Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

NJIT New Jersey's Science & Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

SHORT-TERM CRASH RISK PREDICTION CONSIDERING PROACTIVE, REACTIVE, AND DRIVER BEHAVIOR FACTORS

by
Sina Darban Khales

Providing a safe and efficient transportation system is the primary goal of transportation engineering and planning. Highway crashes are among the most significant challenges to achieving this goal. They result in significant societal toll reflected in numerous fatalities, personal injuries, property damage, and traffic congestion. To that end, much attention has been given to predictive models of crash occurrence and severity. Most of these models are reactive: they use the data about crashes that have occurred in the past to identify the significant crash factors, crash hot-spots and crash-prone roadway locations, analyze and select the most effective countermeasures for reducing the number and severity of crashes. More recently, the advancements have been made in developing proactive crash risk models to assess short-term crash risks in near-real time. Such models could be applied as part of traffic management strategies to prevent and mitigate the crashes. The driver behavior is found to be the leading cause of highway crashes. Nevertheless, due to data unavailability, limited studies have explored and quantified the role of driver behavior in crashes. The Strategic Highway Research Program Naturalistic Driving Study (SHRP 2 NDS) offers an unprecedented opportunity to perform an in-depth analysis of the impacts of driver behavior on crashes events.

The research presented in this dissertation is divided into three parts, corresponding to the research objectives. The first part investigates the application of advanced data modeling methods for proactive crash risk analysis. Several proactive models for segment level crash risk and severity assessment are developed and tested, considering the proactive data available to most transportation agencies in real time at a regional network scale. The data include roadway geometry characteristics, traffic flow characteristics, and weather condition data. The analysis methods include Random-effect Bayesian Logistics Regression, Random Forest, Gradient Boosting Machine, K-Nearest Neighbor, Gaussian

Naïve Bayes (GNB), and Multi-layer Feedforward Deep Neural Network (MLFDNN). The random oversampling technique is applied to deal with the problem of data imbalance associated with the injury severity analysis. The model training and testing are completed using a dataset containing records of 10,155 crashes that occurred on two interstate highways in New Jersey over a period of two years. The second part of the study analyzes the potential improvement in the prediction abilities of the proposed models by adding reactive data (such as vehicle characteristics and driver characteristics) to the analysis. Commonly, the reactive data is only available (known) after the crash occurs. In the proposed research, the crash analysis is performed by classifying crashes in multiple groupings (instead of a single group), constructed based on the age of drivers and vehicles to account for the impact of reactive data on driver injury severity outcomes. The results of the second part of the study show that while the simultaneous use of reactive and proactive data can improve the prediction performance of the models, the absolute crash probability values must be further improved for operational crash risk prediction. To this end, in the third part of the study, the Naturalistic Driving Study data is used to calibrate the crash risk models, including the driver behavior risk factors. The findings show significant improvement in crash prediction accuracy with the inclusion of driver behavior risk factors, which confirms the driver behavior to be the most critical risk factor affecting the crash likelihood and the associated injury severity.

SHORT-TERM CRASH RISK PREDICTION CONSIDERING PROACTIVE, REACTIVE, AND DRIVER BEHAVIOR FACTORS

by
Sina Darban Khales

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Transportation

John A. Reif Jr. Department of Civil and Environmental Engineering

August 2021

Copyright © 2021 by Sina Darban Khales

ALL RIGHTS RESERVED

APPROVAL PAGE

SHORT-TERM CRASH RISK PREDICTION CONSIDERING PROACTIVE, REACTIVE, AND DRIVER BEHAVIOR FACTORS

Sina Darban Khales

Dr. Branislav Dimitrijevic, Dissertation Advisor — Date
Assistant Professor of Civil and Environmental Engineering, NJIT

Dr. Lazar N. Spasovic, Committee Member — Date
Professor of Civil and Environmental Engineering, NJIT

Dr. Steven I-Jy Chien, Committee Member — Date
Professor of Civil and Environmental Engineering, NJIT

Dr. Janice R. Daniel, Committee Member — Date
Professor of Civil and Environmental Engineering, NJIT

Dr. Taha F. Marhaba, Committee Member — Date
Professor of Civil and Environmental Engineering, NJIT

Dr. Joyoung Lee, Committee Member — Date
Associate Professor of Civil and Environmental Engineering, NJIT

BIOGRAPHICAL SKETCH

Author: Sina Darban Khales

Degree: Doctor of Philosophy

Date: August 2021

Undergraduate and Graduate Education:

- Doctor of Philosophy in Transportation,
  New Jersey Institute of Technology, Newark, NJ, 2021

- Master of Science in Transportation Engineering,
  Eastern Mediterranean University, Famagusta, Cyprus, 2013

- Bachelor of Science in Geodesy and Geomatics,
  Khaje Nasir University of Technology, Tehran, Iran, 2011

Major: Transportation

Presentations and Publications:

Khales, S. D, Kunt, M. M., & Dimitrijevic, B. (2019). Analysis of the impacts of risk factors on teenage and older driver injury severity using random-parameter ordered probit. *Canadian Journal of Civil Engineering*, *47*(11), 1249-1257.

Khales, S. D., Chien, S. I., Lee, J., & Dimitrijevic, B. (2020). Analysis of the effects of visibility and warning devices on driver injury severity at highway-rail grade crossings considering temporal transferability of data. *International Journal of Injury Control and Safety Promotion*, *27*(2), 243-252.

Khales, S. D. (2019). Planning factors affecting metropolitan traffic improvement (case study: metropolis of Tabriz). *Gênero & Direito*, *8*(2), 108-125.

Dimitrijevic, B., Khales, S. D., Asadi, R., Lee, J., Kim, K., & Weiss, J. (2020). *Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling* (Technical Paper. Report No. CAIT-UTC-NC62). Rutgers CAIT.

Dimitrijevic, B., Khales, S. D, Chien, S. I., & Spasovic, L. N. (2021). Regional Road-Level RST Estimation Model Using RWIS Dataset: A Comparison of Ordinary Kriging, Empirical Bayesian Kriging and Regression Kriging. In *100th Transportation Research Board Annual Meeting*, Washington, DC.

Dimitrijevic, B., Khales, S. D, Asadi R., Lee, J., Kim, K. (2021). Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling. In *100th Transportation Research Board Annual Meeting*, Washington, DC.

Khales, S. D, Dimitrijevic, B. (2021). Short Term Segment-Level Crash Risk Prediction Using Advanced Data Modeling with Proactive and Reactive Crash Data. *Applied Science* (2021) (Under Review).

Mosaberpanah, M. A., & Khales, S. D. (2013). The role of transportation in sustainable development. In *ICSDEC 2012: Developing the Frontier of Sustainable Design, Engineering, and Construction* (pp. 441-448) ASCE, Fort Worth, Texas.

I dedicate my dissertation to my wife, Saeedeh, and my family. They endured with support and faith in my academic endeavor, every step of the way. This dissertation is their success as much as it is mine.

ACKNOWLEDGMENT

I would like to express my deepest gratitude and appreciation to my Dissertation Advisor, Professor Branislav Dimitrijevic, for sharing with me his vast knowledge and experience. I am also grateful to him for giving me so many exciting opportunities over the years to conduct innovative research, for helping me pursue my academic interests, and instilling in me the passion for research, scholarship, and teaching. Over the years, he encouraged me to explore diverse scientific disciplines, broaden my knowledge and skills, and be creative in my research pursuits. I am sure this dissertation would never have been finished if it wasn't for his persistent support, guidance, and faith in me.

I owe gratitude and thanks to my Dissertation Committee members as well, including Dr. Lazar Spasovic, Dr. Steven Chien, Dr. Janice Daniel, Dr. Taha Marhaba, and Dr. Joyoung Lee. Their constructive critique, insightful suggestions and guidance helped me immensely in improving my dissertation, as well as presenting the outcomes of my dissertation research.

Finally, I would also like to thank the New Jersey Department of Transportation Intelligent Transportation System (ITS) Resource Center at NJIT for the financial support and allowing me to participate in several projects. The skills and experience I gained through those projects were of irreplaceable importance for me.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF ACRONYMS

NHTSA	National Highway Traffic Safety Administration
FHWA	Federal Highway Administration
PDO	Property Damage Only
AVI	Automatic Vehicle Identification
UBI	Usage-Based Insurance
SWITRS	Statewide Integrated Traffic Records System
CALTRANS	California Department of Transportation
SV	Single Vehicle
MV	Multi Vehicle
GIS	Geographic Information System
PeMS	Highway Performance Measurement System
NCDC	National Climate Data Center
CART	Classification and Regression Tree
OP	Ordered Probit
MNL	Multinomial Logit
CDOT	Colorado Department of Transportation
OCEA	Orange County Expressway Authority
BP	Binary Probit
RF	Random Forest
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic

BIC	Bayesian Information Criterion
ML	Machine Learning
DL	Deep Learning
NNC	Nearest Neighbor classification
DT	Decision Tree
KNN	K-Nearest Neighbor
NMVCCS	National Motor Vehicle Crash Causation Survey
CNC	Crash/Near-Crash
NDS	Naturalistic Driving Study
XGB	Extreme Gradient Boosting
GBM	Gradient Boosting Machine
BA	Bagging Average
DLNN	Deep-Learning Neural Network
MFNN	Multilayer Feedforward Neural Network
t-SNE	t-Distributed Stochastic Neighbor Embedding
WZ	Work Zone
NWZ	Non-Work Zone
GNB	Gaussian Naïve Bayes
OOB	Out-of-Bag
MDA	Mean Decrease Accuracy
TP	True Positive
TN	True Negative
FP	False Positive

FN	False Negative
NJDOT	New Jersey Department of Transportation
NJCMS	New Jersey Congestion Management System
TMC	Traffic Management Channel
NWS	National Weather Service
LCD	Local Climatological Data
NOAA	National Oceanic and Atmospheric Administration
ROSE	Random oversampling examples
SMOTE	Synthetic Minority Oversampling Technique
NAS	National Academy of Science
RADAR	Radio Detection and Ranging
GPS	Global Positioning System
VTTI	Virginia Tech Transportation Institute
MCMC	Markov Chain Monte Carlo
DIC	Deviance Information Criteria
DAS	Data Acquisition Systems
BAC	Blood Alcohol Concentration
RBF	Radial Basic Function

LIST OF SYMBOLS

u_j	Random effect variables
K_{H_j}	Kernel function
H_j	Smoothing matrix
h_{glare}	Horizontal angle between the Sun and the vehicle
v_{glare}	Vertical angle between the Sun and the vehicle
$\emptyset$	Azimuth angle of the Sun
θ'	Slope angle of driveway
$mtry$	Number of factors randomly sampled at each split

CHAPTER 1

INTRODUCTION

1.1 Research Problem Background

The primary goal and purpose of highway transportation agencies is to provide a safe and efficient highway transportation system. Highway crashes are the most significant challenge to this goal. They result in significant societal toll reflected in numerous fatalities, personal injuries, and property damage. According to the National Highway Traffic Safety Administration, over 6.78 million people were involved in reported highway crashes in the United States in 2019. Among these, there were 36,096 fatalities, and over 2.74 million people were injured, some sustaining incapacitating injuries (National Center for Statistics and Analysis, 2020). Highway crashes are also a major cause of traffic congestion, accounting for about 25% of non-recurring delays. More severe crashes, especially those occurring during peak commuting hours or in adverse weather conditions, may result in prolonged roadway closures and excessive traffic backups, thus affecting the ability of highway operating agencies to efficiently respond to and manage the clearance of crashes. Ability to predict when and where the crashes would occur (or are likely to occur) would enable the highway authorities to implement proactive traffic management strategies that anticipate and preempt incidents, rather than react to them.

Besides the crash occurrence location and time, understanding the anticipated severity of crashes beforehand can also be beneficial. The traffic impact and disruption resulting from a crash are directly proportional to the crash severity, and so is the associated road-user and societal cost. A recent crash costs analysis published by the Federal Highway

Administration (FHWA) (Harmon, Bahar, & Gross, 2018) found that crash costs vary mainly based on their severity. According to the report, the recommended national comprehensive crash unit cost to be used in the FHWA benefit cost analysis is $11,900 for property damage only (PDO) crashes, $11,295,400 for fatal crashes, and it ranges between $125,600 and $655,000 for injury crashes, in 2016 dollars. Having an accurate injury severity prediction can help the hospitals and emergency health care providers to prepare adequate medical care resources and supplies in advance. The insurance companies also have an interest in accurate prediction of crash frequency and severity to properly assess their cost and be able to translate the cost of crashes to insurance premiums.

The common interest among all these stakeholders is to improve highway safety and reduce the frequency and severity of crashes. The key to reducing frequency and severity of highway crashes is in better understanding of how, why, when, and where the highway crashes occur. With this knowledge, one can ascertain the necessary actions and strategies for reducing the probability of crash occurrence and reducing their severity. The problem of highway crash mitigation has been a subject of numerous research studies resulting in a variety of crash risk assessment and cash prediction models. Most of these efforts and models are taking a reactive approach to the crash analysis: they analyze the data about crashes that have already occurred and were reported with a sufficient level of detail. The main goal of such analysis is to identify the significant crash factors, identify crash hot-spots or crash-prone roadway locations, and to evaluate and select the most effective countermeasures for reducing the frequency and severity of crashes.

More recently, there has been a great level of interest in proactive crash modeling, which utilizes the data collected in near-real time to assess the short-term crash risks and

severity risk of crashes. The goal of these models is to identify the roadways with higher crash risk and assist in selecting the traffic management strategies to prevent the occurrence of highway crashes and mitigate their negative effects on the overall traffic safety and mobility. The analytics resulting from such models can help the highway agencies to strategically plan the deployment of assets dedicated to traffic incident management and take preemptive traffic management actions targeting the locations with elevated crash risk. The underlying assumption of these models is that the roadway geometry, real-time traffic, environmental and weather conditions can characterize crash risks at any roadway segment over time. These models are focused on identifying crash precursors that are likely to lead to crash occurrence in dynamic traffic environment using high-resolution traffic data (such as traffic volume, speed, and density data for 5–10 min intervals), real-time and forecasted weather data, and roadway geometry data. The methods and techniques employed in analyzing dynamic crash risk include regression analysis models, Bayesian network models, data envelop analysis, and more recently the machine learning modeling approaches, such as supervised and deep learning modeling.

Different modeling techniques have different advantages and shortcomings. One common shortcoming of the proactive crash modeling analysis is the lack of consideration of driver-specific and vehicle-specific characteristics, which have been shown to be among the most significant crash factors (Darban Khales, Kunt, & Dimitrijevic; Guo et al., 2017). This research will address these shortcomings by analyzing the potential improvement in the prediction abilities of the proposed models by the simultaneous use of proactive and reactive data. Similar shortcoming is associated with the reactive crash injury severity studies, where the proactive traffic-related parameters were mostly neglected, and the

models were developed based on the police-reported crash databases, which do not reflect the traffic condition at the time of the crash. Having an understanding of the impact of proactive traffic conditions on the severity of crashes can help the decision makers to design more effective countermeasures to reduce the severity of crashes.

This research also takes into account advantages and disadvantages of various data modeling methods. Based on the performance of different statistical and machine learning models applied in the dissertation research, it is possible to identify the models that yield the best results in terms of predictive power. This research presented in this dissertation also includes an application of an effective sampling methodology to deal with the data imbalance problem associated with the crash likelihood and crash injury severity analysis.

1.2 Research Problem Statement

Despite the recent advancement in operational (near-real-time) analysis of highway crash risks, there are serious shortcomings pertaining to the previous studies:

1. All the existing studies dealing with the real-time crash risk prediction are based on the real-time traffic counts and density collected from Automatic Vehicle Identification (AVI) and real-time weather data collected from nearby weather stations. However, this kind of data is mostly available for a relatively small, well-instrumented roadways, without a coverage of a larger regional scope. Application of models on a limited local scale where such data is available, even if they were highly accurate, would present a challenge in making regional operations decisions and achieving the main objective of dynamically monitoring the crash risk (in terms of crash likelihood and severity) at a network level.

2. All the existing proactive crash risk assessment strategies ignore the impact of reactive data, such as driver characteristics and vehicle characteristics, in predicting the injury severity associated with crashes. This identifies a need for a model that would consider the importance of incorporating the reactive data, as well as the proactive data, in the analysis.

3. Considering the wide range of different methodological approaches in the field of crash risk analysis, there is a need for comparing different models to identify those that demonstrate the best performance in terms of accuracy and reliability of the cash risk and severity prediction.

4. Considering the problem of low frequency of killed/injured cases in the crash severity models, it is necessary to apply an appropriate method for dealing with the highly imbalanced datasets.

The research questions that constitute the problem statement for this dissertation are the following:

- What modeling framework should be applied that would enable a dynamic categorization of roadway segments in a network based on their associated crash risk, considering both crash reactive and crash proactive data?

- Which modelling approach yields the best result for the short-term crash risk prediction, considering the available data, including both proactive and reactive data?

- What data processing steps must be taken to prepare the data for crash risk analysis, including the appropriate methods for dealing with the imbalanced input data?

- How can the proposed modeling approach be implemented for traffic management and operations purposes?

The impetus and motivation for the proposed research is the interest in developing and evaluating effectiveness of a crash risk prediction model for a regional highway network, which would quantify the crash risks at a highway segment level. Such a model would be useful to regional and State transportation agencies by providing intelligence for a proactive decision making related to traffic incident management and law enforcement, especially at the outset of specific conditions with adverse effects on highway traffic safety, such as adverse weather conditions during peak commute hours.

1.3 Dissertation Research Objectives and Scope

The dissertation research has three specific objectives:

1. Develop a framework for segment-level crash risk assessment using proactive data, available for the real-time crash risk analysis. The parameters considered in the analysis include roadway geometry characteristics and dynamic parameters affecting the crash risk, including temporal characteristic (e.g., season, day of the week, time of day), traffic flow characteristics (e.g., vehicle volume, average speed, deviation of speed from speed limit), and weather conditions (e.g., precipitation and visibility).

2. Develop a modeling framework and evaluate the effectiveness of simultaneous use of proactive and reactive data (such as driver and vehicle characteristics) in predicting the crash risk and injury severity. Assess any improvements achieved due to inclusion of the reactive data in the crash prediction models.

3. In developing the modeling frameworks for crash likelihood and crash severity prediction, evaluate the statistical and machine learning modeling methods and select the one or a combination of methods that yield the best performance in terms of accuracy of prediction. In doing so, select the most effective sampling methodology that minimizes the effects of data imbalance associated with the crash likelihood and crash severity analysis.

In developing the modeling framework, the historical crash data from selected roadways in in a regional highway network were analyzed to identify important patterns and statistical significance of various contributing factors. The data considered in this part of the analysis was limited to information currently available to transportation agencies in real time, at the roadway segment level, and with network-wide coverage for major regional roadways. This was done purposely, aiming to include the data that could be used for dynamic short-term crash prediction on a regional scale. The proposed modeling framework was implemented and tested using the dataset for selected regional roadways in the State of New Jersey. Different modeling methods (or techniques) were considered and evaluated, aiming to select the one or a combination of techniques that yield the best crash risk assessment results. Ultimately, the aim of the study is to utilize the findings in advancing the development of analytical models and tools to predict relative crash risk and

their severity for a given roadway segment under the given traffic and weather conditions or provide a ranking of roadway segments based on their relative crash risk under a given set of conditions. The crash risk ranking, or other safety performance measures, could then be used to select and prioritize crash and crash-related congestion mitigation strategies and actions by the highway operations agencies.

This study also demonstrated the effectiveness of simultaneous use of reactive and proactive data in predicting the crash injury severity in conjunction with crash risk analysis. This has not been reported in the literature so far (to the best of the author's knowledge). To this end, the utility of different analytical models was investigated, comparing the model results under two different conditions: 1) using proactive data only; and 2) using the combination of reactive and proactive data. The results of this comparison can help to acquire better knowledge of the optimal set of input data for analyzing and predicting the crash frequency and crash injury severity on highways. This will ultimately allow the transportation agencies and decision makers to make more precise and effective decisions, as well as design appropriate countermeasures aimed at reducing the frequency and severity of crashes (Yahaya et al., 2020). The findings of this study can also be used to perform a more accurate crash risk assessment of the roadways at the roadway segment level. The same traffic safety modeling can be used by the insurance companies to improve their Usage-Based Insurance (UBI) system, which uses vehicle telemetry data to determine the risk level of individual drivers.

Both in analyzing crash risk and crash severity, the input data is highly imbalanced, contrasting crashes to non-crash outcomes, or severe crashes (with injuries or fatalities) to property-damage-only (PDO) crashes. In both cases, the crashes and severe crash outcomes

represent rare events, or vastly underrepresented minority classes in the analysis sample. This makes it critical to adopt an effective sampling method to overcome the sample disbalance in predicting these rare events. To overcome the data imbalance problem, this study applied a transformation technique that generates new artificial instances from the original classes to achieve a more balanced dataset. The results of the analysis showed an improvement in the prediction performance of the models for the minority classes after alleviating the data imbalance problem.

1.4 Dissertation Organization

The dissertation is presented in six chapters. The first chapter provides a brief introduction of the research problem background and defines the research problem, research questions, objectives of the dissertation, and it outlines the research scope for addressing the stated research problem. Chapter 2 provides the literature review followed by a summary of literature review findings. Chapter 3 presents the research approach and methodology for implementing advanced data modeling for crash likelihood and severity prediction using both real-time traffic and weather data, and the recorded data from the naturalistic driving study. The case studies that demonstrate implementations of the proposed models are presented in Chapter 4. Chapter 5 discusses the model results and provides a summary of the practical implications of demonstrated models. Lastly, the research conclusions along with a summary of the research contributions and future studies are discussed in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

The most important aspects of analyzing the crash risk are the crash likelihood and crash injury severity. Conceptually, the crash likelihood and severity are influenced by a set of factors related to driver performance, roadway characteristics, vehicle characteristics, and environmental factors. The data describing the crash factors and circumstances are commonly collected and documented by the law enforcement officers after the crash occurrence as part of crash investigation and reporting. Advances in Intelligent Transportation Systems (ITS) and data collection technologies have vastly improved the ability of transportation agencies to collect and analyze traffic and road performance data in real time, such as segment-level travel time, speed, volume, occupancy, and road-weather data. Nevertheless, the challenges in this respect remain as the data collection is often focused on specific sections of major roadways, with a limited coverage of the reginal transportation network. At the same time, numerous studies have been conducted with the goal of analyzing the data collected in real-time to assess the likelihood of crashes and their severity, which provides an excellent basis for development of crash prediction models with a regional scope in mind.

2.1 Crash Likelihood Analysis

The great majority of the previous studies define crash likelihood analysis as a binary classification problem, differentiating between crash and non-crash outcomes. Previous studies applied a variety of statistical models to analyze the crash likelihood , among which

the Bayesian logistic regression (Ahmed, Abdel-Aty, Lee, & Yu, 2014; Ahmed, Abdel-Aty, & Yu, 2012; Ahmed & Abdel-Aty, 2011; Wang, Shi, & Abdel-Aty, 2015) and conditional logistic regression (Kwak & Kho, 2016; Yuan & Abdel-Aty, 2018) are the models most commonly used.

Ahmed and Abdel-Aty (2011) performed a matched case-control binary logistic regression analysis to examine the crash precursors. Two datasets were used in this study: speed data collected by AVI systems and the corresponding crash data from the crash database maintained by the Florida Department of Transportation for year 2008. The parameters considered in the model were average speeds, standard deviations of the speed, and the logarithm of the coefficient of variation in speed, all aggregated into 5-minute intervals. In addition to the crash location segment, the speed data were also obtained for three upstream and three downstream segments closest to the crash segment. The findings of the study showed that the speed parameters obtained from AVI systems within 1.5 mile of the crash location were statistically significant, while the speed parameters obtained from devices that were more than 3 miles far from the crash location were statistically insignificant to model the likelihood of crash.

Xu, Liu, Yang, and Wang (2016) developed a random-effect logit model to predict the secondary crashes on freeways. The study area included a 35-mile section on the I-880 freeway in the State of California. The section is equipped with 134 loop detectors, which provided high resolution traffic data including count, speed, and detector occupancy for each lane every 30 seconds. Crash data was also obtained from the Statewide Integrated Traffic Records System (SWITRS), which is maintained by the California Department of Transportation (Caltrans). The real-time traffic data was further aggregated into 5-minute

intervals and the data for the 5-10 minute prior to the crash was used to represent the traffic condition at the time of the crash occurrence. The reason for selecting this time frame was to account for the potential inaccuracies in the reported crash time. A comparative analysis was performed for the models with and without the traffic variables. Two likelihood ratio tests were conducted to assess the effect of including the traffic variables and the random-effect parameter on the performance of the models. The results showed that the inclusion of both the traffic variables and the random-effect parameter improved the performance of the models. Finally, AUC value was used to evaluate the predictive performance of the models. The model with both the traffic variables and the random-effect parameter provided an AUC value of 0.83, which was 7% higher than the value obtained for the model without the traffic variables and random-effect parameter.

Wang et al. (2015) conducted a study to predict crashes on expressway ramps. Three expressways in Central Florida were included in the study area: SR-408 (14.2 mi), SR-417 (26.9 mi), and SR-528 (7.6 mi). The crash prediction was based on the data recorded from July 2013 through March 2014. The data used in the study included: (1) crash data from the Florida DOT statewide crash database, (2) traffic flow data provided by the Central Florida Expressway Authority, (3) roadway geometry data derived from the roadway Geographic Information System (GIS), and (4) weather data from the National Climate Data Center. To reduce the noise in data, the traffic data was aggregated into 5-minute intervals, and the period of 5-to-10 minutes prior to the time of crash was selected to represent the traffic conditions. Compared with the traffic data within the 5-minutes period before the crash, it was discovered that the period of 5-10 minutes prior to the time of crash provided better model performance and was also sufficient enough to disseminate

warning information to the drivers. The non-crash cases for the model calibration were generated by randomly selecting 0.05% of the 11,270,808 5-minute intervals (12 intervals per hour * 24 hours * 276 days * 141 ramps). The final dataset was further divided into two parts based on the crash type (single vehicle vs. multi-vehicle). The dataset for each crash type was also split into training and validation datasets with a ratio of 70:30. The Pearson correlation test was performed before the model development to detect potential correlations between the explanatory variables. The Bayesian logistic regression was used to establish the prediction models for a single vehicle (SV) and multi-vehicle (MV) crashes. Five variables were found to be significant in the SV crash prediction model: logarithm of the vehicle count in 5-minute intervals, speed, ramp configuration, road surface condition, and visibility. The AUC for the training and validation were also found to be 0.9346 and 0.9710, respectively. In addition, the overall accuracy was 0.89 for the training set and 0.904 for the validation set. All the significant variables in the SV model, except the speed, were found to be significant in the MV model as well. The AUCs for the training and validation were 0.7644 and 0.76, respectively, and the overall accuracy was 0.643 for the training set and 0.764 for the validation set.

Xu, Tarko, Wang, and Liu (2013) developed a model to predict the crash likelihood at three different severity levels. The study area covered a 29-mile segment on the I-88 freeway in San Francisco. The model inputs included 22 traffic flow variables derived from the vehicle count, occupancy, and speed data collected 30-second intervals from the adjacent data collection stations upstream and downstream from of the crash site. The data was obtained from the Highway Performance Measurement System (PeMS). The traffic data was aggregated into 5-minute intervals. The data for the period 5-10 minutes prior to

crash at the upstream and downstream detectors was used as representative of the traffic condition at the time of crash. In addition, the data for nine roadway-geometry variables were also included in the dataset, such as width of the roadway, number of lanes, and the roadway type. The weather condition data (clear vs. adverse) was obtained from the National Climate Data Center (NCDC). For each crash case, 20 non-crash cases were randomly selected. Traffic data, roadway geometry data, and weather data were assigned to all crash cases and non-crash cases in model development. A three-stage sequential binary logit model was used to assess the likelihood of crashes at each severity level. The 20-fold cross-validation was also performed to evaluate the model's performance. The findings of the study showed that the traffic flow characteristics contributing to crash likelihood were substantially different at each severity level.

Yu and Abdel-Aty (2013) studied the real-time crash risk by analyzing a 15-mile mountainous freeway section of I-70 in Colorado. The datasets used in the study included: (1) crash data provided by Colorado DOT, and (2) real-time traffic data collected from the Remote Traffic Microwave Sensor (RTMS) radars. The RTMS data included speed, volume, and occupancy recorded in 30-second intervals. This data was further aggregated into 5-minute intervals and assigned to each crash from the nearest downstream detector. Similar to the Xu et al. (2013), the data aggregated for the period 5-10 minutes prior to the time of crash was selected to represent the traffic condition at the time of the crash. For each crash case, the average and standard deviation of the upstream and downstream speeds, traffic volume, and occupancy were calculated. This makes for the total number of 18 traffic-related explanatory variables associated with each observation. Furthermore, for each crash case, four non-crash cases were identified and matched for the same location,

day of the week, and time of day, two weeks before and two weeks after the crash occurrence. For the modeling part, firstly, a classification and regression tree (CART) was developed to estimate the significant variables to be used as inputs for the crash likelihood models. The selected variables included: downstream average speed, crash location average speed, crash location standard deviation of occupancy, and crash location standard deviation of volume. The correlation matrix was also calculated to find potential correlations between the identified variables. The dataset was then split into a training set (70%), and three testing sets with varying sample sizes (30%, 20%, and 10%). Three Bayesian logistic regression models were applied using the training set: (1) Bayesian fixed-parameter logistic regression, (2) Bayesian random-parameter logistic regression considering seasonal variation, and (3) Bayesian random-effect logistic regression considering the segment level heterogeneity. Comparing the DIC values for the three models demonstrated that the Bayesian fixed-parameter model had better performance than the other two models. Next, two SVM models, one with linear kernel and one with RBF kernel, were employed and tested using different testing sets. The results were compared to the results produced by the Bayesian logistic regression, using the Area under the ROC curve (AUC). The findings of the study showed that the SVM with RBF kernel models was superior, and therefore, concluded that some non-linear relationships existed between the dependent variable and independent variables in the real-time crash risk model.

2.2 Crash Severity Analysis

Similar to crash-likelihood studies, much research has been done analyzing the injury severity of crashes. The most frequently adopted approach to exploring the crash injury

severity employed discrete choice modeling. Discrete choice statistical models can be classified into fixed-parameter models and random-parameter models. Despite the extensive use of fixed-parameter models such as ordered probit (OP) and multinomial logit (MNL) in the past, random-parameter models started gaining more traction mostly due to their capability to account for the unobserved heterogeneity among observations (Milton, Shankar, & Mannering, 2008). A comparative study by Darban Khales, Kunt, and Dimitrijevic (2019) showed that random-parameter ordered probit model outperformed the fixed-parameter ordered probit model in studying the driver injury severity for teenage and older drivers. Similar conclusion was reached in the findings of Kim, Ulfarsson, Kim, and Shankar (2013) which showed that the mixed logit model outperformed the multinomial logit model. Haleem and Gan (2013) also concluded that mixed logit model is superior to the standard logit model in terms of the parameters' interpretation and goodness of fit by studying the effect of driver's age and side of impact on crash injury severity on urban freeways.

With the help of the recent technological advancements, studying the real-time crash injury severity has also gained lots of attention among researchers. The normal logit models and Bayesian logit models have been among the most popular modelling techniques for analyzing the severity of crashes in (near)real-time. Yu and Abdel-Aty (2014b) applied Bayesian models to classify and compare the non-severe crashes and severe crashes on two high-speed facilities: I-70 freeway in Colorado and State Road 408 (SR-408) in Orlando, Florida. Four datasets were utilized to study the severity of crashes on I-70: (1) crash data for I-70 provided by Colorado DOT, (2) roadway segment geometry data from the roadway characteristics inventory, (3) real-time weather data from six

weather stations located along the study area, and (4) real-time traffic data collected by the automated vehicle identification (AVI) detectors. The real-time traffic data was aggregated into 6-minute intervals. The mean, standard deviation, and coefficient of variation of the speeds for 6-12 minutes prior to each crash were calculated to represent the traffic conditions before the crashes happened. The visibility conditions recorded at the closest weather station prior to the time of crash was also assigned to each crash to represent the weather conditions at the time crash. Other explanatory variables in the I-70 model included: two binary indicator variables (snow season vs. dry season and longitudinal grade $\geq$ 4% vs. longitudinal grade < 4%), one real-time traffic variable (standard deviation of speed), and two joint variables (visibility * snow season and visibility * dry season). To analyze the severity of crashes on SR-408, crash data from the crash analysis reporting (CAR) system, and real-time AVI data from the Orange County Expressway Authority (OOCEA) were used. The same approach as in the I-70 model was implemented to aggregate and assign the traffic data for each crash. Three binary indicator variables (passenger car vs. non-passenger car, daytime vs. nighttime, and whether the impact point is the driver side), one roadway geometry variable (shoulder width), and one real-time traffic variable (standard deviation of speed) were defined as the explanatory variables in the crash severity models for SR-408. Four different models were used to analyze the crash injury severities for the two studied roadways: regular binary probit (BP) with maximum likelihood estimation, Bayesian BP, segment level random-effect hierarchical Bayesian BP, and crash-level random-effect Bayesian BP. The Bayesian models were compared based on the deviance information criterion (DIC). First, the results of the BP model were compared to the Bayesian BP, showing that for both roadways the Bayesian BP model

outperformed the regular BP model in terms of the number of significant variables. Second, the Bayesian BP model was compared to the segment level random-effect hierarchical Bayesian BP model: the lower value of DIC in random-effect Bayesian models indicated that they were superior as they better accounted for the unobserved heterogeneity in the data that was not captured in the Bayesian BP model. Finally, the comparison between the two hierarchical Bayesian BP models showed that the model performance can be improved by the crash level random effect model as it allowed for a more flexible error term.

In another study, Yu and Abdel-Aty (2014a) used similar data sources to develop crash injury severity models for the I-70 freeway. First, Random Forest (RF) algorithm was used to rank the variables by significance: the steep grade indicator, speed standard deviation, temperature, and snow season indicators were found to be the most important factors. Then, a Bayesian fixed-parameter binary logit model was developed to model the injury severity (severe vs. non-severe). The results of the model showed that the temperature was not statistically significant. To account for the potential non-linearity between the injury severity levels and independent explanatory variables, a Support Vector Machine (SVM) model with radial basic function (RBF) kernel was performed. The effect of the explanatory variables was also quantified through the sensitivity analyses. Next, a random parameter logit model with an unrestricted variance-covariance matrix was used to model the injury severities by considering the unobserved heterogeneities and correlation between the input variables. Finally, the three models were compared based on the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC values). The results indicated that the SVM model and the Bayesian logit model with random parameters provided better results than the binary logit model with fixed parameters.

2.3 Combined Studies

Several recent studies presented the models combining the prediction of both crash likelihood and their associated injury severity based on real-time data inputs. In one such study, Theofilatos (2017) investigated crash likelihood and severity by incorporating real-time traffic and weather data for urban arterials in Athens, Greece. To build the dataset, traffic data from the nearest upstream loop detector and weather data from the closest weather station were matched to each archived crash. The traffic and weather data were aggregated into 1-hour intervals and used for model training. For every crash case, two non-crash cases were generated for the same location and same time of day one week before and one week after the crash occurrence. The traffic and weather data were assigned to non-crash cases using a similar method as for the crash cases. For the crash likelihood modeling, a random forest (RF) method was used to select the significant variables. Five variables were found to be significant and were included in the final model: 1-hour coefficient of variation of the upstream traffic flow, 1-hour standard deviation of the upstream occupancy, 1-hour standard deviation of the upstream speed, 1-hour coefficient of variation of the upstream speed, and 1-hour coefficient of variation of the upstream occupancy. Next, a correlation matrix was built to assess the correlation between the significant variables to avoid multicollinearity problem. Lastly, a Bayesian logistic regression was used to model the likelihood of crashes. The model outputs showed that the standard deviation of occupancy and the coefficient of variation of traffic flow had the highest impact on the likelihood of crashes. A similar approach was undertaken for the crash severity modeling, where an RF model identified the following significant variables: 1-hour average traffic flow upstream, crash type, 1-hour coefficient of variation of the

upstream traffic flow, 1-hour average upstream speed, as well as 1-hour coefficient of variation of the upstream speed. The correlation matrix was also generated to find the possible correlations between these variables. Two different methods were utilized to model the crash severity: (1) a finite mixture logit (latent class) model, and (2) a mixed effect logit model. The results revealed that the finite mixture model showed a better fit and superior performance as the latent classes are optimally chosen by the model based on the Bayesian Information Criterion (BIC).

2.4 Application of Machine Learning in Analyzing the Real-Time Crash Risk

The interpretation of the outputs in statistical models is straightforward and fairly easy. The statistics associated with coefficients for each explanatory variable quantify the strength and "direction" of the relationship between the explanatory and dependent variables, such as crash likelihood and/or crash injury severity. However, the statistical models also suffer from serious limitations. One, they require assumptions about the distribution of the data. In addition, they assume a linear relationship between the explanatory variables and the dependent variable. To overcome these shortcomings of the statistical models, machine learning (ML) models have been applied as an alternative by the researchers. The ML models do not have any pre-assumptions about the nature of the relationship between the explanatory and dependent variables and are also reported to provide better fitting than the statistical models in traffic crash analysis.

Theofilatos, Chen, and Antoniou (2019) compared the performance of machine learning (ML) and deep learning (DL) methods in predicting crash occurrence. The models were demonstrated using a case study of an urban motorway in Greece (Attica Tollway).

For the analysis purposes, real-time traffic data and weather data were obtained and matched to the crash and non-crash cases. A 1:2 ratio of crash cases to non-crash cases was selected for this study. In addition, the raw data were aggregated to obtain the average, standard deviation, and coefficient of variation of traffic-related parameters. To develop the models, the data was first split into a training set (75%), and a validation set (25%), and various ML methods were employed to predict the crash likelihood using the training set. The ML models considered in the study included: k-nearest neighbor, Naïve Bayes, decision tree (DT), RF, SVM, and shallow neural network. These models were compared based on the performance metrics, including accuracy, sensitivity, specificity, and AUC. An RF model was first applied to identify important variables. Afterward, a binary logistic regression model was generated with the input variables to check and confirm the degree of significance for each of them. The result of the binary logistic model indicated that the standard deviation of speed 0-15 minutes before the crash time, and the total amount of rainfall were the only significant variables, and they were used as inputs for the ML and DL models. The results of the study showed that the DL model outperformed the ML models as it provided a relatively balanced performance among all metrics.

Iranitalab and Khattak (2017) compared the performance of four statistical and ML models in predicting the crash injury severity: MNL, Nearest Neighbor classification (NNC), SVM, and RF. The study found the MNL to have the weakest performance among all models, while the NNC outperformed the other models in terms of overall accuracy. Similar conclusions were reached by Zhang, Li, Pu, and Xu (2018) who compared the ability of two statistical models, OP and MNL, with four popular ML methods: KNN, DT, RF, and SVM, to correctly predict the crash injury severity outcomes. The results of the

study showed that the RF model had the best prediction performance. It was also found that the ML methods had higher accuracy than the statistical methods in general.

2.5 Nature of the Input Data for Crash Analysis

Besides the selection of the most appropriate and the best performing modeling method, other important considerations in developing the real-time crash models are related to the source and temporal aspect of the data used in model calibration. Most of the previous injury severity studies are based on reactive data, i.e., the data related to a crash collected after the crash occurrence (Abdelwahab & Abdel-Aty, 2001; Chen, Zhang, Qian, Tarefder, & Tian, 2016; Chen, Zhang, Yang, & Milton, 2016; Delen, Sharda, & Bessonov, 2006; Iranitalab & Khattak, 2017; Zeng & Huang, 2014). Despite the wide application of reactive data in injury severity analyses, there are serious shortcomings of such approach. Reactive data requires a long period of observation to achieve a reasonable statistical significance (A. Chang, Saunier, & Laureshyn, 2017). Also, in the "reactive" injury severity analyses the injury had already occurred, which makes it difficult for the decision makers to identify and understand the operational (real-time) impact of the contributing factors on the injury severity of crashes, which ultimately could help with preventing (or mitigating) crashes in real-time. Thus, the need emerged for proactive methods for roadway safety analyses that do not rely solely on the data describing the crashes that have already occurred (Saunier & Sayed, 2010).

Unlike the reactive data, the proactive data is collected before the crash occurrence. As mentioned before, until recent years, historical reactive data had been the universal metric for crash severity analyses. Nevertheless, due to the recent technological

advancement, traffic management authorities are becoming more interested in proactive traffic management strategies. These strategies are mainly focused on identifying crash precursors that are likely to lead to crash occurrence in dynamic traffic environment using proactive high-resolution traffic data collected from loop detectors, and weather characteristics collected from weather stations. Using proactive data enables the decision-makers to observe the progression of factors leading to different severity levels in real-time. In that context, a predictive data analysis can help identify the warning signs for the decision makers to first prevent the crashes occurrence, and next, to take proper actions to reduce the severity of potential crashes.

Recently, there has been a handful of studies applying statistical and ML models using proactive data for crash injury severity prediction. Some examples are provided in Section 3.2. Even though these studies have confirmed the important impact of the dynamic traffic and weather variables in predicting the crash injury severity outcomes, there is a common limitation associated with these proactive strategies. Namely, all the existing studies in the real-time crash risk prediction field are based on the real-time traffic counts and density collected from Automatic Vehicle Identification (AVI) and real-time weather data collected from weather stations. However, this kind of data is mostly available on specific, well-equipped roadway segments, without a coverage of a larger regional scope. Application of models on a limited local scale where such data is available, even if they were highly accurate, would present a challenge in making regional operations decisions and achieving the main objective of dynamically monitoring the crash risk (in terms of likelihood and severity). Besides, looking at the findings of the previous studies using proactive data, one cannot ignore the important influence of reactive variables on

predicting the crash injury severity. Therefore, as remarked by Reiman & Pietikäinen (2012), using both reactive and proactive data can be more useful for the organizations and decision makers. This is also confirmed in a study by Sarkar, Pramanik, Maiti, & Reniers (2020), which showed the effectiveness of using a combination of active data and proactive data in predicting the injury severity of accidents in workplaces.

2.6 Driver Behavior and Naturalistic Driving Study

The literature review revealed an important role of a factor which cannot be either grouped as proactive or reactive. Driver behavior, on which the data is not commonly collected either before or after the crash occurrence, has been found to have a key influence on the probability of crash risk. The research by Treat et al. (1979) has indicated that human errors are the main cause of 93% of crashes. The same study found that environmental characteristics and vehicle characteristics account for only 12-34% and 4-13% of crashes, respectively. This was confirmed by another study conducted by the National Motor Vehicle Crash Causation Survey (NMVCCS), which found that the critical reason, which is defined as the last event of the crash causal chain, was assigned to the drivers in about 94% of crashes (Singh, 2018). This number is as small as 2% for both the vehicle component failure and the environmental characteristics (e.g., slick roads, weather, etc.). The same report also found that the driver recognition errors such as inattention, internal and external distractions, and inadequate surveillance were the most frequently assigned driver-related reasons for crash (41%). In addition, it was found that about 33% of crashes with driver-related causation could be attributed to driver decision errors, such as driving too fast given the environmental conditions or roadway geometry (e.g., in sharp horizontal

curves), false assumption of other drivers' actions, or illegal maneuvers and misjudgment of the gap between the vehicles given their speeds.

The majority of the previous studies of short-term crash risk focused on the role of traffic conditions (speed, volume, etc.), roadway geometry characteristics, and environmental characteristics as the contributing or crash risk factors. Thus, more attention needs to be given to consideration of driver-related factors, including driver behavior, which have not been well addressed due to the difficulty of data collection. Since the conventional sources of crash data (e.g., crash reports filed by the law enforcement) do not provide the detailed driver behavior information, alternative data collection techniques have been used to obtain this information. Driving behavior questionnaire, on-road data collections, and driving simulation are the most commonly used methods to this end (Bärgman, 2016). All three methods are relatively inexpensive and require short collection time. However, there are serious limitations associated with each one of them. The validity of driving behavior questionnaires is challenged by many researchers (Agramunt, 2018). The data collected on-road does not provide a rigorous understanding of driver behavior (van Schagen & Sagberg, 2012), and it was found that the actual driver behavior differs from what was observed in a simulated environment (Zöller, Abendroth, & Bruder, 2019)

The recent advancements in information technologies and data collection techniques have enabled driver monitoring in natural driving conditions and recording of the microscopic driver behavior and vehicle performance prior to safety critical events. This new capability provided a great opportunity for traffic safety researchers to perform an in-depth analysis of crash/near-crash (CNC) contributing factors. One example of collecting such data is the naturalistic driving study (NDS), in which the vehicles are

instrumented with on-board data acquisition systems such as cameras, sensors, and radars that automatically and continuously collect driver's and vehicle's parameters. In general, NDS data has made five main improvements over police-reported crash records, which has long been the main source of data in the crash risk studies:

1. NDS data has minimized the human error in data entry and processing, which is a common shortcoming associated with police-reported crash records.

2. NDS data includes information on near-crash events as well as crash events, which helps to alleviate the challenge of imbalance between the non-crash events and comparatively small number of reported crash events.

3. NDS data includes baseline events, which enables the comparison of the influence of different factors associated with crash and non-crash events in the crash likelihood models.

4. NDS data provides detailed information on driver behavior, which is usually lacking in police-reported crash records.

5. Unlike police reported crashes, the detailed data extracted for each event in NDS are video recorded and can be repeatedly verified and analyzed. This makes the NDS data more reliable than police-reported data.

Considering the above, NDS data has been used in a number of recent studies for crash risk analysis. Arvin and Khattak (2020) studied the impact on the CNC probability of the driver distraction caused by performing secondary tasks 0-15 seconds prior to a crash or non-crash event. The study also investigated the association of impaired driving with the CNC risk. The dataset used in the study contained the data from 9,239 trips taken by 1,546 drivers, with 7,396 baseline (non-crash) events, 1,228 near-crashes, and 617 crashes. Four combinations of fixed and random-parameter logistic regression models were developed in the study: cellphone-oriented distraction duration model, object-oriented distraction duration model, activity-oriented distraction duration model, and impaired driving model. Along with the duration of distraction and impaired driving, the traffic

density and vehicular movement were also considered in the model development as control factors. The results of the study showed a significant relationship between the duration of distraction and impaired driving and the increased chance of CNC involvement.

Bakhit, Guo, and Ishak (2018) also studied the risk associated with CNC events when drivers are distracted by a secondary task. To this end, they first showed a significant correlation between the engagement in a secondary task and CNC risk by using a bivariate probit model. Next, they developed two different models to quantify the increased risk associated with each secondary task: a multinomial logit model and an association analysis. The results from both models revealed that reading while driving and reaching for an object are the most significant contributors to a crash/near-crash occurrence.

Wu and Xu (2018) used the NDS data to analyze the impact of familiarity with the traveled road on the secondary task engagement. The data used in the study comprised of 557 trips including 501 trips on familiar roads and 56 trips on unfamiliar roads. All trips were completed by a group of 155 drivers during daytime and under fine weather conditions. The impact of unfamiliarity was explored by comparing the frequency and duration time of different distracted driving activities on familiar and unfamiliar roads. The findings of the study showed the higher chance of involvement in secondary tasks while driving on familiar roads. In addition, duration of distracted driving was also found to be higher on familiar roads compared to unfamiliar roads.

Bharadwaj, Edara, and Sun (2019) developed a logistic regression model to estimate crash risk in work zones based on NDS data. The risk factors investigated in the model encompassed duration of secondary task, driving behavior, traffic density, locality, traffic control, and lighting condition. The logistic regression model was found to have an

acceptable goodness of fit (Chi-squared = 7.69, and p-value = 0.46) and a good predictive performance (AUC = 0.8897). A matched case-control model along with the odd ratios were used to quantify the risk of different factors. According to the model results, the driver behavior is the most critical risk factor in work zone crashes. More specifically, the driver inattention was found to be the most significant among the driver behaviors that increase the risk of crash/near-crash events in work zones.

Mousa, Bakhit, and Ishak (2019) compared the performance power of four machine learning models in predicting the CNC events: extreme gradient boosting (XGB), gradient boosting (GB), bagging average (BANN), and deep-learning neural network (DNN). The study used two distinct sets of data from NDS database: driver characteristics records and event records. The driver characteristics records included the following driver attributes: age, gender, employment status, marital status, years of driving, average annual miles traveled, education level, household income, and State. The event records included both the CNC and baseline events, each with the following attributes: environmental conditions, road conditions, and driving behavior associated with them. In addition, the oversampling technique was used to alleviate the data imbalanced problem. A total of 22 variables was employed to train the models. The results of the study showed that XGB outperformed all other models with a classification accuracy of 84.9%. Also, it was demonstrated that driver behavior and intersection influence had the highest impact on CNC occurrence, accounting for 53.80% and 20.39% of the detection accuracy of the model, respectively.

Y. Chang, Bharadwaj, Edara, and Sun (2020) used NDS data to: (1) evaluate the risk of CNC events by employing a large set of driver characteristics and event characteristics using logistic regression models, and (2) Predict and classify CNC events

using three distinct machine learning methods: RF, DNN, multilayer feedforward neural network (MFNN), and t-distributed stochastic neighbor embedding (t-SNE). The focus of the study was on the CNC events that occurred around work zones. The logistic regression models were developed for both work zone (WZ) and non-work zone (NWZ) datasets to predict CNC against baseline events. The results of the models indicated that driving behavior, secondary task duration, maneuver judgement, and traffic density were the most significant contributing factors of CNC events, both in the WZ and NWZ areas. The calculated odds ratios demonstrated that odds of CNC events for different risk factors followed similar trends for both WZ and NWZ. However, the duration of secondary task and traffic density resulted in increased risk of involvement in a CNC event in WZ compared to NWZ. The AUC values for the WZ and NWZ models were 0.8414 and 0.8564 respectively, indicating the satisfying prediction ability of the models. In addition, two scenarios were considered in this study for event classification: in Case I, the events were classified into safety critical events (crash or non-crash) and baseline events; in Case II, the classification was between crash and near crash events. In both cases, 11 driver indicator variables were utilized to develop the models. In addition, 30 other pre-incident variables were considered in developing the model for Case I, and 61 other pre-incident variables in Case II. For WZ events, the RF model was found to have the best predictive performance with successfully predicting 86.3% of events in Case I and 91.2% of events in Case II. However, the DNN model outperformed the other three models in predicting crash and near-crash events in NWZ.

Osman, Hajij, Bakhit, and Ishak (2019) performed a comparative analysis for predicting the near-crash events from NDS vehicle kinematics data (speed, longitudinal

acceleration, lateral acceleration, yaw rate, and pedal position) using several machine learning methods: KNN, RF, SVM, DT, GNB, and AdaBoost. The hypothesis of this study was that vehicles experience a change in their kinematic pattern before a near-crash occurrence. The dataset used in the study contained 250 near-crashes and 250 baseline events. The findings of the study showed that AdaBoost outperformed all other models with the recall value of 100%, precision of 98%, and F1-score of 99%.

In conclusion, most of the aforementioned studies have confirmed driver behavior to be a leading indicator of crashes.

2.7 Summary of the Literature Review Findings

The summaries of selected crash likelihood and crash severity analysis studies reviewed as part of the literature search are provided in Table 2.1 and Table 2.2, respectively.

Table 2.1 Summary of the Selected Crash Likelihood Prediction Studies

Article	Data Sources	Modeling Method*	Top Performing Model Metric(s)
Ahmed and Abdel-Aty (2011)	• Speed data from AVI systems • Crash data from FDOT	• Binary logistic regression	• Avg. Sensitivity: 0.68 • Avg. Specificity: 0.53
Xu, Liu, Yang, and Wang (2016)	• High resolution traffic data (count, speed, occupancy) • Crash data from Caltrans	• Fixed-effect logistic regression • **Random-effect logistic regression**	• AUC: 0.83
Wang et al. (2015)	• Traffic data from MVDS • Roadway geometry data • Weather data from NCDC • Crash data from FDOT	• Bayesian logistic regression	• SV crashes model: AUC: 0.97; Overall accuracy: 0.9 • MV crash model: AUC: 0.76; Overall accuracy: 0.76
Xu, Tarko, Wang, and Liu (2013)	• Traffic data from loop detectors • Roadway geometry data from PeMS • Weather data from NCDC • Crash data from Caltrans	• Sequential binary logit model	• PDO crashes: Overall accuracy: 0.75 • Non-incapacitating and possible injury (BC) crashes: Overall accuracy: 0.67 • Fatal and incapacitating injury (KA) crashes: Overall accuracy: 0.76
Yu and Abdel-Aty (2013)	• Traffic data from RTMS • Crash data from CDOT	• Bayesian fixed-parameter logistic regression • Bayesian random-parameter logistic regression • Bayesian random-effect logistic regression • SVM with linear kernel function • **SVM with RB kernel function**	• AUC: 0.77
Theofilatos et al. (2019)	• Traffic data from inductive loop detectors • Weather data from the Hydrological Observatory of Athens website • Crash data from Greek crash database	• KNN • NB • DTs • RF • SVM • SNN–Shallow Learning • **DFNN**	• Overall accuracy: 0.68 • Sensitivity: 0.52 • Specificity: 0.77 • AUC: 0.64

* The top performing modeling method is shown in bold font.

Table 2.2 Summary of the Selected Crash Severity Prediction Studies

Article	Data Sources	Modeling Method*	Top Performing Model Metric(s)
Yu and Abdel-Aty (2014b)	• Traffic data from AVI detectors • Roadway Geometry data from RCI • Weather data from weather stations • Crash data from CDOT and CAR	• Regular BP model • Bayesian BP model • **Random-effect hierarchical Bayesian BP model** • Random-effect Bayesian BP	-
Yu and Abdel-Aty (2014a)	• Traffic data from AVI detectors Roadway • Geometry data from RCI • Weather data from weather stations • Crash data from CDOT	• Bayesian fixed-parameter binary logit model • SVM with RB kernel function • **Random-parameter logit model with unrestricted variance-covariance matrix**	AUC: 0.83
Zhang, Li, Pu, and Xu (2018)	Crash data from FDOT	• OP • MNL • KNN • DT • **RF** • SVM	Overall accuracy: 0.53

* The top performing modeling method is shown in bold font.

Similarly, the summary of selected crash risk analysis studies based on the NDS data is provided in Table 2.3.

Table 2.3 Summary of the Selected Crash Risk Studies Based on the NDS Data

Article	Modeling Method*	Significant factors	Top Performing Model Metric(s)
Arvin and Khattak (2020)	• Fixed-parameter logistic regression • **Random-parameter logistic regression**	• Duration of distraction • Impaired driving	McFadden's R-Squared: 0.159
Bakhit, Guo, and Ishak (2018)	• MNL • Association analysis	• Secondary task engagement • Geometry data from RCI	
Bharadwaj, Edara, and Sun (2019)	• Logistic regression	• Driver behavior	AUC: 0.88
Mousa, Bakhit, and Ishak (2019)	• **XGB** • GBM • BANN • DNN	• Driver behavior • Intersection influence	Overall accuracy: 84.9%

* The top performing modeling technique is shown in bold font.

CHAPTER 3

METHODOLOGY

This chapter introduces the analysis methods, model performance criteria, and data sources employed in the dissertation research study. The data preparation techniques for the real-time crash risk analysis are also discussed in this chapter. These techniques include generating non-crash cases for the crash likelihood modeling, determining the significant variables, and dealing with the data imbalance problem.

3.1 General Modeling Methodology

Two separate methodologies are introduced in this dissertation for crash risk modeling: first, the short-term (near-real-time) crash likelihood and crash severity prediction models were developed using generally available input data; second, the crash likelihood and crash severity prediction models were developed using NDS data to ascertain the impact of driver behavior in such models and discuss practical implications of using crash modeling in traffic management. The flowchart of the overall data analysis and modeling methodology for the real-time crash risk and NDS analysis are shown in Figure 3.1 and 3.2, respectively.

As illustrated in Figure 3.1, the model development methodology for real-time crash risk and severity using the commonly available data includes seven main steps:

1. In the first step, data are extracted from different data sources to create the input dataset.

2. In the second step, two data preprocessing methods are applied to prepare the data for the analysis: (a) intercorrelation analysis to find the potential correlation between the explanatory variables, and (b) random forest (RF) variable importance analysis to determine the significant variables.

3. In the third step, the data is split into training and testing sets and the random oversampling examples (ROSE) method is applied on the training set to deal with the data imbalance problem in the crash severity dataset.

4. In the fourth step, the selected modeling methods are applied and tuned to predict crash risk at a road segment level, based on the training set.

5. Next, in the fifth step, the testing set is used to evaluate the models' performance and calculate the model performance metrics.

6. In the sixth step, the best model is identified based on the calculated performance metrics.

7. Lastly, the candidate model is proposed for real-world application.

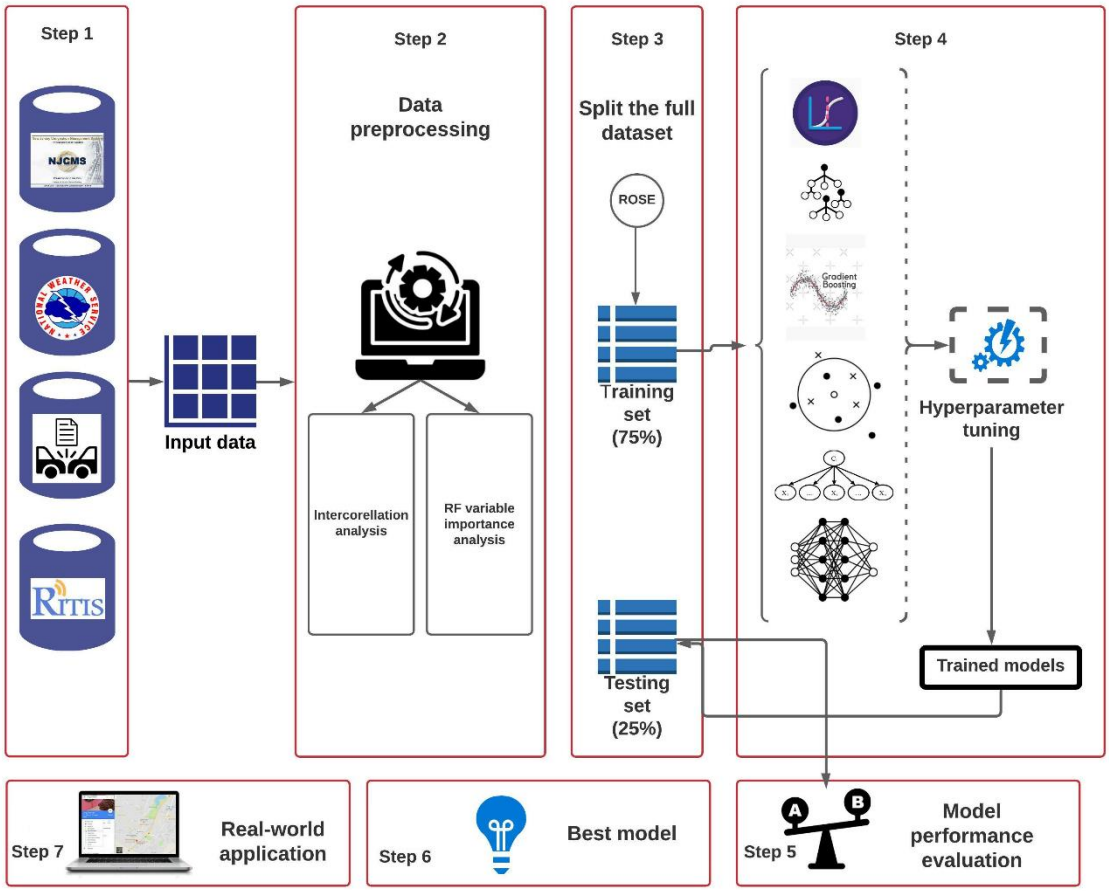


Figure 3.1 Flowchart represinting real-time crash risk analysis.

The modeling methodology for the NDS analysis is very similar to the real-time crash risk analysis using generally available data, with minor differences. Unlike real-time crash risk analysis, NDS analysis does not require extracting data from different data sources. Also, the preprocessing step is not part of the NDS methodology. Figure 3.2 illustrates the steps in the model development using NDS data:

1. In the first step, before the model training, the NDS input data is partitioned into training and testing sets.
2. In step two, the models are developed (tuned) using the training data.
3. In the third step, the model performance is assessed using the test data to find the best model in terms of predictive performance.
4. In the fifth step, the best model is identified based on the calculated performance metrics.
5. Lastly, the candidate model is proposed for real-world application.

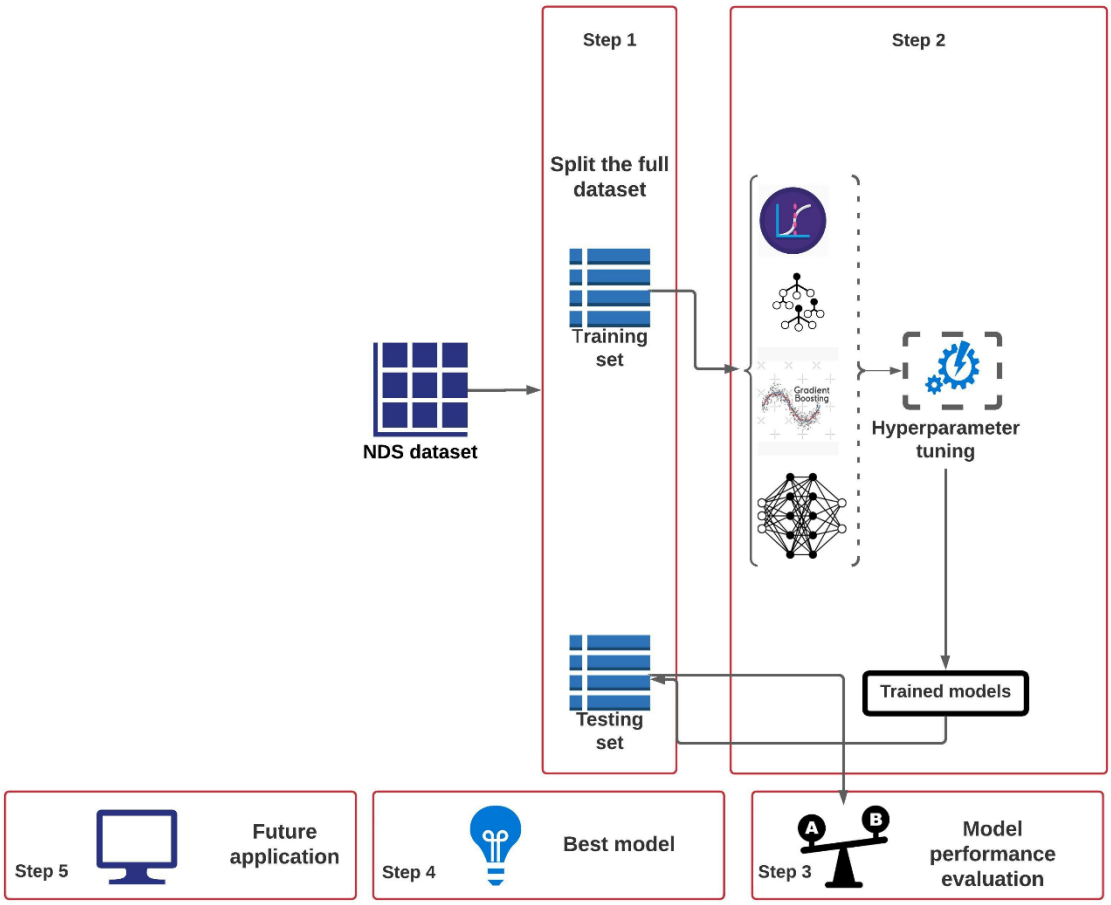


Figure 3.2 Flowchart representing NDS analysis.

3.2 Crash Analysis Modeling Methods

Based on the literature review, a number of modeling methods was considered, including regression models and machine learning models. Considering the objectives and the scope of this study, as well as findings of the previous studies documented in literature, the following methods were selected for the analysis and prediction of crash likelihood and severity: Random Effects Bayesian Logistics Regression (BLR), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), K-Nearest Neighbor (KNN),

Gaussian Naïve Bayes (GNB), and Multi-layer Feedforward Deep Neural Network (MLFDNN). Each method is briefly explained in the following subsections.

3.2.1 Random Effects Bayesian Logistic Regression

This study applied the random effects Bayesian logistic regression model to predict the crash risk. The main difference between the standard logistic regression models and Bayesian models is that in the former models the regression coefficients are fixed, while the Bayesian models assume that the coefficients follow a probability distribution. The other advantage of Bayesian models over the standard models is their capability of avoiding the odds ratio overestimation problem.

In general, the prior distributions of coefficients can be categorized into two main groups based on the availability of prior information about the possible values of the coefficients: informative priors and non-informative priors. While informative priors are used when the possible values of the coefficients are known, non-informative priors are used when little or nothing is known about the values of the coefficients.

In the crash models considered in this study, the binary outcomes are $y_i = 1$ and $y_i = 0$: in the crash likelihood model, "1" represents a crash and "0" represents a non-crash case; in the crash severity model, "1" represents an injury/fatal crash, and "0" represents a property-damage-only (PDO) crash. The probabilities associated with the binary events are p_i and $1 - p_i$, respectively. Thus, applying the Bayes theorem, the random effects Bayesian logistic regression is built as follows:

$$y_i \sim Bernoulli(p_i) \tag{3.1}$$

$$\text{logit}(p) \;=\; \log\left(\frac{p_i}{1-p_i}\right) = X\beta + \; u_j(i) \tag{3.2}$$

where the probability of each observation (y_i) is assumed to follow a Bernoulli distribution, X is the vector of explanatory variables, and β is the vector of coefficients associated with them. The u_j is the random effect variables which accounts for the unobserved heterogeneity in the input data, e.g., associated with the geometric characteristics of road segments not considered in the model, such as grades, work zones, and pavement condition of a road segment.

3.2.2 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that can be defined as the combination of Breiman's bagging idea, (Breiman, Friedman, Stone, & Olshen, 1984) and random feature selection. The basic idea behind RF is to build a collection of decision trees by bootstrapping the sample and use a random subset of input factors for splitting at each node. Thus, an RF consists of multiple decision trees where each of them presents a model (e.g., classification) with a subset of features. The RF outputs are generated as the averages of all decision trees in the forest, which is referred to as voting. The RF models often outperform the traditional classification and regression trees (CART) in terms of accuracy and capability of providing unbiased error. The other advantage of RF over CART is that it obviates the need for a separate cross-validation dataset. The RF is a common method used in different crash likelihood studies (Theofilatos, 2017; Theofilatos et al., 2019).

During the training procedure, about one-third of the training data is held out and is not used in model development. These cases are referred to as the out-of-bag (OOB) data (Breiman, 2000). The main objective of RF is to tune the primary model by selecting the

optimal values of hyperparameters that minimize the OOB error. For example, reducing the number of randomly sampled variables available for splitting at each tree node (*mtry),* reduces both the correlation and the strength. Therefore, an important step in model development is to find the optimal number of *mtry*s. OOB error is a function of the correlation between each pair of trees in the forest and the strength of each individual tree. There is a positive relationship between the inter-tree correlation and OOB error, while the relationship between the strength of the individual trees and OOB error is negative.

The OOB data can further be used to quantify the variable importance. The importance of a variable can be explained by examining the change in the prediction error when that variable is permutated or excluded in the OOB data, while all the other variables remained unchanged. After obtaining the new OOB error, the variable importance can be determined by calculating Mean Decrease Accuracy (MDA) as an average difference in the new error and the initial error over all trees in the random forest (Nicodemus, 2011). Higher values of MDA indicate greater relative importance of a variable. Another variable importance measure is Mean Decrease Gini, which is defined as the average across the forest of the decrease in Gini impurity indicator for a factor (Nicodemus, 2011). While both methods have been used in the literature, MDA was chosen for variable ranking in this study.

3.2.3 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is a powerful ML method, proposed by Friedman (2001). Like RF, GBM is also an ensemble method, using decision trees as the base modeling approach. However, unlike RF, which creates large trees, GBM grows a sequence of small trees such that each tree tries to capture those parts of the training set

which were missed in the preceding tree (Hastie, Tibshirani, & Friedman, 2009). To this end, GBM identifies the missing parts using the gradient of some differentiable loss function on random subsamples of the training set with different sizes. In this study the multinomial deviance is used as the loss function.

3.2.4 K-Nearest Neighbor (KNN)

KNN is a machine learning method that classifies observations of interest based on the labels of its k-th nearest neighbors, identified based on some measure of multi-dimensional distance. As all the *K* neighboring observations do not normally belong to the same class, the class label of the majority of them is selected as the class label of the unclassified observation (Bishop, 2006). Two decisions need to be made with regards to KNN: the value of *K* and the distance function. Normally, the best value of *K* is achieved through an iterative process in which different values are examined and the one that results in the best model performance in terms of the selected performance metric is chosen. Small values of *K* may create weak models unable to properly classify the features in the model, while large values of *K* can lead to overfitting. In addition, as a rule of thumb, when there are only two classes, which is the case in our study, *K* should be an odd integer to avoid ties (Cigdem & Ozden, 2018). With respect to the distance function, Euclidean distance, weighted Euclidean distance, and cosine method are the most commonly used in KNN models. In this study, the Euclidean distance was used as the distance function. Euclidean distance can be formulated as:

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{3.3}$$

where $dist(x_i, x_j)$ denotes the distance between observation i and j, and x_{ik} and x_{jk} are the value of the Kth factor for i and j, respectively.

3.2.5 Gaussian Naïve Bayes (GNB)

The Naïve Bayes (NB) algorithm is one of the probabilistic classification methods based on Bayes' theorem, which assumes that the features are strongly independent of each other. This method has been used in various road safety studies (Shanthi & Ramani, 2011; Theofilatos et al., 2019). Using the Bayes theorem, the posterior probability of a class target y occurs given the attribute vector X, $x_i \in X$, $1 \leq i \leq n$, which is calculated as follows:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(X)} \tag{3.4}$$

where $P(y|X)$ denotes the posterior probability that class y occurs given feature x, and $P(X|y)$ denotes the likelihood probability of x given class y. The $P(y)$ and $P(X)$ represent the prior probabilities of class y and X respectively, which occur independently. In this study, the Gaussian Naïve Bayes (GNB) method uses the Gaussian likelihood function for posterior probabilities:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\pi\sigma_y^2}\right) \tag{3.5}$$

where the parameters σ_y, and μ_y are estimated using maximum likelihood.

3.2.6 Multi-layer Feedforward Deep Neural Network (MLFDNN)

The Multi-layer Feedforward (MLF) neural network consists of model neurons, that are ordered into three main groups of layers: one input layer, one output layer, and one or more

hidden layers (Figure 3.3). The neurons receive information from a user-provided input. Hidden layers are where the majority of learning takes place, and the role of output layer is to project the results. The MLF neural networks can generally be classified into two groups: shallow neural networks with a single hidden layer, and deep neural networks with a structure that consists of multiple hidden layers. Incorporating multiple hidden layers allows for a more sophisticated buildup from simple elements to more complex ones and enables the analysis of high-dimensional data. The Multi-layer Feedforward Deep Neural networks (MLFDNNs) are densely connected layers in which the inputs influence each successive hidden layer with different connection weights. These weights are calculated and adjusted based on different learning rules, such as back propagation (Svozil, Kvasnicka, & Pospichal, 1997).

This study employs MLFDNN to train a function that maps a set of input variables X (including the explanatory variables of the crash and non-crash events) to an output variable y (crash outcome or crash severity) with gradient descent, using back propagation.

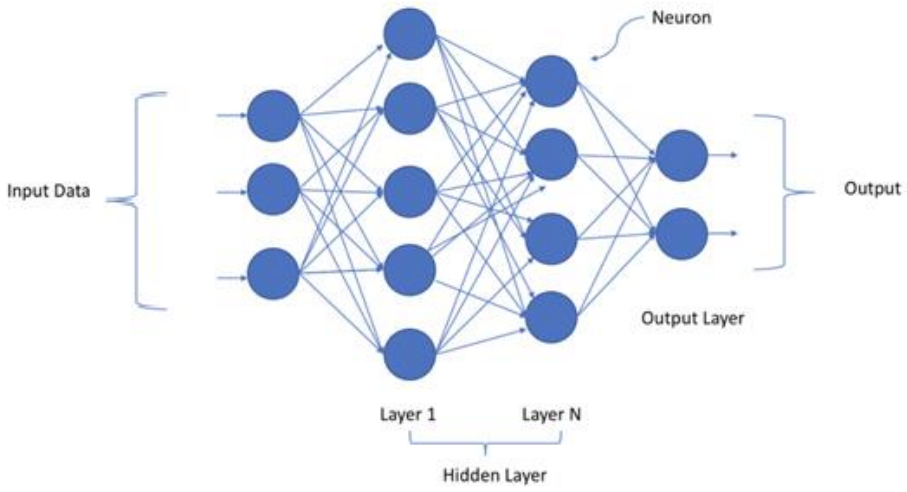


Figure 3.3 Multilayer feedforward neural network.

3.3 Model Performance Criteria

The quality of the predictions provided by different models considered in this study was evaluated based on the confusion matrices and their related performance measures: overall accuracy, sensitivity, specificity, precision, F1-score, as well as the AUC value. Calculating these metrics requires obtaining the true positive (TP), the true negative (TN), the false positive (FP), and the false negative (FN) predictions first. The definitions of these values are provided as follows:

- *TP*: True positive value is defined as the number of crash cases (injury/fatality cases in the injury severity model) that are correctly predicted as crash cases (injury/fatality cases).
- *TN*: True negative value is defined as the number of non-crash cases (PDO cases in the injury severity models) that are correctly predicted as non-crash cases (PDO cases).
- *FP*: False positive value is defined as the number non-crash cases (PDO cases) that are falsely predicted as crash cases (injury/fatal cases).
- *FN*: False negative value is defined as the number of crash cases (injury/fatality cases) that are falsely predicted as non-crash cases (PDO cases).

Having TP, TN, FP, and FN, the performance measures are formulated as:

$$\text{Overall accuracy} = \frac{TP+TN}{Total\ crashes} \tag{3.6}$$

$$\text{Sensitivity (True Positive Rate, Recall)} = \frac{TP}{TP+FN} \tag{3.7}$$

$$\text{Specificity (True Negative Rate)} = \frac{TN}{TN+FP} \tag{3.8}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.9}$$

$$\text{F1-score} = \frac{2*TP}{2*TP+FP+FN} \tag{3.10}$$

Calculating TP, TN, FP, and FN requires the consideration of a threshold for determining whether the outcome is positive or negative (or 1 vs. 0. Changing this threshold, which is normally selected as 0.5, would change the value of the calculated overall accuracy, sensitivity, and specificity. In that sense, a performance metric, which is independent of the threshold's value is desired. Receiver operating characteristic (ROC) curve (Figure 3.4) is a probability curve that plots *sensitivity* versus *1-specificity* over a wide range of possible threshold values (Fawcett, 2006). The Area Under the ROC Curve (AUC) is a collective measure of the model's ability in correctly distinguishing between classes and is used as a summary of the ROC curve. In this study, AUC, which is an evaluation metric for binary classification problems, is used as another measure to summarize the model's performance.

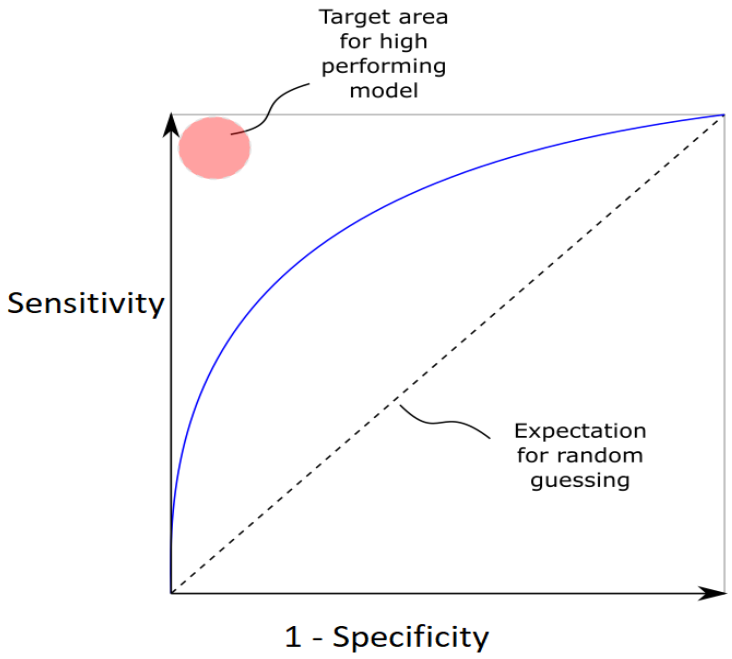


Figure 3.4 Receiver operating characteristic (ROC) curve.
Source: (Dei, 2019)

The closer the values of each of these measures is to 1, the better the prediction. However, very often the prediction models would provide better performance relative to one of these measures, and comparably worse performance relative to the other measure(s). Understanding the implications of the balance (or rather imbalance) of these measures in the model output is one of the critical aspects of interpreting the modeling results.

3.4 Real-time Crash Risk Analysis

3.4.1 Data Sources

The initial model development and analysis were conducted using the data commonly available to transportation agencies for planning and operational analysis, which can be applied in near-real time for a short-term analysis. In this study the data was obtained for a

sample of roadways in the State of New Jersey. In identifying the data sources and datasets to be collected and used in the analysis, the following types of data were of interest:

- Historical crash records – needed to obtain a record of crashes and their severity as the outcomes to be predicted by the crash likelihood and crash severity models.

- Roadway characteristics dataset – providing roadway geometry data.

- Traffic condition datasets – providing real-time data on speeds, travel times, and vehicle volume at a roadway segment level.

- Weather conditions data – providing real-time weather information, such as temperature, precipitation, visibility, wind, etc.

Following the detailed search and review of the datasets available for the major highway facilities managed by the New Jersey Department of Transportation, the following were selected as the data source for the crash prediction model development and testing:

1. NJDOT Crash Records Database, which contains records of all crashes reported by the Police Departments in the State of New Jersey using the NJTR-1 Accident Report Form. The data provides detailed information about the crash characteristics, roadway condition, environmental (ambient) conditions, vehicle characteristics, as well as the condition and characteristics of all participants in a crash. The crash data for the period January 2017 through December 2018 were acquired from the NJDOT website and used in the analysis.

2. NJDOT Congestion Management System (NJCMS), a dataset that provides estimated, synthesized hourly volume and congestion levels (expressed in terms of average speed and volume-to-capacity ratio) at a roadway segment level for all highways in NJDOT jurisdiction. This dataset also provides the basic roadway geometry data, such as number of lanes, median types, and shoulder, which were also acquired and used in developing the analysis dataset for this study. The datasets with 2012 and 2016 vehicle volume data were used as the baseline for calculating 2017 and 2018 hourly volumes for all roadway segments in this study. Moreover, the seasonal traffic factors were applied to calculate vehicle volumes specific to each month of the year.

3. Probe-vehicle mobility data at roadway segment level, which provides the actual prevailing vehicle speeds and travel times aggregated from the probe vehicles and recorded in 1-minute increments. The data was obtained from the RITIS system for the sample of roadway segments and the time periods analyzed in the study. In spatial terms the speeds and travel times are aggregated

and reported in RITIS for traffic management channel (TMC) links. The limits of TMC links do not coincide with the roadway segments defined in the NJCMS dataset, and therefore it was necessary to match and conflate the speed records from the RITIS dataset to the roadway segments defined in the NJCMS dataset for the roadways included in the analysis.

4. Historical weather data from the National Weather Service (NWS) dataset. The historical weather observation data was obtained from the NWS dataset for the locations of reported crashes and time intervals prior to the reported crash time (e.g., 15-30-minute interval). This data provides additional insight into ambient conditions at the time of crash and non-crash cases included in the model dataset. The Local Climatological Data (LCD) was identified as the most complete and reliable dataset that provides local weather information from permanent weather stations in 15-minute increments. The data record for each location and time stamp contains the ambient temperature, air pressure, visibility, hourly precipitation, hourly visibility, and average wind speed. LCD data were obtained from the National Oceanic and Atmospheric Administration (NOAA). Hourly visibility and hourly precipitation are considered as the most prominent weather variables affecting the crash likelihood and severity.

In the next step the data from the above listed data sources was reviewed and key explanatory variables were identified for inclusion in the crash likelihood and crash severity models.

3.4.2 Explanatory Variables

The explanatory variables that were identified as the most critical and informative for crash likelihood and crash severity analysis are listed in Table 3.1. In this study, the proactive data are defined as the type of data that comply with the following conditions: 1) data should be available in real-time and can be collected before the crash occurrence; and 2) data should be available for all sections of the major roadways in the State of New Jersey. In that sense, the traffic characteristics, roadway characteristics, and weather characteristics can be grouped as proactive data. The reactive data on the other hand, are the type of data that are generally available after the crash occurrence and will be used for the crash injury

severity analysis only. Driver characteristics and vehicle characteristics were classified as reactive data. All reactive data were extracted from the NJDOT Crash Records Database.

Table 3.1 Definition of the Explanatory Variables Used in the Real-time Crash Risk Study

Variable	Type	Class	Source	Description
LANES	Categorical	Proactive	NJCMS	Number of lanes {2, 3, 4, or 5}
Hour	Categorical	Proactive	NJCMS	Time of the crash or non-crash [hour]
Month	Categorical	Proactive	NJCR	Time of the crash or non-crash event [month]
MEDIAN_TY	Binary	Proactive	NJCR	Median type {protected or non-protected}
Weekend	Binary	Proactive	NJCR	Time of the crash or non-crash {weekend or weekday}
Sun glare	Binary	Proactive	NJCMS	The effect of sun glare {0: no effect and 1: Sun glare existed}
CAPLINK	Continuous	Proactive	NJCMS	Link capacity [vehicles/hour]
VC_RATIO	Continuous	Proactive	NJCMS	Volume-to-capacity ratio at the highway section during a given hour of the day and month [unitless]
Vol16_Tr	Continuous	Proactive	NJCMS	Hourly Truck volume ratio [unitless]
HourlyPrecipitation	Continuous	Proactive	NWS	Hourly precipitation at the highway section during the hour of the crash or non-crash event obtained from the weather records for the closest weather station [inches/hour]
HourlyVisibility	Continuous	Proactive	NWS	Hourly visibility at the highway section during the hour of the crash or non-crash event obtained from the weather records for the closest weather station [miles]
speed_avg_1015	Continuous	Proactive	RITIS	Average speed on the highway section [miles/hour]. It is calculated for each crash and non-crash event as an average of 1-minute prevailing speeds for the pertinent highway section over a 10-minute period (5-15 minute prior to the crash) preceding the crash or non-crash event.
speed_sd_1015	Continuous	Proactive	RITIS	Standard deviation of speed on the cash location highway section [miles/hour]. It is calculated as a standard deviation of 1-minute prevailing speeds for the pertinent highway section over a 10-minute period (5-15 minute prior to the crash) preceding the crash or non-crash event.

Table 3.2 Definition of the Explanatory Variables Used in the Real-time Crash Risk Study (Continued)

Variable	Type	Class	Source	Description
speedup_sd_1015	Continuous	Proactive	RITIS	Standard deviation of speed on the upstream highway section [miles/hour]. It is calculated as a standard deviation of 1-minute prevailing speeds for the pertinent highway section over a 10-minute period (5-15 minute prior to the crash) preceding the crash or non-crash event.
speeddown_sd_1015	Continuous	Proactive	RITIS	Standard deviation of speed on the downstream highway section [miles/hour]. It is calculated as a standard deviation of 1-minute prevailing speeds for the pertinent highway section over a 10-minute period (5-15 minute prior to the crash) preceding the crash or non-crash event.
speedup_dif_1015	Continuous	Proactive	RITIS	Speed deviation from the speed limit [miles/hour]. Calculated as the difference between the average speed (speed_avg) and the speed limit (obtained for the upstream roadway segment from the NJCMS dataset) for each crash and non-crash event at the given highway section.
speeddown_dif_1015	Continuous	Proactive	RITIS	Speed deviation from the speed limit [miles/hour]. Calculated as the difference between the average speed (speed_avg) and the speed limit (obtained for the downstream roadway segment from the NJCMS dataset).
Shape_Leng	Continuous	Proactive	RITIS	Length of the segment
Age	Categorical	Reactive	NJCR	Driver's age {age≤25, 25<age≤60, or age>60}
Veh_age	Categorical	Reactive	NJCR	Driver's age {0<age≤5, 5<age≤10, or age>10}

3.4.3 Generating Non-crash Cases for the Crash Likelihood Modeling

This study employed a matched case–control methodology in developing the dataset of crash and non-crash cases for the crash likelihood modeling. In the matched case-control methodology, non-crash cases are introduced in the analysis to match the crash cases in terms of crash characteristics such as location and time. To that end, for every crash case,

four non-crash cases were generated for the same location, day of the week and time of day, including one each in the week before, two weeks before, a week after, and two weeks after the crash occurrence. The 1:4 ratio of crash cases to non-crash cases was recommended by Ahmed and Abdel-Aty (2011) who found this value to provide slightly better results when compared to other crash to non-crash case ratios. In addition, according to the finding of another study by S. Kuhn, Egert, Neumann, and Steinbeck (2008), negligible improvement can be achieved by adding non-crash cases beyond 1:3 ratio. It should be noted that the matched case–control methodology employed in this study only accounted for the location (roadway) and time as the crash factors; the other factors, such as vehicle, driver, and environmental characteristics were not considered in the case-control matching.

3.4.4 Determination of Significant Variables

In this study, Random Forest (RF) model was used to determine relative importance of variables to be used in the crash likelihood and crash severity models. This allows to only include the significant variables in models such as KNN, which can produce misleading results in high-dimensional space. In both the crash likelihood and crash severity model datasets, the Mean Decrease in Accuracy (MDA) was used as the criterion in determining the relative variable importance. The mean decrease in accuracy for a variable is calculated based on the out of bag (OOB) error. The importance of a variable can be explained by examining the change in the prediction error when that variable is permutated or excluded in the OOB data, while all the other variables remained unchanged. After obtaining the new OOB error, the variable importance can be determined by calculating MDA as an average difference in the new error and the initial error over all trees in the random forest

(Nicodemus, 2011). Higher values of MDA indicate greater relative importance of a variable.

3.4.5 Dealing with the Data Imbalance Problem

To overcome the problem of a low frequency of fatal crashes, the fatality class was initially combined with the instances in the injury class. However, even after undertaking this action, 79% of the cases were non-injury crashes (8,016 PDO crashes out of the total of 10,155 crashes in the dataset) and 21% of crashes (total of 2,139) were crashes with an injury or a fatal outcome. In the case of training the model with a skewed distribution of classes, the traditional accuracy maximizer techniques are not adequate and normally tend to perform better in favor of the prevalent class. Therefore, it is advantageous to transform the dataset so as to achieve a more balanced training dataset.

Random oversampling examples (ROSE) is a random bootstrapped-based technique, introduced by Menardi and Torelli (2014), which can alleviate the data imbalance issue in the binary classification problems. ROSE combines random oversampling and random undersampling by generating new artificial instances from the original classes based on a smoothed bootstrapped approach (Tibshirani & Efron, 1993).

Consider a training set of size *n*, consisting of a binary response variable y, with class labels Y_j and a set of input data for each class, $x_{ij}, i = 1, \dots, n_j$, where $n_j < n$ is the number of cases in class *j*. For each *x* belonging to the class Y_j, ROSE generates samples from a multivariate kernel density estimate of $f(x \mid y = Y_j)$ as follows:

$$\widehat{f}(x \mid y = Y_j) = \sum_{i=1}^{n_j} p_i \Pr(x \mid x_{ij}) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(x - x_{ij}) \qquad (3.11)$$

where K_{H_j} denotes an estimated kernel function and its smoothing matrix H_j is:

$$H_j = diag(h_1^{(j)}, \dots, h_d^{(j)}) \tag{3.12}$$

where d is the number of explanatory variables and

$$h_q^{(j)} = (\frac{4}{(d+2)n})^{1/(d+4)} \hat{\sigma}_q^{(j)}, q = 1, \dots, d \tag{3.13}$$

where $\hat{\sigma}_q^{(j)}$ is the estimated standard deviation of the qth variable. According to Bowman and Azzalini (1997), the smoothing matrix minimizes the Asymptotic Mean Integrated Squared Error under the assumption that the true conditional densities underlying the data follow a Normal distribution.

The practical implementation of ROSE encompasses the following steps:

1. select $y^* = Y_j$ with probability π_j;
2. select x such that $y_k = y^*, k = 1, \dots, n$ with probability $\frac{1}{n_j}$;
3. sample x^* from the estimated kernel function.

Repeating steps 1 to 3 yields a newly generated training set of size *m*, with the probability of each class to be π_j.

Implementing the newly created dataset using the ROSE method is expected to provide better results than using the original imbalanced dataset. In addition, the findings of a study by Menardi and Torelli (2014) showed that ROSE outperformed other well-known oversampling methods, such as synthetic minority oversampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), by providing higher values of the area under ROC curve (AUC) in the logistic regression and classification tree models.

In this study, the ROSE method was applied to the training set for the crash severity models to generate synthetic training set.

3.5 Crash Risk Modeling Using NDS Data

3.5.1 Data Sources

Recently, Naturalistic Driving Study (NDS) was performed under the second Strategic Highway Research Program (SHRP 2) of the National Academy of Science (NAS). The NDS has emerged as an alternative source of data for evaluating driving behavior. The advantage of NDS data over the traditional datasets is its ability to reflect the actual driver behavior. This identification is made by monitoring the driver's actions in a natural setting via several on-board devices and for a relatively long period.

The main objective of NDS is to answer the need for reducing the toll of motor vehicle crashes by filling the gaps of the previous studies, which investigated the driver reaction to different scenarios using driving simulators or test vehicles. Approximately 4 petabytes of video and sensor data were collected and stored in the NDS dataset. The dataset is comprised of over 50 million miles of travel, 900,000 hours of in-vehicle time, and 5.5 million trips taken by instrumented vehicles of about 3,500 volunteer drivers in six states (Washington, New York, Pennsylvania, North Carolina, Florida, and Indiana). The following data categories were collected during NDS (Antin et al., 2019):

- Videos and images.
- Time-series data: Include vehicle kinematics (e.g., 3D acceleration and deceleration), forward Radio Detection and Ranging (RADAR), Global Positioning System (GPS) data, turn signal usage, seat belt usage, and presence of alcohol in the cabin.

- Crash/Near-crash (CNC) data: Crashes are recorded whenever the vehicle hits another vehicle or object, and Near-crashes are recorded whenever a severe evasive maneuver is made to avoid a crash.

- Driver assessments: All drivers were asked to participate in different assessments and questionnaires. Assessments addressed the driver's cognitive, perceptive, and physical abilities, and questionnaires addressed their attitudes (e.g., health status and medication, perception of risk, and sensation seeking), driving history and knowledge of driving rules and regulations.

- Vehicle features: Include information about make, model, condition, and onboard features (e.g., safety features, adaptive cruise control, navigation system, and voice recognition) of the participating vehicles.

- Crash investigations: Contain two levels of crash investigations. Level I crash investigations provide as much information that can be obtained about a crash without visiting the crash site. This information may include police crash reports, photographs of the vehicle involved in the crash after being removed from the crash scene, Google Earth images of the crash location, and interview with the participant driver. Level II crash investigations provide all information included in Level I, as well as the information collected in a crash site visit.

- Cell phone records: Collected from those drivers who consented to contribute the records of their cell phones. These records include call durations and the time spent for reading the text messages sent to or from the same phone.

- Roadway information: Contained in the SHRP 2 Roadway Information Database (RID) and includes detailed information about roadway geometry (e.g., curves, medians, number of lanes, shoulder, locations of intersections), lighting condition, signs, and rumble strips for 12,000 miles of road travelled by the study participants.

- Supplemental data: Also stored in RID and include other safety-related geospatial data collected mostly by transportation agencies. Such data include weather, work zones, and safety programs at the study locations.

It should be noted that NDS and RID datasets can be linked through the GPS coordinates – latitude and longitude (McLaughlin & Hankey, 2015).

3.5.2 Explanatory Variables

In addition to crash and near-crash (CNC) events, the processed dataset provides more than 19,991 baseline (non-crash) events selected using case-cohort and case-crossover random

sampling techniques, stratified by drivers and driving time (Hankey, Perez, & McClafferty, 2016). Baseline events are critical for crash risk analysis as they provide information on normal driving and typical driver behavior. This dataset has the format similar to the CNC events dataset, and it is used in developing the crash risk models.

For each event, data are available for 76 different variables that can be grouped in several categories. Two distinct datasets of the SHRP2-NDS data were merged and employed in this study:

- **Event characteristics dataset:** Contains all the events (crash, near-crash, and baseline) and the associated severity level for crash events. In addition, environmental characteristics, roadway geometry characteristics, and driving behavior information are also provided as part of the event dataset.

- **Driver characteristics dataset:** Contains the socioeconomic characteristics such as age, gender and education level of the drivers who participated in the program.

Table 3.2 provides a list of the explanatory variables used in the model development. These variables include indicators for driver characteristics, vehicle characteristics, environmental characteristics, and roadway characteristics. Some variables listed in Table 3.2 require additional explanations are as follows:

- **Maneuver judgement:** A vehicle kinematic measure-based variable that describes the legality and safety of a pre-incident maneuver.

- **Driver behavior:** The drivers' actions that cause or contribute to a crash or near-crash event. These include the state or behavior of the driver either within seconds prior to a CNC event or those resulting from the context of the driving environment. In order to provide enough cases belonging to each behavior category, the driver behavior categories in this study were merged into eight larger groups, namely: normal driving, aggressive driving, avoiding other vehicles/pedestrians, distracted/drowsy/fatigued, inattention, sign/signal violation, speed violation, and unnecessary risky driving actions.

- **Driver impairment:** The apparent reason for the observed driver behavior and judgment. In this study, driver impairment was classified into four categories: no impairment, drowsy/fatigued, emotional state, and alcohol or drug.

- **Secondary task:** Includes any observable driver engagement in a secondary task. This does not include tasks that are part of the driving task, such as speedometer checking, mirrors checking, blind spot checking, and gear shifting. For CNC events the secondary tasks begin at any point within 5 seconds prior to the precipitating event time and continue through the end of conflict. For baseline events, secondary tasks are coded for the last 6 seconds of the baseline epoch, which includes 5 seconds prior to Event Start through one second after (to the end of the baseline).

Table 3.2 Definition of the Explanatory Variables Used in the NDS Crash Risk Study

Variable	Type	Category	Availability in real-time	Description
maneuverJudgment	Categorical	Driver characteristics	Not available	Maneuver Judgement
Behavior	Categorical	Driver characteristics	Not available	Driver Behavior
Impairment	Categorical	Driver characteristics	Not available	Driver Impairment
SecondaryTask1	Categorical	Driver characteristics	Not available	Secondary Task engagement
SecondaryDur	Continuous	Driver characteristics	Not available	Secondary task duration
Seatbelt	Binary	Driver characteristics	Not available	Seatbelt Usage
ageGroup	Categorical	Driver characteristics	Not available	Age
educ	Categorical	Driver characteristics	Not available	Education
Male	Binary	Driver characteristics	Not available	Gender
lighting	Categorical	Environmental characteristics	Available	Lighting condition
surfaceCondition	Categorical	Environmental characteristics	Available	Surface Condition
traddicDensity	Categorical	Roadway characteristics	Available	Traffic Density
intersectionInfluence	Categorical	Roadway characteristics	Available	Intersection Influence
grade	Categorical	Roadway characteristics	Available	Roadway grade
Curve	Binary	Roadway characteristics	Available	Roadway alignment
WorkZone	Binary	Roadway characteristics	Available	Presence of work zone
vehClass	Categorical	Vehicle characteristics	Not available	Vehicle classification
Adv.Tech	Binary	Vehicle characteristics	Not available	Advanced vehicle technology
Int.Cell	Binary	Vehicle characteristics	Not available	Vehicle integrated cellphone system

CHAPTER 4
CASE STUDY MODEL IMPLEMENTATIONS

4.1 Case Study of I-80 and I-287 in New Jersey

The case study for developing the initial short-term (real-time) crash likelihood and crash severity prediction models focused on two interstate highways in New Jersey: I-80 and I-287. The interstate I-80 has a west-to-east alignment, and the New Jersey section is 68.5 miles long. The interstate I-287 has a south-to-north alignment and the New Jersey section is 67.5 miles long. Both roadways are located in the northern part of the State and had the highest number of crashes among the interstate highways in the State. The location of I-80 and I-287 on the map of New Jersey is shown in Figure 4.1.

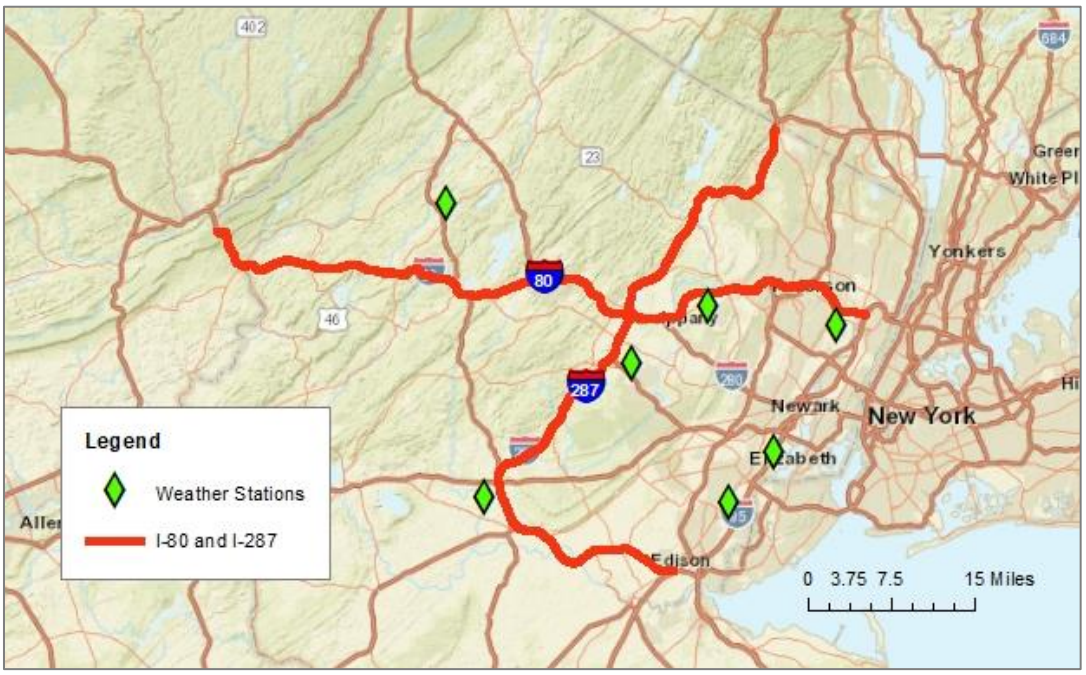


Figure 4.1 The study area with the location of I-80, I-287, and weather stations.

The weather data was obtained from the LCD database from seven weather stations located in the proximity of I-80 and I-287. For each roadway segment the closest LCD station was identified based on the Euclidian distance. The locations of LCD weather stations that provided data for the study area are shown in Figure 4.1. All stations are located at the regional airports.

4.1.1 Discussion of the Data Inputs

The dataset included the total of 10,155 crashes that were recorded along interstate I-80 and interstate I-287 during the period January 2017 – December 2018. Each crash was matched to a corresponding NJCMS record based on the unique road identifier (standard road identifier, or SRI) and milepost. The matching NJCMS record provided the segment-level roadway data, such as speed limit, hourly vehicle volume, v/c ratio, number of lanes and type of median.

The traffic speed data at the crash location prior to the time of crash was obtained from the RITIS dataset. The RITIS data was also matched to the NJCMS segment based on route name and milepost and added to the record of each crash. As previously indicated, the average speed for each segment in RITIS dataset is reported at a 1-minute interval. Nevertheless, to reduce the noise and the impact of human error in reporting the exact time of the crash, the speed data was extracted for a period of 10 minutes, between 5 and 15 minutes prior to the crash occurrence, and then aggregated to calculate the average speed, the standard deviation of speed, the coefficient of variation of speed, and the deviation from the speed limit over the same 10-minute period. For each crash these speed indicators were used as model inputs.

In addition, the weather data was extracted from the LCD data recorded at the weather station closest to the crash location for the time interval matching the date and time of crash. The weather data extracted from the LCD dataset included hourly precipitation and hourly visibility observed during the hour of the crash.

Lastly, the effect of sun glare on crash occurrence was also considered in this study. For this, the position of the sun (sun elevation θ and azimuth angle $\emptyset$) was estimated accurately based on the location (*lat*, *long*) and time (*t*) for each case (crash or non-crash) using Pysolar Python library (Stafford, 2018). As can be seen in Figure 4.2, the horizontal angle between the Sun and the vehicle can be calculates as:

$$h_{glare} = |\emptyset - \varphi| \tag{4.1}$$

where $\emptyset$ is the azimuth angle of the Sun. Similarly, the vertical angle between the Sun and the vehicle can be calculated using the following equation:

$$v_{glare} = |\theta - \theta'| \tag{4.2}$$

where θ' is the slope angle of driveway. The horizontal and vertical angle of the vehicle was estimated using the horizontal and vertical angle of the road segment where the vehicle was located at time *t*. Finally, after calculating the h_{glare} and v_{glare}, if both were below 25 degrees, sun was found to cause glare to the driver (Li et al., 2019).

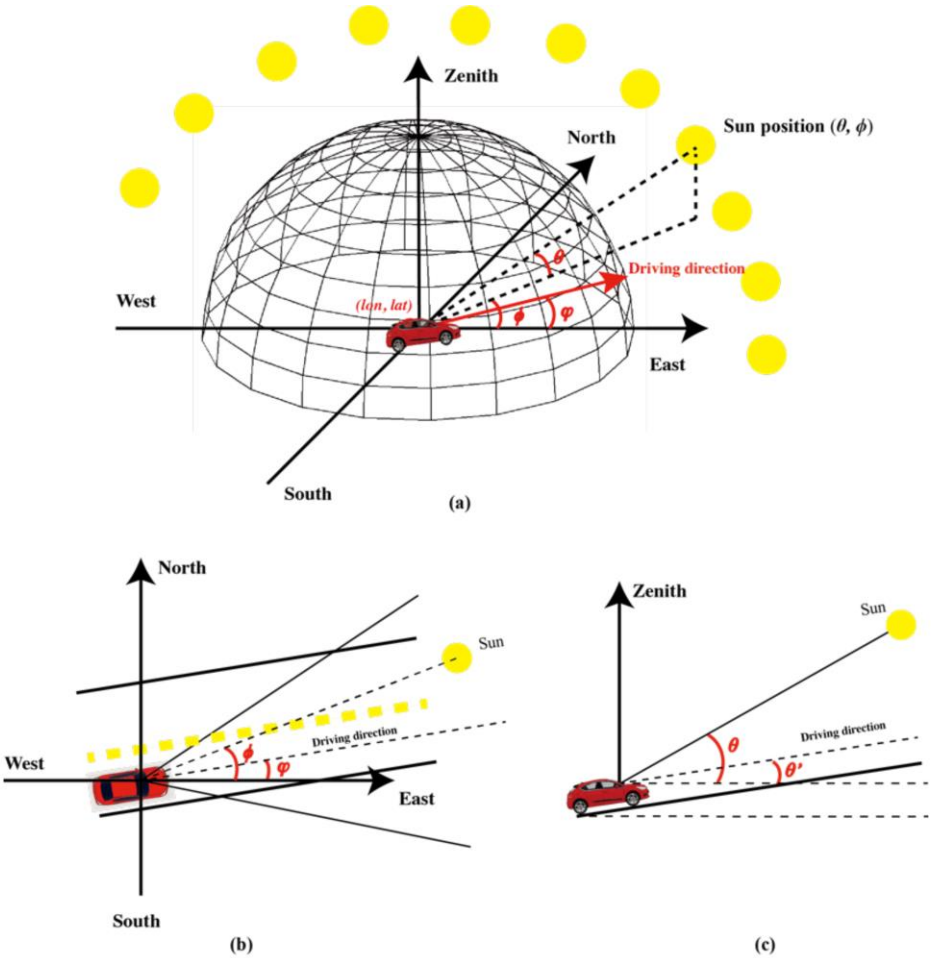


Figure 4.2 The geometric model of the relative position of the Sun and the vehicle.
Source: (Li, Cai, Qiu, Zhao, & Ratti, 2019)

The descriptive characteristics of I-80 and I-287 roadway datasets relevant to this study are summarized in Table 4.1. The continuous explanatory variables used in crash likelihood and crash severity models for I-80 and I-287 are listed in Table 4.2 along with the basic statistics from the input data. Likewise, the categorical (binary) explanatory model variables and the corresponding descriptive statistics are summarized in Table 4.3.

Table 4.1 Summary of the Roadway Segment Characteristics (Including Crash Statistics)

Characteristic	I-287	I-80	Total
Number of crashes (total)	1,267	8,888	10,155
Number of injury/fatal crashes	236	1,903	2,139
Number of PDO crashes	1,031	6,985	8,016
Roadway length (in miles)	67.5	68.5	136
Number of roadway segments (both ways)	116	164	280
Minimum length of a roadway segment (in miles)	0.020	0.100	0.020
Maximum length of a roadway segment (in miles)	5.140	4.020	5.140
Average length of a roadway segment (in miles)	1.218	0.936	1.053

Table 4.2 Summary of Basic Statistics for the Continuous Variables

Variable	Min	Max	Mean	Median
CAPLINK	3268	8570	6138	6856
VC_RATIO	0.032	1.599	0.600	0.577
Vol16_Tr	0.032	1.450	0.576	0.554
HourlyPrecipitation	0.000	0.720	0.002	0.000
HourlyVisibility	0.000	74.00	8.898	10.000
speed_avg_1015	2.00	83.00	61.56	64.80
speed_sd_1015	0.00	25.23	1.29	0.89
speedup_dif_1015	0.00	63.00	8.72	6.20
speeddown_dif_1015	0.00	63.00	8.23	5.80

Table 4.3 Summary of Basic Statistics for the Binary/Categorical Variables

Variable	Description	n	%
LANES	two-lane	679	6.69
	three-lane	4119	40.56
	four-lane	5245	51.65
	five-lane	111	1.09
MEDIAN_TY	protected	9442	92.98
	unprotected	713	7.02
HOUR	0:00	9	0.09
	1:00	137	1.35
	2:00	148	1.46
	3:00	129	1.27
	4:00	139	1.37
	5:00	273	2.69
	6:00	522	5.14
	7:00	719	7.08
	8:00	804	7.92
	9:00	517	5.09
	10:00	401	3.95
	11:00	399	3.93
	12:00	438	4.31
	13:00	434	4.27
	14:00	546	5.38
	15:00	626	6.16
	16:00	780	7.68
	17:00	982	9.67
	18:00	769	7.57
	19:00	373	3.67
	20:00	280	2.76
	21:00	292	2.88
	22:00	226	2.23
	23:00	217	2.14
MONTH	Jan	761	7.49
	Feb	733	7.22
	Mar	803	7.91
	Apr	659	6.49
	May	812	8.00
	Jun	801	7.89
	Jul	856	8.43
	Aug	887	8.73
	Sep	839	8.26
	Oct	1083	10.66
	Nov	1044	10.28
	Dec	876	8.63
Weekend	Yes	1981	19.51
	No	8174	80.49
Sunglare	Yes	1078	10.62
	No	9077	89.38

4.1.2 Data Preprocessing

As explained in Chapter 3, to balance the crash cases in the crash likelihood prediction model, non-crash cases were generated using the matched case-control method. After creating the non-crash cases, the same procedure that was applied to crashes was used to match the traffic flow, speed, and weather data to each non-crash case. After completing this step, the study dataset for the crash likelihood model had additional 40,620 records representing non-crash cases (four non-crash cases for each of the 10,155 crash records).

Before selecting the explanatory variables that should enter the models, it is important to check for correlation between the explanatory variables in the analysis dataset. To that end, the correlation matrix was created using Pearson correlation coefficient to identify the correlated variables, as shown in Figure 4.3.

It should be noted that while this method is unable to detect the non-linear dependencies among the variables, this does not present problems with developing the ML and DL models applied in this study for two reasons: first, due to the regularization parameters within the models, and second, the fact that neural networks and tree-based models are robust to multicollinear problem (Garg & Tai, 2012).

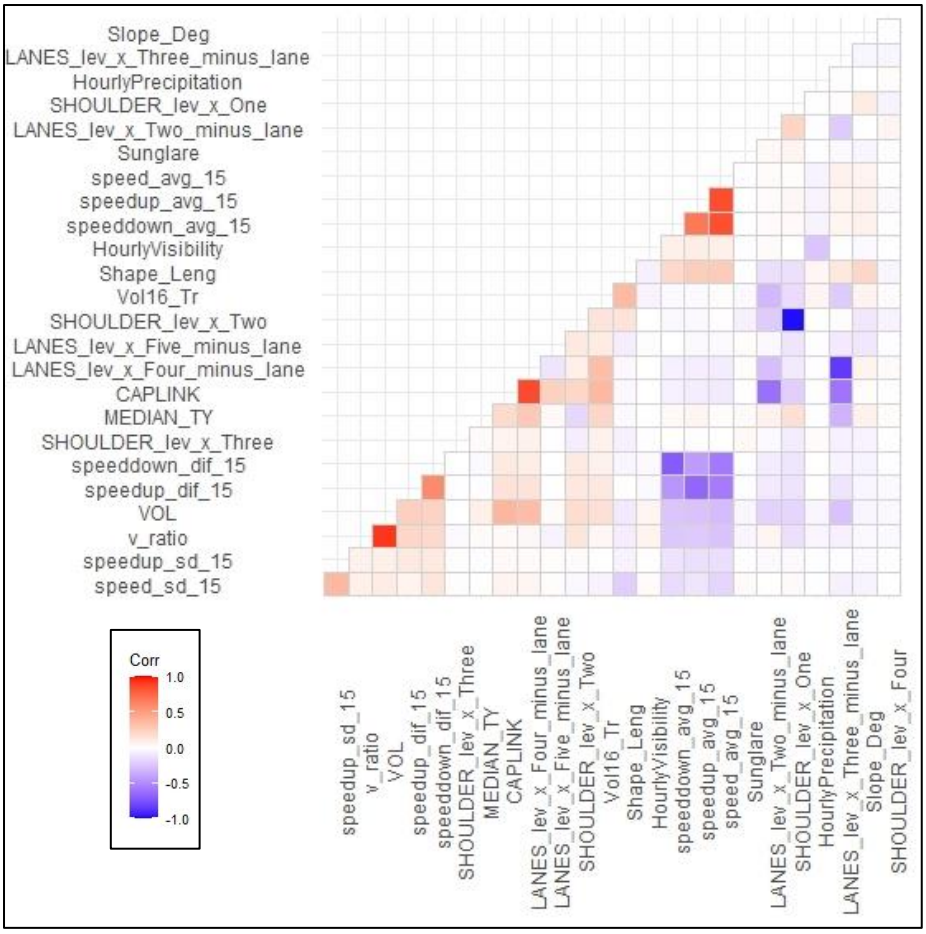


Figure 4.3 Correlation matrix for the crash likelihood analysis dataset.

Based on the correlation matrix, it was decided to exclude from further consideration the highway capacity variable (CAPLINK) as it was correlated with the number of lanes (LANES), as well as hourly volume (VOL) since it was correlated with the v-c ratio (v_ratio). In addition, the average speeds for the upstream and downstream segments were also found to be highly correlated with the average speed of the segment where the crash happened and therefore were excluded from the model.

An RF model for the crash likelihood analysis dataset was then used to determine the relative importance of the explanatory variables. The RF model had mtry = 2 (number of factors randomly sampled at each split), number of trees = 500, split.rule = Extra trees,

and node.size = 1 (minimum number of observations in each terminal node). The ranking of the relative variable importance in the crash likelihood model assessed using the RF model is illustrated in Figure 4.4. The vertical red line denotes a cordon between the significant variables that should be considered (on the right-hand side) and variables that should be excluded as insignificant (on the left-hand side of the cordon line). The line was placed where the gap between variables was relatively large in terms of MDA.

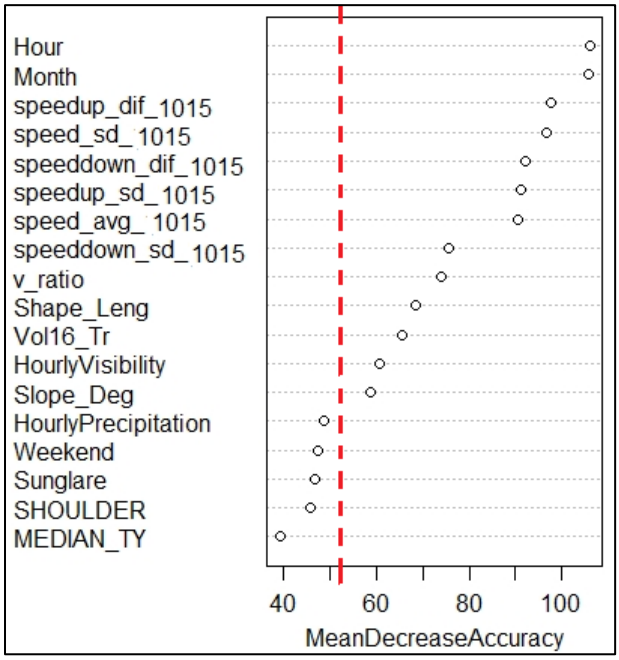


Figure 4.4 RF variable importance plot for the crash likelihood model.

As it can be observed, the variables HourlyPrecipitation, Weekend, Sunglare, SHOULDER, and Median_TY were not significant. Thus, the final list of decision variables to be used in modeling the crash likelihood included: Hour, Month, speed_sd_15, speeddown_dif_15, speedup_sd_15, speed_avg_15, speedup_dif_15, speeddown_sd_15, v_ratio, Shape_Leng, Vol16_Tr, HourlyVisibility, and Slope_Deg.

An RF model was also used to determine the relative importance of explanatory variables in the crash severity modeling dataset. The RF had mtry = 16 (number of factors randomly sampled at each split), number of trees = 500, split.rule = gini, and node.size = 1 (minimum number of observations in each terminal node). The ranking of the relative variable importance in the crash severity dataset is illustrated in Figure 4.5. The red line marks the separation between the variables with high degree of importance (right-hand side) and variables with low or insignificant importance. (left-hand side).

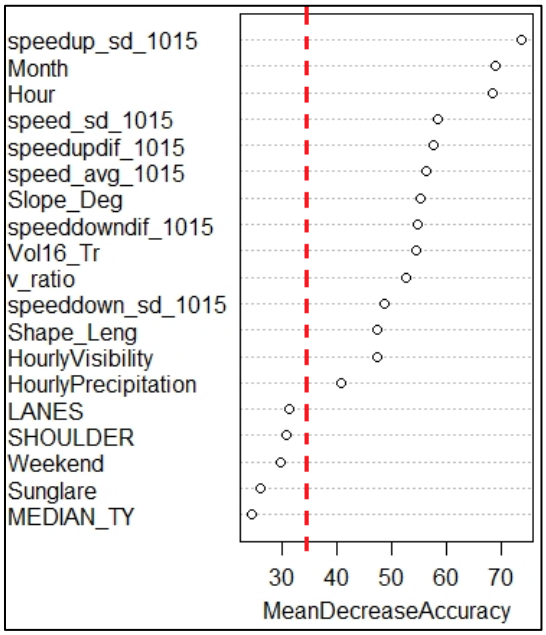


Figure 4.5 RF variable importance plot for the crash severity model.

Thus, the final list of decision variables to be used in modeling the crash severity included: speeup_sd_1015, Month, Hour, speed_sd_1015, speedupdif_1015, speed_avg_1015, Slope_Deg, speeddowndif_1015, Vol16_Tr, v_ratio, speeddown_sd_1015, Shape_Leng, HourlyVisibility, and HourlyPrecipitation.

4.1.3 Preparation of the Training and the Testing Datasets

For the analysis purposes, it was first necessary to split both the crash likelihood and the crash severity datasets into two subsets each: (a) training dataset, containing 75% of features (data records), and (b) testing dataset, containing 25% of features. A stratified sampling technique was used for splitting the datasets to ensure that there is the same proportion of output class labels in both the training set and testing set, as in the original data.

As explained in Chapter 3, the ROSE transformation was applied to the training dataset for the crash severity model to address the imbalance between the severe crashes (with injuries and/or fatalities) and other (PDO) crashes. Following the ROSE methodology, different probability values for the minority classes in each dataset were evaluated (e.g., 0.2, 0.3, 0.4, 0.5, 0.6). The evaluation showed that the probability of 0.5 yielded best results in terms of sensitivity and AUC in the crash injury severity training datasets. A visual representation of the dataset before and after applying ROSE is shown in Figure 4.6, displaying the example of the data reflecting the average speed vs. v/c ratio. The visual representation shows more balanced distribution of the severe (injury/fatal) vs. other (PDO) crashes.

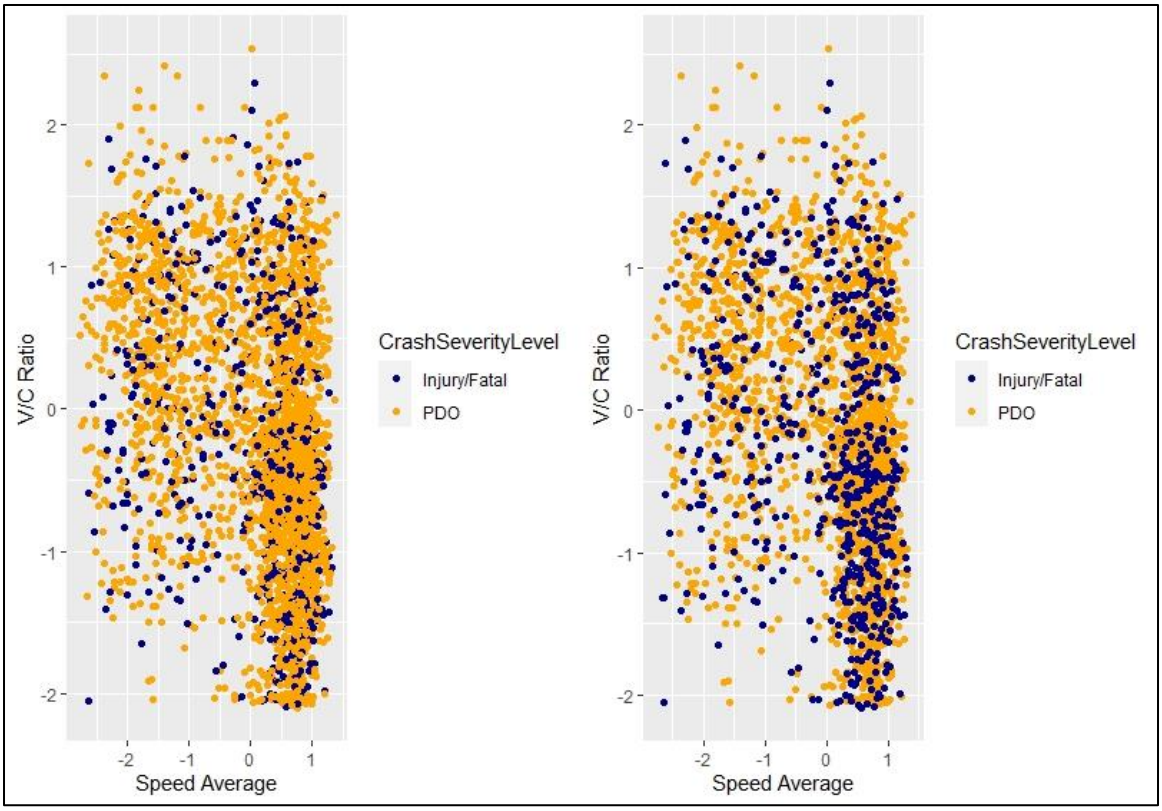


Figure 4.6 Average speed vs. v/c ratio in the crash injury severity dataset: before ROSE (left) and after ROSE (right).

The number of crash records (features) in the training dataset for each class before and after the ROSE transformation, as well as the size of each class in the testing dataset are summarized in Table 4.4.

Table 4.3 Size of Input Datasets for the Crash Severity Models

Models / Corresponding Classes	Training Dataset		Testing Dataset
	Before ROSE	After ROSE	
Crash Severity Dataset	7616	12719	2359
PDO Crashes	6009	6614	2003
Injury/Fatal Crashes	1607	6105	536

4.1.4 Model Tuning and Application

The random effect Bayesian Logistic Regression models were calibrated in WINBUGS statistical software. All fixed and random effect parameters are set to follow non-informative priors. The fixed-effect variables are assumed to be normally distributed as $\beta \sim Normal(0, 0.000001)$ where the first parameter is the mean and the second parameter is the precision (the reciprocal of the variance), so the variance is one million. The random effect variable is also set to have a normal distribution as $u_j \sim N(0, t)$, where the precision parameter t has a gamma prior with Gamma distribution as $t \sim Gamma(0.001, 0.001)$ so that the mean is 1 and the variance is 1000.

Full Bayesian inference was employed based on the Markov Chain Monte Carlo (MCMC) simulation. Unlike the previous crash risk analysis studies which did not give initial values to the variables, this study employs an ordinary logistic regression to assign the initial values to the variables. 20,000 iterations are set up and the first 5,000 samples are considered as burn-in. To consider the explanatory variable as significant, 95% Bayesian Credible Interval (BCI) should be reached (Gelman, 2003). The explanatory variable is statistically significant if zero is not included in the range of 95% credible interval of the coefficient (Lunn et al., 2012). To evaluate the Bayesian models, deviance information criteria (DIC) are one of the factors utilized for model complexity and fit. The DIC measures goodness-of-fit in the model corresponding to the negative likelihood of the model as well as a penalty term corresponding to the number of coefficients. DIC's penalty term is measured by the deviation between the expected log-likelihood and the log-likelihood at the posterior mean point. The Bayesian logistic model with smaller values of DIC is preferable. In this project, the random-effect Bayesian logistic regression models of

crash severity and crash likelihood are estimated separately. The models are fitted on the training datasets and were then evaluated on the test dataset to derive the performance metrics.

The ML models were implemented in R statistical software using CARET package version 6.0-86 (M. Kuhn et al., 2020) and the DL model was executed in R using h2o package version 3.20.0.8 (LeDell et al., 2018). A 10-fold cross validation was performed for all models to evaluate their performance. In addition, the preprocessing step included centering and scaling of all the continuous variables used in the models.

In developing and tuning the ML and DL models, several parameters (referred to as hyperparameters) are considered and calibrated for the RF, GBM, KNN, and MLFDNN models. The set of tuning parameters that were found to yield the highest AUC value for the RF, GBM, KNN, and MLFDNN models are summarized in Table 4.5.

Table 4.5 Summary of the Hyperparameters for the RF, GBM, KNN, and MLFDNN Models

Model	**Hyperparameters for the crash likelihood analysis**	**Hyperparameters for the crash injury severity analysis**
RF	*mtry* = 4	*mtry* = 16
	split.rule = gini	*split.rule* = gini
	node.size = 1	*node.size* = 1
	sample.size = full training set	*sample.size* = full training set
GBM	*n.trees* = 50	*n.trees* = 50
	interaction.depth = 1	*interaction.depth* = 1
	shrinkage = 0.1	*shrinkage* = 0.1
	n.minobsinnode = 10	*n.minobsinnode* = 10
KNN	*K* = 13	*K* = 5
MLFDNN	*epochs* = 15	*epochs* = 15
	hidden.layer1 = 50	*hidden.layer1* = 50
	hidden.layer2 = 50	*hidden.layer2* = 50

In the RF model, after preparing the training data, the OOB sample and 10-fold cross-validation based experimental design were used separately, to determine the optimal hyperparameters for the RF. Similar results were achieved through OOB error minimization and cross-validation. For the crash likelihood analysis, both approaches found that the combination of *mtry* = 4, *split.rule* = gini, *node.size* = 1, and *sample.size* = full training set, yield the model with the lowest OOB error and highest AUC value. Using a similar approach for the injury severity analysis, the parameters *mtry* = 16, *split. rule* = gini, *node.size* = 1, and *sample.size* = full training set, were found to yield the best result in terms of the AUC value.

In the GBM model, an important factor is the selection of the number of trees. Finding the optimal number of trees (*n.trees*) is a challenging task: larger number of trees contributes to good learning, while it might also increase the risk of overfitting (Opitz & Maclin, 1999). The size of the trees is another parameter which is indicated by *interaction.depth* in the R model and accounts for the order of predictor-to-predictor interaction captured in the model (Hastie et al., 2009). The learning rate or *shrinkage* is another hyperparameter pertaining to GBM, which determines the effect of each tree on the output result and takes values between 0 and 1. Overall, lower learning rates provide better results by adding more trees to the iteration (Friedman, 2001). Finally, the parameter *n.minobsinnode* defines the minimum number of observations allowed per node. In general, larger values of *n.minobsinnode* generate smaller trees that are less impacted by noise. Using a 10-fold cross-validation, the set of parameters *n.trees* = 50, *interaction.depth* = 1, *shrinkage* = 0.1, and *n.minobsinnode* = 10 was found to yield the result with the highest AUC value for the crash likelihood analysis. In the crash injury severity model, the

set of parameters *n.trees* = 250, *interaction.depth* = 5, *shrinkage* = 0.1, and *n.minobsinnode* = 10 was found to return the best model in terms of the AUC value.

To tune the KNN model, one should find the optimal number of neighbors (*K*). The 10-fold cross-validation results showed that *K*= 13 and *K*= 5 produced the model with the highest AUC value in the crash likelihood and the crash injury severity analysis, respectively.

Lastly, to tune MLFDNN, one should find the number of iterations (*epochs*), the number of hidden layers and the number of neurons at each hidden layer. Reducing the number of training epochs contributes to the mitigation of the overfitting problem (Panchal, Ganatra, Shah, & Panchal, 2011). There is no well-defined approach for choosing the number of hidden layers and the number of nodes within them. In general, adding more layers and nodes increases the opportunity for new features to be learned during model training. In this study, the result of the 10-fold cross-validation showed that *epochs* = 15, employing two hidden layers, *hidden.layer 1 = 50*, and *hidden.layer 2 = 50*, returns the best model in terms of AUC for the crash likelihood model. The same number of layers and nodes with *epochs* = 12 was found to give the best result in the crash severity analysis.

4.2 Case Study of the NDS Dataset with Driver Behavior Explanatory Variables

The SHRP2 NDS includes 3,500 driving participants from New York, Washington, Florida, Indiana, Pennsylvania, and North Carolina. An NDS data subset resulted from a data reduction processed performed by the Virginia Tech Transportation Institute (VTTI) was used in this study. The VTTI, which is the custodian and publisher of the NDS dataset,

has developed reduced or aggregated files to provide easier access to the NDS data (Precht, Keinath, & Krems, 2017).

4.2.1 Discussion of the Data Inputs

Prior to model development, the dataset was reduced and cleaned to exclude any biases that might affect the crash risk estimates. The data cleaning included removing the records with missing or unknown values and merging the categories for some of the factors. Also, due to the limited number of crashes in the dataset, near crash events were combined with the crash events to create a single variable named "crash or near-crash event, or "CNC". This classification resulted in 8,136 CNC event records and 18,909 baseline event records in the final dataset. It should be noted that using near-crashes as a surrogate measure and combining them with crash events has been adopted by many researchers in the literature.

Similarly, to overcome the problem of low frequency of cases belonging to severe crash categories, the "most severe" crashes were combined with records in the "Police-reportable" class to create a more balanced sample for the crash severity analysis. Table 4.6 provides the events statistics along with crash events that were counted and classified in terms of severity.

For each event, data were extracted for 19 different variables. Table 4.7 provides a summary of basic statistics for the binary/categorical variables used in the analysis. The only continuous variable in the analysis dataset is *SecondaryDur* (duration of the secondary task performed by driver).

Table 4.6 Summary of Response Variables for NDS Crash Risk Analysis

Characteristic	Total
Number of crash events	1,724
Number of near-crash events	6,412
Number of baseline events	18,909
Number of police-reportable crashes	260
Number of minor crashes	1,464

Table 4.7 Summary of Basic Statistics for the Binary/Categorical Variables

Variable	Description	Number	Percentage
ManeuverJudgment	Safe and legal	23,724	87.72
	Unsafe and illegal	1,809	6.69
	Unsafe but legal	781	2.89
	Safe but illegal	731	2.70
Behavior	Normal	21,446	79.30
	Distracted/Drowsy/Fatigued	2,115	7.82
	Risky driving	1,120	4.14
	Sign/Signal violation	1,037	3.83
	Speed violation	860	3.18
	Inattention	302	1.12
	Aggressive	90	0.33
	Avoiding	75	0.28
Impairment	No impairment	26,375	97.52
	Drowsy/Fatigued	426	1.58
	Emotional state	213	0.79
	Alcohol/Drug	31	0.11
SecondaryTask1	None	12,254	45.31
	Activity oriented	8,152	30.14
	Object oriented	4,026	14.89
	Cellphone oriented	2,440	9.02
	Other	173	0.64
Seatbelt	Yes	25,675	94.93
	No	1,370	5.07
AgeGroup	(16-19)	4,015	14.85
	(20-24)	6,419	23.73
	(25-49)	7,429	27.47
	(50-69)	5,094	18.84
	(+70)	4,088	15.12
Educ	High school	11,919	44.07
	College degree	10,323	38.17
	Graduate degree	4,723	17.46
Male	Yes	13,440	49.69
	No	13,605	50.31

Table 4.7 Summary of Basic Statistics for the Binary/Categorical Variables (Continued)

Variable	Description	Number	Percentage
Lighting	Daylight	21,019	77.72
	Darkness lighted	3,534	13.07
	Darkness not lighted	1,418	5.24
	Dusk	727	2.69
	Dawn	347	1.28
surfaceCondition	Dry	22,538	83.34
	Wet	4,234	15.66
	Snowy/Icy	273	1.01
traddicDensity	LOS A1	9,197	34.01
	LOS A2	6,969	25.77
	LOS B	7,958	29.43
	LOS C	1,897	7.01
	LOS D	692	2.56
	LOS E	281	1.04
	LOS F	51	0.19
intersectionInfluence	No junction	18,474	68.31
	Traffic signal	3,252	12.02
	Interchange/Intersection	1,572	5.81
	Parking/Driveway entrance	1,379	5.10
	Uncontrolled	1,068	3.95
	Stop sign	969	3.58
	Yes, other	331	1.22
Curve	No	23,150	85.60
	Yes	3,895	14.40
grade	Level	22,747	84.11
	Grade up	2,817	10.42
	Grade down	1,481	5.48
WorkZone	No	25,807	95.42
	Yes	1,238	4.58
vehClass	Car	19,545	72.27
	SUV/Crossover	5,210	19.26
	Pickup/Truck	1,376	5.09
	Van/Minivan	914	3.38
Adv.Tech	No	25,371	93.81
	Yes	1,674	6.19
Int.Cell	No	21,705	80.26
	Yes	5,340	19.74

4.2.2 Preparation of the Training and Testing Datasets

Similar to the real-time crash risk analysis, it is also important to check for correlation between the decision variables in the NDS sample dataset. Figure 4.7 provides a graphical representation of the correlation matrix created, using Pearson correlation. Based on the results, no significant correlation (>0.7) was found among the explanatory variables.

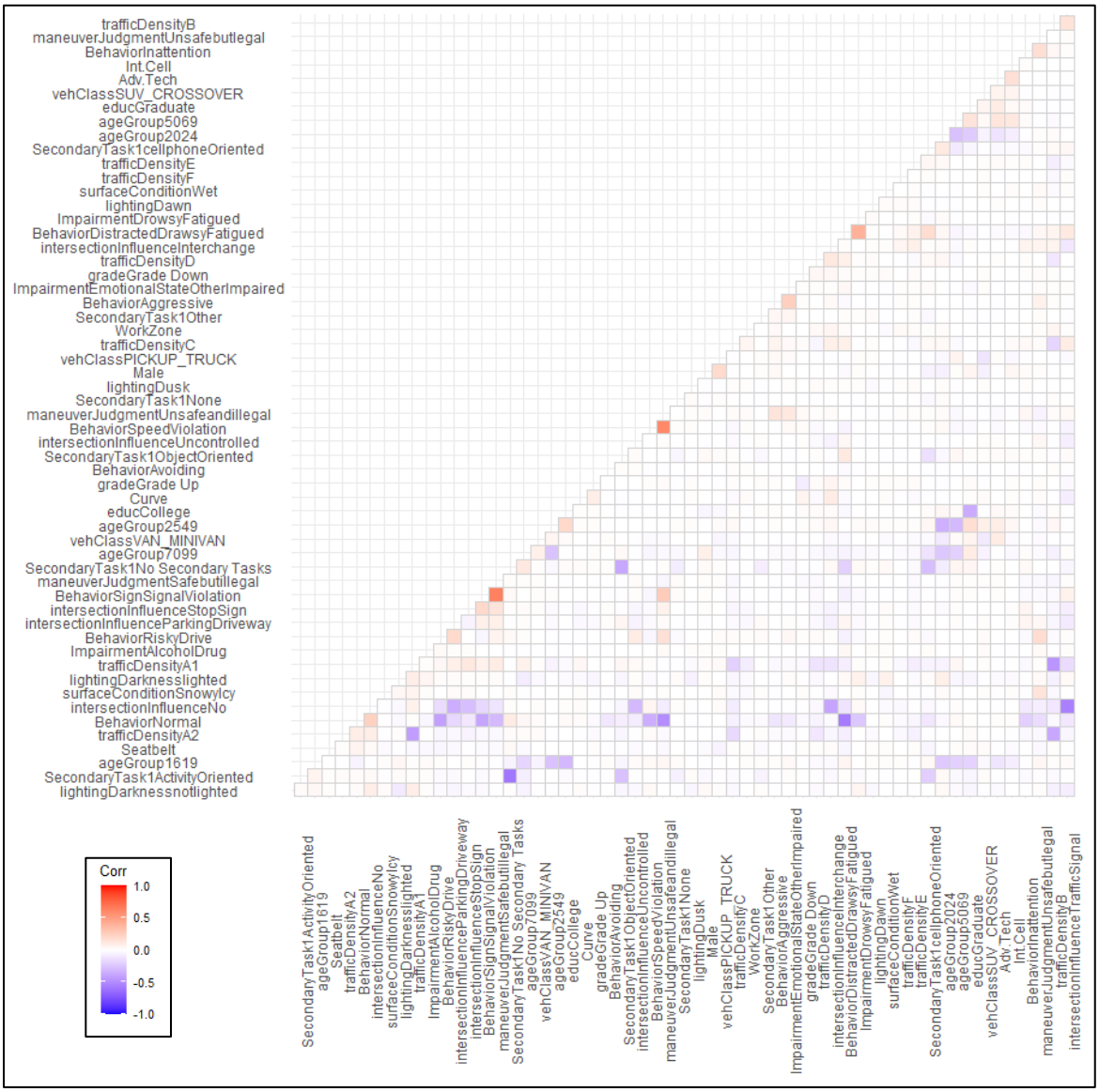


Figure 4.7 Correlation matrix for the crash likelihood analysis dataset.

After checking for the correlations, the next step was to divide the full data set into training and testing set using the same proportions as in 4.1.5. All models are fitted on the training datasets and were then evaluated using the test dataset to derive the performance metrics.

4.2.3 Model Tuning and Application

The fixed effect Bayesian Logistic Regression models were calibrated in R statistical software using *rjags* package (Plummer, Stukalov, Denwood, & Plummer, 2019). All parameters are assumed to be normally distributed as $\beta \sim Normal(0, 0.000001)$ which indicates that they follow non-informative priors. Full Bayesian inference was employed based on the Markov Chain Monte Carlo (MCMC) simulation. This study employs an ordinary logistic regression to assign the initial values to the variables. 20,000 iterations are set up and the first 5,000 samples are considered as burn-in. Also, to consider the explanatory variable as significant, 95% Bayesian Credible Interval (BCI) should be reached. The explanatory variable is statistically significant if zero is not included in the range of 95% credible interval of the coefficient.

The ML models were implemented in R statistical software using *CARET* package and the DL model was executed in R using *h2o* package. A 10-fold cross validation was performed for all models to evaluate their performance. In developing and tuning the machine learning models, several parameters (referred to as hyperparameters) are considered and calibrated for the RF, GBM and MLFDNN models. The set of tuning parameters that were found to yield the highest AUC value for the models using the NDS data are summarized in Table 4.8.

Table 4.8 Summary of the Hyperparameters for the RF, GBM, and MLFDNN

Model	**Hyperparameters for the crash likelihood analysis**	**Hyperparameters for the crash injury severity analysis**
RF	*mtry* = 6	*mtry* = 5
	split.rule = gini	*split.rule* = gini
	node.size = 4	*node.size* = 4
	sample.size = full training set	*sample.size* = full training set
GBM	*n.trees* = 104	*n.trees* = 153
	interaction.depth = 3	*interaction.depth* = 5
	shrinkage = 0.3	*shrinkage* = 0.01
	n.minobsinnode = 15	*n.minobsinnode* = 10
	bag.fraction = 0.8	*bag.fraction* = 0.65
MLFDNN	*epochs* = 15	*epochs* = 15
	hidden.layer1 = 50	*hidden.layer1* = 50
	hidden.layer2 = 50	*hidden.layer2* = 50

CHAPTER 5

DISCUSSION OF THE MODEL RESULTS

5.1 Real-time Crash Likelihood Model

The outputs of the BLR model estimation are summarized in Table 5.1. The standard deviation of speed (Speed_sd_1015), average speed (speed_avg_1015), hourly precipitation and visibility (HourlyPrecipitation, HourlyVisibility, respectively) and v/c ratio (v_ratio) are found to be significant at the 95% Bayesian credible interval (BCI). As shown in the table, hourly precipitation, average speed, and standard deviation of speed have positive correlation to the crash occurrence, while v/c ratio and hourly visibility have negative correlation to the crash occurrence. The Bayesian model has the deviance information criterion (DIC) of 8578.467, and AUC of 0.67. The DIC value is lower than the null model, indicating that explanatory variables improve the model fit.

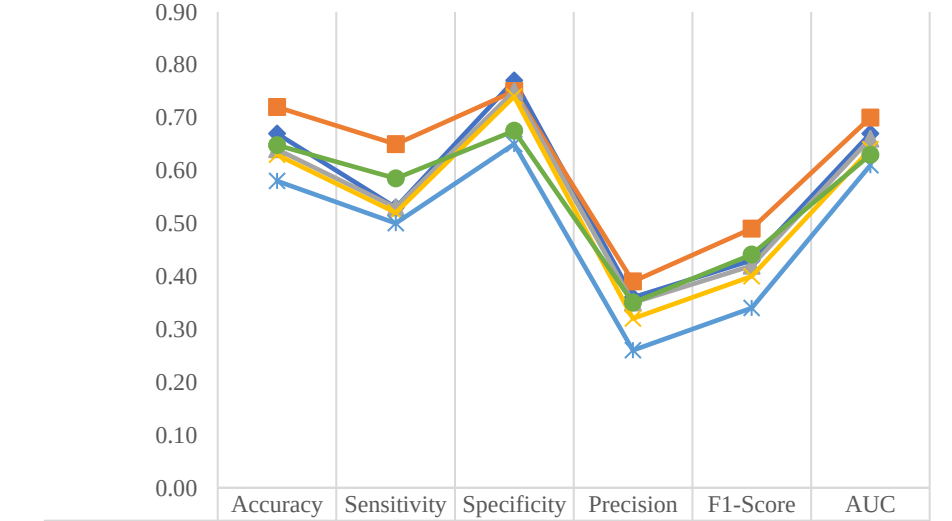
Table 5.1 Summary of the Random Effect BLR Model for Real-time Crash Likelihood

Variables	Mean	Std. Err	95% BCI
speed_sd_1015	0.069	0.022	(0.028, 0.111)
speed_avg_1015	0.32	0.024	(0.415, 0.295)
HourlyPrecipitation	0.125	0.033	(0.082, 0.179)
HourlyVisibility	-0.118	0.026	(-0.167, -0.071)
V_ratio	-0.138	0.027	(-0.192, -0.080)
Constant	-0.148	0.026	(-0.206, 0.103)

The performance statistics for the BLR, RF, GBM, NB, KNN, and MLFDNN models in terms of the overall accuracy, sensitivity, specificity, and the AUC values is

summarized in Figure 5.1. It should be noted that lager values for all metrics indicate better performance of the models.

In terms of specificity, which reflects the ability of the models to correctly predict non-crash cases, the BLR has the highest value of 0.77, while the values for GBM and RF are just slightly lower. The lowest specificity has the KNN model (0.65), which balances the sensitivity value (0.50). Therefore, most of the models tend to favor majority class (non-crash cases) over the minority class (crash cases). Overall, RF appears to demonstrate the best performance of all tested models. It has the highest overall accuracy, sensitivity, precision, F1-score, and AUC value and the specificity is comparable to slightly higher value achieved by the BLR.



	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
BLR	0.67	0.53	0.77	0.36	0.43	0.67
RF	0.72	0.65	0.75	0.39	0.49	0.70
GBM	0.64	0.53	0.75	0.35	0.42	0.66
GNB	0.63	0.52	0.74	0.32	0.40	0.64
KNN	0.58	0.50	0.65	0.26	0.34	0.61
MLFDNN	0.65	0.59	0.68	0.35	0.44	0.63

Figure 5.1 I80/I-287 crash likelihood model results.

Nevertheless, even that performance of the RF model can be categorized as "unsatisfactory" as it correctly predicts only 65% of crash occurrences in the testing set. With such performance, this model would not be applicable in practice as it would not be sufficiently effective in detecting the conditions that (may) lead to crashes.

5.2 Real-time Crash Severity Model

A random effects BLR model was calibrated for the real-time prediction of crash severity, based on the dataset for I-20 and I-287 in New Jersey. The summary of the output statistics for calibrated BLR crash severity model is provided in Table 5.2. The results show that an increase in the average speed, hourly visibility, and the existence of sun glare result in an increase in crash severity. The DIC and AUC values of this model are equal to 4850.40 and 0.59, respectively. Compared to the null model, the DIC value of this model is lower, which means that the explanatory variables help the model fit.

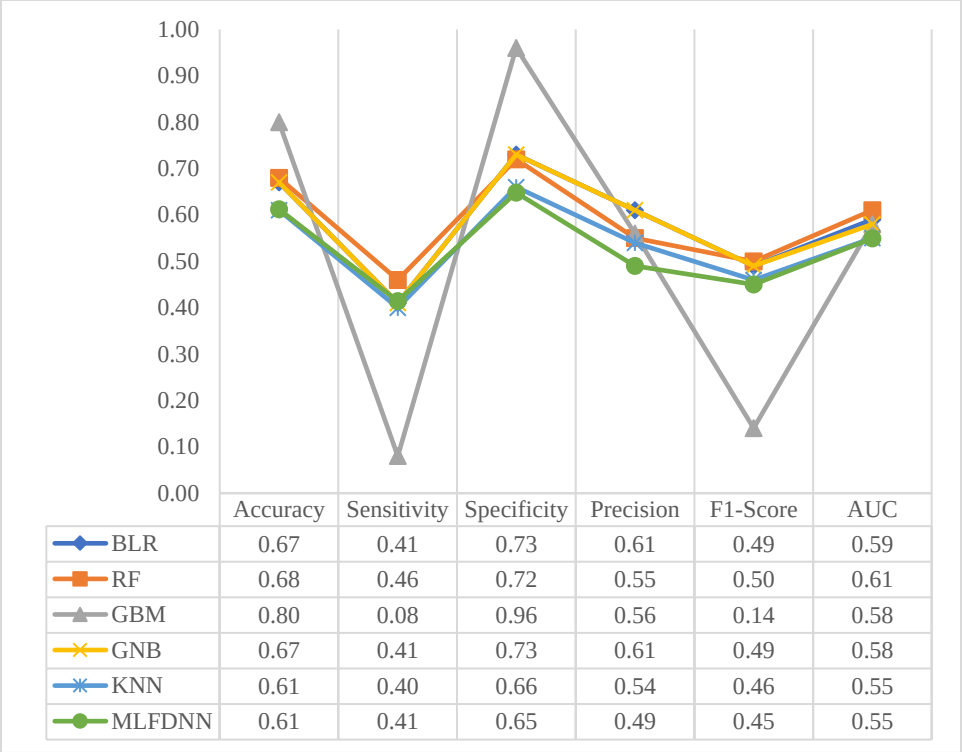
Table 5.2 Summary of the Random Effect BLR Model for Real-time Crash Severity

Variables	Mean	Std. Err	95% BCI
Speed_avg_1015	0.2	0.08	(0.05, 0.4)
HourlyVisibility	0.02	0.009	(0.001, 0.03)
Sunglare	0.01	0.006	(0.005, 0.02)
Constant	0.02	0.001	(0.004, 0.04)

The performance metrics for the BLR, RF, GBM, GNB, KNN, and MLFDNN models for the crash injury severity analysis is summarized in Figure 5.2. It can be observed that the AUC values range between 0.55 and 0.61, with RF having the highest AUC value. In relation to sensitivity, which indicated the capability of the models to correctly predict

the injury crashes, RF has the highest sensitivity value (0.46), followed by GNB, BLR, and MLFDNN which all provide the similar sensitivity value of 0.41. In fact, BLR and GNB have identical performance across the overall accuracy (0.67), sensitivity (0.41), specificity (0.73), precision (0.61), and F1-score (0.49). In terms of specificity, which reflects the ability of the models to correctly predict PDO cases, GBM provides the highest values (0.96), followed by GNB and BLR (0.73), and RF (0.72). MLFDNN also provides the lowest specificity value (0.65) among all investigated models. It is noteworthy that despite the high specificity value achieved by GBM, it cannot be recommended for predicting the severity of crashes as it provides the lowest sensitivity value among all investigated models.

Overall, as in the analysis of crash likelihood, RF appears to demonstrate the best performance of all tested models. It has the highest AUC value, and the overall accuracy is the second highest after GBM. Nevertheless, looking at the sensitivity values, the overall performance of all the models can be categorized as “weak” and it can be concluded that none of the models is adequate in terms of predicting the crash severity.



	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
BLR	0.67	0.41	0.73	0.61	0.49	0.59
RF	0.68	0.46	0.72	0.55	0.50	0.61
GBM	0.80	0.08	0.96	0.56	0.14	0.58
GNB	0.67	0.41	0.73	0.61	0.49	0.58
KNN	0.61	0.40	0.66	0.54	0.46	0.55
MLFDNN	0.61	0.41	0.65	0.49	0.45	0.55

Figure 5.2 I-80/I-287 case study crash severity model results.

5.3 Real-time Combined Driver Severity Model

The low sensitivity and AUC values obtained from the crash severity models triggered the idea of adding the reactive crash data as inputs, in addition to the proactive data. The reactive data had been widely used in developing metrics for crash severity analyses. However, despite the critical impact of the factors described by reactive data on the crash severity outcomes, the main challenge of using the reactive data for operational crash prediction is their unavailability in real-time. To overcome this problem, crash records were analyzed by dividing them in groups considering the age of the drivers and vehicles. Table 5.3 provides a summary of the driver age and vehicle age characteristics of the crash groupings, along with the number of crash records in the case study data set in each group. It should be noted that as one aims to investigate the impact of driver age and vehicle age

on crash severity in the combined study, the driver injury severity level should be considered as the dependent variables rather than the crash injury severity. The case study dataset contained 12,566 driver records, with 11,059 (88%) records of non-injury cases and 1,507 (12%) records of injury cases.

Table 5.3 Summary Statistics of Crash Records Considering Driver and Vehicle Age

Group #	Variable	n	%
1	DrAge[1] < 25 & VehAge[2] < 5	1096	8.72
2	DrAge < 25 & 5 ≤ VehAge < 10	627	4.99
3	DrAge < 25 & 10 ≤ VehAge	716	5.69
4	25 ≤ DrAge < 70 & VehAge < 5	5927	47.17
5	25 ≤ DrAge < 70 & 5 ≤ VehAge < 10	1991	15.84
6	25 ≤ DrAge < 70 & 10 ≤ VehAge	1832	14.57
7	70 ≤ DrAge & VehAge < 5	198	1.58
8	70 ≤ DrAge & 5 ≤ VehAge < 10	89	0.71
9	70 ≤ DrAge & 10 ≤ VehAge	90	0.72

1: DrAge = driver age; 2: VehAge = vehicle age.

The RF method was used to predict the driver injury severity for each age group. The RF method was selected over the other models as it outperformed other investigated models in the initial assessment. The performance statistics for the RF models considering the driver and vehicle age is summarized in Table 5.4.

Table 5.4 Results of the Driver Injury Severity RF Model for Each Driver-Vehicle Age Group

Group #	Accuracy	Sensitivity	Specificity	AUC
1	0.61	0.60	0.61	0.62
2	0.58	0.66	0.55	0.63
3	0.62	0.55	0.63	0.66
4	0.68	0.54	0.79	0.68
5	0.69	0.52	0.77	0.64
6	0.63	0.55	0.72	0.64
7	0.68	0.42	0.89	0.66
8	0.65	0.52	0.76	0.66
9	0.62	0.52	0.67	0.61
Average	0.64	0.54	0.71	0.64

It can be observed from the result that the average AUC value incased to 0.64, a 4-percentage point increase from the crash severity model that did not account for driver and vehicle age. Similar improvement was achieved in terms of model sensitivity, which increased by 8-percentage points, from 0.46 in the crash severity model, to 0.54 in the driver severity model. The overall performance of the model can be categorized as “weak”, with limited practical applicability for crash severity prediction.

It should be states that the information about individual drivers and vehicles is not known in real time. In fact, as noted earlier, this information is only known after the crashes occur about the drivers and vehicles participating in reported crashes. Nevertheless, it is possible to predict the driver injury severity outcome if relative shares of drivers by age and vehicles by age can be estimated in a total driver population and vehicle fleet respectively for a given road or an analysis area. Having the estimated share of each group of drivers (e.g., by age) and vehicles (by age) travelling on a road segment, the probability of a crash having a certain driver injury severity outcome along that segment can be calculated using the law of total probability:

$$P_i(S_j) = \sum_k P_i(G_k)\, P_i(S_j|G_k) \tag{5.1}$$

where $P_i(S_j)$ is the probability that a crash on segment *i* will result in driver injury severity outcome *j*, $P_i(G_k)$ is the proportion of drivers and vehicles belonging to group *k* ($\sum_k P_i(G_k) = 1$), and $P_i(S_j|G_k)$ is the conditional probability of driver injury severity outcome *j* for group *k*.

Combined, the crash likelihood prediction model and the crash injury severity prediction model can be applied to estimate the probability of crash and the expected severity of a crash (if/when the crash occurs) at a given roadway segment with a given set of roadway, traffic, and environmental characteristics, and the assumed (estimated) composition of drivers by age and vehicles by age. Nevertheless, despite the improved performance when including driver age and vehicle age, the overall accuracy and predictive power of the resulting models is found to be relatively poor and must be further improved to be used for a meaningful operational crash risk prediction.

5.4 Crash Risk with Driver Behavior Explanatory Variable

It can be concluded from the results obtained with previous models that the input dataset is lacking and does not provide sufficient "information" for either statistical or machine learning models to successfully predict crash likelihood or severity in real-time, operational context. Based on the previous research, the missing information is likely related to driver behavior factor. However, it is understood that such data is not readily available, especially not in real time and with a sufficient coverage and sample to provide reliable source for operational analysis.

Nevertheless, despite the general unavailability of driver behavior data for operational analysis, using such data available to researchers for crash modeling provides an unprecedented opportunity to shed light on how factors such as driver behavior can influence the risk of crash occurrence and severity. One such dataset is the NDS dataset. Knowing that NDS data is obtained using vehicles instrumented with advanced onboard data acquisition systems (DAS), it is not yet possible to collect such data for all or majority of vehicles in a network. Thus, unlike the data collected for the I-80 and I-287 in New Jersey, NDS data cannot be applied to develop real-time operational crash risk models. However, as more vehicle manufacturers are installing vehicle telemetry and driver monitoring sensors in the newer vehicle models, it is quite plausible that the data similar to NDS would soon become more available, and at some point, available in the prevailing share of vehicles operating on the roads. Furthermore, besides detailed information on driver behavior and other driver-related factors, NDS dataset contains additional information such as presence of work zone, intersections, and roadway geometry (e.g., roadway alignment, presences of curve, and grade), not only for CNC events, but also for baseline conditions (i.e., normal driving). For these reasons, additional prediction models were developed using the NDS dataset to demonstrate if and how the inclusion of driver behavior factors would improve the accuracy and predictive power of the crash prediction models.

It should be mentioned that based on the results obtained from the initial set of models for the I-80/I-287 case study, it was decided to exclude the KNN and GNB methods from consideration in the NDS crash risk analysis, due to their poor performance. Also, as the NDS data does not support calculation of the crash risk at the roadway segment level,

the random-effect Bayesian Logistic Regression was replaced with the normal Bayesian Logistics Regression model by excluding the random parameter.

5.4.1 Crash Likelihood

The estimation of the BLR model is summarized in Table 5.5. The table provides the summary for all explanatory variables significant at the 95% Bayesian credible interval (BCI). The explanatory variables that were found not to be statistically significant include:

- lighting dawn,
- lighting dusk,
- seatbelt usage,
- wet road surface,
- LOS F,
- gender,
- education level,
- activity-oriented secondary task,
- drivers older than 50,
- driving SUV/crossover,
- driving pickup/truck,
- vehicle advanced technology, and
- integrated cellphone system.

According to the calculated odd ratios, driver behavior indicator (*Behavior*) has the highest influence on increasing the risk of CNC occurrence, followed by intersection influence (*intersection influence*) and maneuver judgment (*maneuver judgment*). Based on

the sample, the Odds Ratios suggest that distracted and drowsy driving presents a risk of CNC event 411.44 times higher than normal driving. Other risky driving behaviors are inattention and risky actions during driving, which increase the CNC risk by 64.99 and 23.70 times, respectively. In addition, the impact of a parking lot or driveway entrance/exit has an odds ratio of 10.45, indicating the positive correlation between this factor and the probability of a CNC event. Another influencing factor in this regard is the level of service (LOS). It is found that the odds ratio of LOS D is the highest at 10.12, indicating that the drivers are more likely to be involved in a CNC events when high density exists, but the flow is stable. Unsafe and/or illegal maneuver judgments are also found to increase the odds of CNC event, with the corresponding odds ratios higher than 8.00, relative to safe and legal maneuvers. In terms of the impact of driver age, drivers belonging to the 16-19 age group are more at risk of being involved in the CNC events. Finally, horizonal curves, work zones, lighting condition, snow/icy road surface, roadway grade, engagement in a secondary task, and secondary task duration, all increase the odds of a CNC event.

Table 5.5 Summary of the BLR Model for CNC Likelihood (NDS Data) (Continued)

Variable	Mean	SD	References	95% BCI	Odds ratio
Constant	-3.073	0.126	-	(-3.316,-2.810)	0.05
Behavior (Aggressive)	2.875	0.871	vs. normal	(1.403,4.730)	13.69
Behavior (Avoiding other vehicles/objects)	1.400	0.330	vs. normal	(0.776,2.050)	4.05
Behavior (Distracted/Drowsy/Fatigued)	6.092	0.325	vs. normal	(5.492,6.745)	411.44
Behavior (Inattention)	4.244	0.391	vs. normal	(3.471,4.986)	64.99
Behavior (Risky action)	3.192	0.141	vs. normal	(2.913,3.466)	23.70
Behavior (Sign/Signal violation)	0.712	0.186	vs. normal	(0.325,1.061)	2.02
Behavior (Speed violation)	-0.709	0.185	vs. normal	(-1.054,-0.313)	0.50
Impairment (Drowsy/Fatigued)	-5.465	0.353	vs. no impairment	(-6.184,-4.830)	0.10
Impairment (Emotional state)	0.893	0.270	vs. no impairment	(0.355,1.400)	2.43
Safe but illegal maneuver	-1.469	0.199	vs. safe and legal	(-1.871,-1.075)	0.23
Unsafe and illegal maneuver	2.091	0.170	vs. safe and legal	(1.770,2.438)	8.00
Unsafe but legal maneuver	2.320	0.143	vs. safe and legal	(2.044,2.610)	10.03
Cellphone oriented secondary task	-0.215	0.088	vs. no secondary task	(-0.391,-0.044)	0.81
Object oriented secondary task	0.513	0.058	vs. no secondary task	(0.401,0.626)	1.67
Dark but lighted	0.235	0.064	vs. daylight	(0.115,0.360)	1.26
Dark and unlighted	0.321	0.098	vs. daylight	(0.127,0.512)	1.38
Snowy/Icy surface condition	1.193	0.181	vs. dry	(0.840,1.558)	3.27
LOS A2	0.167	0.063	vs. LOS A1	(0.041,0.290)	1.18
LOS B	1.186	0.057	vs. LOS A1	(1.072,1.292)	3.24
LOS C	2.023	0.080	vs. LOS A1	(1.868,2.179)	7.47
LOS D	2.327	0.127	vs. LOS A1	(2.075,2.570)	10.12
LOS E	1.488	0.195	vs. LOS A1	(1.121,1.884)	4.39
Presence of curve	0.216	0.058	vs. straight	(0.109,0.332)	1.24
Grade (Down)	0.743	0.085	vs. level	(0.584,0.914)	2.10
Grade (Up)	0.509	0.065	vs. level	(0.377,0.630)	1.66
Presence of Work zone	0.372	0.089	vs. non-work-zone	(0.194,0.548)	1.45

Table 5.5 (Continued) Summary of the BLR Model for CNC Likelihood (NDS Data)

Variable	Mean	SD	References	95% BCI	Odds ratio
Intersection influence (Intersection/Interchange)	1.341	0.073	vs. No influence	(1.197,1.480)	3.81
Intersection influence (Parking/Driveway)	2.356	0.087	vs. No influence	(2.193,2.530)	10.45
Intersection influence (Stop Sign)	0.888	0.114	vs. No influence	(0.655,1.099)	2.43
Intersection influence (Traffic Signal)	1.126	0.062	vs. No influence	(1.001,1.242)	3.07
Intersection influence (Uncontrolled intersection)	2.224	0.094	vs. No influence	(2.037,2.405)	9.17
Driver age (16-19)	0.382	0.077	vs. 25-49	(0.232,0.530)	1.46
Driver age (20-24)	0.220	0.061	vs. 25-49	(0.104,0.340)	1.24
Van/Minivan	-0.640	0.140	vs. car	(-0.912,-0.367)	0.53
Secondary task duration	0.048	0.023	-	(0.003,0.093)	1.05

The performance statistics for the evaluated models (BLR, GBM, RF, and MLFDNN) in terms of the overall accuracy, sensitivity, specificity, precision, F1-score, and the AUC values, is summarized in Figure 5.3. The results reveal that all models perform relatively well, and markedly better than the models excluding the driver behavior characteristics. However, GBM outperforms all candidate models, followed by RF and MLFDNN. The overall accuracy, sensitivity, specificity, precision, F1-score, and AUC values achieved by the GBM model are highest at 86.5%, 83.4%, 87.7%, 73.7%, 0.783, and 0.934, respectively.

Furthermore, compared to the initially presented real-time crash risk models, a significant improvement is achieved in the model's performance where the AUC value obtained by the GBM model shows a 23.4 percentage point increase from what was obtained in the RF crash likelihood model developed in the case study using the New Jersey roadway and traffic data without consideration of driving behavior factors. Considering the

model performance measures, the overall performance of the presented GBM model can be characterized as “very strong”.

	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
BLR	0.826	0.771	0.850	0.697	0.732	0.880
GBM	0.865	0.834	0.877	0.737	0.783	0.934
RF	0.851	0.827	0.860	0.709	0.764	0.918
MLFDNN	0.812	0.798	0.818	0.656	0.720	0.886

Figure 5.3 CNC likelihood model performance summary (NDS Data).

In general, there can be two potential reasons for this significant improvement: first, the inclusion of driver behavior and driver-related factors; and second, the completely different nature of the NDS dataset compared to the dataset utilized in the real-time crash risk models for the I-80/I-287 case study. To address the uncertainty related to the difference between the datasets, an additional model was developed based on NDS data, this time using only the variables that can be obtained in real-time, and comparable to the variables used in the initial crash likelihood and crash severity models calibrated for the New Jersey case study. The list of the variables used in the model development included

lighting condition, surface condition, traffic density (which can be computed from the speed and flow rate), intersection influence, roadway alignment, and presence of work zone. The GBM model was used for the comparison as it outperformed other investigated models in the previous NDS analysis. Table 5.6 presents a side-by-side comparison of the model's performance indicators. It is clear from the comparison that while using the NDS dataset improves the model performance in general, this improvement is much higher after adding driver behavior and driver-related factors. The comparison reveals that using the full NDS dataset (with driver behavior variables) improves the AUC value by more than 34%. Around 21% of this improvement can be explained by driver behavior and driver-related factor, and the remaining 13% can be associated with the difference between the input datasets.

Table 5.6 Comparison of Model Performance Indicators

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
Real-time crash likelihood model (RF)	0.722	0.651	0.749	0.394	0.491	0.704
CNC likelihood model without driver behavior variables (GBM)	0.752	0.674	0.781	0.564	0.614	0.789
CNC likelihood model with driver behavior variables (GBM)	0.865	0.834	0.877	0.738	0.783	0.934

This conclusion can be elaborated by identifying the important variables in a detection process, wherein the variables are ranked based on their relative influence in developing the GBM model (Figure 5.4). During this process, which is very similar to the RF variable importance ranking discussed in Subsection 3.4.4, the relative importance of the variables is determined by a variable's average relative influence across all trees

generated by the GBM algorithm (Friedman, 2001; Ridgeway, 2007). The variable importance is scaled on a scale of 0 to 100, where a higher number represents higher importance. As can be seen in the figure, only the intersection influence (62.83%) and traffic density (33.70%) are found to be significant predictors. This implies that the prediction performance of the real-time models can be significantly improved by adding these two factors to the input dataset, especially at locations other than freeways (where the impact of traffic controls and intersections is minimal or none). Also, the real-time dataset having more variables compared to the NDS dataset, the real-time models' underperformance can be mainly attributed to the existence of noise in the input data. The potential causes of this noise are the use of synthesized hourly volumes rather than the actual real-time volumes, as well as the crash time reporting error.

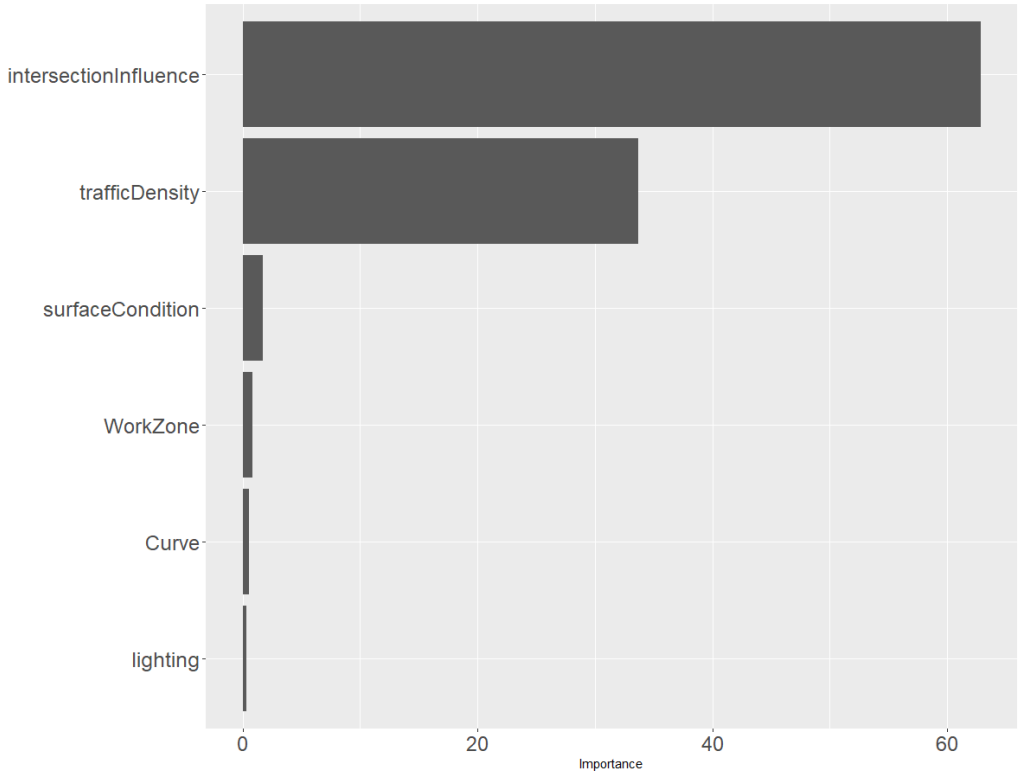


Figure 5.4 GBM variable importance plot for CNC likelihood with selected variables.

A similar approach was undertaken to detect the important variables for the full model (Figure 5.5). The results of the variable importance analysis demonstrate that driver behavior, secondary task duration, intersection influence, traffic density, maneuver judgement, and impairment are the most influential factors to CNC occurrence, accounting for 39.66%, 28.69%, 12.03%, 7.67%, 5.87%, and 3.05% of the GBM model's detection accuracy, respectively. On the other hand, surprisingly, roadway geometry characteristics (i.e., grade and curvature), driver characteristics (i.e., age, gender, and education), and environmental conditions (i.e., lighting and surface condition) are found to have no significant influence on model's accuracy.

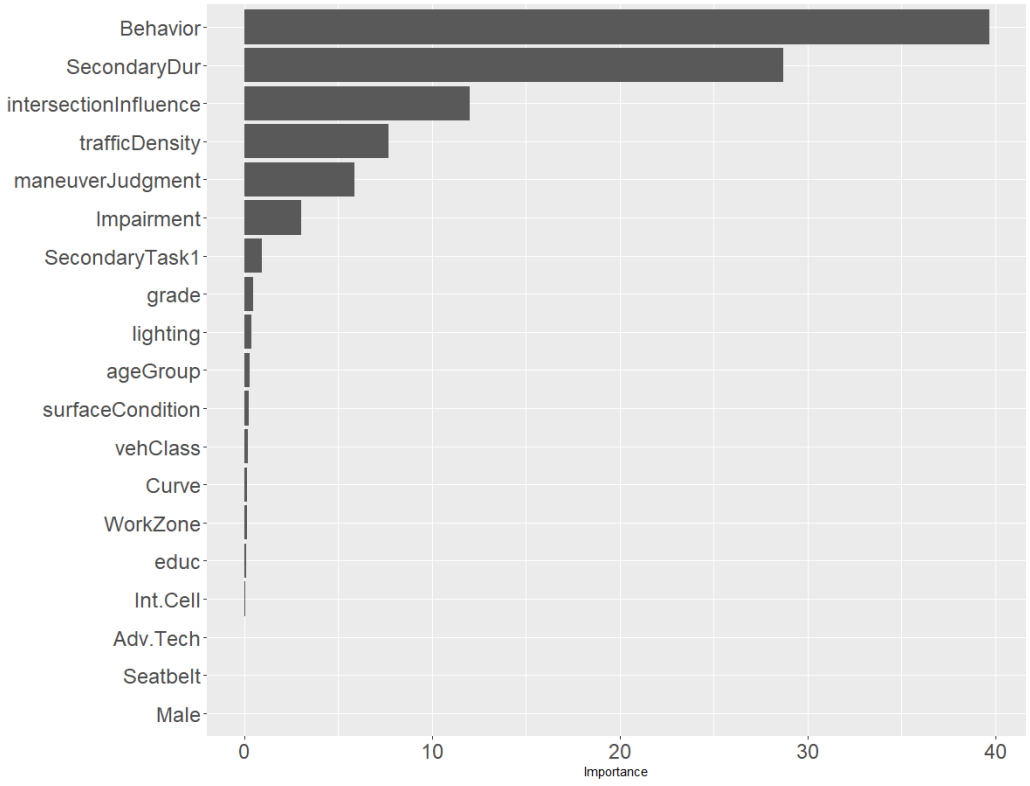


Figure 5.5 GBM variable importance plot for CNC likelihood with all variables.

5.4.2 Crash Severity

The estimation of the BLR model for the crash severity prediction using the NDS data is summarized in Table 5.7. Based on the model results, traffic density and driver behavior have the highest impact on the increased probability of a severe crash (classified as a police-reportable crash in the NDS dataset). According to the odd ratios, traffic density is the most contributing factor to police-reportable crashes, where LOS E, LOS D, LOS C, LOS B, and LOS A2 all increase the odd of police-reportable crash compared to LOS A1. Among the driver behavior factors, inattention driving has the highest influence with a risk of 13.76 times higher than normal driving. Engagement in cellphone oriented secondary task, wet road surface, presence of stop sign or traffic signal, uncontrolled intersection, and male driver are also found to increase the odds of police-reportable crashes. On the other hand, college education, and driving pickup/truck or SUV/crossover reduce the severity of crashes.

Table 5.7 Summary of the BLR Model for Crash Severity (NDS Data)

Variable	Mean	SD	References	95% BCI	Odds ratio
Constant	-3.147	0.557	-	(-4.15,-2.07)	0.04
Behavior (Inattention)	2.652	0.809	vs. normal	(1.08,4.26)	13.76
Behavior (Risky action)	2.367	0.385	vs. normal	(3.13,1.63)	0.1
Cellphone oriented secondary task	0.703	0.311	vs. no secondary task	(0.07,0.29)	2.02
Wet surface condition	0.621	0.237	vs. dry surface	(0.16,1.08)	1.87
LOS A2	0.742	0.304	vs. LOS A1	(0.13,1.32)	2.11
LOS B	1.766	0.259	vs. LOS A1	(1.24,2.26)	5.81
LOS C	2.398	0.385	vs. LOS A1	(1.63,3.13)	11
LOS D	2.567	0.662	vs. LOS A1	(1.32,3.93)	12.94
LOS E	4.052	1.056	vs. LOS A1	(1.99,6.09)	54.44
Intersection influence (Stop Sign)	0.753	0.406	vs. No influence	(0.01,1.59)	2.13
Intersection influence (Traffic Signal)	0.980	0.268	vs. No influence	(0.44,1.50)	2.65
Intersection influence (Uncontrolled intersection)	0.945	0.375	vs. No influence	(0.19,1.66)	2.58
Male driver	0.437	0.202	vs. female	(0.06,0.84)	1.54
College education	-0.615	0.265	vs. high school	(-1.12,-0.08)	0.54
Pickup/Truck	-1.149	0.620	vs. car	(-2.40,-0.01)	0.33
SUV/Crossover	-0.820	0.280	vs. car	(-1.34,-0.29)	0.44

The performance statistics for the investigated models are provided in Figure 5.6. At the first glance, one can note that GBM outperforms other candidate models in terms of all performance metrics, followed by the BLR and RF. Also, MLFDNN has the lowest performance among the investigated models. Compared to the metrics obtained from the crash likelihood models for the NDS dataset, the values of precisions and F1-scores are lower for all investigated models illustrated in Figure 5.6. This can be explained by the very low number of police-reportable crashes in the dataset. It should be noted that in crash likelihood and severity analysis it is more important to correctly detect the positive cases (crashes in the crash likelihood analysis and severe/police-reportable crashes in the crash

severity analysis). Based on the definition, precision is implied as the number of correctly predicted positive cases (police-reportable crashes in the crash severity model) from all predicted positive cases while sensitivity is the number of correctly predicted positive cases from all the actual positive cases. Therefore, sensitivity being more important than precision in evaluating the models' performance, the overall performance of BLR, GBM, and RF are satisfactory.

	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
BLR	0.802	0.722	0.814	0.357	0.478	0.850
GBM	0.810	0.809	0.810	0.421	0.554	0.866
RF	0.794	0.722	0.803	0.345	0.467	0.813
MLFDNN	0.738	0.698	0.744	0.319	0.438	0.788

Figure 5.6 Crash severity model results (NDS Data).

It is also noteworthy that, as in the CNC likelihood analysis, using NDS data improves the predictive performance of injury severity models, significantly. This can be concluded by looking at the AUC values, where the newly obtained value (0.866) is 37% higher than the AUC value obtained in the real-time combined driver severity model (0.641).

The graphical representation of the GBM-based variable importance ranking for the crash severity dataset is presented in Figure 5.7. Overall, the results confirm the findings of the BLR model, where traffic density, driver behavior, intersection influence, secondary task duration, and surface condition are the most significant variables, accounting for 36.36%, 32.89%, 28.69%, 11.09%, 4.48%, and 4.11% of the GBM model's detection accuracy, respectively.

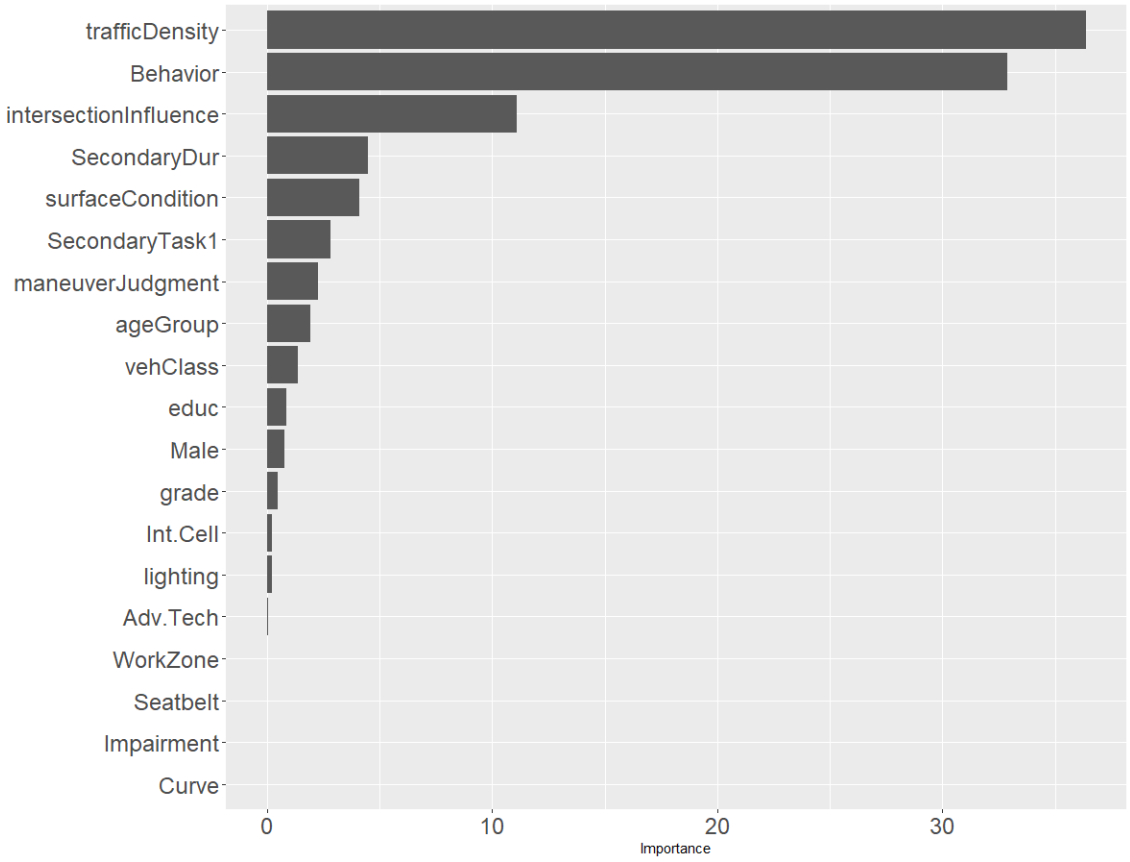


Figure 5.7 GBM variable importance plot for crash severity (NDS data).

5.4.3 Impact of Different Factors on Driver Behavior Fault

Based on the findings of this study, the driver behavior has the highest impact on the likelihood of CNC occurrence. Therefore, a separate analysis was conducted to examine the correlation between driver behavior fault and other factors. A clear understanding of the impact of other factors on driver behavior would help to evaluate and select the most effective countermeasures to reducing the number of crashes. In addition, it is hard to constantly watch and assess driver behavior; instead, more attention should be given to the attempts towards controlling the factors that contribute to driver behavior fault. To serve this objective, a GBM-based model was developed where the response variable was either 0, when the driver acted normal and 1, when the driver acted faultily.

The developed model had an overall good performance in terms of all metrics, where the overall accuracy, sensitivity, specificity, and AUC were found to be 71%, 65%, 72%, and 0.73, respectively. The result of the variable importance ranking also indicate that intersection influence (44.73%) and driver impairment (42.2%) have the highest impact on driver behavior fault (Figure 5.8).

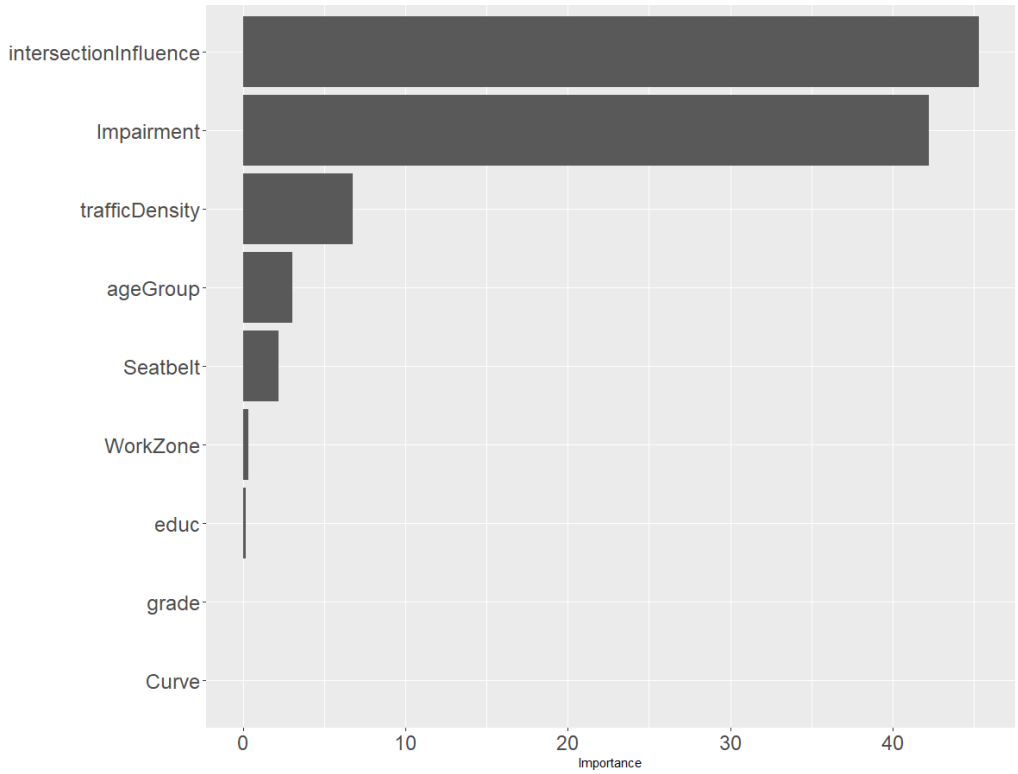


Figure 5.8 GBM variable importance plot for driver behavior analysis.

5.5 Practical Implications of Demonstrated Models

The outcomes of this research can be implemented in designing an operational traffic safety management system that can predict the relative short-term (e.g., next 5-15 minutes) crash risk for all regional roadways at the roadway segment level. Also, to account for the relatively poor model performance, it is suggested to use relative crash risks instead of absolute probability values, for operational purposes. To this end, the probability of having a crash and its associated injury severity level is calculated for each road segment. These values are then clustered into multiple groups based on pre-defined thresholds to represent the relative risk of crash. To exemplify, the following values can be used to categorize the crash risk at a road segment level:

$$\begin{cases} Low\ risk & if \quad P_i \leq 0.3 \\ Moderate\ risk & if \quad 0.3 < P_i \leq 0.6 \\ High\ risk & if \quad 0.6 < P_i \leq 0.75 \\ Extremely\ high\ risk & if \quad 0.75 < P_i \end{cases} \tag{5.1}$$

Besides, to facilitate monitoring the roadway's safety condition in real-time, a map-based system in which the road segments are colored/labeled based on their associated relative crash risks is proposed to be developed (Figure 5.9). This is expected to help the traffic operations management authorities to take proactive traffic management strategies such as utilizing variable speed limit, variable message signs, and coordinated warning signals to mitigate crash risks for high-risk locations.

Figure 5.9 Real-time crash risk map-based system.

The results of the NDS data analysis indicated that driver behavior and driver distraction are the most significant contributing factors to crashes. Therefore, as it is not yet possible to track and monitor driver behavior and distraction directly, encouraging more drivers to use mobile apps or devices which are able to collect data from the vehicle telematics system is suggested as a potential solution to reduce the likelihood and severity of crashes. These apps or devices should emphasize on safe driving and also be capable of notifying the drivers when they are speeding or apply harsh braking to be penalized.

To sum up, the results of the study showed satisfactory performance of the models in predicting crash risk, when including driver behavior and driver-related variables. This

hints that focusing on driver-related factors instead of controlling for speed and volume may yield a higher return on investment, in terms of reduction in number and severity of crashes, as well as indirect negative effects of traffic crashes.

CHAPTER 6

CONCLUSION, RESEARCH CONTRIBUTION, AND FUTURE RESEARCH

6.1 Conclusion

This main goal of this study was to apply advanced data analytics methods to develop and evaluate crash severity and crash likelihood prediction models that can be used in near-real time. For this purpose, the models were built using the data that is available to regional transportation agencies in real time and provides coverage of all major highway facilities on a reginal or statewide scale. The dataset applied in the study consisted of data collected for two interstate highways in New Jersey – I-80 and I-287 and included detailed crash data from the New Jersey State DOT crash records database, basic roadway geometry data, synthetic vehicle volume and capacity data, probe-vehicle traffic speed data, and weather data from the National Weather Service. All data is available in real time and is provided on a roadway segment level, which range in length between 0.02 miles and 5.14 miles. The crash records dataset consisted of 10,155 crashes, including 2,139 crashes with an injury or fatal outcome, and 8,016 PDO crashes. For the crash likelihood model additional records were created to represent non-crash cases following the matched case–control methodology. To deal with the data imbalance between the crash cases and non-crash cases in the crash likelihood model, as well as between PDO and injury/fatality crashes in the crash severity model the study employed the random oversampling examples (ROSE) method. The relative importance of explanatory variables was evaluated using RF model and they were ranked based on mean decrease accuracy.

The BLR model further revealed (or rather confirmed) the significance of each explanatory variable in both crash likelihood and crash injury severity analyses. The crash likelihood model had five significant explanatory variables, including the standard deviation of speed 5-15 minutes preceding the crash, average speed 5-15 minutes prior to the crash, hourly precipitation, hourly visibility, and v/c ratio. On the other hand, the crash severity model had three significant explanatory variables, two of which were the same as in the crash likelihood model (average speed 5-15 minutes prior to the crash and hourly visibility), and sun glare as an additional significant factor.

The Odds Ratios were calculated for all explanatory variables and showed that while hourly precipitation, average speed, and standard deviation of speed increase the odds of crash occurrence; v/c ratio and hourly visibility were found to reduce the chance of crash involvement. Also, speed average, hourly visibility, and sun glare were found to increase the odds of injury crashes. In addition to the BLR model, five additional machine learning (ML) and Deep learning (DL) methods were implemented for crash likelihood and crash severity prediction. A 10-fold cross-validation method was applied for tuning all ML and DL models, which produced optimal combination of the hyperparameters for each model, as applicable. The prediction accuracy of all models was evaluated using the performance metrics including the overall accuracy, sensitivity, specificity, and the AUC value. The estimation results for the crash likelihood and crash severity models revealed that the RF model outperformed all the other investigated models in terms almost of all performance metrics. In conclusion, even the best performing model of crash likelihood and crash severity could be characterized as having limited predictive value based on the performance metrics.

The results of the analysis hint that the data used in this study is not sufficient or sufficiently informative to enable satisfactory separation of crash outcome and severity classes in the crash dataset. In that sense, and considering results of numerous previous studies and literature, the present study tried to bridge the gap between the use of proactive and reactive crash factors by developing a combined model that includes the data reflecting both proactive and reactive factors. The study examined the potential improvement in the predictive performance of the injury severity models by incorporating the reactive data on driver age and vehicle age. This was achieved through implementation of a modeling framework that evaluated injury severities for different crash groupings based on the age of drivers and age of vehicles. The results indicated that while inclusion of driver age and vehicle age can improve the predictive performance of the severity model, the results were still far from satisfactory.

To tackle this problem, this dissertation developed additional models using NDS data, which includes driver behavior indicators, to identify the most important risk factors contributing to crash/near-crash (CNC) events. To this end, different statistical, ML, and DL models were developed to find the linear and non-linear correlations between a large set of explanatory variables and CNC occurrence. Among the candidate models, GBM was found to be superior in terms of almost all performance metrics, indicating the model's ability to correctly classify the data into CNC and baseline events based on pre-event variables. The results from the BLR and GBM models confirmed driver behavior to be the most critical factor to CNC occurrence. Also, among all types of driver behavior, distracted and drowsy driving was found to have the highest CNC risk. Developing the real-time driver monitoring systems that are capable of providing reliable feedback to drivers when

apparent signs of distraction and drowsiness are detected, can be proposed as a solution to reduce the influence of this factor on the increased risk of crash. Furthermore, the variable importance analysis by the GBM model showed intersection influence and traffic density to be among the most significant risk factors to CNC events, emerging the need for enhanced safety by legislators and transportation agencies at the risk-prone locations. These treatments might include improving the roadway designs in term of geometry and operational indices and applying stricter policies in the areas with high CNC risk.

Similar analysis was also carried out to find the key contributing factors to different crash severity levels where the results from the BLR and GBM models indicated that traffic density, driver behavior, and intersection influence are the most important contributing factors to more severe crashes. The GBM model was found to have the best predicting performance among all investigated models.

It was obvious from the findings that a clearer understanding of driver behavior's role in the occurrence of CNC events would help to analyze and implement pre-crash safety measures and develop enforcement policies, infrastructure design, and advanced vehicle safety systems. This would not be possible without investigating the influence of other factors on driver behavior fault. The presented study developed a GBM model to explore the impact of factors contributing to faulty driver behavior. The results of the model demonstrated that intersection influence and impaired driving have the highest impact on driver behavior fault, meaning that a driver is more at risk of acting irregularly when impaired or approaching an intersection/interchange.

At the outset of the study, the aim was at developing models that would allow the transportation agencies and decision makers to assess the crash likelihood and anticipated

severity of crashes in near-real-time, using the data already available to them. That in turn would allow them to make more effective operational decisions and implement operational countermeasures and tactics to reduce the likelihood and severity of crashes. Some examples would include proactive activation of advanced warnings on variable message sign (VMS), adjustments of variable speed limits (VSL) and ramp metering (RM), as well as tactical deployment of highway safety patrols and other traffic operations and management assets. Similarly, benefiting from the NDS data, this study provides the knowledge required by transportation agencies and decision makers to find the important risk factors in order to properly allocate funds to safety programs. The availability of microscopic information collected by well-instrumented vehicles on real-time driving behavior and instantaneous decisions of drivers via NDS has enabled investigation of the correlation between driver behavior and crash risk. The results showed that while it is hard to control for driver behavior, which was found to have the highest impact on CNC risk, it is expected that crash frequency and its associated injury severity would be reduced significantly by adding advanced safety features to more vehicles on the roads. For the crash prediction capability, advent of connected and automated vehicles will be critical, as those vehicles will likely have the capability of providing data on driver behavior and vehicle telemetry, similar to the NDS data.

6.2 Research Contributions

The main contributions of this dissertation are listed as follows:

1. Developed a modeling framework that would allow the transportation agencies and decision makers to assess the crash likelihood and anticipated severity of crashes in near-real-time, using the data available at the highway-network scale. Despite the relatively weak performance of the models developed in the case

study of New Jersey roadways, the modeling framework can be readily applied in determining the relative crash risk at a network scale. Findings of the analysis using the Naturalistic Driving Study (NDS) data suggest that adding driver behavior and driver-related factors in the same modeling framework significantly improve the models' performance for operational purposes. The datasets that reflect these factors, similar to the NDS dataset, are expected to become available in real-time or near-real-time with the advancements in vehicle technologies and the advent of connected and autonomous vehicles in near future.

2. Demonstrated the use of Random Over-Sampling Examples (ROSE) method to deal with the data imbalance problem. This method has not been previously demonstrated in literature in crash risk modeling applications. The presented application of ROSE in crash severity prediction models was found to greatly improve their sensitivity.

3. Demonstrated the application of a combined model that includes the data reflecting both proactive and reactive crash factors, including driver age and vehicle age. The model can be readily applied in operational analysis with the inclusion of inferred or statistically representative reactive factor data. The factors such as driver age, education, vehicle type and age, can be estimated based on the statistics of driver and vehicle composition in a given analysis area, a highway corridor, or roadway segment, depending on the availability of driver population and vehicle usage data.

4. Compared the predictive performance of various statistical, machine learning and deep learning methods under a validation framework to explore the linear and non-linear relationship between a large set of contributing factors and crash risk, in terms of both likelihood and severity. The presented analysis demonstrated varying performance of crash likelihood and crash severity models developed using different modeling methods and different datasets. As such, this study contributed to filling the methodological gap and added to the current knowledge by comparing this large set of models. In particular, the analysis demonstrated and quantified critical improvement of predictive performance when including driver behavior data in the prediction models. While the driver behavior data used in this study is not currently available in real- or near-real-time, or with sufficient spatial and temporal coverage, the presented analysis demonstrates the methodology of including such data in crash prediction models and determining the relative influence of different driving behavior factors.

5. The presented research can be readily applied in traffic management information systems to identify roadway segments with relatively high crash risk based on the available data provided as model inputs. The presented models can be implemented in map-based computer applications to visualize crash risks on roadway segments. They can also be used in traffic safety decision support

systems (DSS) to assist the traffic management authorities proactively manage the traffic safety risks using active traffic management strategies and tools. It is expected that such crash risk visualization and DSS tools will gain more practical value with greater availability of vehicle telemetry data and driver behavior data provided from connected and autonomous probe vehicles.

6.3 Future Studies

The present study will provide models to quantify the relative risks of a crash at a given highway segment and the expected injury severity level when the crash occurs. The proposed models will provide a basis for further research in crash prediction, considering the emerging datasets, such as driving behavior records collected by the UBI systems, which are already being offered to be used by some insurance companies. In addition, NDS and connected vehicles (CVs) sharing similar data format, the increasing adoption of CV technology and the share of CVs on the roads will provide an opportunity to utilize driver behavior data in near real time. This can significantly improve the performance of the crash risk prediction models in the future, using the modeling framework presented in this study.

Last but not least, driver behavior was found to be the most critical risk factor in both crash likelihood and crash severity models, clearly indicating that the attention should be given to the monitoring of driver's behavior to reduce the crashes. There are some recent attempts in this regard. As an example, to increase safety for commercial vehicles, automakers have partnered with NHTSA to launch the Driver Alcohol Detection System for Safety (DADSS) program. The goal of this public-private partnership is to develop novel technology to passively detect drivers with a Blood Alcohol Concentration (BAC) over the legal limit. This technology, once completed, will be licensed to anyone for commercial applications (Alliance for Automotive Innovation, 2021).

As another example, The ATTENTION ASSIST system developed by Mercedes-Benz measures more than 70 parameters that are analyzed to detect fatigue. This continual monitoring is important for recognizing the gradual transition from alertness to tiredness, providing sufficient time to initiate timely alerts and warnings to the driver. Based on a variety of data, ATTENTION ASSIST creates an individual driver profile during the first few minutes of each journey and compares this with sensor data and the driving situation as recognized by the vehicle's electronic control unit. Alongside values such as speed and longitudinal/lateral acceleration, the Mercedes system also measures indicator and pedal usage, for example, as well as specific operations and external factors such as crosswinds and the unevenness of the roads. If the system detects drowsiness, it emits an audible warning signal and flashes up an unequivocal message on the display in the instrument cluster: "ATTENTION ASSIST. Break!" (Euro NCAP Advanced, 2021). Similar systems are also planned to be deployed by other car manufacturers. In 2019, Volvo announced that it will fit all its new cars with driver-facing cameras and other sensors to detect distracted driving. Volvo claimed that this system not only can alert the driver, but also have the power to reduce the car’s speed, or even slow down, park and call the assistance service (Volvo Cars, 2019). Another interesting technology offered by Nuance can use voice commands and eye movement to control certain vehicle systems. For example, the driver can close the window or ask information about a point of interest outside of the car and be assisted by the car’s advanced voice assistance system. Lastly, Bosch is using cameras to switch between human and different levels of computer drive. This system decides to control the vehicle if driver distraction is identified.

Despite the commercial availability, there are still serious concerns about the effectiveness of these new technologies and more importantly the fully automated driving systems. As a result, the Alliance for Automotive Innovation, which represents auto suppliers and manufacturers producing nearly 99% of new cars and light trucks sold in the U.S., has recently released several safety principles such as adoption of camera-based driving monitoring systems in vehicles with automated driving or driver-assist systems. It is believed that similar models as developed in this study can help such organizations to assess the effectiveness of advanced technologies by performing different analyses with before-and-after experimental designs.

Considering these advancements and vehicle technologies already on the market, alongside the advent of connected and autonomous mobility, there will be more data similar to that provided by the NDS, which will allow the transportation safety researchers to perform more accurate crash risk analyses. When such data become available in real time or near real time, the modeling framework presented in this study can be further refined to predict unsafe conditions in near-real-time considering the variety of risk factors, including driver behavior factors. Finally, according to the findings of this study, it is expected that the deployment of automated vehicles will potentially results in a massive decline in crash frequency. But even then there will be other factors, such as computational errors or computer system failures, which may become a new focus of concern and risk factor in assessing the crash risks.

REFERENCES

Abdelwahab, H. T., & Abdel-Aty, M. A. (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record, 1746*(1), 6-13.

Agramunt, S. (2018). *Naturalistic driving behaviour and self-regulation practices among older drivers with bilateral cataract: a prospective cohort study* [Doctoral dissertation, Curtin University, Perth, Australia].

Ahmed, M. M., Abdel-Aty, M., Lee, J., & Yu, R. (2014). Real-time assessment of fog-related crashes using airport weather data: A feasibility analysis. *Accident Analysis & Prevention, 72*, 309-317.

Ahmed, M. M., Abdel-Aty, M., & Yu, R. (2012). Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. *Transportation Research Record, 2280*(1), 60-67.

Ahmed, M. M., & Abdel-Aty, M. A. (2011). The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on Intelligent Transportation Systems, 13*(2), 459-468.

Alliance for Automotive Innovation. (2021). *Working to prevent impaired driving*. Retrieved June 15, 2021, from https://www.autosinnovate.org/initiatives/safety/impaired-driving.

Antin, J. F., Lee, S., Perez, M. A., Dingus, T. A., Hankey, J. M., & Brach, A. (2019). Second strategic highway research program naturalistic driving study methods. *Safety Science, 119*, 2-10.

Arvin, R., & Khattak, A. J. (2020). Driving impairments and duration of distractions: Assessing crash risk by harnessing microscopic naturalistic driving data. *Accident Analysis & Prevention, 146*, 105733.

Bakhit, P. R., Guo, B., & Ishak, S. (2018). Crash and near-crash risk assessment of distracted driving and engagement in secondary tasks: a naturalistic driving study. *Transportation Research Record, 2672*(38), 245-254.

Bärgman, J. (2016). *Methods for analysis of naturalistic driving data in driver behavior research* [Doctoral dissertation, Chalmers University of Technology, Gothenburg, Sweden].

Bharadwaj, N., Edara, P., & Sun, C. (2019). Risk factors in work zone safety events: a naturalistic driving study analysis. *Transportation Research Record, 2673*(1), 379-387.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin, Germany: Springer.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations* (Vol. 18). Oxford, England: Oxford University Press.

Breiman, L. (2000). *Some infinity theory for predictor ensembles* (Technical Report. Report No. 579). Statistics Dept. University of California, Berkeley, California.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks.

Chang, A., Saunier, N., & Laureshyn, A. (2017). *Proactive methods for road safety analysis* (Technical Paper. Report No. WP-0005). Society of Automotive Engineers International.

Chang, Y., Bharadwaj, N., Edara, P., & Sun, C. (2020). Exploring Contributing Factors of Hazardous Events in Construction Zones Using Naturalistic Driving Study Data. *IEEE Transactions on Intelligent Vehicles, 5*(3), 519-527.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention, 90*, 128-139.

Chen, C., Zhang, G., Yang, J., & Milton, J. C. (2016). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident Analysis & Prevention, 90*, 95-107.

Cigdem, A., & Ozden, C. (2018). Predicting the severity of motor vehicle accident injuries in Adana-turkey using machine learning methods and detailed meteorological data. *International Journal of Intelligent Systems and Applications in Engineering, 6*(1), 72-79.

Darban Khales, S., Kunt, M. M., & Dimitrijevic, B. Analysis of the impacts of risk factors on teenage and older driver injury severity using random-parameter ordered probit. *Canadian Journal of Civil Engineering, 47*(11), 1249-1257.

Dei, M., (2019). *11 evaluation metrics data scientists should be familiar with lessons from a high rank Kagglers' new book*. Retrieved June 18, 2021, from https://towardsdatascience.com/11-evaluation-metrics-data-scientists-should-be-familiar-with-lessons-from-a-high-rank-kagglers-8596f75e58a7.

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention, 38*(3), 434-444.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Garg, A., & Tai, K. (2012, June). Comparison of regression analysis, artificial neural network and genetic programming in handling the multicollinearity problem. In *2012 Proceedings of International Conference on Modelling, Identification and Control* (pp. 353-358) IEEE, Wuhan, China.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, *71*(2), 369-382.

Guo, F., Klauer, S. G., Fang, Y., Hankey, J. M., Antin, J. F., Perez, M. A., . . . & Dingus, T. A. (2017). The effects of age on crash risk associated with driver distraction. *International Journal of Epidemiology*, *46*(1), 258-265.

Haleem, K., & Gan, A. (2013). Effect of driver's age and side of impact on crash severity along urban freeways: A mixed logit approach. *Journal of Safety Research, 46*, 67-76.

Hankey, J. M., Perez, M. A., & McClafferty, J. A. (2016). *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets* (Technical Report). Virginia Tech Transportation Institute.

Harmon, T., Bahar, G. B., & Gross, F. B. (2018). *Crash costs for highway safety analysis* (Technical Report. Report No. FHWA-SA-17-071). U.S. Department of Transportation Federal Highway Administration.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, New York: Springer.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention, 108*, 27-36.

Kim, J.-K., Ulfarsson, G. F., Kim, S., & Shankar, V. N. (2013). Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention, 50*, 1073-1081.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . & Benesty, M. (2020). Package 'caret'. *The R Journal*, 223.

Kuhn, S., Egert, B., Neumann, S., & Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics, 9*(1), 400.

Kwak, H.-C., & Kho, S. (2016). Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. *Accident Analysis & Prevention, 88*, 9-19.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., . . . & Kurka, M. (2021). *Package 'h2o'. dim*. https://cran.r-project.org/web/packages/h2o/h2o.pdf.

Li, X., Cai, B. Y., Qiu, W., Zhao, J., & Ratti, C. (2019). A novel method for predicting and mapping the occurrence of sun glare using Google Street View. *Transportation Research Part C: Emerging Technologies, 106*, 132-144.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book, a practical introduction to bayesian analysis*. Boca Raton, Florida: CRC Press.

McLaughlin, S. B., & Hankey, J. M. (2015). *Naturalistic driving study: linking the study data to the roadway information database* (SHRP 2 Report. Report No. S2-S31-RW-3). Transportation Research Board of the National Academies.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*(1), 92-122.

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis & Prevention, 40*(1), 260-266.

Mousa, S. R., Bakhit, P. R., & Ishak, S. (2019). An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: a naturalistic driving study. *Canadian Journal of Civil Engineering, 46*(8), 712-721.

National Center for Statistics and Analysis. (2020, October). *Preview of motor vehicle traffic fatalities in 2019* (Research Note. Report No. DOT HS 813 021). National Highway Traffic Safety Administration.

Nicodemus, K. K. (2011). Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics, 12*(4), 369-373.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169-198.

Osman, O. A., Hajij, M., Bakhit, P. R., & Ishak, S. (2019). Prediction of near-crashes from observed vehicle kinematics using machine learning. *Transportation Research Record, 2673*(12), 463-473.

Panchal, G., Ganatra, A., Shah, P., & Panchal, D. (2011). Determination of over-learning and over-fitting problem in back propagation neural network. *International Journal on Soft Computing, 2*(2), 40-51.

Plummer, M., Stukalov, A., Denwood, M., & Plummer, M. M. (2019). *Package 'rjags'*. https://cran.r-project.org/web/packages/rjags/rjags.pdf.

Precht, L., Keinath, A., & Krems, J. F. (2017). Effects of driving anger on driver behavior–Results from naturalistic driving data. *Transportation Research Part F: Traffic Psychology and Behaviour, 45*, 75-92.

Reiman, T., & Pietikäinen, E. (2012). Leading indicators of system safety–monitoring and driving the organizational safety potential. *Safety Science, 50*(10), 1993-2000.

Sarkar, S., Pramanik, A., Maiti, J., & Reniers, G. (2020). Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. *Safety Science, 125*, 104616.

Saunier, N., & Sayed, T. (2006, June). A feature-based tracking algorithm for vehicles in intersections. In *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)* (pp. 59-59) IEEE, Quebec, Canada.

Shanthi, S., & Ramani, R. G. (2011). Classification of vehicle collision patterns in road accidents using data mining algorithms. *International Journal of Computer Applications, 35*(12), 30-37.

Singh, S. (2018). *Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey* (Traffic Safety Facts Crash· Stats. Report No. DOT HS 812 506). National Highway Traffic Safety Administration.

Stafford, B. (2018). *Pysolar documentation*. https://buildmedia.readthedocs.org/media/pdf/pysolar/latest/pysolar.pdf.

Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems, 39*(1), 43-62.

Theofilatos, A. (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. *Journal of Safety Research, 61*, 9-21.

Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation Research Record, 2673*(8), 169-178.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability, 57*, 1-436.

Treat, J. R., Tumbas, N., McDonald, S., Shinar, D., Hume, R. D., Mayer, R., . . . Castellan, N. (1979). *Tri-level study of the causes of traffic accidents: final report. Executive summary.* (Technical Report. Report No. DOT HS 805 099). Indiana University, Bloomington, Institute for Research in Public Safety.

Van Schagen, I., & Sagberg, F. (2012). The potential benefits of naturalistic driving for road safety research: Theoretical and empirical considerations and challenges for the future. *Procedia-Social and Behavioral Sciences, 48*, 692-701.

Volvo Cars. (2019). *Volvo cars to deploy in-car cameras and intervention against intoxication, distraction*. Retrieved June 18, 2021, from https://www.media.volvocars.com/global/en-b/media/pressreleases/250015/volvo-cars-to-deploy-in-car-cameras-and-intervention-against-intoxication-distraction.

Wang, L., Shi, Q., & Abdel-Aty, M. (2015). Predicting crashes on expressway ramps with real-time traffic and weather data. *Transportation Research Record, 2514*(1), 32-38.

Wu, J., & Xu, H. (2018). The influence of road familiarity on distracted driving activities and driving operation using naturalistic driving study data. *Transportation Research Part F: Traffic Psychology and Behaviour, 52*, 75-85.

Xu, C., Liu, P., Yang, B., & Wang, W. (2016). Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. *Transportation Research Part C: Emerging Technologies, 71*, 406-418.

Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention, 57*, 30-39.

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., & Jiang, X. (2020). A machine-learning method for improving crash injury severity analysis: a case study of work zone crashes in Cairo, Egypt. *International Journal of Injury Control and Safety Promotion*, *27*(3), 266-275.

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention, 51*, 252-259.

Yu, R., & Abdel-Aty, M. (2014a). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science, 63*, 50-56.

Yu, R., & Abdel-Aty, M. (2014b). Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accident Analysis & Prevention, 62*, 161-167.

Yuan, J., & Abdel-Aty, M. (2018). Approach-level real-time crash risk analysis for signalized intersections. *Accident Analysis & Prevention, 119*, 274-289.

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis & Prevention, 73*, 351-358.

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access, 6*, 60079-60087.

Zöller, I., Abendroth, B., & Bruder, R. (2019). Driver behaviour validity in driving simulators–Analysis of the moment of initiation of braking at urban intersections. *Transportation Research Part F: Traffic Psychology and Behaviour, 61*, 120-130.