Report #24-UD-159

NSTSCE

National Surface Transportation Safety Center for Excellence

Risk Factors Re-evaluation with Bayesian Network Using SHRP 2 Data

Shu Han • Feng Guo

Submitted: September 11, 2024

Technology

Impairment

Mobility

Infrastructure

Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Zac Doerzaph from the Virginia Tech Transportation Institute; John Capp from General Motors Corporation; Terri Hallquist from the Federal Motor Carrier Safety Administration; Mike Fontaine from the Virginia Department of Transportation and the Virginia Transportation Research Council; and Melissa Miles from State Farm Insurance.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

EXECUTIVE SUMMARY

Traffic safety is a complex system influenced by numerous factors, including human behavior, road design, vehicle technology, and environmental conditions. Each of these factors can impact the safety of the transportation system in unique ways, and all factors could interact with each other in complex ways. The goal of this study was to evaluate the joint contribution of multiple risk factors to traffic safety by examining the interactions among different factors. This study considered 24 potential risk factors that reflect different perspectives in the analysis, including driver demographics, driving behavior, environmental conditions, road characteristics, traffic context, vehicle kinematics within a 5-second window of each event, and cell phone ban policies. There were two aspects to this study: first, it explored the relationships between traffic safety risk factors using unsupervised learning models with data from the Second Strategic Highway Research Program Naturalistic Driving Study. Second, with supervised learning models, the study developed a robust data-driven Bayesian network model, evaluated impacting risk factors, quantified their corresponding importance on driving risk, and consequently identified high-risk scenarios. The main findings are described below.

INTERRELATIONSHIPS AMONG RISK FACTORS

The results from unsupervised learning methods were consistent between association analysis and clustering analysis; that is, factors with higher Cramer's V values tend to be grouped into the same clusters (e.g., "age" and "risk score," and different types of distractions). The maximum weight spanning tree (MWST) structure reveals the relationship among five clusters as shown in Figure E1:

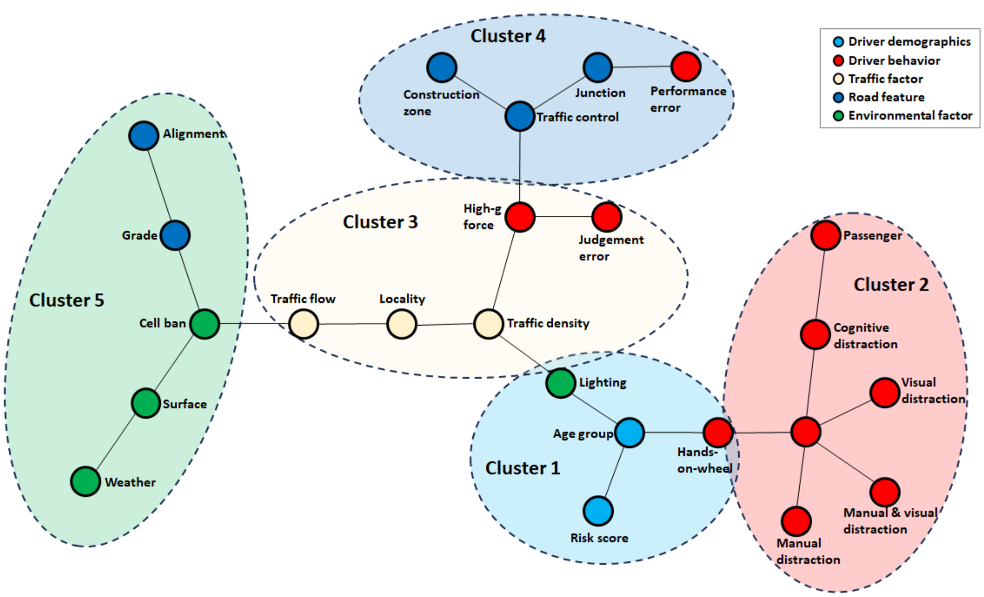
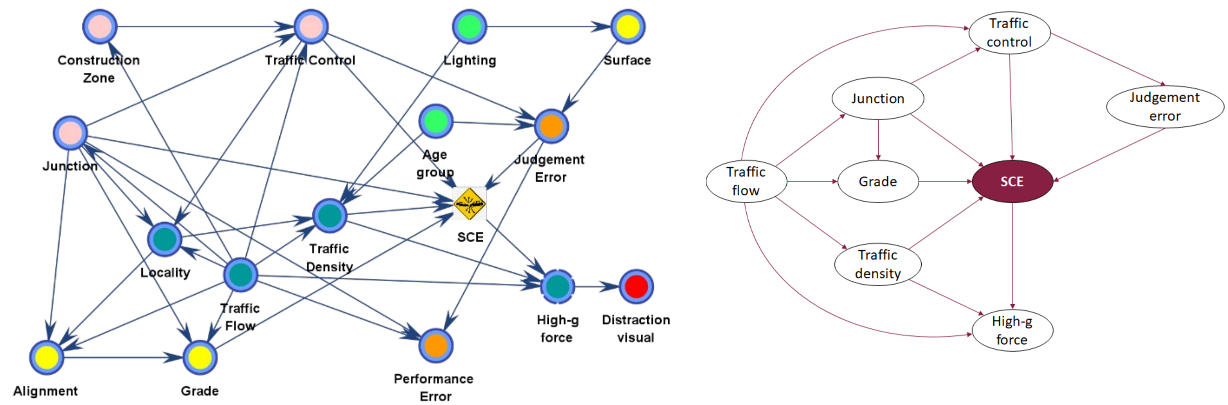


Figure E1. Diagram. Variable relationship via the MWST structure.

ESTIMATED BAYESIAN NETWORK STRUCTURE

The estimated network structure, which consists of 15 out of 24 variables, is shown in Figure E2(a), and the Markov blanket of a safety-critical event (SCE) outcome is shown in Figure

E2(b). The Markov blanket can encapsulate all the direct influences on a node and capture all the information needed to make that node conditionally independent of the rest of the network given its Markov blanket. There are five variables that are the parents of the safety outcome (i.e., SCE): judgment error, traffic density, traffic control, junction, and grade. The high g-force is the child of the safety outcome, traffic density, and traffic flow. Traffic flow is also included in the Markov blanket of the target node via being the spouse (i.e., coparent of high g-force).



(a) Estimated Bayesian network (b) Markov blanket of SCE

Figure E2. Diagrams. Estimated Bayesian network structure.

MARGINAL EFFECT OF RISK FACTORS

The conditional probability ratio was used as a measure of marginal effect.

$$conditional\ probability\ ratio = \frac{P(SCE = Yes|Test\ scenario)}{P(SCE = Yes|Reference\ scenario)}$$

For the marginal effect, the reference scenario is the low-risk scenario. The top five risk factors ranked by conditional probability ratio are judgment error, traffic density, traffic control, junction, and distraction (visual).

HIGH-RISK SCENARIOS

The scenario with the highest ratio occurs when all five parent nodes of SCEs are "Yes," i.e., $\mathrm{P(SCE = yes|Judgment\ error = yes, traffic\ density = heavy, traffic\ control = control, junction = yes, grade = nonlevel)}$. The conditional probability ratio for this scenario is 9.69 compared to when all five parent nodes are at a "safe" level. The four next riskiest scenarios are (1) $\mathrm{Judgment\ error = yes, traffic\ density = heavy, traffic\ control = control, junction = no, grade = nonlevel}$; (2) $\mathrm{Judgment\ error = yes, traffic\ density = heavy, traffic\ control = control, junction = yes, grade = level}$; (3) $\mathrm{Judgment\ error = yes, traffic\ density = heavy, traffic\ control = noncontrol, junction = yes, grade = nonlevel}$; and (4) $\mathrm{Judgment\ error = yes, traffic\ density = heavy, traffic\ control = noncontrol, junction = yes, grade = level}$.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

DAG	directed acyclic graph
DAS	data acquisition system
EM	Expectation-Maximization
EOR	eyes-off-road
EQ	Expectation-Maximization with Quantum Concepts
MDL	minimum description length
MWST	maximum weight spanning tree
NB	negative binomial
NDS	naturalistic driving study
PCA	principal component analysis
PSEM	probabilistic structural equation model
SC	structural coefficient
SCE	safety-critical event
SEM	structural equation model
SHRP 2	Second Strategic Highway Research Program
STE	standardized total effect
TAN	tree augmented naïve Bayes
TE	total effect

CHAPTER 1. INTRODUCTION

Traffic safety is a complex system that is influenced by numerous factors, including human behavior, road design, vehicle technology, and environmental conditions. Each of these factors can impact the safety of the transportation system in unique ways, and all these factors could interact with each other in complex ways. For example, driver behavior is often considered one of the most significant factors affecting traffic safety, and often drivers can make choices to decrease the risk of crashes, such as avoiding distracted driving, speeding, and the influence of drugs or alcohol, or by following traffic signals. Other factors such as road design, vehicle technology, and environmental conditions can influence driver behavior and hence the likelihood of crashes. There is a large body of research on the topic of crash risk factors, but the majority of the work used Poisson and negative binomial (NB) regression or logistic regression models, which can quantify the impact of individual factors (Dingus et al., 2016). However, driving risk is generally related to multiple factors rather than a single behavior or mistake. Understanding the complex interrelationship of factors contributing to traffic safety risks is crucial for developing effective strategies to mitigate road incidents and improve overall road safety. Traditional approaches to studying traffic safety risk factors have relied heavily on retrospective analysis of historical crash records, survey data, and simulator studies. While these methods have provided valuable insights, they often suffer from limitations, such as data biases, subjective self-reporting, and controlled environments that may not fully capture real-world driving behavior. In recent years, naturalistic driving studies (NDSs) have revolutionized the traffic safety research field by enabling researchers to collect rich, high-resolution data on actual driving behavior in real-world settings. Leveraging advances in sensor technology and data analytics, NDSs allow the collection of detailed information on driver demographics, driving behaviors, environmental factors, vehicle characteristics, and road infrastructure.

A Bayesian network is a probabilistic graphic model that represents the relationships between different variables in a system. It offers a powerful tool for modeling and analyzing the complex relationships among various factors, which can help to identify the key risk factors that influence traffic safety. In addition, a Bayesian network can (1) incorporate categorical data, continuous data, and time-series data and (2) make inferences about causal relationships from observational data possible by estimating the strength of the connections between the variables from the data. Caution should be taken with causal inference, as it requires additional assumptions and methods to deal with the possibility of confounding variables and other sources of bias. Several studies have been conducted on the application of Bayesian networks in traffic safety. Song et al. (2022) developed a Bayesian network model to evaluate the safety of automated vehicles, and the relationships between safety outcomes and operational design domain (ODD) attributes (e.g., driver behavior, weather, lighting condition, intersection geometry, traffic control device) were described. Keshan et al. (2015) compared over 10 different methods and showed that a Bayesian network outperformed other supervised learning algorithms when discerning between two different emotion levels with electrocardiogram signals.

This study has focused on evaluating the joint contribution of multiple risk factors to traffic safety by examining the interactions among different factors. There are two aspects to this study, as shown in Figure 1. First, it explored the relationships between traffic safety risk factors using unsupervised learning models with data from the Second Strategic Highway Research Program (SHRP 2) NDS. Unsupervised learning techniques, including association analysis, clustering, and a probabilistic structural equation model (PSEM), offer a data-driven approach to uncovering patterns and relationships within large, complex datasets. Second, with supervised learning models, the study aimed to (1) develop a robust data-driven Bayesian network model; (2) evaluate impacting risk factors and quantify corresponding importance on driving risk; and (3) identify high-risk scenarios.

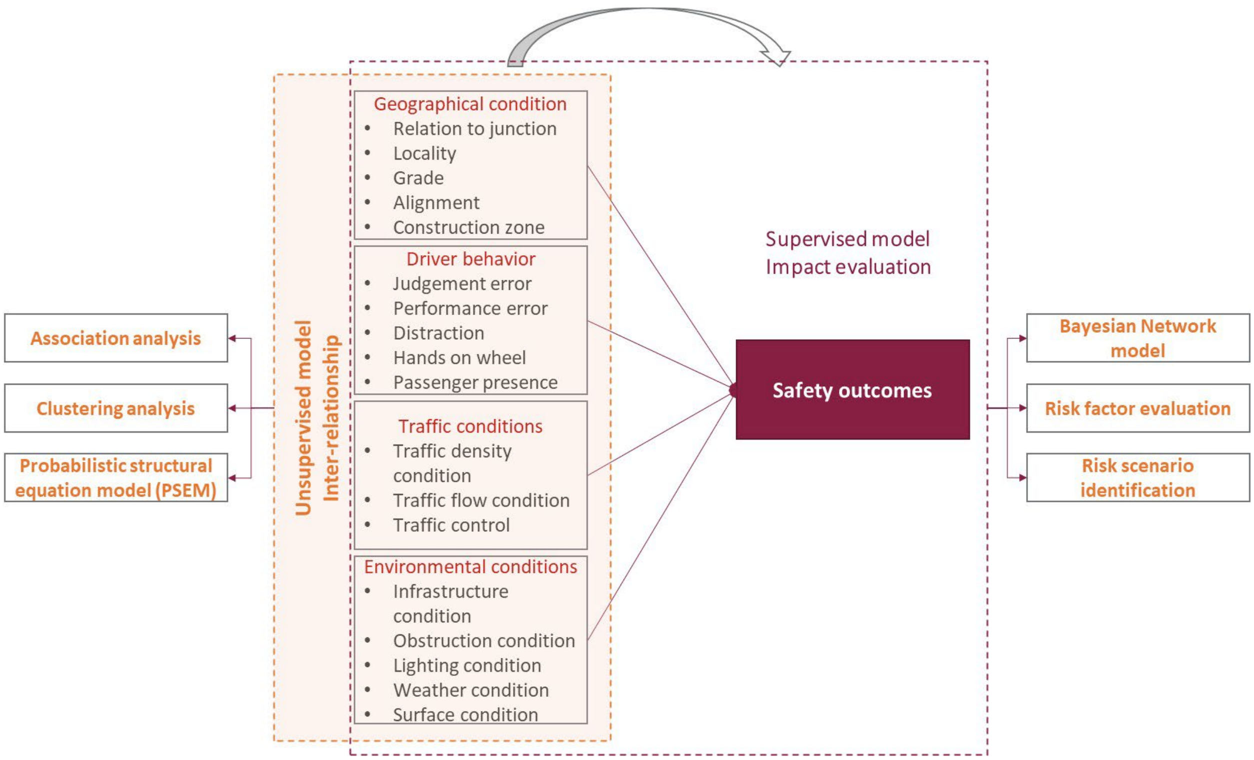


Figure 1. Diagram. Analysis framework.

This report is organized as follows: existing studies on traffic risk factors evaluation are reviewed, with unsupervised learning and supervised learning models reviewed separately; data and methodology are introduced; the results are presented for unsupervised learning (i.e., association analysis, clustering, relationship among all risk factors, and joint effect on safety outcomes) and supervised learning (i.e., model selection and inference); and the report closes with a summary and conclusion.

Potential applications of the method proposed in this study are as follows: (1) corresponding strategies can be developed regarding the identified risk factors; and (2) crash risk can be predicted for a certain scenario with certain combinations of risk factors, and warnings/alerts can then be pushed to drivers in advance via advanced driver assistance systems.

CHAPTER 2. RELATED WORK

This section first reviews commonly used risk factors, then introduces the structural equation model (SEM) to explore the interrelationship among risk factors, followed by an extensive review of supervised learning models used to evaluate the impact of risk factors on traffic safety outcomes.

REVIEW OF RISK FACTORS

The traffic system is complex, and a wide range of factors could contribute to driving risk, such as driver demographic characteristics, driver behavior, vehicle attributes, traffic conditions, environmental factors, law enforcement, and interventions. For example, Stavrinos et al. (2023) examined how adolescent drivers' attention to the driving task develops with both age and driving experience. Tapia et al. (2024) assessed the effects of cognitive training on young drivers' performance with a simulator, while Ivers et al. (2009) used questionnaire data to study risky driving behaviors and risk perceptions and their association with crash risk in young novice drivers.

Driving Behavior

Driving behavior, considered one of the key factors for driving risk, is also the most frequently assessed risky driving factor (Dingus et al., 2016; Wang & Xu, 2019; Zhu et al., 2017). Klauer et al. (2009) evaluated general risky driving behavior (e.g., traveling at inappropriate speeds, improper braking) in relation to crash/near-crash rates. Dingus et al. (2016) evaluated each crash risk factor individually with SHRP 2 NDS data, while Guo et al. (2022) conducted a similar study with Shanghai NDS data. Some researchers evaluated drivers' attention via eye-glance movement. The relationship between drivers' eye-glance patterns during lane change and crash risk was explored via an Embedding Kernel algorithm (Xu et al., 2024). Han et al. (2023) conducted an in-depth analysis on crash risk associated with eyes-off-road (EOR) duration on different road types. Anderson et al. (2024) investigated the EOR behavior of drivers when traveling on uncontrolled access roadways in vehicles equipped with SAE Level 2 automated features. Drivers' traffic violations (e.g., signal violation, drunk driving, speeding, lane drifting) have been used as an important component of assessing driver behavior for crash risk (Mallia et al., 2015; Park et al., 2019; Zhang et al., 2013). Among these behaviors, speeding and driving under the influence have been investigated thoroughly and found to be strongly correlated with fatal crashes. Negative emotions, including anger and fright, have been shown to significantly affect a driver's reaction time (Zimasa et al., 2019), perception ability (Lu et al., 2013; Mesken et al., 2007), driving performance (Precht et al., 2017; Zhang et al., 2016), and even attitude toward danger (Jeon et al., 2014). Zhang et al. (2019) investigated how anger while driving and aberrant driving behaviors are related to crash risk. Dozio et al. (2024) investigated how different emotional states affect drivers' attention in a simulator. Tang et al. (2023) integrated emotional factors into trajectory prediction. Fatigue is also a sub-category of impaired driving. Soccolich et al. (2024) explored the association between fatigue prevalence, secondary tasks, and corresponding risk, and de Winkel et al. (2024) investigated the relationship between fatigue/drowsiness, distraction/inattention, intoxication, sudden incapacitation, and speeding with crash risk. Health conditions can also have a substantial impact on driving risk, especially for commercial drivers (Hickman et al., 2020). Swain (2023) assessed whether diabetes-related visual function changes the associations of at-fault motor vehicle collisions.

Environmental Factors

As for environmental factors, some researchers focused on the impact of weather, region, and road type on driving risk (Brown, 2016; Sagar et al., 2020). Lee et al. (2018) used weather and crash records to analyze the extent to which rainfall intensity and level of water depth are responsible for traffic accidents. Ogungbire et al. (2023) compared different methods for weather-related crash severity predictions.

Familiarity with the environment was assessed by Özkan (2023), who used simulated driving to explore the impact of driving in left-side traffic on drivers who are familiar with right-side traffic.

Other Factors

With the rapid development of information technology, the collection of robust data with a wide array of variables has become increasingly available. Zhang et al. (2024) obtained time-invariant kinematic features (speed and speeding) and established their relationships with crash risk. Kinematic features and road characteristics were extracted from traffic camera data via computer vision algorithms, and their effect on crash risk was evaluated both individually and jointly (Sonth et al., 2023). The relationships between temporal driving behavior and individual crash risk were investigated by Zhang et al. (2023) with deep learning models.

Law and enforcement can also influence drivers' behavior and crash risk. Truelove et al. (2023), in a study conducted in Australia, investigated the impact of these elements on phone use while driving from the perspective of law enforcement. Findings indicated that drivers frequently attempt to avoid punishment for this behavior, with those who are successful in evading detection for using their phone while driving being significantly more likely to continue engaging in that behavior (Truelove et al., 2019; Truelove et al., 2021). Thompson and Wundersitz (2024) confirmed the effectiveness of mobile phone detection cameras based on the experiences from the aforementioned Australian study. However, Demir et al. (2024) claimed that the success of intervention varied by sociodemographic group (e.g., age, gender, driving experience).

REVIEW OF THE SEM

As these different approaches to driver behavior evaluation indicate, driver characteristics are complex and interrelated, and the relationships among variables are difficult to identify. A SEM can concurrently address the complex connections among endogenous and exogenous variables (Lee et al., 2008). Hamdar et al. (2008) developed a quantitative intersection aggressiveness propensity index using SEM techniques to observe environmental, situational, and driving behavior variables. Zhao et al. (2019) investigated the relationships between illegal driving behaviors and driver characteristics with a SEM based on a driving simulator. The influence of environmental and road characteristics, situational factors, and individual characteristics on speeding have also been explored (Javid et al., 2022; Sadia et al., 2018).

Some studies focused on multiple human factors. X. Song et al. (2021) explored the relationships among drivers' demographic characteristics (gender, age, and cumulative driving years), sensation seeking, risk perception, and risky driving behaviors with SHRP 2 NDS data via a SEM. Young drivers were assessed from different perspectives. For example, Scott-Parker et al. (2013) investigated relationships between the characteristics of drivers and risky driving of young drivers, while Mirón-Juárez et al. (2020) assessed the relationship of impulsivity, road anger expression, and the control of impulsivity and anger with driving risk in young drivers. Other researchers focused on older drivers; Liang et al. (2022) examined the relationships among age, gender, living status, and functional declines in older adult drivers with respect to driving exposure. Wong et al. (2018) integrated sociodemographic and driving-related variables (driving space, dependency on other drivers, health, and driving performance) and psychosocial appraisals (driving confidence, attitudes and beliefs towards driving) to predict driving self-regulation.

Some studies integrated multiple factors and evaluated their relationship. Kim et al. (2011) examined the severity of crashes in terms of human, vehicle, and roadway factors, etc., along with accessibility measures. Xu et al. (2022) used a SEM and Shanghai NDS data to analyze the structural relationships between driving risk and driver characteristics, vehicle kinematic features, and environmental factors. Bao et al. (2023) pulled out key features (road type, surrounding vehicle, kinematics of ego vehicle, environmental factors, and weather condition) from driving video clips, aiming to understand how different participants perceive risks during driving. Hai et al. (2023) concluded a meta-analysis on factors that influence driving behaviors and incorporated 43 factors into a SEM. The findings revealed that the combination of certain risk factors such as frontal collision, no lighting, and young drivers can cause more serious crashes (De Oña et al., 2011; Wang et al., 2019).

Summary

As presented above, a wide range of risk factors have been examined prior to this study. However, there are several limitations in these recent studies. First, most data sources are either survey data (Truelove et al., 2023), historical crash records (Lee et al., 2018), or simulator data (Papantoniou et al., 2021), all of which suffer various limitations. Second, some studies used illegal driving behavior (Zhao et al., 2019), traffic violations (Zhang et al., 2013), risk perception level (X. Song et al., 2021), or speeding (Zhang et al., 2023) as proxies for crash risk, approaches that need further justification. Finally, most previous research has focused on the impact of risk factors on driving risk but still lacks a systematic evaluation of the relationships among risk factors. This study aims to bridge the gap by applying a data-driven unsupervised learning approach to explore the potential risk factors collected from NDS data, including driver demographics, driving behaviors, surrounding environmental traffic situations, road infrastructure, vehicles' kinematic characteristics, and mobile phone ban policies. The detailed variables are listed in the next chapter.

REVIEW OF SUPERVISED LEARNING MODELS

Statistical Models

Generalized linear models including Poisson and NB models and their variations are most commonly used in traffic safety modeling where risk factors are treated as covariates (Guo, 2019). This method is very flexible in that either single or multiple factors can be included. For example, Fitch et al. (2013) examined the driving risk associated with cell phone distractions with a mixed-effect Poisson model, and Chen et al. (2016) investigated the impact of truck drivers' sleep patterns adjusted by driver demographics with an NB model. Furthermore, Antin, Guo, Fang, Dingus, Hankey, & Perez (2017); Antin, Guo, Fang, Dingus, Perez & Hankey (2017), and Guo et al. (2015) studied health conditions of senior drivers and corresponding annualized mileage and driving risk with NB models. Ouimet et al. (2014) used a mixed-effect longitudinal Poisson regression model to evaluate the effect of cortisol hormones on teen drivers' driving risk over time. A series of risk factors including driver demographics, personality, and driving characteristics were also measured with NB models (Guo & Fang, 2013). Statistical models have good interpretability, which allows researchers to understand the impact of each factor on driving risk. However, many statistical models assume linear relationships between predictor factors and the outcome, which may perform poorly on complex, high-dimensional datasets with nonlinear relationships and interactions, limiting the prediction accuracy.

Machine Learning Models

Machine learning models, on the other hand, such as decision trees and neural networks, can capture complex, nonlinear relationships and iterations, which allows more accurate prediction power. These models can also automatically learn relevant features from raw data, which reduces the need for manual feature engineering and potentially uncovers hidden patterns that may not be evident to manual processes. Finally, machine learning models often achieve higher predictive performance compared to traditional statistical models. Therefore, this technique has become popular in the research area of driving safety and crash risk prevention. Support vector machines have been used to study the temporal and spatial dilemma zone at a U-turn intersection (Khan & Mohapatra, 2023) and to predict crashes (Li et al., 2008; Mohamed, 2014; Sharma et al., 2016). Yang et al. (2019) investigated time-dependent crash risk by using the dangerous driving event data captured by smartphones with a multivariate conditional autoregressive model. Smartphone sensors were also used to predict driving behaviors such as aggressive driving and drowsy driving using machine learning methods (Ferreira et al., 2017; Li et al., 2020; Mantouka et al., 2021). Trajectory data were used to evaluate car-following risk with a feature learning model (Shi et al., 2019). Machine learning models were also employed to predict a vehicle's lane-change maneuvers (Xie et al., 2019) and a driver's secondary tasks (Osman et al., 2019). Drivers' biological characteristics including age, gender, and health status were considered through a machine learning method to predict driving risk (Ding et al., 2022; Koohestani et al., 2018). Li et al. (2023) used a stacked sparse autoencoder neural network, a type of artificial neural network that can capture complex relationships in the data to predict crashes. Other driving risk prediction methods are ensemble learning algorithms (Jamal et al., 2021; Umer et al., 2020), long short-term memory recurrent neural network models (Jiang et al., 2020; Ma et al., 2023; Yuan et al., 2019), deep learning techniques (Azhar et al., 2023; Dong et al., 2018; Ma et al., 2021), and gradient boosted decision trees (Chung, 2013; Dong et al., 2022; Zheng et al., 2018). Despite the popularity and performance of machine learning models, some machine learning models, particularly deep neural networks, are often considered “black box” models and not easily interpretable. Moreover, they usually require large amounts of training data to achieve optimal performance, which also requires significant computational resources.

Bayesian Network Methods

Network analysis offers a unique approach that can complement both statistical models and machine learning models. The approach focuses on modeling relationships and interactions between different elements, including drivers, vehicles, road characteristics, and environmental factors, as interconnected nodes and edges in a network graph. A Bayesian network is a type of robust probabilistic model with a graphic structure that presents a set of features and specifies their conditional dependencies using a directed acyclic graph (DAG; Pearl, 2009). Bayesian networks have been widely applied in risk assessment, considering their capability to conduct comprehensive and precise analysis on a sophisticated system (Kabir & Papadopoulos, 2019). Several existing studies have applied Bayesian networks to evaluate the effectiveness of this method on tunnel risk assessment (Borg et al., 2014), hazardous material transportation (Sun et al., 2022; Zhao et al., 2012); black spots identification (Gregoriades & Mouskos, 2013); driver behavior risk (Zhu et al., 2017); violation risk (Joo et al., 2022); rear-end crash severity (Chen et al., 2018; Chen et al., 2015); traffic crash severity (De Ona et al., 2013; De Oña et al., 2011; Mujalli & De Oña, 2011; Y. Song et al., 2021); and two-wheeler fatal crashes (Liu et al., 2022). Kinematic data have also been included in some recent studies. Zhu et al. (2017) integrated GPS data (road type, speed, high g-force events, and speeding), driver characteristics (age and gender), and driving experience (vehicle mileage traveled and driving experience) into a Bayesian network model to evaluate crash risk. In addition, Peng et al. (2021) established a vehicle-based Bayesian network risk model with detailed vehicle parameters and kinematics, as well as time and space headway, to the leading vehicle based on simulator data. Chen et al. (2022) conducted a similar study about on-road crashes (i.e., excluding intersections, ramps, etc.) by developing a predefined hierarchy Bayesian network structure with SHRP 2 NDS data. Sheehan et al. (2017) used kinematic data from a Level 3 semi-autonomous vehicle to estimate crash risk with a Bayesian network for the application of motor insurance risk pricing.

There are various novel extensions that incorporated Bayesian networks with other methods to improve the model performance and accuracy. For example, Hossain and Muromachi (2012) applied a two-step model: a random multinormal logit model to identify the most important impacting factors, and a Bayesian network model to conduct real-time crash prediction within the next 4 to 9 minutes. The performance evaluation results reflected that, at an average threshold value, the model can successfully classify 66% of future crashes with a false alarm rate less than 20%. Chen et al. (2015) and Lalika et al. (2022) adopted a similar approach, with the former evaluating crash severity and the latter focusing on older pedestrian fatalities/severe injuries. Zong et al. (2019) combined information entropy and a Bayesian network method to analyze crash severity. Zywiec et al. (2021) developed a novel methodology to incorporate a neural network metamodel into Bayesian network-based probabilistic risk assessment for industrial facilities. With the various temporal features becoming available, some researchers have applied dynamic Bayesian network models for real-time crash prediction. For example, Sun and Sun (2015) used speed data only to build a dynamic Bayesian network, and Song et al. (2022) incorporated traffic density data in addition to speed data to estimate secondary rear-end accidents with simulator data. Ma et al. (2019) integrated time-series kinematic and eye-movement characteristics in their dynamic Bayesian network models to predict real-time crash risk with experimental route data.

Summary

Above all, Bayesian networks have shown advantages in modeling uncertainty explicitly using probability theory, which can represent complex stochastic relationships. These models can also incorporate knowledge from domain experts, apply informative priors, update beliefs about variables based on observed data, and handle small sample size and/or missing data issues. However, the existing studies have often used historical crash records, which lack control samples (Chen et al., 2015; Joo et al., 2022; Lalika et al., 2022; Liu et al., 2022; Y. Song et al., 2021; Sun et al., 2022; Zhang et al., 2013; Zhao et al., 2012; Zong et al., 2019). Some researchers relied on time-series kinematic data (Ma et al., 2019; Sheehan et al., 2017; Sun & Sun, 2015), but these often lack contextual information; simulators can collect detailed data from various aspects, but the validity and generalization of conclusions still need to be justified (Gregoriades & Mouskos, 2013; Peng et al., 2021; Song et al., 2022). NDS data utilize unobtrusive data gathering equipment to record information about the driver, the vehicle, and the surroundings (Dingus et al., 2016), allowing researchers to evaluate driving risk under real-world scenarios (Van Schagen et al., 2011). With SHRP 2 NDS data, this paper employs a data-driven approach to build a robust Bayesian network model to evaluate impacting factors and high-risk scenarios.

CHAPTER 3. DATA AND METHODOLOGY

This section introduces SHRP 2 NDS data and variables considered in the study, unsupervised learning methods, and supervised learning methods.

DATA

Overview of SHRP 2 NDS Data

The SHRP 2 NDS (Hankey et al., 2016) is the largest study of naturalistic driving behaviors to date, with approximately 3,400 participant drivers, more than 1 million hours of driving, and 49.5 million travel miles of naturalistic driving data collected between 2010 and 2013. Data were collected from six sites around the U.S.: Seattle, Washington; Tampa, Florida; Buffalo, New York; Durham, North Carolina; State College, Pennsylvania; and Bloomington, Indiana. Participants' vehicles were instrumented with a data acquisition system (DAS) that collected four video views (i.e., driver's face, driver's hands, forward roadway, rear roadway), vehicle network information (e.g., speed, brake, accelerator position), and information from additional sensors included with the DAS (e.g., forward radar, accelerometers). In addition to in-vehicle collected data, the SHRP 2 NDS also collected comprehensive participant assessment, roadway, vehicle, and cell phone records data, as shown in Figure 2.

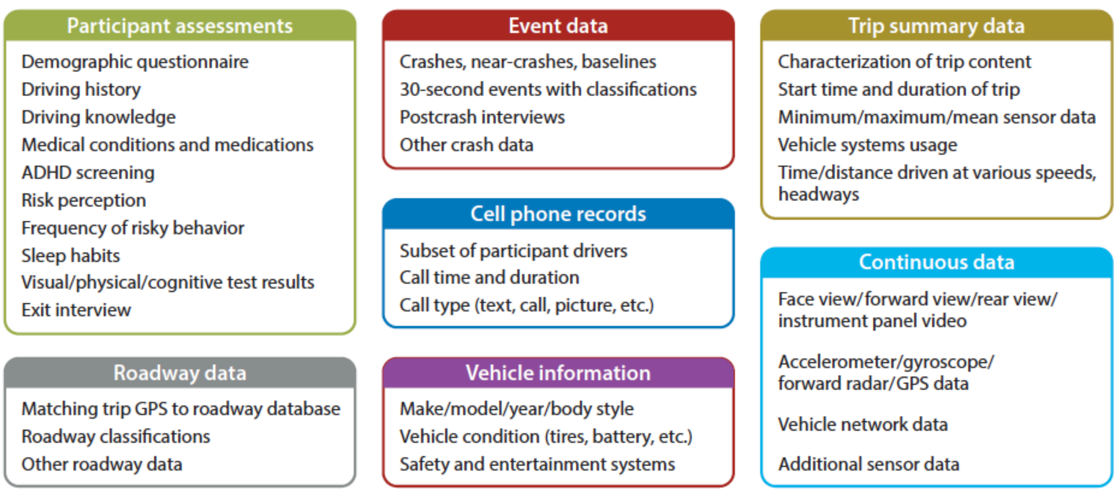


Figure 2. List. SHRP 2 NDS data elements (Guo (2019).

Detailed observation and coding of over 8,000 crashes/near-crashes and over 19,000 balanced baselines were conducted, which can reveal critical information related to traffic safety. This information includes event type, corresponding driving behavior, environmental conditions, road characteristics, and traffic context, which are considered as potential risk factors in this study. In addition, each driver's age and risk score based on questionnaire data, vehicle kinematics within a 5-second window of each event, and cell phone ban policies at different sites are also incorporated; these can reflect risk factors from different perspectives. Overall, 24 potential risk factors are included in the analysis, as shown in Figure 3.

Age	Judgement error	Performance error	Distraction
Cognitive distraction	Manual distraction	Visual distraction	Manual & visual distraction
Traffic density	Locality	Traffic flow	Lighting
Weather	Road surface	Traffic control type	Junction type
Road alignment	Road grade	Construction zone	Hands-on-wheel
Interaction with passenger(s)	Kinematic feature	Risk score	Cell ban policy

Figure 3. Text box. Potential risk factors.

Variable Descriptions

Twenty-five available factors are shown in Table 1, including one target variable (i.e., safety outcomes, or safety-critical events [SCEs]) and 24 risk factors selected from reduced SHRP 2 NDS data. Detailed definitions for each factor are available in Hankey et al. (2016). Among these risk factors, six categories are considered that cover driver demographics, driver behavior, driver distraction, traffic, environmental factors, road geographic characteristics, and policy. Most variables have two levels (Yes vs. No) apart from age group (teen/young, middle-aged, and senior), risk score (low, medium, and high risk), kinematic features (low, medium, and high values), and policy (less, moderately, and more strict). The criteria for age groups are based on a previous SHRP 2 study (Guo et al., 2017). Two continuous variables (i.e., risk score and kinematic features) are discretized with a normalized equal distance method (Kotsiantis & Kanellopoulos, 2006).

Table 1. Variable list (SHRP 2 NDS).

Category	Factor	Level	State(s)	Sample Size
Target	SCE	2	Crash/near-crash	8,409
			Baseline	19,184
Driver demographics	Age group	3	Teen/Young (16-29)	13,655
			Middle (30-64)	8,236
			Senior (65+)	5,702
	Risk score	3	Low risk score (<=1.71)	19,789
			Medium risk score (1.71~2.419)	7,424
			High risk score (2.419~3.129)	380
Driver behavior	Kinematic feature	3	Low value (<=0.437)	20,902
			Medium value (0. 437~2.742)	6,414
			High value (>2.742)	277
	Judgment error*	2	No	25,665
			Yes	1,928
	Performance error*	2	No	26,271
			Yes	1,322
	Hands on wheel	2	Both hands	10,550

Category	Factor	Level	State(s)	Sample Size
			Non-both hands	17,043
Driver distraction	Distraction (overall)	2	No	12,259
			Yes	15,334
	Distraction cognitive	2	No	21,435
			Yes	6,158
	Distraction manual	2	No	26,569
			Yes	1,024
	Distraction visual	2	No	23,376
			Yes	4,217
	Distraction manual & visual	2	No	23,841
			Yes	3,752
	Passenger	2	No	7,846
			Yes	19,747
Traffic	Traffic density	2	Light	16,340
			Heavy	11,253
Environmental factors	Light condition	2	Daylight	21,468
			Non-daylight	6,125
	Weather	2	No adverse	24,895
			Adverse	2,698
	Surface type	2	Dry	22,950
			Non-dry	4,643
	Construction zone	2	No	26,329
			Yes	1,264
Road geographic characteristics	Traffic flow	2	Non-divided	13,651
			Divided	13,942
	Locality	2	Residential	6,216
			Non-residential	21,377
	Traffic control	2	Non-control	22,000
			Control	5,593
	Junction type	2	Non-junction	12,700
			Junction	14,893
	Alignment	2	Straight	23,623
			Non-straight	3970
	Grade	2	Level	23,066
			Non-level	4,527
Policy	Cell ban**	3	Less strict	8,184
			Moderately strict	6,418
			More strict	12,991

* Both performance errors and judgement errors contain driver behavior information about what the driver did to cause or contribute to the crash or near-crash, but the focus of the two types of error is different. Performance errors include the following 10 types: apparent inexperience; blind spots errors; improper turn; right-of-way error; signal violation (apparently did not see

signal); stop yield violation (apparently did not see stop sign); wrong side of road; driving too slow; sudden or improper brake or stop; and fail to signal. Judgement errors include the following six categories: aggressive driving; speeding; illegal passing; following too close; intentional signal violation; and intentional stop yield violation.

** Cell ban while driving is based on each state's policy during 2010 to 2013 (i.e., data collection period) in the corresponding site. The less strict group contains Florida and Pennsylvania, which had no handheld bans or novice driver cell use bans; the moderately strict group consists of North Carolina and Indiana, with texting bans for all drivers and novice drivers prohibited from all cell use; the more strict group included Washington and New York, where texting, handheld use, and novice driver cell use were all banned prior to 2010.

Three metrics—standard deviation of longitudinal (x), lateral (y), and vertical (z) acceleration within the 5-second window of each event—are used as indicators of high g-force events (Shi et al., 2022) and thus were selected to represent kinematic features in this study. The 3-dimensional acceleration data were linearly transformed onto a new coordinate system via principal component analysis (PCA; Abdi & Williams, 2010) such that the main directions (principal components) capturing the largest variation can be used as input to explore the relationship among impacting factors.

The results include two main components: the first component explains 57.5% of variance while the second component explains 28.6%, as shown in Figure 4. All three standard deviations of acceleration point to the same side of the first component's axis (Dim1), which implies a positive relationship. The positive trend remains between standard deviation of longitudinal acceleration and the second component (Dim2), but the standard deviations of lateral and vertical acceleration show a negative relationship. The projected length of each variable vector represents the contribution to the corresponding component. For example, standard deviation of lateral acceleration contributes the most to the first component, followed by vertical and longitudinal. The ranking of contribution to the second component from highest to lowest is longitudinal, vertical, and lateral. Note that only the first component is used as input for Bayesian network model development, such that key kinematic features can be considered but not dominate compared to other attributes.

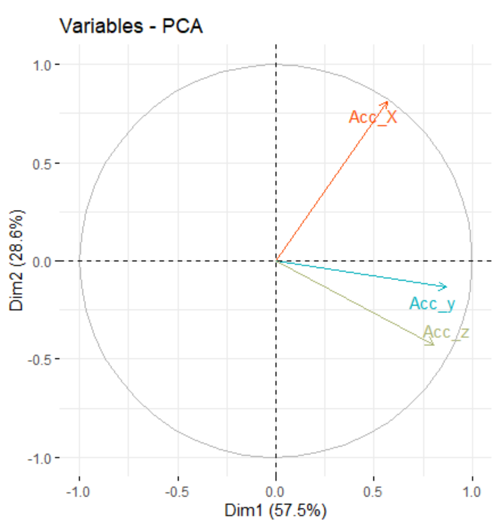


Figure 4. Graph. PCA results of 3-dimensional acceleration data.

METHODOLOGY

Unsupervised Learning Methods

The three unsupervised learning methods employed in this study were association analysis (Cramer's V), clustering analysis (Kullback-Leibler [KL] divergence), and PSEM.

Strength of Association

The strength of association between two categorical variables is measured by Cramer's V (McHugh, 2013). Cramer's V ranges from 0 to 1, with 0 indicating there is no association between the two variables, and 1 indicating a perfect association between the two variables. It is computed by taking the square root of the χ^2 statistic and dividing it by the sample size n and the minimum dimension minus 1, as shown below:

$$Cramer's\ V = \sqrt{\frac{\chi^2/n}{\min\,(k-1, r-1)}}$$

where χ^2 is the chi-squared statistic.

Cramer's V is a quantitative measure of association strength, which is particularly useful when the sample size is small.

Variable Cluster

Variables are clustered via a hierarchical agglomerative clustering algorithm that computes the distance between nodes and cluster variables with KL divergence (Hershey & Olsen, 2007). KL divergence can measure the strength of the relationship between two variables by measuring the difference between their distributions P and Q. For discrete variables in this study, the formula of KL divergence is defined as follows:

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)}\right)$$

where P and Q are discrete probability distributions defined on the same sample space $\mathcal{X}$.

At the beginning of the variable clustering process, each variable is treated as a distinct cluster. The clustering algorithm proceeds iteratively by merging the "closest" clusters into a new cluster. Several criteria are often used for determining the number of clusters:

- Stop threshold: a minimum value below which clusters are not merged.
- Maximum cluster size: the maximum number of variables per cluster.
- Cluster number: the number of clusters.

PSEM

A PSEM extends the traditional SEM framework by explicitly incorporating probabilistic dependencies and uncertainties. The traditional SEM (Bowen & Guo, 2011) consists of a measurement model and a structural model, as follows:

Measurement model:

$$X = \Lambda_X \xi + \delta; Y = \Lambda_Y \eta + \varepsilon$$

where X and Y are exogenous and endogenous variables; Λ_X and Λ_Y are corresponding structural coefficients; ξ and η are observations of exogenous and endogenous variables; and δ and ε are error terms.

Structural model:

$$F = B\eta + \Gamma\xi + \zeta$$

where F is the structural function; η is the latent endogenous latent variables ($(\mathrm{m} \times 1)$ vector); ξ is the latent exogenous variables ($(\mathrm{n} \times 1)$ vector); and ζ is the error term ($(\mathrm{m} \times 1)$ vector). Both B and Γ are the structural coefficients of the model while B is an $m \times m$ coefficient matrix for the latent endogenous variables, and Γ is an $m \times n$ coefficient matrix for the latent exogenous variables.

The pros of PSEM (Conrady & Jouffe, 2015) compared to the traditional SEM are (1) PSEM explicitly incorporates probabilistic dependencies between variables unlike the deterministic relationship in SEM; (2) the observed and latent variables are also described with probability distributions, which can be parametric (e.g., Gaussian, binomial) or non-parametric (e.g., kernel density estimators) and reflect the uncertainties and variability in the data; and (3) parameter estimations in PSEM are typically done with Bayesian methods rather than maximum likelihood estimations or generalized least squares in SEM. This makes PSEM more flexible and suitable for modeling systems where uncertainties and variability play a significant role, and PSEM is therefore employed in this study to explore the relationship among potential risk factors.

Supervised Learning Methods

In this section, both supervised learning and semi-supervised learning methods used in this study are introduced; then, a model selection criterion is presented.

Supervised Learning Models

Supervised learning models belong to a machine learning paradigm where the algorithm learns from labeled data, and each data point consists of an input (variables or features) and an output (or target) variable.

A naïve Bayes structure (Perez et al., 2006) is a Bayesian network with only one parent, the target node. It assumes the variables are conditionally independent given the class label. The only arcs in the graph are those directly connecting the target node to a set of other nodes (i.e., impacting factors). Mutual information (Kraskov et al., 2004) is used to measure the amount of information gained on the target node Y by observing an impacting factor X. The formula is as follows:

$$I(X,Y) = \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 \frac{p(y|x)}{p(x)}$$

The Markov blanket of a variable consists of its parents, children, and other parents of its children. It represents a set of variables that are conditionally independent of the variable given its Markov blanket. The Markov blanket method is to learn the Markov blanket of the target variable in the Bayesian network (Bui & Jun, 2012). Compared to naïve Bayes, supervised Markov blanket learning offers a more efficient and interpretable approach to identifying relevant variables for predicting the target.

Within these two basic methods, two upgraded versions (Conrady & Jouffe, 2015)—augmented and tree augmented—are also compared in this study. The augmented algorithms relax the independence assumption among variables in the basic algorithms, while the tree augmented algorithms specify a tree structure dependence among variables. For example, augmented naïve Bayes extends the basic naïve Bayes model by relaxing the assumption of variable independence. Analogous to regression models, interaction terms are also taken into consideration rather than main effect only. By allowing more flexibility in modeling the relationships between variables, augmented naïve Bayes can potentially achieve higher accuracy. Similarly, tree augmented naïve Bayes (TAN) is an extension of the naïve Bayes classifier that introduces a tree structure to model dependencies between features, overcoming the oversimplified independence assumption of the traditional naïve Bayes classifier. While both augmented naïve Bayes and TAN aim to overcome the limitations of the traditional naïve Bayes classifier, TAN specifically introduces a tree structure to model dependencies between features. This structured approach may offer advantages in terms of interpretability and capturing complex dependencies in the data, compared to the more general enhancements made in augmented naïve Bayes.

Semi-supervised Learning Models

Semi-supervised learning (Van Engelen & Hoos, 2020; Zhu, 2005) is a machine learning method in which a model is trained on a dataset that contains both labeled and unlabeled data. The main idea behind semi-supervised learning is to leverage the abundance of unlabeled data to improve the model's performance. By incorporating unlabeled data during training, the model can learn more about the underlying structure of the data distribution and generalize better to unseen examples. The following three algorithms (Conrady & Jouffe, 2015) are employed in semi-supervised learning:

- **Expectation-Maximization with Quantum Concepts (EQ)** is developed by BayesiaLab. It extends the Expectation-Maximization (EM) algorithm by incorporating "quantum-like" concepts and allows for more efficient and effective learning in complex datasets.
- **Taboo** is an extension of the EQ algorithm that incorporates additional constraints or optimization techniques to prevent convergence to suboptimal solutions.

- **Taboo order** is an extension of Taboo that focuses on learning the order of variables within the Bayesian network structure. This can potentially impact the performance and accuracy of the resulting model.

The above algorithms can be combined with different orders, such as Taboo, EQ + Taboo, Taboo + EQ, and Taboo order + Taboo. While all of them aim to learn a Bayesian network in a semi-supervised learning setting, they differ in focus within the learning process. For example, Taboo EQ combines the Taboo algorithm with the EQ algorithm, where EQ may be primarily responsible for the learning process, with Taboo providing additional constraints or optimization techniques. Taboo order + Taboo focuses on utilizing two instances of the Taboo algorithm, with one instance dedicated to learning the variable order and the other for parameter learning.

In addition to considering a variety of learning algorithms, to avoid overfitting and increase the robustness of the developed model, 90% of the dataset was set as the learning set with the remaining 10% as the test set. In addition, data perturbation was used to perturb each observation by multiplying a random perturbation value, which is drawn from a normal distribution. The purpose of data perturbation is to introduce variability into the data and to help escape from local minima during the learning process.

Model Selection

The "best" model is selected that best predicts the target variables while considering the complexity of the resulting network. The network's complexity is measured by structural coefficient (SC; (Reise et al., 2013). The default value of SC is set to 1, which reliably prevents the learning algorithms from overfitting the model to the data. When there are relatively small sample sizes, SC will be decreased, which is equivalent to increasing the number of observations via resampling. An SC value of 0 would create a fully connected network. On the other hand, an SC with value greater than 1 is equivalent to sampling the dataset, which can help manage the complexity of networks learned from large datasets. In this study, the original dataset was split into a learning set (90%) and test set (10%), and the prediction performance was evaluated as a function of varying SC levels between 0.5 and 1 due to small sample size, especially in certain scenarios (e.g., the number of SCEs under manual distraction and high-risk score group is only 8). The SC value was set as 0.8 to balance the network complexity and avoid overfitting the data.

Multiple learning algorithms are compared via the minimum description length (MDL) score (Hansen & Yu, 2001), which is commonly used in statistics and machine learning for model selection. The MDL is a two-component score that is based on the idea that the best model is the one that compresses the data most effectively, balancing the complexity of the model with its ability to explain the data.

The MDL principle states that the best model is the one that minimizes the total description length required to encode both the model and the data. The MDL score is the sum of these two components:

$$MDL(B, D) = \alpha \times DL(B) + DL(D|B)$$

where:

- α represents the preset SC, which changes the weight of the structural part of the model. Lower value in α results in a more complex network.
- MDL ($DL(B)$): This part of the description length measures the complexity of the model itself. It quantifies the number of bits required to encode the model structure and parameters. The minimum value of $DL(B)$ is obtained via the simplest structure, i.e., the fully unconnected network.
- Data description length ($DL(D|B)$): This part of the description length measures how well the model compresses the data. It quantifies the number of bits required to encode the data given in the model. The minimum value of $DL(D|B)$ is obtained from the fully connected network.

Therefore, minimizing the MDL score finds the best trade-off between both model and data. To use the MDL score for model selection, a set of candidate models is considered, and the MDL score is computed for each model. The model with the lowest MDL score is then selected as the best model.

CHAPTER 4. RELATIONSHIP EXPLORATION OF TRAFFIC SAFETY RISK FACTORS VIA UNSUPERVISED LEARNING MODELS

Using Cramer's V, a pair-wise relationship was assessed among 24 potential risk factors. A variable cluster was then conducted via a hierarchical agglomerative clustering algorithm. Using the clustered results, the relationship of variables within and across clusters was investigated. Lastly, the causal path between clusters and safety outcomes was explored.

STRENGTH OF ASSOCIATION

The pair-wise strength of association among 24 potential risk factors is illustrated in Figure 5. The matrix is symmetric, and the diagonal is left blank because the association within the same variable is not very meaningful. The color and size of dots represent the value of Cramer's V between the pairs, with the darker and bigger dots indicating a higher Cramer's V.

Figure 5. Graph. Strength of association among 24 potential risk factors.

The highest strength of association occurs between weather condition and surface type (0.7). Different types of distractions show a certain level of association, with Cramer's V values ranging from 0.2 to 0.5. Age appears to be associated with several variables, such as judgment error, hands-on-wheel, risk score, distraction, and others.

Though Figure 5 reveals the strength of association among all 24 risk factors, it only describes the pair-wise relationship. Results from PSEM are presented in the following section.

CLUSTER OUTPUT

Cluster results and the distance between variables within a cluster are demonstrated with the number of clusters set at five, as shown in Figure 6. The five clusters can be described as driver-, distraction-, traffic-, road-, and environment-dominated attributes. As for cluster 1, "Age group" is grouped with "Risk score," which are both driver demographic characteristics whose association has been established by existing studies (Guo & Fang, 2013; Ryan et al., 1998; Wong et al., 2018). All distraction behaviors are grouped into cluster 2. "Distraction" is grouped with "Distraction-cognitive" first, then "Distraction-visual," "Distraction-visual & manual," "Passenger," and finally "Distraction-manual." Traffic-related factors are dominant in cluster 3 and include traffic density, traffic flow, and locality. "High g-force" is more likely to be driver behavior, but the association between "High g-force" and "Traffic density" is shown in Figure 6. In addition, roadway speed has been shown to be the most important predictor of all types of acceleration rates. Road feature factors are dominant in cluster 4, with the exception of performance error; however, drivers tend to have performance errors under complicated conditions, such as in construction zones or at junctions (Devlin et al., 2011). Environmental factors are dominant in cluster 5. In this cluster, "Weather" and "Surface" are grouped together, which is consistent with the Cramer's V results; "Cell ban policy" is grouped with "Grade," possibly because "Cell ban policy" is defined by the level of cell ban while driving in different states, which may have very different geographic characteristics and hence substantial differences in road grade.

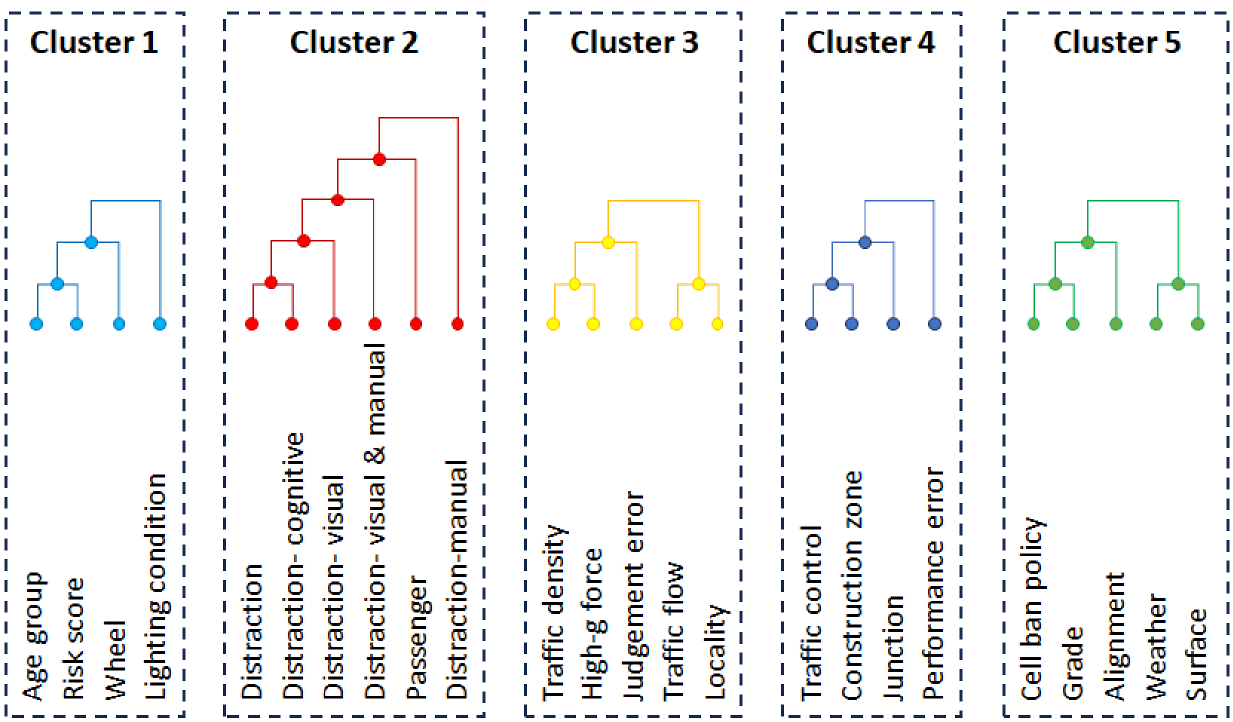


Figure 6. Diagram. Variable clusters.

RELATIONSHIP AMONG VARIABLES

Unsupervised structural learning, a form of knowledge discovery without hypotheses on possible relationships among variables, was used to explore the interrelationship among all 24 potential factors simultaneously. The maximum weight spanning tree (MWST) (Conrady & Jouffe, 2015; Pettie & Ramachandran, 2002) is used to learn a tree structure (i.e., one parent per node). MWST can provide a simple but insightful structure without too much computing burden. The tree structure is shown in Figure 7.

The structure in Figure 7 is consistent with Figure 6, but reveals the interrelationship among variables across different clusters. For example, the association between "Weather" and "Surface" is shown through Cramer's V, while "Cell ban" is associated with "Grade," and "Grade" is connected with "Alignment," and all are in cluster 5, where the environmental features play a main role. Meanwhile, "High g-force" is connected with "Traffic density" and "Judgment error" within cluster 3, and this cluster is linked with cluster 1, cluster 4, and cluster 5. Driver characteristics, locality, and driving errors are associated, as shown in Papantoniou et al. (2019). Cluster 1 and cluster 2 are connected, which suggests an association between driver distraction and driver demographic characteristics, as shown in Guo et al. (2017). Note that the tree structure is a data-driven approach and extracted automatically without any prior domain knowledge.

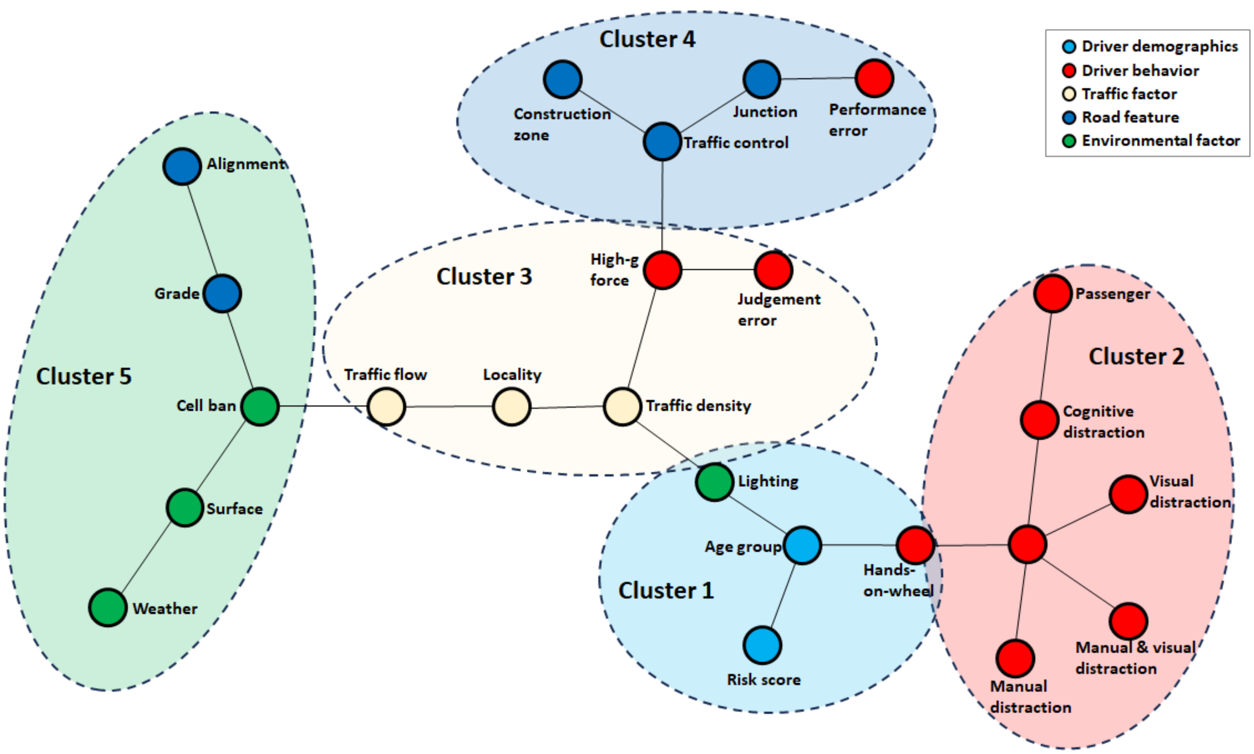


Figure 7. Diagram. Variable relationship via the MWST structure.

JOINT EFFECT ON SCEs

With the clustered factors, the structure of all 24 variables on SCEs is shown with a PSEM, depicted in Figure 8. Original variables are listed in rectangles, and different colors indicate different categories. Values after each original variable represent the contribution of these variables to the corresponding cluster. Three clusters—cluster 2 (distraction-dominated), cluster 3 (traffic-dominated), and cluster 4 (road-dominated)—have a direct impact on SCEs. Cluster 1 (driver-dominated) and cluster 5 (environmental-dominated) have an indirect impact on SCEs via cluster 2 (distraction-dominated) and cluster 3 (traffic-dominated). The overall network performance was evaluated by splitting the dataset into a learning set (90%) and test set (10%). The contingency table fits, which compares the entropy of the current naive Bayes structure to the entropy of a fully connected structure, are 63% and 51% correspondingly. This performance is acceptable in practice.

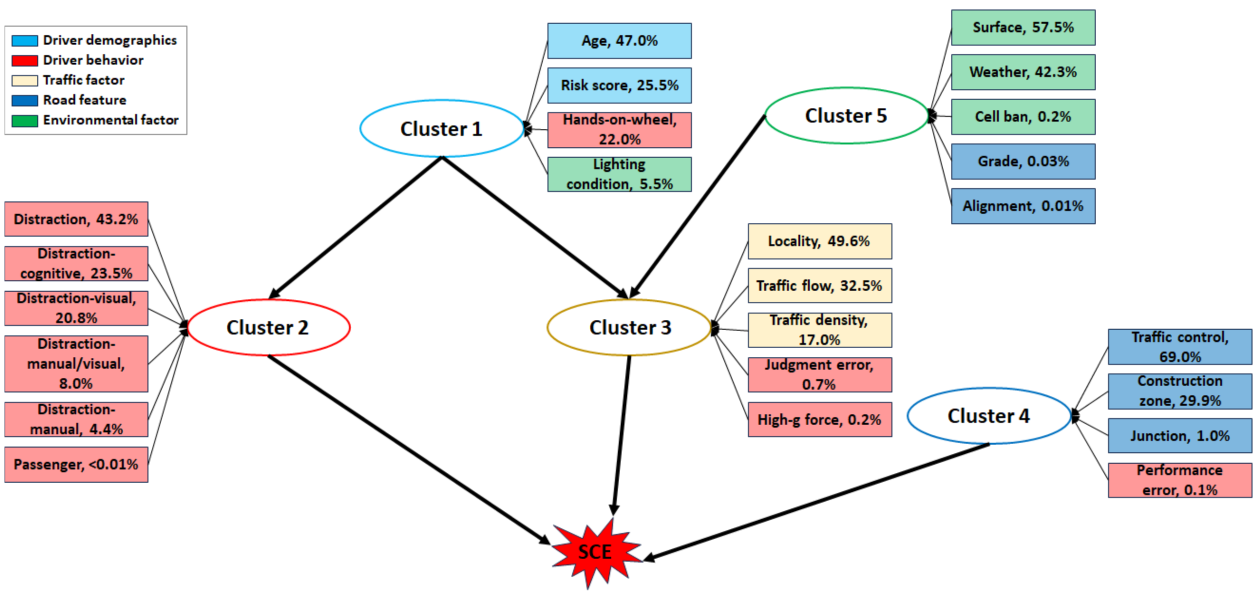


Figure 8. Diagram. Impact structure of clustered factors on SCE.

CHAPTER 5. ASSESSMENT ON DRIVING RISK FACTORS AND HIGH-RISK SCENARIOS VIA BAYESIAN NETWORK MODELS

BayesiaLab (Conrady & Jouffe, 2015), a software platform designed for probabilistic reasoning and decision-making using Bayesian networks, was used in this study to develop Bayesian network models. The results are presented in two parts: Bayesian network model selection and inference. The inference section contains three subsections, including feature importance based on total effect, and the marginal and joint effects of contributing factors based on conditional probability ratios.

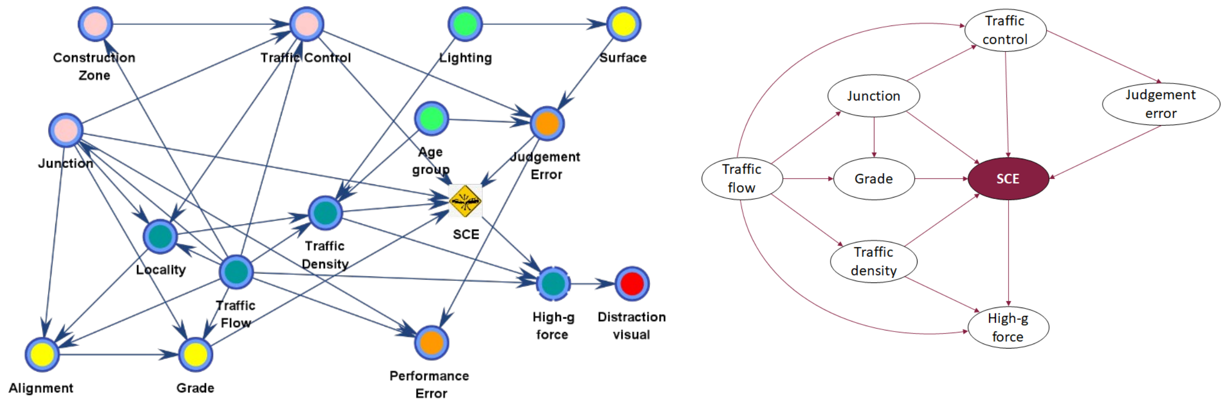
MODEL SELECTION

Ten algorithms were compared as shown in Table 2. Overall, semi-supervised learning algorithms outperform supervised methods, and MDL scores with perturbated data are close to those without perturbation, which implies the stability of the estimated network. Semi-supervised EQ + Taboo was selected as the final model in the study.

Table 2. MDL score of different algorithms.

Algorithm	MDL Score	MDL Score (Data Perturbation)
Naïve Bayes	640,469.385	640,547.558
Augmented Naïve Bayes	632,125.815	627,907.054
Tree Augmented Naïve Bayes	632,797.999	629,146.715
Markov Blanket	640,365.706	640,490.301
Augmented Markov Blanket	632,209.322	627,937.825
Tree Augmented Markov Blanket	632,689.668	628,825.439
Semi-supervised (Taboo)	600,035.021	**567,030.589**
Semi-supervised (EQ+Taboo)	**566,647.73**	567,316.191
Semi-supervised (Taboo EQ)	567,395.655	567,394.517
Semi-supervised (Taboo order+Taboo)	566,659.058	576,807.562

The estimated network structure, which consists of 15 out of 24 variables, is shown in Figure 9(a), and the Markov blanket of SCE is shown in Figure 9(b). The Markov blanket is made of a set of nodes that fully shields a node from the influence of all other nodes in the network, including parents (nodes that have direct edges leading into the node), children (nodes that have direct edges leading out from the node), and spouses (nodes that are coparents of the child nodes). It can encapsulate all the direct influences on a node and capture all the information needed to make that node conditionally independent of the rest of the network given its Markov blanket. There are five variables that are the parents of the safety outcome (i.e., SCE): judgment error, traffic density, traffic control, junction, and grade. The high g-force is the child of the safety outcome, traffic density, and traffic flow. Traffic flow is also included in the Markov blanket of the target node via being the spouse (i.e., coparent of high g-force).



(c) Estimated Bayesian network (d) Markov blanket of SCE

Figure 9. Diagrams. Estimated Bayesian network structure.

INFERENCE

Despite SCEs having five parents that have direct impact on the safety outcomes, other nodes in the developed Bayesian network have indirect impact. This section quantifies the feature importance based on corresponding total effect. Corresponding marginal effect is also assessed. Note that input uses reduced events that include balanced baselines, which is a case-control study. As a result, the estimated probability of having SCEs is not very meaningful. Conditional probability ratios are proposed to evaluate high-risk scenarios.

Feature Importance

The importance of each variable is evaluated by the total effect (TE), which estimates the expected change in the target node associated with a change in an impacting factor, considering both direct and indirect paths without necessarily implying causation. It is the derivative of the target node with respect to the impacting factor, as follows:

$$TE(X,Y) = \frac{\delta_Y}{\delta_X}$$

where X is the impacting factor, and Y is the target node.

The standardized total effect (STE) takes the variation into consideration, representing the total effect multiplied by the ratio of the standard deviation of the impacting variable and the standard deviation of the target node as follows:

$$STE(X,Y) = \frac{\delta_Y}{\delta_X} \times \frac{\sigma_X}{\sigma_Y}$$

The importance of each variable in the developed Bayesian network is shown in Table 3. The top four most important variables are judgment error, traffic density, traffic control, and junction. Note high g-force is excluded because this factor is a consequence (child) rather than a cause (parent) of the safety outcomes. All 15 variables are statistically significant at $\alpha = 0.05$ except for surface condition and road alignment.

Table 3. Importance of impacting factors.

Node	Total Effect	Standardized Total Effect	G-test	Degree of Freedom	*p*-value
High g-force	0.427	0.772	18,718	2	<0.001
Judgment error	0.383	0.212	1,119	1	<0.001
Traffic density	0.234	0.250	1,713	1	<0.001
Traffic control	0.159	0.139	511	1	<0.001
Junction	0.123	0.134	498	1	<0.001
Distraction visual	0.101	0.079	166	1	<0.001
Grade	0.093	0.075	151	1	<0.001
Construction zone	0.093	0.042	47	1	<0.001
Locality	0.071	0.065	118	1	<0.001
Performance error	0.055	0.025	17	1	<0.001
Traffic flow	0.036	0.039	42	1	<0.001
Surface	0.007	0.006	1	1	0.338
Alignment	0.006	0.004	1	1	0.468
Age group	-0.025	-0.042	51	2	<0.001
Lighting	-0.031	-0.028	21	1	<0.001

The importance of each feature can be illustrated in a quadrant plot as shown in Figure 10. The *x*-axis indicates the normalized mean value of each impacting factor, while the *y*-axis reflects the computed total effect, i.e., the importance of each variable with respect to the safety outcome.

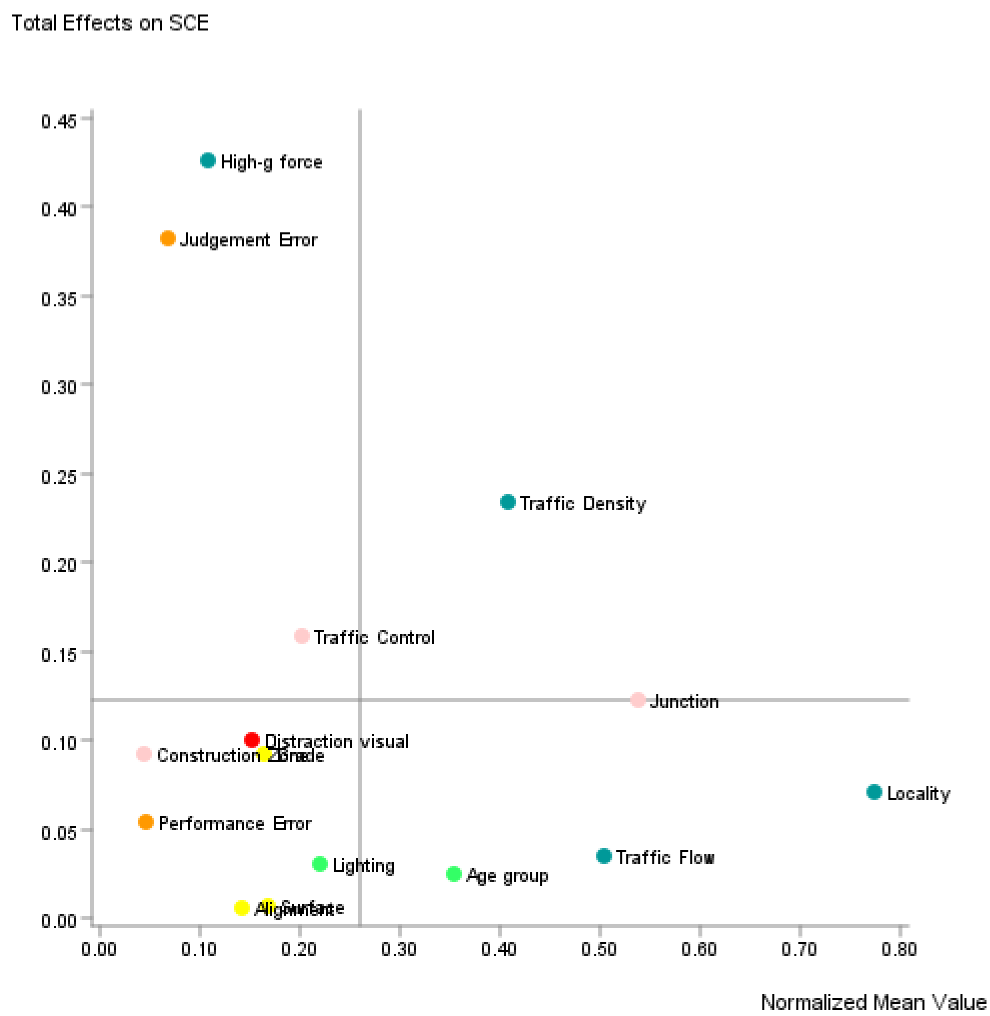


Figure 10. Graph. Feature importance by total effect.

The relationship is further explored by the response curves of the safety outcomes as a function of the values of each impacting factor, as shown in Figure 11. This illustrates not only the strength of each variable, but also the pattern over different values of each variable. For instance, high g-force showed the largest impact on SCEs, but not a linear relationship. The association of high g-force and SCE decreases as the high g-force value increases. On the other hand, there is a negative association between SCE and age group, i.e., teen/young drivers have the largest SCE risk while senior drivers have the lowest. A possible reason is that age group is the parent of judgment error, which has a direct effect on safety outcomes. Existing studies show that the experienced drivers tend to cope with hazard scenarios better than novice drivers (Mao et al., 2023; Upahita et al., 2018), which implies that senior drivers are less likely to perform judgment errors.

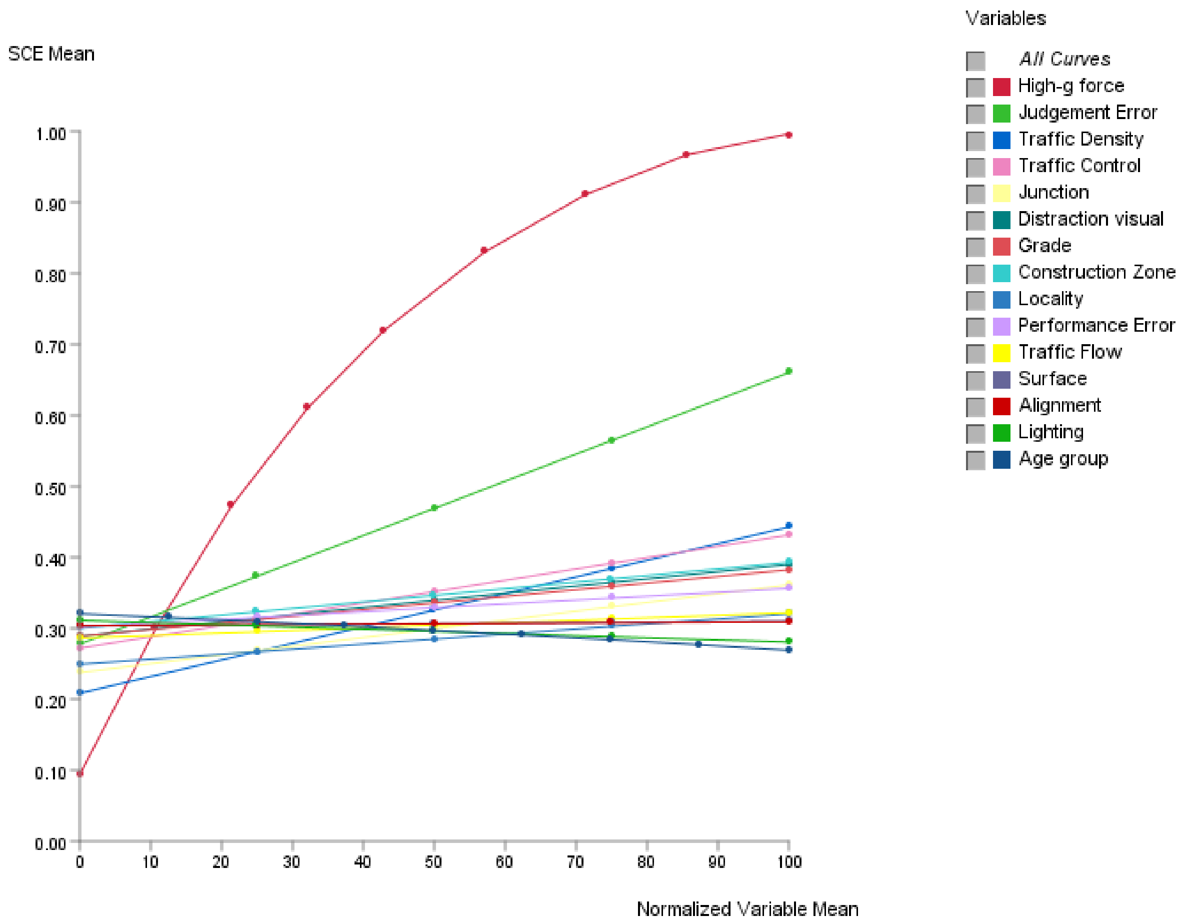


Figure 11. Graph. Response curve of impacting factors on safety outcomes.

Marginal Effect of Impacting Factors

There are two ways to quantify the crash risk of impacting factors: the marginal effect of individual factor, and the joint effect of multiple factors. The conditional probability ratio is used as a measure with the formula:

$$conditional\ probability\ ratio = \frac{P(SCE = Yes|Test\ scenario)}{P(SCE = Yes|Reference\ scenario)}$$

The rationale is that this is a case-cohort study, and the absolute probability does not reflect the actual probability of having SCEs.

For the marginal effect, the reference scenario is the low-risk scenario. For example, no judgment error is the reference scenario, while the presence of judgment error is the test scenario. The factor states in Table 1 can provide further details of the different scenarios. Results are sorted by conditional probability ratio in Table 4.

Table 4. Marginal effect of individual factor

Impacting Factor	Reference Level	Test Level	Ratio
Judgment error	27.67%	66.02%	2.39
Traffic density	20.86%	44.26%	2.12
Traffic control	27.21%	43.14%	1.59
Junction	23.77%	36.11%	1.52
Distraction visual	28.89%	38.99%	1.35
Grade	28.90%	38.23%	1.32
Construction zone	30.01%	39.29%	1.31
Locality	24.93%	32.03%	1.28
Performance error	30.17%	35.62%	1.18
Traffic flow	28.63%	32.20%	1.12
Surface	30.31%	31.02%	1.02
Alignment	30.35%	30.92%	1.02
Lighting	31.11%	28.05%	0.90

Overall, the rank of marginal effect is consistent with feature importance. Judgment error ranks highest with the largest ratio (2.39), which implies that the odds of having SCEs are 2.39 times greater in a scenario with the presence of judgment error than no occurrence of judgment error. The "crude" odds ratio (i.e., without the Bayesian network) of having SCEs under judgment error is 5.32, which is much higher. The same was found for the other four parents; the "crude" odds ratio for traffic density, traffic control, junction, and grade are 3.22, 2.17, 1.79, and 1.67, and the conditional probability ratio is lower than the "crude" odds ratio for all four. A possible reason is due to the synergistic interactions and dependencies between the variables. This phenomenon is common in complex systems such as traffic safety systems where multiple factors interact with each other to produce an outcome. Note that the Bayesian network method is very versatile in that it can quantify the effect not only of parent nodes that have direct impact on the target node, but also of non-parent nodes that have indirect impact. The conditional probability ratio of distraction (visual), performance error, construction zone, and others are listed in Table 4.

Joint Effect of Impacting Factors

The joint effect of these five parent nodes on safety outcomes has been evaluated, and high-risk scenarios are ranked by ratio, $\mathrm{P(SCE = yes|test\ scenario)/P(SCE = yes|reference\ scenario)}$, as shown in Table 5. The reference level is when the state of all five parent nodes is "No," which is highlighted in gray in the last row. The scenario with the highest ratio occurs when all five parent nodes are "Yes," i.e., $\mathrm{P(SCE = yes|Judgment\ error = yes, traffic\ density = heavy, traffic\ control = control, junction = yes, grade = nonlevel)}$ vs. $\mathrm{P(SCE = yes|Judgment\ error = no, traffic\ density = light, traffic\ control = noncontrol, junction = no, grade = level)}$. The conditional probability ratio for this scenario is 9.69 times, with conditional probability values of 93.18% and 9.62%, respectively. The four next riskiest scenarios are (1) $\mathrm{Judgment\ error = yes, traffic\ density = heavy, traffic\ control = control, junction = no, grade = nonlevel}$; (2) $\mathrm{Judgment\ error = yes, traffic\ density = heavy, traffic\ control = control, junction = yes, grade = level}$; (3) $\mathrm{Judgment\ error =}$

$\text{yes}, \text{traffic density} = \text{heavy}, \text{traffic control} = \text{noncontrol}, \text{junction} = \text{yes}, \text{grade} = \text{nonlevel};$ and (4) $\text{Judgment error} = \text{yes}, \text{traffic density} = \text{heavy}, \text{traffic control} = \text{noncontrol}, \text{junction} = \text{yes}, \text{grade} = \text{level}.$

Table 5. High SCE risk scenarios regarding five parent nodes.

Judgment Error	Traffic Density	Traffic Control	Junction	Grade	P(SCE= Yes)	Ratio
Yes	Yes	Yes	Yes	Yes	93.18%	9.69
Yes	Yes	Yes	No	Yes	92.31%	9.60
Yes	Yes	Yes	Yes	No	91.78%	9.54
Yes	Yes	No	Yes	Yes	85.53%	8.89
Yes	Yes	No	Yes	No	83.26%	8.65
Yes	No	Yes	Yes	Yes	79.17%	8.23
Yes	No	Yes	No	No	78.26%	8.14
Yes	Yes	Yes	No	No	77.78%	8.09
Yes	No	Yes	Yes	No	74.68%	7.76
Yes	Yes	No	No	No	73.88%	7.68
Yes	Yes	No	No	Yes	73.49%	7.64
No	Yes	Yes	Yes	Yes	72.16%	7.50
Yes	No	Yes	No	Yes	70.00%	7.28
No	Yes	Yes	No	Yes	64.84%	6.74
No	Yes	No	Yes	Yes	55.99%	5.82
Yes	No	No	Yes	Yes	55.17%	5.73
Yes	No	No	Yes	No	54.31%	5.65
No	Yes	Yes	Yes	No	51.96%	5.40
Yes	No	No	No	Yes	51.25%	5.33
No	Yes	No	No	Yes	48.20%	5.01
No	No	Yes	Yes	Yes	44.16%	4.59
No	Yes	Yes	No	No	43.32%	4.50
No	Yes	No	Yes	No	40.61%	4.22
Yes	No	No	No	No	38.46%	4.00
No	Yes	No	No	No	31.48%	3.27
No	No	Yes	Yes	No	28.85%	3.00
No	No	Yes	No	Yes	24.29%	2.52
No	No	No	Yes	Yes	23.49%	2.44
No	No	Yes	No	No	20.89%	2.17
No	No	No	Yes	No	20.84%	2.17
No	No	No	No	Yes	15.32%	1.59
No	No	No	No	No	9.62%	1.00

Furthermore, the inverse probability problems have been explored under one or more factors regardless of their roles in the network. Due to the numerous combinations of factors and their levels, only one example is provided in this paper: the conditional probability ratio of having

judgment error under the occurrence of SCEs and different age groups (i.e., $P(Judgement\ error = Yes|SCE = Yes, Age = Teen\ or\ Young)/$
$P(Judgement\ error = Yes|SCE = No, Age = Senior)$). The value is 13.34, which implies the odds of having judgment errors are 13.34 times for a teen/young driver when there is occurrence of SCEs compared to a senior driver when there is no occurrence of SCEs.

CHAPTER 6. SUMMARY AND CONCLUSION

The traffic system is complex, and a range of factors contribute to driving risk, such as driver demographic characteristics, driver behavior, vehicle attributes, traffic conditions, environmental factors, law enforcement, and interventions. Each factor can influence the safety of the transportation system in unique ways, and all factors interact with each other in complex ways. Understanding the complex interrelationship of factors contributing to traffic safety risks is crucial for developing effective strategies to mitigate road accidents and improve overall road safety. Traditional approaches to studying traffic safety risk factors have relied heavily on retrospective analysis of historical crash records, survey data, and simulator studies. While these methods have provided valuable insights, they often suffer from limitations, such as data biases, subjective self-reporting, and controlled environments that may not fully capture real-world driving behavior. In addition, existing studies either focused on single or very few factors to evaluate crash risk with traditional statistical models, or they have integrated multiple factors into considerations for crash prediction via machine learning, which leads to a lack of interpretability. With SHRP 2 NDS data, this study aimed to (1) explore the relationships among traffic safety risk factors using unsupervised learning models by taking a data-driven approach to uncovering patterns and relationships within large, complex datasets, and (2) develop a robust Bayesian network model that contains extensive driver personality, behavioral, contextual, and kinematic information.

We considered 24 potential risk factors, including driver demographics, driving behavior, environmental conditions, road characteristics, traffic context, vehicle kinematics within a 5-second window of each event, and cell ban policies, which can reflect risk factors from different perspectives.

With unsupervised learning methods, the pair-wise strengths of associations were assessed via Cramer's V, and all 24 factors were clustered into five groups with a hierarchical agglomerative clustering algorithm. Based on the clustering results, the interaction and joint effect of these factors were explored by using PSEM. The results were consistent between association analysis and clustering analysis; that is, factors with higher Cramer's V values tend to be grouped into the same clusters (e.g., "age" and "risk score," and different types of distractions). The tree structure reveals the relationship among five clusters. The distraction-dominated cluster (cluster 2) is related to the driver-dominated cluster (cluster 1), while the traffic-dominated cluster (cluster 3) is connected to the driver-dominated cluster (cluster 1), road-dominated cluster (cluster 4), and environment-dominated cluster (cluster 5). The joint effect on safety outcomes suggests that the distraction-dominated cluster (cluster 2), traffic-dominated cluster (cluster 3), and road-dominated cluster (cluster 4) have a direct impact on safety outcomes, while the driver-dominated cluster (cluster 1) and environment-dominated cluster (cluster 5) show indirect impact via the distraction-dominated cluster (cluster 2) and traffic-dominated cluster (cluster 3).

With the Bayesian network method, the SC was chosen by splitting the original dataset into a learning set (90%) and test set (10%) to achieve an appropriate level of network complexity. Multiple supervised and semi-supervised learning algorithms were used on both original data and perturbation data to select the most robust structure. The final model balances the complexity of the model with its ability to explain the data. Results showed that the model based on a semi-supervised (EQ+Taboo) algorithm outperformed others regarding the lowest MDL score, and the estimated network structure consists of 15 out of 24 factors from input data. Among these, five factors (judgment error, traffic density, traffic control, junction, and grade) are parent nodes of SCEs, i.e., have direct impact on safety outcomes. The rank of feature importance is consistent with parent nodes, with the exception of visual distraction, which ranks higher than grade but is not included in the developed network.

The response curves on SCE by different factors show both the strength of each variable and the pattern over different values of each variable. The association of high g-force and SCE decreases as the high g-force value increases. Note that high g-force represents the first component of standard deviation from 3-dimensional acceleration instead of raw 3-dimensional acceleration values. In addition, there is a negative association between SCE and age group, i.e., teen/young drivers have the largest SCE risk while senior drivers have the lowest. The marginal effect of each factor is consistent with feature importance, but the magnitude of the conditional probability ratio is lower than the "crude" odds ratio, the main reason being the synergistic interactions and dependencies between the variables. The joint effect of five parent nodes provides high-risk scenarios, the highest risk scenario occurs when all five parent nodes are at the "risky" level, and the conditional probability ratio is 9.69 compared to when all five parent nodes are at a "safe" level. An example of this inverse probability is that the odds of having judgment errors are 13.34 times higher for a teen/young driver when there is occurrence of SCEs compared to a senior driver when there is no occurrence of SCEs.

Despite the insightful findings, there is some future work to be considered. First, drivers' behavior is reflected via distractions, judgment errors, and performance errors. Sub-types of distractions such as visual, manual, cognitive, etc., have been considered in previous studies (Antin et al., 2023; Hanowski et al., 2017). However, judgment errors and performance errors have not been categorized into sub-types. Speeding could be substantially different from fatigued driving in relation to traffic context, and hence more details should be included to delineate risks from different aspects. Second, driving is a dynamic task, with most factors changing constantly. This study employed data information extracted by data reductionists within a relatively short window (20 seconds for crashes and near-crashes, and 6 seconds for baselines), but time-series features should be incorporated to better reflect the dynamic pattern. This is particularly promising with the development of video recognition. The time-series eye-glance data should also be included in the risk model, which reflects drivers' situation awareness level. Third, high g-force was found to be the child of the safety outcome, which suggests that high g-force could be the consequence rather than the cause of SCEs. However, this needs further investigation. Finally, most factors had two levels in this study for sample size purposes, but more detailed levels may be considered to further explore the nonlinear relationship among different levels.

Nevertheless, by applying a variety of unsupervised learning methods and developing a robust, data-driven Bayesian network with NDS data, this study identified underlying structures and associations among various risk factors via a data-driven approach; corresponding strategies can be developed regarding the identified risk factors; and crash risk can be predicted for a certain scenario with certain combinations of risk factors. Warnings/alerts can be pushed to drivers in advance with advanced driver assistance systems with these findings, which shed light on these factors' interactions and contributions to overall traffic safety.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

Anderson, G. T., Glaser, Y., & Klauer, C. S. (2024). *Evaluation of eyes off road during l2 activation on uncontrolled access roadways*.

Antin, J., Klauer, C., Han, S., & Fincannon, T. (2023). *Exploring driver distraction in adaptation to lower levels of automation: Older adult driver comparisons*. 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV), National Highway Traffic Safety Administration,

Antin, J. F., Guo, F., Fang, Y., Dingus, T. A., Hankey, J. M., & Perez, M. A. (2017). The influence of functional health on seniors' driving risk. *Journal of Transport & Health*, *6*, 237-244.

Antin, J. F., Guo, F., Fang, Y., Dingus, T. A., Perez, M. A., & Hankey, J. M. (2017). A validation of the low mileage bias using naturalistic driving study data. *Journal of Safety Research*, *63*, 115-120.

Azhar, A., Rubab, S., Khan, M. M., Bangash, Y. A., Alshehri, M. D., Illahi, F., & Bashir, A. K. (2023). Detection and prediction of traffic accidents using deep learning techniques. *Cluster Computing*, *26*(1), 477-493.

Bao, N., Carballo, A., Tsukada, M., & Takeda, K. (2023). *Personalized causal factor generalization for subjective risky scene understanding with vision transformer*. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC),

Borg, A., Bjelland, H., & Njå, O. (2014). Reflections on Bayesian Network models for road tunnel safety design: A case study from Norway. *Tunnelling and Underground Space Technology*, *43*, 300-314.

Bowen, N. K., & Guo, S. (2011). *Structural equation modeling*. Oxford University Press.

Brown, K. T. (2016). *A safety analysis of spatial phenomena about the residences of drivers involved in crashes* Clemson University].

Bui, A. T., & Jun, C.-H. (2012). Learning Bayesian network structure using Markov blanket decomposition. *Pattern Recognition Letters*, *33*(16), 2134-2140.

Chen, C., Liu, X., Chen, H.-H., Li, M., & Zhao, L. (2018). A rear-end collision risk evaluation and control scheme using a Bayesian network model. *IEEE Transactions on Intelligent Transportation Systems*, *20*(1), 264-284.

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., & Guan, H. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis & Prevention*, *80*, 76-88.

Chen, G. X., Fang, Y., Guo, F., & Hanowski, R. J. (2016). The influence of daily sleep patterns of commercial truck drivers on driving performance. *Accident Analysis & Prevention*, *91*, 55-63.

Chen, T., Wong, Y. D., Shi, X., & Wang, X. (2022). Optimized structure learning of Bayesian Network for investigating causation of vehicles' on-road crashes. *Reliability Engineering & System Safety*, *224*, 108527.

Chung, Y.-S. (2013). Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. *Accident Analysis & Prevention*, *61*, 107-118.

Conrady, S., & Jouffe, L. (2015). *Bayesian networks and BayesiaLab: A practical introduction for researchers* (Vol. 9). Bayesia USA Franklin.

De Ona, J., López, G., Mujalli, R., & Calvo, F. J. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention*, *51*, 1-10.

De Oña, J., Mujalli, R. O., & Calvo, F. J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention*, *43*(1), 402-411.

de Winkel, K. N., Christoph, M., & van Nes, N. (2024). Towards a framework of driver fitness: Operationalization and comparative risk assessment. *Transportation Research Interdisciplinary Perspectives*, *23*, 101030.

Demir, B., Demir, S., Wickens, C. M., Nooshabadi, M. H., Rodak, T., & Donmez, B. (2024). Exploring the Behaviour Change Wheel and the Theoretical Domains Framework in interventions for mobile phone driver distraction: A scoping review. *Accident Analysis & Prevention*, *195*, 107369.

Devlin, A., Candappa, N., Corben, B., & Logan, D. (2011). Designing safer roads to accommodate driver error. *Psychopharmacology*, *10*, 193-212.

Ding, H., Ghazilla, R. A. R., Singh, R. S. K., & Wei, L. (2022). Deep learning method for risk identification under multiple physiological signals and PAD model. *Microprocessors and Microsystems*, *88*, 104393.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, *113*(10), 2636-2641.

Dong, C., Shao, C., Li, J., & Xiong, Z. (2018). An improved deep learning model for traffic crash prediction. *Journal of Advanced Transportation*, *2018*, 1-13.

Dong, S., Khattak, A., Ullah, I., Zhou, J., & Hussain, A. (2022). Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley

Additive exPlanations. *International Journal of Environmental Research and Public Health*, *19*(5), 2925.

Dozio, N., Bertoni, M., & Ferrise, F. (2024). Driving emotions: Using virtual reality to explore the effect of low and high arousal on driver's attention. *Virtual Reality*, *28*(1), 1-11.

Ferreira, J., Carvalho, E., Ferreira, B. V., de Souza, C., Suhara, Y., Pentland, A., & Pessin, G. (2017). Driver behavior profiling: An investigation with different smartphone sensors and machine learning. *PloS One*, *12*(4), e0174959.

Fitch, G. M., Soccolich, S. A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., Perez, M. A., Hanowski, R. J., Hankey, J. M., & Dingus, T. A. (2013). *The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk.*

Gregoriades, A., & Mouskos, K. C. (2013). Black spots identification through a Bayesian Networks quantification of accident risk index. *Transportation Research Part C: Emerging Technologies*, *28*, 28-43.

Guo, F. (2019). Statistical methods for naturalistic driving studies. *Annual Review of Statistics and Its Application*, *6*, 309-328.

Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, *61*, 3-9.

Guo, F., Fang, Y., & Antin, J. F. (2015). Older driver fitness-to-drive evaluation using naturalistic driving data. *Journal of Safety Research*, *54*, 49. e29-54.

Guo, F., Han, S., & Hankey, J. (2022). *The Shanghai Naturalistic Driving Study (SHNDS)*. http://hdl.handle.net/10919/111703

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

Hai, D. N., Minh, C. C., & Huynh, N. (2023). Meta-analysis of driving behavior studies and assessment of factors using structural equation modeling. *International Journal of Transportation Science and Technology*.

Hamdar, S. H., Mahmassani, H. S., & Chen, R. B. (2008). Aggressiveness propensity index for driving behavior at signalized intersections. *Accident Analysis & Prevention*, *40*(1), 315-326.

Han, S., Guo, F., & Klauer, C. (2023). *In-depth analysis of crash risk associated with eyes-off-road duration by road control type and intersection type*. http://hdl.handle.net/10919/113544

Hankey, J. M., Perez, M. A., & McClafferty, J. A. (2016). *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets*. http://hdl.handle.net/10919/70850

Hanowski, R. J., Olson, R. L., Hickman, J. S., & Bocanegra, J. (2017). Driver distraction in commercial motor vehicle operations. In *Driver Distraction and Inattention* (pp. 141-156). CRC Press.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*(454), 746-774.

Hershey, J. R., & Olsen, P. A. (2007). *Approximating the Kullback Leibler divergence between Gaussian mixture models*. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07,

Hickman, J. S., Mabry, J. E., Marburg, L., Guo, F., Huiying, M., Hanowski, R. J., Whiteman, J., & Herbert, W. (2020). *Commercial Driver Safety Risk Factors (CDSRF)*. Federal Motor Carrier Safety Administration. https://rosap.ntl.bts.gov/view/dot/49620

Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention*, *45*, 373-381.

Ivers, R., Senserrick, T., Boufous, S., Stevenson, M., Chen, H.-Y., Woodward, M., & Norton, R. (2009). Novice drivers’ risky driving behavior, risk perception, and crash risk: Findings from the DRIVE study. *American Journal of Public Health*, *99*(9), 1638-1644.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., & Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International Journal of Injury Control and Safety Promotion*, *28*(4), 408-427.

Javid, M. A., Ali, N., Shah, S. A. H., & Abdullah, M. (2022). Structural equation modeling of drivers’ speeding behavior in Lahore: Importance of attitudes, personality traits, behavioral control, and traffic awareness. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, *46*(2), 1607-1619.

Jeon, M., Walker, B. N., & Yim, J.-B. (2014). Effects of specific emotions on subjective judgment, driving performance, and perceived workload. *Transportation Research Part F: Traffic Psychology and Behaviour*, *24*, 197-209.

Jiang, F., Yuen, K. K. R., & Lee, E. W. M. (2020). A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. *Accident Analysis & Prevention*, *141*, 105520.

Joo, Y.-J., Kho, S.-Y., Kim, D.-K., & Park, H.-C. (2022). A data-driven Bayesian network for probabilistic crash risk assessment of individual driver with traffic violation and crash records. *Accident Analysis & Prevention*, *176*, 106790.

Kabir, S., & Papadopoulos, Y. (2019). Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. *Safety Science*, *115*, 154-175.

Keshan, N., Parimi, P., & Bichindaritz, I. (2015). *Machine learning for stress detection from ECG signals in automobile drivers*. 2015 IEEE International Conference on Big Data (Big Data),

Khan, T., & Mohapatra, S. S. (2023). Identification of spatial and temporal dilemma zone at mid-block median openings: A gap acceptance based approach. *Transportation Research Record*, *2677*(3), 160-175.

Kim, K., Pant, P., & Yamashita, E. (2011). Measuring influence of accessibility on accident severity with structural equation modeling. *Transportation Research Record*, *2236*(1), 1-10.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2009). *Comparing real-world behaviors of drivers with high versus low rates of crashes and near crashes*.

Koohestani, A., Kebria, P., Khosravi, A., & Nahavandi, S. (2018). *Drivers performance evaluation using physiological measurement in a driving simulator*. 2018 Digital Image Computing: Techniques and Applications (DICTA),

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 47-58.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*(6), 066138.

Lalika, L., Kitali, A. E., Haule, H. J., Kidando, E., Sando, T., & Alluri, P. (2022). What are the leading causes of fatal and severe injury crashes involving older pedestrian? Evidence from Bayesian network model. *Journal of Safety Research*, *80*, 281-292.

Lee, J.-Y., Chung, J.-H., & Son, B. (2008). Analysis of traffic accident size for Korean highway using structural equation models. *Accident Analysis & Prevention*, *40*(6), 1955-1963.

Lee, J., Chae, J., Yoon, T., & Yang, H. (2018). Traffic accident severity analysis with rain-related factors using structural equation modeling–A case study of Seoul City. *Accident Analysis & Prevention*, *112*, 1-10.

Li, H., Gao, Q., Zhang, Z., Zhang, Y., & Ren, G. (2023). Spatial and temporal prediction of secondary crashes combining stacked sparse auto-encoder and long short-term memory. *Accident Analysis & Prevention*, *191*, 107205.

Li, P., Abdel-Aty, M., Cai, Q., & Islam, Z. (2020). A deep learning approach to detect real-time vehicle maneuvers based on smartphone sensors. *IEEE Transactions on Intelligent Transportation Systems*, *23*(4), 3148-3157.

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, *40*(4), 1611-1618.

Liang, D., Lau, N., & Antin, J. (2022). A structural equation model of older adults' driving exposure and avoidance using objective driving records. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1, pp. 6-7). Sage Publications.

Liu, L., Ye, X., Wang, T., Yan, X., Chen, J., & Ran, B. (2022). Key factors analysis of severity of automobile to two-wheeler traffic accidents based on Bayesian network. *International Journal of Environmental Research and Public Health*, *19*(10), 6013.

Lu, J., Xie, X., & Zhang, R. (2013). Focusing on appraisals: How and why anger and fear influence driving risk perception. *Journal of Safety Research*, *45*, 65-73.

Ma, Y., Qi, S., Fan, L., Lu, W., Chan, C.-Y., & Zhang, Y. (2019). Dynamic Bayesian network approach to evaluate vehicle driving risk based on on-road experiment driving data. *IEEE Access*, *7*, 135050-135062.

Ma, Y., Xie, Z., Chen, S., Qiao, F., & Li, Z. (2023). Real-time detection of abnormal driving behavior based on long short-term memory network and regression residuals. *Transportation Research Part C: Emerging Technologies*, *146*, 103983.

Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis & Prevention*, *160*, 106322.

Mallia, L., Lazuras, L., Violani, C., & Lucidi, F. (2015). Crash risk and aberrant driving behaviors among bus drivers: The role of personality and attitudes towards traffic safety. *Accident Analysis & Prevention*, *79*, 145-151.

Mantouka, E., Barmpounakis, E., Vlahogianni, E., & Golias, J. (2021). Smartphone sensing for understanding driving behavior: Current practice and challenges. *International Journal of Transportation Science and Technology*, *10*(3), 266-282.

Mao, Y., Wang, X., He, W., & Pan, G. (2023). Improving hazard perception for drivers based on driving styles with visual alerts. *Transportation Research Part F: Traffic Psychology and Behaviour*, *97*, 367-382.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, *23*(2), 143-149.

Mesken, J., Hagenzieker, M. P., Rothengatter, T., & De Waard, D. (2007). Frequency, determinants, and consequences of different drivers' emotions: An on-the-road study using self-reports,(observed) behaviour, and physiology. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(6), 458-475.

Mirón-Juárez, C. A., García-Hernández, C., Ochoa-Ávila, E., & Díaz-Grijalva, G. R. (2020). Approaching to a structural model of impulsivity and driving anger as predictors of risk behaviors in young drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *72*, 71-80.

Mohamed, E. A. (2014). Predicting causes of traffic road accidents using multi-class support vector machines. *Journal of Communication and Computer*, *11*(5), 441-447.

Mujalli, R. O., & De Oña, J. (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. *Journal of Safety Research*, *42*(5), 317-326.

Ogungbire, A., Kalambay, P., Gajera, H., & Pulugurtha, S. S. (2023). *Deep learning, machine learning, or statistical models for weather-related crash severity prediction*. https://scholarworks.sjsu.edu/mti_publications/480/

Osman, O. A., Hajij, M., Karbalaieali, S., & Ishak, S. (2019). A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accident Analysis & Prevention*, *123*, 274-281.

Ouimet, M. C., Brown, T. G., Guo, F., Klauer, S. G., Simons-Morton, B. G., Fang, Y., Lee, S. E., Gianoulakis, C., & Dingus, T. A. (2014). Higher crash and near-crash rates in teenaged drivers with lower cortisol response: an 18-month longitudinal, naturalistic study. *JAMA Pediatrics*, *168*(6), 517-522.

Özkan, B. (2023). *The effect of driving in unfamiliar traffic flow to simulated driving performance* [Master's thesis, Middle East Technical University].

Papantoniou, P., Vlahogianni, E. I., & Yannis, G. (2021). Are driving errors and driving performance correlated? A dual structural equation model. *Advances in Transportation Studies*, *53*, 37-50.

Papantoniou, P., Yannis, G., & Christofa, E. (2019). Which factors lead to driving errors? A structural equation model analysis through a driving simulator experiment. *IATSS Research*, *43*(1), 44-50.

Park, H.-C., Joo, Y.-J., Kho, S.-Y., Kim, D.-K., & Park, B.-J. (2019). Injury severity of bus–pedestrian crashes in South Korea considering the effects of regional and company factors. *Sustainability*, *11*(11), 3169.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peng, Y., Cheng, L., Jiang, Y., & Zhu, S. (2021). Examining Bayesian network modeling in identification of dangerous driving behavior. *PloS One*, *16*(8), e0252484.

Perez, A., Larranaga, P., & Inza, I. (2006). Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, *43*(1), 1-25.

Pettie, S., & Ramachandran, V. (2002). An optimal minimum spanning tree algorithm. *Journal of the ACM (JACM)*, *49*(1), 16-34.

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

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and psychological measurement*, *73*(1), 5-26.

Ryan, G. A., Legge, M., & Rosman, D. (1998). Age related changes in drivers' crash risk and crash type. *Accident Analysis & Prevention*, *30*(3), 379-387.

Sadia, R., Bekhor, S., & Polus, A. (2018). Structural equations modelling of drivers' speed selection using environmental, driver, and risk factors. *Accident Analysis & Prevention*, *116*, 21-29.

Sagar, S., Stamatiadis, N., Wright, S., & Cambron, A. (2020). Identifying high-risk commercial vehicle drivers using sociodemographic characteristics. *Accident Analysis & Prevention*, *143*, 105582.

Scott-Parker, B., Watson, B., King, M. J., & Hyde, M. K. (2013). A further exploration of sensation seeking propensity, reward sensitivity, depression, anxiety, and the risky behaviour of young novice drivers in a structural equation model. *Accident Analysis & Prevention*, *50*, 465-471.

Sharma, B., Katiyar, V. K., & Kumar, K. (2016). Traffic accident prediction model using support vector machines with Gaussian kernel. *Proceedings of Fifth International Conference on Soft Computing for Problem Solving: SocProS 2015, Volume 2* (pp. 1-10). Springer.

Sheehan, B., Murphy, F., Ryan, C., Mullins, M., & Liu, H. Y. (2017). Semi-autonomous vehicle motor insurance: A Bayesian Network risk transfer approach. *Transportation Research Part C: Emerging Technologies*, *82*, 124-137.

Shi, L., Qian, C., & Guo, F. (2022). Real-time driving risk assessment using deep learning with XGBoost. *Accident Analysis & Prevention*, *178*, 106836.

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

Soccolich, S., Hammond, R., Camden, M., & Walker, S. (2024). *Streamlining drowsiness assessment: An In-Depth Review of ORD and PERCLOS Methods*.

Song, X., Sun, Y., & Tao, P. (2022). A dynamic bayesian network model for real-time risk propagation of secondary rear-end collision accident using driving risk field. *IEEE Access*, *10*, 72429-72443.

Song, X., Yin, Y., Cao, H., Zhao, S., Li, M., & Yi, B. (2021). The mediating effect of driver characteristics on risky driving behaviors moderated by gender, and the classification model of driver's driving risk. *Accident Analysis & Prevention*, *153*, 106038.

Song, Y., Kou, S., & Wang, C. (2021). Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach. *Journal of Safety Research*, *76*, 64-72.

Sonth, A., Sarkar, A., Jain, S., Bhagat, H., & Doerzaph, Z. R. (2023). *Real-time risk prediction at signalized intersections using a graph neural network*. https://hdl.handle.net/10919/116786

Stavrinos, D., McManus, B., Mrug, S., Anthony, T., Underhill, A. T., & Pawar, P. (2023). Behind the wheel: Unraveling the impact of experience over age over the first 18 months of licensure. *Journal of Pediatric Psychology*, *48*(12), 1030-1037.

Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, *54*, 176-186.

Sun, M., Zhou, R., Jiao, C., & Sun, X. (2022). Severity analysis of hazardous material road transportation crashes with a Bayesian network using Highway Safety Information System data. *International Journal of Environmental Research and Public Health*, *19*(7), 4002.

Swain, T. A. (2023). *The impact of diabetes on at-fault motor vehicle collisions in older adults: Visual function, medication, and crash characteristics* [Doctoral dissertation, The University of Alabama at Birmingham].

Tang, L., Yue, L., Fu, A., Zhou, X., Sun, J., & Li, Z. (2023). *Interactive vehicle trajectory prediction considering abnormal emotion based on SOR cognitive framework*. 2023 7th CAA International Conference on Vehicular Control and Intelligence (CVCI) (pp. 1-7). IEEE.

Tapia, J. L., Duñabeitia, J. A., & Borda, D. S. (2024). *The effects of cognitive training on driving performance*.

Thompson, J., & Wundersitz, L. (2024). Mobile phone detection cameras in South Australia. *Prevention*, *119*, 225-236.

Truelove, V., Freeman, J., & Davey, J. (2019). "I Snapchat and Drive!" A mixed methods approach examining snapchat use while driving and deterrent perceptions among young adults. *Accident Analysis & Prevention*, *131*, 146-156.

Truelove, V., Oviedo-Trespalacios, O., Freeman, J., & Davey, J. (2021). Sanctions or crashes? A mixed-method study of factors influencing general and concealed mobile phone use while driving. *Safety Science*, *135*, 105119.

Truelove, V., Stefanidis, K., Mills, L., & Oviedo-Trespalacios, O. (2023). Police and public perspectives on the use and impacts of technology that expose enforcement locations for phone use while driving. *Safety Science*, *164*, 106155.

Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N., & Madni, H. A. (2020). Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction. *arXiv preprint arXiv:2010.14921*.

Upahita, D. P., Wong, Y. D., & Lum, K. M. (2018). Effect of driving experience and driving inactivity on young driver's hazard mitigation skills. *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 286-297.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373-440.

Van Schagen, I., Welsh, R., Backer-Grondahl, A., Hoedemaeker, M., Lotan, T., Morris, A., Sagberg, F., & Winkelbauer, M. (2011). Towards a large-scale European Naturalistic Driving study: Final report of PROLOGUE. *SWOV Institute for Road Safety Research, Leidschendam, Netherlands*.

Wang, C., Xu, C., & Dai, Y. (2019). A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. *Accident Analysis & Prevention*, *123*, 365-373.

Wang, X., & Xu, X. (2019). Assessing the relationship between self-reported driving behaviors and driver risk using a naturalistic driving study. *Accident Analysis & Prevention*, *128*, 8-16.

Wong, I. Y., Smith, S. S., & Sullivan, K. A. (2018). Validating an older adult driving behaviour model with structural equation modelling and confirmatory factor analysis. *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 495-504.

Xie, D.-F., Fang, Z.-Z., Jia, B., & He, Z. (2019). A data-driven lane-changing model based on deep learning. *Transportation Research Part C: Emerging Technologies*, *106*, 41-60.

Xu, J., Qian, C., Han, S., & Guo, F. (2024). Detecting critical mismatched driver visual attention during lane change: An embedding kernel algorithm. *IEEE Transactions on Intelligent Transportation Systems*.

Xu, W., Wang, J., Fu, T., Sobhani, A., & Nabavi Niaki, M. (2022). Investigating contributing factors on aggressive driving based on a structural equation model. *Journal of Advanced Transportation*, *2022*.

Yang, D., Xie, K., Ozbay, K., Yang, H., & Budnick, N. (2019). Modeling of time-dependent safety performance using anonymized and aggregated smartphone-based dangerous driving event data. *Accident Analysis & Prevention*, *132*, 105286.

Yuan, J., Abdel-Aty, M., Gong, Y., & Cai, Q. (2019). Real-time crash risk prediction using long short-term memory recurrent neural network. *Transportation Research Record*, *2673*(4), 314-326.

Zhang, G., Yau, K. K., & Chen, G. (2013). Risk factors associated with traffic violations and accident severity in China. *Accident Analysis & Prevention*, *59*, 18-25.

Zhang, R., Wen, X., Cao, H., Cui, P., Chai, H., Hu, R., & Yu, R. (2023). Critical safety management driver identification based upon temporal variation characteristics of driving behavior. *Accident Analysis & Prevention*, *193*, 107307.

Zhang, R., Wen, X., Cao, H., Cui, P., Chai, H., Hu, R., & Yu, R. (2024). High-risk event prone driver identification considering driving behavior temporal covariate shift. *Accident Analysis & Prevention*, *199*, 107526.

Zhang, T., Chan, A. H., Ba, Y., & Zhang, W. (2016). Situational driving anger, driving performance and allocation of visual attention. *Transportation Research Part F: Traffic Psychology and Behaviour*, *42*, 376-388.

Zhang, T., Chan, A. H., Xue, H., Zhang, X., & Tao, D. (2019). Driving anger, aberrant driving behaviors, and road crash risk: Testing of a mediated model. *International Journal of Environmental Research and Public Health*, *16*(3), 297.

Zhao, L., Wang, X., & Qian, Y. (2012). Analysis of factors that influence hazardous material transportation accidents based on Bayesian networks: A case study in China. *Safety Science*, *50*(4), 1049-1055.

Zhao, X., Xu, W., Ma, J., Li, H., & Chen, Y. (2019). An analysis of the relationship between driver characteristics and driving safety using structural equation models. *Transportation Research Part F: Traffic Psychology and Behaviour*, *62*, 529-545.

Zheng, Z., Lu, P., & Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. *Journal of Safety Research*, *65*, 115-124.

Zhu, X., Yuan, Y., Hu, X., Chiu, Y.-C., & Ma, Y.-L. (2017). A Bayesian Network model for contextual versus non-contextual driving behavior assessment. *Transportation Research Part C: Emerging Technologies*, *81*, 172-187.

Zhu, X. J. (2005). *Semi-supervised learning literature survey*. http://digital.library.wisc.edu/1793/60444

Zimasa, T., Jamson, S., & Henson, B. (2019). The influence of driver's mood on car following and glance behaviour: Using cognitive load as an intervention. *Transportation Research Part F: Traffic Psychology and Behaviour*, *66*, 87-100.

Zong, F., Chen, X., Tang, J., Yu, P., & Wu, T. (2019). Analyzing traffic crash severity with combination of information entropy and Bayesian network. *IEEE Access*, *7*, 63288-63302.

Zywiec, W. J., Mazzuchi, T. A., & Sarkani, S. (2021). Analysis of process criticality accident risk using a metamodel-driven Bayesian network. *Reliability Engineering & System Safety*, *207*, 107322.