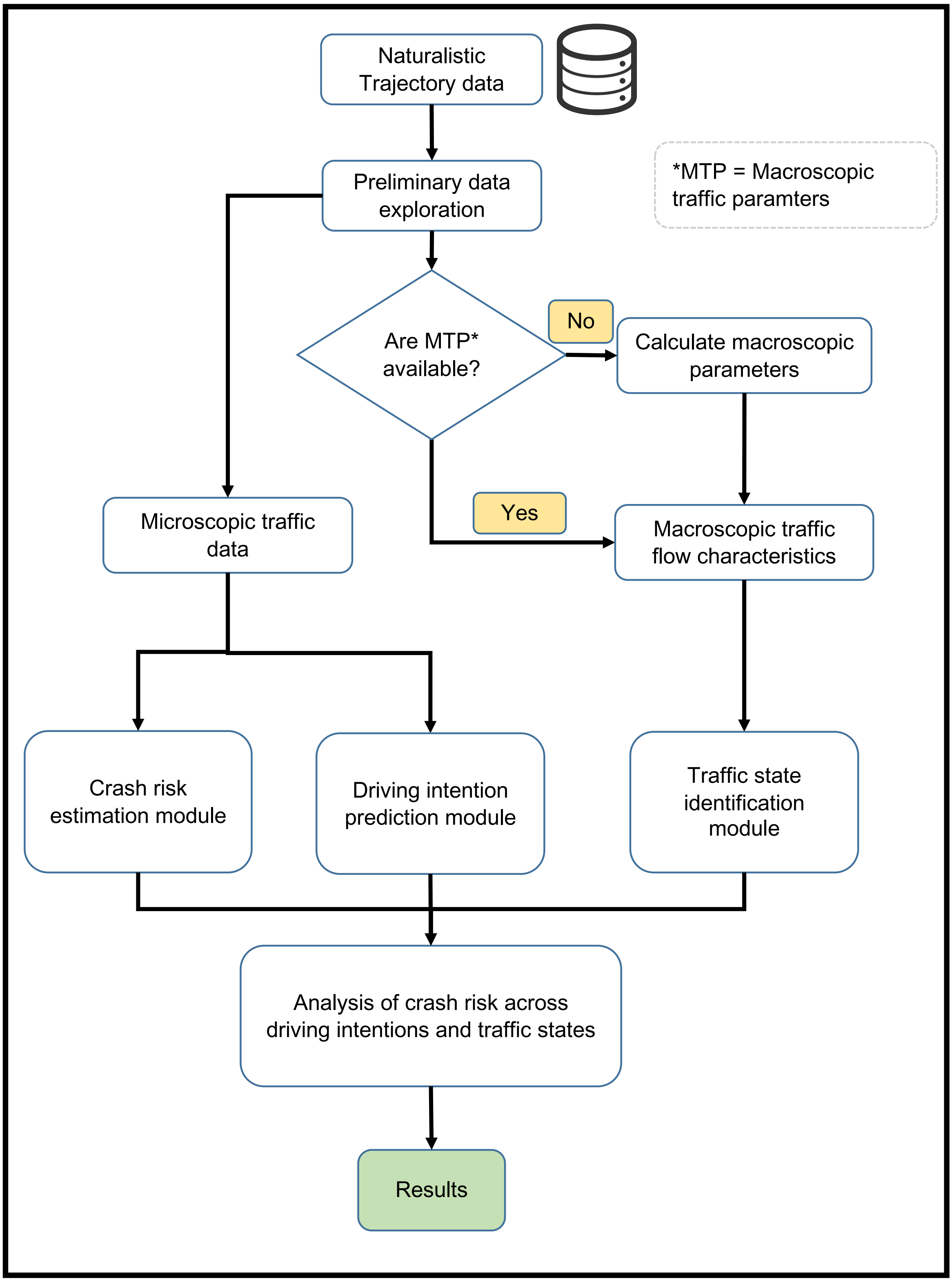
Naturalistic Trajectory data
Preliminary data exploration
*MTP = Macroscopic traffic paramters
Are MTP* available?
No
Calculate macroscopic parameters
Yes
Microscopic traffic data
Macroscopic traffic flow characteristics
Crash risk estimation module
Driving intention prediction module
Traffic state identification module
Analysis of crash risk across driving intentions and traffic states
Results

Figure 3.1.: Overview of the methodology

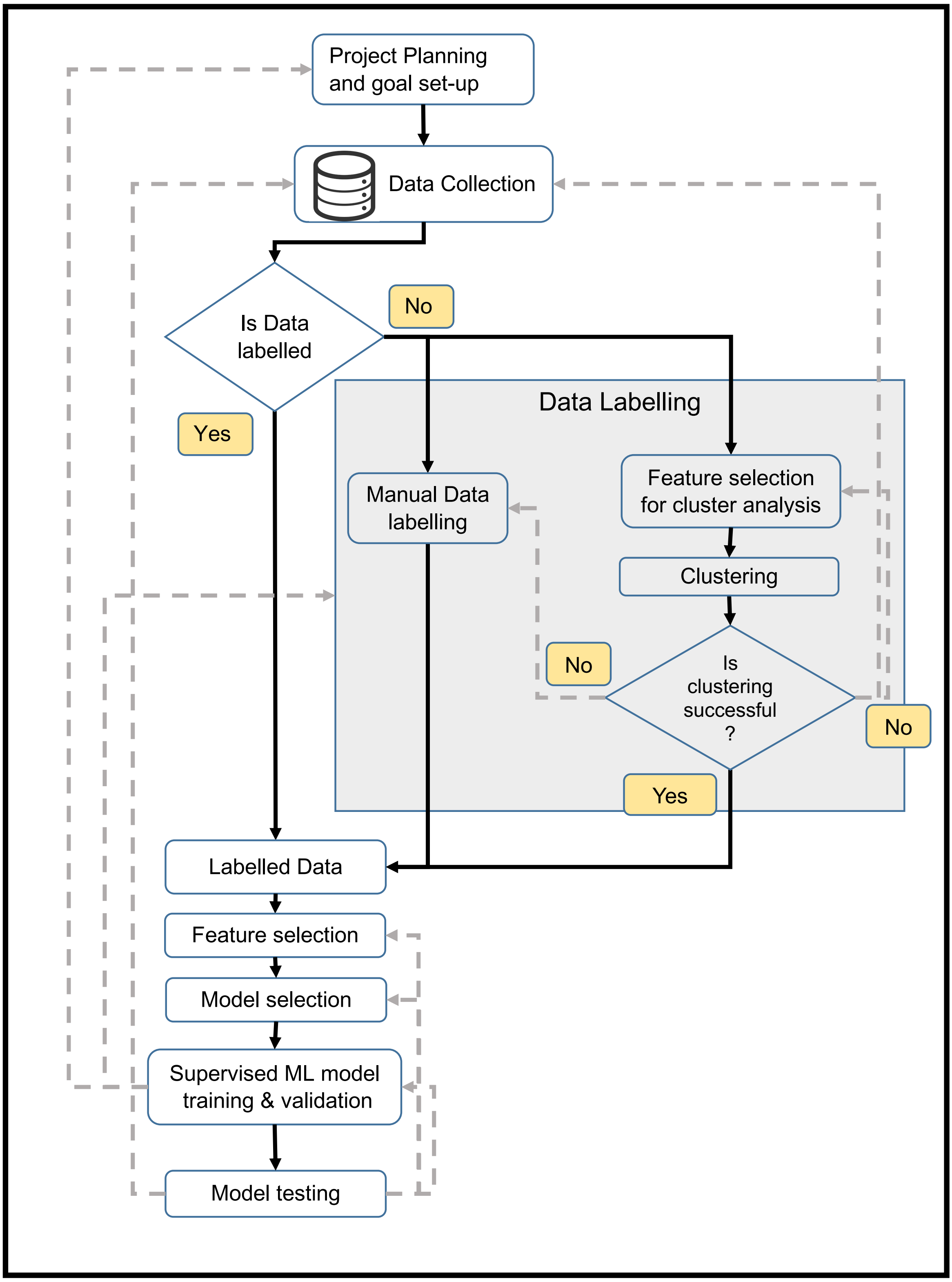

Figure 3.2.: General Machine learning pipeline

2. **Data Collection**: In this step, researchers look for the data required for the ML project. Open data is available free of cost whereas proprietary data needs to be purchased. In case the suitable data for ML project is not available, then the project stakeholders (researchers or practitioners) have to collect data either through in-house resources or by outsourcing the data collection process. During data collection, it is to be kept in mind that the data coverage should account for all the scenarios expected during inference/ testing of model. This is because theoretically the ML model cannot make accurate predictions for the data which was not seen during modeling and thus will give unreliable predictions.

3. **Data Labelling**: This is done to annotate the data with respect to the ground truth. Generally the data labelling is kept in mind during data collection and both are done simultaneously. Data labeling is a critical step in a ML project because the accuracy of the data labelling plays a crucial role in determining the accuracy of the machine learning model. Generally, there is a trade off between the quality of the data labelling and the costs involved since good labeling will need high quality control and more resources. In case the collected data is unlabelled, Manual labelling is a common yet expensive practice.

   Alternatively, unsupervised ML methods such as clustering are applied to identify the different groups and subsequently label the data. This is equivalent to a classification ML project in itself, that is why clustering is preceded and succeeded by feature extraction and model evaluation respectively. Few clustering algorithms are briefly discussed below:

   a) K-means: This algorithm clusters data into specified number of groups of equal variance by minimizing ″within-cluster sum-of-squares criterion″. This is a first choice because of its simple algorithm but it performs poorly on clusters with unequal variance or uneven shapes. Also the number of clusters are to be manually specified and thus it is highly sensitive to interpretation.

   b) Density-Based Spatial Clustering of Applications with Noise (DBSCAN): This algorithm can separate areas of high density points from low density points. The DBSCAN is effective in discovering clusters of arbitrary shape with least domain knowledge (Daszykowski and Walczak, 2009; Schubert et al., 2017). The tunable parameters for this type of clustering are maximum distance for neighborhood point (eps), minimum number of points to be considered in neighborhood to define a core point (minimum samples) and metric for calculating distances.

   c) Gaussian Mixtures (GM): This algorithm assumed that the data comes from a mixture of finite number of normal distributions. Like k-means algorithm, the number of clusters are to be specified manually, but GM is an improvement over k-means because ″it can generalize over varying shapes because it incorporates information about the covariance structure of the data and centers of the latent Gaussians″ (Scikit-learn, 2019a). The main parameters for GM models are number of clusters and covariance type. The covariance-type parameter specifies whether the Gaussian components have a shared or a unique covariance matrix.

d) Agglomerative clustering: This is a type of bottom-up hierarchical clustering wherein clustering starts from bottom and successively merges the clusters (Scikit-learn, 2019b). The main parameters is the linkage criteria which specifies which metric is used for merging strategy. Agglomerative clustering can capture clusters of uneven shapes but it is a computationally expensive algorithm.

Different algorithms and parameters settings are tried until good clustering is achieved. In case all the above methods for data labeling fail to achieve good clustering, then data may need to be either manually labelled or collected again with labels or more features.

4. **Data Processing**: This is an important step to ensure the quality of the data for modelling. The crucial steps in data processing are mentioned below:

   a) EDA: To obtain various insights about the data through visualization and calculating summary statistics such as mean, median, variance and quantiles.

   b) Data cleaning: This is used to discard the outliers, incomplete, erroneous, irrelevant, etc. data.

   c) Feature Selection: The important features are selected based on their significance in model performance and the correlated features are dropped. In cases where labels are estimated from the data using clustering, feature selection is done simultaneously with clustering.

   d) Feature Compression: If the data is high dimensional, feature compression or dimensionality reduction technique such as Principal Component Analysis (PCA) is used. This is necessitated due to the need for visualizing the clusters and selecting only the significant features for clustering.

   e) Feature Scaling: Certain ML models including deep learning models need all the features to be scaled before modeling. This process called as feature scaling is used to feature-wise transform the data to a common range of [-1, 1] or [0, 1] for effective performance. This is generally done by using Min-Max scaling or Standardization.

5. **Modeling**: During modeling, Machine learning model/ algorithm is selected and it is trained on the labelled data. Deep learning is a relatively new branch of machine learning which consists of computation models with multiple layers, which are capable of learning representations at different hierarchical levels (LeCun et al., 2015). Few supervised classification algorithm from classical machine learning and deep learning domain are discussed below:

   a) SVM: It is used to separate a multidimensional data into different classes by generating a boundary called ”hyper plane” (Vapnik, 1998). The parameters for SVM are the kernel function and regularization term C.

b) Random forest (RF): This a decision tree based classifier. RF fits a number of decision trees to the data and averages the results for the final prediction. This is why, RF is effective against over-fitting. Gini impurity is used to measure the quality of the split in RF. (Breiman, 2001). The main parameters for random forests classifier are number of estimators, maximum depth of the tree and measurement criteria for quality of the split.

c) Recurrent Neural Networks: RNNs are a special kind of neural network (Rumelhart et al., 1986) consist of chain like structure consisting of multiple neural networks. This chain like structure can be quite used for modelling temporal dependencies in a sequence. RNN can model temporal dependencies well but in practice they face difficulties in modelling long dependencies (Bengio et al., 1994; Hochreiter, 1991).

d) Long Short Term Memory: LSTMs networks are special kind of Recurrent Neural Networks which are capable of learning long term dependencies (Hochreiter and Schmidhuber, 1997). LSTMs also have a chain like structure similar to RNN but with modifications in the individual unit. LSTMs consist of a memory cell shown in figure 3.3 and control the flow of information by using input, forget and output gate layers which discard the non-essential information and memorize only essential information for the purpose. These operations in the LSTM cell can be represented by the following set of equations (Olah, 2019):

$$forget\ gate, f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3.1}$$

$$input\ gate, i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.2}$$

$$\bar{C_t} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3.3}$$

$$cell\ update, C_t = f_t * C_{t-1} + i_t * \bar{C_t} \tag{3.4}$$

$$output\ gate, \bar{o_t} = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3.5}$$

$$\bar{h_t} = o_t * tanh(C_t) \tag{3.6}$$

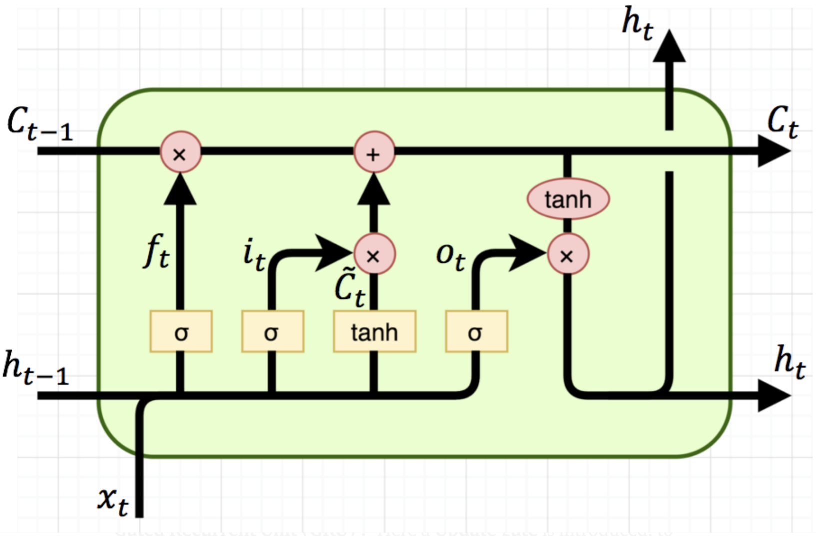

Figure 3.3.: LSTM Cell. Source: Olah (2019)

where W_f, W_i, W_o, W_c and b_f, b_i, b_o, b_c represent the weights and biases during the LSTM operations.

These cells are stacked in a layer like RNN to learn the sequential dependencies. Due to this reason, LSTMs have been successful in many tasks such as language translation (Sutskever et al., 2014), handwriting recognition (Graves and Schmidhuber, 2008) and image captioning (Kiros et al., 2014). The main tunable parameters of RNN/ LSTM are mentioned below:

i. Number of LSTM layers and dense layers stacked in the model. A dense layer is a fully connected layer in which all the inputs are connected with all the outputs.

ii. Number of Memory cells in each LSTM layer

iii. Learning rate: This determines the step-size while changing the weights of the model parameters based on the value of cost function. Large learning rate might not lead to discovery of local minima, whereas small learning rate might real to a very slow convergence.

iv. Optimizer: The optimization function used to train the data. For example Stochastic Gradient Descent (SGD), ADAM (Kingma and Ba, 2014), etc .

Advantages and Disadvantages of deep learning models Deep learning models offer advantages such as reduced focus on hand designing the features which takes some skill and time (LeCun et al., 2015). Deep learning models learns the underlying data distributions faster using the concept of distributed representations due to increased computing (LeCun et al., 2015). The deep learning models like RNN and

LSTM are certainly more complicated than the classical/ common ML models like SVM and RF, which makes it harder to train and tune these models. These models are also computationally expensive and take more time during modeling since they require large data to train.

The model selection is followed by model training to find the best model parameters which can represent the data distribution. The domain knowledge and state-of-the-art can help to narrow down the possible models for application. This can also help to avoid starting from the scratch and help to evaluate the results with respect to the established baselines in literature.

Before training, the dataset is divided into 2 parts namely training data and validation data. The training data is used to fit the model to the underlying distribution. The validation data comes in play to avoid over-fitting i.e., model should fit the training data distribution instead of the training data. Model training and validation is done using a measure of model performance called as cost or error function. A cost function is a mathematical formulation of difference between the true value and predicted value of the model. The aim of the model training is to achieve the optimum (lowest) value for this cost function by changing the model parameters. The choice of cost function is based on the nature of the problem and priorities. For example. for a regression problem, loss function can be Root Mean Squared error whereas for a classification problem, the cost/ error function can be Logarithmic loss or Categorical Cross Entropy (equation 3.7). The selection of cost function can also be done with the help of literature to identify the most suitable.

If model training and validation fails to achieve good results, model selection, feature selection, data labelling or even data collected may need to be re-investigated.

$$CE = -\sum_{i}^{C} C_i * \log(s_i) \tag{3.7}$$

where C_i and s_i are the ground truth and model prediction for each target class i in C.

6. **Evaluation**: During clustering, the ground truth or true labels are not available, therefore the evaluation of the clusters is performed using one of the internal indices which measure the goodness of the clustering structure without external information (Tseng and Wong, 2005; Wang, Wang, et al., 2009. One such internal index is the silhouette coefficient, with which each cluster is described by its silhouette based on the comparison of its separation and tightness (Rousseeuw, 1987). The silhouette coefficient is calculated using the mean intra-clustering distance and mean nearest-cluster distance for each sample. Silhouette score has a range of -1 to 1, with scores closer to 1 indicating good clustering performance.

   The classification algorithms such SVM/RF/LSTM are evaluated using the precision, recall and accuracy scores as defined in equations 3.8, 3.9 and 3.10.

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

Where,TP: True Positive,TN: True Negative,FP: False Positive, FN: False Negative

The trained is model is evaluated on the new data or test data. If the model fails on the test data, there might be problems related to model over fitting or the data used might not be fully representative of the phenomenon to be learned.

3.2. Driving Intention Estimation and Prediction

The method flowchart is shown in 3.4. The dicussions on the steps in the flowchart follows in subsequent subsections.

3.2.1. Goal Definition

The goal is to develop a ML model for estimating and predicting the driving intention from disaggregate traffic data using automatic labelling. The lateral driving maneuvers namely lane keeping and lane changing are selected for the study. Thus, the ML task is to classify disaggregate traffic data into these two binary classes of maneuvers.

3.2.2. Data Collection

The disaggregate/ trajectory data is collected generally using GPS, ground based video cameras, drone video recording, etc. The raw data from the vehicle sensors contains multiple time series for different features, also referred as a multi-variate time series for each of the vehicle. The features in the trajectory data generally characterizes vehicle motion through its position, velocity and acceleration. In the disaggregate data, the state (S_t^n) of a vehicle n at time t can be represented by equation 3.11. When these states are collected over a period of equal time intervals, it becomes a time series data as shown in equation 3.12. The n simply denotes the

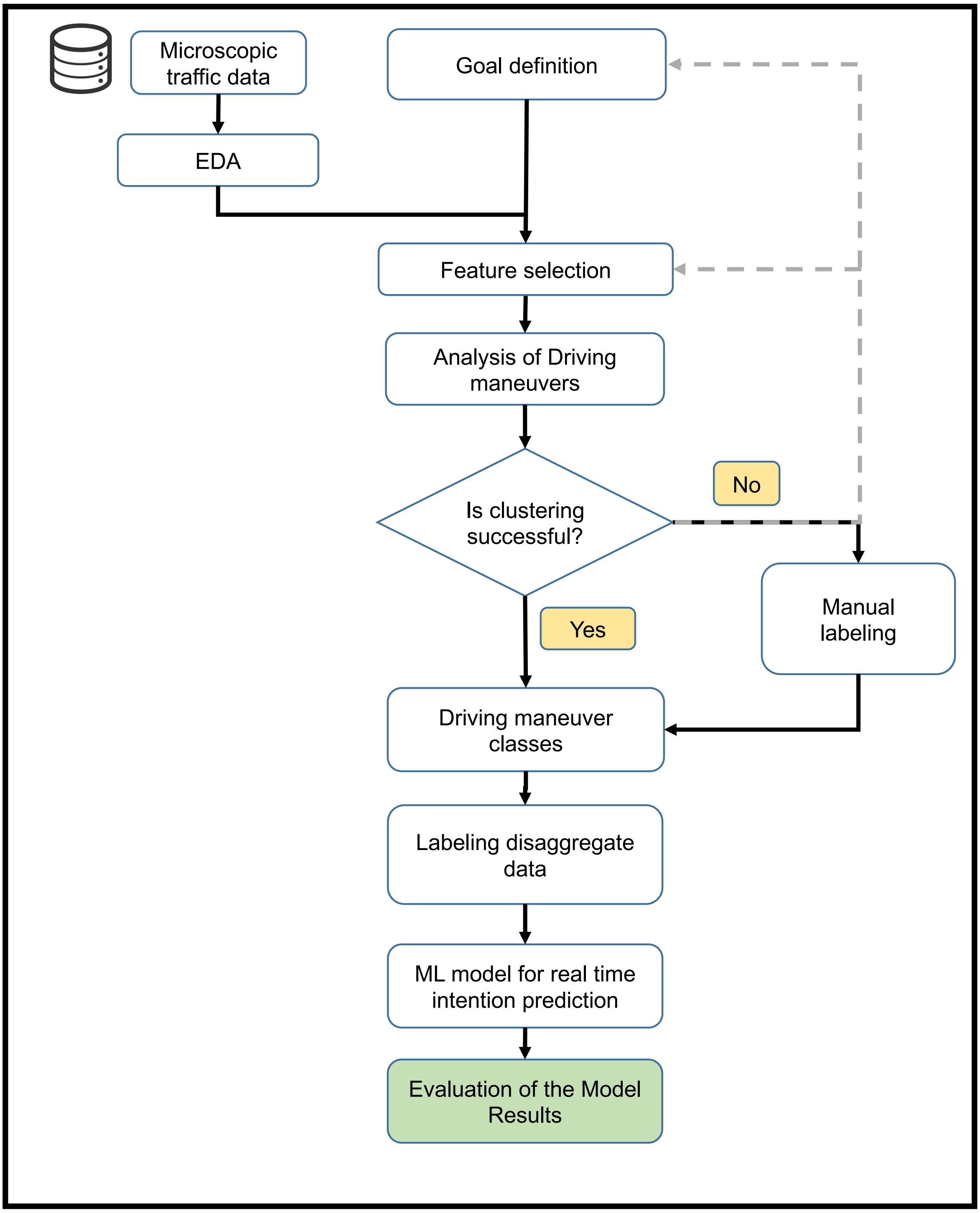
Microscopic traffic data
Goal definition
EDA
Feature selection
Analysis of Driving maneuvers
Is clustering successful?
No
Yes
Manual labeling
Driving maneuver classes
Labeling disaggregate data
ML model for real time intention prediction
Evaluation of the Model Results

Figure 3.4.: Driving intention prediction module

index of the vehicle and does not represent any mathematical power operation.

$$(S_t)^n = (t, f_t^1, f_t^2, \ldots f_t^{k-1}, f_t^k)^n \tag{3.11}$$

Where, t: time, f_t^k: feature/ variable of time series at time t, k: number of features in time series, n: index of the vehicle

$$T^n = (S_1, S_2, S_3, \ldots S_{t-2}, S_{t-1}, S_t)^n, \tag{3.12}$$

In addition to the vehicle motion features, the data may contain additional information such as vehicle dimensions, vehicle type (car/ truck), other details like surrounding vehicles, road layout, etc.

3.2.3. Data Labelling

The data labelling can be done on the raw disaggregate data as well as aggregated data. If the labeling is done on the disaggregate data, the the labelling can be shown by equation 3.13. This equation shows that the state at for each vehicle at time t can be mapped to a unique maneuver label (lane changing or lane keeping).

$$(S_t)^n \rightarrow (l_t)^n, \tag{3.13}$$

Where $(l_t)^n$ is the maneuver assigned to state S of vehicle n at time t

Data labelling can also be done on the spatially or temporally aggregated features as well. For example, if vehicle is labelled in terms of whether it changed lane or not during observation. In this case, temporally aggregated features from the trajectory of vehicles are used and can be represented as equation 3.14. This equation represents the mapping of the trajectory of each vehicle to a unique label. The trajectory data is not labeled in terms of driving maneuvers (lane changing. lane keeping).Therefore data labelling using clustering need to be explored.

$$T^n \rightarrow L^n, \tag{3.14}$$

Where: L^n is the label assigned to the Trajectory features of vehicle n.

3.2.4. Feature Selection

The high frequency of trajectory data collection and nature of vehicle dynamics/ traffic features can cause the features as well as the time series points to be correlated. EDA is used to estimate descriptive statistics and identify such correlation, if any.

Instant/ disaggregate speed and acceleration of a vehicle can be used as features. The aggregated features can also be calculated from data. A lane change maneuver can be observed through lateral displacement of a vehicle due to its lateral acceleration and velocity. This behaviour may also be visible in the aggregated features such as mean (equation 3.15) and standard deviation (equation 3.16) of the speed and accelerations.

$$\mu_k^n = \frac{1}{N^n} \sum_{t=t_b}^{t=t_e} (f_t^k)^n \tag{3.15}$$

$$\sigma_k^n = \sqrt{\frac{1}{N^n} \sum_{t=t_b^n}^{t=t_e^n} ((f_t^k)^n - \mu_f^n)^2} \tag{3.16}$$

Where: N^n is the length of bi-variate time series for vehicle n,
t_b^n is the time when the nth vehicle enters the area/ section under observation,
t_e^n is the time when the nth vehicle exits.

If needed, a dimensionality reduction technique such as PCA is applied to compress the m raw features into k compressed features as shown by equation 3.17.

$$P_1, P_2 \ldots P_d = F(f_1, f_2, f_3, \ldots f_k) \tag{3.17}$$

Where: $F(x)$ is the dimensionality reduction algorithm,
d is the number of compressed features,
k is the number of raw features form the dis-aggregate or aggregate data,
$d < k$,

3.2.5. Unsupervised Classification

Clustering, a type of unsupervised ML is used here to identify patterns and groups in the data as shown in equation 3.18). The data in the identifiable groups is labelled according to its respective class. Clustering algorithms mentioned in 3.1 are used to discover the clusters by hit and trial approach. The clustering method which is able to discover the cluster groups with a good quality is selected as the candidate.

$$C = G(P_1, P_2 \ldots P_k) \tag{3.18}$$

Where: $G(x)$ is the classification algorithm ,
C belongs to the set of clusters corresponding to the target maneuver class.

3.2.6. Automated Data labelling

A supervised classifier such as SVM is trained on the identified clusters to learn the cluster boundaries so that the classifier can be used for online/ automated data labelling. After obtaining the labels from either of the methods discussed above, the disaggregate/ aggregate data is labelled as shown in equation 3.19.

$$[X, y]^n = [(x^{(t_1)}, x^{(t_2)} \dots x^{(t-1)}, x^{(t)}), (L^{(t_1)}, L^{(t_2)} \dots L^{(t-1)}, L^{(t)})]^n] \tag{3.19}$$

Where, X is the set of independent features,
y is the dependent/ target variable/ labels

3.2.7. Modeling

The next step is to formulate a time series prediction problem such that the temporal relationships among the features and labels can be learned. Since the maneuver label is a discrete variable, the prediction is formulated as a time series classification problem, in which the part or complete time series is used to assign a class. A moving window time series classification can be defined with the help of following parameters:

1. Frame frequency, F is the frequency at which the raw data has been collected.
2. Frame granularity, f is the sampling rate of the time series data for using as an input for the model. If f=2, this means the effective frame frequency is F/2 Hz.
3. Prediction horizon time, p_t is the time in future at which the prediction is being made.
4. Prediction horizon length, p is the time series steps corresponding to p_t and f. Thus, p = F*p_t/f
5. Look-back time, k_t is the past time (and corresponding data) used to make the prediction.
6. Look-back length, k is the time series steps corresponding to k_t and f. Thus, k = F*k_t/f
7. Discard Threshold, T is the length below which data is not considered for training. Further T $>$ p+k or T = p+k+b where b is the buffer length.

Based on the above parameters, the time-series data (equation 3.20) is transformed into rolling window data such that for time t = αf where α is an integer. The resulting training data is formulated as equation 3.21. In other words, the past time series data of constant length is used to predict the class label at a future time step. The data is split into training and validation data set so that vehicles in the training data are not included in the validation data.

$$[X, y]^n = [(x^{(t_b)}, x^{(t_b+f)} \dots x^{(t_b+N-2f)}, x^{(t_b+N-f)}), (L^{(t_b)}, L^{(t_b+f)} \dots L^{(t_b+N-2f)}, L^{(t_b+N-f)})]^n] \tag{3.20}$$

Where: $N = F * (t_e - t_b + 1)$

$$[X, y]^n = [(x^{(t_b-k*f)}, x^{(t-(k-1)*f)} \dots x^{(t_b-2f)}, x^{(t-f)}, x^t)^n, (M^{(t+p*f)})^n] \tag{3.21}$$

Model selection and training

The RNN/LSTM are suitable for the sequence prediction problem as seen during the previous discussion in 3.1. Further LSTM is more capable than RNN in learning long-term dependencies, therefore LSTM is selected as the primary prediction model. Adam optimizer is used for adapting the learning rate of LSTM. The training is stopped when validation accuracy does not improve over 5 consecutive iterations/ epochs to avoid over-fitting problem. The maneuver class with higher probability score is the predicted class.Categorical cross entropy (equation 3.7) is used as the cost function during LSTM training. Random Forest is selected as the baseline model for comparison with LSTM model.

3.2.8. Evaluation

The machine learning pipeline for intention prediction consists of multiple steps such as clustering, labeling and prediction. Thus, the performance of the complete model is dependent on the performance of these individual stages. Clustering, label classification and time-series prediction are evaluated independently based on the common performance metrics for these algorithms as discussed above in section 3.1. The performance of the complete model is also evaluated based on the correct classification and detection of the maneuvers in the trajectory.

In addition to aforementioned classification metrics, Advance detection time is used to assess the predictive range of the model, as it shows the how much in advance a prediction is made before a vehicle actually crosses the lane line (i.e. Point C in Figure 2.1). It should be noted here that advance detection time is different from the aforementioned prediction horizon time in subsection 3.2.7 . The advance detection time measures the performance of labeling and prediction model, which in turn is governed by both the SVM and LSTM models respectively.

3.2.9. Model testing

Model testing is a two-stage process. First the learned labelling classifier is used for learning the target maneuver classes and subsequently used for labelling the raw measurements in real time. The labeled data is used to predict the intention using the predictor model.

3.3. Traffic State Identification

The method flowchart is shown in fig. 3.5. The discussion on the steps in the flowchart follows in subsequent subsections.

3.3.1. Goal Definition

The goal is to develop a ML model for identifying and labeling the traffic states into free-flow and congestion from aggregate data. A free-flow traffic state is a state where driver experiences no constrains from other drivers on the road and travel at free speeds which is determined according to the road design, speed limits, etc. At congestion state, queuing of vehicle starts, and vehicles can no longer travel at free speeds resulting in link speed drop. The traffic state labeling is to be achieved by using clustering algorithms to identify the groups corresponding to different traffic states.

3.3.2. Data Collection

The procedure for data collection is similar as discussed in subsection 3.2.2. The trajectory data is used to calculate the aggregate traffic flow parameters such as traffic flow rate, density and average velocity for individual lane as well as complete link/ section. The traffic flow rate is calculated by placing a virtual detector in the middle of the section being observed. In case of longer sections, the virtual detectors can be placed on more than one places to capture the variation along the length of the section.

The traffic related features are estimated by temporal and spatial aggregation of the disaggregate data. The lane traffic flow is equal to the number of vehicles of a given lane passing a particular cross-section of a road in a time interval as shown in equation 3.22. The total traffic flow is simply the summation of the traffic flow of the individual lanes. Using this equation, lane flow and total flow can be calculated.

$$q = \frac{n}{T} \tag{3.22}$$

Where: n is the number of vehicles passing the lane/ road section, T is the duration

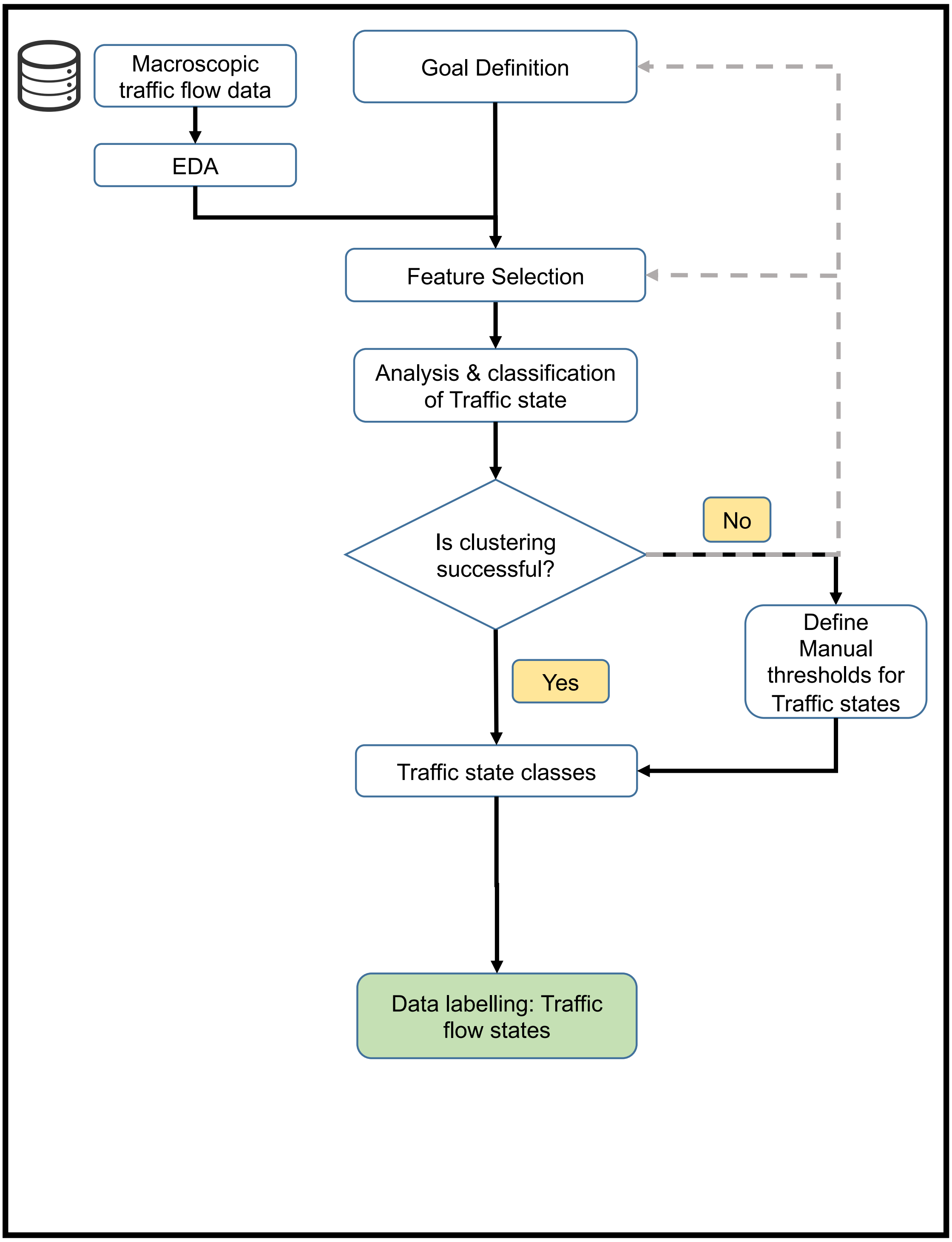
Macroscopic traffic flow data
EDA
Goal Definition
Feature Selection
Analysis & classification of Traffic state
Is clustering successful?
No
Yes
Define Manual thresholds for Traffic states
Traffic state classes
Data labelling: Traffic flow states

Figure 3.5.: Traffic state identification module

of the time interval

The lane traffic density traffic density is calculated by simply counting the number of vehicles on a particular lane at a given time instant divided by the length of the road. This is shown in equation 3.23. The total traffic density is simply the summation of the traffic density of individual lanes. The average lane density is the average of the traffic density of individual lanes. The average density over a time interval is calculated by taking an average of the traffic densities during that time, as shown in equation 3.24.

$$k = \frac{n}{L} \tag{3.23}$$

Where: n is the number of vehicles passing the detector, L is the distance of the section

$$K = \frac{\sum_{t=t_b}^{t_e} k_t}{(t_e - t_b + 1)} \tag{3.24}$$

The lane/ link speed is calculated by taking the average of speeds of vehicles on that particular lane/ link at a given time instant. This is shown in equation 3.25. The average link speed over a time interval is calculated by taking an average of the link speeds during that interval as shown in equation 3.26.

$$v_{average}^{t} = \frac{\sum_{t=1}^{N} v_i}{(N)} \tag{3.25}$$

Where: n is the number of vehicles on a road lane/ link

$$v_{average}^{\Delta t} = \frac{\sum^{\Delta t} v_{average}^{t}}{(\Delta t)} \tag{3.26}$$

3.3.3. Data Labelling

The aggregated traffic flow features calculated from the data are to be labelled. The labelled data is represented as equation 3.27.

$$(v^{\Delta t}, k^{\Delta t}, q^{\Delta t}) \rightarrow L^{\Delta t}, \tag{3.27}$$

Where: $Q^{\Delta t}$ are the aggregated traffic parameters during the time interval Δt.

3.3.4. Feature Selection

The parameters in traffic Fundamental Diagram (FD) namely, density and average road link speed are used as features for clustering because the characteristics of these parameters are distinct in free and congestion flow.

3.3.5. Unsupervised Classification

The method is same as mentioned in subsection 3.2.5. The traffic data is labelled in terms of the traffic states- free-flow and congestion by using clustering algorithms.

3.4. Crash Risk Estimation

The flowchart is for risk estimation is shown in fig. 3.6. The discussion on the steps in the flowchart follows in subsequent subsections.

3.4.1. Goal Definition

The goal is to develop a ML model for identifying the levels of rear-end crash risk such as high risk and low risk from the data. The risk quantification is done in terms of likelihood and severity of a crash.

3.4.2. Data Collection

The procedure for data collection is similar as discussed in subsection 3.2.2.

3.4.3. Feature Estimation

The data does not contain the features for aggregate crash risk and therefore, it is calculated from the trajectory data. The aggregate crash risk is composed of two components i.e, Crash likelihood and Crash severity. Two Surrogate safety measures are used to quantitatively estimate each of these components.

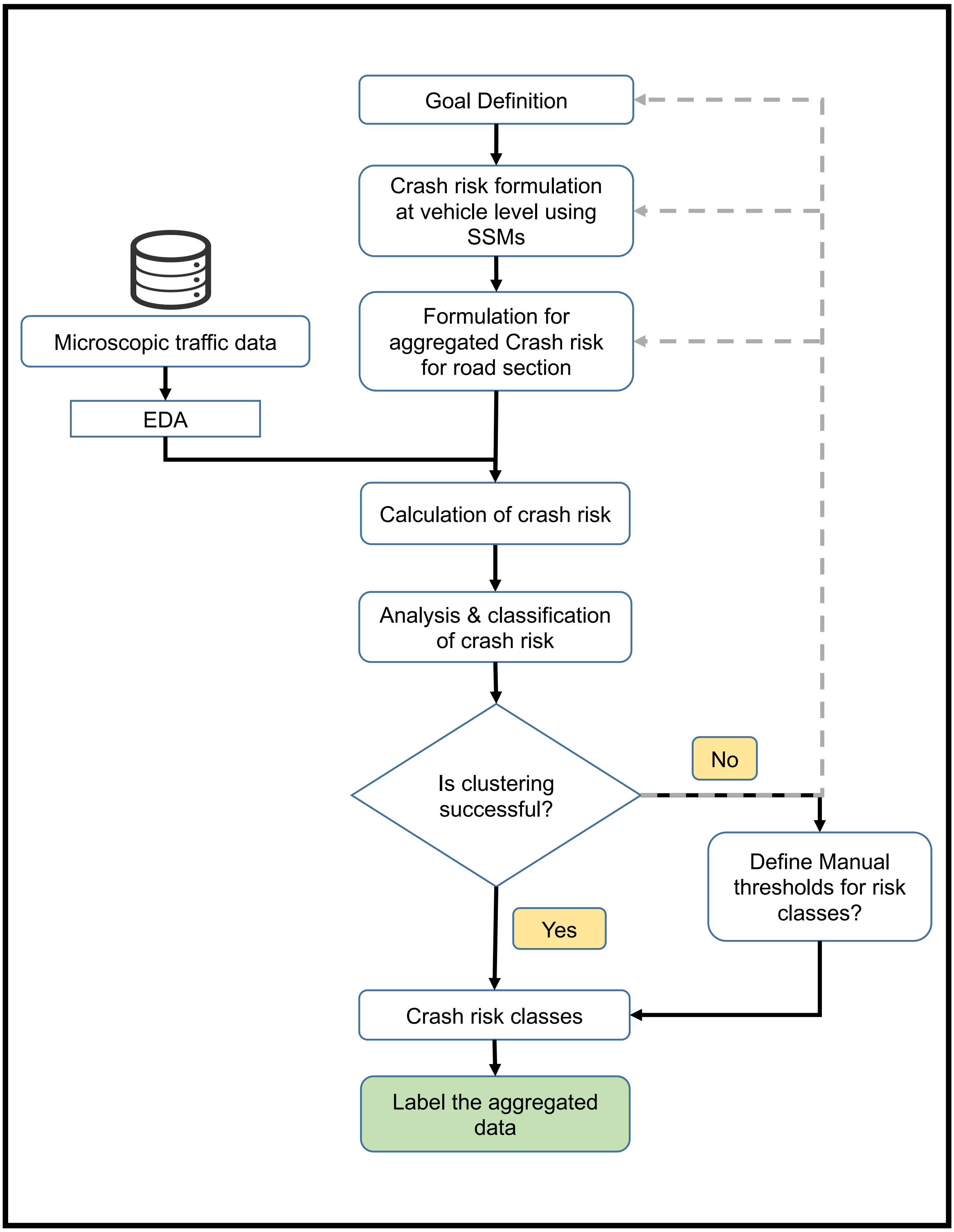
Goal Definition
Crash risk formulation at vehicle level using SSMs
Formulation for aggregated Crash risk for road section
Microscopic traffic data
EDA
Calculation of crash risk
Analysis & classification of crash risk
Is clustering successful?
No
Define Manual thresholds for risk classes?
Yes
Crash risk classes
Label the aggregated data

Figure 3.6.: Crash risk estimation module

Crash Likelihood

Surrogate safety measure - MTTC is estimated between the subject vehicle and preceding vehicle at each time step. The crash likelihood is defined as a negative exponential function of the the modified time to collision as shown in equation 3.28. A similar function to calculate the likelihood of a crash has been used by Weng and Meng (2014) and Yang (2012). According to this equation, high positive values of MTTC will lead to low chances of a crash and vice-versa.

$$p_{crash} = \begin{cases} e^{-MTTC/\lambda} & \text{if preceding vehicle is present and MTTC} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.28}$$

where, λ is a parameter and its value is 3.5 seconds for this study based on Yang (2012)

The crash likelihood is calculated at each time step from the disaggregate trajectory data. The crash likelihood of all the vehicles observed during the time interval is summed to obtain an aggregate indicator of the crash likelihood for a highway section as shown in equation 3.29. The obtained P_{crash} is divided by the flow of vehicles on the road section to obtain the Average Crash Likelihood (ACL) per vehicle on a road section for a given duration as shown in equation 3.30. This formulation is inspired by Ozbay et al. (2008) and Kuang et al. (2017) for estimating the crash index density and average risk respectively.

$$P_{crash} = \sum_{n}^{N} \sum_{t=t_b}^{t=t_e} (p_{crash} \cdot \delta_n^t) \tag{3.29}$$

$$ACL = \frac{P_{crash}}{\sum_n^N \sum_{t=t_b}^{t=t_e} (\delta_n^t)} \tag{3.30}$$

where, $\delta_n^t = 1$ if vehicle n is within the observed section during the time interval (t_b, t_e),
else $\delta_n^t = 0$

Crash Severity

A safety surrogate to calculate the severity/ impact of a crash (independent of the likelihood) is not found in the literature as already discussed in section 2.3. A severity surrogate called CRash IMpact (CRIM) is defined as shown in equation 3.31. The impact of rear-end crash is a product of speed of following vehicle and the speed differential between following and preceding vehicle. This is based on the similar hypothesis used by Chan (2006) that high speed of the following vehicle will lead to a severe crash. The use of differential term is incorporated in context of a rear-end crash. In case speed of the following vehicle is much greater than preceding vehicle, the crash will have more impact as compared to the case when the difference of velocities is smaller. This concept is somewhat similar to the differential term in the DRAC formulation which was discussed in section 2.3. In an extreme case scenario, a high speed vehicle rear crashing into a stopped vehicle will cause maximum damage. If the following vehicle is slower than the preceding vehicle, then the value of CRIM is negative which implies no impact. It is to be noted that the impact of a crash depends on the product of velocity terms, which is proportional of the kinetic energy per unit mass.

$$CRIM = v_{following} \cdot (v_{following} - v_{preceding}) \tag{3.31}$$

A positive exponential function is used to calculate the crash severity for a front collision (equation 3.32). Here the assumption is made that the severity of the crash increase exponentially with the increase in CRIM. The $CRIM_{max}$ is the maximum possible value of CRIM as shown in equation 3.33 and is used to normalize the CRIM. The $CRIM_{max}$ corresponds to the impact when a vehicle is travelling at the maximum speed (free speed) and the front vehicle is stationary.

$$S_{crash} = \begin{cases} e^{CRIM/CRIM_{max}} & \text{if preceding vehicle is present,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.32}$$

$$CRIM_{max} = v_{max} \cdot v_{max} \tag{3.33}$$

where, v_{max} is a parameter for a usual maximum speed on the freeway. Its value is selected as 108 km/h for this study.

The crash severity of all the vehicles observed during the time interval is summed to obtain an aggregate indicator of the crash severity for a highway section as shown in equation 3.34. The obtained S_{link} is divided by the flow of vehicles on the road section to obtain the Average Crash Impact (ACI) per vehicle on a road section for a given duration as shown in equation 3.35.

$$S_{link} = \sum_{n}^{N} \sum_{t=t_b}^{t=t_e} (S_{crash} \cdot \delta_n^t) \tag{3.34}$$

$$ACI = \frac{S_{link}}{\sum_{n}^{N} \sum_{t=t_b}^{t=t_e} (\delta_n^t)} \tag{3.35}$$

Total Risk

The total risk can be calculated by the product of average likelihood and average severity as shown in equation 3.36.

$$Total\ Risk = ACL \cdot ACI \tag{3.36}$$

3.4.4. Classification and Labelling

Clustering is attempted on the aggregate crash likelihood and severity indicators (ACL and ACI) to group the data into risk classes. In case, the clustering is not feasible, classes are labeled by applying manual thresholds/ boundaries based on the data distribution. This study only considers binary classification of the risk into high and low for both the likelihood ad severity.

The four scenarios depicting the likelihood and severity of a rear-end crash on a highway are shown in fig. 3.7. When the preceding and following vehicles are travelling at almost same low speeds and/or are farther from each other, the crash has low likelihood and low severity (1). If the vehicles have small value of TTC but they are traveling at low speeds, the crash likelihood high but the severity is still low (2). Whereas if the speed of following vehicles is high along with large speed differential, the severity of the conflict is also high (3). As the following vehicles closes in on the preceding vehicle with a large speed differential, both the likelihood and severity of a crash is high (4).

3.5. Statistical tests for Analysis

The data is labelled in terms of aggregate crash risk, traffic states and driving maneuvers. The analysis is conducted by calculating risk difference and odds ratios for classes of crash likelihood and crash severity. The effect of the driving intention (lane changing) and traffic state (congestion) on the change of risk from low class to high class is compared and tested for significance of proportions. The method for testing proportions is based on the concepts of case-control study Kirkwood and Sterne (2003). It is important to define few relevant terms before discussing the proportion tests. The definitions are given below:

1. Outcome variable: The subject of the study. In this study, outcome variable is the crash likelihood and crash severity with binary possibilities for each of them i.e., high or low.
2. Exposure Variable: The factor whose effect is to be studied i.e., driving intention and traffic condition. The exposure variable in terms of driving intention is lane changing i.e., lane changing vehicle vs lane keeping vehicle (which did not changes lanes). The exposure for traffic condition is congestion i.e., congestion vs no congestion (free-flow).

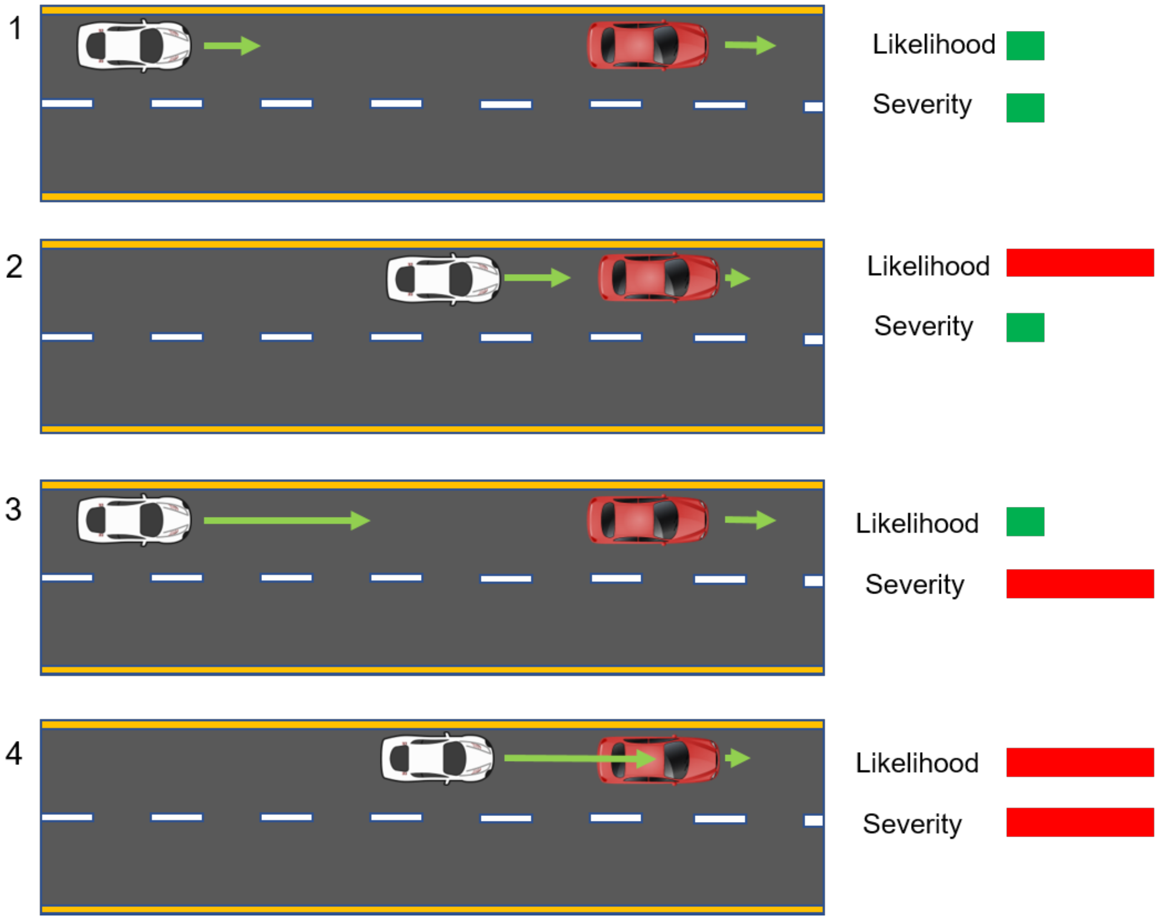

Figure 3.7.: Different scenarios of crash likelihood and crash severity: 1. Low likelihood and low severity, 2. High likelihood and low severity, 3. Low likelihood and high severity and 4. High likelihood and high severity

Table 3.1.: Contingency Table: source :Kirkwood and Sterne, 2003

	Exposure = congestion or lane changing		
Outcome = high likelihood or severity	Exposed	Not Exposed	Total
Experience Event	d_1	d_0	d
Did not Experienced Event	h_1	h_0	h
Total	n_1	n_0	n

3. Sampling Condition: The condition to satisfy while sampling the data for proportion tests. For example, select only the risk outcomes during Free-flow conditions, select only risk outcomes for lane changing vehicles

4. Contingency table: This a table wherein the samples are tabulated depending on the binary outcomes and exposure as shown in table 3.1.

5. Risk Difference: It is defined as the difference of high crash likelihood or severity proportions between exposed and unexposed samples as shown in equation 3.37. It gives an indication of the outcome due to the exposure.

$$Risk\ difference = p_1 - p_0 \tag{3.37}$$

   where $p_1 = d_1/n_1$ and $p_0 = d_0/n_0$

6. Odds Ratio (OR): It is the ratio of the odds of the outcome event (high likelihood/ severity) in the exposed group to the odds of the outcome event in the unexposed group, as shown in equation 3.38

$$Odds\ ratio, OR = \frac{d_1 h_0}{d_0 h_1} \tag{3.38}$$

The Confidence Interval (CI) are calculated for Risk difference and Odds ratio. The null hypothesis for testing is formulated as: "*there is no difference between the proportions in the population from which the sample was drawn*" (Kirkwood and Sterne, 2003). The formulae for CI and hypothesis testing (z) for risk difference are based on Kirkwood and Sterne (2003) and are presented in 3.39, 3.40 and 3.41.

$$standard\ error, s.e. = \sqrt{(\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0})} \tag{3.39}$$

$$CI = (p_1 - p_0) \pm z' \cdot s.e. \tag{3.40}$$

where z′= 1.96 for 95% CI

$$z = \frac{p_1 - p_0}{\sqrt{(p(1-p)(1/n_1 + 1/n_0)}} \tag{3.41}$$

where $p = d/n$

The formulae for CI and hypothesis testing for OR are based on Kirkwood and Sterne (2003) and are presented in 3.42, 3.43 and 3.44.

$$standard\ error, s.e.(log(OR)) = \sqrt{1/d_1 - 1/n_1 + 1/d_0 - 1/n_0} \tag{3.42}$$

$$CI(OR) = exp(log(OR) \pm z' \cdot s.e.) \tag{3.43}$$

where z′= 1.96 for 95% CI

$$z = \frac{log(OR)}{s.e.(log(OR))} \tag{3.44}$$

3.6. Conclusion

The methodology presented here provides the general methods and formulation to identify/ predict the driving intention, traffic state and crash risk from the disaggregate data, for further statistical analysis. The methods are motivated by the data-driven and machine learning approaches which makes it an iterative process, till the objectives are achieved.

4. Data Collection

This study uses the HighD trajectory dataset (Krajewski et al., 2018). The dataset is freely available for non-commercial purposes. The dataset can be obtained by sending a request form to the data owners (Dataset, 2019). The Chair of Transportation Systems Engineering, Technical University Munich provided the data to the author for this study.

4.1. Meta-details of the data

HighD Dataset is a naturalistic vehicle trajectory dataset collected on German highways using drone videography. The dataset consists of 60 recordings over 16.5 hours from six locations (labelled 1 to 6) at a frame frequency of 25 Hz. The dataset covers 4 lane (2 per direction) and 6 lane (3 per direction) highways with central dividing median and hard shoulders on outer edge. HighD dataset is recorded on highways with either no speed limit or with limit of 120 Km/h and 130 Km/h. Recording of the data took place on weekdays between 08:00 in the morning and 19:00 in the evening. In the dataset, each recording has a meta file which has the information on the location, time, duration and other characteristics of the recording. The attributes in the meta file are listed out in Table 4.1. The meta file gives an overview of the recordings and some relevant statistics of the highD dataset are shown in Table 4.2, which are extracted from the meta-data of each recording with features. The total distance driven by 110,000 vehicles (81% Cars and 19% trucks) in the dataset is 45,000 Km with 5600 complete lane changes (Krajewski et al., 2018).

4.2. Data attributes

Krajewski et al. (2018) applied tracking algorithms and post processing to retrieve smooth position, velocities and accelerations in both x (longitudinal/ driving direction) and y (lateral/ perpendicular to the driving direction) axes. The list of attributes in the dataset is shown in Table 4.3. At vehicle level, the data has trajectory and physical features. The trajectory of the vehicle is characterized by its position, velocity and acceleration in longitudinal and lateral direction. The physical features in the data are the width and the length of the vehicle.

The data has attributes which tell us about the surroundings of the subject vehicle.

Table 4.1.: Attributes in the Meta file (Source: Krajewski et al., 2018)

Attributes	Description	Unit
id	The id of the recording. Every recording has a unique id.	
frameRate	The frame rate which was used to record the video.	[hz]
locationId	The id of the recording location.	
speedLimit	The speed limit of the driving lanes	[m/s]
month	The month the recording was done.	
weekDay	The week day the recording was done.	
startTime	The start time at which the recording was done.	[hh:mm]
duration	The duration of the recording.	
totalDrivenDistance	The total driven distance of all tracked vehicles.	
totalDrivenTime	The total driven time of all tracked vehicles.	[s]
numVehicles	The number of vehicles tracked including cars and trucks.	
numCars	The number of cars tracked.	
numTrucks	The number of trucks tracked.	
upperLaneMarkings	The y positions of the upper lane markings.	[m]
lowerLaneMarkings	The y positions of the lower lane markings.	[m]

Table 4.2.: Summary of the highD dataset

Location id	No. of recordings	Speed Limit (Km/h)	Duration (min.)	Total vehicles	Cars	Trucks	Lanes
1	37	120	664	85962	69751	16211	3+3
2	3	No limit	49	3074	2400	674	2+2
3	3	130	57	3747	2710	1037	3+3
4	4	No limit	56	4751	3799	952	3+3
5	10	No limit	143	10079	8192	1887	2+2
6	3	120	30	2903	2287	616	4+3

Table 4.3.: Attributes in the trajectory data (Source: Krajewski et al., 2018)

Name	Description	Unit
frame	current frame	[-]
id	vehicle's id	[-]
x	x position of the upper left corner of the vehicle's bounding box.	[m]
y	y position of the upper left corner of the vehicle's bounding box.	[m]
width	width of the bounding box of the vehicle (length of the vehicle)	[m]
height	height of the bounding box of the vehicle (width of the vehicle)	[m]
xVelocity	longitudinal velocity in the image coordinate system.	[m/s]
yVelocity	lateral velocity in the image coordinate system.	[m/s]
xAcceleration	longitudinal acceleration in the image coordinate system.	[m/s]
yAcceleration	lateral acceleration in the image coordinate system	[m/s]
frontSightDistance	distance to the end of the recorded highway section in driving direction from the vehicle's center.	[m]
backSightDistance	distance to the end of the recorded highway section in the opposite driving direction from the vehicle's center.	[m]
dhw	distance Headway (= 0, if no preceding vehicle exists)	[m]
thw	time Headway (= 0, if no preceding vehicle exists)	[s]
ttc	Time-to-Collision (= 0, if no preceding vehicle or valid TTC exists)	[s]
precedingXVelocity	longitudinal velocity of the preceding in the image coordinate system	[-]
precedingId	id of the preceding vehicle in the same lane	[-]
followingId	id of the following vehicle in the same lane	[-]
leftPrecedingId	id of the preceding vehicle on the left adjacent lane in driving direction	[-]
leftAlongsideId	id of the adjacent vehicle on the left adjacent lane in driving direction	[-]
leftFollowingId	id of the following vehicle on the left adjacent lane in driving direction	[-]
rightPrecedingId	id of the preceding vehicle on the right adjacent lane in driving direction	[-]
rightAlsongsideId	id of the adjacent vehicle on the right adjacent lane in driving direction	[-]
rightFollowingId	id of the following vehicle on the right adjacent lane in driving direction	[-]
laneId	id start at 1 and are assigned in ascending order	[-]

There are 8 attributes for the id of the vehicles surrounding the subject vehicle namely, front-preceding, rear-following, left (preceding, along, following) and right (preceding, along and following). In case, there is no vehicle present for a particular position, it is assigned a value of 0 in the data. The lane position of the vehicle can also be ascertained from the attribute 'laneId' from the dataset. Some additional attributes have been already calculated in the data namely distance headway, time headway and time-to-collision with respect to the front preceding vehicle. There is a aerial photograph of the road section for each recording in the data and one such photograph is shown in figure 4.1.

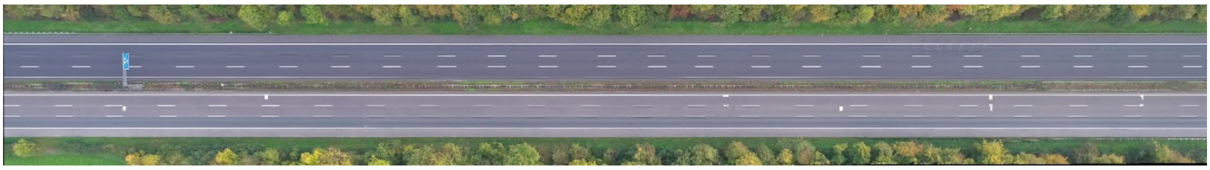

Figure 4.1.: Aerial snapshot the road section for recording no. 25 in the highD dataset

Table 4.4.: Comparison of HighD and NGSIM based on Krajewski et al. (2018)

Feature	NGSIM dataset	HighD dataset
Total duration	1.5 hours	16.5 hours
Number of locations	2	6
Average duration of recording	45 minutes	15 - 20 minutes
Lanes per direction	5-6	2-3
Total vehicles	9206	110,000
Truck percentage	3%	23%
Mean speeds	< 75 km/h	Cars(120 km/h), trucks (80 km/h)
Lane changes per vehicle	0.45	0.10
Average traffic density	higher	lower
Recording frame frequency (Hz)	10	25
Length of road section	0.50 - 1.00 km	0.42 Km

4.3. Comparison with similar dataset

A comparison of the HighD dataset with the Next Generation SIMulation (NGSIM) dataset (FHWA, 2019) has been already done and the same is reflected in table 4.4. The HighD dataset is larger as compared to NGSIM in terms of number of observed vehicles and duration of the recordings. The length of the observed road section and average recording duration is shorter in HighD when compared to NGSIM. The average number of lane changes are less in HighD possibly due to lesser number of lanes per direction and lesser congestion. The average speeds in HighD are much greater than NGSIM. HighD dataset has low errors in trajectories due to use high resolution cameras and application of advanced post processing algorithms.In the area of driving intention and traffic risk, there are multiple examples of NGSIM use-cases (Hu et al., 2018; Kuang et al., 2017; Woo et al., 2017), but similar studies on HighD Dataset are not found.

4.4. Conclusion

The HighD dataset has large number of recordings of vehicles over different time and days of the week. This captures driving behaviour at different times of traffic and their interaction with other vehicles, which is significant for the intention and risk analysis. The HighD data is a relatively new dataset and as per the knowledge of author, this dataset has not been used for the investigation of driving intention and rear-end crash risk estimation. Therefore, this dataset offers a good opportunity for this study.

5. Data Analysis

This chapter provides details on the tools, programming language and libraries used to perform the analysis. Subsequently, it shows the results of the preliminary analysis/ EDA on the data to gain some insights about the data.

5.1. Software and Tools

Python programming language has been used to perform most of the analysis in this study. Python has an extensive collection of libraries for data handling, computation and visualization. The machine learning frameworks are mature in python and its use is commonly among the scientific community. The list of the prominent python libraries and tools used in this study is given below:

1. Computation: Numpy
2. Data handling and manipulation: Pandas
3. Parallel processing: Pathos, Multiprocessing
4. Visualization and plotting: Matplotlib, Seaborn
5. Machine learning: Scikit-learn, Keras (Tensorflow backend) (Chollet et al., 2015)

The hardware used for the study is a 2018 MacBook Pro with i7 processor and 16 GB RAM with occasional use of a Desktop PC with same specifications.

5.2. Exploratory Data Analysis

EDA helps in gaining preliminary understanding of the data, discovering patterns and confirming assumptions. The detailed microscopic, macroscopic and criticality analysis of the highD dataset is done by Kruber et al. (2019), and is recommended to the reader for an elaborate understanding of the highD data. In this report, the first sub-section: Microscopic Traffic Analysis shows the results of EDA on disaggregate trajectory data. The second sub-section: Macroscopic Traffic Analysis presents the results for the traffic flow parameters from the data.

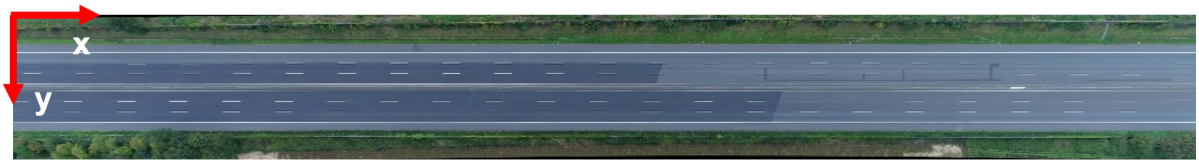

Figure 5.1.: Road section for track 4

5.2.1. Microscopic Traffic Analysis

The HighD dataset consists of multiple recordings corresponding to different locations and duration of the day (Table 4.2). A individual recording is referred to as a ”track” for further analysis. A track contains the data at a location and is continuous without any recording interruptions. The ”track 4” is selected as a sample to explore the microscopic traffic variables and understand the data. The road section corresponding to track 4 along with adopted coordinate system is shown in shown in figure 5.1. This track has 3 lanes in each direction with right direction (longitudinal) as positive x-axis and bottom direction (lateral) as positive y-axis. The positions of all the vehicles are positive due to this coordinate system. Since Germany follows a right side driving rule, the vehicles travelling in right direction (bottom 3 lanes) have positive longitudinal velocities whereas the vehicles travelling in left direction (top 3 lanes) have negative longitudinal velocities. When a vehicle makes a lateral maneuver such that it moves towards in positive y-axis, it has a positive y velocity and vice-versa.

The summary statistics of the vehicle features are shown in Table 5.1 and 5.2. Each of these tables correspond to the two driving directions i.e., positive x (bottom 3 lanes) and negative x (top 3 lanes), respectively. Since the data is recorded at a frame frequency of 25 Hz, one second of data for a vehicle would generate 25 data points for all the features. This tracks has data for 1163 vehicles and the count represents the number of the data points corresponding to all these vehicles. The x position varies between 0 to 415 m (length of the road section). The y-values range from 12 m to 35 m corresponding to top lane and bottom lane respectively. These x and y position in the data correspond to the top left corner of the vehicle bounding box. The longitudinal speeds range from 73 km/h to 177 km/h with a mean close to 110 km/h. The statistics for lateral speeds, longitudinal acceleration and lateral acceleration can also be seen in these tables. The length and width of vehicles also varies between (2.9 m, 21 m) and (1.7 m, 3.0 m) which shows different types of vehicle such as cars and trucks in the data.

The plot of the trajectory of a vehicle with Id. no. 200 on the track is shown in figure 5.2. The evolution of its longitudinal and lateral motion (position, velocity and acceleration) is shown in figure 5.3 and 5.4 respectively. It can be seen from these plots that the velocity and acceleration have first order and second order differential relationship respectively with the position data. The plots for other tracks are not shown here to avoid repetition.

Table 5.1.: Disaggregate features: Positive longitudinal Direction

	x	y	w (m)	L (m)	v_x (Km/h)	v_y (Km/h)	a_x (m/s^2)	a_y (m/s^2)
count	193375							
mean	206.62	31.84	2.17	8.48	110.68	-0.05	0.03	-0.02
std	117.72	2.64	0.30	5.27	20.13	0.75	0.18	0.09
min	1.58	26.33	1.72	2.93	73.26	-5.00	-1.36	-0.53
25%	104.94	30.54	1.92	4.65	90.11	-0.29	-0.06	-0.06
50%	206.32	31.13	2.02	5.05	112.10	-0.04	0.03	-0.01
75%	308.34	34.53	2.50	14.25	128.38	0.18	0.11	0.03
max	414.50	35.84	3.03	20.62	177.05	5.18	1.53	0.56

x, v_x, a_x: longitudinal position, velocity and acceleration; y, v_y, a_y: lateral position, ; velocity and acceleration, w: width of vehicle, L: length of vehicle; 25%, 50% and 75% represent the first, second and third quartiles respectively

Table 5.2.: Disaggregate features: Negative longitudinal Direction

	x	y	w (m)	L (m)	v_x (Km/h)	v_y (Km/h)	a_x (m/s^2)	a_y (m/s^2)
count	197312							
mean	196.25	16.57	2.18	8.68	-105.66	-0.03	0.05	-0.02
std	117.41	2.64	0.28	5.46	20.37	0.86	0.26	0.11
min	-20.88	12.44	1.72	3.39	-176.80	-7.49	-1.57	-1.02
25%	94.83	13.80	1.92	4.65	-122.58	-0.29	-0.07	-0.06
50%	195.45	17.32	2.02	5.05	-101.95	0.00	0.03	-0.01
75%	297.38	17.83	2.50	15.06	-87.30	0.25	0.14	0.03
max	407.33	22.04	2.93	21.02	-74.52	6.30	2.42	1.18

x, v_x, a_x: longitudinal position, velocity and acceleration; y, v_y, a_y: lateral position, velocity and acceleration, w: width of vehicle, L: length of vehicle

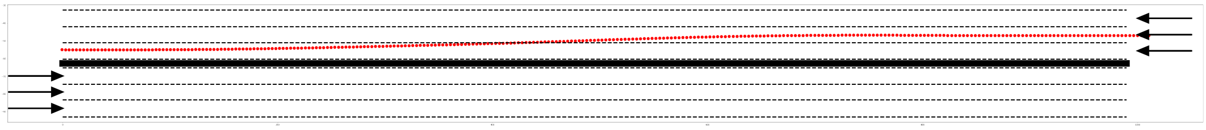

Figure 5.2.: Sample trajectory of a vehicle

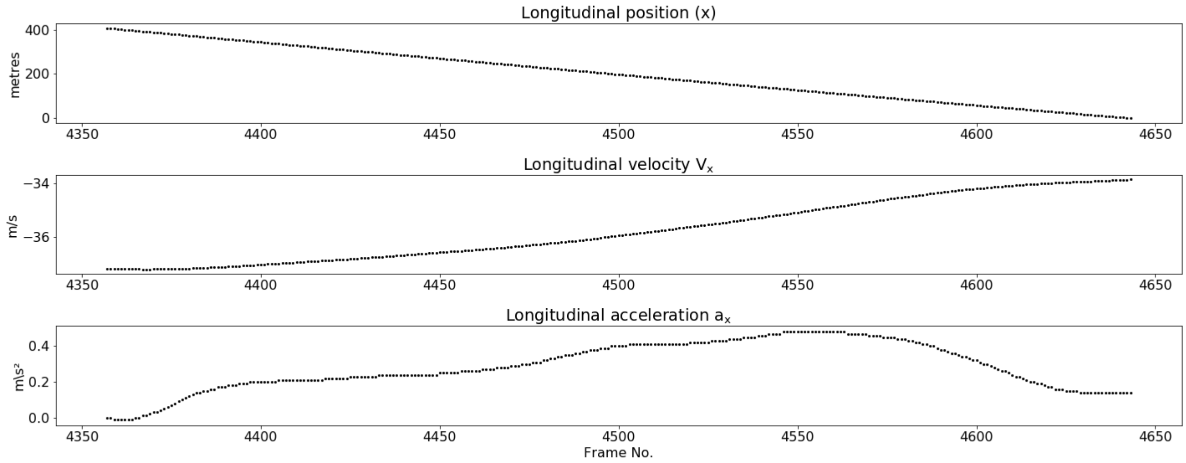

Figure 5.3.: Longitudinal motion variables of a sample vehicle

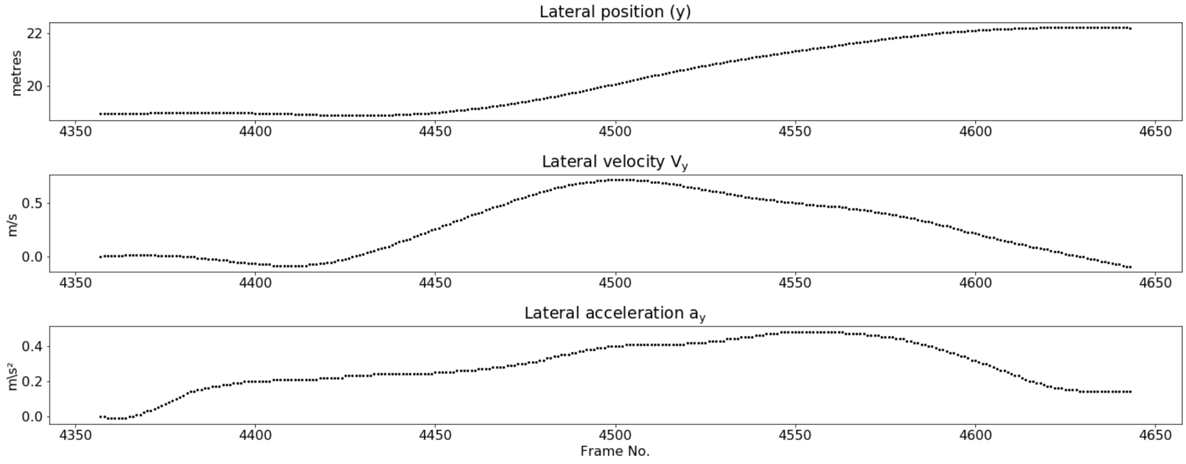

Figure 5.4.: Lateral motion variables of a sample vehicle

Safety Surrogates

In the data, safety surrogates namely, TTC, time headway (TH) and distance headway (DH) are already available in the data as mentioned in table 4.3. In addition to these, likelihood and severity indicators i.e., MTTC and CRIM are calculated for comparison. The pair plot of these indicators is shown in fig. 5.5. Only, the instances with MTTC between 0 and 5 seconds is used for this plot. The negative MTTC are not significant for safety analysis of rear-end crashes. The values higher than 5 seconds are less critical and thus dropped to make the plots clean and readable.

The TH and DH are linearly correlated with Pearson correlation coefficient close to 1. MTTC and TTC are also positively correlated but the correlated is not strong. It can be seen that for the MTTC values upto 5s, TTC varies upto 20s with most of the values in the range upto 10s. This reiterates the fact that MTTC covers more conflict scenarios. The CRIM has a high positive correlation with TH and DH. The possible reason is following vehicles have large speed differences when they are far from one another. The correlation of MTTC and CRIM is not very strong which points to the fact that they measure separate aspects of a conflict i.e., likelihood and severity. The distribution of MTTC is uni-modal and skewed to the left, indicating that instances of low MTTC (high likelihood) are fewer. On the other hand, the distribution of CRIM is skewed to the right with multiple peaks which indicate the instances of high severity are fewer. This makes sense because high risk instances are lesser compared to the normal or low risk instances. Crash severity has few negative values in case where preceding vehicle is faster than the following vehicle.

5.2.2. Macroscopic Traffic Analysis

The macroscopic traffic analysis is carried only for the 6 lane freeway sections with 3 lanes in each direction. This selection is done because data for 6 lane (3+3) road section constitute majority of the recordings (44 locations) as compared to other road configurations (2+2, 3+4) in the highD data. This also allows for uniformity in terms of road geometry for the comparison of macroscopic traffic parameters. The summary of the selected data is given in table 5.3. In terms of speed limit, location 4 has no speed limit, whereas data from location 1 and 3 has a speed limit of 120 km/h and 130 km/h respectively.

Table 5.3.: Details of the data used for Macroscopic traffic characteristics

Location id	No. of recordings	Speed Limit (Km/h)	Duration (min.)	Total vehicles	Cars	Trucks	Lanes
1	37	120	664	85962	69751	16211	3+3
3	3	130	57	3747	2710	1037	3+3
4	4	No limit	56	4751	3799	952	3+3

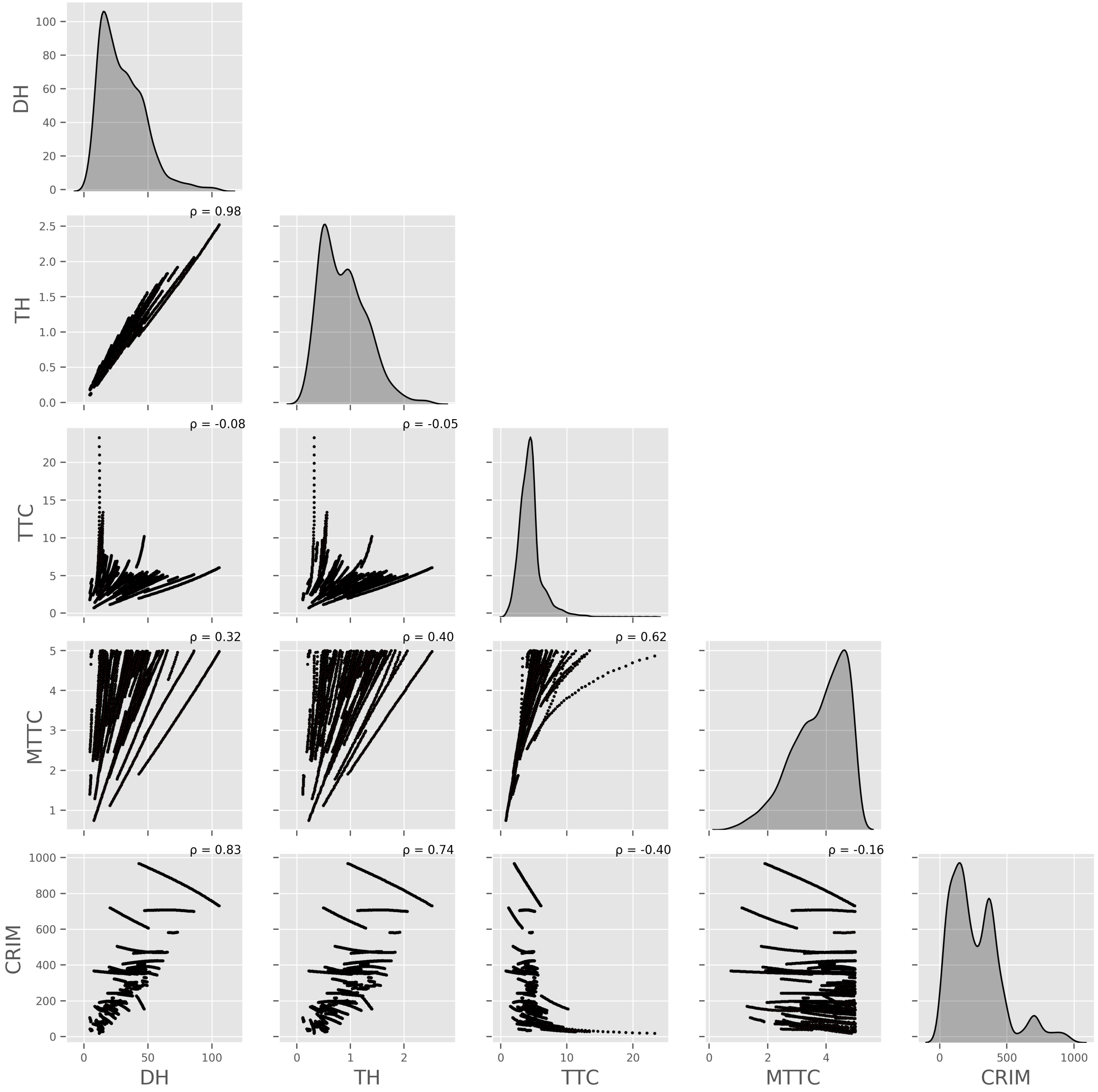

Figure 5.5.: Pair plot of the Safety Surrogates for disaggregate data

The traffic flow is calculated at the approximate middle of the road section at x = 200 m. The vehicles crossing are aggregated for 30 s interval for each lane to get the Traffic flow. The obtained traffic flow is scaled to hourly values by multiplying with a factor of 120 (3600/30). Similarly, lane density, total density, average lane speeds and average speed are calculated for 30 s data aggregation by applying the procedure in 3.3.2.

The pair-plots of the fundamental diagram (FD) is shown in fig. 5.6. In figure, it can be seen that at low densities ranging from 0-60 vehicles/km, the average speed is close to the speed limit in the range of 100-120 km/h. The maximum traffic flow of around 7000 vehicles/h occurs at a speed between 80-100 km/h. The flow in high speed regime ranges between 1000-6000 vehicles/h whereas the flow decreases when there is a drop in the speed. The relationship between flow and density is also evident. Initially, the flow increases upto the road capacity with increase in density, but a further increase in traffic density results in a drop in the traffic flow.

The analysis of the track-wise/ location wise trends of the macroscopic traffic parameters is not possible in the fig. 5.6. To analyze the location wise trends, the box plots in fig. 5.7, 5.8 and 5.9 shows the variation of total flow, total density and average speeds for different tracks. The mean section speed in majority of the tracks is between 100 km/h - 120 km/h indicating high speed traffic conditions (free-flow). The few tracks such as track no. 11, 12, 13, 14, 25, 26, 27 and 46 shows significant drop in the average speeds which change in traffic conditions. In these few tracks, increase in traffic density also points to the development of congestion like conditions. The mean flow in the track nos. 04, 05, 06, 07, 08, 09 and 10 is lower than than the mean flow in other tracks. Majority of the recordings correspond to high speed traffic flow. A few recordings capture the lower speeds as well. The pre-domianance of the free-flow states in the highD data was also found by Kruber et al. (2019). Only few recordings have high variation in the traffic parameters. This may be due to the reason that the recordings correspond to a short duration (< 20 minutes) with majority showing no traffic transition.

The box plots in fig. 5.10,5.11 and 5.12 shows the variation of lane flow, lane density and average lane speeds for different tracks for both the driving directions. The average speeds on the right lanes in both directions is lower than the average speeds on the middle and left lanes. This is because left lane is normally reserved for over-taking purposes only. Another reason is that the heavy vehicles like trucks generally prefer to use right lanes due to their low speeds compared to the cars. The average lane flow on right lane is observed to be lower than other lanes due to lower speeds and longer vehicles (trucks).

The recordings showing large variation in the speed are further analyzed to see the trend of the parameters with time. The time series plot of the traffic density, flow and velocity for the selected recordings showing large variation i.,e recording no 12, 25, 26 and 46 is further inspected. The plots of traffic parameters for upper (Westward direction) and lower lanes (Eastward) are shown in fig. 5.13 and 5.14. It can be seen that upper lanes in tracks no. 12 and 46 show speed increase from around 50 km/h to 100 km/h. In lower lanes on tracks no 25 and

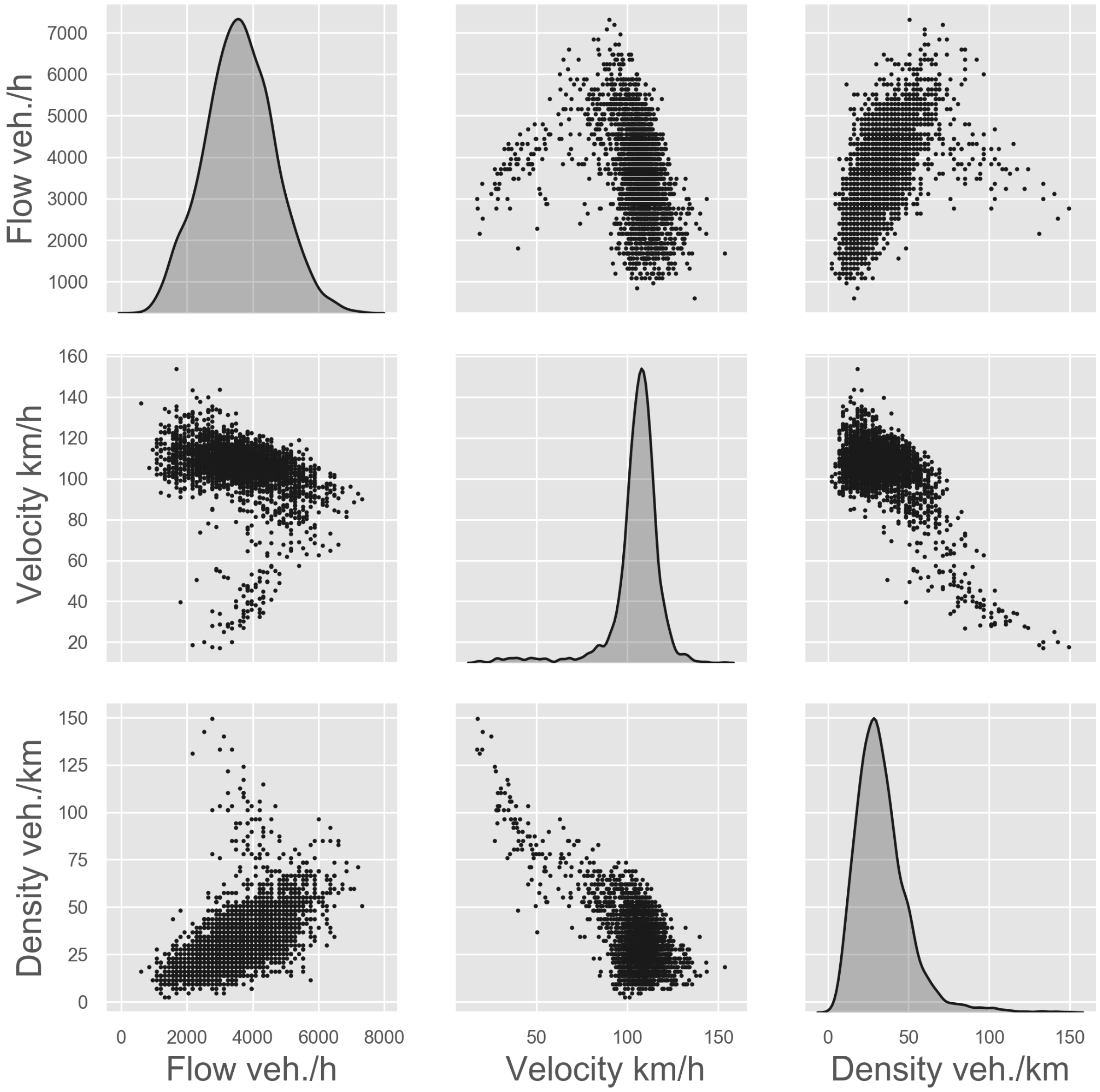

Figure 5.6.: Pair plots of the Traffic FD

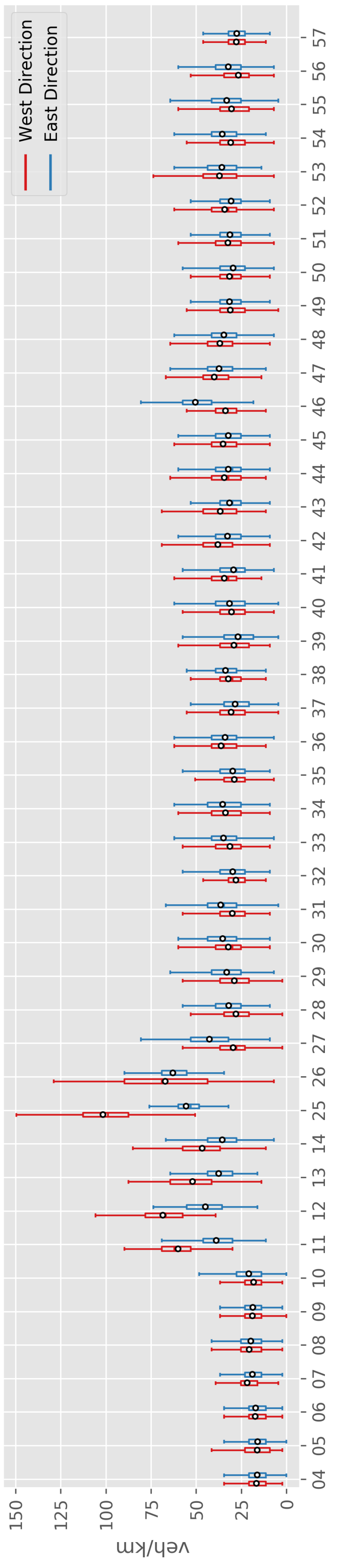

Figure 5.7.: Average Density

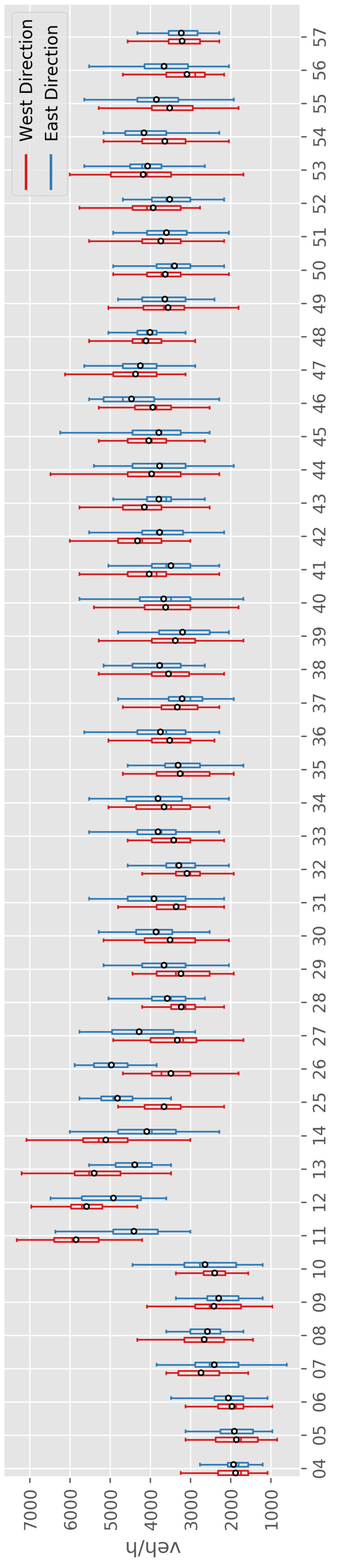

Figure 5.8.: Average flow

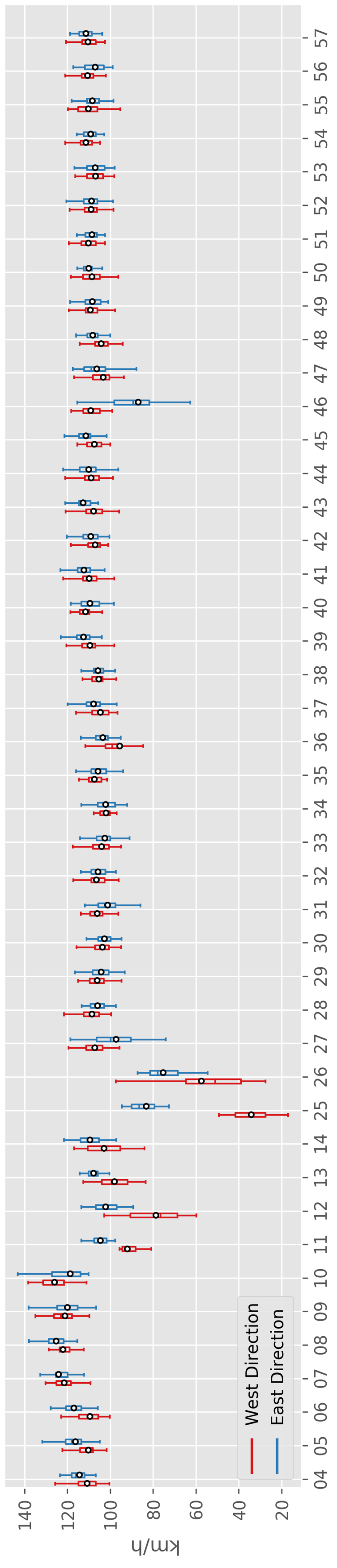

Figure 5.9.: Average Speed

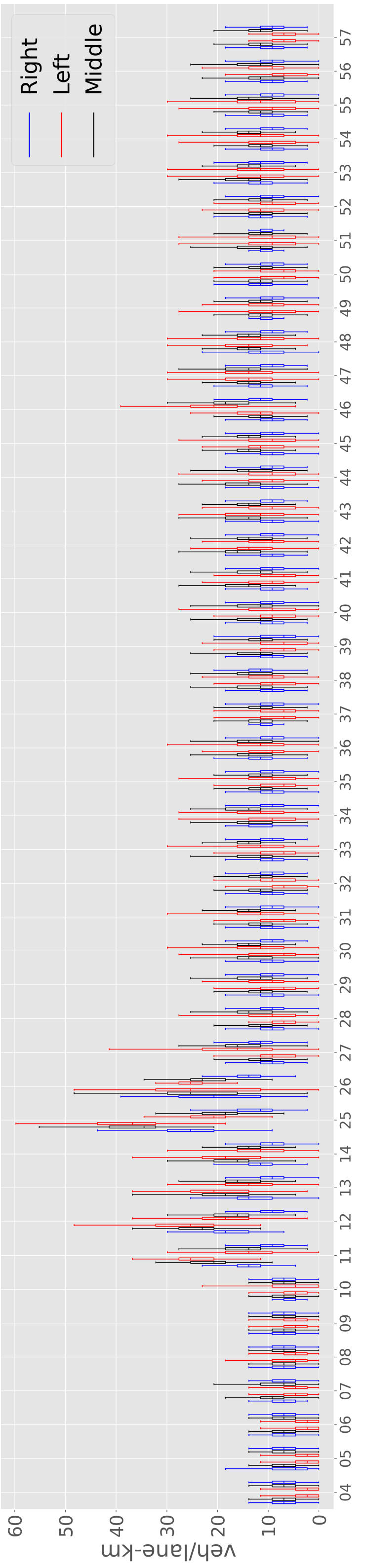

Figure 5.10.: Average lane-wise density

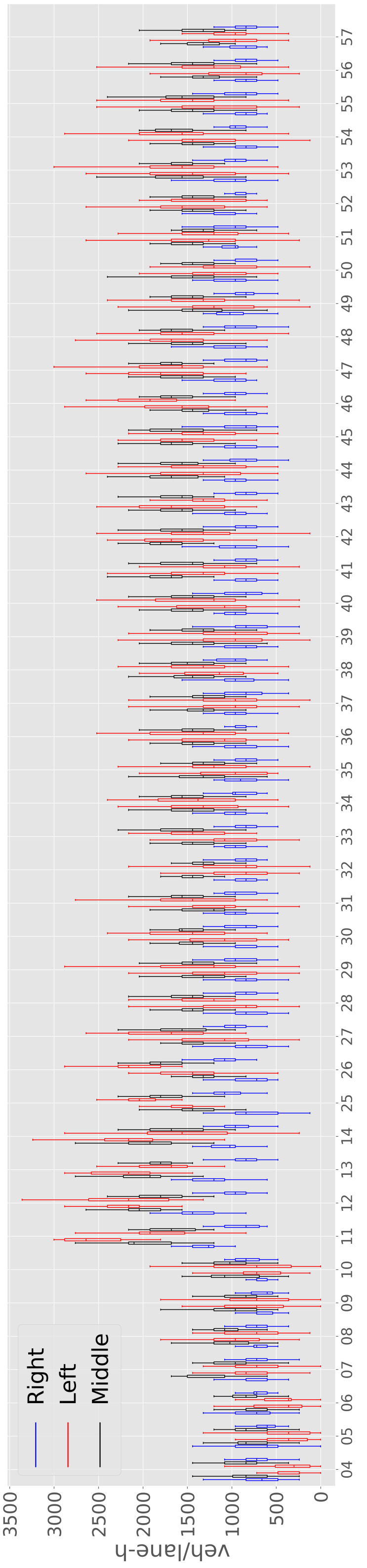

Figure 5.11.: Average lane-wise flow

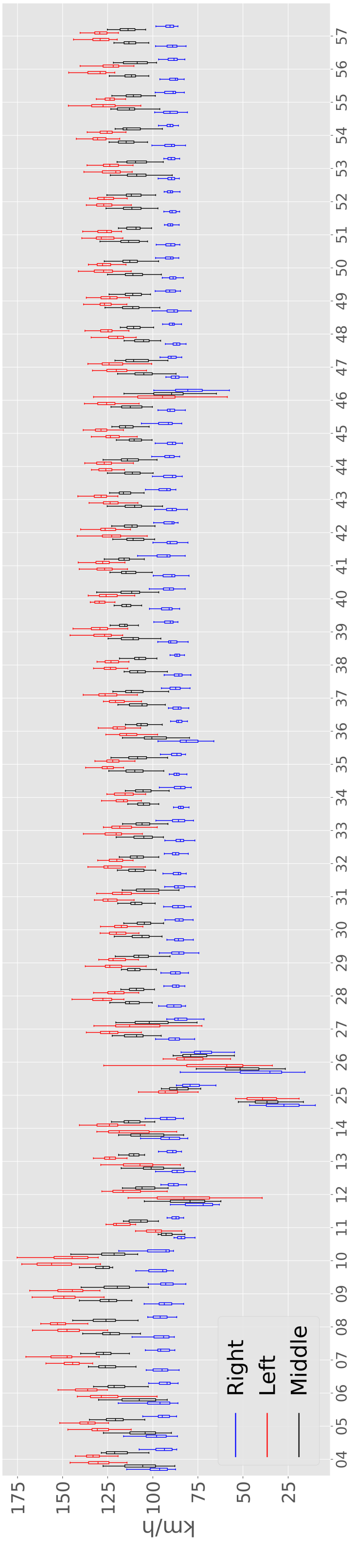

Figure 5.12.: Average lane-wise Speed

26, initially a speed increase is observed which is follower by the speed drop towards the end. The lower lanes in track no. 46 show a consistent increase in speed 25 km/h to around 125 km/h. These trends give an indication of the flow transition from free-flow to congestion or vice-versa.

5.3. Conclusion

The data analysis has brought out some interesting insights about the data. MTTC and Crash severity are found to be not correlated as intended for separate estimation of likelihood and severity at disaggregate level. The data corresponds majorly to free-flow traffic but contains few tracks during which transition of traffic state is observed, which shows that there is some variability in the traffic conditions.

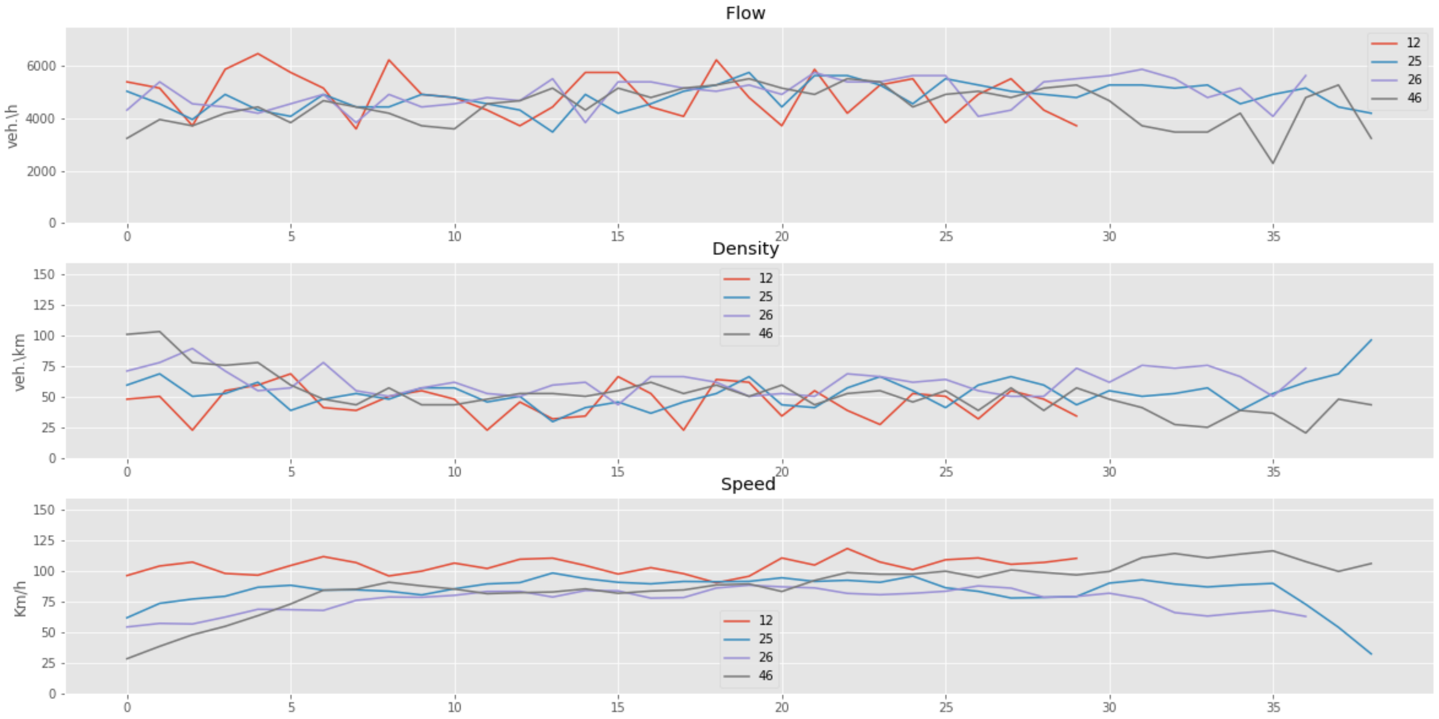

Figure 5.13.: Time series of the traffic variables for Lower lanes for track no 12, 25, 26, 46

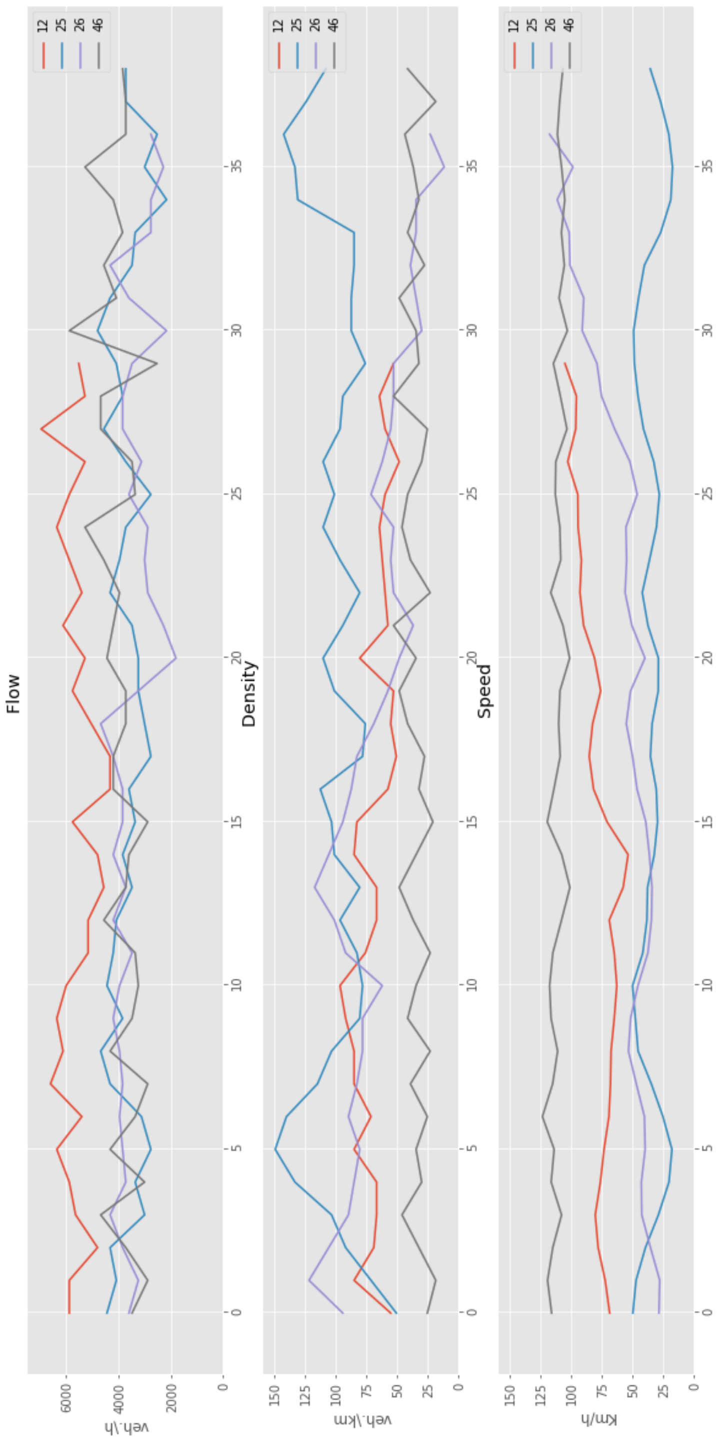

Figure 5.14.: Time series of the traffic variables of Upper Lanes for Track no 12, 25, 26, 46

Part III.

Results and Conclusion

6. Results

This chapter starts with the results of the driving maneuver labeling and intention prediction. Thereafter, the results of traffic state clustering are presented. Finally, the chapter deals with crash risk estimation and hypothesis testing results.

6.1. Driving Intention

Only the track no. 45 with 3 lane configuration in each direction is selected for the labeling of driving maneuver i.e., lane changing or lane keeping. This dataset contains 2449 vehicles recorded in 18 minutes and contains 334 vehicles which executed a lane change. The selection of this is made randomly but the results of this track are tested on tracks from other locations in the dataset to ensure transferability of results.

6.1.1. Maneuver analysis on aggregated data

The results of PCA to obtain two Principal components (PCs) on the aggregated features of vehicle motion (velocity and acceleration) for longitudinal motion, lateral motion and Net motion are shown in table 6.1. The two principal components explain 97% of the variance in aggregated lateral motion and the principal component PC_1 explains up for 94% of the variance. The obtained PCs are plotted with respect to the ground truth of lane change from the data i.e., If vehicle executed at least one lane change or Did not execute a lane change at all. The plots are shown in figures 6.1, 6.2 and 6.3, where it can be seen that lateral motion explains the distinction of lane changing from No-lane changing more clearly as evident from the separation of the two groups in fig. 6.1. This trend is also confirmed on other tracks as shown in figure 6.4. This shows that lateral velocity and lateral acceleration are significant features to distinguish between lane changing and no lane changing maneuvers. This forms the basis of the selection of only the lateral motion for further analysis for driving intention identification from the disaggregate data.

Table 6.1.: PCA on aggregated metrics

Motion type	Features	No. of aggregated features (μ and σ)	Variance PC_1 (%)	Variance PC_2 (%)
Lateral motion	v_y, a_y	4	94	3
Longitudinal motion	v_x, a_x	4	52	24
Net motion	v_y, a_y, v_x, a_x	8	49	24

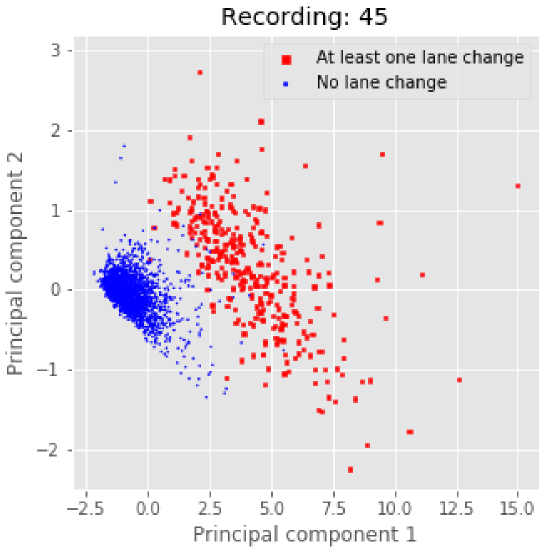

Figure 6.1.: PC obtained from aggregate features of lateral motion

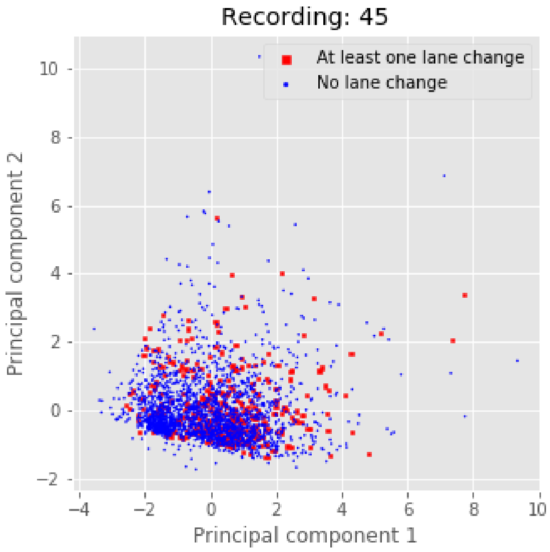

Figure 6.2.: PC obtained from aggregate features of longitudinal motion

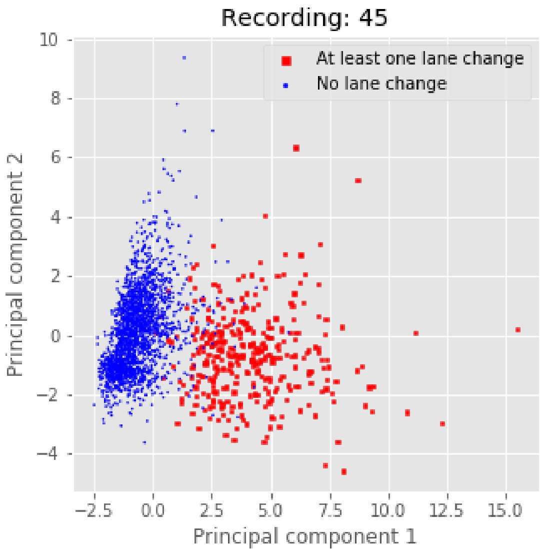

Figure 6.3.: PC obtained from aggregate features of net motion

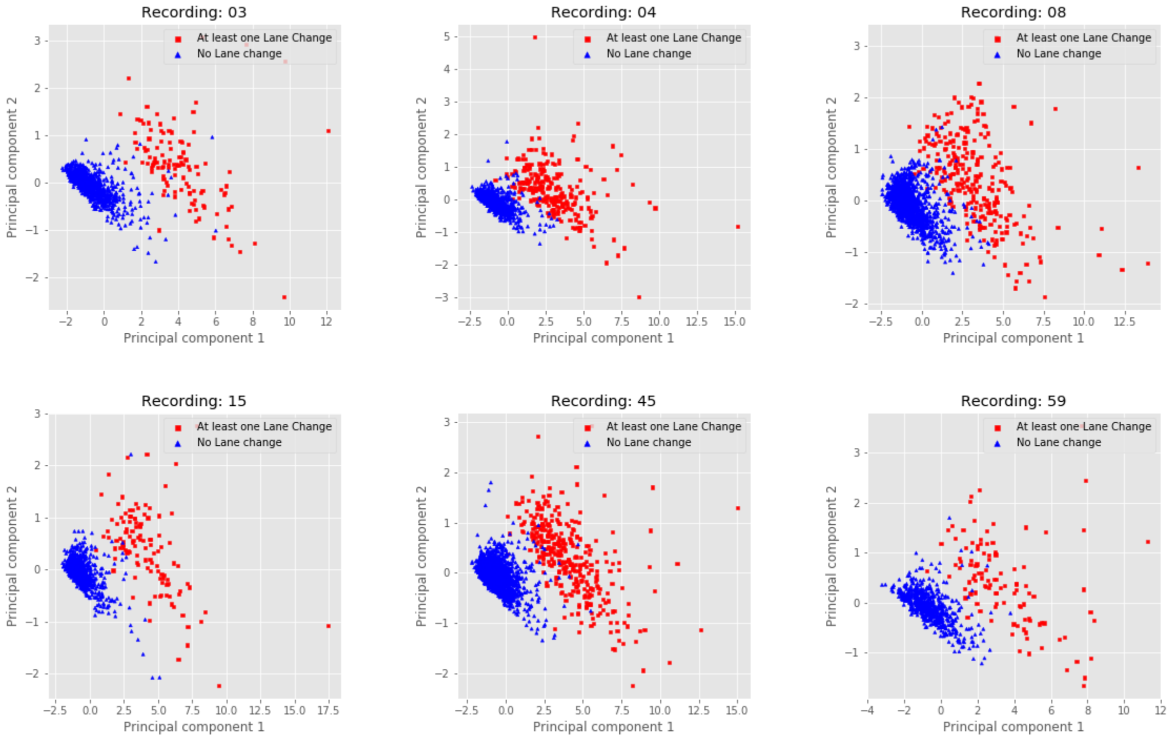

Figure 6.4.: Scatter plot of the Principal components of the Trajectories: At least one lane change (red) and No lane change (Blue)

6.1.2. Driving Maneuver clustering from real-time motion

The results of the different clustering algorithms K-means, agglomerative clustering, GM and DBSCAN on the two lateral motion features v_y and a_y are shown in fig 6.5. The agglomerative clustering is used for different linkage algorithms (single, complete, ward). The DBSCAN clustering is used with eps = 0.05 and different values of minimum samples of for clustering (10, 50 and 80). The DBSCAN and GM algorithms show good results. The GM is able to identify two clusters but the boundaries are not well defined since lot of points of high density region are lying outside the inner cluster. The reason DBSCAN performs well is that the points of low density on the outside represents extreme values of v_y and a_y during lane change whereas the region of high density core corresponds to the points when the vehicle is not changing lanes, thus exhibiting negligible lateral motion. The change of DBCAN parameters also does not significantly changes the clustering (fig. 6.5) which points to the fact that it is less sensitive and more robust for the current task.

The result of the density-based clustering of the manoeuvres on the bi-variate time series of v_y and a_y are labelled as shown in fig. 6.6. The low density points (in red) are labelled as lane changing maneuver and high density core (blue) is labelled as lane keeping maneuver. The parameters for clustering are eps =0.05 and minimum samples = 80 with Euclidean metric for distance calculation. These parameters are obtained by tuning the clustering with respect to Silhouette score. The Silhouette score for the two identified clusters corresponding to two manoeuvre classes is 0.77 which indicates good clustering performance. Each point in this scatter plot corresponds to the lateral acceleration and lateral velocity of a vehicle at a time instant. The distribution plot of transverse acceleration and transverse velocity for lane keeping manoeuvre follow a uni-modal distribution with a mean close to 0 and narrow deviation. This is due to the reason that during lane keeping there are small deviations with respect to the intended path. The distribution of lane changing has bi-modal distribution with wide deviation. This is due to the reasons that a lane change manoeuvre involves a significant acceleration and velocity during the lateral movement. The two peaks on the velocity distribution for lane changing is due to the left and right-side lane changes.

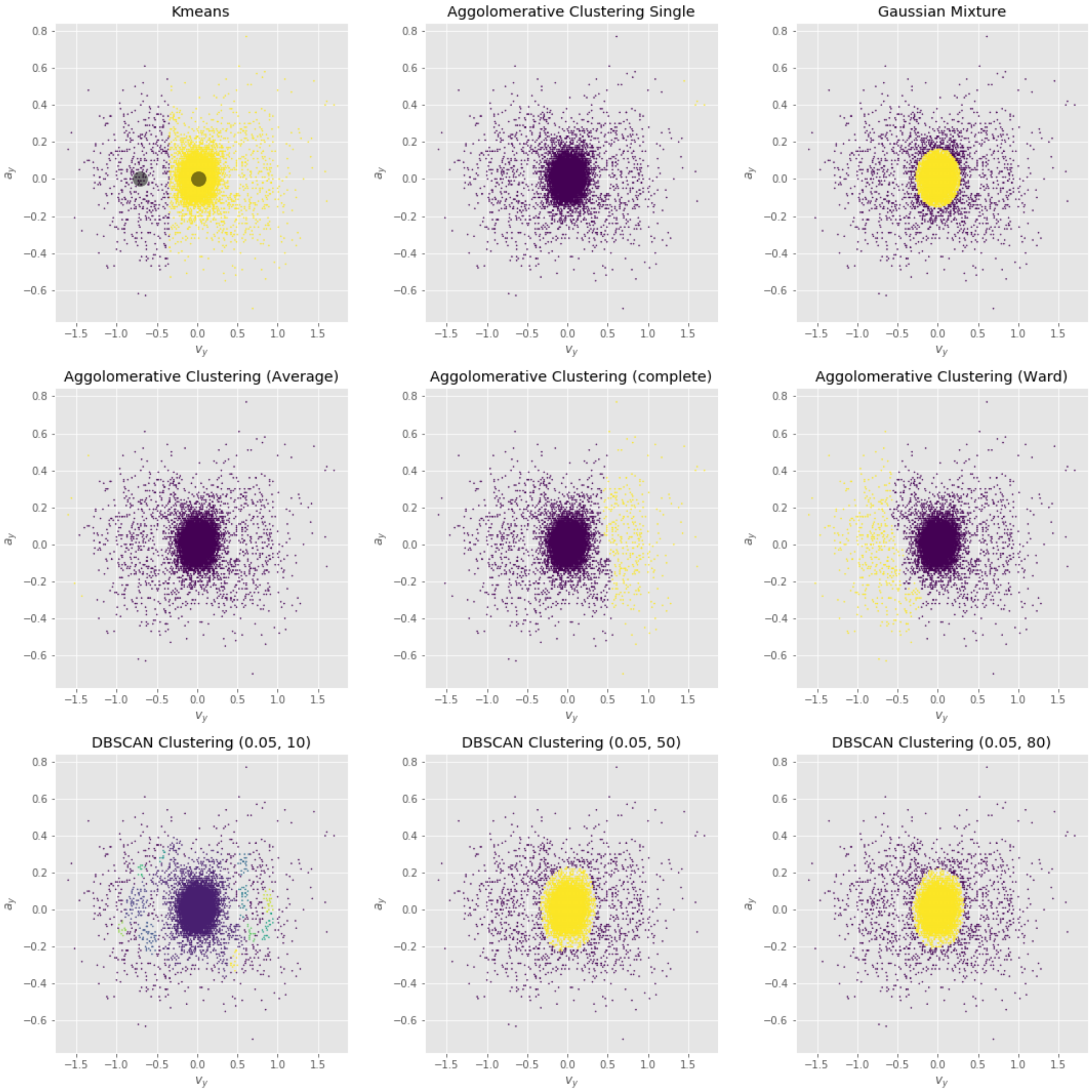

Figure 6.5.: Results of the different clustering algorithms

Table 6.2.: Evaluation metrics of the SVM in manoeuvre classification

Manoeuvre Class	Precision	Recall
Lane Changing	0.99	0.99
Lane Keeping	1	1

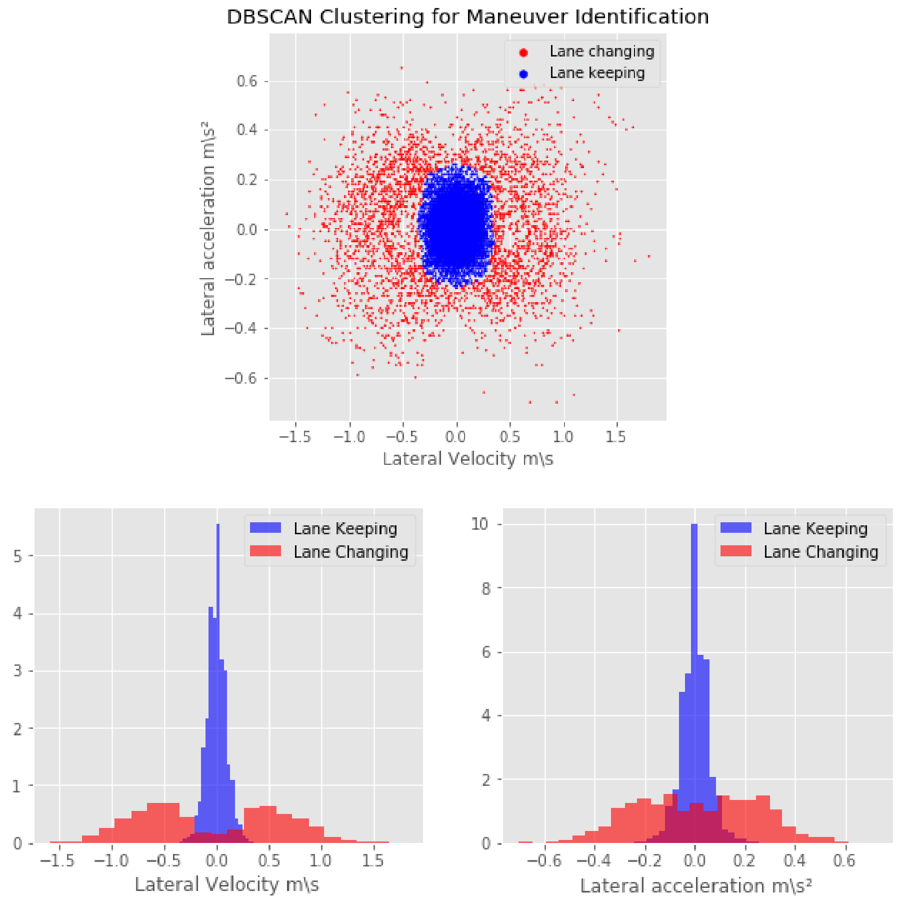

Figure 6.6.: Top: Density-based clustering of the manoeuvres: Lane change (red) and Lane Keeping (blue); Bottom: Distribution of the Lane keeping and Lane changing Manoeuvres

The data is split into training and validation data (80:20) for training the maneuver labelling classifier. A SVM classifier is trained on the clustered labels to learn the representations of the two manoeuvres and thus label the bi-variate time series data during the inference stage. The best parameters for SVM are C = 0.5 with a radial basis function as kernel. Table 6.2 shows the performance of the SVM manoeuvre classifier on the validation data. The manoeuvre classifier achieves an accuracy of 99.8% with high recall and precision, which shows the excellent performance of the SVM in distinguishing both classes. Based on the labelling, the average duration of a full lane change is about 5.68 seconds with a standard deviation of 0.77 seconds.

6.1.3. Driving Intention Prediction

The data for the prediction model is split in 80:20 proportion for training and validation in terms of the vehicle trajectories. The test data consists of 10715 trajectories (6198 trajectories with full lane change and 4517 trajectories with no lane change), sampled from all the six locations in the data-set.

In this study, the best configuration of the LSTM predictor was identified by its performance on the validation dataset and the training time. The adopted model consists of two LSTM layers, each containing 50 units as shown in fig. 6.7. The layers are stacked on top of each other to enable the model to learn higher level temporal dependencies. The second LSTM layer return its output to a dense layer with 20 neurons. This dense layer is further connected with two dense layers with 20 and 10 neurons. The final layer is also a dense layer but with softmax activation and the output of the final layer is the probability of the two manoeuvre classes. The random forest classifier is trained with estimators=10 and maximum depth of the tree=15.

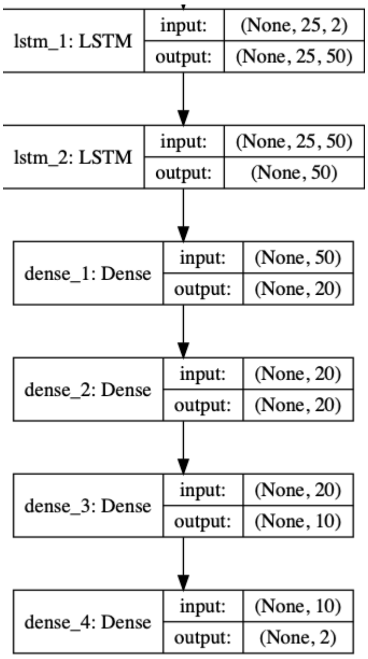

Figure 6.7.: Architecture of the LSTM model

The look back period greater than 1 second did not seem to have a significant improvement on the performance of the model. It was found that sampling the data with frame granularity, f $>$ 1 results in longer training times, whereas training the model at the raw frame frequency i.e., f =1 is faster and achieves the convergence much earlier. Thus, values of f =1 and k_t = 1 second are used for the model training and evaluation. The loss and accuracy curves of LSTM model for training data and validation data for different values of prediction horizon time pt is shown in Figure 6.8. It can be seen that accuracy decreases with increase in prediction horizon since the dependencies are too long to be learned even by the LSTM model.

Table 6.3.: Evaluation metrics for LSTM and Random Forest model

Look-back time (sec)	Prediction horizon(s)	Accuracy (%)	
		Random Forest	LSTM
1	0.5	97.2	98.8
1	1	94.5	97.6
1	2	88	93
1	3	83	88

Table 6.3 compares the results of intention prediction from the RF and LSTM models. At small prediction horizons, it can be seen that RF is as accurate as LSTM. However, when the prediction horizon increases, LSTM performs better than RFs. This is because LSTM is better capable to model the sequence or temporal dependencies. The LSTM model achieves an accuracy of above 97% for prediction horizon upto 1 second which is very significant for highway driving scenarios. The accuracy drops significantly for prediction time greater than 1 second. As a result, the LSTM model with prediction horizon of 0.5 seconds is used for the test data due to its high accuracy. Thus, the LSTM model can predict the lane change manoeuvre 0.5 seconds before the start of manoeuvre with high accuracy.

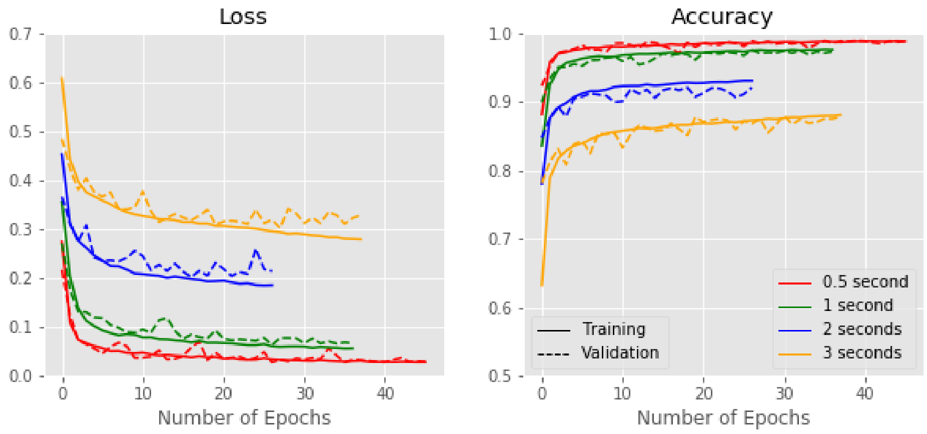

Figure 6.8.: Loss and accuracy during training and validation of LSTM model for different prediction horizons

The complete lane change detection model is evaluated on the test data by using recall, precision and advance detection time. The model is able to detect the lane change maneuver with a recall of 0.99 and precision of 0.98. The advance detection time has a mean of 3.18 seconds and standard deviation of 0.98 seconds. The distribution of the advance detection time is shown in Figure 6.9. The 90^{th} and 99^{th} percentile time for advance detection is 4.04 s and 6 s respectively. Few examples of the manoeuvre prediction are shown in Figure 6.10. In these figures, the actual trajectory is color coded as per the prediction and actual status of the manoeuvre. The initial length of trajectory is colored in black since no data is available to make the predictions for this length. The correct prediction for lane changing and lane keeping is shown in green and blue color respectively. It can be seen in fig. 6.10 that model performs accurately for different lane change scenarios, such as left lane change, right lane change, passing or return, etc. The few instances of misclassifications of lane change and lane keeping are shown in red and orange color. Thus, the model is able to predict the lane changing intention very well in advance before the vehicle crosses into adjacent lane. The mean detection of about 3 seconds achiever after automatic labeling is better/ comparable to the ones established in the previous studies as shown in Table 2.1.

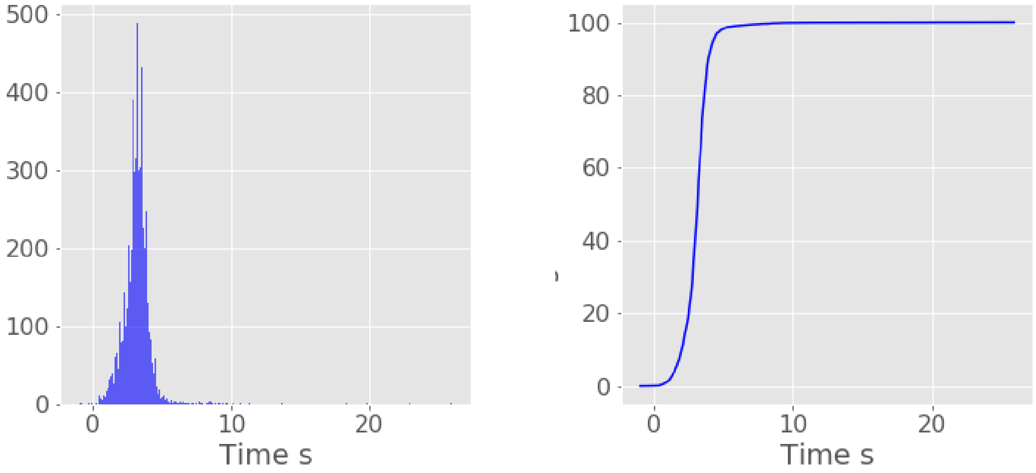

Figure 6.9.: Distribution and cumulative distribution of advance detection time

6.2. Traffic State

The results of the traffic state clustering using GM algorithm on speed and density data is shown in the fig. 6.11. The parameters for clustering are covariance type ('tied') and number of components (2). Based on the clustering, the boundary between the free and the congestion state is identifiable at the average speed of 70 km/h. The traffic flow above this speed falls in free-flow cluster and vice-versa for congestion cluster. The labeled traffic state pair-plots are shown in fig. 6.12.

The summary statistics of the two traffic states are shown in table 6.4 and table 6.5. The mean speeds in free-flow and congestion flow are 107 km/h and 46 km/h respectively. The minimum traffic speed in congestion is 17 km/h whereas the maximum speed in the free-flow state is 154 km/h. The count of free-flow is about 29 times the count of congestion states as the occurrence of congestion states is way less than the free flow states in the HighD Dataset.

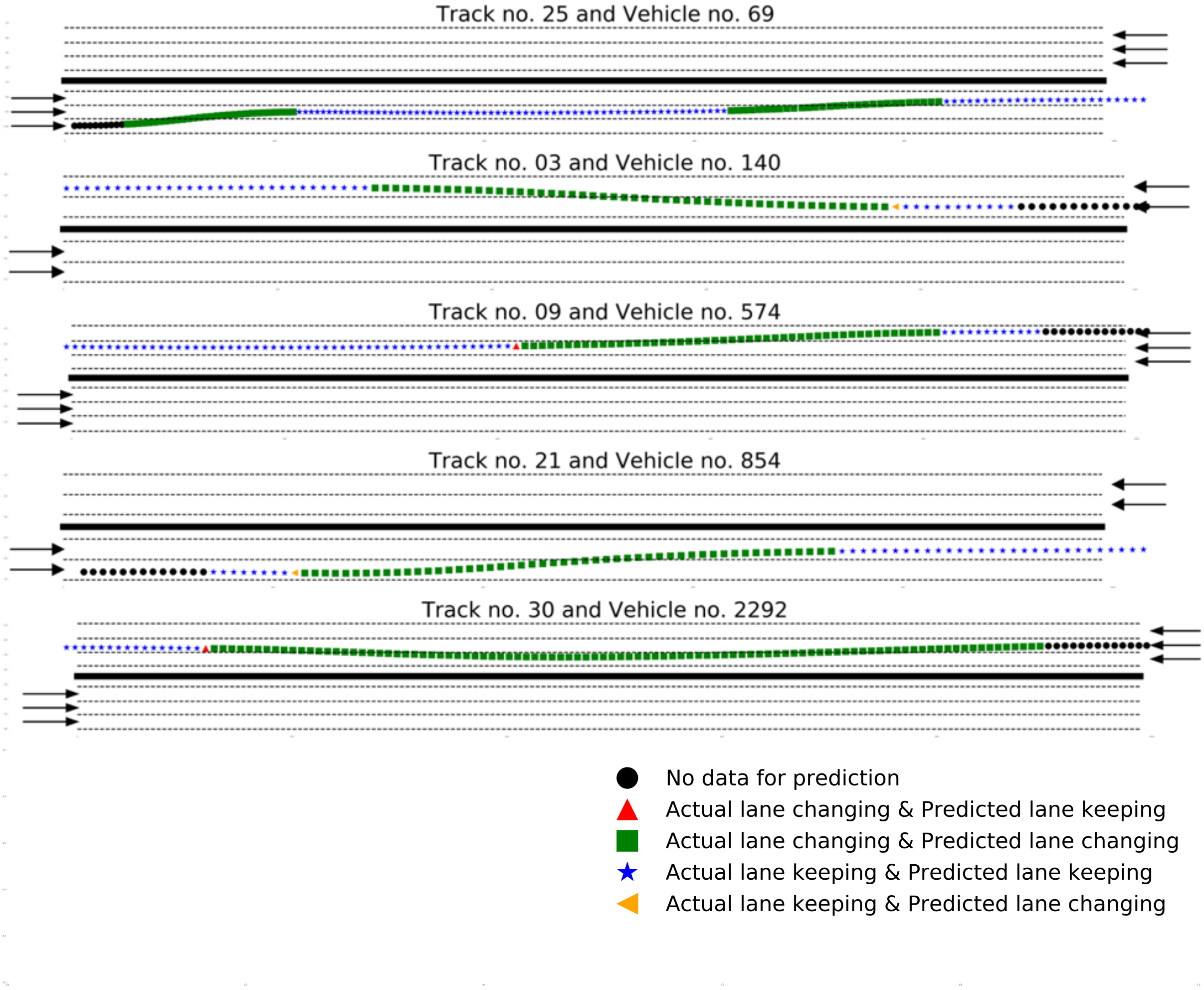

Figure 6.10.: Examples of model predictions of the manoeuvre class on test data

Table 6.4.: Summary of Free flow Cluster

	Flow (veh/h)	Speed (km/h)	Density (veh/km)
count	2958	2958	2958
μ	3594	107	32
σ	1057	9	13
min	600	70	2
median	3600	107	30
max	7320	154	83

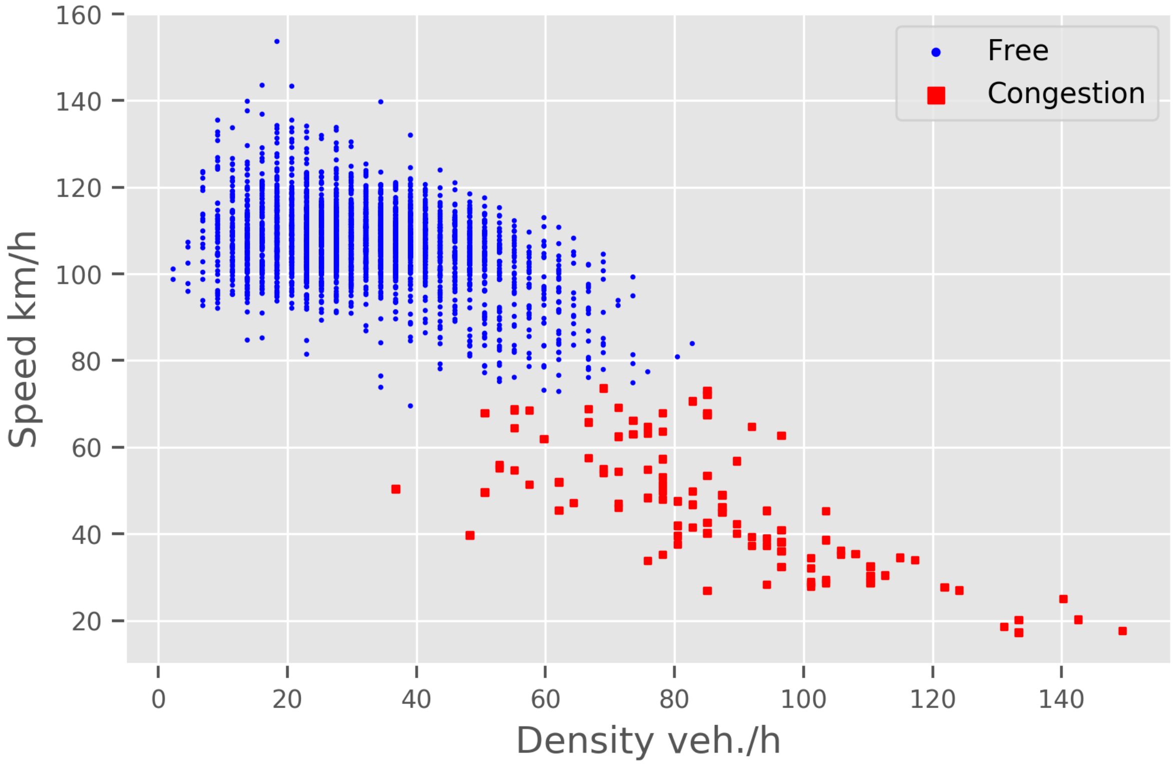

Figure 6.11.: Results of GM clustering for Traffic state

Table 6.5.: Summary of Congestion Cluster

	Flow (veh/h)	Speed (km/h)	Density (veh/km)
count	102	102	102
μ	4091	46	86
σ	941	15	22
min	1800	17	37
median	3960	46	85
max	6600	74	149

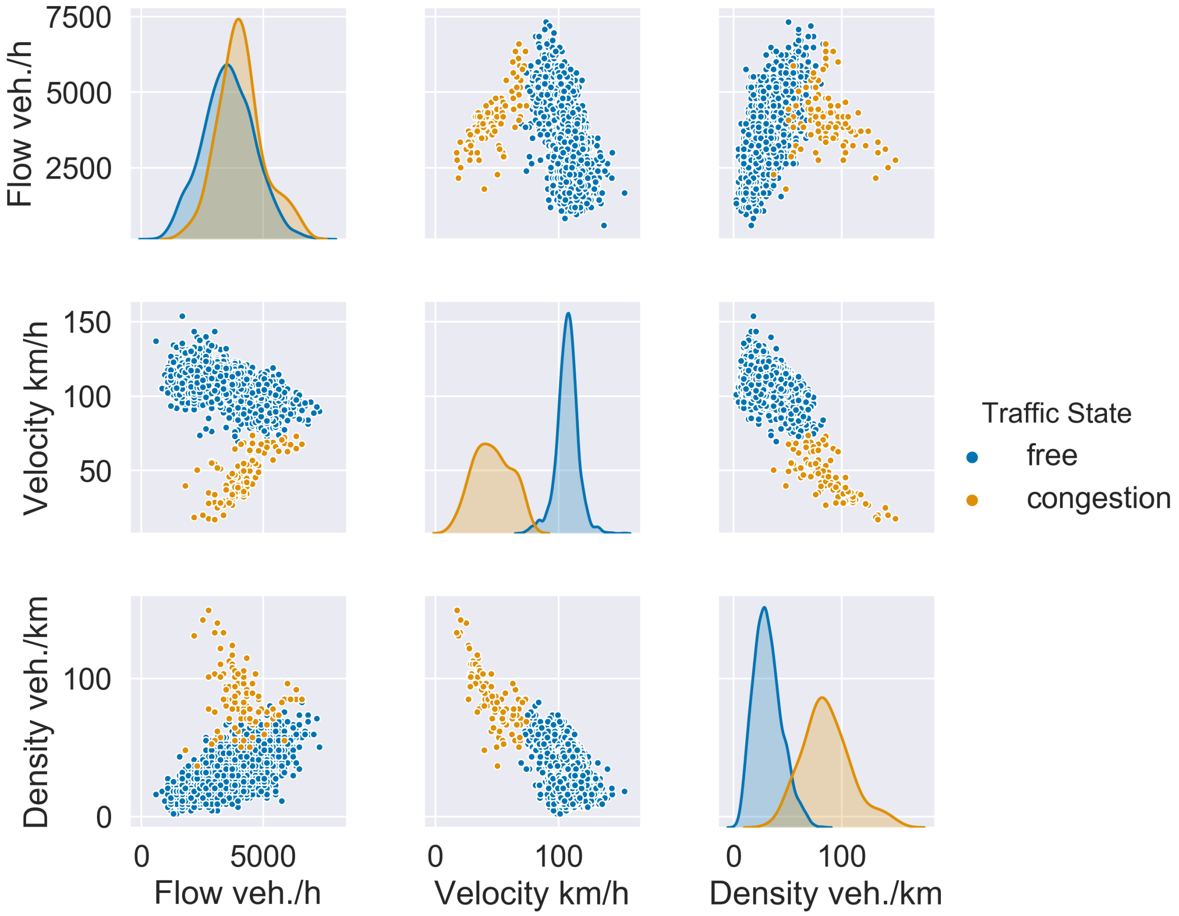

Figure 6.12.: Pair plots of the traffic state clustering

Table 6.6.: Crash Risk Estimates

	P_{Crash}	S_{Crash}	Flow	ACL	ACI	Total Risk
μ	217.14	8927.87	30.09	6.91	287.48	3.45E+03
σ	342.21	4896.25	8.81	12.40	127.78	2.04E+04
min	0.49	460.58	5.00	0.07	92.12	1.02E+01
Median	146.75	8142.09	30.00	5.01	270.06	1.31E+03
max	5470.90	46882.26	61.00	251.89	2350.61	5.92E+05

6.3. Crash Risk

The minimum value of MTTC is 0.02s and is observed in track no. 25. The maximum value of CRIM is observed track no 10, when the subject vehicle no. 845 is traveling at a speed of 205 km/h whereas the speed of preceding vehicle is 115 km/h. The maximum values of the ACL and ACI is 251 and 2350 and both are observed in track no. 25 during a lane change event. The summary statistics of the aggregate crash likelihood and crash severity for all the traffic states is shown in table 6.6. The ACL and ACI have a mean 7 and 287 respectively.

The pair-plot of the crash likelihood and severity is shown in fig. 6.13. In the plot, the ACL, ACI and total risk are scaled down by dividing each one of them by its maximum obtained values. It can be seen that the value of the ACL and ACI have high values in the congestion traffic state, whereas they have low values in the free-flow states. The ACL and ACI have a strong positive correlation with each other. The correlation of the likelihood, severity and total risk with speed is negative, implying higher crash risk during speed drops on a freeway.

The distribution of the crash likelihood and crash severity is analyzed for extraction of low/ normal vs high risk cases. There are not distinct clusters in the likelihood and severity distribution due to the reason that their formulation is based on continuous mathematical functions and hence the discrete risk groups cannot be clustered. The manual thresholds are defined to divide the likelihood and severity distribution into two groups of high vs low. The likelihood and severity show uni-modal distributions with a long right tail (left skewed). This is evident in the pair-plots in fig. 6.13. It is assumed that normal behaviour or low risk cases follows a symmetrical normal distribution and the instances in the long tail belong to abnormal behaviour or high risk cases. Using this approach, a threshold is identified to separate the low risk cases from high risk cases. The thresholds for crash likelihood is at scaled ACL = 0.2 and that for crash severity is at scaled ACI = 0.4. The distribution for the crash likelihood and crash severity below these thresholds follows close to a symmetrical distribution i.e., with equal left and right tails. The instances above the selected thresholds show a decreasing trend. The distribution of the crash likelihood and severity after dividing into two classes is shown in fig. 6.14 and 6.15.

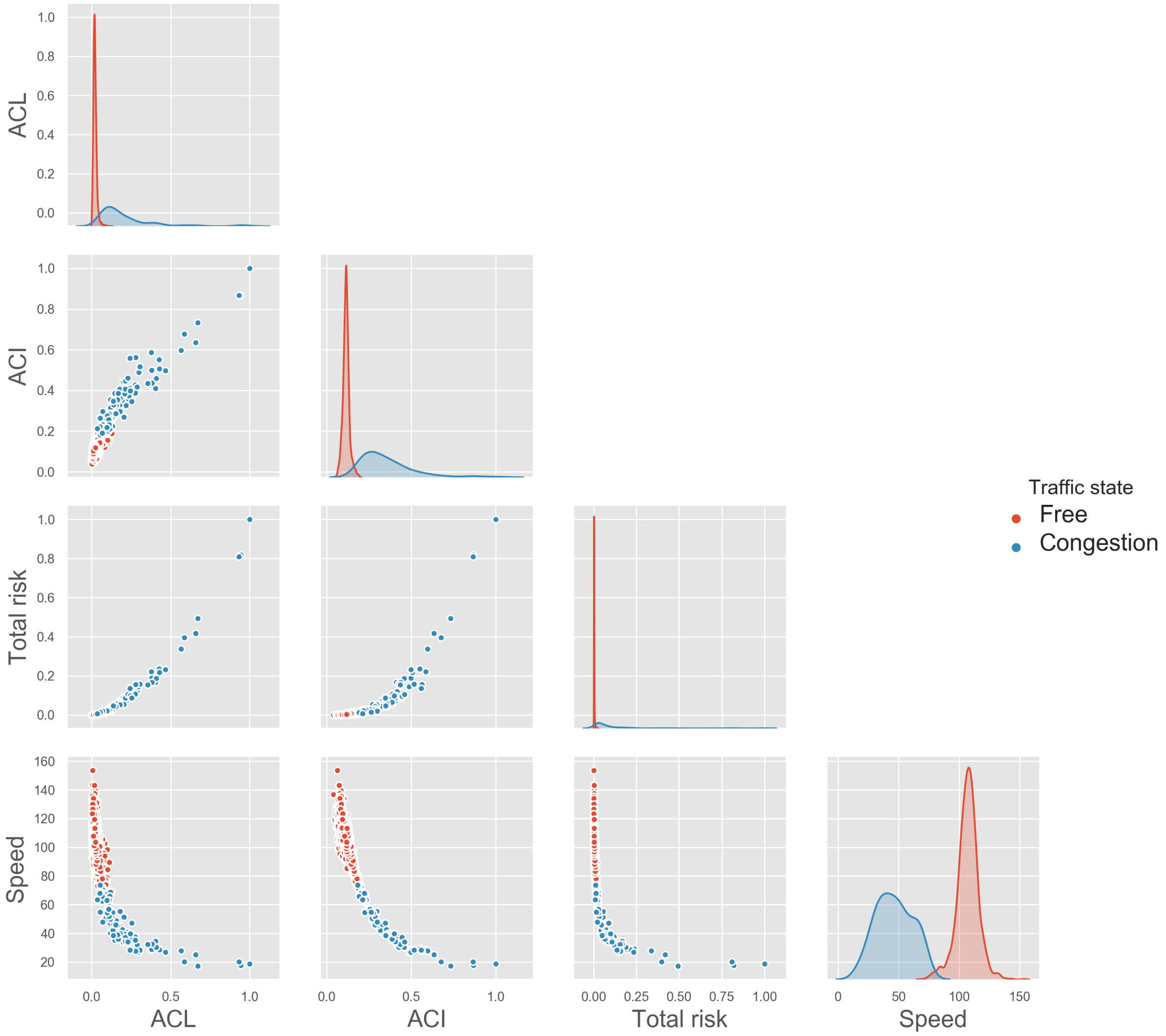

Figure 6.13.: Pair plots of the likelihood, Severity and risk of a crash. The ACI, ACL and total risk values are divided by their maximum.

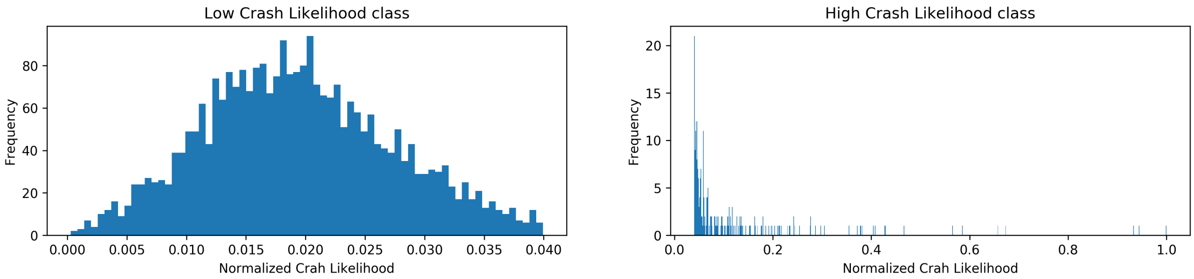

Figure 6.14.: Distribution of crash likelihood classes

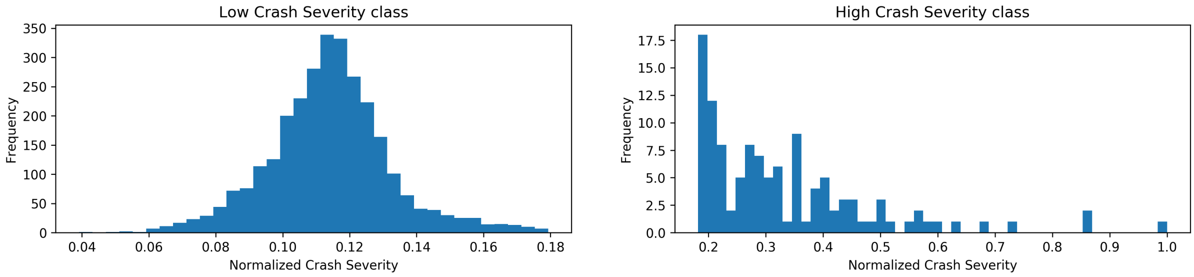

Figure 6.15.: Distribution of crash severity classes

The grouping of the estimated risk into high and low class for both the likelihood and severity gives four combinations of the crash risk. These combinations are high crash likelihood & high crash severity (HL-HS), low crash likelihood & high crash severity (LL-HS), high crash likelihood & low crash severity (HL-LS) and low crash likelihood & low crash severity (LL-LS). These four risk categories are plotted on the speed-flow diagram as shown in 6.16. Here, the four risk classes are assigned an overall risk i.e., Extreme, High, Medium and Low for for the combinations HL-HS, LL-HS, HL-LS and LL-LS respectively. It can be seen that total risk increases as the speed on the highway drops which implies congestion like conditions. Now, it also is possible to analyze the risk dependence on the driving intention and traffic states. The crash likelihood and severity is separately analyzed with respect to its correlation with the traffic states and individual driving intention in the nexts section.

6.4. Effect of Intention and Traffic on risk

There are significant differences in the distribution of the crash risk across the driving intention and traffic states. Firstly, due to the same aggregation level of 30s used for the estimation of crash estimate and traffic states, it is possible to visualize them on the same plot. The speed-flow diagram for the traffic classes and the crash classes (likelihood and severity) is shown is shown in fig 6.17. It can be seen that majority of the high crash likelihood and all the high crash severity instances fall in the congestion/ unstable traffic flow conditions. A few instances of high crash likelihood fall in the free flow traffic conditions.

The sample occurrences of the crash risk classes with respect to the state of the traffic are shown in table 6.7. Further, the sample occurrences are also shown by traffic state for lane changing vehicles and lane keeping vehicles separately in table 6.8 and 6.9.

The sample occurrences of the crash risk classes with respect to the driving intention are shown in table 6.10. Further, the sample occurrences are also shown for free traffic state and congestion traffic state separately in table 6.11 and 6.12.

The proportion of the crash likelihood and crash severity from the sample are analyzed for the (risk) difference and odds ratio and their 95% confidence interval and its signifi-

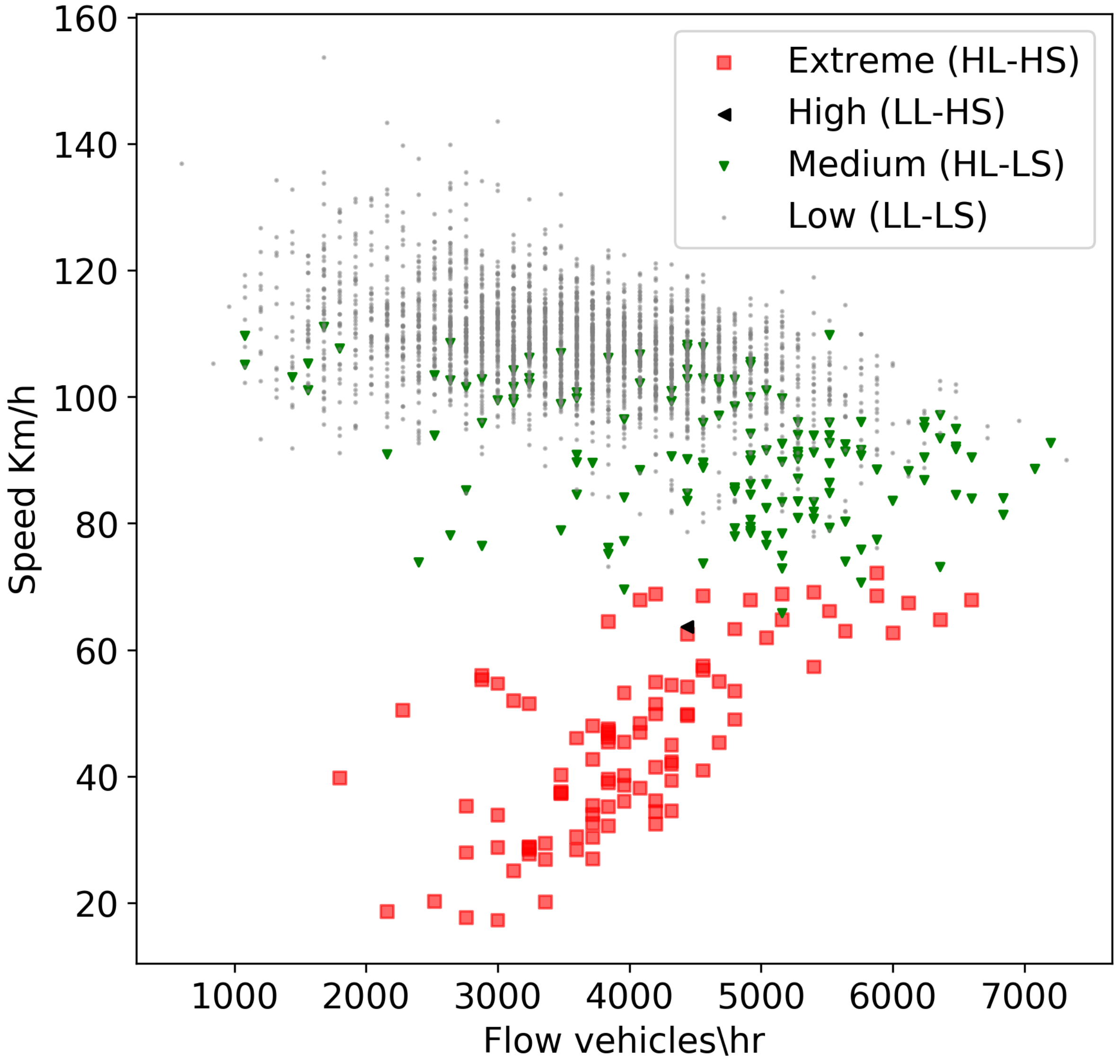

Figure 6.16.: Distribution of Risk classes

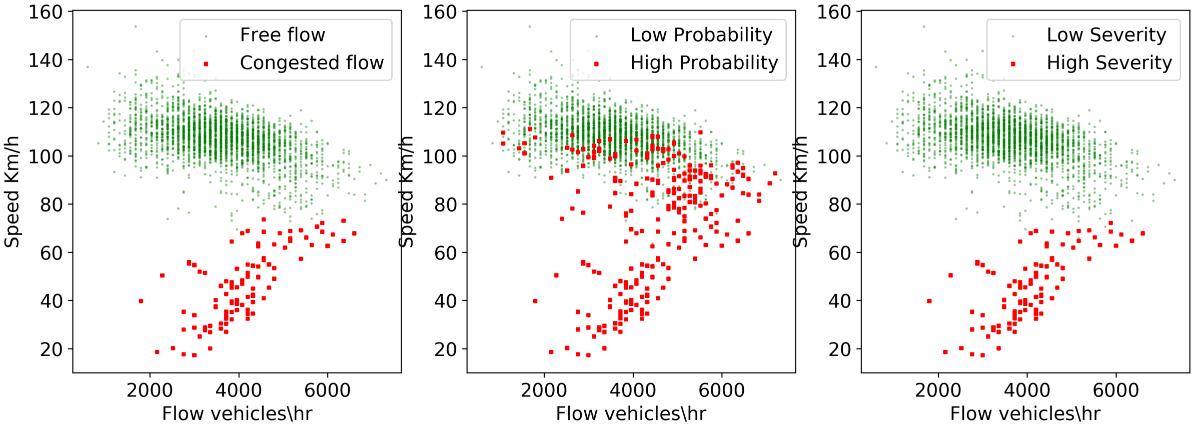

Figure 6.17.: Crash classes on speed-flow diagram

Table 6.7.: Crash classes for all vehicles by Traffic State

Crash risk class	Free-flow	Congestion
High Likelihood and High Severity	0	97
High Likelihood and Low Severity	147	4
Low Likelihood and High Severity	0	1
Low Likelihood and Low Severity	2811	0

Table 6.8.: Crash classes for Lane Changing vehicles by Traffic State

Crash risk class	Free-flow	Congestion
High Likelihood and High Severity	34	76
High Likelihood and Low Severity	1247	16
Low Likelihood and High Severity	7	2
Low Likelihood and Low Severity	1570	2

Table 6.9.: Crash classes for lane keeping vehicles by Traffic State

Crash risk class	Free-flow	Congestion
High Likelihood and High Severity	0	98
High Likelihood and Low Severity	90	3
Low Likelihood and High Severity	0	1
Low Likelihood and Low Severity	2868	0

Table 6.10.: Crash classes for all traffic states by driving intention

Crash risk class	Lane changing	Lane keeping
High Likelihood and High Severity	110	98
High Likelihood and Low Severity	1263	93
Low Likelihood and High Severity	9	1
Low Likelihood and Low Severity	1572	2868

Table 6.11.: Crash classes for free-flow state by driving intention

Crash risk class	Lane changing	Lane keeping
High Likelihood and High Severity	34	0
High Likelihood and Low Severity	1247	90
Low Likelihood and High Severity	7	0
Low Likelihood and Low Severity	1570	2868

Table 6.12.: Crash classes for congestion state by driving intention

Crash risk class	Lane changing	Lane keeping
High Likelihood and High Severity	76	98
High Likelihood and Low Severity	16	3
Low Likelihood and High Severity	2	1
Low Likelihood and Low Severity	2	0

cance. The hypothesis testing for the crash likelihood and crash severity is done. The exposure variable are the traffic congestion and lane change intention. In case of outcome variable as crash likelihood and exposure as traffic congestion, the null hypothesis is: *Traffic congestion does not effects the likelihood of crashes*. In case the lane change intention is selected as exposure, the null hypothesis for testing is: *Crash likelihood is same during lane changing and lane keeping*. Similarly, In case of outcome variable as crash severity and traffic congestion as exposure, the null hypothesis is: *Traffic congestion does not effects the severity of crashes*. In case the individual intention is selected as exposure, the null hypothesis for testing is: *Lane changing maneuvers does not cause increase in the severity of the crashes*. The contingency tables for hypothesis testing are obtained using the data in table 6.7 to table 6.12. These tables are reproduced in Appendix A.2.

The table 6.13 shows the results of the hypothesis testing for the crash likelihood. From the table, it can be seen that that the effect of traffic congestion is significant for all intentions although the magnitude of the effect is different as reflected in risk difference and odds ratio. The traffic congestion results in increased crash likelihood overall, i.e., for all the vehicles. The reason might be that during a congestion, there is formation of a shock-wave and there is a high chance of a crash due to speed difference between the congestion flow and flow at the upstream of the congestion. During congestion, the odds for the increased likelihood for the lane keeping vehicles are more than those for the lane changing vehicles. The fig. 6.18 shows the risk classes for the lane keeping and lane changing vehicles on two different plots. It can be seen that effect of congestion is more prominent for lane keeping vehicles as evident by the increase of risk with capacity drop. The reason might be that the lane keeping vehicles reduce their speeds upto the congestion speed at the upstream end of the congestion whereas the lane changing vehicles change their lanes to select the optimum lane (if possible) during a congestion flow. Therefore, the speed differences between following and preceding vehicles exists for a longer duration for the lane keeping vehicles.

The lane changing intention is found to be significant for increased crash likelihood. The effects of the lane changing is more pronounced during free flow conditions as compared to that during congestion as shown in fig. 6.18. The reason could be the larger variability in lane speeds during free conditions because the left lane is used for overtaking purposes in Germany whereas the right lane is used otherwise. Thus, a lane changing vehicle will need to increase or decrease its speed from its initial lane to the final lane due to which there is more friction and higher chances of a rear crash. The variability in lane speeds and the required

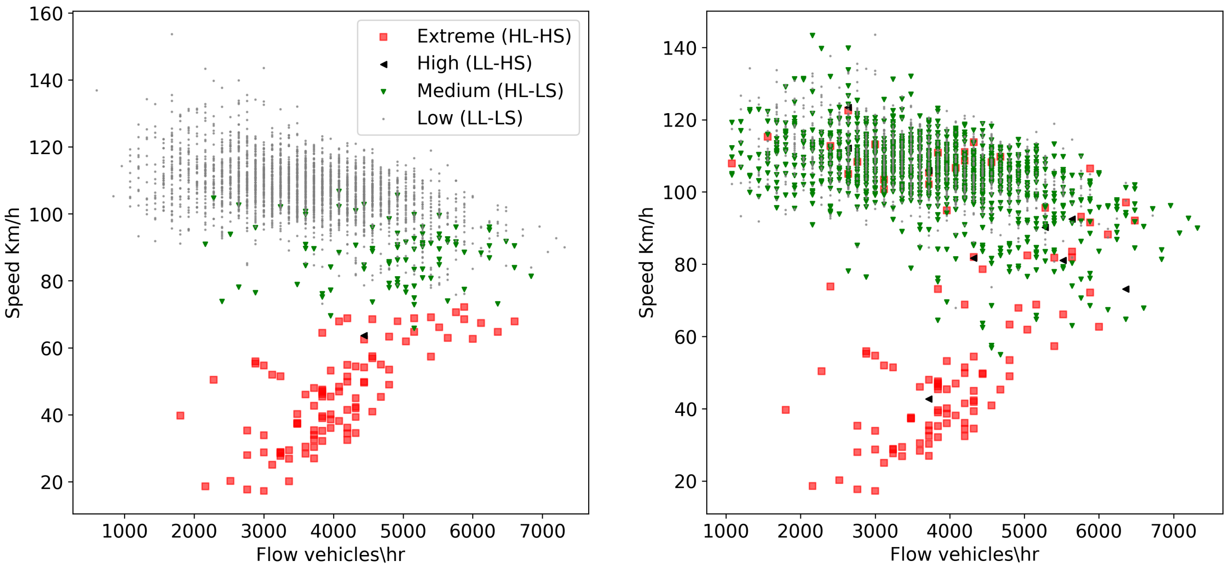

Figure 6.18.: Risk for Lane Keeping (Left) and Lane Changing (Right) intention

Table 6.13.: Crash Likelihood: risk difference and odds ratio

Sample Control	Exposure	Outcome: Crash Likelihood					
		Risk Difference			Odds Ratio		
		Difference	95% CI	p value	Ratio	95% CI	p value
-	Traffic congestion	0.94	0.92-0.96	<0.001	1931	267-13941	<0.001
Intention: Lane changing	Traffic congestion	0.51	0.46-0.55	<0.001	28	10-77	<0.001
Intention: Lane Keeping	Traffic congestion	0.96	0.94-0.98	<0.001	3218	444-23330	<0.001
-	Lane changing	0.40	0.38-0.42	<0.001	13	11-15	<0.001
Traffic: Free	Lane changing	0.41	0.39-0.43	<0.001	25	20-32	<0.001
Traffic: Congestion	Lane changing	-0.032	0.076-0.012	0.144	0.2	0 - 2	0.169

speed increase/ decrease for changing lanes is smaller during the congestion as compared to the free-flow conditions. That is why, effect of lane changing on crash likelihood during congestion is not found to be significant as observed by its high p values in table 6.13.

The table 6.14 shows the results for the Crash Severity. From the table, it can be seen that the effect of traffic congestion is significant for all intentions, although the magnitude of effect is different as reflected in risk difference and odds ratio. This is again due to the reason that speed adjustment is needed at the upstream end of the congestion. Due to this speed differential, the congested flow shows greater odds for a severe rear-end crashes. The odds ratio for the effect of congestion on lane keeping vehicles is shown as infinity due to absence of high severity outcomes during the free flow conditions. As far as intention is concerned, lane changing is found to be significant for increase in crash severity during the free flow conditions. The fig. 6.18 also shows that due to lane changing, the high severity points can be seen during the free-flow traffic as well. This might be again due to larger lane-speed variability between left and right lanes during free flow conditions. The lane changing is found to be associated with reduction in crash severity during congestion based on the negative risk

Table 6.14.: Crash Severity: risk difference and odds ratio

Sample Control	Exposure	Outcome: Crash Severity					
		Risk Difference			Odds Ratio		
		Difference	95% CI	p value	Ratio	95% CI	p value
-	Traffic congestion	0.96	0.92-0.98	<0.001	∞	-	-
Intention: Lane changing	Traffic congestion	0.80	0.72-0.87	<0.001	297	163-541	<0.001
Intention: Lane Keeping	Traffic congestion	0.97	0.93-1.00	<0.001	∞	-	-
-	Lane changing	0.00	0.00-0.01	0.103	1.2	0.9-1.6	0.104
Traffic: Free	Lane changing	0.14	0.01-0.02	<0.001	∞	-	-
Traffic: Congestion	Lane changing	-0.06	-0.24-0.07	0.001	0.131	0.037-0.462	0.003

difference and less than 1 odds ratio.

The consequences of congestion on the crash likelihood are in agreement with studies by Golob, Recker, and Alvarez (2004), Kuang et al. (2017), and Marchesini and Weijermars (2010) who have too found positive correlation of rear-end crash risk with speed drops, speed variability or congestion. Regarding Severity, congestion conditions are perceived to result in less severe crashes (Chang and Xiang, 2003; Marchesini and Weijermars, 2010). The congestion in outer lanes is also found to be associated with low likelihood of injury crashes by Golob, Recker, and Pavlis (2008). They also found that transition from free flow to congestion has no impact on severity. This anomaly might be due to the fact that their severity analysis is based on accident and resulting severity of the injury. The conflict cases where a possible severe crash is avoided due to human action/ response are not reflected in these studies. Another possible reason could be short length of the road section used in this study due to which only partial observation of congestion and its effects is possible.

6.5. Main Findings

Firstly, The density based clustering is efficient for clustering of the driving maneuvers from the disaggregate data. The intention prediction model based on these automatic labels is efficient in predicting driving intentions on an average 3 seconds before lane change with only few false alarms. Secondly, The results show that the congestion/ speed drop has an increased likelihood as well as increased severity as compared to the crash risk under free flow. In terms of intention, vehicles executing a lane changing manoeuvre are at an increased likelihood and severity of a crash as compared those not changing lanes during the free-flow traffic conditions. The effect of lane changing during congestion is somewhat significant for rear-end crash severity but not significant for increase in likelihood.

7. Conclusion

Lane changing and lane keeping are the two primary lateral driving maneuvers on highways. Their timely detection is one of the keys to highway safety and cooperative driving, however, data-driven prediction of these manoeuvres has so far been constrained by the manual labelling of training data-sets. This study bridges the gap in the literature by demonstrating that only with lateral movement data in terms of velocity and acceleration can distinguish whether a vehicle will carry out a lane change or not. A density based clustering approach and a SVM classifier demonstrated good results for automatically identifying and labelling manoeuvres as lane keeping or lane changing.

Crash risk has two components namely, likelihood and severity. MTTC and a newly introduced impact surrogate: *CRIM* give an estimation for the rear crash risk by quantifying its likelihood and severity separately. The estimation of aggregate risk and traffic states shows that the congestion causes increased likelihood and severity of the rear-end crash. Lane changing intention is found to be significant for increase in crash likelihood and severity during free flow traffic conditions. Although, this study only analyses rear-end crashes for risk estimation, but it is possible apply this approach to other crash types by incorporating crash specific likelihood and severity surrogates.

The unsupervised labeling approach is generic and can easily be transferred to other locations or scenarios for developing end-to-end intention detection models. Furthermore, the developed LSTM model for predicting manoeuvres over trajectories from different highway locations shows significant performance and can detect lane changes at an average 3 seconds before a vehicle enters into a new lane. The results of the study are envisioned to enhance highway safety, as the successful and timely prediction can lead to better coordination among vehicles and proactive alerts to the drivers in the near automated future. The insights from the effect of intention and traffic on change in crash likelihood and severity provides an evidence for real time traffic management system policies.

7.0.1. Limitations and Future Work

The presented research is not without its own limitations. The data recorded were obtained from a short highway segment, which consequently limits the number and nature of the identified manoeuvres and traffic interactions. Observing a longer road segment could help in observing more interactions. Finally, this study does not differentiate maneuver labels into left or right lane changes, and as a result the utilization of the direction of longitudinal and lateral velocity with respect to a frame of reference should be further researched. The model needs validation in real driving scenario to check its robustness with respect to noise in sensor measurements, unlike the trajectory data which is post-processed and is free of such noise.

Also, data has small variability in traffic states since most traffic conditions belong to free-flow. Further research can be done on the validation of the crash severity estimation method for more diverse traffic scenarios where the accident severity data is also available. The crash risk method is only an estimation from intention and traffic states and not a real time prediction. Thus prediction of traffic states based on long duration data can be helpful in predicting crash risk in real time. This means that ability to predict traffic states and driving intention can help in eventually predicting crash risk, which can be explored in future.

Bibliography

Abdel-Aty, M. and Pande, A. (2008). “The Viability of Real-time Prediction and Prevention of Traffic Accidents”. *Efficient Transportation and Pavement Systems, U.K.: Taylor & Francis*, 215–226.

Abdel-Aty, Mohamed and Pande, Anurag (2007). “Crash Data Analysis: Collective vs. individual Crash Level Approach”. *Journal of Safety Research* 38.5, 581–587.

Ahmed, Kazi Iftekhar (1999). “Modeling Drivers ′ Acceleration and Lane Changing Behavior”. *Transportation* Ph.D, 189.

Alexiadis et al. (2004). “The Next Generation Simulation Program”.

Andrews, Scott (2012). “Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) Communications and Cooperative Driving”. *Handbook of Intelligent Vehicles*. Ed. by Azim Eskandarian. London: Springer London, pp. 1121–1144.

Aoude, G. S. et al. (2011). “Behavior Classification Algorithms at Intersections and Validation using Naturalistic Data”. *2011 IEEE Intelligent Vehicles Symposium (IV)*, pp. 601–606.

Azevedo, Carlos Lima, Cardoso, João L., and Ben-Akiva, Moshe E. (2018). “Probabilistic Safety Analysis using Traffic Microscopic Simulation”. *ArXiv* abs/1810.04776.

Beglerovic, Halil et al. (2018). “Deep Learning Applied to Scenario Classification for Lane-Keep-Assist Systems”. *Applied Sciences* 8.12.

Bengio, Y., Simard, P., and Frasconi, P. (1994). “Learning Long-term Dependencies with Gradient Descent is Difficult”. *IEEE Transactions on Neural Networks* 5.2, 157–166.

Breiman, Leo (2001). “Random Forests”. *Machine Learning* 45.1, 5–32.

Chan, Ching-Yao (2006). “Defining Safety Performance Measures of Driver-Assistance Systems for Intersection Left-Turn Conflicts”. *2006 IEEE Intelligent Vehicles Symposium*, pp. 25–30.

Chang, G-L and Xiang, H (2003). *The Relationship between Congestion Levels and Accidents*. Publication RR-8379. Maryland State Highway Administration, Baltimore.

Chen, C., Liu, L., et al. (2018). “Driver’s Intention Identification and Risk Evaluation at Intersections in the Internet of Vehicles”. *IEEE Internet of Things Journal* 5.3, 1575–1587.

Chen, Peng, Zeng, Weiliang, et al. (2017). “Surrogate Safety Analysis of Pedestrian-Vehicle Conflict at Intersections Using Unmanned Aerial Vehicle Videos”. *Journal of Advanced Transportation* 2017.

Chen, Tianyi, Shi, Xiupeng, and Wong, Yiik Diew (2019). “Key Feature Selection and Risk Prediction for Lane-changing Behaviors based on Vehicles Trajectory Data”. *Accident Analysis & Prevention* 129, 156–169.

Cheung, E. et al. (2018). “Identifying Driver Behaviors Using Trajectory Features for Vehicle Navigation”. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3445–3452.

Chollet, François et al. (2015). *Keras*. https://keras.io.

Dang, H. Q. et al. (2017). “Time-to-lane-change Prediction with Deep Learning”. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–7.

Daszykowski, M. and Walczak, B. (2009). “2.29 - Density-Based Clustering Methods”. *Comprehensive Chemometrics*. Ed. by Steven D. Brown, Romfffdfffd Tauler, and Beata Walczak. Oxford: Elsevier, pp. 635–654.

Dataset, HighD (2019). *Application for Access*. URL: https://www.highd-dataset.com/#download (visited on 01/05/2019).

Dimitriou, Loukas, Stylianou, Katerina, and Abdel-Aty, Mohamed A. (2018). “Assessing Rear-end Crash Potential in Urban Locations based on Vehicle-by-vehicle Interactions, Geometric Characteristics and Operational Conditions”. *Accident Analysis & Prevention* 118, 221–235.

European Union (2019). *Road Safety*. visited on: 2019-08-21. URL: https://ec.europa.eu/transport/road_safety/home_en.

FHWA (2004). “NGSIM Task E.1-1: Core Algorithms Assessment, Final Report, Cambridge Systematic, Inc., Massachusetts”, 164.

FHWA, USDOT (2019). *Next Generation Simulation (NGSIM)*. URL: https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm (visited on 05/01/2019).

Fitch, Gregory M. and Hanowski, Richard J. (2012). “Using Naturalistic Driving Research to Design, Test and Evaluate Driver Assistance Systems”. *Handbook of Intelligent Vehicles*. Ed. by Azim Eskandarian. London: Springer London, pp. 559–580.

Gayko, Jens E. (2012). “Lane Departure and Lane Keeping”. *Handbook of Intelligent Vehicles*. Ed. by Azim Eskandarian. London: Springer London, pp. 689–708.

Gettman, Douglas and Head, Larry (2003). “Surrogate Safety Measures from Traffic Simulation Models”. *Transportation Research Record* 1840.1, 104–115. eprint: https://doi.org/10.3141/1840-12.

Gipps, P. G. (1986). "A Model for the Structure of Lane-changing Decisions". *Transportation Research Part B* 20.5, 403–414.

Golob, Thomas F., Recker, Wilfred W., and Alvarez, Veronica M. (2004). "Tool to Evaluate Safety Effects of Changes in Freeway Traffic Flow". *Journal of Transportation Engineering* 130.2, 222–230.

Golob, Thomas F., Recker, Will, and Pavlis, Yannis (2008). "Probabilistic models of freeway safety performance using traffic flow data as predictors". *Safety Science* 46.9, 1306–1333.

Graves, Alex and Schmidhuber, Jürgen (2008). "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks". *Proceedings of the 21st International Conference on Neural Information Processing Systems*. NIPS'08. Vancouver, British Columbia, Canada: Curran Associates Inc., pp. 545–552.

Guido, Giuseppe et al. (2012). "Estimation of Safety Performance Measures from Smartphone Sensors". *Procedia - Social and Behavioral Sciences* 54. Proceedings of EWGT2012 - 15th Meeting of the EURO Working Group on Transportation, September 2012, Paris, 1095–1103.

Hamdar, Samer (2012). "Driver Behavior Modeling". *Handbook of Intelligent Vehicles*. Ed. by Azim Eskandarian. London: Springer London, pp. 537–558.

Herty, Michael and Visconti, Giuseppe (2018). "Analysis of Risk Levels for Traffic on a Multilane Highway". *IFAC-PapersOnLine* 51.9. 15th IFAC Symposium on Control in Transportation Systems CTS 2018, 43–48.

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München*.

Hochreiter, Sepp and Schmidhuber, Jürgen (1997). "Long Short-Term Memory". *Neural Comput.* 9.8, 1735–1780.

Hossain, Moinul and Muromachi, Yasunori (2013). "A Real-time Crash Prediction Model for the Ramp Vicinities of Urban Expressways". *IATSS Research* 37.1, 68–79.

Hourdos, John N. et al. (2006). "Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations". *Transportation Research Record* 1968.1, 83–91. eprint: https://doi.org/10.1177/0361198106196800110.

Hu, Yeping, Zhan, Wei, and Tomizuka, Masayoshi (2018). "Probabilistic Prediction of Vehicle Semantic Intention and Motion". *2018 IEEE Intelligent Vehicles Symposium (IV)*, 307–313.

Hyden, C. (1996). "Traffic Conflicts Technique: State-of-the-art, Traffic Safety Work with Videoprocessing". *Green Series, 43, University Kaiserlauten. Transportation Department*.

Ioannou, P. A. and Stefanovic, M. (2005). "Evaluation of ACC Vehicles in Mixed Traffic: Lane Change Effects and Sensitivity Analysis". *IEEE Transactions on Intelligent Transportation Systems* 6.1, 79–89.

Johnsson, Carl, Laureshyn, Aliaksei, and Ceunynck, Tim De (2018). "In Search of Surrogate Safety Indicators for Vulnerable Road Users: A Review of Surrogate Safety Indicators". *Transport Reviews* 38.6, 765–785. eprint: https://doi.org/10.1080/01441647.2018.1442888.

Katrakazas, Christos et al. (2015). "Real-time Motion Planning Methods for Autonomous Onroad Driving: State-of-the-art and Future Research Directions". *Transportation Research Part C: Emerging Technologies* 60, 416–442.

Kingma, Diederik P. and Ba, Jimmy (2014). "Adam: A Method for Stochastic Optimization". *The Computing Research Repository* abs/1412.6980.

Kirkwood, Betty R. and Sterne, Jonathan A C (2003). "Comparing Two Proportions". *Essential Medical Statistics*. Blackwell Publishing, pp. 148–164.

Kiros, Jamie Ryan, Salakhutdinov, Ruslan, and Zemel, Richard S. (2014). "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models". *ArXiv* abs/1411.2539.

Krajewski, Robert et al. (2018). "The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems". *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* 2018-Novem.November, 2118–2125.

Kruber, Friedrich et al. (2019). *Highway traffic data: macroscopic, microscopic and criticality analysis for capturing relevant traffic scenarios and traffic modeling based on the highD data set*. arXiv: 1903.04249 [eess.SP].

Kuang, Yan, Qu, Xiaobo, and Yan, Yadan (2017). "Will Higher Traffic Flow lead to more Traffic Conflicts? A Crash Surrogate Metric based Analysis". *PLOS ONE* 12.8, 1–11.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey (2015). "Deep learning". *Nature* 521, 436 EP -.

Lee, Chris, Hellinga, Bruce, and Saccomanno, Frank (2003). "Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic". *Transportation Research Record* 1840.1, 67–77. eprint: https://doi.org/10.3141/1840-08.

Lee, S. E., Olsen, C. B., and Wierwille, W. W. (2004). ""A Comprehensive Examination of Naturalistic Lane Changes"". *Report DOT HS 809702 NHTSA U.S. Department of Transportation*.

Lefèvre, Stéphanie, Laugier, Christian, and Ibañez-Guzmán, J (2013). "Intention-Aware Risk Estimation for General Traffic Situations, and Application to Intersection Safety". *INRIA Research Report No. 8379* October.

Lefèvre, Stéphanie, Vasquez, Dizan, et al. (2014). “A Survey on Motion Prediction and Risk Assessment for Intelligent Vehicles”. *ROBOMECH Journal* 1, 1.

Leonhardt, V. and Wanielik, G. (2017). “Neural Network for Lane Change Prediction Assessing Driving Situation, Driver Behavior and Vehicle Movement”. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6.

Liu, Shixia et al. (2017). “Towards Better Analysis of Machine Learning Models: A visual analytics perspective”. *Visual Informatics* 1.1, 48–56.

Mahmud, S. M.Sohel et al. (2017). “Application of Proximal Surrogate Indicators for Safety Evaluation: A Review of Recent Developments and Research Needs”. *IATSS Research* 41.4, 153–163.

Mandalia, Hiren M. and Salvucci, Mandalia Dario D. (2005). “Using Support Vector Machines for Lane-Change Detection”. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49.22, 1965–1969. eprint: https://doi.org/10.1177/154193120504902217.

Marchesini, P. and Weijermars, W. (2010). “The Relationship between Road Safety and Congestion on Motorways: A Literature Review of Potential Effects”. *SWOV Institute for Road Safety Research, Leidschendam*, 1–28.

McCall, Joel C. and Trivedi, Mohan M. (2007). “Driver Behavior and Situation Aware Brake Assistance for Intelligent Vehicles”. *Proceedings of the IEEE* 95.2, 374–387.

Meng, Qiang and Qu, Xiaobo (2012). “Estimation of Rear-end Vehicle Crash Frequencies in Urban Road Tunnels”. *Accident Analysis & Prevention* 48. Intelligent Speed Adaptation + Construction Projects, 254–263.

Minderhoud, Michiel M. and Bovy, Piet H.L. (2001). “Extended Time-to-collision Measures for Road Traffic Safety Assessment”. *Accident Analysis & Prevention* 33.1, 89–97.

Morris, Brendan Tran, Doshi, Anup, and Trivedi, Mohan Manubhai (2011). “Lane Change Intent Prediction for Driver Assistance: On-road Design and Evaluation”. *2011 IEEE Intelligent Vehicles Symposium (IV)*, 895–901.

Naji, Jamil A. and Djebarni, Ramdane (2000). “Shortcomings in Road Accident Data in Developing Countries, Identification and Correction: A Case Study”. *IATSS Research* 24.2, 66–74.

National Highway Traffic Safety Administration (2016). *2016 Fatal Traffic Crash Data*. URL: https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data.

Oh, Jun-Seok et al. (2005). “Real-Time Estimation of Accident Likelihood for Safety Enhancement”. *Journal of Transportation Engineering* 131.5, 358–363.

Olah, Christopher (2019). *Understanding LSTM Networks*. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (visited on 03/07/2019).

Oxford Dictionary (2019). *Machine Learning*. URL: https://tinyurl.com/y4oq3r6b (visited on 01/05/2019).

Ozbay, Kaan et al. (2008). "Derivation and Validation of New Simulation-Based Surrogate Safety Measure". *Transportation Research Record* 2083.1, 105–113. eprint: https://doi.org/10.3141/2083-12.

Özgüner, Ümit, Stiller, Christoph, and Redmill, Keith (2007). "Systems for Safety and Autonomous Behavior in Cars: The DARPA Grand Challenge Experience". *Proceedings of the IEEE* 95, 397–412.

Pande, Anurag, Abdel-Aty, Mohamed A., and Miyake, Hitoshi (2010). "A Classification Tree Based Modeling Approach for Segment Related Crashes on Multilane Highways." *Journal of safety research* 41 5, 391–7.

Pande, Anurag, Abdel-Aty, Mohamed, and Das, Abhishek (2010). "A classification tree based modeling approach for segment related crashes on multilane highways". *Journal of Safety Research* 41.5, 391–397.

Park, Hyunjin et al. (2018). "Development of a Lane Change Risk Index Using Vehicle Trajectory Data". *Accident Analysis & Prevention* 110, 1–8.

Rousseeuw, Peter J. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Journal of Computational and Applied Mathematics* 20, 53–65.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. (1986). "Learning Representations by Back-propagating Errors". *Nature* 323.6088, 533–536.

Russo, Francesca, Busiello, Mariarosaria, and Dell'Acqua, Gianluca (2016). "Safety Performance Functions for Crash Severity on Undivided Rural Roads". *Accident Analysis & Prevention* 93, 75–91.

Sayed, Tarek, Brown, Gerald, and Navin, Francis (1994). "Simulation of traffic conflicts at unsignalized intersections with TSC-Sim". *Accident Analysis & Prevention* 26.5, 593–607.

Schubert, Erich et al. (2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". *ACM Trans. Database Syst.* 42.3, 19:1–19:21.

Scikit-learn (2019a). *Gaussian Mixture Models*. URL: https://scikit-learn.org/stable/modules/mixture.html#mixture (visited on 10/03/2019).

— (2019b). *Hierarchical Clustering*. URL: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering (visited on 10/03/2019).

Stipancic, Joshua et al. (2018). “Surrogate Safety and Network Screening: Modelling Crash Frequency using GPS Travel Data and Latent Gaussian Spatial Models”. *Accident Analysis & Prevention* 120, 174–187.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. (2014). “Sequence to Sequence Learning with Neural Networks”. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’14. Montreal, Canada: MIT Press, pp. 3104–3112.

Treat, J. R. et al. (1979). “Tri-level Study of the Causes of Traffic Accidents. Volume I: Causal Factor Tabulations and Assessments”. *U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington D.C.*

Tseng, George C. and Wong, Wing H. (2005). “Tight Clustering: A Resampling-Based Approach for Identifying Stable and Tight Patterns in Data”. *Biometrics* 61.1, 10–16. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2005.031032.x.

Vapnik, Vladimir N. (1998). “THE Support Vector Method For Estimating Indicator Functions”. *Statistical Learning Theory*. Wiley, pp. 401–442.

Wang, Chao, Quddus, Mohammed A., and Ison, Stephen G. (2009). “Impact of Traffic Congestion on Road Accidents: A Spatial Analysis of the M25 Motorway in England”. *Accident Analysis & Prevention* 41.4, 798–808.

Wang, Kaijun, Wang, Baijie, and Peng, Liuqing (2009). “CVAP: Validation for Cluster Analyses”. *Data Science Journal* 8, 88–93.

Wang, X., Murphey, Y. L., and Kochhar, D. S. (2016). “MTS-DeepNet for Lane Change Prediction”. *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 4571–4578.

Weng, Jinxian and Meng, Qiang (2014). “Rear-end Crash Potential Estimation in the Work Zone Merging Areas”. *Journal of Advanced Transportation* 48.3, 238–249. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/atr.211.

WHO (2018). *Road traffic injuries*. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (visited on 03/07/2019).

Wissing, Christian et al. (2017). “Lane Change Prediction by Combining Movement and Situation based Probabilities”. *IFAC-PapersOnLine* 50.1. 20th IFAC World Congress, 3554–3559.

Woo, H. et al. (2017). “Lane-Change Detection Based on Vehicle-Trajectory Prediction”. *IEEE Robotics and Automation Letters* 2.2, 1109–1116.

Xing, Zhou and Xiao, Fei (2018). “Predictions of Short-term Driving Intention Using Recurrent Neural Network on Sequential Data”. *ArXiv* abs/1804.00532.

Xu, Chengcheng et al. (2013). "Development of a Crash Risk Index to Identify Real Time Crash Risks on Freeways". *KSCE Journal of Civil Engineering* 17.7, 1788–1797.

Xue, Qingwen et al. (2019). "Rapid Driving Style Recognition in Car-Following Using Machine Learning and Vehicle Trajectory Data". *Journal of Advanced Transportation* 2019.1.

Yang, Hong Guang (2012). "Simulation-based Evaluation of Traffic Safety Performance Using Surrogate Safety Measures".

Yang, Xue et al. (2018). "Automatic Change Detection in Lane-level Road Networks using GPS Trajectories". *International Journal of Geographical Information Science* 32.3, 601–621. eprint: https://doi.org/10.1080/13658816.2017.1402913.

Zheng, Zuduo (2012). "Empirical Analysis on Relationship between Traffic Conditions and Crash Occurrences". *Procedia - Social and Behavioral Sciences* 43. 8th International Conference on Traffic and Transportation Studies (ICTTS 2012), 302–312.

— (2014). "Recent Developments and Research Needs in Modeling Lane Changing". *Transportation Research Part B: Methodological* 60, 16–32.

Zheng, Zuduo, Ahn, Soyoung, and Monsere, Christopher M. (2010). "Impact of Traffic Oscillations on Freeway Crash Occurrences". *Accident Analysis & Prevention* 42.2, 626–636.

Zyner, Alex, Worrall, Stewart, and Nebot, Eduardo M. (2018). "Naturalistic Driver Intention and Path Prediction using Recurrent Neural Networks". *CoRR* abs/1807.09995.

Appendix

A. Data Visualizations and Additional Data

A.1. Data Visualization

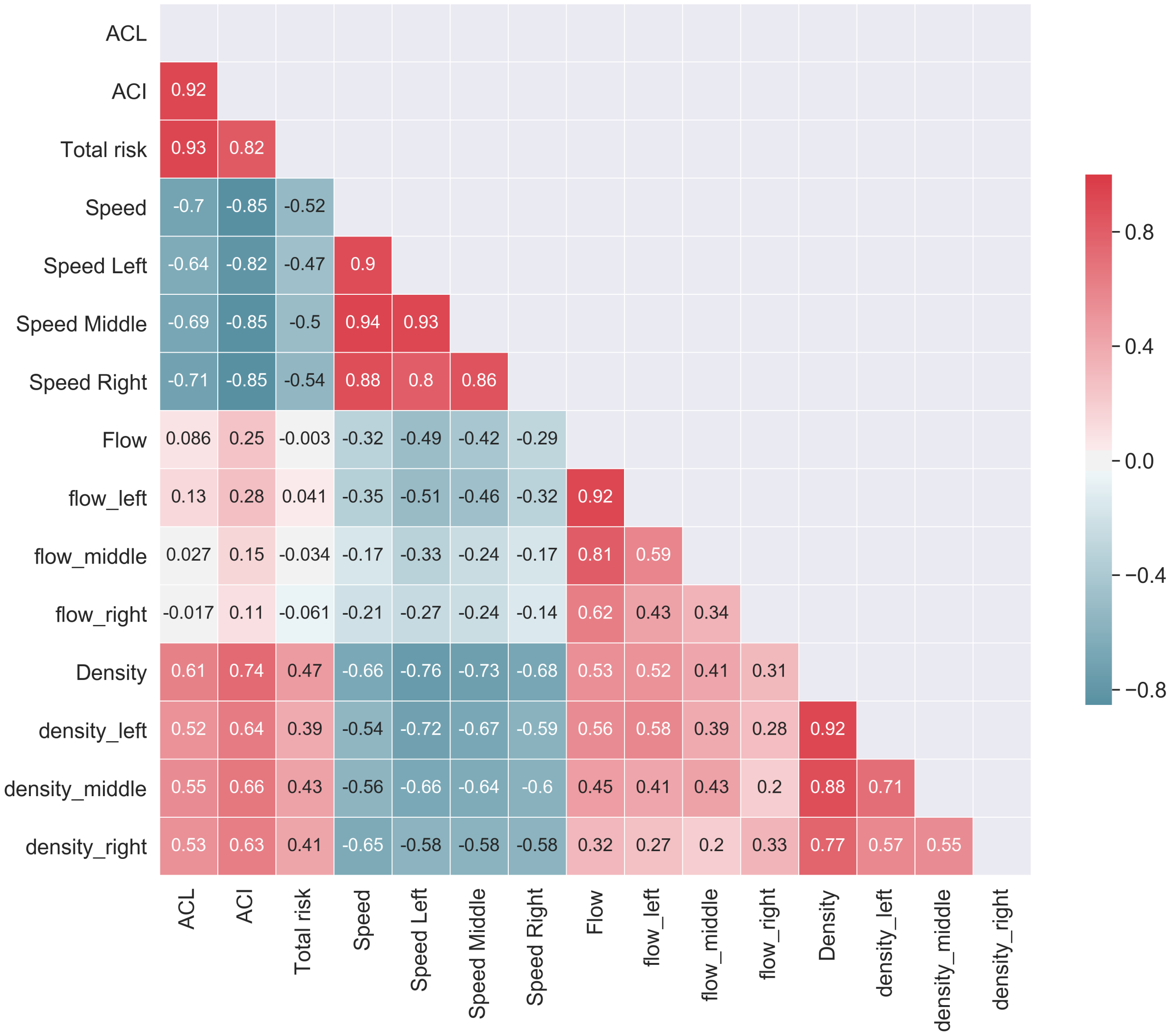

Figure A.1.: Correlation plots of the risk indicators with lane-wise traffic parameters

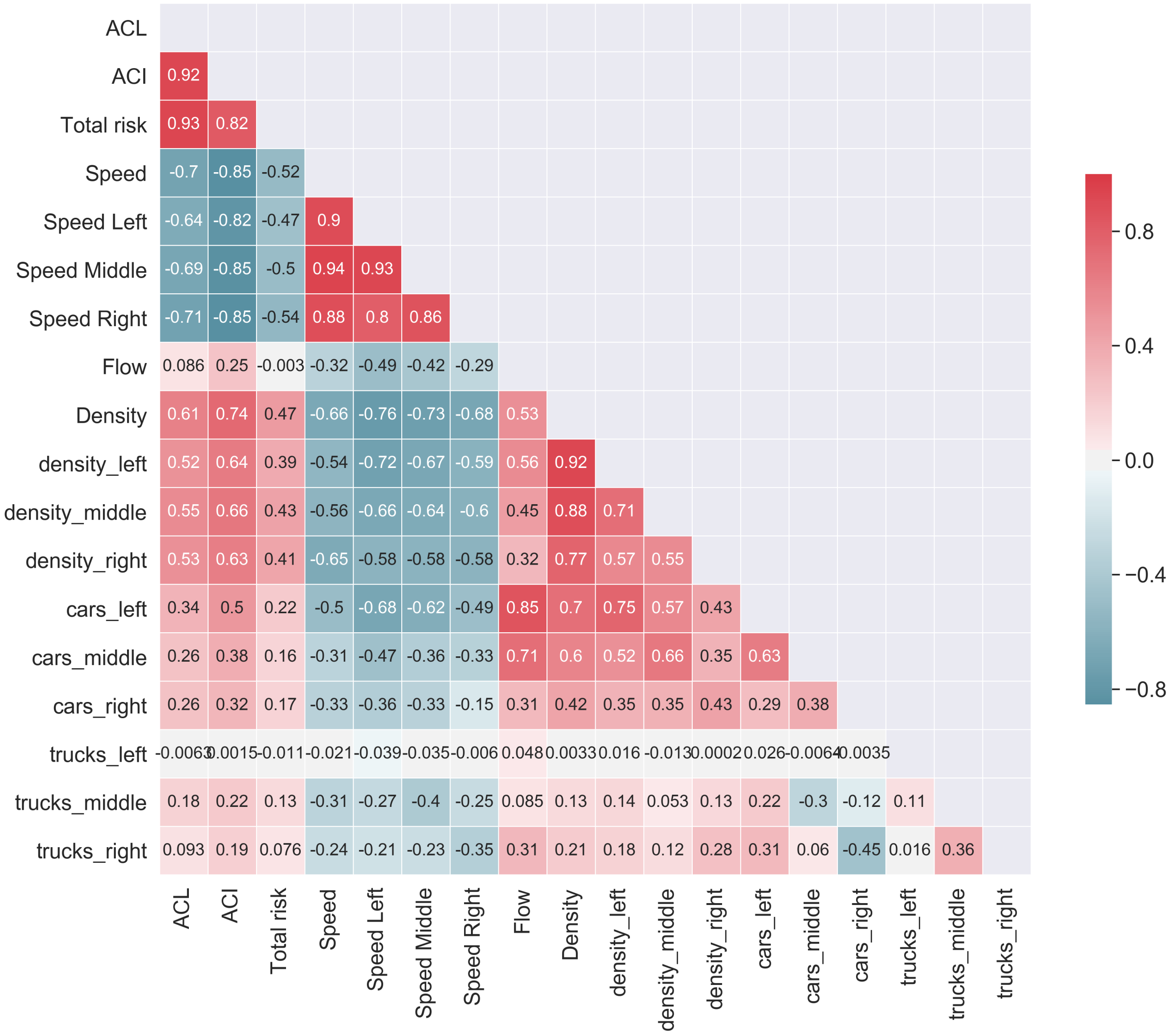

Figure A.2.: Correlation plots of the risk indicators with lane-wise cars and trucks

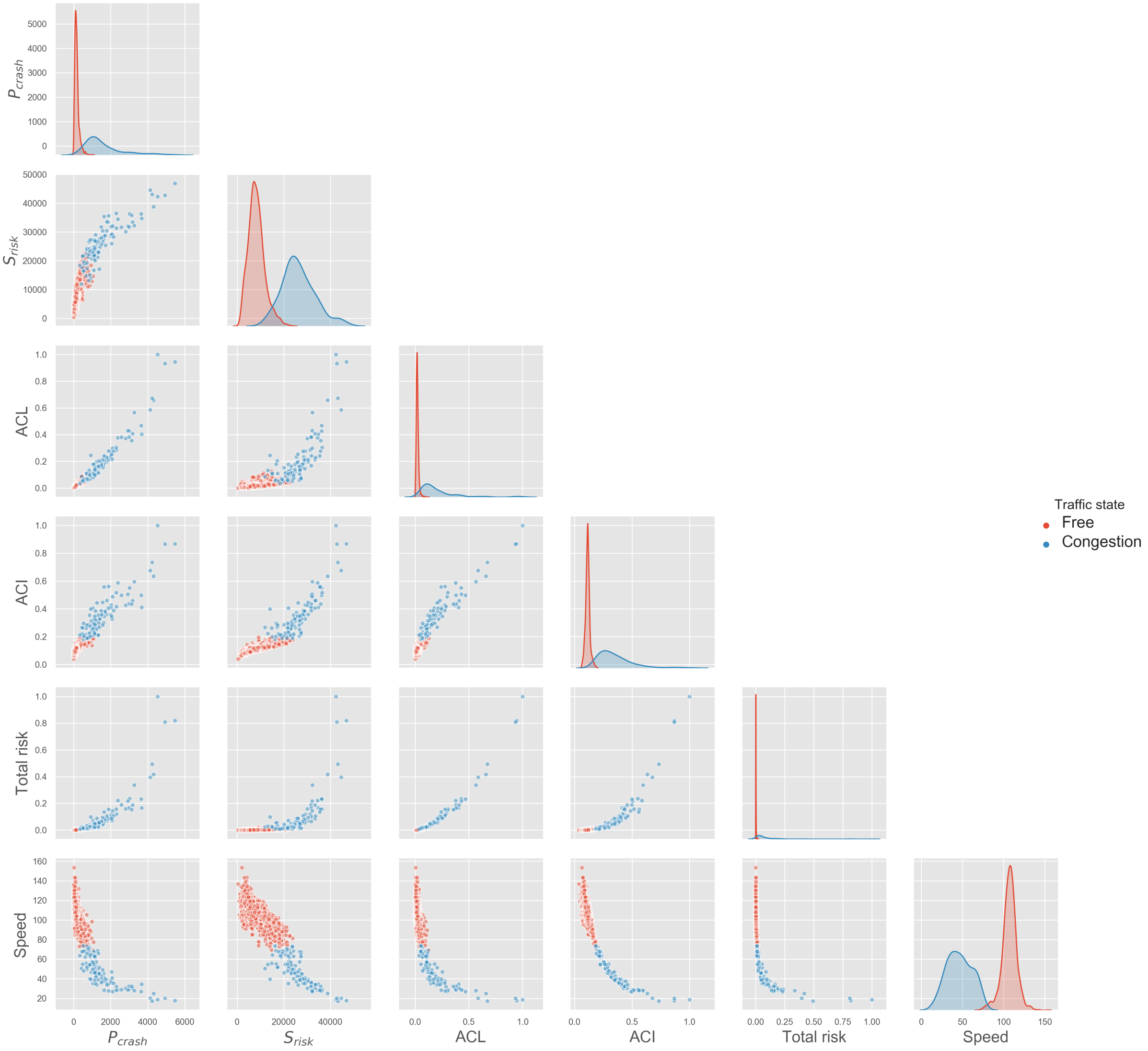

Figure A.3.: Pair plots of the likelihood, Severity and risk of a crash. The ACI, ACL and total risk values are divided by their maximum.

A.2. Contingency tables for Hypothesis testing

Table A.1.: Likelihood outcome with traffic congestion

Likelihood outcome	Congestion	No Congestion
High	101	147
Low	1	2811

Table A.2.: Severity outcome with traffic congestion

Severity outcome	Congestion	No Congestion
High	98	0
Low	4	2958

Table A.3.: Likelihood outcome with Driving Intention

Likelihood outcome	Lane-Change	No Lane-Change
High	1373	191
Low	1581	2869

Table A.4.: Severity outcome with Driving Intention

Severity outcome	Lane-Change	No Lane-Change
High	119	99
Low	2835	2961

Table A.5.: Likelihood outcome with Driving Intention in Free Traffic

Likelihood outcome	Lane-Change	No Lane-Change
High	1281	90
Low	1577	2868

Table A.6.: Severity outcome with Driving Intention in Free Traffic

Severity outcome	Lane-Change	No Lane-Change
High	41	0
Low	2817	2958

Table A.7.: Likelihood outcome with Driving Intention in Congestion

Likelihood outcome	Lane-Change	No Lane-Change
High	92	101
Low	4	1

Table A.8.: Severity outcome with Driving Intention in Congestion

Severity outcome	Lane-Change	No Lane-Change
High	78	99
Low	18	3

Table A.9.: Likelihood outcome with Traffic for Lane changing

Likelihood outcome	Congestion	No Congestion
High	92	1281
Low	4	1577

Table A.10.: Severity outcome with Traffic for Lane changing

Severity outcome	Congestion	No Congestion
High	78	41
Low	18	2817

Table A.11.: Likelihood outcome with Traffic for No-Lane changing

Likelihood outcome	Congestion	No Congestion
High	101	90
Low	1	2868

Table A.12.: Severity outcome with Traffic for No-Lane changing

Severity outcome	Congestion	No Congestion
High	99	0
Low	3	2958