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

ABSTRACT

CRASH INJURY SEVERITY PREDICTION WITH ARTIFICIAL NEURAL NETWORKS

by
Rima Abisaad

Motor vehicle crashes are one of our nation's most serious social, economic and health issues. They are the leading cause of death among children and young adults, killing approximately 1.35 million people each year. Providing a safe and efficient transportation system is the primary goal of transportation engineering and planning. To help reduce traffic fatalities and injuries on roadways, crash prediction models are used to forecast the injury severity of potential crashes and apply precautionary countermeasures accordingly. Most of these models are reactive as they use historical crash data to categorize crash-related factors. Recently, advancements have been made in developing proactive crash prediction models to measure crash risk in real-time.

Crash occurrence and the resulting injury severity are influenced by several stochastic factors including driver behavior characteristics, roadway characteristics, vehicle characteristics, traffic volumes, environmental conditions, and time conditions. The objective of this research is to develop a data-driven model for crash injury severity prediction using the aforementioned factors intended to support highway safety improvement projects. The model interacts with various data sources in effective and efficient manners, which are expected to support state and local traffic management agencies in planning and operations to reduce crash injury severity.

This research explores several types of data and modeling techniques used in crash studies. The data associated with crashes on New Jersey freeways in 2017 are collected

along with INRIX reported speeds. The weighted speed variance across the traffic stream before crash occurrence is introduced as a potential variable affecting crash injury severity in the prediction model. An Artificial Neural Network (ANN) is developed to estimate crash injury severity based on potential risk parameters suggested by previous studies and data availability for New Jersey freeways. A linear regression model (LRM) is also developed using the same dataset and the performance of both models are compared and discussed. While both models have advantages and limitations, the ANN outperforms the LRM for all levels of injury severity. In addition, the traffic speed and the weighted speed variance are two variables that highly influence the injury severity level resulting from a crash.

The model can be used both proactively and reactively. It can be integrated into the State Strategic Highway Safety Plan (SHSP) to allow highway safety programs and partners in the State to work together to align goals, leverage resources and collectively address the State's safety challenges. The ability to estimate crash injury severity in real-time allows transportation agencies to deploy active countermeasures to increase safety and reduce crashes and associated delays and costs. These countermeasures include increasing service patrol coverage, implementing stricter speed rules and lowering dynamic speed limits under critical conditions to avoid crash resulting injuries and fatalities and enhance emergency response time in case of a crash.

CRASH INJURY SEVERITY PREDICTION WITH ARTIFICIAL NEURAL NETWORKS

by
Rima Abisaad

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

December 2021

.

APPROVAL PAGE

CRASH INJURY SEVERITY PREDICTION WITH ARTIFICIAL NEURAL NETWORKS

Rima Abisaad

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

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

Dr. Branislav Dimitrijevic, Committee Member — Date
Assistant Professor of Civil and Environmental Engineering, NJIT

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

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

BIOGRAPHICAL SKETCH

Author: Rima Abisaad

Degree: Doctor of Philosophy

Date: December 2021

Undergraduate and Graduate Education:

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

- Bachelor of Engineering in Civil Engineering,
  Lebanese American University, Jbeil, Lebanon, 2014

Major: Transportation

Presentations and Publications:

Abisaad, R., & Chien, S., "Investigating the Effect of Speed Variance, Weather Conditions, and Time of Day on Crash Occurrence and Severity". Transportation Research Board 97th Annual Meeting, Washington, DC, January 2018.

Abisaad, R., Chien, S., & Ting, C-J., "Estimating Freeway Crash Injury Severity Using Artificial Neural Networks". Under Review by the Transportation Research Board.

Khoury, J., Amine, K., & Abisaad, R., "An Initial Investigation of the Effects of a Fully Automated Vehicle Fleet on Geometric Design," Journal of Advanced Transportation, 2019, 1–10. doi: 10.1155/2019/6126408

I dedicate my dissertation to my family. A special appreciation to my parents, Joseph and Salam Abisaad whose affection and encouragement are my everyday motivation. My brother, Mazen stood by my side and offered mental support and research expertise. My supportive husband, Tony was there for me throughout the entire doctorate program; and my wonderful best friend, Micha's inspiring words kept me going during tough times. You have been my best supporters, every step of the way.

ACKNOWLEDGMENT

I wish to thank my dissertation advisor Professor Steven Chien for being a generous, patient, and supportive mentor throughout my studies. His dedication and expertise helped me through the research and writing process and steered me in the right direction. He is an esteemed professional and I am proud to have been mentored by him.

I would also like to thank the members of my dissertation committee: Professor Taha Marhaba, Professor Janice Daniel, Dr. Joyoung Lee and Dr. Branislav Dimitrijevic for their time, effort, insightful comments and encouragement.

I would like to acknowledge and thank the John A. Reif, Jr. Department of Civil and Environmental Engineering for providing technical and financial support. Special thanks go to the faculty and staff, especially Professor Taha Marhaba for his continued encouragement and support.

Finally, I would like to thank my colleague and friend, Celina Semaan. Her technical expertise and constant encouragement made the completion of this research possible.

TABLE OF CONTENTS

Chapter		Page
1	INTRODUCTION……..	1
1.1	Background and Problem Statement ……..	1
1.2	Objective and Work Scope……..	6
1.3	Organization…….	7
2	LITERATURE REVIEW…….....	8
2.1	General Overview of Crash Injury Severity……	8
2.2	Data Collection Methods and Injury Severity Risk Factors……	12
2.2.1	Data Collection for Crash Prediction Methods……..	12
2.2.2	Crash Risk Factors…….....	18
2.3	Crash Injury Severity Prediction Models…….....	24
2.3.1	Parametric Models…….	25
2.3.2	Non-Parametric Models……	33
2.4	Summary…….	37
3	DATA ACQUISITION……	40
3.1	Database Overview…….	40
3.2	Data Sources Description……	41
3.2.1	Crash Record Database……..	41
3.2.2	NJ-SLD Database……...	42
3.2.3	NJCMS Database……...	43
3.2.4	Floating-car Database……	43

TABLE OF CONTENTS
(Continued)

Chapter			Page
3.3	Data Processing…….…….		44
3.4	Calculation of Traffic Volume and Weighted Speed Variance……		45
	3.4.1	Traffic Volume……	45
	3.4.2	Weighted Speed Variance……	46
3.5	Final Database……		48
3.6	Summary……		52
4	METHODOLOGY…….		53
4.1	Linear Regression Model Development……		53
	4.1.1	Dependent Variable and Explanatory Variables……..	53
	4.1.2	Step Procedure for Model Development……..	58
4.2	Artificial Neural Network (ANN) …….		69
	4.2.1	ANN Types……...	71
	4.2.2	Weight and Bias Initialization …….	75
	4.2.3	Network Training…….	75
	4.2.4	Network Validation and Final Structure……...	76
4.3	Key Takeaways from Model Development……		78
5	MODEL EVALUATION……		81
5.1	Numerical Evaluation and Sensitivity Analysis…….....		81
	5.1.1	Numerical Evaluation……	81
	5.1.2	Sensitivity Analysis……...	86

TABLE OF CONTENTS
(Continued)

Chapter **Page**

5.2 Model Applications…….. 92

5.2.1 Network Screening…….. 92

5.2.2 Real-time Traffic Management.…….. 95

6 CONCLUSIONS AND FUTURE RESEARCH…… 98

6.1 Research Contributions …… 100

6.2 Research Limitations...…….. 102

6.3 Future Research……. 103

REFERENCES…….. 104

LIST OF TABLES

Table		Page
2.1	Crash-related Factors in Previous Studies…….	17
2.2	Correlation between Speed Measures and Crashes in Previous Studies……	23
2.3	Summary of Parametric Models and Key Findings …….	32
2.4	Key Findings of Non-parametric Models…….	36
3.1	Numerical Codes for Most Severe Physical Injury……	42
3.2	Spatiotemporal Weighted Speed Variance with and without Crash……	48
3.3	Crash Distribution on Freeways by Level of Injury Severity for 2017……	48
3.4	Descriptive Statistics for Data Collected on I-80…….....	51
4.1	Description of Injury Severity Levels…….	56
4.2	Description of Explanatory Variables……	58
4.3	Pearson's Chi-square Test Results…….	63
4.4	Correlations of Explanatory Variables……..	65
4.5	Stepwise Regression Results…….	68
4.6	RMSEs of Various ANN Models……..	77
5.1	Test Samples Classified by Injury Severity Level and Peak Period……..	82
5.2	RMSE of Test Samples by Injury Severity Level and Peak Period……...	83
5.3	Test Samples Classified by Injury Severity Level and Weather Condition…	84
5.4	RMSE of Test Samples by Injury Severity Level and Weather Condition…	84
5.5	ANN Precision by Injury Severity Level……..	86
5.6	Percent Change in ISI after Speed Perturbation, Segments 1-4…….	88
5.7	Percent Change in ISI after WSV Perturbation, Segments 1-4……..	91

LIST OF FIGURES

Figure		Page
2.1	Summary of crash injury severity modeling techniques……	11
3.1	Process of database consolidation from four data sources…….	41
3.2	Crash numbers by year or NJTP and GSP……..	49
4.1	Injury severity levels categories……..	55
4.2	Step-by-step procedure for LRM development……	60
4.3	P-values of explanatory variables……	64
4.4	Stepwise regression process……	67
4.5	General structure of an ANN……..	70
4.6	Workflow steps for ANN design…….	71
4.7	ANN computation procedure……..	74
4.8	Simplified structure of ANN training procedure…….	76
4.9	Final configuration of proposed ANN…….	78
5.1	Change in ISI vs. speed on segments 1-4…….	90
5.2	Change in ISI vs. weighted speed variance on segments 1-4……..	92
5.3	HSM 6-Step roadway safety management process.…….	93
5.4	Heat map of predicted ISL on different segments by time of day……..	97
5.5	Color-coded freeway segments based on real-time injury severity risk. ……	99

CHAPTER 1

INTRODUCTION

1.1 Background and Problem Statement

Motor vehicle crashes are one of the world's most serious social, economic and health issues. According to the global status report on road safety published by the World Health Organization (WHO), crashes are the leading cause of death among children and young adults aged 5–29 years, and are the cause of death of approximately 1.35 million people each year (WHO, 2018). The report also notes that simple prevention measures can significantly reduce the number of deaths and serious injuries resulting from crashes. In the United States, vehicle crashes resulted in 37,461 deaths, more than 4.6 million injuries and property damage totaling of $432 billion. In 2019, over 6.78 million people were involved in highway crashes in the United States. There were 36,096 deaths and more than 2.74 million injuries (National Center for Statistics and Analysis, 2020). Although substantial efforts have been invested to improve traffic safety, crashes continue to be a major problem worldwide and in the United States.

According to the National Highway Traffic Safety Administration (NHTSA), 10,111 lives were lost due to speed-related accidents in 2016, up 4% from 9,723 in 2015. 2018 recorded 569 speed-related deaths, a 5.7% decrease from the previous year (NHTSA, 2019). USDOT has made safety its top priority, and crash fatality for the first 9 months of 2019 reduced 2.2% compared to a year before. An estimated 26,730 people died in motor vehicle crashes through September 30, making the third quarter of 2019 the eighth consecutive year-to-year quarterly decline in fatalities since the fourth quarter of 2017

(USDOT, 2019). The NHTSA stated that "The path forward calls for a combination of policies, research, and action that requires committed and sustained effort from State, local, and Federal governments; and from highway safety partners, schools, and communities – all committed to reducing fatalities on our Nation's roads". The NHTSA also stated that speed-related crashes cost Americans an average of $40.4 billion per year, and that speeding was a major contributing factor in 29% of all fatal crashes (NHTSA, 2019).

The crash injury severity is defined by the NHTSA as the severity of a crash based on the most severe injury affecting any person involved. Crashes may lead to property damage only (PDO), injury, disability and/or death as well as financial costs to both the society and the parties involved. In addition, the normal flow of traffic is disrupted by impedance in the travel lanes. Closing a lane or even a shoulder of a road segment will disrupt traffic, especially during peak hours, thus leading to reduced travel time reliability and increased delay costs in addition to direct medical and property damage costs. With more uncertainty of traffic conditions, an unpredicted failure of highway mobility brings challenges for transportation authorities, first responders, and motorists responding to the unexpected disruptions.

The Federal Highway Administration (FHWA) office of safety states that understanding the most prevalent safety problems on our roadways is the first step in solving them and recommends the use of scientific methods and data-driven decisions to reduce the number and severity of crashes on roadways (FHWA, 2018). In addition, together with stakeholders, partners, and other USDOT agencies, the office of safety is committed to the vision of zero deaths and serious injuries on the nation's roadways by identifying safety

needs and delivering programs focusing on roadway safety designs and policies, technologies, and analytic processes that improve highway safety performance.

To date, more than 40 states including New Jersey have incorporated zero-based traffic safety efforts. With the aim to reduce fatalities and injuries resulting from crashes, the development of a data-driven methodology that predicts the crash injury severity under real-time conditions can be integrated into the state Strategic Highway Safety Plan (SHSP) to allow highway safety programs and partners in the state to align goals, leverage resources and collectively address the state's safety challenges. The ability to estimate crash injury severity allows transportation agencies to deploy active measures to increase safety and reduce crash severity and associated delays and costs. This in turn can help relieve non-recurrent congestions and aid in traffic management operations, decision-making and future planning.

In order to reduce the number of people killed and/or injured in traffic crashes, many studies have been conducted to identify the risk factors that significantly influence the injury outcomes of crashes. Several contributing factors can be classified into roadway, vehicle, driver, traffic and environmental characteristics. A good approach can illustrate the simultaneous influence of these factors on the crash likelihood and resulting injury severity. Previous research has adopted different models to explore the relationship between crashes and various factors. Because the injury severity outcome of crashes is regarded as a random event, statistical models have been extensively employed to explore the factors contributing to crash injury severity. The logistic regression model and the ordered choice model are the most commonly used models in crash analysis (Wang, 2005). Other models include multivariate Poisson lognormal regression models (Karlis, 2003; Ma, 2006), negative

binomial distributions (Chang, 2005; Lord and Mannering, 2010), zero-inflated Poisson and zero-inflated negative binomial models (Ma et al., 2017) and generalized additive modeling and random-parameters models (Fountas and Anastasopoulos, 2017). Regression analysis gained prominence in crash analysis because it is efficient, easily interpretable, does not require many computational resources or tuning, and it outputs well-calibrated predicted probabilities. A linear regression model predicts the target as a weighted sum of the inputs. The linearity of the relationship makes its interpretation clear and well-defined. Linear regression models have long been used to solve quantitative problems such as crash analysis and they have proven to yield decent predictions in previous studies (Lui and McGee, 1988; Al-Ghamdi, 2002; Kononen et al., 2011; Chen et al., 2016). However, most linear functions were formulated under assumptions and predefined underlying simplified relationships between dependent and explanatory variables. If these assumptions are violated, the model could result in biased estimations. In general, the stochastic nature of crashes is poorly described by linear functions with independent variables.

Other models were explored such as fuzzy logic and artificial neural networks (ANNs) as they exhibit better nonlinear approximation properties than traditional linear models (Chang, 2005; Lord and Mannering, 2010). While ANNs originated in biology and psychology, it rapidly advanced into other areas including business and economics, medicine, construction, information technology and transportation engineering (i.e., travel behavior, flow and incident management). ANNs can relate input with output and can automatically generate identifying characteristics from the learning material that they process and use it to deal with new situations. ANNs allow the inclusion of many variables, where irrelevant variables readily show negligible weight values, while relevant variables

show significant weight values. In addition, no assumptions are required regarding the functional form of the relationship between predictor and response variables as the case is with the statistical methods. As machine learning techniques can recognize patterns and adjust dynamically with gained prominence and maturity, ANNs used in crash modelling can predict desired results despite limited data sets. Traffic forecasting complications involving random and complex variables can therefore be successfully resolved.

Regression models and exhibit strengths and weaknesses in different scenarios: regression models are straight forward and easily understandable, and ANNs tend to achieve a better fit and forecast by catching sophisticated non-linear integrating effects. Therefore, a base linear regression model (LRM) is developed in this study in addition to an ANN to serve for comparison and assessment purposes. Limitations of previously developed models are addressed in this research by improving data quality and investigating the effect of additional explanatory variables on crash injury severity. Whereas previous studies used speed measures such as travelling speed, mean speed, speed limit, and speed limit deviation, this study introduces and analyzes the speed variance as a new factor affecting crash injury severity. It is important to note that speed variance has previously been used to model crash frequency and/or probability, but its use in crash injury severity modelling is novel in this research (Kockelman and Ma, 2010). In addition, traffic speed and count data in most previous models are collected by conventional loop detectors. As a variety of massive traffic data from infrastructure sensors and floating cars has become available with technology, new comprehensive data made way for big data analytics as an emerging method for predicting crash injury severity. The traffic data collection technologies utilizing floating-car concepts have improved rapidly in the past few years, in terms of geographic coverage,

sample size, accuracy in detecting vehicle location and data processing algorithms. Such improvements provide an opportunity to address the challenges of the existing models and increase the accuracy of the predictions generated by the proposed models. Therefore, this research strives to fill the related research gaps and contribute to ANN applications in improving the accuracy of crash injury severity prediction and traffic incident management in general.

1.2 Objective and Work Scope

The objective of this study is to develop an ANN for crash injury severity prediction, considering roadway characteristics, speed measures including speed variance, traffic flow, and environmental conditions. In addition to the ANN, a base LRM is also developed and serves as a basis of comparison and differentiation between the two models. Both models are developed using the same set of data and their performances are compared to highlight the advantages and limitations of each model. The proposed models interact with various real-time data sources in effective and efficient manners, which is expected to support state and local traffic management agencies in operations to reduce crash injury severity levels. An accurate prediction of the crash injury severity will allow for a more effective traffic management and mitigation plans. This in turn will result in reduced travel delay and induced medical cost. To achieve the objective of this study, various crash likelihood prediction approaches and crash injury severity models will be thoroughly reviewed.

1.3 Organization

This dissertation is organized into six chapters. Chapter 1 introduces the research problem, the background of the research and the need for this study. It also presents the research objectives and work scope. Chapter 2 comprises a thorough review of the current literature about crash prediction models. Chapter 2 is divided into three sections. The first section covers a general overview of crash injury severity research; the second section covers data collection methods used in crash injury severity prediction methods; the third section summarizes previously developed crash injury severity prediction methods including parametric and non-parametric models, and the fourth section presents a summary of the literature review findings and highlights the need for the comprehensive model developed in this research. Based on conclusions drawn from Chapter 2, the data collection procedure in discussed in Chapter 3. The detailed development of the working database is explained, including a database overview, data sources description, data processing, and final database. In Chapter 4, the model formulation is presented, and the weighted speed variance is introduced to the data acquired in Chapter 3. Consequently, a base LRM is developed using qualified freeway crash data in 2017. With the same data, an ANN is developed and evaluated in Chapter 4. Chapter 5 discusses the model results and provides a summary of the practical implications of the developed models. Chapter 6 presents conclusions and discusses the strengths and limitations of the developed models as well as the potential and need for future studies.

CHAPTER 2

LITERATURE REVIEW

This discussion is a comprehensive overview of crash prediction methods and data used to formulate them. The first section presents a general synopsis of the research around crash injury severity. The second section summarizes the types of data used in developing crash prediction models as well as suggested variables proven to affect crashes, and identifies a new variable to be added in this research: weighted speed variance. In the third section, crash prediction models are presented under two categories: parametric models and non-parametric models including artificial neural networks. The advantages and disadvantages of these models are also discussed. The fourth section summarizes the findings of the literature review and justifies the proposed data, variables and model used in this study.

2.1 General Overview of Crash Injury Severity Research

Roadway crashes are the cause of substantial economic loss and human life loss, and the injury severity of crashes is of utmost importance on freeways, where high speed limits lead to higher injury and fatality rates (Florence et al. 2013). Moreover, freeway lane closures due to crashes is the leading cause of nonrecurring congestion and commuter delays, particularly significant when crashes occur on busy roadways. Crash prediction models can be used to forecast the injury severity of crashes likely to occur and are very important safety tools that help remedy economic and social loss. Given the importance of roadway safety, substantial effort has been invested into estimating factors associated with crashes and their resulting injury severity levels. These include driver behavior characteristics,

roadway characteristics, vehicle characteristics, traffic volumes, environmental conditions, and time conditions. This section presents a summary of research efforts invested in this area and the corresponding findings.

Previous research has examined crash injury severity using different modeling techniques. These techniques can be classified into four groups: discrete outcome models, data mining techniques, soft computing techniques and other techniques (Mujalli and Ona, 2013).

1. Discrete outcome models (DOMs) are used to model the probability of a specific outcome based on risk factors or characteristics. In general, DOMs cannot be calibrated using standard curve-fitting techniques (Ortu´zar and Willumsen, 2001). Some examples of DOMs are described below:

    - Logit models (LMs) or logistic regression are a special form of general linear regression which assumes that a response variable follows the logit-function. The logistic model is an approach used to describe the relationship of single or several independent variables to a binary outcome variable. The multinomial logit model (MNL) is used when the outcome variable has more than two unordered categories.

    - Probit models (PMs) deal with the limitations of LM: they can handle random variation; they allow any pattern of substitution; and they are applicable to panel data with temporally correlated errors (Train, 2009). The most used type of PM in the analysis of accident severity is the ordered probit model (OPM). The OPM is a generalization of the PM to the case of more than two outcomes of an ordinal dependent variable.

    With regression models, the dependent variable is continuous: it has an infinite number of possible outcomes. In that case, discrete choice models cannot be applied. Discrete choice models can be used to estimate specific limited outcomes from a set of two or more discrete choices. Linear regression is more natural and easier. The choice between a regression and a discrete choice model is governed by the research objectives and the available data.

2. Data-mining techniques are defined as the process of discovering patterns in data. Data mining is popular in different fields of science, economy, engineering and more specifically traffic studies. They have also been

employed to model the injury severity levels resulting from a crash. Data mining techniques include Decision trees and Bayesian networks.

3. Soft computing techniques are based on artificial intelligence and natural selection that provides quick and cost-effective solution to very complex problems for which analytical formulations might not exist. Some examples of soft computing techniques are briefly discussed below:

    - Neural networks are interrelated assemblies of simple processing elements, units or nodes. The processing ability of the network is contained in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns (Gurney, 1997).

    - Evolutionary algorithms (EAs) mimic natural evolution to find an optimal solution to a problem (Brameier and Banzhaf, 2007). These algorithms exploit differential fitness advantages in a population of solutions to gradually improve the state of that population.

    - Genetic programming (GP) is defined as any direct evolution or breeding of computer programs for the purpose of inductive learning. Unlike other EAs, GP can complete missing parts of an existing model.

4. Generalized linear models (GLMs) are modeling techniques where the response variable can have an error distribution other than a normal distribution. The log-linear model is a general form of GLMs where there is no distinction between independent and dependent variables; all variables are treated as response variables.

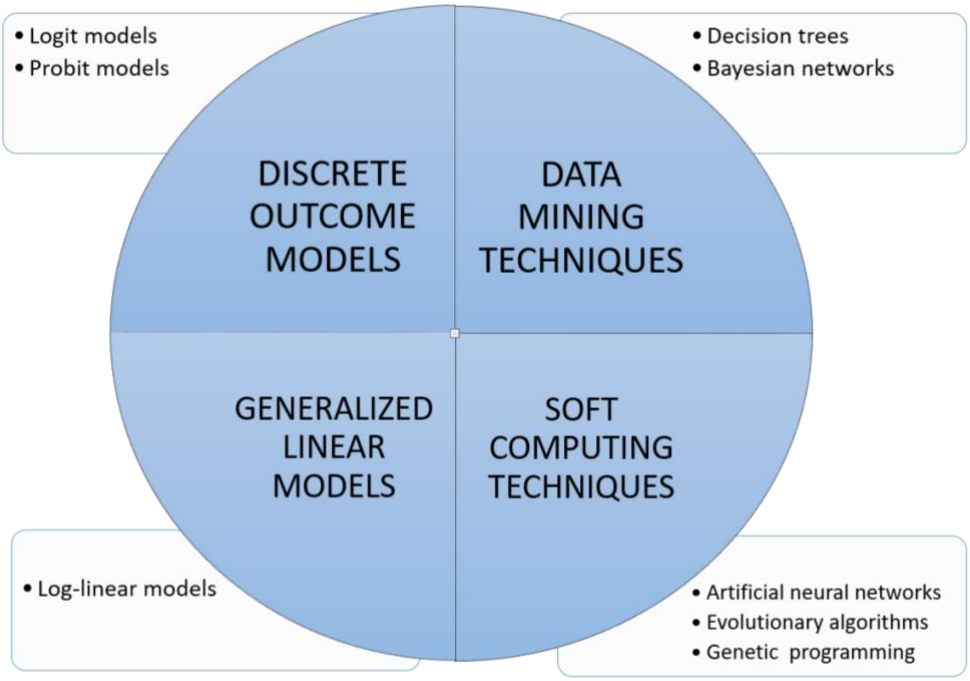


Figure 2.1 Summary of crash injury severity modeling techniques.

The state-of-the-art research suggests that statistical models can predict reliable estimates by relating crash aggregates to various explanatory measures of speed, flow, site characteristics and road geometry, and the reliability of crash prediction models is important for the improvement of traffic safety management and the prevention of traffic injury.

Based on previous research, several stochastic variables affect crash injury severity in different degrees under different circumstances. The objective of this study is to develop a comprehensive and accurate model for crash injury severity prediction, considering roadway characteristics, speed measures including weighted speed variance, traffic flow, and environmental conditions. The proposed model interacts with various data sources in effective and efficient manners, which is expected to support state and local traffic management agencies in planning and operations to reduce both roadway crashes and their

injury severity levels. The following sections deliver a thorough background about the current literature and in-depth overview of modeling techniques summarized in this section.

2.2 Data Collection Methods and Injury Severity Risk Factors

This section provides a more detailed look into the crash injury severity prediction methods by identifying the types of data used in model development and the risk factors associated with crash injury severity. The purpose of this section is to identify a good type of data to be collected and a list of potential parameters affecting injury severity to be used for the model development.

2.2.1 Data Collection for Crash Prediction Methods

Previous studies have investigated factors affecting crash likelihood and/or injury severity over several types of roadways and using different data types. The types of data analyzed in crash prediction models are grouped into panel data (or historic data) and real-time data crash models. Panel data crash models mainly focus on modeling longitudinal data resulting from yearly repeated observations and are therefore unable to capture the effects of contributing factors that vary within a year. For example, when it comes to traffic flow and weather information, panel data crash studies represent their effects with long-term aggregated and/or averaged variables such as annual average daily traffic (AADT) volume and number of days with rainfall over a year. Real-time data crash models focus on the relative crash risk with real-time traffic and environmental conditions prior to crashes. In today's fast changing world, impromptu decisions are made by the second to accommodate unpredictable traffic conditions, hence lies the importance of real-time data in crash prediction methods.

Panel Data Crash Studies

Panel data gives insight about the past trends and helps analyze mistakes and circumstances to be avoided. Corrective actions can be taken accordingly to increase the efficiency of the prediction process. The findings from studies using panel data for crash prediction are discussed and summarized below.

Abdel-Aty and Radwan (2000) used historical data and a negative binomial distribution model to predict crash frequency as a function of AADT, horizontal curvature, section length, lane, shoulder and median widths. Accident data over 3 years, including 1,606 accidents on a principal arterial in central Florida, were used to build the model. The results showed that crash frequency increases with AADT, horizontal curvature and section length, while it decreases with lane, shoulder and median width.

Other studies using historical data showed that traffic volume was a significant variable affecting crashes. Greibe (2003) developed accident prediction models based on data from 1,036 junctions and 142 km road links in urban areas. Generalized linear modelling techniques were used to relate accident frequencies to explanatory variables such as AADT, speed limit, number of lanes and road width. The findings illustrated that the AADT, highly correlated with crash frequency, was the most important and powerful variable in the model. The modelling also showed that road links with high-speed limits tend to have lower accident risk. Ye et al. (2018) developed a multivariate Poisson regression model to model head-on crashes and analyze crash frequency by collision type using crash data for 165 rural intersections in Georgia. The results showed that posted speed limit and traffic volume on both the major and minor roads had a positive effect on the number of head-on crashes.

Garder (2005) examined the simultaneous influence of different variables on the crash severity with historical head-on crash data on two-lane rural highways in Maine. The ordered probit model results showed that most of the crashes were due to either driver inattention and/or distraction, excessive speeds, fatigue, or alcohol/drug use. In addition, higher speed limits, more travel lanes, wider shoulder widths and higher AADT were found to contribute positively to crash severity. Similar conclusions were drawn by Bham et al. (2012); they investigated single-vehicle and multi-vehicle collisions using a multinomial logistic regression model. Historical data collected on urban highways in Arkansas between 2005 and 2007 were used to determine the impacts of several factors on crash outcomes for six collision types. The study concluded that slowing or stopping and driving under the influence of alcohol were found to be significantly associated with head-on collisions. In addition, the authors noted that head-on collisions contribute to a higher risk of severe injuries compared to other crash types.

Other research evaluated the effect of roadway geometry on crash frequency and severity. Yan et al. (2011) analyzed Beijing historical crash data over four years to understand the relationship between crash patterns and injury severity. Their results showed that injury severity levels could be elevated by crash patterns including head-on and angle collisions, nighttime, undivided roads, higher speed limit and heavy vehicle involvement. The study suggested installation of median, improvement of illumination on road segments, and reduced speed limit at roadway locations with high traffic volume. Similar variables related to geometric design were found to be correlated with crash injury severity by Ma et al. (2017). They used a negative binomial model and a random effect negative binomial model. The accident data was retrieved on a 50km long expressway in China, including 567

crash records between the years 2006 and 2008. Three explanatory variables, including longitudinal grade, road width, and ratio of longitudinal grade and curve radius, were found as significantly affecting crash frequency.

Real-time Data Crash Studies

This section presents studies using real-time data for crash prediction. Unlike panel data crash studies where the AADT is a leading contributing factor, real-time data studies show that speed data is the primary factor affecting crashes.

Ahmed and Abdel-Aty (2012) examined freeway locations with high crash occurrence using real-time speed data collected from automatic vehicle identification (AVI) systems. Travel time, space mean speed data, and crash data of a total of 78 miles on the expressway network in Orlando in 2008 were collected. The results of the random forest technique showed that the likelihood of a crash is statistically related to speed data obtained from AVI segments within an average length of 1.5 miles.

Real-time data crash studies also concluded that roadway geometry contributes to crashes. Yu and Abdel-Aty (2013) presented a multi-level analysis for single- and multi-vehicle crashes using crash data from a 15-mile mountainous freeway section on I-70. They developed an aggregate model using five years of crash data, and a disaggregate model using one year of crash data along with real-time traffic and weather data. The model results indicated that the effects of the selected variables on crash occurrence vary across seasons and that geometric characteristic variables contribute to the segment variations.

Effati et al. (2015) presented a comprehensive geospatial approach based on the fuzzy classification and regression tree (FCART) to predict motor vehicle crashes and their severity on two-lane, two-way roads. They applied a bagging algorithm in the FCART

model to account for high-variance crash data and improve the performance of the learning process. They also conducted a sensitivity analysis to determine the importance of input factors. The results showed that vehicle failure, drivers wearing seat belts, and weather condition factors are some of the most important factors contributing to crash severity. In addition to these factors, geographical factors such as proximity to curves and adjacent facilities and land use have a significant effect on crash severity.

Chen et al. (2016) developed crash prediction models with hourly recorded data to describe the time-varying nature of these crash-contributing factors. They developed an unbalanced panel data mixed logit model to analyze hourly crash likelihood of highway segments, and incorporated temporal driving environmental data, including road surface and traffic condition, obtained from the Road Weather Information System (RWIS). Their results showed that weekend indicator, November indicator, low speed limit and long remaining service life of rutting indicator are found to increase crash likelihood, while 5-am indicator and number of merging ramps per lane per mile are found to decrease crash likelihood. Table 2.1 summarizes the studies discussed in Sections 2.1.1 and 2.1.2 and the factors found to be affecting crashes.

Table 2.1 Crash-related Factors in Previous Studies

Authors (year)	Model	Data Type	Factors affecting crash
Abdel-Aty and Radwan (2000)	Negative binomial regression	Panel	AADT, horizontal curvature, section length, lane, shoulder, and median widths
Greibe (2003)	Generalized linear model	Panel	Traffic volume
Garder (2005)	Ordered probit model	Panel	Driver distraction, excessive speeds, AADT, alcohol/drug use
Bham et al. (2012)	Multinomial logistic regression	Panel	Slowing or stopping and driving under the influence
Ahmed and Abdel-Aty (2012)	Multiple models	Real-time	Real-time traffic variables and visibility conditions
Yu and Abdel-Aty (2013)	Multi-level aggregate model	Real-time	Geometric characteristics
Effati et al. (2015)	Fuzzy classification and regression trees	Real-time	Vehicle failure, seat belt usage, and weather condition
Chen et al. (2016)	Unbalanced panel data mixed logit model	Real-time	Weekend indicator, November indicator, and low speed limit
Shi et al. (2016)	Multi-level Bayesian framework	Real-time	Speed and speed variation
Ma et al. (2017)	Negative binomial regression	Panel	Longitudinal grade, road width, and ratio of longitudinal grade and curve radius
Chiou et al. (2017)	Clustering and multivariate approaches	Panel	Geometric and environmental factors in the afternoon, traffic in the morning
Ye et al. (2018)	Multivariate Poisson regression model	Panel	Speed limit and traffic volume

A major difference between crash studies is the type of data used in the research as discussed in this section. Panel data crash studies mainly focus on modeling longitudinal data resulting from yearly repeated observations, whereas real-time data crash studies focus on the relative risk with real-time traffic and environmental conditions prior to crashes. This

study is intended to provide reliable predictions for crash injury severity to help reduce crash fatalities/injuries, relieve non-recurrent congestions, and aid state and local traffic management agencies in traffic management operations using real-time data. As a variety of massive traffic data from infrastructure sensors and floating cars has become available with technology, this comprehensive data is a promising tool for predicting freeway crash injury severity. The traffic data collection technologies utilizing floating-car concepts have improved rapidly in the past few years, in terms of geographic coverage, sample size, accuracy in detecting vehicle location, and data processing algorithms, and such improvements provide an opportunity to address the challenges of the existing models and increase the accuracy of the predictions generated by the proposed model.

2.2.2 Crash Risk Factors

Motor vehicle crashes cause more than 1.2 million deaths worldwide and a greater number of injuries yearly. To improve road safety, extensive information about the causes of crashes is needed. Police reports prepared on the crash scene are the main source of data used for collecting data on the causes of crashes. The degree of crash injury depends on the relationship between physical injuries and crash mechanisms, but understanding is often limited by complicated crash mechanics (Carlson, 1979).

In addition to roadway, vehicle, driver, traffic and environmental characteristics discussed in the previous section, speed measures such as average speed, speed limit, and speed variance are major contributors to crash occurrence and/or injury severity. According to the NHTSA, speeding has been involved in approximately one-third of all motor vehicle fatalities for more than two decades. Speed also affects the motorist's safety even when driving at the speed limit but too fast for road conditions, such as during bad weather, when

a road is under repair, or in a dark area. The exact relationship between crashes and speed measures depends on several factors.

Generally, the safety office of the FHWA states that if on a road the driven speeds become higher, the crash rate will increase. In addition, the crash rate is also higher for an individual vehicle that drives at higher speed than the other traffic on that road. As speeds get higher, crashes also result in more serious injury, for the driver who caused the crash as well as for the crash opponent (SWOV, 2012). The World Health Organization suggests that speed is the key risk factor in road traffic injuries, influencing both the risk of a road crash, as well as the severity of the resulting injuries (WHO, 2004). An increase in average speed of 1 km/h typically results in a 3% higher risk of a crash involving injury, with a 4–5% increase for crashes that result in fatalities. When a collision occurs; for car occupants in a crash with an impact speed of 80 km/h, the likelihood of death is 20 times as that with the impact speed of 30 km/h.

The correlation between speed and crashes has been widely discussed in previous research. This section presents the main findings of studies analyzing the relationship between crashes and speed measures. Solomon (1974) studied reported accidents on two-lane and four-lane rural highways. A significant correlation was drawn between speed limit deviation and the probability of a crash involvement. In addition, the greater the variation in speed of any vehicle from the average speed of all traffic, the greater its chance of being involved in an accident. The results suggested that travel speed of many vehicles involved in a crash had deviated widely from the speed limit.

Garber and Gadiraju (1989) investigated the influence of different factors on speed variance and quantified the relationship between speed variance and accident rates. It was

concluded that speed variance is minimum if the posted speed limit is between 5 and 10 mph lower than the design speed. Outside this range, speed variance increases with an increasing difference between the design speed and the posted speed limit. It was also found that drivers tend to drive at increasing speeds as the roadway geometric characteristics improve, regardless of the posted speed limit, and that accident rates do not necessarily increase with an increase in average speed but do increase with an increase in speed variance. They concluded that higher speed does not necessarily lead to more crashes, but higher speed variation does, and higher speed affects the severity of the crash rather than its likelihood.

Kloeden et al. (2001) investigated the relationship between free travelling speed and the risk of involvement in a casualty crash in 50 mph or greater speed limit zones in rural South Australia. Free travelling speed was defined as "the speed of a moving vehicle, not closely following another vehicle, and not slowing to leave a road, or accelerating on entering one." They reported that a vehicle traveling six mph above the speed limit doubles its risk of being involved in an injury-type crash. The risk becomes nearly six times as great when travelling twelve mph above the speed limit. It was shown that even small reductions in travelling speeds have the potential to greatly reduce crash and injury. In a subsequent study, Kloeden et al. (2002) used a modified logistic regression to model the risk of being involved in a casualty crash based on free travelling speed in an urban roadway with 60 km/h speed limit in South Australia. They reported that travelling speeds directly affect crash frequency as opposed to other factors such as the type of drivers who choose to travel at different speeds or the variance in travelling speeds. They also concluded that a small reduction in travel speed (such as 1 km/h or less) can significantly decrease casualty crashes.

Kockelman et al. (2006) concluded that the total number of crashes increased by around 3% when the speed limit increased from 55 mph to 65 mph, while the probability of a fatal crash increased by 24%. Ma and Kockelman (2006) conducted a cost-benefit analysis and suggested that raising speed limits results in substantial travel time saving. The results indicated that increasing speed limit (from 55 mph to 65 mph) would save hours of vehicle travel (equivalent to $1,607,455), whereas the cost of additional crashes was $437,964.

Malyshkina and Mannering (2007) studied the influence of the posted speed limit on the severity of accidents using Indiana accident data from 2004 (the year before speed limits were raised) and 2006 (the year after speed limits were raised) on rural Interstates and some multilane non-Interstate highways. The unordered-probability approach that includes multinomial, nested, and mixed logit models estimated the injury severity of accidents on various roadway classes. The results showed that speed limit affects incident severity on some non-Interstate highways whereas it does not affect it on Interstate highways.

While some studies indicate that travelling speed directly affects the crash outcome, others indicate that its effect is minimal and speed variation has a greater effect on crashes. Cooper (1997) studied the relationship between speeding behavior and crash involvement. The author differentiated between two speeding convictions: excessive speed (driving 25 mph or more above the speed limit) and exceeding the speed limit (driving 5 to 10 mph above the speed limit). The results showed that the presence of speeding convictions was significantly related to the risk of crash involvement. However, of these two classes of speeding conviction, only excessive speed became a more important crash-involvement predictor as the severity of subsequent crash events increased. Another observation was that

having only speeding convictions of the exceeding-speed-limit type (and no excessive speed convictions) seemed to be associated with speed-related crash risk at a very similar level to that associated with having non-speeding convictions.

Golob and Recker (2002, 2003) investigated the effect of speed variation on crashes. They studied freeway crashes in Orange County, California, in which traffic flow regimes were classified based on speed variation. The highest crash rates during morning peak appeared under the conditions of heavy flow, low mean speed, and low speed variation whereas the lowest crash rates were found near capacity conditions during low-speed variations and high speeds.

Abdel-Aty et al. (2004) analyzed the relationship between crash likelihood and traffic characteristics using matched case-control logistic regression and found that the most significant factors influencing the likelihood of crash occurrence were average occupancy observed at the upstream station and coefficient of variation in speed at the downstream station. Abisaad and Chien (2018) investigated the effect of speed profiles among other factors on freeway crashes. Several factors were proven to affect crash injury severity under different conditions. The findings showed that the speed variance paired with adverse weather conditions and/or other factors are a good indicator for crash injury severity and can be used as guidelines for traffic monitoring centers. Depending on weather conditions, posted speed limit, and time of day, the traffic patterns should be monitored for potential crashes yielding high injury severity levels. Mitigation measures can be applied by increasing service patrol coverage, implementing stricter speed rules, and lowering dynamic speed limits to increase safety and enhance emergency response time in case of a crash.

Table 2.2 summarizes the studies discussed in this section and correlation between speed measures and crashes.

Table 2.2 Relationship between Speed Measures and Crashes from Previous Studies

Authors (year)	**Key Findings**
Solomon (1974)	Speed limit deviation and variation from the average speed of all traffic increases chance of crash involvement
Garber and Gadiraju (1989)	Increase in speed variance affects the severity of the incident rather than its likelihood
Cooper (1997)	Previous speeding convictions is significantly related to the risk of crash involvement
Kloeden et al. (2001)	Small reductions in travelling speeds greatly reduce crash and injury
Kloeden et al. (2002)	A small reduction in travel speed can significantly decrease casualty crashes
Golob and Recker (2002, 2003)	The highest crash rates appear under heavy flow and low speed variation whereas the lowest crash rates happen during low-speed variations and high speeds
Abdel-Aty et al. (2004)	The most significant factors influencing the crash occurrence are average occupancy and coefficient of variation in speed
Oh et al. (2005)	The most significant variable influencing crash likelihood is the standard deviation of speed
Kockelman et al. (2006)	The total number of crashes and the probability of a fatal crash both increase when the speed limit increases
Ma and Kockelman (2006)	Increasing speed limit saves hours of vehicle travel equivalent to an amount higher than the cost of additional crashes
Malyshkina and Mannering (2007)	Speed limit affects incident severity on some non-Interstate highways whereas it does not affect it on Interstate highways
Kononov et al. (2012)	Once a combination of speed and density threshold is exceeded, the crash rate rises rapidly.
Abisaad and Chien (2018)	The speed variance paired with adverse weather conditions and/or other factors are a good indicator for crash injury severity.

The proposed model in this research considers big data such as crash information, road geometry, directional traffic volumes, and floating-car data. It confirms the unique and significant impacts on crash executed by the real-time weather, road surface, and traffic conditions. As for the correlation between speed measures and crashes, some studies indicate that travel speed directly affects crashes, others argue that its effect is minimal and speed variation has a greater effect on crashes, and other studies concluded that speed variance does not affect crashes. The literature has contradicting theories on the speed-crash debate, which is due to the stochastic nature of factors that lead to incidents, and further research presented in this study is required to better understand and identify the factors associated with crashes.

2.3 Crash Injury Severity Prediction Models

After identifying the types of data used in crash prediction methods, crash-related factors as well as the effect of speed measures on crashes in the previous sections, this section reviews the models used in crash injury severity prediction. Injury severity studies institute the statistical relationship between the dependent variable, injury severity, and several explanatory variables relating to driver characteristics, roadway characteristics and environmental conditions. Parametric and non-parametric models used to predict crash injury severity are discussed in this section.

A parametric model is a distribution that can be described using a finite set of parameters. Given the parameters, predictions are independent of the observed data used in developing the original model. Therefore, the complexity of the model is bounded even if the amount of data is unbounded, and this feature limits the flexibility of parametric

models compared to non-parametric models. Non-parametric models assume that the data distribution cannot be defined in terms of a finite set of parameters, but rather by assuming an infinite dimensional or function (Schmidt-Burkhardt, 2011). The amount of information that this function can capture about the data can grow as the amount of data grows, and this feature provides them with more flexibility in accommodating data sets.

Parametric models (e.g., polynomial regression, logistic regression, Poisson model, etc.) assume a specific form for the model. They are simpler and are interpretable but require a greater number of data points. They work well if the assumption is correct. Non-parametric models (e.g., artificial neural network model, support vector machine model, Gaussian process, etc.) do not assume anything about the data and learn from the data gradually and are slower. They typically require less data than what is required for parametric models. The application of ANNs in engineering science has been proven highly efficient in recent years, because of their capability to predict and present desired results despite limited data sets. ANNs were recently introduced to the transportation field and their use is addressed in this section.

2.3.1 Parametric Models

Regression analysis has been the most popular technique in developing crash injury severity prediction models, and other models described in Section 2.1 have also proven the ability to predict crash injury severity (i.e., logit, probit, loglinear, etc.). The reviewed studies on crash injury severity using parametric models are grouped into regression model studies and other studies.

Regression Models

Lui and McGee (1988) used a logistic regression to analyze the probability of fatal outcomes of accidents given that the crash has occurred. They obtained data for accidents including at least one fatality and modeled the probability of a fatality as a target variable dependent on driver's age and gender, impact points, car deformation, use of restraint system and vehicle weight. Their findings reveal that a heavier car weight can greatly reduce the driver's risk of dying in a two-car crash, because larger cars' frames can better absorb energy from an impact, or the fact that the small and lighter cars tend to roll over more easily. Driver's age and/or gender was also investigated by Farmer et al. (1997). The study examined the impact of vehicle and crash characteristics on injury severity in two-vehicle side-impact crashes using a binomial regression model. Their results revealed that rollover or ejection from the vehicle increases the likelihood of a serious injury or death and that light-duty trucks were fourteen times more likely to roll than cars, when struck on the side. While gender was not a statistically significant factor in their results, the oldest drivers (aged 65 and over) were estimated to be more at risk for serious injury.

Mao et al. (1997) assessed the factors affecting the severity of motor vehicles traffic crashes involving young drivers in Ontario using unconditional logistic regression. Their results show that factors significantly increasing the risk of fatal injury crashes include drinking and driving, impairment by alcohol, exceeding speed limits, not using seat belts, full ejection from vehicle, intersection without traffic control, bridge or tunnel, road with speed limit 70-100 km/hour, bad weather, head-on collision, and overtaking. Results of the same model applied to major and minor injury crashes demonstrated consistent but weaker associations with decreasing levels of crash severity. Another study by Al-Ghamdi (2002) using the logistic regression approach examined the contribution of individual variables to

the injury severity level resulting from a crash. The study evaluated 560 accidents obtained from the police records in Riyadh, Saudi Arabia. The dependent variable was modeled as a dichotomous variable that could only take values of fatal or non-fatal crash outcomes. According to the logistic regression results, out of nine independent variables used in this study, only two were found to be statistically significant with respect to the injury severity: location and cause of accident. Moreover, the odds of being in a fatal accident at a non-intersection location are 2.64 higher than those at an intersection, and the odds of severe injury increases on accidents caused by over-speeding and entering the wrong way traffic.

Similar conclusions were drawn by Kononen et al. (2011). They developed a multivariate logistic regression model, based upon National Automotive Sampling System Crashworthiness Data System (NASS-CDS) data for calendar years 1999–2008 to predict the probability that a crash-involved vehicle will result in serious or incapacitating injuries. Model input parameters included: crash direction, change in velocity, multiple vs. single impacts, belt use, presence of at least one older occupant (≥55 years old), presence of at least one female in the vehicle, and vehicle type. Model sensitivity and specificity were 40% and 98%, respectively. The results indicated that seat belt use and crash direction were the most important predictors of serious injury resulting from a crash.

Ma et al. (2015) developed a generalized ordered logit model by using police-reported crash records of selected freeway tunnels in China. They reported five factors significantly related to injury severity. The season, time of day, location, tunnel length, and adverse weather were found to affect injury severity on freeway tunnels. In addition, less fatal injuries occurred during the summer season, and less injury crashes occurred during night-time. Subsequently, Ma et al. (2017) developed a method to explore the relationship

between various explanatory variables and crash injury severity based on 673 crash records collected on rural two-lane highways in China. A partial proportional odds model examined factors influencing crash injury severity, and an elasticity analysis was conducted to quantify the marginal effects of each contributing factor. The results showed that nine explanatory variables, including at-fault driver's age, at-fault driver having a license or not, alcohol usage, speeding, pedestrian involved, type of area, weather condition, pavement type, and collision type, significantly affect injury severity.

Behnood and Mannering (2017) applied a random parameters logit model to investigate the effects of drug and alcohol consumption on driver injury severities. Using data from single-vehicle crashes in Cook County, Illinois, from January 1, 2004, to December 31, 2012, separate models for unimpaired, alcohol-impaired, and drug-impaired drivers were estimated. A wide range of variables potentially affecting driver injury severity was considered, including roadway and environmental conditions, driver attributes, time and location of the crash, and crash-specific factors. The results showed that unimpaired drivers are more responsive to variations in lighting, adverse weather, and road conditions. Age and gender were found to be important determinants of injury severity. Moreover, unimpaired drivers tend to have more heterogeneity in their injury outcomes under adverse weather and road surface conditions. In contrast, alcohol-impaired and drug-impaired drivers have far less heterogeneity in the factors that affect injury severity, resulting from the decision-impairing substance.

Ji and Levinson (2020) adopted the energy loss-based vehicular injury severity (ELVIS) to explain the effects of the energy absorption of two vehicles in a collision. A multivariate ordered logistic regression with multiple classes was estimated and the results

showed that occupants in heavy vehicles absorb less impact from the crashes, suffering less significant injuries. Moreover, the number of vehicle occupants is positively correlated to the most severe injuries in one vehicle.

Other Models

Chang and Wang (2006) used the 2001 accident data for Taipei, Taiwan to develop a CART model that establishes the relationship between injury severity and driver/vehicle characteristics, highway/environmental variables, and accident variables. The results indicated that the most important variable associated with crash severity is the vehicle type. Pedestrians, motorcycle, and bicycle riders were identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents. Qi et al. (2007) used a discrete response ordered probit model to predict accident likelihood. The results illustrated that the model performs well in identifying factors associated with traffic accidents. In addition, when applied in a predictive setting, the model provided benefits in forecasting the likelihood of accidents based on both time-varying and site-specific parameters. Hao and Daniel (2014) applied an ordered probit model to explore the causes of driver injury severity under various control measures at highway-rail grade crossing in the United States. Their analysis found that peak hour factor, visibility, car speed, train speed, driver's age, area type, traffic volume and highway pavement impact driver injury severity at both active and passive highway-rail crossings. In another study, Hao and Daniel (2015) applied a mixed logit model to explore the determinants of driver injury severity under different weather conditions at highway–rail grade crossing. A reduction in speed limit during inclement weather conditions could be particularly effective in controlling injury severity, allowing more reaction time for maneuvering and braking before impacts.

Fountas and Anastasopoulos (2017) used a random threshold hierarchical ordered probit model with random parameters to analyze highway accident data collected in the State of Washington, between 2011 and 2013. They found seven variables affecting crash injury severity. These include geometric characteristics (vertical curve length); traffic characteristics (AADT – per lane); driver-specific characteristics (use of alcohol/drugs); and accident-specific characteristics (indicator for out-of-control vehicle, speed, indicator for vehicle going straight ahead at the time of the accident, and pedestrian involvement indicator). Osman et al. (2018) analyzed the injury severity of commercially licensed drivers involved in single-vehicle crashes using the ordered response modeling framework. The effect of driver's age on all other factors was examined by segmenting the parameters by driver's age group. The empirical analysis was conducted using four years of the Highway Safety Information System (HSIS) data that included 6247 commercially-licensed drivers involved in single-vehicle crashes in the state of Minnesota. Their results showed that important factors affecting the crash severity of crashes for commercially licensed drivers across all age groups include lack of seatbelt usage, collision with a fixed object, speeding, vehicle age of 11 years or more, wind, nighttime, weekday, and female drivers.

Chen and Jovanis (2000) modelled the relationship between crash severity and associated factors using a loglinear model. They analyzed 408 observations considering bus crashes in a freeway in Taiwan between 1985 and 1993. Frontal impact collisions and driving during late night hours or early morning hours were among the factors affecting crash severity.

Several studies developed more than one model and compared the model performances to illustrate advantages and disadvantages of different approaches. Jacob and

Anjaneyulu (2013) conducted a study using data of more than 500 kilometers of National and State Highways of Kerala, India to highlight the influence of various roadway and traffic conditions on traffic safety. They found that factors that cause injury crashes are significantly different from those that cause fatal crashes. Four models were developed: a multiple linear regression model, a Poisson regression model, a negative binomial regression model and a zero-inflated Poisson regression model. The comparison of models based on the percentage root mean square error showed that the Poisson regression model yielded the most accurate results for predicting fatal and injury crashes.

Mooradian et al. (2013) also compared several parametric models using the same data. They proposed a partial proportional odds (PPO), a type of logistic regression, to predict vehicular crash severities on Connecticut state roads using data from 1995 to 2009. The PPO model was compared to ordinal and multinomial response models on the basis of adequacy of model fit, significance of covariates, and out-of-sample prediction accuracy. The study results show that the PPO model has adequate fit and performs best overall in terms of covariate significance and holdout prediction accuracy. Wang et al. (2016) used an intersection data inventory of 36 safety relevant parameters for three- and four-legged non-signalized intersections along state routes in Alabama to study the importance of intersection characteristics on crash rate and the interaction effects between key characteristics. Four different models were developed and compared: Poisson regression, negative binomial regression, regularized generalized linear model, and boosted regression trees. The boosted regression tree model significantly outperformed the other models and identified several intersection characteristics as having strong interaction effects. Table 2.3 summarizes the key findings of studies using parametric models discussed in this section.

Table 2.3 Summary of Parametric Models and Key Findings

Authors (year)	Model	Key Findings
Lui and McGee (1988)	Logistic regression	Heavier cars reduce the risk of a fatality
Farmer et al. (1997)	Binomial regression	Rollover or ejection increases the likelihood of injury or death
Mao et al. (1997)	Unconditional logistic regression	Factors increasing the risk of fatal crashes include speeding, not using seat belts, ejection, higher speed limit and bad weather
Chen and Jovanis (2000)	Loglinear regression	Driving during late night or early morning affects crash severity
Al-Ghamdi (2002)	Logistic regression	The odds of severe injury increase by over-speeding
Chang and Wang (2006)	Classification and regression trees	Pedestrians, motorcycle, and bicycle riders have higher risks of injury than other drivers
Kononen et al. (2011)	Multivariate logistic regression	Seat belt use and crash direction are predictors of serious injury
Jacob and Anjaneyulu (2013)	Multiple models	Poisson regression yields the most accurate results for predicting fatal and injury crashes
Mooradian et al.(2013)	Multiple models	The PPO model performs best overall in terms of prediction accuracy
Hao and Daniel (2014)	Ordered probit model	Peak hour factor, visibility, car speed, train speed, driver's age, area type and traffic volume impact driver injury severity
Hao and Daniel (2015)	Mixed logit model	A speed limit reduction in inclement weather is effective in controlling injury severity
Ma et al. (2015)	Generalized ordered logit model	Season, time of day, location, tunnel length and adverse weather affect injury severity
Wang et al. (2016)	Multiple models	The boosted regression tree model outperformed the other models
Behnood and Mannering (2017)	Random parameters logit model	Age and gender are the most important determinants of injury severity
Fountas and Anastasopoulos (2017)	Random threshold hierarchical ordered probit model	Factors affecting injury severity include AADT per lane, speed, and vertical curve length
Ma et al. (2017)	Partial proportional odds model	Alcohol usage, speeding, weather, pavement and collision type affect injury severity

Osman et al. (2018)	Ordered response model	Factors affecting the crash severity include lack of seatbelt, speeding, wind, nighttime, weekday and female drivers

2.3.2 Non-parametric Models

This section presents non-parametric models, specifically ANN research done in crash prediction. ANNs have been widely applied in the transportation field. Their success is due to their ability to emulate the human brain and learning power. ANNs allow the inclusion of many variables, where irrelevant variables readily show negligible weight values, while relevant variables show significant weight values. In addition, no assumptions are required regarding the functional form of the relationship between predictor and response variables as the case is with the parametric methods.

Delen et al. (2006) investigated the injury severity experienced by drivers in crashes without limiting the study to any specific geographic area of the United States. They developed eight binary neural models to classify accidents by level of injury severity from no-injury to fatality and conducted sensitivity analysis to identify the prioritized importance of crash-related factors. The models investigated several factors related to injury severity level such as driver age, gender, alcohol consumption, seat-belt use, daylight/no daylight, rollover occurrence, type of crash, and weekends/weekdays. In general, the use of a restraint system like a seat belt, use of alcohol or drugs, age and gender, and vehicle role in the accident were found to have an important influence on the outcome of the crash. Also, weather conditions or the time of the crash did not seem to affect its severity. This result was deemed surprising by the authors who suggested it needs further study.

Other significant crash contributing factors were investigated by Moghaddam et al. (2010). They used a series of artificial neural networks to estimate crash severity and to

identify significant crash-related factors on urban highways. The results illustrated that highway width, head-on collision, type of vehicle at fault, ignoring lateral clearance, following distance, inability to control the vehicle, violating the permissible velocity and deviation by drivers are the most significant factors that increase crash severity in urban highways. Their study showed that feedforward backpropagation neural networks yield the best results. Fatalities and injuries were modelled together, and property damage crashes were modelled separately. Their findings suggest that changes in crash severity does not occur necessarily by any single dependent parameter but occurs as a simultaneous result of changes of these parameters.

Akin and Akbas (2010) also developed an ANN to predict intersection crashes in Macomb County in Michigan by grouping the crashes into three types: fatal, injury and PDO accidents. They modelled the relationship between the crash type and crash properties such as time, weather, light condition, surface condition and driver and vehicle characteristics using 16,000 crash records. The results showed that the likelihood of being involved in a crash is highest at intersections, the last working day of the week witnesses the highest crash probability, crash occurrence in the afternoon peak is almost twice as high was that in the morning peak, and crash occurrence increases with heavier traffic volumes.

The ANN's advantage over traditional parametric models was investigated by Zeng and Huang (2014). They proposed a convex combination (CC) algorithm to train a neural network (NN) model for crash injury severity prediction, and a modified NN pruning for function approximation (N2PFA) algorithm to optimize the network structure. The proposed approach was compared with the NN trained by traditional backpropagation (BP) algorithm using a two-vehicle Florida crash dataset from 2006. The results showed that the

CC algorithm outperformed the BP algorithm both in convergence ability and training speed. Both neural networks had better fitting and predicting performance than the ordered logit model, which again demonstrates the NN's superiority over statistical models for predicting crash injury severity.

Several studies developed parametric and non-parametric models using the same set of data to compare model performances. Abdelwahab and Abdel-Aty. (2001) examined the relationship between driver injury severity and driver, vehicle, roadway, and environment characteristics using the multilayer perceptron (MLP) and fuzzy adaptive resonance theory (ART) neural networks. Accident data focusing on two-vehicle accidents at signalized intersection for 1997 for the Central Florida area was used. They classified an accident into one of three injury severity levels using the readily available crash factors. The percentage of correct classifications of MLP neural network was compared to that of the ordered logit model. Their results revealed that MLP accurately classified 65.6% and 60.4% of cases for the training and testing phases, respectively, whereas the ordered logit model correctly classified 58.9 and 57.1% of cases for the training and testing phases, respectively. Results showed that female drivers are more likely to experience a severe injury than are male drivers, and male drivers are more likely to experience fatalities. In addition, drivers at fault are less likely to experience severe injury than those not at fault, and drivers in passenger cars are more likely to experience a greater injury severity level than drivers of vans or pickup trucks.

Iranitalab and Khattak (2017) compared the performance of four statistical and machine learning methods including multinomial logit (MNL), nearest neighbor classification (NNC), support vector machines (SVM) and random forests (RF), in

predicting traffic crash severity. Two-vehicle crashes were extracted as the analysis data from the 2012–2015 crash data from Nebraska. The proposed approach showed that NNC had the best prediction performance overall and in more severe crashes. RF and SVM had sufficient performances, and MNL was the weakest performing method. Table 2.4 summarizes the key findings of studies using non-parametric models discussed in this section.

Table 2.4 Key Findings of Non-parametric Models

Authors (year)	**Model**	**Key Findings**
Abdelwahab and Abdel-Aty (2001)	MLP and ART neural networks	Female drivers are more likely to experience a severe injury, and male drivers are more likely to experience fatalities; drivers in passenger cars are more likely to experience a greater injury severity
Delen et al. (2006)	Binary neural model	The use of a seat belt, use of alcohol or drugs, age and gender, and vehicle role in the accident influence the outcome of the crash. Weather conditions or the time of the accident do not affect the severity
Akin and Akbas (2010)	Artificial neural networks	The likelihood of a crash increases on the last working day of the week, in the afternoon peak, and with heavier traffic volumes
Moghaddam et al. (2010)	Artificial neural networks	Highway width, head-on collision, type of vehicle at fault and speeding are significant factors that increase crash severity in urban highways
Codur and Tortum (2015)	Artificial neural networks	The degree of vertical curvature is the most important parameter that affects the number of accidents on highways.
Iranitalab and Khattak (2017)	Multiple machine learning models	Nearest Neighbor Clarification (NNC) had the best prediction performance overall and in more severe crashes, and Multinomial Logit (MNL) was the weakest performing method.

2.4 Summary

This chapter presented a thorough literature review on crash injury severity prediction models and data used in crash analysis. The findings and conclusions of the comprehensive literature review are summarized herein.

Various factors, in the areas of geometric design, traffic flow, driver and environment, have been taken into consideration to predict either crash likelihood, crash injury severity, or both. Efforts to improve traffic safety are helped by mathematical models that allow researchers to better assess the effect of those factors on crashes. There is no unanimity on how to evaluate crash injury severity and the literature has contradicting assessments on the best model fit to predict crash outcomes, which is due to the random nature of factors and major differences in environments and datasets.

Panel data crash studies and real-time data crash studies were evaluated and discussed. Based on the findings and the fast technology advancement, the proposed model uses real-time data as it intends to deliver reliable predictions for crash injury severity. The traffic data collection technologies utilizing floating-car concepts have improved rapidly in the past few years, and such improvements provide an opportunity to address the challenges of the existing models and increase the accuracy of the predictions generated by the proposed model.

Crash prediction models are either parametric or non-parametric, and while the parametric conventional models such as regression, logit, and Poisson models have been widely explored in previous research, non-parametric models are still being discovered and calibrated to accommodate crash analysis. Statistical regression models have been extensively employed to analyze injury severity of crashes. However, most regression

models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables, and if these assumptions are violated, the model could result in erroneous estimations. Machine learning techniques can recognize patterns and adjust dynamically with gained prominence and maturity. ANNs specifically can relate input with output and automatically generate identifying characteristics from the learning material that they process. ANNs used in crash modelling, discussed in section 2.3.2, are capable to predict and present desired results despite limited data sets, and traffic forecasting complications involving random and complex variables can therefore be successfully resolved.

The thorough review presented in this chapter reveals that the perfect model for crash prediction does not exist, and different models present benefits and limitations depending on the data availability and study environment. For predicting dichotomous outcomes, logistic regression has become known as the statistical method of choice. In general, parametric models are simple and transparent. The magnitude of the influence of factors on crashes is directly determined by coefficient weights provided by a regression analysis. ANNs represent a newer technique and a potential alternative to regression analysis. Non-parametric models have several advantages over conventional parametric models and are good data-driven approaches to predict crash injury severity. The theoretical advantage of ANNs is that relationships need not be specified in advance since the method itself establishes relationships through a learning process. Research has been done to compare the performances of ANN and traditional statistical models (Kumar, 2005; Pao, 2006; Wang & Elhag, 2007; Zhang, 2001; etc.). Most researchers find that ANNs can outperform linear models under a variety of situations.

Therefore, a base statistical model (LRM) is developed in this study in addition to the ANN to serve for comparison and assessment purposes. Although discrete choice models are better at predicting a specific injury severity level, linear regression is the most used model in quantifying cash risks. It is also more natural and easier, and its use and effectiveness will be further investigated in this study.

The proposed model introduces the use of speed variance in the crash injury severity prediction. The literature has opposing views on the effect of speed and speed variation on crash occurrence and injury severity. The ongoing discussion about the speed-crash relationship is not conclusive due to the multitude of factors that affect crashes under different circumstances, and additional research is needed to gain more insight on this debate. Both models are developed using the same data to investigate their performances experimentally in a data-driven context. An assessment of the results is conducted based on each model's capacity and their advantages and disadvantages are discussed thoroughly.

CHAPTER 3

DATA ACQUISITION

3.1 Database Overview

To develop a sound model for predicting crash injury severity, significant amount of data is required under different categories as discussed in Chapter 2. The data in this study was collected from four different databases to form one comprehensive working database. The first round of data collection was carried out using data for the year 2016 on New Jersey freeways. After collecting crash data from the Plan4Safety database, the number of crash records and the corresponding information was not enough to draw reliable conclusions from model development, training and testing in terms of unsatisfactory number of crashes. The NJDOT crash records retrieved from the NJDOT accident webpage: (https://www.state.nj.us/transportation/refdata/accident) were therefore used as the new source of crash data for this study during the second round of data collection. The data were collected for the year 2017 on NJ freeways using the following data sources:

- Crash records: Four summary reports retrieved from NJDOT were combined to obtain comprehensive crash related information on NJ freeways in 2017.
- Road geometry data: The New Jersey straight line diagrams (SLD) database includes road type, road characteristics, number of lanes, posted speed limit and median type.
- Traffic volume data: The New Jersey Congestion Management System (NJCMS) database includes passenger-car and truck volumes collected from the sources of big data.
- Floating-car data: The INRIX database provides traffic speeds for freeway segments under normal and crash conditions, which are collected by floating-car technologies.

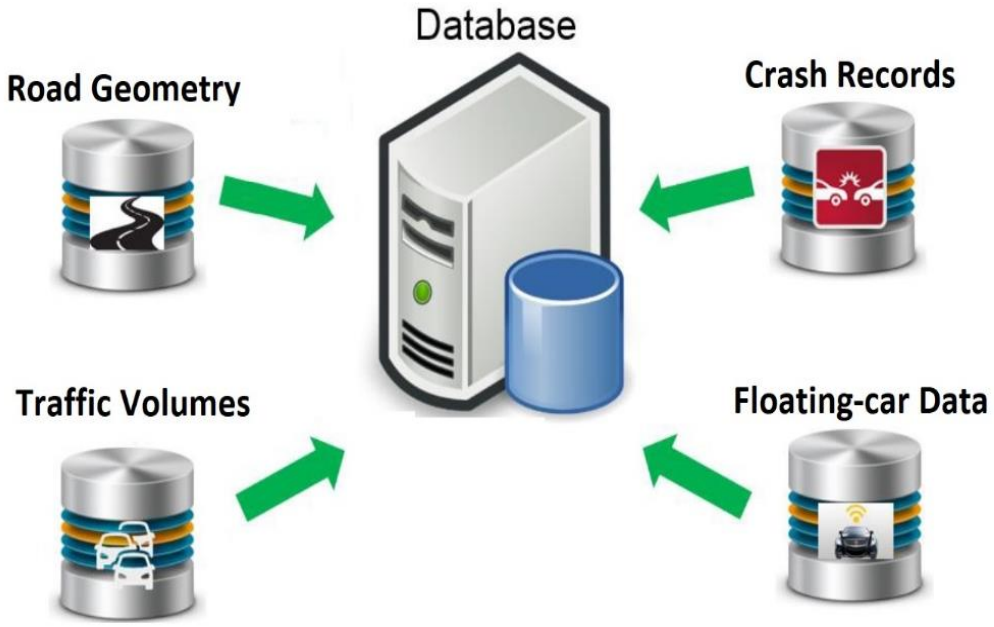


Figure 3.1 Process of database consolidation from four data sources.

3.2 Data Sources Description

This section presents an overview of the four databases used to develop the final working database. We note that the process to define the working database is based on the availability and applicability to predict the crash injury severity as required by the proposed prediction models. The database was developed using the advanced computing resources, which provided adequate data storage and computing processing to handle the large data resources necessary to process and execute the proposed prediction model.

3.2.1 Crash Record Database

The crash data was obtained from the NJDOT summary reports, which provide detailed police reports of crashes occurring in New Jersey for a specific year. Four summary reports were retrieved from NJDOT for freeway crashes in 2017: crash table, driver table, vehicle table and occupant table. These summary reports are available to the public for years

between 2001 and 2019, and the NJDOT is currently working on the 2020 data. Combined through an identical case number, the four tables are merged into a crash record database including a total of 122 crash related entries.

The crash injury severity in the state of New Jersey is classified into three levels: Fatal crash – Any crash that results in one or more fatal injuries, injury crash – Any crash that results in one or more non-fatal injuries, and property damage crash – Any crash that does not result in injuries or fatalities. When a crash results in one or more injuries, the injury level is specified using the entry "type of most severe injury entry". There are eight possible numerical values for the type of injury shown in Table 3.1. Similarly, other entries in the crash database are assigned numerical codes. The interpretation of these codes is provided by the NJDOT.

Table 3.1 Numerical Codes for Most Severe Physical Injury

Numerical Code	Most Severe Physical Injury
1	Amputation
2	Concussion
3	Internal
4	Bleeding
5	Contusion/Bruise
6	Burn
7	Fracture/Dislocation
8	Complaint of Pain

Source: https://www.state.nj.us/transportation/refdata/accident/pdf/NJTR-1_Overlays.pdf. Retrieved on January 19, 2019.

3.2.2 NJ-SLD Database

The roadway inventory and geometry data of each crash event (e.g., total number of lanes and posted speed limit), was retrieved from the NJDOT SLD. The SLD, initially designed as a planning tool, is a one-dimensional graphical depiction of a section of roadway and its

related data which includes the Interstate freeways, the US highways, and the State routes. The SLD information management system, including the data repository and software, is maintained by NJDOT's Bureau of Transportation Data Development (BTDD). By using mileposts, the main geometric characteristics of the crash location such as the posted speed limit and the number of lanes at the can be identified.

3.2.3 NJCMS Database

The traffic flow data, necessary for the analysis of crash impacts, were obtained from the NJCMS database. The NJCMS is a data management and data analysis system used primarily by the Bureau of Systems Planning to forecast congestion and propose mitigation measures for New Jersey roadways. The roadway links in the NJCMS tables are identified by SRI or Route Name (e.g., I-80, or I-195), and by start and endmileposts. The link information stored in NJCMS was tied to crashes identified in the crash record database using these unique link identifiers.

3.2.4 Floating-car Database

The traffic speed data used for model development are historical speed data from INRIX. The historical INRIX speed data is anonymously collected from GPS-enabled vehicles and mobile devices through Traffic Message Channel (TMC) and compiled into 1-minute-average speed. This historic 1-minute speed data were aggregated into 15-minute speed data for each TMC upstream of each crash. There are nearly 1,200 directional predefined TMCs on the New Jersey interstate freeways. The INRIX raw data was collected for 24 hours a day, over a 1-year period from January 2017 to December 2017. This period, including

weekdays, weekends, peak and non-peak hours, reflects real traffic conditions before, during and after a crash occurred.

3.3 Data Processing

As technology advances and safety becomes more of a priority for DOTs, the process of data collection before and after crashes becomes more inclusive. In recent years, new parameters are gathered by police reports and floating-car technologies that could be included in new research for better crash prediction accuracy. These parameters include accurate real-time speed measures and traffic counts, safety equipment available in the vehicle, safety equipment used at the time of the crash, airbag deployment, detailed occupant information, charges and summons, hazmat involvement, alcohol use and cell phone use. A larger database can be used to develop comprehensive crash prediction models with increased precision.

The process to define the final database that is utilized by the models developed in the following chapter was based on the applicability to predict the crash injury severity level as required by the proposed prediction model. The data was evaluated in terms of data structure, compatibility and usability for the models. Since the models are intended to support state and local traffic management agencies during operations to reduce crash injury severity levels while using real-time data, only the parameters that are known to the agencies in real-time before the crash are included in the working database.

The major issues encountered during data processing are described below.

1. Crash Record: Discrepancies in time and even the location of observed crashes reported in the database were observed. When speed measures were added, a few inconsistent records showed normal speed under a severe crash condition.

These entries were screened out manually and neglected in the model development process.

2. INRIX: While INRIX reported speed was on a TMC basis, which has only starting and ending coordinates, the corresponding crash in Crash Record is based on the SLD. These two data sources could not be cross-referenced with each other. Therefore, a conversion methodology to associate INRIX TMC information and SLD information based on SRI and mileposts was developed, and the database merged the INRIX TMC data to the SRI-based Crash Record data.

3. NJCMS: The SRI in NJCMS is not directional. Since the SRI is a key factor for merging databases and for model development, the NJCMS records had to be manually updated.

3.4 Calculation of Traffic Volume and Weighted Speed Variance

As discussed in previous studies (Solomon, 1974; Lave, 1985; Garber and Gadiraju, 1989; Kloeden et al., 2001; Kloeden et al., 2002), many factors affect crashes including roadway design, traffic speeds, traffic density, and vehicle mix and speed variance. Speed averages alone are not enough to predict the likelihood and/or crash injury severity of a crash, and speed variance also plays an important part in crash occurrence and severity.

In addition to the data entries processed from several databases described in this chapter, two factors are calculated and added to the working database: the traffic volume and the weighted speed variance. The calculation procedure and assumptions are retrieved from a previous study (Abisaad and Chien, 2018) and discussed in this section.

3.4.1 Traffic Volume

The volume counts used in this study are obtained from NJCMS. For each TMC, the passenger car count is reported as well as a separate truck count for vehicles with more than two axles. The volume is assumed to be equally divided over one-minute intervals since no

other data is available. The volume of passenger cars is added to the volume of trucks to obtain a total volume:

$$X_i = V_{Ci} + V_{Ti} \tag{3.1}$$

Where V_{Ci} and V_{Ti} are respectively the volumes of cars and trucks in the traffic stream at minute i, and X_i is the total volume of both cars and trucks at minute i.

3.4.2 Weighted Speed Variance

As discussed in Chapter 2, speed variance plays an important role in crash occurrence and injury severity. In order to identify the best speed variance generated from a specific time interval, the intervals were classified into five categories prior to the time of the incident: two minutes, four minutes, six minutes, eight minutes and ten minutes. The following assumptions are made:

- The speed profile does not vary greatly during these small two-minute intervals. This assumption is reasonable since traffic patterns generally require longer than a few minutes to shift, such as from peak to off-peak conditions.
- Speed variation patterns more than ten minutes before a crash occurs do not have a direct effect on its likelihood and severity.
- The beginning timestamp of the incident recorded in the database is the moment when the crash occurred. This is a bold assumption since this data is logged manually by the individual filing the report and no advanced technology confirms that timestamp.

The 2-minute, 4-minute, 6-minute, 8-minute and 10-minute average speeds are computed as follows:

$$V_n = \frac{\sum_{i=1}^{n} v_i X_i}{\sum_{i=1}^{n} X_i} \quad (3.2)$$

Where V_n is the speed average over n minutes before a crash occurred, n = 2, 4, 6, 8 or 10. For example, V_2 is the average speed over the two minutes before the crash occurred, while V_{10} is the average speed over the ten minutes before the crash occurred, and v_i is the INRIX reported speed at minute i.

Since the length of TMC varies, speed variances computed over shorter TMCs will have a greater effect than those computed on larger TMCs. Different TMC lengths should be accounted for in the computation of speed variance by dividing it by the length of the TMC. The weighted speed variances are then calculated for each specific time interval as follows:

$$S_n = \frac{\sum(V_n - v_i)^2}{(n-1)*l} \quad (3.3)$$

Where l is the length of a TMC where the speed and volume were reported. The four-minute weighted speed variance is used to create the final database for model development.

To illustrate that the weighted speed variance before the crash is larger than that in a similar non-crash time, a historical crash example was used, and speed variances were calculated on a four-minute interval basis with and without crash. A crash was reported on the eastbound Interstate 287 on April 8 at 15:01. It occurred under dry weather conditions and resulted in property damage only and a right shoulder closure for thirty minutes. Table 3.2 illustrates the spatiotemporal speed variances for the crash situation versus a non-crash

situation. The non-crash situation speed variance is calculated at the same timestamps during a typical day where no crash occurred. The table shows the highest speed variances on the crash day, and lower speed variances on the regular day. The table also shows that speed variance was high prior to a crash, and monitoring speed in the traffic stream can raise a flag about potential crashes.

Table 3.2 Spatiotemporal Weighted Speed Variance with and without Crash

	Weighted Speed Variance (With Crash)					
	Timestamp					
Distance to Accident (miles)	**14:57-15:00**	**14:56-14:59**	**14:55-14:58**	**14:54-14:57**	**14:53-14:56**	**14:52-14:55**
0.73	5.02	4.56	4.03	3.81	4.02	3.96
3.43	1.82	1.25	1.62	1.19	1.24	0.84
3.86	2.1	1.28	1.67	1.24	0.9	0.67
4.11	0.59	0.87	1.24	0.24	0.56	0.87
5.5	0.47	0.24	0.37	0.92	1.28	1.57
	Weighted Speed Variance (Without Crash)					
0.73	0.45	0.63	0.06	0.17	0.24	0.99
3.43	0.49	0.87	0.12	0.29	0.41	1.69
3.86	0.62	1.07	0.36	0.29	0.41	1.24
4.11	0.83	0.83	1.25	0.72	0.69	0.89
5.5	0	0.25	0.19	0.2	0.2	0.19

3.5 Final Database

The final database includes all New Jersey freeway crashes documented in 2017. A total of 16,649 crashes were distributed over nine freeways as shown in Table 3.3.

Table 3.3 Crash Distribution on Freeways by Level of Injury Severity for 2017

Freeway Location	Miles in NJ	Number of crashes	Fatalities	Injuries	PDO
I-278	1.3	5	0	1	4
I-676	6.9	204	3	56	145
I-76	3.08	492	1	99	392
I-280	17.85	1605	2	350	1253
I-195	34.17	631	2	140	489
I-295	68.1	2407	12	562	1833
I-78	67.83	3326	13	732	2581
I-287	67.50	3618	5	698	2915
I-80	68.54	4261	16	942	3303
Total	335.27	16549	54	3580	12915

New Jersey is a regional corridor for transportation since it is located between two major metropolitan centers: New York City and Philadelphia. Its freeways carry large volumes of interstate and intrastate traffic and goods. The Interstate system includes 431 miles and carries around 20 percent of vehicle travel in NJ.

The NJTP and the GSP are not included in New Jersey's Interstate highway network. They are however two of the busiest highways in the United States (Meyer, 2018). These two toll roads are maintained by the New Jersey Turnpike Authority (NJTA). On both roadways, crashes remain a critical issue as they withstand a high number of crashes per year. Figure 3.2 shows the number of crashes by injury severity between 2013 and 2017.

Figure 3.2 Crash numbers by year or NJTP and GSP.

Due to the difference of configurations and patterns on New Jersey freeways, crash data gathered in different locations is too large to be accommodated by one model. The factors associated with crashes are arbitrary and complicated by nature and combining all freeway data under the same model adds complication to the model's predictions and leads to inaccurate results. Therefore, crash data gathered on I-80, the freeway with the most crashes, were chosen for model development and evaluation in the following chapters. The crash data recorded on one freeway is homogenous in terms of travel configurations which increases the accuracy of predictions.

The Interstate 80 (I-80) in an east-west coast-to-coast freeway that runs from downtown San Francisco, California, to Teaneck, New Jersey. It is the second-largest

Interstate Highway in the United States, following I-90. The segment of I-80 that runs through New Jersey is also known as the Christopher Columbus Highway and the Bergen-Passaic Expressway. This segment runs for 65.84 miles from the Delaware Water Gap Toll Bridge at the Pennsylvania state line to Teaneck, Bergen County. I-80's designated end in four miles short of New York City, where the New Jersey Turnpike northbound begins. I-80 runs through rural areas of Warren and Sussex counties and continues through suburban surroundings in Morris County, and urban areas of Passaic and Bergen counties.

Table 3.4 Descriptive Statistics for Data Collected on I-80

Factors	Type	Description	Descriptive Statistics
Injury Severity Level	Nominal	0 = PDO; 1 = Mild Injury; 2 = Moderate Injury; 3 = Fatal/Severe Injury	86.8% (n = 3699); 8.0% (n = 340); 2.6% (n = 110); 2.6% (n = 112)
Month	Nominal	1 = January; 2 = February; 3 = March; 4 = April; 5 = May; 6 = June; 7 = July; 8 = August; 9 = September; 10 = October; 11 = November; 12 = December	7.2% (n = 308); 6.5 % (n = 277); 7.9% (n = 335); 6.0 % (n = 254); 8.6 % (n = 366); 8.1 % (n = 346); 7.4% (n = 316); 7.9 % (n = 338); 8.3% (n = 353); 11.8 % (n = 501); 10.6 % (n = 454); 9.7 % (n = 413)
Day	Nominal	1 = Monday; 2 = Tuesday; 3 = Wednesday; 4 = Thursday; 5 = Friday; 6 = Saturday; 7 = Sunday	14.9 % (n = 634); 16.0 % (n = 682); 16.6% (n = 706); 15.3 % (n = 654); 17.0 % (n = 724); 10.0 % (n = 428); 10.2 % (n = 433)
Peak Period	Binary	0 = Non-peak; 1 = Peak	45.1 % (n = 1923); 54.9 % (n = 2338)
Direction	Binary	0 = Westbound; 1 = Eastbound	50.4 % (n = 2148); 49.6 % (n = 2113)
Horizontal Alignment	Binary	0 = Curved; 1 = Straight	18.3 % (n = 780); 81.7% (n = 3481)
Road Grade	Binary	0 = Level; 1 = Grade	80.8 % (n = 3441); 19.2% (n = 820)
Surface Type	Binary	0 = Blacktop; 1 = Otherwise	91.2 % (n = 3885); 8.8% (n = 376)
Surface Condition	Binary	0 = Dry; 1 = Otherwise	75.8 % (n = 3231); 24.2% (n = 1030)
Light Condition	Binary	0 = Daylight; 1 = Otherwise	69.7 % (n = 2971); 30.3% (n = 1290)
Road divided	Binary	0 = Barrier Median; 1 = Other	85.5 % (n = 3643); 14.5 % (n = 618)
Environmental condition	Binary	0 = Clear; 1 = Other	76.8 % (n = 3274); 23.2 % (n = 987)
Temporary Traffic Control Zone	Binary	0 = TTCZ; 1 = None	1.7 % (n = 74); 98.3% (n = 4187)
Speed	Nominal	Speed (mph)	Std Dev. = 17.58; Mean = 50.97
Average Speed	Nominal	Speed (mph)	Std Dev. = 10.81; Mean = 56.45
Weighted Speed Variance	Nominal	(m/h^2)	Std Dev. = 0.32; Mean = 4.47
Speed Limit	Nominal	Speed limit (mph)	50, 55 or 65 mph
Milepost	Nominal	Milepost	Std Dev. = 16.84; Mean = 44.84
Number of Lanes	Nominal	One-way number of lanes: 2, 3 or 4	2.2 % (n = 91); 26.5% (n = 1130); 71.3% (n = 3040);

3.6 Summary

This chapter presented the data collection procedure followed to create an inclusive working database that will be used for model development in the next chapter. The final data was obtained from combining four resources: crash records, NJ-SLD, NJCMS and floating-car data. These databases were introduced and explained, and the challenges faced in the data collection procedure were discussed. The traffic volume and the weighted speed variance. The data development process yielded a final database consisting of crash information on New Jersey's interstates. This comprehensive data will be used to develop a base LRM and an ANN in Chapter 4.

CHAPTER 4

METHODOLOGY

This chapter explains the general structure and development of the Linear Regression Model (LRM) and Artificial Neural Network (ANN). The first section is dedicated to the LRM, the dependent and explanatory variables used to develop it and its final structure. The second section describes the procedure to develop the ANN along with its final proposed configuration.

4.1 Linear Regression Model Development

4.1.1 Dependent Variable and Explanatory Variables

This study is investigating the possible effect of different traffic-related factors on crash injury severity level. The explanatory variables or independent variables are used to predict or explain the behavior of the response variable or dependent variable. This section introduces the dependent variable and explanatory variables used in the model development. The same data are used for the LRM and the ANN development process discussed in subsequent sections. As such, the dependent and explanatory variables discussed in this section are applicable to both models.

Dependent Variable

The dependent variable is the one being measured and assessed. It represents the outcome resulting from changing input in the explanatory variables. The purpose of the model developed in this study is to predict the output of a crash in terms of injury severity, hence the dependent variable is “injury severity level”.

The "KABCO" injury scale was developed by the National Safety Council (NSC) and is frequently used by law enforcement for classifying injuries:

- **K –** Fatal;
- **A –** Incapacitating injury;
- **B –** Non-incapacitating injury;
- **C –** Possible injury; and
- **O –** No injury.

In New Jersey, these categories are defined by the State as follows:

- Killed: Victim is deceased. (Must check "Fatal" box at the top of the report)
- Incapacitated: Victim has a non-fatal injury. Cannot walk, drive or normally continue the activities that they could perform before the crash
- Moderate Injury: An evident injury, other than fatal and incapacitating. Injury is visible, such as a lump on head, abrasion, bleeding or lacerations
- Complaint of Pain: A reported or claims of injury that is not fatal, incapacitating or moderate. Injury is not visible to the investigating officer.
- No injury: No reported injury.

In the raw data retrieved from the NJDOT crash records website and discussed in Chapter 3, there are ten possible levels of injury severity resulting from a crash. In addition to the types of most severe injury listed in Table 3.1, there are two additional possibilities: no injury, also known as PDO, and fatality. In this model, the injury severity levels were narrowed down from ten to four categories. Several types of injury were combined under one category because there was not enough data to represent each type in a separate category. The definitions follow those provided by New Jersey except that the "K" & "A" (Killed and Incapacitated) categories were merged together under "Severe Injury".

Figure 4.1 Injury severity levels categories.

As shown in Figure 4.1, the dependent variable "injury severity level" can take four different values ranging from no injury to severe injury as follows: 0 = Property Damage Only (PDO); 1 = Mild Injury (or complaint of pain); 2 = Moderate Injury; 3 = Severe Injury. The first category (PDO) is when there are no reported injuries nor complaint of pain by anyone involved in the crash, but only injury to property resulting from the crash. The second category (mild injury) is when a complaint of pain is reported on the scene by a party involved in the crash. The third category (moderate injury) is when any of the following is reported on the scene: contusion, bruise, abrasion, burn, fracture or dislocation. The fourth category (severe injury) is when the crash leads to one or more fatalities or an amputation, concussion, or internal injury. The four categories are summarized in Table 4.1.

Table 4.1 Description of Injury Severity Levels

Injury severity level	Description	Reported injury
0	PDO	No injuries, no complaint of pain
1	Mild Injury	Complaint of pain/no visible injury
2	Moderate Injury	Contusion, bruise, abrasion, burn, fracture or dislocation
3	Severe Injury	amputation, concussion, internal injury or death

Since the LRM is incapable of predicting integers, its output will be a real number referred to as Injury Severity Index (ISI) in this research. The ISI will be converted into a real number representing one of the four categories cited in Table 4.1 by rounding to the nearest integer.

Explanatory Variables

An explanatory variable is a variable that is assumed to have an effect on the dependent variable. A change in the explanatory variable inflicts a change in the dependent variable. As discussed in Chapter 2, there is a multitude of factors that can cause crashes and/or lead to a higher injury severity level resulting from a crash.

Eighteen explanatory variables are tested to be included in the model development: month, day, peak period, direction, horizontal alignment, road grade, surface type, surface condition, light condition, road divider, environmental condition, temporary traffic control zone, speed, average speed, weighted speed variance, speed limit, milepost and number of lanes. These variables were chosen based on suggestions by the literature reviewed in Chapter 2 as well as the availability and applicability of the data discussed in Chapter 3. It is important to note that the prediction model developed in this research is intended to help traffic management centers during operations, and only real-time information available to decision makers prior to a crash is considered during model development. More detailed

data available after the crash is discarded from the analysis. Such data include information about the driver and the passenger(s), safety equipment available, safety equipment used, previous crash involvement, cellphone usage, alcohol usage, etc.

Like the case of the dependent variable, the possible values of explanatory variables were also merged to avoid the under representation of some rare categories. An example is given to further explain this procedure: the light condition can take seven different values according to the State of New Jersey police crash investigation report provided by NJDOT: daylight, dawn, dusk, dark (streetlights off), dark (no streetlights), dark (streetlights on, continuous), and dark (streetlights on, spot). These categories were reduced to only two categories: daylight and otherwise, separating crashes that happened under daylight conditions from all others, no matter the condition of streetlights.

The explanatory variable "Temporary traffic control zone" was not included among the potential explanatory variables used for model development. The majority of the data entries (98.3%) did not occur in a temporary traffic control zone and the remaining crash entries (1.7%) are not statistically sufficient for the explanatory variable to show a substantial effect on the injury severity level. Table 4.2 summarizes the final list of explanatory variables used in the following sections.

Table 4.2 Description of Explanatory Variables

Explanatory Variables	Type	Description
Month	Nominal	1 = January; 2 = February; 3 = March; 4 = April; 5 = May; 6 = June; 7 = July; 8 = August; 9 = September; 10 = October; 11 = November; 12 = December
Day	Nominal	1 = Monday; 2 = Tuesday; 3 = Wednesday; 4 = Thursday; 5 = Friday; 6 = Saturday; 7 = Sunday
Peak period	Binary	0 = Non-peak; 1 = Peak
Direction	Binary	0 = Westbound; 1 = Eastbound
Horizontal alignment	Binary	0 = Straight; 1 = Curved
Road grade	Binary	0 = Level; 1 = Grade
Surface type	Binary	0 = Blacktop; 1 = Other
Surface condition	Binary	0 = Dry; 1 = Other
Light condition	Binary	0 = Daylight; 1 = Other
Road divider	Binary	0 = Barrier Median; 1 = Other
Environmental condition	Binary	0 = Clear; 1 = Other
Speed	Nominal	Speed (mph)
Average speed	Nominal	Speed (mph)
Weighted speed variance	Nominal	(m/h^2)
Speed limit	Nominal	Speed (mph)
Milepost	Nominal	Milepost
Number of lanes	Nominal	One-way number of lanes: 2, 3 or 4

4.1.2 Step Procedure for Model Development

This section summarizes the step-by-step procedure followed to develop the LRM. A linear regression is a simple linear approach to modeling the relationship between a dependent variable and one or more explanatory variables. When only one independent or explanatory variable exists, the model is called simple linear regression. When there are several explanatory variables, the model is called multiple linear regression.

Linear regression is an attractive model because its representation is very simple and straight forward. The representation is a linear equation that combines a set of input values (x), the solution to which is the predicted output value (y). As such, both the input values

and the output value are numeric. The linear equation assigns a coefficient to each input value. If the goal is prediction or forecasting, linear regression can be used to fit a predictive model to an observed data set. The general equation for a multiple linear regression is formulated in Equation (4.1).

$$y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n \quad (4.1)$$

Where:

- **y** = the predicted value of the dependent variable
- $\mathbf{B_0}$ = the y-intercept (value of y when all other parameters are equal to 0)
- $\mathbf{B_1X_1}$ = the regression coefficient (B_1) of the first explanatory variable ($\mathbf{X_1}$)
- $\mathbf{B_2X_2}$ = the regression coefficient (B_2)of the first explanatory variable ($\mathbf{X_2}$)
- $\mathbf{B_nX_n}$ = the regression coefficient (B_n)of the last independent variable ($\mathbf{X_n}$).

Developing a linear regression model means estimating the values of the coefficients used in the representation with the available data. Once the regression model is developed, if additional values of the explanatory variables are collected, the fitted model can be used to predict the response. The use of many variables to predict one particular outcome is one of the most useful prediction techniques. In this section, a multiple linear regression model is developed where several explanatory variables are used to predict a single quantitative outcome. The procedure followed to develop the LRM is depicted in Figure 4.2.

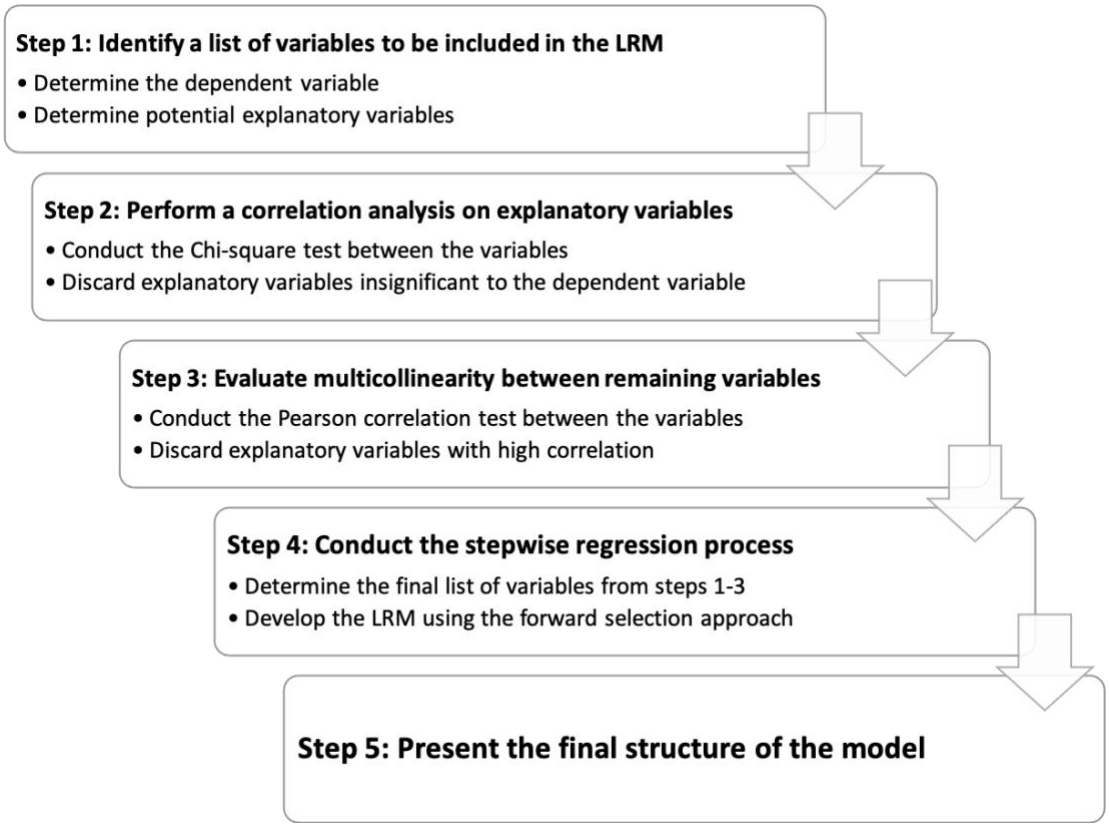


Figure 4.2 Step-by-step procedure for LRM development.

Data Filtering

The 5-step procedure depicted in Figure 4.2 was followed for a first attempt at the LRM development conducted using the entire crash database with a total of 4,261 crashes. This first model (LRM 1) yielded an R-squared value of 0.165. R-squared explains to what extent the variance of one variable explains the variance of the second variable. For instance, if the R^2 of a model is 0.90, then approximately 90% of the observed variation can be explained by the model's inputs.

There are no rules regarding what the minimum R-squared value should be as this varies between research areas. However, a value of 0.165 was considered low in this research as less than 20% of the observed variation was explained by the model, and a data

sorting process was conducted to narrow down data entries to be used in a second model (LRM 2) explained in detail in the following sections.

Data entries with inconsistent speed measures were removed from the 4,261 initial crashes to increase the R-squared value. The speed under crash conditions should be less than the speed under normal conditions, and if the data does not show it, it is assumed that the speed measure reported from INRIX is not accurate and hence will affect the accuracy of the model. Crashes where the speed measure was more than 5 mph higher than the average speed in normal conditions were discarded from the following analysis and development of LRM2

After removing the above-mentioned data entries, the remaining 2,966 crashes were tested to develop a new regression model. 2,966 crashes in 2017 on I-80 were randomly divided into three groups (i.e., 70%, 20%, and 10% of total crashes, respectively) for training, validation, and testing purposes. The model development process using the new filtered data is discussed next.

Correlation Analysis

This section discusses step 2 of the model development: correlation analysis between the dependent and explanatory variables.

Careful consideration of the correlation between explanatory variables and the dependent variable is required in regression analysis. Correlation analysis is a statistical method used to evaluate the relationship between two variables. A high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related. This method is strictly connected to the linear regression analysis.

The correlation is tested by conducting the Pearson's chi-squared test. Pearson's chi-square test (X^2) is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. Pearson's chi-square test was performed to evaluate the relationship between each potential explanatory variable and the dependent variable: injury severity level. The computational process for the chi-square test includes the following steps:

1. Calculate the chi-square test statistic X^2 as follows:

$$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{4.2}$$

Where f_0 = the observed frequency (the observed counts in the cells) and f_e = the expected frequency if no relationship existed between the variables.

2. Determine the degrees of freedom, df, of that statistic.

3. Select a desired level of confidence (significance level, p-value) for the result of the test.

4. Compare X^2 to the critical value from the chi-square distribution with degrees of freedom and the selected confidence level.

5. Sustain or reject the null hypothesis that the observed frequency distribution is the same as the theoretical distribution based on whether the test statistic exceeds the critical value of X^2.

6. If the test statistic exceeds the critical value of X^2, the null hypothesis can be rejected, and the alternative hypothesis (H1 = there is a difference between the distributions) can be accepted. If the test statistic is less than the critical X^2 value, then no clear conclusion can be reached, and the null hypothesis (H0 = there is no difference between the distributions.) is sustained but not necessarily accepted.

The chi-square statistic is computed using cross tabulation in SPSS. It can be evaluated by examining the p-value provided by the software. To make a conclusion about

the hypothesis with 95% confidence, the p-value of the chi-square statistic should be less than 0.05. If so, we can conclude that the variables are not independent of each other and that there is a statistical relationship between the explanatory variables. Assumptions are made when the chi-square test is conducted:

- The tested data is randomly picked from the population.
- The categories are mutually exclusive; each subject cannot fit in more than one category. For example, a crash that occurred on Monday cannot be duplicated on Tuesday.

Table 4.3 below shows the results of Pearson's chi-square tests on the dataset. Figure 4.3 shows the P-values of different independent variables.

Table 4.3 Pearson's Chi-square Test Results

Independent Variable	**Likelihood Ratio Tests**		
	Chi-square	**df**	**P-value**
Month	4.468	1	0.035
Day	0.004	1	0.95
Peak period	8.306	1	0.004
Direction	1.713	1	0.191
Milepost	0.652	1	0.419
Speed limit	1.018	1	0.313
Number of lanes	2.7	1	0.1
Horizontal alignment	0.76	1	0.383
Road grade	1.27	1	0.26
Surface type	5.538	1	0.019
Surface condition	9.689	1	0.002
Light condition	4.649	1	0.031
Environmental condition	2.275	1	0.131
Road divider	1.17	1	0.279
Speed	29.82	1	0.001
Average speed	0.458	1	0.499
Weighted speed variance	10.638	1	0.001

After discarding the variables that did not show significance at the 0.05 level, there are seven remaining significant explanatory variables. The explanatory variables with a P-value less than 0.05 are: Month, peak period, surface type, surface condition, light condition, speed, and weighted speed variance. These explanatory variables shown in Figure 4.3 are considered for model development.

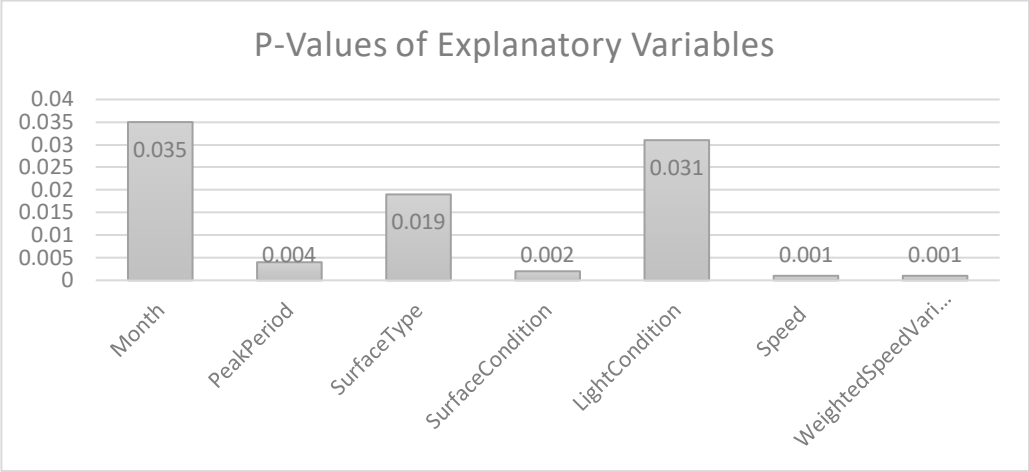


Figure 4.3 P-values of explanatory variables.

Multi Collinearity Analysis

The next step is to evaluate multi collinearity among the variables. Multi collinearity is an occurrence where one explanatory variable in a multiple regression model can be linearly predicted from other explanatory variables accurately. It creates redundant information and might offset the results when trying to determine how well each explanatory variable can be used most effectively to predict the dependent variable. An example of multi collinearity might exist between speed and traffic volume. When the traffic volume increases, the speed automatically decreases and vice versa, hence those two variables are dependent of each

other. The statistical implications from a model with multi collinearity may not be dependable.

The Pearson's rank correlation coefficients among the remaining variables are summarized in Table 4.4. The only slightly significant correlation exists between Light condition and Surface type. However, the Pearson correlation of 0.385 is not exceedingly high and both variables can still be included in the model. All seven explanatory variables will be considered for model development.

Table 4.4 Correlations of Explanatory Variables

Correlations								
		Month	Peak period	Surface type	Surface condition	Light condition	Speed	Weighted speed variance
Month	Pearson Correlation	1	0.18	0.085	-0.26	0.2	-0.01	0.026
	Sig. (2-tailed)		0.22	0.558	0.068	0.163	0.96	0.858
	N	2076	2076	2076	2076	2076	2076	2076
Peak period	Pearson Correlation	0.175	1	-0.185	-0.135	-0.074	-0.13	0.057
	Sig. (2-tailed)	0.224		0.199	0.35	0.612	0.35	0.693
	N	2076	2076	2076	2076	2076	2076	2076
Surface type	Pearson Correlation	0.085	-0.19	1	0.25	0.385**	-0.22	0.207
	Sig. (2-tailed)	0.558	0.2		0.08	0.006	0.12	0.148
	N	2076	2076	2076	2076	2076	2076	2076
Surface condition	Pearson Correlation	-0.26	-0.14	0.25	1	0.128	0.02	0.047
	Sig. (2-tailed)	0.068	0.35	0.08		0.374	0.88	0.746
	N	2076	2076	2076	2076	2076	2076	2076
Light condition	Pearson Correlation	0.2	-0.07	0.385**	0.128	1	-0.23	0.181
	Sig. (2-tailed)	0.163	0.61	0.006	0.374		0.11	0.209
	N	2076	2076	2076	2076	2076	2076	2076
Speed	Pearson Correlation	-0.007	-0.13	-0.224	0.023	-0.227	1	-0.04
	Sig. (2-tailed)	0.961	0.35	0.117	0.876	0.113		0.785
	N	2076	2076	2076	2076	2076	2076	2076
Weighted speed variance	Pearson Correlation	0.026	0.06	0.207	0.047	0.181	-0.04	1
	Sig. (2-tailed)	0.858	0.69	0.148	0.746	0.209	0.79	
	N	2076	2076	2076	2076	2076	2076	2076
**** Correlation is significant at the 0.01 level (2-tailed).**								

Stepwise Regression

The next step is to conduct the stepwise regression procedure with the explanatory variables left after eliminating multi collinearity in the previous step.

The regression model uses the stepwise method in which the choice of explanatory variables used in the final model is carried out by an automatic procedure. It involves adding or removing potential explanatory variables consecutively and testing for statistical significance after each iteration. In each step, a variable is considered for addition to or subtraction from the set of independent variables. The forward selection approach involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent. Figure 4.4 shows the iterative process of the stepwise regression used to reach the final model.

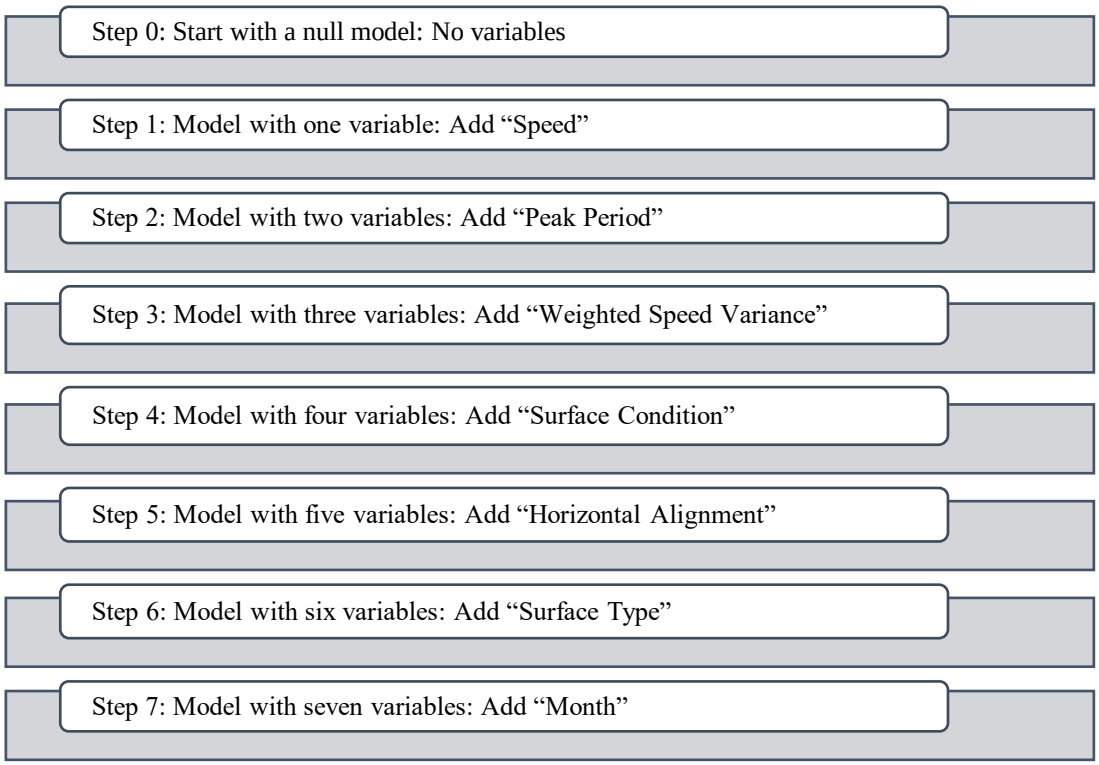


Figure 4.4 Stepwise regression process.

The results of the stepwise regression are shown in Table 4.5. The predictors from steps 1 through 7 are respectively: speed, peak period, weighted speed variance, surface condition, horizontal alignment, surface type and month. These are the seven explanatory variables that were found significant enough to the change in the dependent variable. The R-squared that the model reached using all variables is 0.422, meaning 42.2% of the values are explained by the developed LRM. The initial R-squared is 0.393 in step 1, using only one independent variable and the most significant to the model: speed. The stepwise method automatically adds other variables to the model and calculates the new R-squared based on that addition. It keeps adding explanatory variables until the addition no longer yields statistically significant results.

Table 4.5 Stepwise Regression Results

Step	R-squared	Adjusted R- squared	Std. Error of the Estimate	R- squared Change	Sig. F Change
1	0.393	0.393	1.077	0.012	0
2	0.405	0.404	1.071	0.01	0
3	0.415	0.414	1.066	0.002	0
4	0.417	0.415	1.066	0.003	0.035
5	0.42	0.419	1.064	0.002	0.012
6	0.422	0.419	1.063	0.002	0.041
7	0.422	0.420	1.063	0.002	0.036

LRM Final Equation

The linear regression model is presented in Equation (4.3):

$$y = 0.061 + 0.0212 * Sp - 0.0197 * PP + 0.0189 * WSV + 0.005 * SC + 0.0049 * LC + 0.0048 * ST - 0.0032 * Mo + \quad (4.3)$$

Where:

- y = the predicted value of the injury severity index
- Sp = Speed
- PP = Peak period
- WSV = Weighted speed variance
- SC = Surface condition
- LC = Light Condition
- ST = Surface type
- Mo = Month

While the R-squared of 0.422 is a large increase from the initial R-squared of 0.165, this value is still considered not good enough for a model to make reliable predictions as only less than half the values can be explained by the LRM. This low R-squared can be explained by several reasons: the most important being the data inaccuracy and the inadequacy of the LRM to absorb its complicated patterns:

- Most crash entries are PDO crashes. In that case, a linear regression is modeled to cover the bigger portion of the data and can therefore very poorly predict other crash injury severity categories (mild injury, moderate injury, and severe injury).

- The lowest number of crashes in the database is that of moderate injury crashes and severe injury crashes (2.6% each). The LRM tends to favor majority class over the minority class, and it is very difficult for the LRM equation to predict these injury severities as the speed would have to be exceedingly high.

- The data has inaccuracies which generally led to an insufficient model performance. The overall accuracy and power of the resulting LRM is relatively poor and must be further improved by improving the data quality if the model is to be used for a meaningful operational crash risk prediction.

- The linear regression is poor in handling classification problems, and this is further proven by the low value of R-squared. A better fit would be a discrete choice model as discussed in Chapter 2.

4.2 Artificial Neural Network (ANN)

This section introduces a non-parametric approach for predicting crash injury severity by an artificial neural network (ANN) model. ANNs are computer programs designed to simulate the way in which the human brain processes information. ANNs aggregate their knowledge by recognizing the patterns and relationships in data and learning through experience.

An ANN consists of artificial neurons aggregated into layers. It models the neurons in a human brain by transmitting signals between its artificial neurons. It has been proven that an ANN with one hidden layer can approximate any finite nonlinear function with very high accuracy. This study adopts several forms of ANNs to find the best performing model in predicting crash injury severity. The typical structure of an ANN consists of an input

layer, an output layer, and one or more hidden layers. Figure 4.5 depicts the structure of a typical ANN.

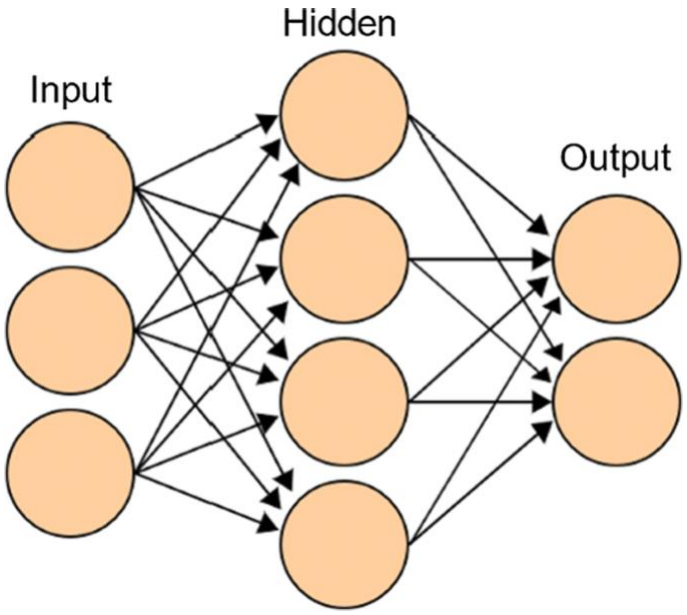


Figure 4.5 General structure of an ANN.

ANN Development

The Neural network toolbox in MATLAB was used for developing the ANN. For consistency purposes, the same crash data that was used in the LRM development are used for the ANN development. 2,966 crashes in 2017 on I-80 were randomly divided into three groups (i.e., 70%, 20%, and 10% of total crashes, respectively) for training, validation, and testing purposes. The model is initially built on training algorithms that adjust ANN weights of all explanatory variables using the training set. The fitted model is used to predict the outcome in the validation set that provides an independent evaluation. The testing set is designed to give an assessment of the ANN's performance when the entire design procedure is completed. Figure 4.6 illustrates the workflow for ANN design.

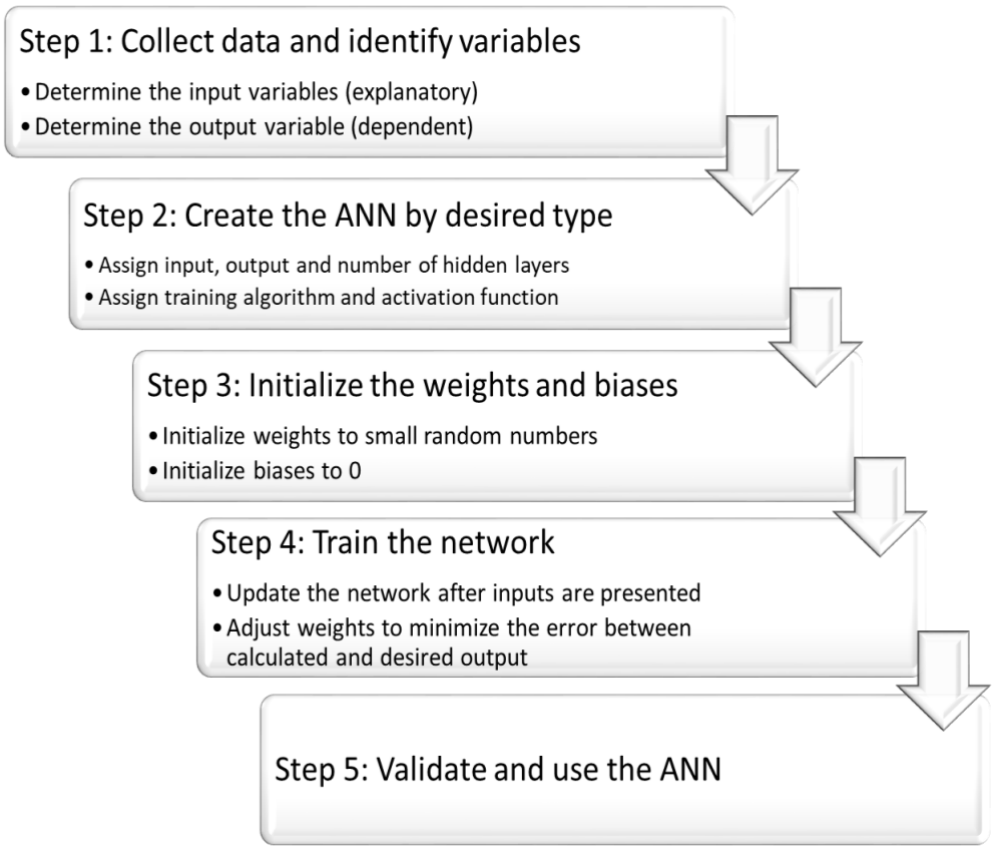


Figure 4.6 Workflow steps for ANN design.

4.2.1 ANN Types

Step 1 of the ANN development includes data collection and variable identification. This was covered in Section 4.1.1. The same dependent and explanatory variables used for the LRM development are used for the ANN development. The next step is to create the desired type of ANN. This includes training algorithm, activation function, input, output and number of hidden layers.

In this study, two types of ANNs are tested: a three-layer Feed-Forward (FF) network consisting of an input layer, a hidden layer, and an output layer, and a four-layer Deep Feed-Forward (DFF) network consisting of an input layer, two hidden layers, and an output layer. By having multiple hidden layers, the ANN can compute more complex

functions. The number of hidden layers determines the depth of the neural network. In general, deeper networks can learn more complex functions. They do however require more computation time.

Training Algorithms

The training (or learning process) of an ANN is carried out by determining the difference between the predicted output of the network and the actual output. This difference is known as the error. The network then adjusts its weighted associations using this error value. Consecutive corrections cause the ANN accuracy to increase until a small acceptable error is reached, and the training can be ended.

Several training algorithms can be used to determine the ANN weights. While every algorithm has advantages and disadvantages, the Resilient Backpropagation (RB) is the fastest on pattern recognition problems, whereas Levenberg-Marquardt (L-M) has the fastest convergence on function approximation problems. Both these algorithms in addition to Scaled Conjugate Gradient (SCG) and Variable Learning Rate Propagation (VLRP) are used in this study.

Activation Functions

Activation functions introduce non-linear properties to the ANN. They convert the input signal of a node to an output signal. That output signal is then used as input in the next layer. Between the i-th neuron of one layer and the j-th neuron of the next layer, the sum of products of inputs and their corresponding weights is calculated in Equation (4.4), and the activation function is applied to it to get the output of that layer and feed it as an input to the next layer as in Equation (4.5). This process repeats itself until the output of the final layer is generated.

$$S_i = \sum_{i=1}^{n} W_{ij} X_i \tag{4.4}$$

$$O_i = F(S_i) \tag{4.5}$$

Where X_i is the input of neuron i

- W_{ij} is the weight coefficient between neuron i of one layer and neuron j of the next layer;
- O_i is the output of neuron I;
- F is the activation function.

The ANN calculates the difference between its calculated output and the desired output and uses backpropagation to minimize the error until it reaches the minimum value. Several activation functions are used in this study: identity, Softplus, Softmax, and ReLU (Rectified linear unit).

Input, hidden and output layers

Based on potential risk parameters suggested by previous studies and data availability, the input layer of the ANN consists of 17 neurons representing the explanatory variables discussed in Section 4.1.1. The number of explanatory variables was not narrowed down like it was in the process of LRM development by conducting statistical tests, as the ANN is a smart and dynamic tool that can recognize explanatory variables that do not have a significant effect on the dependent variable and automatically assign them negligible weights.

The hidden layers are located between the input and output of the network. They perform nonlinear transformations of the entered inputs to compute an output. There are several methods for determining the number of neurons in the hidden layers. The most commonly used are the following:

- The number of hidden neurons should be between the number of neurons in the input layer and the number of neurons in the output layer.
- The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.

The number of hidden neurons used to develop the ANN is 12 as it satisfies both conditions.

The output layer consists of one neuron: the estimated dependent variable "injury severity index". The output of the ANN is a number between 0 and 3 representing the potential crash injury severity index given a crash has occurred at a given location. Figure 4.7 is a simplified chart of the ANN computation procedure example.

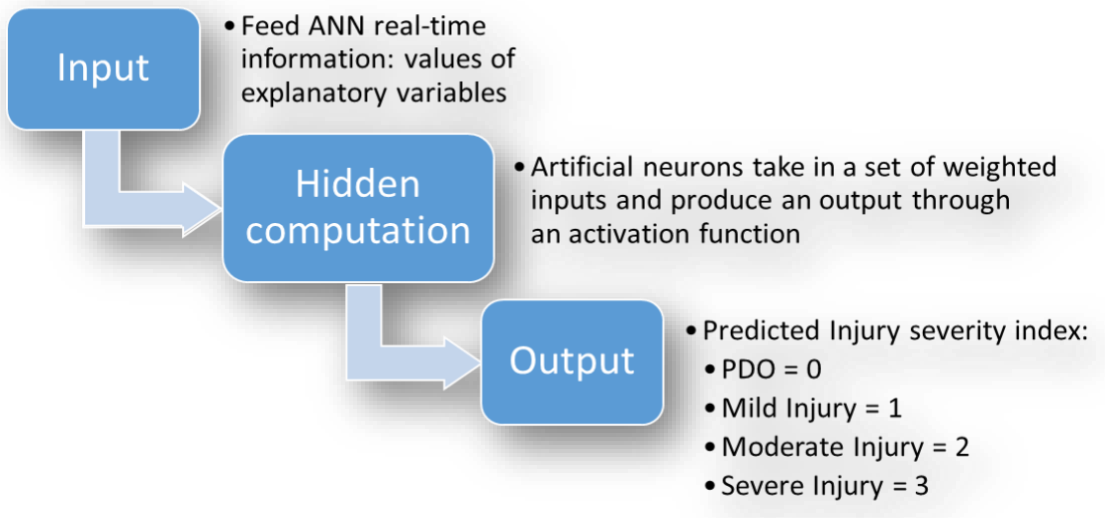


Figure 4.7 ANN computation procedure.

4.2.2 Weight and Bias Initialization

When the inputs are transmitted from layer to layer between neurons, the weights and biases are applied to the inputs. Weights control the signal between two neurons. A weight decides how much influence the input will have on the output. Biases are an additional input into the next layer that always have the value of 1. The bias unit guarantees that even when all the inputs are zeros there will still be an activation in the neuron. In general practice biases are initialized with 0 and weights are initialized with random numbers.

4.2.3 Network Training

ANNs are trained by processing examples that contain a known input and output. This forms an association between input and output which is stored within the ANN. The training of an ANN is conducted by determining the difference between the predicted output of the network and a target output, also known as the error. The network then adjusts its weighted associations according to a learning rule using this error value. This iterative process and adjustments will cause the neural network to produce output which is increasingly similar to the target output. After enough iterations, the training can be terminated based upon certain criteria. The training is conducted using the training algorithms discussed in Section 4.2.1. Figure 4.8 depicts the simplified structure of ANN training.

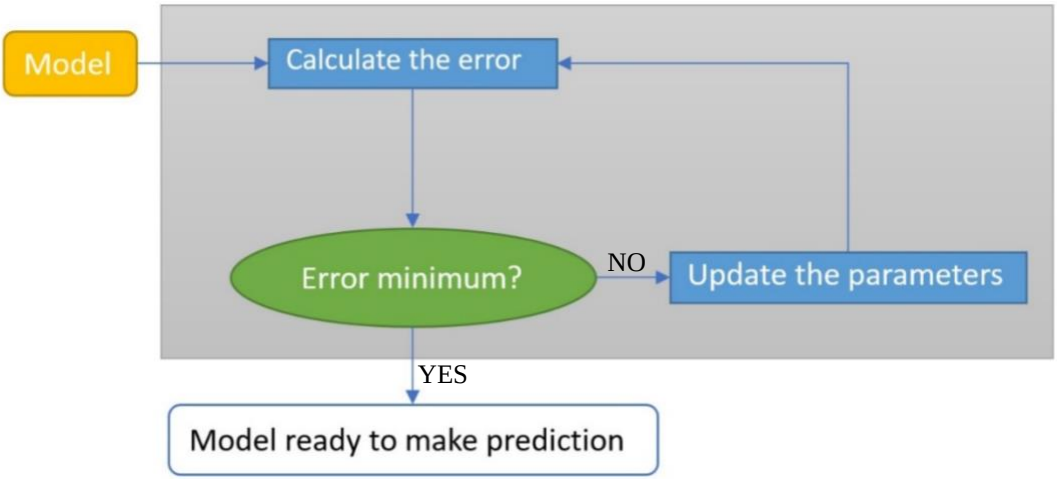


Figure 4.8 Simplified structure of ANN training procedure.

4.2.4 Network Validation and Final Structure

A validation dataset is used to tune the architecture of the ANN. As previously mentioned, I-80 crashes were randomly divided into three groups (i.e., 70%, 20%, and 10% of total crashes, respectively) for training, validation, and testing purposes. The validation dataset functions as a hybrid: it is training data used for testing, but neither as part of the low-level training nor as part of the final testing. Since our goal is to find the network having the best performance on crash data, the simplest approach to the comparison of different networks is to evaluate the error function using data which is independent of that used for training. The performance of the networks is then compared by evaluating the error function using an independent validation set, and the network having the smallest error with respect to the validation set is selected. The Root Mean Square Error (RMSE) was used as an index to determine the optimal ANN type, training algorithm, activation function and numbers of hidden layers. The lower the RMSE value, the better the model performance. Using different ANN models, each having a different number of hidden layers, input variables, training algorithm, and activation function, a total of 11 ANN were tested, and the best

ANN was chosen based on the lowest RMSE shown in Table 4.6. This RMSE was computed from the validation dataset (20% of crash data).

Table 4.6 RMSEs of Various ANN Models

ANN	Hidden layers	Activation function	RMSE
1	1	Identity	0.99
2	1	Softplus	1.05
3	**1**	**Softmax**	**0.81**
4	1	ReLU	0.88
5	1	Identity	0.96
6	1	Softplus	1.11
7	1	ReLU	0.96
8	2	ReLU	0.88
9	2	Softplus	0.95
10	2	Identity	0.98
11	2	Softplus	0.91
12	2	ReLU	0.89

ANN3 and yields the lowest RMSE based on the validation dataset. Adding a second hidden layer does not enhance the performance of the ANN. Hence a single layer FF ANN model is satisfactory to predict crash injury severity with acceptable accuracy along with the benefit of reduced computation time as compared to deep ANN models with two or more hidden layers. ANN4 also performs well but uses a different activation function and returns a real number output which is referred to as the ISI. The ISI is converted into an integer representing one of the four categories cited in Table 4.1 by rounding to the nearest integer. The numerical performance of ANN3 and ANN4 is evaluated in Chapter 5.

The finalized architecture of the proposed ANN model is shown in Figure 4.9. The ANN model consists of an input layer with seventeen neurons representing different explanatory variables, one optimized hidden layer with twelve neurons and an output layer with one neuron representing predicted crash injury severity.

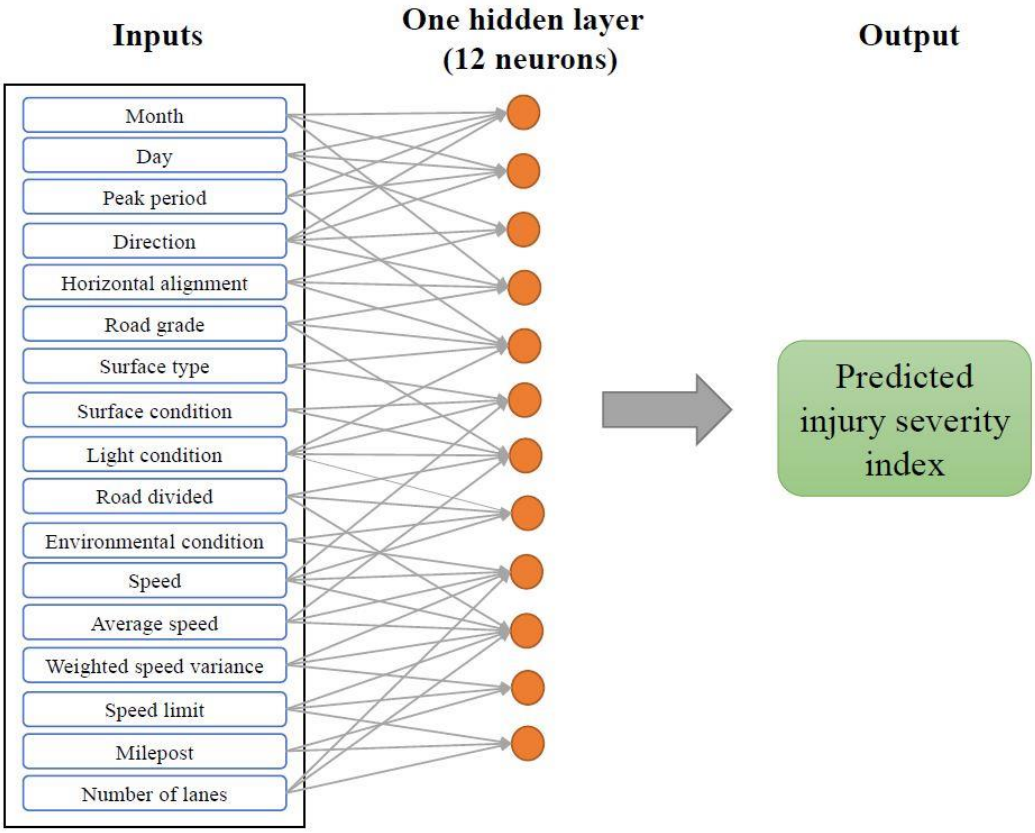


Figure 4.9 Final configuration of proposed ANN.

4.3 Key Takeaways from Model Development

Parametric and nonparametric models exhibit strengths and limitations in different capacities. This section summarizes the key takeaways from the development process of the LRM and ANN.

Explanatory Variables

The LRM development requires more data sorting and statistical analysis. As discussed in Section 4.1, careful consideration of the correlation between explanatory variables and the dependent variable is required in regression analysis to avoid skewed predictions of the dependent variable. Chi-square tests and multi collinearity tests were conducted to eliminate

insignificant explanatory variables from the analysis. The ANN requires less work and can readily accept all explanatory variables as input. That is because every input has an associated weight in the ANN computations. The weights are initialized randomly and updated during the model training process. The ANN then assigns a higher weight to the more important input as compared to the ones considered less important.

Structure Determination

The ANN development requires a trial and error process to determine a good combination of its constituents: number of hidden layers, number of neurons in the hidden layers, activation function and training algorithm. An optimal neural network structure is derived from a series of tests. The LRM on the other hand does not require experimentations, it is always under the form of a linear function and there are no additional parameters that require trials and optimization.

Assumptions

The first assumption of linear regression is that there is a linear relationship between the explanatory variable and the dependent variable. When the assumption of the linearity is not satisfied, a large number of stochastic data entries can lead to the development of an unsatisfactory model and a low R-squared value. To address this issue, the LRM development requires more meticulous data screening and filtering discussed in Section 4.1.2. The ANN is however dynamic and adaptive and can intelligently analyze input and provide reliable output without the need to remove data entries. In this research and for the purpose of consistency, the same data was used for the LRM and ANN development. However, the ANN does not assume any underlying patters within the data and uses an

exhaustive search to capture those patterns. The ANN has the ability to keep refining its outputs until it gets faster and more accurate at what it does. This dynamic learning process, also known as backpropagation, involves taking the network's output, comparing it with an ideal result, and feeding it back into the network from scratch.

Simplicity and Interpretation

As discussed in the previous paragraph, the ANN captures more complicated relationships and hidden patterns than the LRM. It might outperform the LRM since it uses a sophisticated architecture with designed activation functions. However, the output of the LRM is a linear relationship that is clear and well-defined while ANN makes it difficult to verify the why of the output. Limitations of ANNs include its "black box" nature, where the inputs and outputs are only known without detailed knowledge of its internal workings. Even if the ANN has a higher degree of accuracy, it is relatively easy to explain a linear model, its assumptions and why the output is what it is.

CHAPTER 5

MODEL EVALUATION

Two crash injury severity prediction models were developed in Chapter 4. The first section of this chapter presents a detailed analysis conducted to evaluate the overall model performance of the LRM and ANN for predicting crash injury severity. The second section discusses the potential applications of the proposed models to support traffic safety planning and operations on freeways.

5.1 Numerical Evaluation and Sensitivity Analysis

5.1.1 Numerical Evaluation

To evaluate the model performance, the Root-Mean-Square-Error (RMSE) is used to compute the variability between estimated values and observed values as shown in Equation (5.1). A RMSE value of 0 indicates a perfect fit to the data, therefore the smaller the RMSE, the better the model performance. The RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{X}_i - X_i)^2}{N}} \tag{5.1}$$

Where:

- $\widehat{X}_i$ is the observed injury severity level for crash i;

- X_i is the estimated injury severity level derived from the estimated injury severity index for crash i; and
- N is the total number of crashes.

As discussed in Chapter 4, 10% of the data is set aside for testing the model. The test dataset is used once a model is trained using the training and validation sets. The procedure to test the model performance is explained below.

Step 1: Classify the 296 freeway crashes selected for testing by injury severity level and further classify the crashes into peak and non-peak periods per each injury severity level. The corresponding data distribution of the selected crashes is illustrated in Table 5.1.

Table 5.1 Test Samples Classified by Injury Severity Level and Peak Period

Injury severity level	Number of crashes	Peak	Non-peak
PDO	226	132	94
Mild injury	49	29	20
Moderate injury	14	9	5
Severe Injury	7	5	2

Step 2: Run each crash with the LRM and ANN3 models, respectively. Then compute the RMSE based on the predicted injury severity versus the actual injury severity. The RMSE is the square root of the variance of the residuals. It indicates how close the observed data points are to the predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction. The results are summarized in Table 5.2.

Table 5.2 RMSE of Test Samples by Injury Severity Level and Peak Period

	RMSE					
	ANN			**LRM**		
Injury Severity Level	Peak	Non-peak	Overall	Peak	Non-peak	Overall
PDO	0.89	0.73	0.82	1.19	0.89	1.07
Mild injury	1.04	0.92	0.99	1.35	1.33	1.34
Moderate injury	0.82	0.89	0.85	1.19	1.14	1.17
Severe Injury	0.45	0	0.32	2.22	2.01	2.16
Overall	0.90	0.76	0.84	1.25	0.99	1.14

It is found that the ANN model had a lower overall RSME (0.84) and therefore outperformed the LRM for all injury severity levels. The lowest RMSE is recorded by the ANN for severe injury crashes and the highest RMSE was recorded by the LRM also for severe injury crashes. A low value of RMSE means that the difference between the actual value of injury severity and the predicted value of injury severity is small. A high value translates into a larger error caused by the model. The lowest possible RMSE is 0. This value is provided by the ANN for severe injury crashes under non-peak conditions. This means that the ANN correctly predicted all crashes in that category and returned no error. Based on the results displayed in Table 5.2, both the ANN and LRM offer a smaller RMSE with PDO crashes that represent most crashes (226 crash entries). The RMSE of the ANN model ranges between a minimum of 0.32 and a maximum of 0.99. The RMSE of the LRM is generally higher, ranging between 1.07 and 2.16.

An additional analysis is conducted where crashes are further subdivided by peak/non-peak period to evaluate the model performance under different traffic conditions and speed patterns. Furthermore, grouping the crashes into peak and non-peak gives more detailed insight on the performance of the models.

The best performance and lowest RMSE is recorded by the ANN during non-peak conditions for severe injury crashes. The highest RMSE is recorded by the LRM during peak conditions also for severe injury crashes. Both the LRM and the ANN generally perform better in non-peak periods when traffic is flowing smoothly, and speeds follow a standard pattern. The model performance somewhat decreases in peak periods when traffic patterns become more complex and less predictable. The ANN still generally outperforms the LRM. RMSEs calculated from the ANN model also have a smaller variation between peak and non-peak periods, as opposed to the RMSEs calculated from the LRM that have a large variation for severe injury crashes.

Step 3: Crashes are divided by environmental condition to evaluate the model performance under different weather circumstances. The same testing sample is utilized and the data distribution of the selected crashes is shown in Table 5.3. The RMSE distribution of the ANN and the LRM model are shown in Table 5.4.

Table 5.3 Test Samples Classified by Injury Severity Level and Weather Condition

Injury Severity Level	**Number of crashes**	**Clear conditions**	**Other**
PDO	226	184	42
Mild injury	49	32	17
Moderate injury	14	10	4
Severe Injury	7	3	4

Table 5.4 RMSE of Test Samples by Injury Severity Level and Weather Condition

	RMSE					
	ANN			**LRM**		
Injury Severity Level	Clear	Other	Overall	Clear	Other	Overall
PDO	0.83	0.79	0.82	1.18	1.04	1.15
Mild injury	1.12	0.91	1.05	1.44	1.53	1.47
Moderate injury	0.75	0.75	0.75	1.26	1.41	1.30
Severe Injury	0	0.5	0.29	2.17	2.39	2.30
Overall	0.86	0.80	0.84	1.25	1.16	1.24

The best performance and lowest RMSE is recorded by the ANN during clear weather conditions for severe injury crashes. Under these conditions, the RMSE is 0 which means that the ANN correctly predicted the outcome of all tested crashes. The highest RMSE is recorded by the LRM also for severe injury crashes during non-clear weather conditions. Generally, the model performance decreases during inclement weather periods when traffic patterns become more complex and less predictable. As for the ANN, it consistently outperforms the LRM by providing lower RMSEs.

However, the RMSEs calculated in Tables 5.2 and 5.4 are still relatively high and the overall predictive power of the models is found to be comparatively poor and must be further improved to be used for a meaningful operational crash risk prediction. An overall RMSE value of 0.84 is close to 1 and means that the model is missing the prediction by an average of 1. For example, if the actual crash injury severity is mild injury (ISL= 1), the model could have predicted either a PDO or a moderate injury crash (ISL = 0 or ISL= 2). The best performance is witnessed for severe injury crashes when the RMSE hits 0 under certain conditions. Even when the RMSE for severe crashes is at its highest of 0.5 in Table 5.4, it means the model is missing the prediction by 0.5. This result is considered acceptable as this range still translates into a severe injury crash and the model can alert decision makers to potentially unsafe conditions.

In addition to the RMSE, the quality of the ANN predictions is evaluated based on its precision. The precision calculation requires the true positive (TP) and the false positive (FP) numbers for each injury severity level category.

- TP: True positive value is defined as the number of crash cases under a specific injury severity category whose outcome is correctly predicted.

- FP: False positive value is defined as the number of crash cases whose outcome is falsely predicted as another injury severity level.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.2}$$

The results are grouped by injury severity level and displayed in Table 5.5:

Table 5.5 ANN Precision by Injury Severity Level

Injury Severity Level	Number of crashes	TP	FP	Precision
PDO	226	141	85	0.623
Mild injury	49	34	15	0.694
Moderate injury	14	10	4	0.714
Severe Injury	7	6	1	0.857

The precision of a model represents the percentage of the results that were correctly estimated. The results show that the precision is highest for severe injury crashes where the model accurately predicted 86% of the testing sample. The precision decreases as the injury severity level decreases. It is lowest for PDO crashes where the model accurately predicted 62% of the data. A good precision depends on the model objective and data type, but the ANN precision for severe injuries is considered good as most crashes are accurately predicted. The model performs more poorly for less severe injury crashes, and this is due to the stochastic nature of crash risk factors and data randomness and inaccuracies.

5.1.2 Sensitivity Analysis

The ANN captures more complicated relationships and hidden patterns than the LRM, and therefore outperforms the LRM as the results showed. However, it is relatively easy to explain the assumptions and output of the LRM determine countermeasures accordingly. A limitation of the ANN is its "black box" nature, where the output is known without detailed

knowledge of its internal workings. A sensitivity analysis is conducted in this section to understand how explanatory variables affect the output of the ANN. The model used in this analysis is ANN4 as it returns a continuous output, and it is easier to discern slight ISI changes due to input perturbation. The input perturbation technique consists of changing an input of the ANN and measuring the corresponding change in the output. The perturbation is applied on one input at a time while others are fixed. The analysis is performed on four segments. These lane configurations and posted speed limits are typical on I-80.

- Segment 1: milepost 1.45-3.59, speed limit = 55mph, 2 lanes per direction;
- Segment 2: milepost 3.59-5.10, speed limit = 55mph, 3 lanes per direction;
- Segment 3: milepost 8.10-21.51, speed limit = 65mph, 3 lanes per direction;
- Segment 4: milepost 34.02-42.46, speed limit = 65mph, 4 lanes per direction;

The ISI is computed for each entry before and after perturbation, and the percent change in ISI is computed as shown in Equation (5.2).

$$\%\ \text{Change in ISI} = \frac{Y'_i - Y_i}{Y_i} * 100 \tag{5.2}$$

- Y_i is the estimated injury severity index of crash i using the initial explanatory variables without perturbation;
- Y'_i is the estimated injury severity level of crash i using the altered explanatory variables with perturbation;

Step 1: The sensitivity between speed and ISI is conducted on segments 1-4. All explanatory variables but the speed are fixed in the ANN. The initial speed is set to 50 mph. The speed is then increased or decreased in increments of 10% up to 50%. The new ISI is calculated and the % change in ISI is computed. The results are displayed in Table 5.5.

Table 5.6 Percent Change in ISI after Speed Perturbation, Segments 1-4

		S1		S2		S3		S4	
% Change in speed	Speed (mph)	ISI	% Change in ISI	ISI	% Change in ISI	ISI	% Change in ISI	ISI	% Change in ISI
-50	25	1.03	-20.16	1.02	-16.28	1.32	-9.59	1.25	-8.76
-40	30	1.04	-19.38	1.02	-16.28	1.32	-9.59	1.26	-8.03
-30	35	1.04	-19.38	1.03	-15.50	1.33	-8.90	1.26	-8.03
-20	40	1.09	-15.50	1.08	-11.63	1.39	-4.79	1.31	-4.38
-10	45	1.16	-10.08	1.15	-6.20	1.39	-4.79	1.32	-3.65
0	50	1.29	0.00	1.23	0.00	1.46	0.00	1.37	0.00
10	55	1.29	0.00	1.24	0.78	1.51	3.42	1.42	3.65
20	60	1.31	1.55	1.25	1.55	1.53	4.79	1.43	4.38
30	65	1.45	12.40	1.27	3.10	1.57	7.53	1.45	5.84
40	70	1.48	14.73	1.29	4.65	1.63	11.64	1.51	10.22
50	75	1.51	17.05	1.34	8.53	1.67	14.38	1.53	11.68

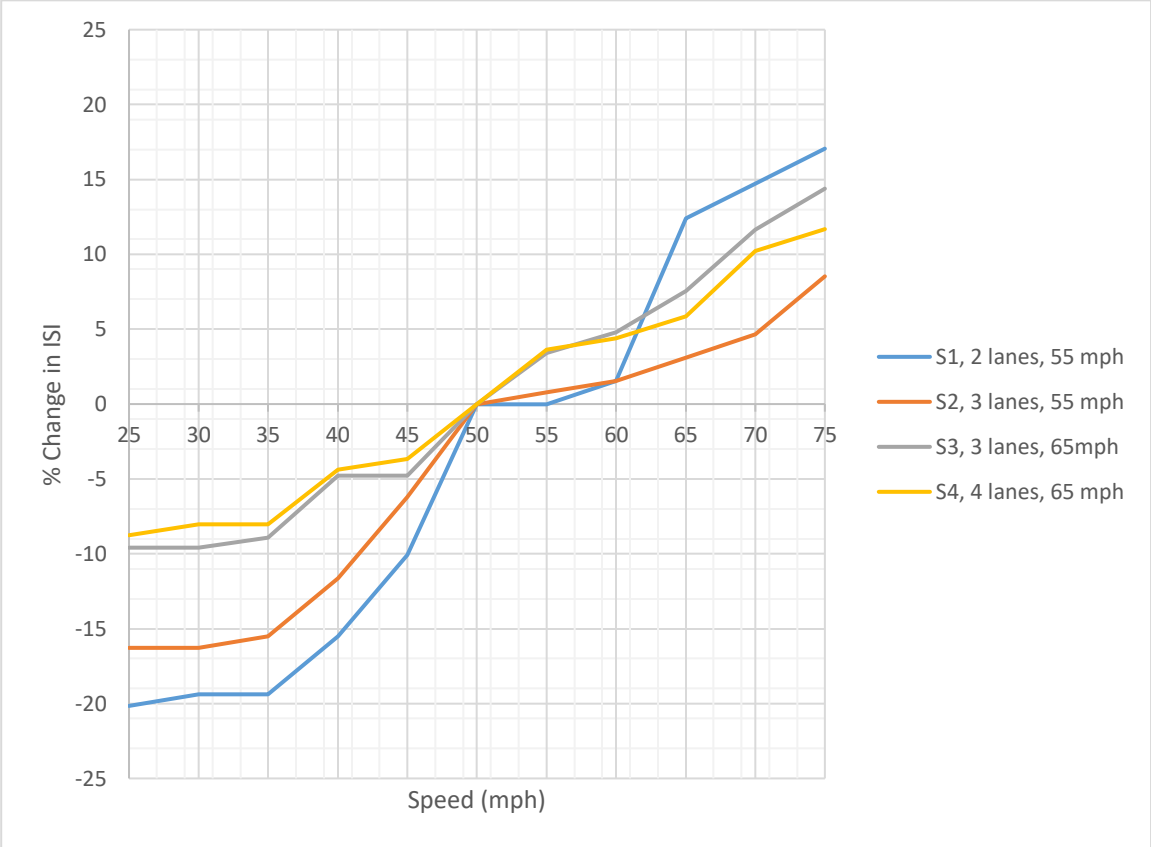
When a speed increase/decrease in increments of 10% is added to the initial speed and a new output is calculated, the general trend on all four segments shows that higher speeds increase the ISI and lower speeds decrease the ISI. This result is expected as discussed in Chapter 2 and confirms the findings of previous studies.

The ISI increases at a faster rate when speeds exceed 60 mph on two-lane highways with a 55 mph speed limit (Segment 1). The ISI also decreases substantially when the speed is decreased from 50 mph to 35 mph, and it reaches a plateau when the speed further decreases beyond 35 mph. On segment 2, the ISI increases slowly with an increase in speed until the speed exceeds 65 mph and the ISI starts increasing faster. The decreasing pattern is similar to that on segment 1, where the ISI decreases briskly until the speed reaches 35 mph and then less intensely for lower speeds.

Even though the ISI is directly proportional to the speed on all segments, the change in ISI on three-lane segments with a speed limit of 55 mph is less extreme than it is on two-lane segments with the same speed limit. A 50% speed increase leading to a 75 mph speed on a three-lane segment increases the ISI by 8.53% compared to 17.05% on a two-lane segment. The injury severity level remains 1.0 (mild injury) even with a large increase in speed on three-lane segments. If the increasing pattern holds on two-lane segments, a minimum of 50% increase in speeds is required to raise the injury severity level to moderate injury (2.0).

On Segments 3 and 4 with a higher speed limit (65 mph), the ISI follows the same pattern with respect to the speed. The ISI starts increasing faster when speeds exceed 65 mph on three-lane and four-lane highways. Similarly to segments 1 and 2, the ISI reaches a plateau when the speed further decreases beyond 35 mph. The ISI is more sensitive to a

speed increase than it is to a speed decrease on segments with 65 mph speed limit. A 75 mph speed (50% increase from 50 mph) on three-lane and four-lane segments rises the ISI by 14.37% and 11.68% respectively, whereas 50% speed reduction decreases the ISI by 9.59% and 8.76% respectively. The opposite is true for segments 1 and 2 with a lower speed limit of 55 mph where the ISI is more sensitive to a speed decrease than it is to a speed increase: A 50% speed decrease on two-lane and three-lane segments lowers the ISI by 20.16% and 16.28% respectively, whereas 50% speed increase raises the ISI by 17.05% and



8.53% respectively. Figure 5.1 shows the change in ISI with respect to the change in speed on all segments.

Figure 5.1 Change in ISI vs. speed on segments 1-4.

Step 2: The sensitivity between weighted speed variance (WSV) and ISI is conducted on segments 1-4. All explanatory variables but the WSV are fixed in the ANN. The initial value is set to 5 m/h^2. It is increased in increments of 5 m/h^2 up to 25 m/h^2. The new ISI and the % change in ISI are calculated. The results are displayed in Table 5.6.

Table 5.7 Percent Change in ISI after WSV Perturbation, Segments 1-4

	S1		S2		S3		S4	
WSV (m/h^2)	ISI	% Change in ISI	ISI	% Change in ISI	ISI	% Change in ISI	ISI	% Change in ISI
5	1.24	0.00	1.16	0.00	1.35	0.00	1.31	0.00
10	1.32	6.45	1.22	5.17	1.38	2.22	1.33	1.53
15	1.36	9.68	1.26	8.62	1.40	3.70	1.35	3.05
20	1.41	13.71	1.28	10.34	1.43	5.93	1.37	4.58
25	1.45	16.94	1.31	12.93	1.47	8.89	1.41	7.63

When an increase in increments of 5 m/h^2 is added to the initial WSV and a new ISI output is calculated, the general trend on all four segments shows that a higher WSV increases the ISI. On segments 1 and 2, the ISI follows a comparable trend: the ISI increases at a steady rate with the increase in WSV. However, a bigger increase in ISI is observed on segment 1 compared to segment 2. A WSV of 25m/h^2 brings an increase to the ISI of 16.94% on a two-lane highway whereas it brings an increase of 12.93% on a three-lane highway with the same speed limit. On segments 3 and 4 (where the speed limit is 65mph), the increase rate of the ISI is slower and more gradual as it reaches a lower maximum of 8.89% and 7.63% increase respectively when the WSV is highest compared to segments 1 and 2 with a speed limit of 55mph.

The increase in ISI due to a WSV increase is however not as extreme as that due to a speed increase. The ISI increases by a maximum of 16.94% when the WSV is multiplied by 5. A close % increase (17.04%) is reached by increasing the speed by 50% on the same segment. Nevertheless, the ISI is directly proportional to the WSV and an increase in WSV leads to an increase in ISI on all four segments. Figure 5.2 shows the change in ISI with respect to the change in WSV on all segments.

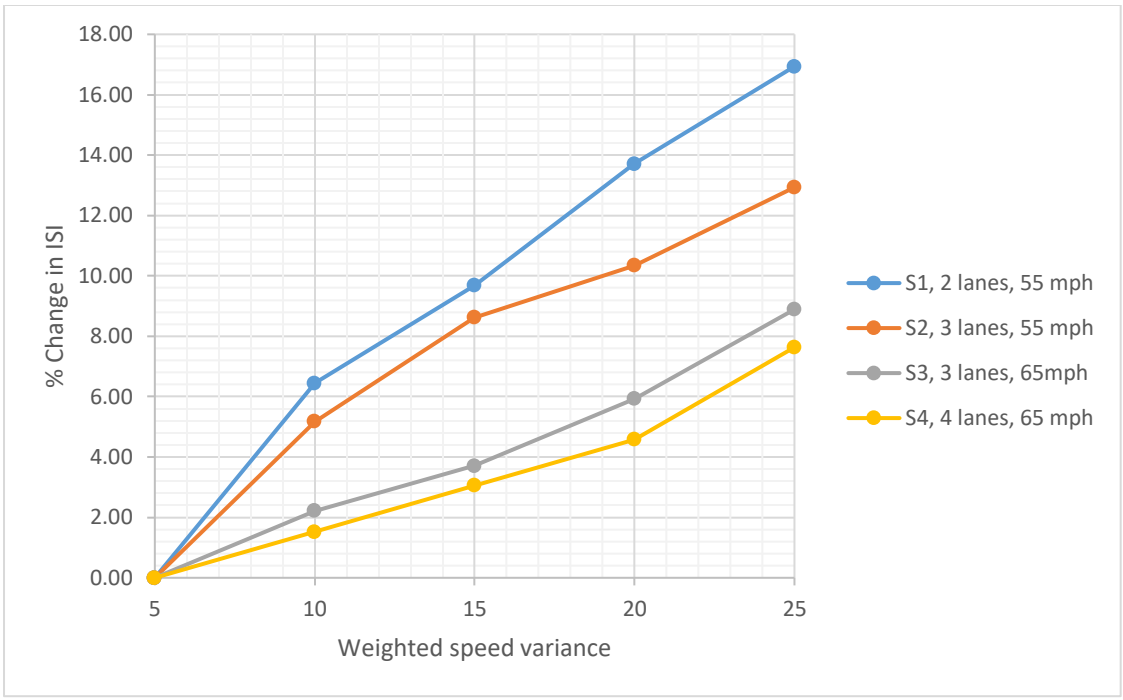


Figure 5.2 Change in ISI vs. weighted speed variance on segments 1-4

5.2 Model Applications

A comprehensive crash injury severity prediction ANN using big data was developed and evaluated in this research. The model offers a variety of practices and can be used to increase safety and optimize budget allocation to mobility by supporting traffic agencies in the planning and operation stages. In this section, potential applications for the proposed model are presented and safety countermeasures are discussed.

5.2.1 Network Screening

The AASHTO Highway Safety Manual (HSM) presents a variety of methods for quantitatively estimating crash frequency or severity at a variety of locations to develop a safer, more efficient roadway transportation system. HSM presents a 6-step Roadway

Safety Management Process (RSMP) shown in Figure 5.3. RSMP is a repetitive process to managing road safety.

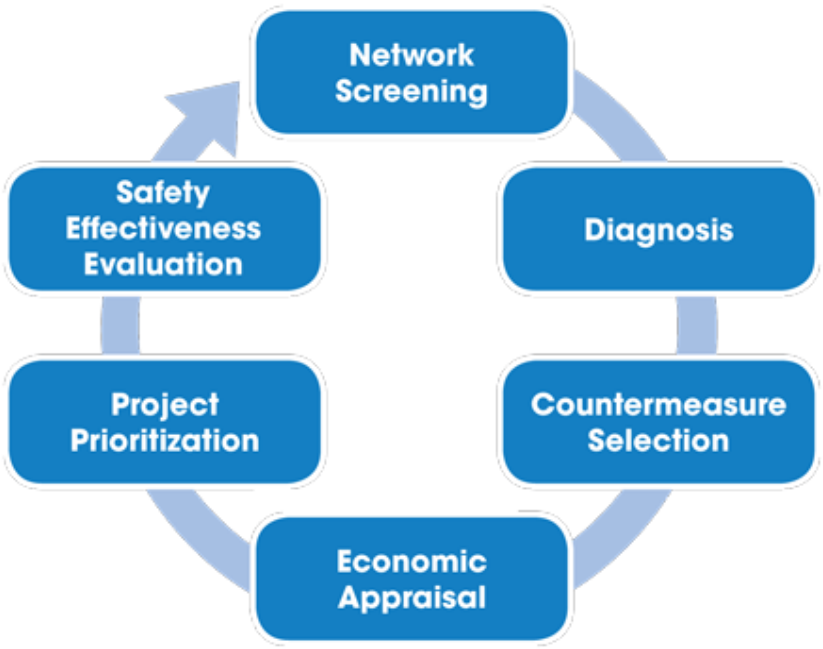


Figure 5.3 HSM 6-Step roadway safety management process.
Source: https://safety.fhwa.dot.gov/tools/crf/resources/cmfs/management.cfm, Retrieved on May 22, 2018.

- Step 1: Network screening is the process of prioritizing sites within a transportation network. Sites are ranked in this step, and the sites that are deemed to be the most dangerous are prioritized.

- Step 2: Diagnosis is the process of investigating a site by statistical and analytical testing to determine the present contributing crash factors. The collected evidence helps prioritize safety countermeasures in the next steps.

- Step 3: Countermeasure selection is the recommendation of means expected to mitigate the crash contributing factors. They include road design changes, public awareness, law enforcement, or EMS policies.

- Step 4: Economic appraisal is a comprehensive economic assessment conducted under the form of a cost-benefit analysis for the proposed countermeasures. This maximizes the potential safety benefit per dollar.

- Step 5: Project prioritization is the creation of a plan that implements the safety countermeasures that provide the maximum road safety benefit within the total allocated budget.

- Step 6: Safety effectiveness evaluation is the assessment of how well a countermeasure performs once it is applied. This step leads back into step 1 with updated site information for the next round of network screening.

Network screening is a very important step because financial resources are limited and transportation networks are large; it is practically impossible to modify an entire network simultaneously. The objective of network screening is to identify emphasis areas for directing investments in location specific road safety improvements. Based on existing conditions and historical data, transportation agencies can recommend methods to support their safety objective in several ways.

Based on the ANN structure shown in Figure 4.9, some of the explanatory variables used as model input vary in real-time such as speed and environmental condition while others represent existing conditions: horizontal alignment, road grade, surface type, road divider, speed limit and one-way number of lanes. These variables can be assessed during network screening.

Speeding is a factor in almost one-third of all fatal crashes, according to the NHTSA. Although agreement is almost general on the relationship between speed and crash severity, that relationship is complicated and changes on different road segments as discussed in Section 5.1.2. Posting speed limits is the most widely used method for managing speed. The MUTCD recommends that speed limits are set within 5 mph of the 85^{th} percentile speed of free-flowing traffic. The MUTCD also lists other risk factors such as road geometry, roadside development, parking practices, pedestrian activity, and crash experience.

There are predominantly two speed limit zones on I-80: 55 mph and 65 mph, in addition to a 1.45-mile, 50 mph segment near the Delaware water gap area as I-80 enters New Jersey from Pennsylvania. During the network screening process, it is recommended

that the segments are screened one at a time to determine if a change in speed limit is required. A traffic study using existing conditions and crash data should be performed to warrant a speed increase/decrease.

Increasing the speed limit reduces travel time and improves economic productivity but it also increases potential injury and/or fatality risks. There are several methods for determining an appropriate speed limit on a roadway section such as the engineering approach, the expert system approach, the optimization approach and the injury minimization or safe system approach**.** In this case, the acceptance of the human body to injury during a crash is the primary factor in identifying appropriate speeds. Speed management involves balancing safety and efficiency in travel and that is why a network screening using ANN cash injury severity simulations on all segments is required to determine appropriate speed limits for combined safety and efficiency.

When developing tools for safety data analysis, it is important to take into consideration the operational needs and abilities of the local authority. New Jersey has adopted the national vision for highway safety – Toward Zero Deaths. This calls for a national goal of reducing the number of traffic fatalities by half by the year 2030. New Jersey's short-term crash reduction goal is to reduce serious injuries and fatalities by 2.5 percent annually with the support of all safety partners, thus every countermeasure contributing to that goal is worth the invested time and resources.

5.2.2 Real-time Traffic Management

In addition to network screening and the assessment of existing conditions discussed in the previous section, the proposed ANN is a helpful tool that can be used to assist the New Jersey Strategic Highway Safety Plan (SHSP) by providing the ability to predict crash injury

severity under real-time conditions and shedding light on areas witnessing high fatality/severe injury rates during operations. With real-time monitoring of traffic and the existence of a comprehensive database reflecting existing conditions on the freeway, crash injury severity predictions can be made dynamically at specific locations under different time of day given traffic volume, speeds, weather conditions, etc. A heat map is generated to visually translate the numbers and point to potential sites considered unsafe. Figure 54 is an illustration of this method. It shows the predicted injury severity level at different segments and during different times of day given real-time conditions.

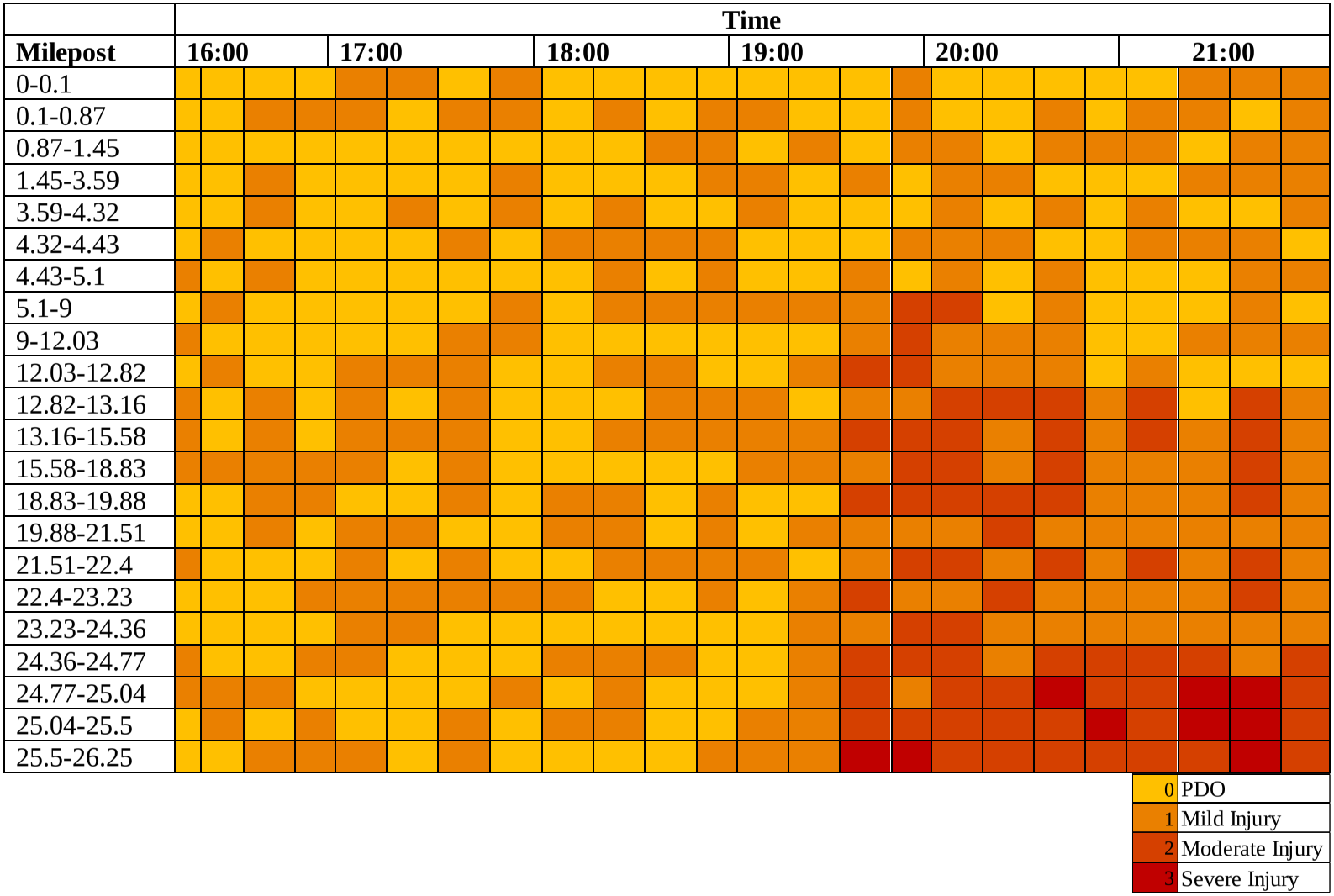


Figure 5.4 Heat map of predicted injury severity level on different segments of I-80 by time of day.

Real-time countermeasures on specific locations shown on the dynamic heat map can be recommended accordingly. Based on the sensitivity analysis results discussed in Section 5.1.2, speed highly affects the consequences of a crash and examples of countermeasures to reduce speed are listed below:

- Dynamic warning signs: the installation of dynamic warning signs reminds drivers of their travel speed and creates a sense of being monitored. These signs are shown to reduce vehicle speeds by up to 5 mph and are effective when used at speed transitions that occur as a driver enters an urban area.

- Enforcement to improve speed compliance: Enforcement can be a preventive method applied before speeding. The physical presence of police serves as a psychological traffic calming method. It is also a corrective method if used after speeding and punishing the violating driver by police officers or other authorized officials.

- Adaptive ramp metering: this strategy consists of deploying traffic signals on ramps to dynamically control the rate vehicles enter a freeway facility. This soothes the flow of traffic onto the mainline, allowing efficient use of existing freeway capacity.

- Dynamic speed limits: This strategy adjusts speed limits based on real-time traffic, roadway, and weather conditions. Dynamic speed limits can be applied to an entire roadway segment or individual lanes. Real-time and anticipated traffic conditions are used to adjust the speed limits dynamically to meet the safety and mobility objectives.

While the ANN model is developed based on I-80 data, the same methodology can be applied on other freeways in New Jersey such as the New Jersey Turnpike (NJTP) and the Garden State Parkway (GSP) as well as Interstates 278, 676, 76, 280, 195, 295, 78 and 287. Real-time ANN computations can be extended to a network scale. With real-time monitoring of traffic along NJ freeway segments, crash injury severity predictions can be made in dynamically covering the entirety of the network. Figure 5.5 is an illustration of this method. It shows an example of potential crash injury severity level based on real-time computations by the ANN along New Jersey's Interstate highway network.

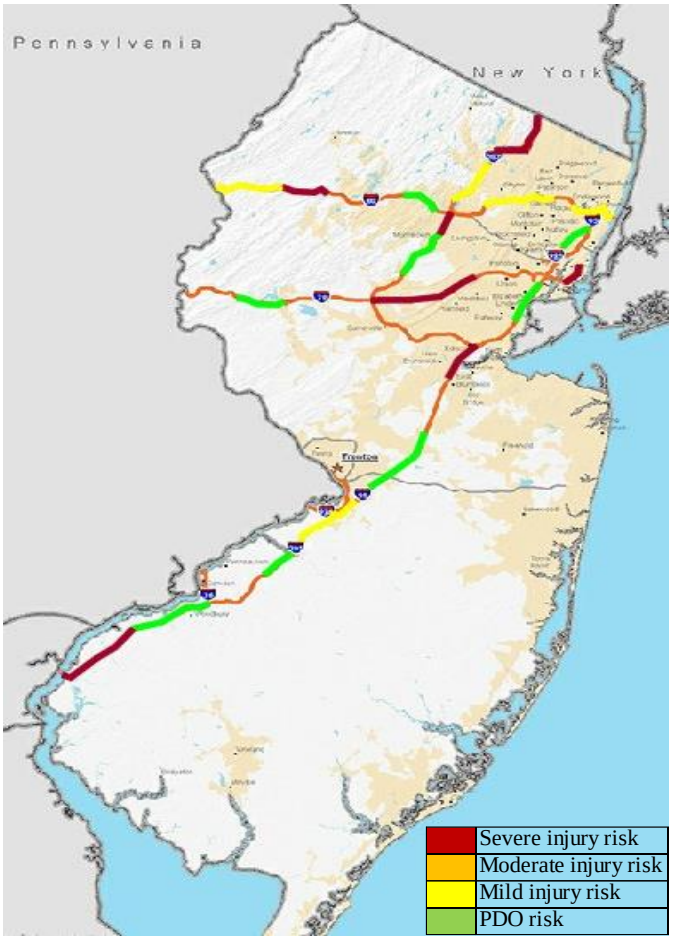


Figure 5.5 Color-coded freeway segments based on real-time injury severity risk.

CHAPTER 6

CONCLUSIONS AND FUTURE RESEARCH

With the increasing need for roadway mobility, motor vehicle crashes continue to be one of the United States' most serious social, economic and health issues. The capability to precisely predict the crash factors and impacts on injury severity is essential to minimize both the cost and traffic congestion induced by crashes. In response to this challenge, an ANN model for predicting the injury severity level resulting from a crash was developed using big data sources in this research. The injury severity level index was predicted using explanatory variables proven to affect the crash outcome based on freeway data from 2017. The ANN performance was analyzed and compared to the performance of a LRM developed with the same set of data.

6.1 Research Contributions

The main contributions of this dissertation are listed below:

1. A modeling framework was developed for decision makers to predict the anticipated crash injury severity in real-time using the available data required for model input. Despite the relatively poor performance of the developed models, the numerical analysis data suggest that driver decisions such as the traveling speed heavily influence the crash injury severity.

2. The weighted speed variance was incorporated in the model as an explanatory variable, which proved to be critical to crash injury severity based on the results from the sensitivity analysis. The crash injury severity increases when the speed variance increases across the traffic stream.

3. The predictive performance of a statistical model and a machine learning model was compared. The results show that ANN (machine learning) can explore the linear and non-linear relationship between a large set of contributing factors. The explored statistical model performs poorly and is not good enough in making reliable predictions.

4. The effect of traveling speed and weighted speed variance on the crash injury severity were demonstrated using sensitivity analysis. The risk of a severe injury increases when either one of these measures increases.

5. The presented model can be readily applied in traffic management to identify roadway segments with relatively high crash injury severity risk based on the available data provided as model inputs. The presented model can be implemented to visualize crash risks on roadway segments.

6. The modeling framework can be applied on other freeways in New Jersey and other states. Real-time ANN computations can also be extended to a network scale. With real-time monitoring of traffic along roadway segments, crash injury severity predictions can be made in dynamically covering the entirety of the network. Accordingly, traffic agencies can implement traffic calming strategies.

7. The model offers a variety of practices and can be used to increase safety and optimize budget allocation to mobility by supporting state and local traffic agencies in the construction, operations, and planning phases. The ANN can be used to:

    - Identify sites with potential for improvement: quantify crash injury severity impacts of alternative highway geometry proposed plans; provide general methods useful for identifying sites with potential for improvement, diagnosis, and countermeasures.

    - Quantify real-time crash injury severity predictions resulting from changes in the traffic stream on freeways. With this information, transportation agencies can monitor traffic operation in real-time and be alert of potential high injury severity risk in specific locations.

    - Modify existing conditions to maintain safe operations: based on the historical data and the generated model results, crash patterns at existing locations can be identified and analyzed to determine potential risk factors that increase the likelihood of high injury severity.

    - Identify needs and program projects: identify sites most likely to benefit from safety improvement, classify targeted crash patterns for the network, and prioritize expenditures for efficiency. Isolate unsafe zones based on existing conditions and propose improvements for crash injury severity reduction.

6.2 Research Limitations

While developing the crash injury severity prediction model, a wealth of insights, challenges, areas of potential improvements, and opportunities available to agencies in the areas of safety assessment, data collection, and performance measurement were identified, all of which are summarized below.

- As each model presents advantages and limitations over the other model, it is difficult to say that one model is better than the other under all circumstances. The ANN captures more complicated relationships and hidden patterns than the LRM, especially when nonlinear relationships are involved. On the other hand, the ANN typically requires a large dataset for its optimization. This study reaffirmed that roadway crashes are stochastic and complex and while a good prediction model is a great tool to monitor traffic conditions and increase safety, prediction models react differently to varied sets of data and one model might not be optimal under diverse conditions.

- With technological advancement, the transportation industry has been experiencing unprecedented massive traffic data obtained from different sources, such as infrastructure sensors, mobile devices, and floating cars. It is important to appropriately manage and interpret this new and rich form of data (big data) in efficient ways. The use of conventional data management tools cannot effectively uncover hidden correlations and intricate patterns and other insights, which would leave a large amount of traffic data underutilized. For the crash injury severity analysis, leveraging big data analytics and advanced prediction methods (e.g., ANN models), the accuracy of predicted crash injury severity can be significantly improved, in comparison to predictions using traditional deterministic approaches with data captured by loop detectors. The ability of big data analytics to work faster and adapt gives transportation agencies an unprecedented edge to enhance mobility, increase safety and reduce crash delays and costs.

- The crash data is retrieved from the NJDOT, and some inaccurate entries negatively influence the predictive power of the model. For example, the beginning time of a crash is manually entered into the database by the police officer who reported to the scene. If this time is inaccurate, the speed information used in the model will challenge the accuracy of the prediction. Increasing the precision of the beginning timestamp is vital to the prediction model. In addition, the traffic counts information at the scenes of crashes are important measures for predicting injury severity. However, this data is missing for most locations. The hourly traffic volumes recorded in NJCMS are used instead for model development.

6.3 Future Research

The advantage of using prediction models to estimate crash injury severity is that it becomes possible to quantify how changes to a specific characteristic of a facility or to the traffic conditions would affect the crash outcome. Creating a good statistical model requires access to a large amount of detailed data about existing facilities and accurate crash statistics. This data may be nonexistent or hard to obtain for some geographic areas and types of projects. In addition, one model is not always the best option for injury severity prediction as discussed in this research. Data analysis and prediction is a complicated process that requires additional research. Future research needed to enhance the prediction models developed in this study shall focus on enhancing the quality of the crash database.

Improving data accuracy is the most important step to develop a better prediction model. Several data entries used in this study can be optimized: more accurate speeds, actual traffic volumes and more punctual crash timestamps will substantially improve the reliability of the developed model and produce more accurate results. It is also worthwhile to develop a self-updating database by gathering data from various sources in an automated manner where and when feasible. Linked databases can be updated separately with more up-to-date survey and information. For example, any change to the roadway geometry or an ongoing work zone can be automated and communicated between databases to reduce the time required for manual processing and improve productivity. As discussed in this study, ANNs are dynamic models that can absorb new data and learn and adapt accordingly. With the ever-changing dynamics of traffic patterns, it is vital to have up-to-date data sources for accurate predictions and results.

Moreover, additional performance measures could be provided to further assess the predictive capability of the ANN. These measures include but are not limited to accuracy, precision, AUC and weighted F1 score. The RMSE was used in this research to compare both models for consistency purposes but including more performance measures can give a deeper understanding of the model reliability.

In addition, the proposed model can be further extended to include the driver characteristics such as age, gender, driving experience, etc. which were not included in this study due to the lack of data availability in real-time. In 2017, major revisions were made to the New Jersey crash report manual (NJTR-1) involving the addition of new values to fields of the crash database. New values were added to distraction due to mobile devices field, and the safety restraints field was expanded with two new categories for child restraints. There were also changes in the environmental factors, roadway conditions and driving under influence (DUI) fields. In future research and as technology advances, more data can be gathered including more parameters such as cellphone usage, drug involvement, and driver related characteristics and behavior.

REFERENCES

Abdel-Aty, M. A., & Radwan, A. (2000). Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention, 32(5), 633–642. doi:10.1016/s0001-4575(99)00094-9.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transportation Research Record: Journal of the Transportation Research Board, 1897(1), 88–95. doi:10.3141/1897-12.

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: Journal of the Transportation Research Board, 1746(1), 6–13. doi:10.3141/1746-02.

Abisaad, R., & Chien, S. (2018). Investigating the effect of speed variance, weather conditions, and time of day on crash occurrence and severity". Transportation Research Board 97th Annual Meeting, Washington, DC, January 2018.

Ahmed, M. M., & Abdel-Aty, M. A. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Transactions on Intelligent Transportation Systems, 13(2), 459–468. doi:10.1109/tits.2011.2171052.

Akin, D. & Akbaç, B. (2010). A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics. Sci Res Essays 5(19):2837–2847.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention, 34(6), 729–741. doi:10.1016/s0001-4575(01)00073-2.

Behnood, A., & Mannering, F. L. (2017). The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. Traffic Injury Prevention, 18(5), 456–462. doi:10.1080/15389588.2016.1262540.

Bham, G. H., Javvadi, B. S., & Manepalli, U. R. R. (2012). Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban U.S. highways in Arkansas. Journal of Transportation Engineering, 138(6), 786–797. doi:10.1061/(asce)te.1943-5436.0000370.

Caliendo, C., Guida, M., & Parisi, A. (2007). A crash prediction model for multilane roads. Accident Analysis and Prevention, 39(4), 657–670. doi:10.1016/j.aap.2006.10.012.

Carlson, W.L. (1979). Crash injury prediction model, Accident Analysis and Prevention, 11, pp. 137-153.

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Science, 43(8), 541–557. doi:10.1016/j.ssci.2005.04.004.

Chang, L.-Y., & Wang, H.-W. (2006). Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention, 38(5), 1019–1027. doi:10.1016/j.aap.2006.04.009.

Chen, F., Chen, S., & Ma, X. (2016). Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models. International Journal of Environmental Research and Public Health, 13(6), 609. doi:10.3390/ijerph13060609.

Chen, W.-H., & Jovanis, P. P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record: Journal of the Transportation Research Board, 1717(1), 1–9. doi:10.3141/1717-01.

Chen, Y., Wang, K., King, M., He, J., Ding, J., Shi, Q., & Li, P. (2016). Differences in factors affecting various crash types with high numbers of fatalities and injuries in China. Plos One, 11(7). doi:10.1371/journal.pone.0158559.

Chiou, Y.-C., Sheng, Y.-C., & Fu, C. (2017). Freeway crash frequency modeling under time-of-day distribution. Transportation Research Procedia, 25, 664–676. doi:10.1016/j.trpro.2017.05.450.

Çodur, M. Y., & Tortum, A. (2015). An artificial neural network model for highway accident prediction: a case study of Erzurum, Turkey. PROMET - Traffic&Transportation, 27(3), 217–225. doi:10.7307/ptt.v27i3.1551.

Cooper, P. J. (1997). The relationship between speeding behavior (as measured by violation convictions) and crash involvement. Journal of Safety Research, 28(2), 83–95. doi:10.1016/s0022-4375(96)00040-0.

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 and Prevention, 38(3), 434–444. doi:10.1016/j.aap.2005.06.024.

Digges, K. & Dalmotas, D. (2001). Injuries to restrained occupants in far-side crashes. Technical Report. SAE Technical Paper.

Effati, M., Rajabi, M. A., Hakimpour, F., & Shabani, S. (2015). Prediction of crash severity on two-lane, two-way roads based on fuzzy classification and regression tree using geospatial analysis. Journal of Computing in Civil Engineering, 29(6), 04014099. doi:10.1061/(asce)cp.1943-5487.0000432.

Farmer, C. M., Braver, E. R., & Mitter, E. L. (1997). Two-vehicle side impact crashes: the relationship of vehicle and crash characteristics to injury severity. Accident Analysis and Prevention, 29(3), 399–406. doi:10.1016/s0001-4575(97)00006-7.

Fountas, G., & Anastasopoulps, P. C. (2017). A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury severities. Analytic Methods in Accident Research, Volume 15, September 2017, Pages 1-16.

Garber, J. J., & Gadiraju, R. (1989). Factors affecting speed variance and its influence on accidents. Transportation Research Record No. 1213, 64-71.

Golob, T. F., & Recker, W. W. (2003). Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. Journal of Transportation Engineering, 129(4), 342–353. doi:10.1061/(asce)0733-947x(2003)129:4(342).

Golob, T. F., & Recker, W. W. (2004). A method for relating type of crash to traffic flow characteristics on urban freeways. Transportation Research Part A: Policy and Practice, 38(1), 53–80. doi:10.1016/j.tra.2003.08.002.

Greibe, P. (2003). Accident prediction models for urban roads. Accident Analysis and Prevention, 35(2), 273–285. doi:10.1016/s0001-4575(02)00005-2.

Gårder, P. (2006). Segment characteristics and severity of head-on crashes on two-lane rural highways in Maine. Accident Analysis and Prevention, 38(4), 652–661. doi:10.1016/j.aap.2005.12.009.

Hao, W., & Daniel, J. (2014). Motor vehicle driver injury severity study under various traffic control at highway-rail grade crossings in the United States. Journal Of Safety Research, 51, 41-48. doi: 10.1016/j.jsr.2014.08.002

Hao, W., & Daniel, J. (2015). Driver injury severity related to inclement weather at highway–rail grade crossings in the United States. Traffic Injury Prevention, 17(1), 31-38. doi: 10.1080/15389588.2015.1034274.

Hassan, H. M., & Abdel-Aty, M. A. (2013). Predicting reduced visibility related crashes on freeways using real-time traffic flow data. Journal of Safety Research, 45, 29–36. doi:10.1016/j.jsr.2012.12.004.

Harwood, D.W., Torbic, D.J., Richard, K.R., & Meyer, M.M. (2010). SafetyAnalyst: tools for safety management of specific highway sites. Turner-Fairbank Highway Center.

Hosseinpour, M., Yahaya, A. S., & Sadullah, A. F. (2014). Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: case studies from Malaysian federal roads. Accident Analysis and Prevention, 62, 209–222. doi:10.1016/j.aap.2013.10.001.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis and Prevention, 108, 27–36. doi:10.1016/j.aap.2017.08.008.

Ji, A., & Levinson, D. (2020). An energy loss-based vehicular injury severity model. Accident Analysis and Prevention, 146, 105730. doi:10.1016/j.aap.2020.105730.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. Journal of Applied Statistics, 30(1), 63–77. doi:10.1080/0266476022000018510.

Kloeden, C.N., McLean, A. J., & Glonek, G. (2002). Reanalysis of traveling speed and the risk of crash involvement in Adelaide South Australia. Road Accident Research Unit. The University of Adelaide. Report No. CR 207 (2002).

Kloeden, C.N., Ponte, G., & McLean, A. J. (2001). Traveling speed and the risk of crash involvement on rural roads. Road Accident Research Unit. Adelaide University. Report No. CR 204 (2001).

Knecht, C., Saito, M., & Schultz, G. G. (2016). Development of crash prediction models for curved segments of rural two-lane highways. International Conference on Transportation and Development 2016. doi:10.1061/9780784479926.072.

Kockelman, K., Bottom, J., Kweon, Y.J., Ma, J., & Wang, X. Safety impacts and other implications of raised speed limits on high-speed roads. National Cooperative Highway Research Program (NCHRP) Report 17-23, 2006.

Kockelman, K. K., & Ma, J. (2010). Freeway speeds and speed variations preceding crashes, within and across lanes. Journal of the Transportation Research Forum. doi:10.5399/osu/jtrf.46.1.976.

Kononen, D. W., Flannagan, C. A., & Wang, S. C. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Analysis and Prevention, 43(1), 112–122. doi:10.1016/j.aap.2010.07.018.

Kononov, J., Durso, C., Reeves, D., & Allery, B. K. (2012). Relationship between traffic density, speed, and safety and its implications for setting variable speed limits on freeways. Transportation Research Record: Journal of the Transportation Research Board, 2280(1), 1–9. doi:10.3141/2280-01.

Krammes, R.A., Rao, K.S., & Oh, H. (1995). Highway geometric design consistency software. Transportation Research Record (1500).

Kumar, U. A. (2005). Comparison of neural networks and regression analysis: a new Expert Systems with Applications, 29(2), 424–430.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice, 44(5), 291–305. doi:10.1016/j.tra.2010.02.001.

Lu, P., & Tolliver, D. (2016). Accident prediction model for public highway-rail grade crossings. Accident Analysis and Prevention, 90, 73–81. doi:10.1016/j.aap.2016.02.012.

Lui, K.-J., Mcgee, D., Rhodes, P., & Pollock, D. (1988). An application of a conditional logistic regression to study the effects of safety belts, principal impact points, and car weights on drivers fatalities. Journal of Safety Research, 19(4), 197–203. doi:10.1016/0022-4375(88)90024-2.

Ma, J. (2006). Bayesian multivariate Poisson-lognormal regression for crash prediction on rural two-lane highways. Ph.D. Dissertation, the University of Texas at Austin.

Ma, J., & Kockelman, K. (2006). Crash frequency and severity modeling using clustered data from Washington State. 2006 IEEE Intelligent Transportation Systems Conference. doi:10.1109/itsc.2006.1707456.

Ma, Z., Zhao, W., Chien, S. I.-J., & Dong, C. (2015). Exploring factors contributing to crash injury severity on rural two-lane highways. Journal of Safety Research, 55, 171–176. doi:10.1016/j.jsr.2015.09.003.

Ma, Z., Zhang, H., Chien, S. I.-J., Wang, J., & Dong, C. (2017). Predicting expressway crash frequency using a random effect negative binomial model: a case study in

China. Accident Analysis and Prevention, 98, 214–222. doi:10.1016/j.aap.2016.10.012.

Malyshkina, N.V., Mannering, F., & Labi, S. (2007). Influence of speed limits on roadway safety in Indiana. doi:10.5703/1288284313353.

Malyshkina, N.V., & Mannering, F. (2008). Effect of increases in speed limits on severities of injuries in accidents. Transportation Research Record: Journal of the Transportation Research Board 2008, doi:10.3141/1665-14.

Mao, Y., Zhang, J., Robbins, G., Clarke, K., Lam, M., & Pickett, W. (1997). Factors affecting the severity of motor vehicle traffic crashes involving young drivers in Ontario. Injury Prevention, 3(3), 183–189. doi:10.1136/ip.3.3.183.

Moghaddam, F.R., Afandizadeh, S., Ziyadi, M. (2010). Prediction of accident severity using artificial neural networks. International Journal of Civil Engineering 9:1.

Mooradian, J., Ivan, J. N., Ravishanker, N., & Hu, S. (2013). Analysis of driver and passenger crash injury severity using partial proportional odds models. Accident Analysis and Prevention, 58, 53–58. doi:10.1016/j.aap.2013.04.022.

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

NHTSA. (2016).USDOT Releases 2016 Fatal Traffic Crash Data, available online at: https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data.

Oh, J.-S., Oh, C., Ritchie, S. G., & Chang, M. (2005). Real-time estimation of accident likelihood for safety enhancement. Journal of Transportation Engineering, 131(5), 358–363. doi:10.1061/(asce)0733-947x(2005)131:5(358).

Osman, M., Mishra, S., & Paleti, R. (2018). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: accounting for unobserved heterogeneity and age group differences. Accident Analysis and Prevention, 118, 289–300. doi:10.1016/j.aap.2018.05.004.

Schmidt-Burkhardt, A. (2011). Learning machines. Maciunas' Learning Machines, 65–84. doi:10.1007/978-3-7091-0480-4_7.

Shi, Q., Abdel-Aty, M., & Yu, R. (2016). Multi-level Bayesian safety analysis with unprocessed automatic vehicle identification data for an urban expressway. Accident Analysis and Prevention, 88, 68–76. doi:10.1016/j.aap.2015.12.007.

Solomon, D. (1964). Accidents on main rural highways related to speed, driver, and vehicle. Federal Highway Administration, Washington, D.C., 1964 (Reprinted 1974).

Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer stacking framework. Accident Analysis and Prevention, 122, 226–238. doi:10.1016/j.aap.2018.10.016.

Traffic safety facts. (2017). U.S. Department of Transportation. NHTSA.

Wang, K., Simandl, J. K., Porter, M. D., Graettinger, A. J., & Smith, R. K. (2016). How the choice of safety performance function affects the identification of important crash prediction variables. Accident Analysis and Prevention, 88, 1–8. doi:10.1016/j.aap.2015.12.005.

Wang, Y. M., & Elhag, T. M. S. (2007). A comparison of neural network, evidential and multiple regression analysis in modelling bridge risks. Expert Systems with Applications, 32(2), 336–348.

World Health Organization. (2004). World report on road traffic injury prevention.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M., & Wu, C. (2011). Motor vehicle–bicycle crashes in Beijing: irregular maneuvers, crash patterns, and injury severity. Accident Analysis and Prevention, 43(5), 1751–1758. doi:10.1016/j.aap.2011.04.006.

Ye, X., Wang, K., Zou, Y., & Lord, D. (2018). A semi-nonparametric Poisson regression model for analyzing motor vehicle crash data. Plos One, 13(5). doi:10.1371/journal.pone.0197338.

Yu, R., & Abdel-Aty, M. (2013). Multi-level Bayesian analyses for single and multi-vehicle freeway crashes. Accident Analysis and Prevention, 58, 97-105. doi:10.1016/j.aap.2013.04.025.

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. Accident Analysis and Prevention, 73, 351–358. doi:10.1016/j.aap.2014.09.006.

Zhang, C., & Ivan, J. N. (2005). Effects of geometric characteristics on head-on crash incidence on two-lane roads in Connecticut. Transportation Research Record: Journal of the Transportation Research Board, 1908(1), 159–164. doi:10.1177/0361198105190800119.

Zhang, P., Parenteau, C., Wang, L., Holcombe, S., Kohoyda-Inglis, C., Sullivan, J., & Wang, S. (2013). Prediction of thoracic injury severity in frontal impacts by selected anatomical morphemic variables through model-averaged logistic regression approach. Accident Analysis and Prevention, 60, 172–180. doi:10.1016/j.aap.2013.08.020.