Project Report

Proactive Traffic Incident Management in Alabama

ALABAMA TRANSPORTATION INSTITUTE

THE UNIVERSITY OF ALABAMA

2024

Proactive Traffic Incident Management in Alabama

A Report to

The Alabama Department of Transportation
1409 Coliseum Boulevard
Montgomery, AL 36110

by

Jun Liu, Ph.D.
Zihe Zhang, Ph.D.
Alex Hainen, Ph.D.
Steven Jones, Ph.D.
Tim Barnett, P.E., PTOE, RSP2i
Steve Burdette

Alabama Transportation Institute
(Center of Transportation, Operations, Planning, and Safety)
The University of Alabama
28 Kirkbride Ln, Tuscaloosa, AL 35401

August 2024

Acknowledgement

This research project, numbered 931-026, was made possible through the financial support of the Alabama Department of Transportation's Research Advisory Committee (RAC). We also extend our gratitude to the Project Advisory Committee members for their invaluable support, including Mr. Stacey Glass, Mr. Kerry NeSmith, Mr. DeJarvis Leonard, Mr. Brad Lindsey, Mr. Virgil Clifton, and Mrs. Kristy Harris. The viewpoints expressed in this report are those of the authors and do not necessarily reflect the official views or policies of the Alabama Department of Transportation.

Summary

The recent availability of the state-wide large-scale real-time traffic database, traffic incident database, and road crash database has enabled more efficient traffic incident management. This project aims to assist traffic incident management (TIM) in Alabama by leveraging three large-scale real-time databases maintained by ALDOT including the ALGO traffic database, CARE crash database, and HERE traffic database. Specially, by integrating and linking the information provided in the three databases, this project focuses on the four main objectives: (1) predicting the occurrence of crashes based on real-time traffic characteristics; (2) detecting the occurrence of the incidents; (3) evaluating the impact of traffic incident and identifying the key associated factors; (4) Identifying and mapping the high-frequency incident-crash segment of interstate in the state of Alabama. Additionally, this project explores the potential of using HERE speed data to improve the crash severity modeling.

An automatic speed matrix extraction tool was developed using Python to extract the speed in 0.1 miles×1-minute resolution in the predefined space and time coverage. The speed matrix was created for every incident/crash stored in the ALGO database and CARE database. Machine learning models including Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) were developed and compared. Separate models were estimated for three major crash types: single-vehicle, rear-end, and sideswipe crashes. The model prediction accuracy indicated that the RF models outperform other models. Models for rear-end crashes are found to have greater accuracy than other models, which implies that rear-end crashes have a significant relationship with pre-crash traffic dynamics and are more predictable. The traffic speed factors that are ranked high in terms of feature importance are the speed variance and speed reduction before crashes. According to partial dependence plots, the rear-end crash risk is positively related to the speed variance and speed reductions. More results are discussed in the paper. Regarding automatic incident detection (AID), three types of AID algorithms are built based on image-like spatial-temporal speed matrices extracted from HERE using Artificial Neural Network (ANN) and Convolution Neural Network (CNN). The results show that incidents (e.g., crashes and vehicle fires) that have significant impacts on traffic can be detected and classified with high confidence by the algorithms developed in this project. Regarding the incident impact modeling, this project created and leveraged three metrics including the maximum queue length, time at the maximum queue length, and volume (spatiotemporal extent of a queue) to measure the traffic impacts. To reduce the estimation bias from any single model, this project applies five machine learning models to interpret nonlinear relationships by averaging the marginal effects on estimating the spatiotemporal impacts of traffic incidents considering traffic flow speed data, traffic incident features, and road environment characteristics. The model results reveal that traffic flow speed dynamics are strongly associated with spatiotemporal impacts of traffic incidents. The findings of this project provide transportation agencies with practical models, and better-informed planning and operational decisions and strategies needed to manage traffic incidents more proactively and efficiently.

Key Words: Proactive Incident Management, Traffic Data, Crash Risk Prediction, Automatic Incident Detection, Interpretable Machine Learning, Deep Learning, Incident Impact.

This page is deliberately left blank.

Table of Contents

List of Figures

List of Tables

1. Introduction

Traffic incidents, including crashes and disabled vehicles, have been identified as a major contributor to increased congestion accounting for account for approximately one-fourth of all traffic delays (FHWA, 2010; ALDOT, 2014). The INRIX 2018 Global Traffic Scorecard shows that Americans lost an average of 97 hours a year due to congestion, costing them nearly $87 billion in 2018, an average of $1,348 per driver (INRIX, 2019). In addition to the mobility impacts, traffic incidents also significantly affect the safety of both motorists and emergency responders by increasing the chances of secondary crashes or incidents. Secondary crashes can make the traffic congestion worse and make it even more difficult for responders to get to and from the incident scene.

Traffic Incident Management (TIM) is a planned and coordinated multidisciplinary process to detect, respond to, and clear traffic incidents so that traffic flow may be restored as safely and quickly as possible (FHWA, 2010). The TIM Program currently administered by the Alabama Department of Transportation (ALDOT) brings together agencies from local, regional, and state transportation and public safety communities to make Alabama highways safer for both incident responders and motorists by reducing the time needed to reopen travel lanes and get traffic moving again (ALDOT, 2014). With effective TIM, incidents can be cleared safely in less time, minimizing congestion and the impacts of traffic incidents on overall mobility and safety.

Effective TIM requires fast and accurate incident detection that involves both the collection and analysis of traffic data obtained from detectors such as inductive loops, video-based technologies, and global position system (GPS) based vehicle tracking systems (Ren et al., 2016; D'Andrea et al., 2017; Sun et al., 2018). Most current TIM responses are reactive because information from these detectors is used to alert transportation managers and incident responders to the occurrence of incidents or crashes. As such, reactive TIM actions may result in unnecessary delays and increase the chances of secondary incidents and crashes. To minimize response delays, proactive TIM operations are needed. The goal of this project is to develop data-enabled tools to support proactive TIM operations in Alabama. Specially, this project takes advantage of three state-wise databases mentioned by ALDOT including the CARE crash database, HERE traffic information database, and ALGO incident database to create tools and models to (1) predict the risk of occurring specific types of incidents based on real-time traffic characteristics; (2) detecting the occurrence of incidents based on spatial-temporal flow information; (3) evaluate the spatial impact, temporal impact, and spatial-temporal impact of traffic incidents on traffic flow and identifying the associated factors. The outcome of this project provides a batch of tools (e.g., traffic characteristics extraction tool, incident risk prediction tool, etc.) and suggest assisting the TIM in the state of Alabama.

2. Review of Related Work

Traffic Incident Management (TIM) is a process of planning and coordination to detect, respond to, and clear traffic incidents, and restore roadway traffic flow as safely and quickly as possible. The TIM is a coordinated effort that brings together agencies from local, regional, and state transportation and public safety communities (FHWA, 2020). Effective TIM could reduce the duration and impact of traffic incidents, and improve the safety of drivers, crash victims, and emergency responders.

This project's literature review summarizes some state or regional TIM practices focusing on the relevant programs, systems, and technologies. Then the second part of this literature review is separated into three subsections to summarize the recent scholarly research on (1) crash risk prediction; (2) Incident detection; and (3) Incident impact analysis. This part focuses on the data source, data processing, model development, and performance.

2.1 Statewide or Regional Practices

In summary, most current TIM practices are a reactive process, as they rely on the information that is already generated from various sources, including 911 calls, or observed traffic patterns (e.g., speed drops). Then the transportation managers and incident responders are alerted about the occurrence of an incident or crash. Reactive TIM actions may result in unnecessary delays and increase the chances of secondary incidents and crashes.

TABLE 1 summarizes the selected traffic incident management (TIM) practices at the state or region level with deployed techniques, platforms, and datasets. The New York City TIM program if formed by the New York State Department of Transportation (NYSDOT). In NYC, there are two important systems, Highway Emergency Local Patrol (HELP) and Citywide Incident Management System (CIMS), play important roles in the NYC TIM program given the unique incident management requirements in NYC (NCHRPTIMPM, 2020; NYC Gov, 2020). The HELP program provides most of the incident information to the traffic management centers (TMCs). These data could then be cross-referenced to the NYSDOT operator logs. The same data could be accessed by various agencies for further analysis (NCHRPTIMPM, 2020). The Citywide Incident Management System (CIMS) establishes roles and responsibilities for the New York City TIM program to perform and support emergency response (NYC Gov, 2020).

TABLE 1 State or Region-Level Practices

State/Region	Coverages	Technologies
New York City, NY	New York City, NY	Highway Emergency Local Patrol (HELP) program, Citywide Incident Management System (CIMS)
Los Angeles, CA	Los Angeles, CA	The Automated Traffic Surveillance and Control (ATSAC) Center, Los Angeles Regional Transportation Management Center (LARTMC)
Nevada	Northern Nevada, Southern Nevada, and rural TIM coalitions	911 calls, CCTV Cameras, Freeway message signs

Florida	Seven FDOT districts and the Florida Turnpike Enterprise	911 calls, TMCs' cameras and sensors, Computer-aided dispatch (CAD) system, Public-facing mobile roadway navigation applications
Delaware Valley Regional Planning Commission	Bucks, Chester, Delaware, Montgomery, and Philadelphia in Pennsylvania; and Burlington, Camden, Gloucester, and Mercer in New Jersey	Interactive Detour Route Mapping (IDRuM), Regional Integrated Multi-Modal Information Sharing (RIMIS)
Wisconsin	State of Wisconsin	CCTV Cameras, Dynamic messaging signs (DMS), Broadcast, WisDOT 511 Mobile App, Twitter, Computer-aided dispatch (CAD) system
Alabama	State of Alabama	ALGO traffic platform, HERE mobility data, HPMS volume data, RTMC incident data

The Los Angeles TIM program is led by the California Department of Transportation (Caltrans) to clear debris and vehicles, keep traffic flowing, and help motorists by cooperating with police, firefighters, and other partners (Caltrans, 2020). The Los Angeles TIM program consists of two centers, the Automated Traffic Surveillance and Control (ATSAC) Center and the Los Angeles Regional Transportation Management Center (LARTMC) (LADOT, 2020; Miyamoto, 2020). The ATSAC center uses real-time detector loops between and at intersections to make signal timing changes according to traffic conditions. The center staff is provided with graphical visualization of the real-time traffic condition. Traffic incident notifications can be automatically generated when the traffic condition becomes abnormal (e.g., traffic accident, police, or fire emergency, etc.). The center staff verified the situation by checking the traffic situation using cameras installed at major intersections (LADOT, 2020). The LARTMC is a high-tech facility dedicated to managing traffic on highly congested roads in the Los Angeles area. The LARTMC supports joint operations and serves as the center of intelligent transportation systems (ITS) and emergency response operations (Miyamoto, 2020).

The Nevada TIM coalition covers Northern Nevada, Southern Nevada, and rural TIM coalitions. Nevada TIM Coalition takes advantage of 911 calls from travelers, and CCTV cameras videos to quickly detect highway incidents and provide timely responses to incidents (Nevada TIM Coalition, 2020a; Nevada TIM Coalition, 2020b).

The Florida TIM program covers seven Florida Department of Transportation (FDOT) Districts and the Florida Turnpike system. In addition to phone calls from travelers to notify law enforcement agencies, the Florida TIM program makes use of CCTV cameras and various electric sensors to capture the traffic flow changes (such as speed drops) due to traffic incidents. Besides, the Florida TIM program has deployed two emerging techniques, the computer-aided dispatch (CAD) system and public-facing mobile roadway navigation applications (Florida TIM Responder, 2020a; Florida TIM Responder, 2020b), to facilitate their traffic incident management. It is possible that public safety agencies may notice an incident in the network before the traffic

management centers (TMCs) (e.g., a phone call to the 911 center). The CAD system was designed to ensure timely communications between public safety agencies and TMCs to implement fast responses from TMCs and bolster roadway clearance (Florida TIM Responder, 2020a). Besides, the Florida TIM program initiated a pilot project to test the use of traveler information platforms such as Waze to facilitate their responses to incidents by connecting the platforms to their incident response vehicles (IRVs). Once an incident occurs, IRVs could receive notifications very quickly from Waze; and then IRVs display corresponding warning messages on its arrow board to alert drivers around the active incident scene (FIGURE 1). This initiative is an example of Infrastructure-to-Vehicle (I2V) communication to improve safety for roadway drivers and incident responders (Florida TIM Responder, 2020b).

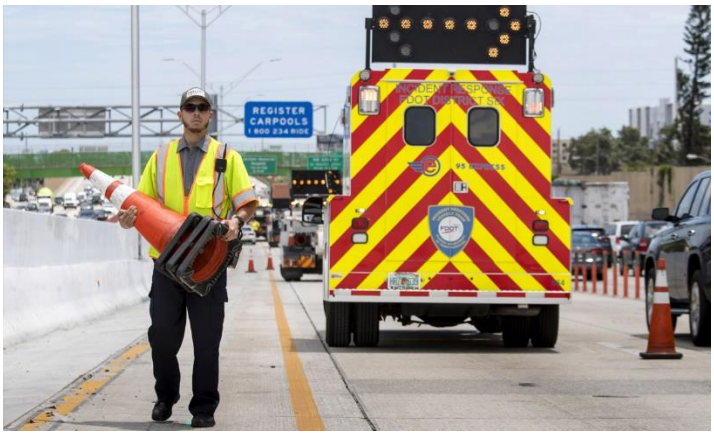


FIGURE 1 Incident Response Vehicles in Florida (Florida TIM Responder, 2020b)

Delaware Valley Regional Planning Commission (DVRPC) covers a diverse nine-county region in Pennsylvania and New Jersey. The DVRPC has developed two applications Regional Integrated Multi-Modal Information Sharing (RIMIS) and Interactive Detour Route Mapping (IDRuM), to improve the efficiency of TIM (Delaware Valley Regional Planning Commission, 2020a; Delaware Valley Regional Planning Commission, 2020b). RIMIS is a web-based application connecting highway operation centers, TMCs, and 911 call centers in the region to enable agencies to receive real-time incidents information. The website consists of transportation system databases (e.g., roadway networks), situational information, and real-time traffic videos (Delaware Valley Regional Planning Commission, 2020a). In addition, IDRuM is a web-based Interactive Detour Route Mapping application to present responders with real-time detour route maps for all limited-access highways (Delaware Valley Regional Planning Commission, 2020b). These two applications are to foster communications and information-sharing between agencies in the region.

The Wisconsin Traffic Incident Management Enhancement (TIME) led by WisDOT initialed to cover the whole state with fostering effective TIM. The Wisconsin TIME has used CCTV cameras and the CAD system to provide fast detection and responses to incidents. Notably, The Wisconsin TIME has used the 511-mobile app and Twitter platform to allow travelers to report roadway incidents. Besides, the 511-mobile app, Twitter platform, broadcast, and dynamic messaging signs (DMS) are used to inform nearby drivers with timely incident information and roadway conditions around incidents (WisDOT TIME, 2017).

In terms of the TIM program in Alabama, The ALDOT Traffic Accident Management (TIM) program brings together all agencies involved in eliminating road accidents. They work together to provide safer incident management for responders and motorists and reduce the delay time and make traffic moving again (Alabama Traffic Incident Management, 2020). ALDOT recognizes the importance of TIM in maintaining the operational safety and efficiency of roads in the state. The TIM program is a comprehensive, multi-agency, multi-disciplinary plan led by the Alabama Department of Transportation and the Alabama Department of Public Safety, focusing on the goals of the three National United Goal (NUG), which are Responder safety; Safe quick clearance; Prompt, reliable incident communications (Alabama Traffic Incident Management, 2020).

Alabama DOT (ALDOT) established the Regional Traffic Management Centers (RTMCs) with monitoring the traffic data and information using the ALGO website (Alabama Traffic Incident Management, 2020; GS&P, 2014). The ALGO website provides real-time information on cameras, speed sensors, and information on related infrastructure. Besides, HERE mobility data purchased by ALDOT provides state-wide real-time traffic flow speed information.

In summary, the state-level or region-level TIM programs are multi-agency collaborated under the management of advanced traffic management centers (TMCs). They are often equipped with 911 call centers and CCTV cameras and videos. Some states/regions have deployed the emerging computer-aided dispatch (CAD) systems to improve regular 911 call centers' response time. Other techniques like interactive mobile apps and web-based information sharing centers to support multi-agencies' responses to incidents. Dynamic messaging signs (DMS), broadcast, and social media platforms are often used to inform nearby drivers and travelers with real-time traffic incident updates.

2.2 Scholarly Research

2.2.1 Crash Risk Prediction

The studies that focused on crash risk prediction have generated significant insights to support proactive traffic incident management. Some earlier studies relied on static data (e.g., average daily traffic, land use type, road geometry characteristics) to identify sites with high crash risks at an aggregate level, e.g., the crash count at a segment with limited sight distance (Abdel-Aty & Radwan, 2000; Khattak et al., 2010). In recent years, researchers have taken advantage of traffic data generated by electronic sensors such as loop detectors and mobile devices and have developed a variety of models to predict the crash risk in a real-time or near-real-time manner. TABLE 2 summarizes selected studies that used pre-crash traffic dynamic information to develop crash risk prediction models.

TABLE 2 Selected studies on crash risk prediction

Authors	**Year**	**Study area**	**Data source**	**Model**	**Key independent variables or features (traffic dynamics)**	**Dependent variable or target features**
Hossain &	2012	11.9 km Shibuya 3	Loop detector	Bayesian	Congestion index, speed,	Crash risk

Muromachi (2012)		and 13.5 km Shinjuku 4 expressways, Tokyo		network (model)	and detector occupancy	
Qu et al. (2012)	2012	9.3-mile I-894, Milwaukee, Wisconsin	Loop detector	Support vector machine	Speed, occupancy, volume	Rear-end crash risk
Xu et al. (2013)	2013	29-mile I-880, San Francisco, California	Loop detector	Sequential logit model	Vehicle count, speed, detector occupancy	Crash risk at different severity levels
Yu & Abdel-Aty (2013)	2013	15-mile I-70, Colorado	Remote Traffic Microwave Sensor	Support vector machine	Speed, occupancy, volume	Crash risk
Qu et al. (2011)	2013	9.3-mile I - 894 Milwaukee, Wisconsin	Loop detector	Support vector machine	Traffic state and variances between adjacent lanes	Side-wipe crash risk
Wang et al. (2015)	2015	22-mile SR 408, Central Florida	Microwave Vehicle Detection System (MVDS) detector	Multilevel Bayesian logistic regression model	Speed, occupancy, volume	Crash risk for weaving segments
Park & Haghani (2016)	2016	51-mile I-695	Probe vehicle data	Neural network model	Speed	Secondary incident occurrences risk
Wu et al. (2018a)	2018	7-mile I-75, 11-mile I-4, Tampa and 11-mile SR-408, Florida	Loop and radar detectors; Microwave Vehicle Detection System (MVDS) sensors	Binary logistic regression	Speed, occupancy, volume	Crash risk
Wu et al., (2018b)	2018	Segment of I-4 in Florida	Remote Traffic Microwave Sensor (RTMS)	Random parameters logistic regression	Speed, volume	Rear-end crash risk
Cai et al. (2020)	2020	24-mile SR 408 in Central Florida	Microwave Vehicle Detection System (MVDS) detector	Convolutional Neural Network (CNN)	Speed, occupancy, volume	Crash risk
Huang et al. (2020)	2020	13.78-mile I-235, Des Moines	Roadside radar sensors	Support vector machine	Speed, occupancy, volume	Crash risk

A large portion of relevant studies relied on the traffic loop detector data (Qu et al., 2012; Qu et al., 2011; Hossain & Muromachi, 2012; Xu et al., 2013). Loop detectors play an essential role in traffic management and have been widely installed at intersections and on freeways to monitor the traffic (FHWA, 2017). The loop detector data can be used to generate common traffic flow characteristics such as volume, speed, and density. The loop detector data are readily available for researchers to explore the relationships between crash risk and traffic flow characteristics. However, such data are limited to sites with loop detectors and do not cover the entire road network. Some studies used sensor data from Microwave Vehicle Detection System (MVDS) (Cai et al., 2020; Wu et al., 2018a), Automated Vehicle Identification (AVI) (Basso et al., 2018), and Bluetooth Detectors (Hossain et al., 2019). Like loop detectors, these sensors are fixed on the road or roadside, and the data from these sensors have the same coverage limitation. Probe vehicle data are not limited to specific road segments or areas. Researchers started to use probe vehicles to understand the relationship between crash risk and traffic flow dynamics. The focus of the studies using probe vehicle data is primarily on the risk of secondary crashes instead the initial or primary crashes (Park & Haghani, 2016).

From the model perspective, researchers have developed a variety of regression-based and also machine learning models (Hossain et al., 2019). The most commonly used regression modeling approach is the logistic model, including binary logistic regression (Wu et al., 2018a), sequential logistic model (Xu et al., 2013), and random parameters logistic regression (Wu et al., 2018b). Among the machine learning models, the Support Vector Machine (SVM) has been frequently used by researchers, and the SVM models produced decent accuracies (Qu et al., 2011). Recently, more sophisticated modeling approaches were introduced for crash risk prediction. For example, Cai et al. (2020) used a convolutional neural network model to predict crash risk based on a deep convolutional generative adversarial network (DCGAN) and achieved good performance. In most of these studies, the crash types were not discussed, and it is likely that their models were to predict the risk of all crashes. Assuming that crashes of different types could have different relationships with the traffic dynamics, some studies built separate models for specific crash types such as rear-end crashes (Qu et al., 2012, Wu et al., 2018b) and sideswipe crashes (Qu et al., 2011).

To the best of the authors' knowledge, crowdsourced probe vehicle data have not been extensively explored by researchers regarding crash risk prediction. Using such data could overcome the limitation of sensor data that covers only road segments where fixed detection units are installed. Besides, models using crowdsourced data could uncover the relationships between traffic dynamics and crash risk over a diverse road or land use environment, and such models may have greater applicability. Using statewide crowdsourced probe vehicle data, the team attempts to develop machine learning models to predict crash risk for the entire freeway network in Alabama. Further, previous studies focused on improving the model performance, especially when the machine learning models are adopted. To improve the transportation systems by reducing the crashes, it is essential to interpret the modeling results and translated models into actionable items such as countermeasures to address the issues at some locations where high crash risks are identified.

2.2.2 Incident Detection

Automatic incident detection (AID) system is an integral part of a TIM system. An AID system consists of two main components: a traffic detection system and an automatic incident detection algorithm. Traffic detection systems provide real-time traffic information such as traffic count/volume, speed, and travel time. Automatic incident detection algorithms are used to identify the traffic anomaly due to the occurrence of the traffic incident. Researchers in academia focus on improving the accuracy of different AID algorithms using various AID data sources. Generally, from the methodology perspective, these methods can be classified as comparative algorithms, statistical algorithms, machine learning algorithms, and deep learning algorithms.

1) Comparative algorithm

Comparative algorithms were widely used in the early stage of traffic incident detection due to the simple algorithm and lower computation requirement. In comparative algorithms, the occurrence of the incident is detected when abnormalities of the traffic flow parameters occur. The abnormalities are identified by comparing the pre-established thresholds with the measured traffic flow parameters. These traffic flow parameters (e.g., speed, occupancy rate, and volume) are often collected from the fixed detectors equipped on adjacent lanes. Two popular comparative algorithms are the California algorithm (Karim et al., 2002) and the McMaster algorithm (Hall, 1993). The drawback of the comparative algorithm is limited detection accuracy and power (Samant, 2000).

2) Statistical algorithms

The different incident detection algorithm is also based on a predefined threshold, but the difference from the comparative algorithms is that statistical models are used to predict the traffic parameters based on the existing traffic parameters. The logic behind this kind of model is that the real value measured by the detector is compared with the predicted value forecasted by the statistical models. An incident is identified when the difference between the actual value and the predicted value is greater than a predefined threshold. For example, Ahmed and Cook (1982). applied autoregressive integrated moving average (ARIMA) to predict the short-term forecast of traffic data, and an incident is detected if the observed occupancy value lies outside the confidence limits of the corresponding point forecast (Ahmed & Cook, 1992). Hawas and Ahmed (2016) used a logit model, coupled with some predefined threshold value, to estimate the probability of incident status at different analysis time steps. A drawback of this method is that it is hard to consider the spatial and temporal variations simultaneously in the model construction.

3) Machine learning algorithms

The essence of incident detection is to recognize anomalous traffic patterns, and it can also be regarded as a classification problem in machine learning models. Machine learning models have been widely used in traffic incident detection these years. The commonly used machine learning models for incident detection include support vector machine (SVM), Random Forest (RF), and their extended models. For example, Yuan & Cheu (2003) developed and tested three different SVMs using data from the I-880 freeway in California. Xiao & Liu (2012) applied the multiple-kernel learning support vector machine (MKL-SVM) in traffic incident detection. The results show that the MKL-SVM avoided the burden of choosing the appropriate kernel function and parameters and achieved better prediction accuracy than the SVM ensemble.

4) Deep learning algorithms

Deep Learning may not require feature extraction manually and can directly take the spatial-temporal traffic pattern variations as input. Recently, a researcher has introduced deep learning algorithms to improve traffic incident detection performance. Specifically, deep learning algorithms contribute to the current state-of-the-art of incident detection studies from two aspects: (a) generative adversarial network (GAN) is used to tackle the data imbalanced issue by expanding the sample size and balance datasets (Li et al., 2020); (b) convolutional neural network (CNN) is used to capture the spatial-temporal variation of the traffic flow parameter and identify the traffic pattern (e.g., incident and non-incident) with high accuracy. The results show that deep learning algorithms such as CNN could achieve relatively high detection accuracy compared to the machine learning model (Huang et al., 2020). However, from the implication aspect, a drawback of deep learning is that it requires high-performance GPUs and lots of data. Practitioners may select the proper AID algorithm based on their conditions.

2.2.3 Incident Impact

Existing studies covering research topics of traffic incidents-related spatiotemporal impacts concentrate on the following three aspects: 1) identifying secondary crashes (Vlahogianni et al., 2010; Zhang and Khattak, 2011; Chung, 2013; Chen et al., 2016; Kitali et al., 2019); 2) determining incident impact areas (Chung and Recker, 2013; Wang et al., 2018; Ou et al., 2019; Zheng et al., 2021); 3) predicting and forecasting the incident impact (Miller and Gupta, 2012; Pan et al., 2015; Liu et al., 2017; Huang et al., 2020). From the perspective of primary and/or secondary crashes, Vlahogianni et al. (2010) developed a Bayesian network to identify the occurrence and estimate the influence of secondary crashes. Results indicate that traffic conditions at the time of an incident and the time needed to respond to and clear the crash scene are the most significant determinants in defining the upstream influence area of a crash. Zhang and Khattak (2011) focused on analyzing time gaps and distances for secondary incidents in the same direction using 2008 incident data in Virginia. Chung (2013) developed a method to define the spatiotemporally different boundaries varying with different crash types.

It should be noted that there remain significant gaps in the current understanding of how traffic incidents' spatiotemporal impacts are correlated with incident-related attributes and speed info. Specifically, two major research gaps the team aims to fill are shown as follows: 1) There is a limited number of works connecting incident-related factors with spatiotemporal impacts; 2) The existing research mainly used loop detector data to cover relatively small urban areas. The team aims to bridge the gaps by taking advantage of a large-scale high-resolution crowdsourced probe vehicle containing minute-by-minute updated traffic flow speed data covering statewide freeways in the state of Alabama. The team trains five machine learning models, including Categorical Naive Bayes (CNB), Support vector machine (SVM), Random Forest (RF), AdaBoost (Boost), and Neural network (NN) to untangle the relationship between contributing factors and measurements of traffic incidents' spatiotemporal impacts.

TABLE 3Selected studies on incident impact analysis.

Literature	Data	Study Area	Method	Research Focus
Vlahogianni et al., 2010	1,746 crash records for 2007 and 2008	Attica Tollway (65.2-km urban motorway) in Greece	Bayesian networks (BNs)	Secondary incident detection, Distance of secondary to primary incident
Zhang and Khattak, 2011	Nearly 80,000 incident records for 2008	Hampton Roads, VA	A deterministic queuing model-based identification method	Secondary incident detection, Distance of secondary to primary incident, Time gap of secondary to primary incident
Chung, 2013	Around 6,200 crashes and 52 million traffic flow records from 2001 to 2002	I-5, I-405, SR-55, SR-57, and SR-91, Orange County, CA	Binary integer programming (BIP)	Secondary incident detection, Distance of secondary to primary incident, Time gap of secondary to primary incident
Chen et al., 2016	Traffic data from loop detectors and 1,377 incident records for 2013	I-15 Northbound corridor (25-mile), Salt Lake City, UT	K-nearest neighbor (KNN)	Secondary incident detection, Distance of secondary to primary incident, Time gap of secondary to primary incident
Kitali et al., 2019	66,756 incidents and speed data from Bluetooth device from 2015 to 2017	I-95 (35-mi), I-10 (21-mi) , and I-295 (61-mi), Jacksonville, FL	Static and dynamic methods	Secondary incident detection, Distance of secondary to primary incident, Time gap of secondary to primary incident
Chung and Recker, 2013	Around 6,200 crashes and 52 million traffic flow records from 2001 to 2002	I-5, I-405, SR-55, SR-57, and SR-91, Orange County, CA	Binary integer programming (BIP)	Spatiotemporal region identification
Wang et al., 2018	1 incident on April 16th, 2016	A 5-km freeway in the North 3rd Ring Road, Beijing, China	Integer programming model	Spatiotemporal region identification
Ou et al., 2019	15 incidents with detector data	I-5, San Diego, CA	Fuzzy clustering	Spatiotemporal region identification
Zheng, 2021	1 incident on March 17th, 2010	I-5, San Diego, CA	Integer programming model	Spatiotemporal region identification

Miller and Gupat, 2012	28 million sensor and 173 incident records in 2009; 32 million sensor and 244 records in 2011	I-5 in Los Angeles; US-101 in San Francisco Bay Area, CA	AdaBoost and K-nearest neighbor (KNN)	Cost of delay or duration prediction
Pan et al., 2015	450 million sensor and 6,811 incident records	I-405 and I-5 in Los Angeles County, CA	Numerical modeling method	Propagation behavior prediction
Liu et al., 2017	2 incident cases	Quyang Road, Shanghai, China	GIS spatiotemporal analysis	Propagation behavior prediction
Huang et al., 2020	Simulated incident scenarios	NA	Conditional deep convolutional GAN (C-DCGAN) model	Propagation behavior prediction
Current study	17,808 incident records and statewide HERE traffic data	All interstates in Alabama	Machine learning-based model	Spatiotemporal extent/volume of a queue, Understanding the correlations of a queue

3. Project Framework and Objectives

3.1 Project Overall Framework

The project scope is limited to the highway segments where adequate data is available. The key data streams include HERE traffic data, ALGO incident data, and CARE crash data as described in the following sections. The project scope is limited to Interstates based on the current availability of HERE traffic data. Besides, the research team also identified the high crash/incident occurrence segments and developed separate models for these sites. This project delivers a set of tools that can be used to support proactive incident management. In addition, the description of the steps of tool development are incorporated in this technical report. The tools and the report can be useful to a range of end users (ALDOT management, ALDOT engineers, external consultants, and others), particularly useful for traffic incident managers and responders. The report is also an important guide for the potential physical implementation of models proposed in this project for proactive traffic incident management. FIGURE 2 presents an overview of the project framework and workflow to develop and deploy the proactive traffic incident management tools.

3.2 Project Main Objectives

The objective of the team is to develop real-time crash risk prediction models by exploiting crowdsourced probe vehicle data that are not limited to a specific environment but cover diverse road environments. The objective of the team is to use a statewide live traffic database from HERE to detect freeway traffic incidents. The objective of the team is to identify the correlations of spatiotemporal impacts of traffic incidents by exploiting crowdsourced probe vehicle data (HERE live traffic speed database), pairing a dataset containing over 10,000 traffic incidents that occurred in 2019, covering the entire freeway network in Alabama.

The overall objective of the project is to develop a set of tools to support proactive traffic operations and incident management on Alabama interstates. In doing so, the project utilizes three sets of big transportation data in Alabama: HERE traffic data, ALGO incident data, and CARE crash data. These data were integrated and examined to:

- Understand the relationships between traffic flows, incidents, and crashes on Interstate Highway in Alabama.
- Support the development of tools to detect and potentially forecast traffic incidents and crashes on Interstate in Alabama.

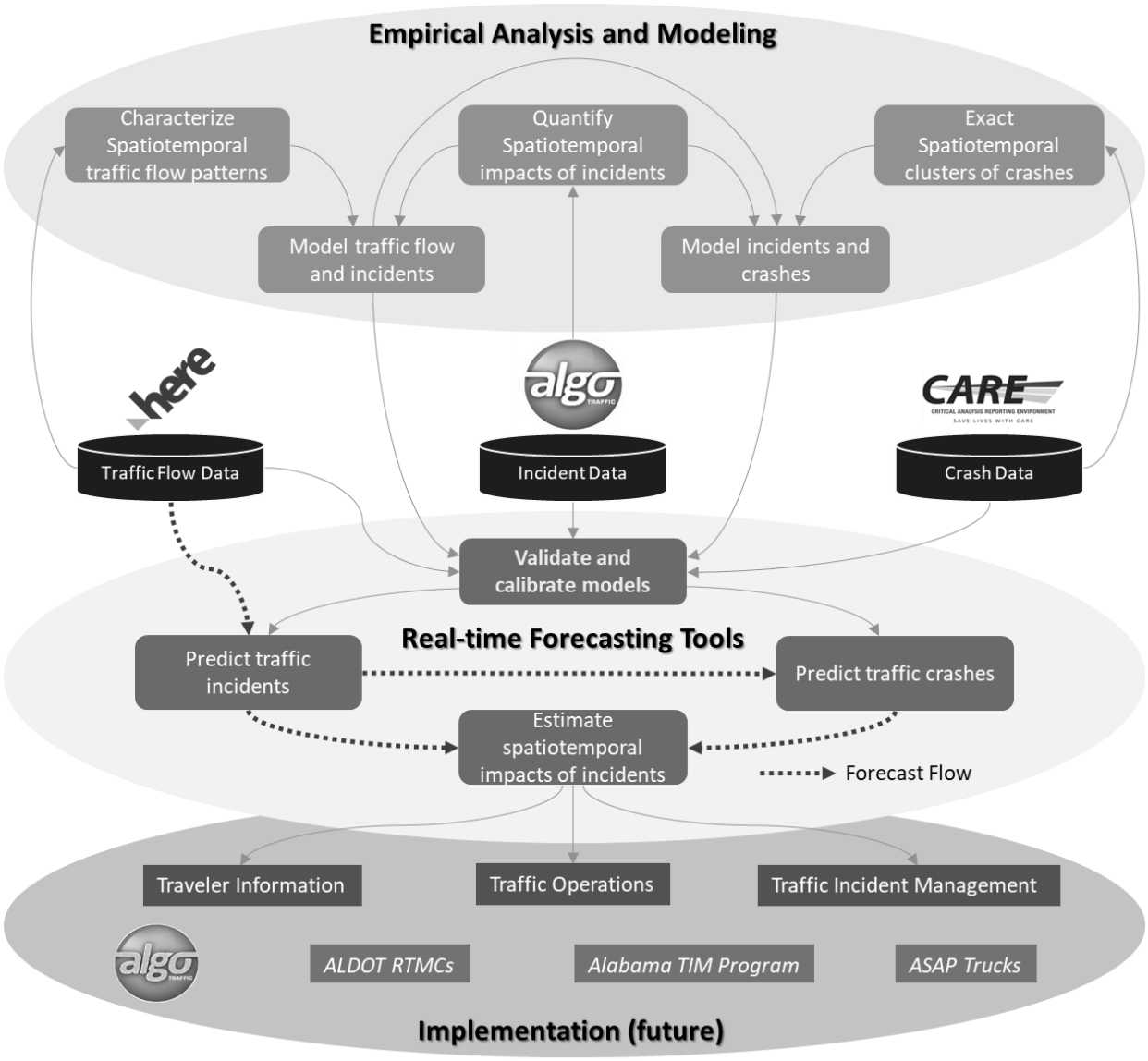


FIGURE 2 Overall Methodology Framework

4. Data

4.1 Data Sources

4.1.1 HERE Traffic Data

To monitor congestion, ALDOT has recently purchased state-wide crowdsourced mobility data from HERE. This data is used in the ALGO Traffic Platform for real-time speed observation, but the data is also continuously collected and stored and can be used to generate performance metrics that provide a stronger quantitative assessment of mobility. These metrics will enable Alabama to follow recent FHWA guidance that supports the Fixing America's Surface Transportation (FAST) Act by measuring and assessing system performance (FHWA, 2017). Across the state of Alabama, there are just over 10,000 traffic message channels (TMCs) that record the average speed of probe vehicles on the segment each minute, as shown in FIGURE 3. UA has been storing this data in an SQL server that has grown to several terabytes to contain data from 2017-current. A sample of this data is shown below in TABLE 4.

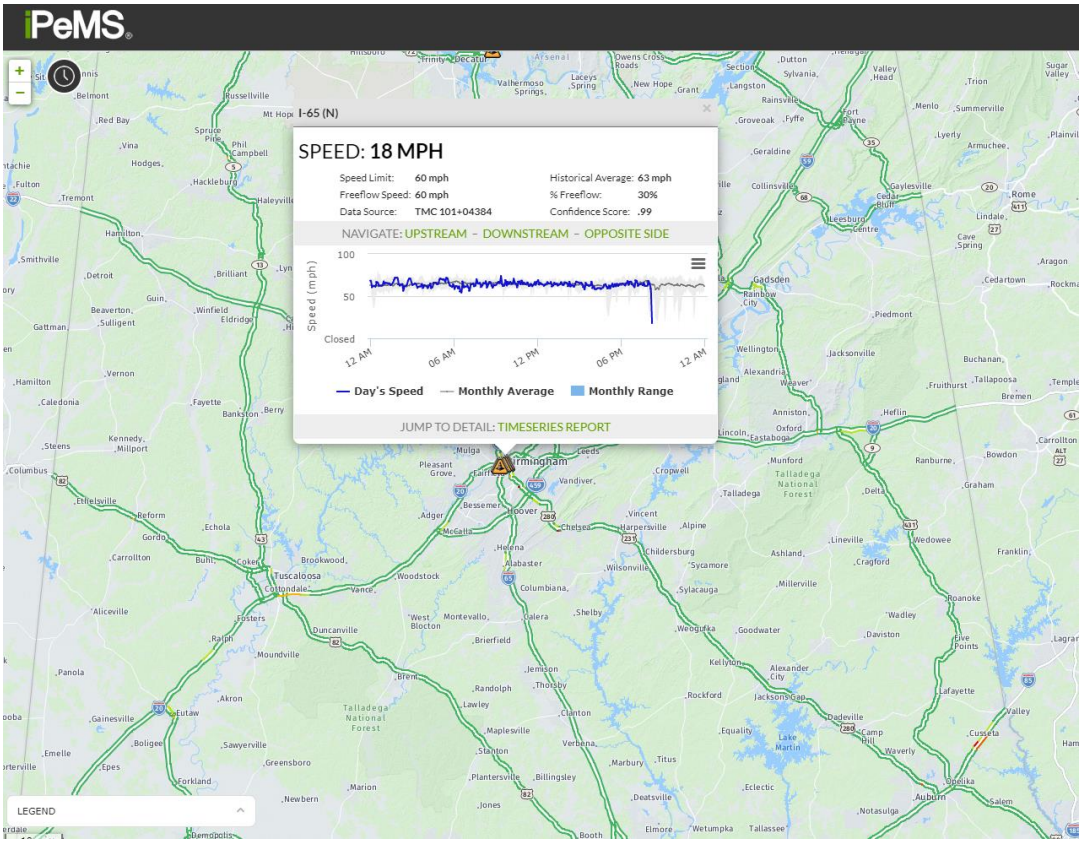


FIGURE 3 Coverage of HERE traffic data in Alabama with sample TMC records

TABLE 4 Sample TMC Speed Data from UA Database

Results | Messages

	tstamp	TMC	linear	isSub	offset	sub_len	length	queue_dir	flow_type	speed_cap	speed	ff_speed	jam_factor	confidence
1	2018-08-01 00:00:46.0000000	102+08361	102+03003	0x00	0	NULL	1.659192	-	TR	66.1964	68.0547	65.75513	0	0.77
2	2018-08-01 00:00:46.0000000	102+08362	102+03003	0x00	0	NULL	3.100385	-	TR	61.9018	63.12616	61.59105	0	0.8
3	2018-08-01 00:00:46.0000000	102+08363	102+03003	0x00	0	NULL	1.77115	-	TR	64.07085	66.9422	63.45556	0	0.79
4	2018-08-01 00:00:46.0000000	102+08364	102+03003	0x00	0	NULL	0.9805283	-	TR	64.63642	64.63642	65.01554	0.05228714	0.7
5	2018-08-01 00:00:46.0000000	102+08365	102+03003	0x00	0	NULL	0.5287135	-	TR	65.25171	65.25171	64.63642	0	0.7
6	2018-08-01 00:00:46.0000000	102+08366	102+03003	0x00	0	NULL	0.8226973	-	TR	64.63642	64.63642	64.63642	0	0.7
7	2018-08-01 00:00:46.0000000	102+08367	102+03003	0x00	0	NULL	0.7036793	-	TR	65.25171	65.25171	64.63642	0	0.7
8	2018-08-01 00:00:46.0000000	102+08368	102+03003	0x00	0	NULL	0.5250093	-	TR	59.04288	59.04288	58.9248	0	0.7
9	2018-08-01 00:00:46.0000000	102+08369	102+03003	0x00	0	NULL	0.658931	-	TR	52.83406	52.83406	52.89	0.01070851	0.7
10	2018-08-01 00:00:46.0000000	102+08370	102+03003	0x00	0	NULL	0.7198446	-	TR	55.31386	55.31386	52.26849	0	0.7
11	2018-08-01 00:00:46.0000000	102+08371	102+03003	0x00	0	NULL	0.4232629	-	TR	52.21255	52.21255	52.70354	0.0859913	0.7
12	2018-08-01 00:00:46.0000000	102+08372	102+03003	0x00	0	NULL	2.600783	-	TR	60.28589	60.28589	60.16159	0	0.7
13	2018-08-01 00:00:46.0000000	102+08373	102+03003	0x00	0	NULL	5.190292	-	TR	65.81107	65.91672	65.19577	0	0.75
14	2018-08-01 00:00:46.0000000	102+08374	102+03003	0x00	0	NULL	3.291343	-	TR	67.45184	67.45184	70.16781	0.3519764	0.81
15	2018-08-01 00:01:46.0000000	102+08361	102+03003	0x00	0	NULL	1.659192	-	TR	66.1964	68.0547	65.75513	0	0.77
16	2018-08-01 00:01:46.0000000	102+08362	102+03003	0x00	0	NULL	3.100385	-	TR	61.9018	63.12616	61.59105	0	0.8
17	2018-08-01 00:01:46.0000000	102+08363	102+03003	0x00	0	NULL	1.77115	-	TR	64.07085	66.9422	63.45556	0	0.79
18	2018-08-01 00:01:46.0000000	102+08364	102+03003	0x00	0	NULL	0.9805283	-	TR	64.63642	64.63642	65.01554	0.05228714	0.7
19	2018-08-01 00:01:46.0000000	102+08365	102+03003	0x00	0	NULL	0.5287135	-	TR	65.25171	65.25171	64.63642	0	0.7

The crowdsourced probe vehicle data used in this project is from the HERE traffic database, which provides live traffic information for the entire freeway network in Alabama (HERE, 2021). FIGURE 4 shows the freeway network in Alabama (I-459, I-59, I-65, I-85, I-10, I-20, I-20/I-59, I-22, I-565). The HERE traffic database provides the live speed information updated every minute for each traffic management channel (TMC). A TMC is a pre-defined section of the road for traffic data reporting (Esri, 2020). As shown in FIGURE 5, the size of TMC can range from under 0.1 mile to a few miles. TABLE 2 shows the descriptive statistics of 635 TMCs for the Alabama freeway network. The total length of these TMCs is 1,976 miles, and the average length is 3.1 miles. Note that the TMCs are specified for each direction. One TMC may be split into several shorter dynamic units to capture the speed information in a higher resolution when the traffic flow speed changes quickly within a TMC.

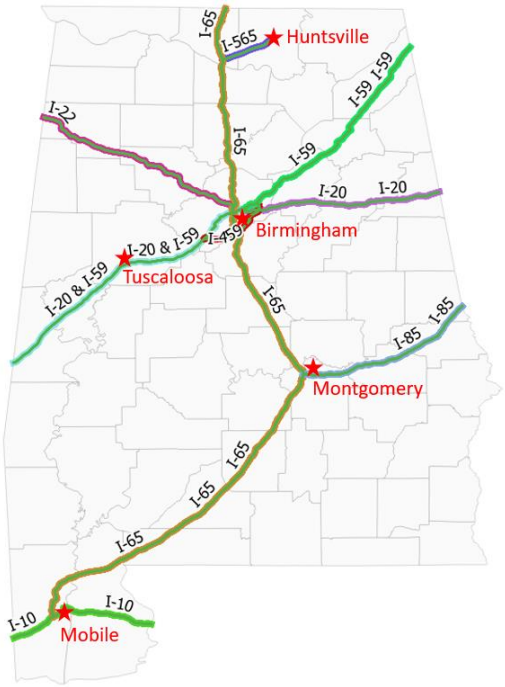


FIGURE 4 Selected Interstate freeways

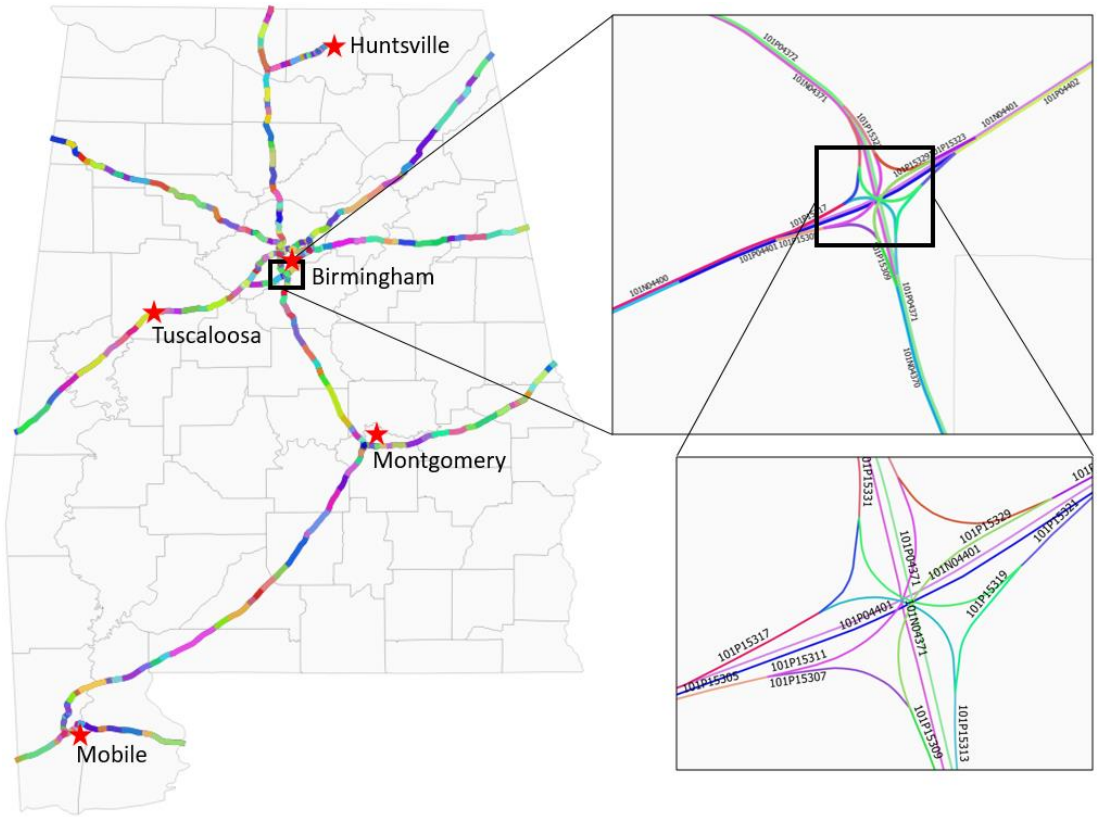


FIGURE 5 Traffic management channels (TMCs) for traffic data reporting
(Note: different colors are used to distinguish different TMCs)

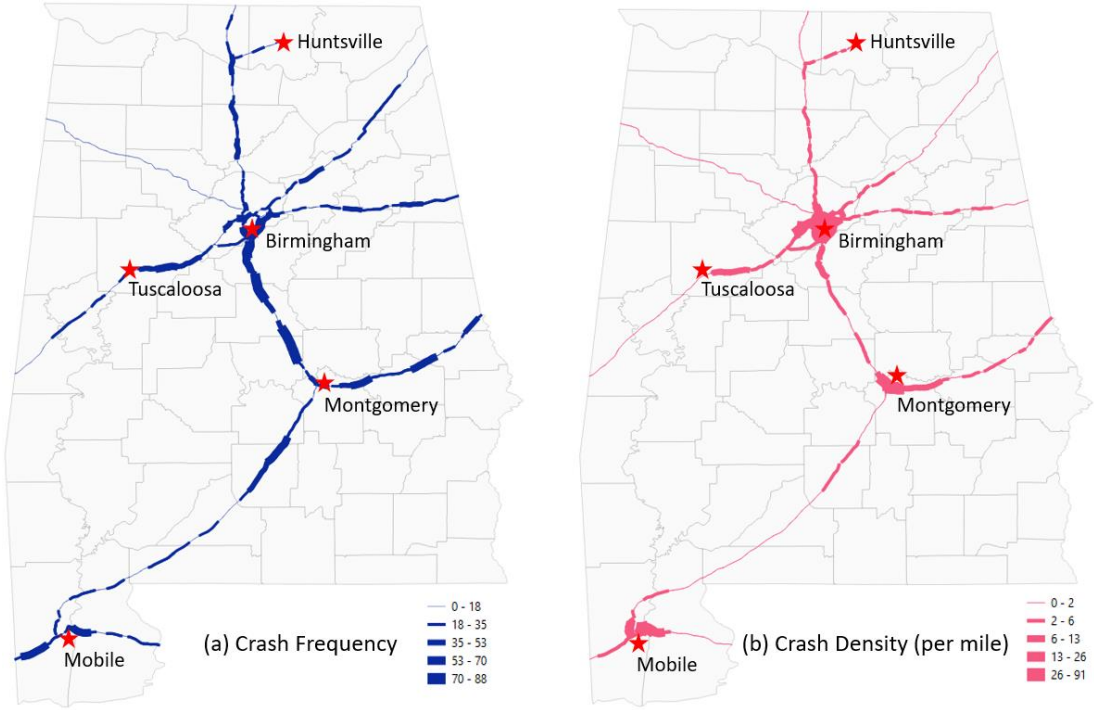


FIGURE 6 Spatial distribution of (a) crash frequency and (b) crash density per mile

FIGURE 7 Spatial distribution of freeway traffic volumes (AADT)

TABLE 5 Descriptive statistics for TMC length (Unit: mile)

Primary Road	Direction	Number of TMC	Total Length of TMC	Min	Max	Mean	Var
I-459	NS	28	65.401	0.082	4.581	2.336	1.850
I-59	NS	54	222.200	0.026	17.046	4.115	18.057
I-65	NS	214	731.130	0.029	13.872	3.416	9.833
I-85	NS	64	159.891	0.064	9.242	2.498	5.924
I-10	WE	43	132.749	0.044	13.426	3.087	10.075
I-20	WE	56	168.769	0.042	7.539	3.014	4.109
I-20/I-59	WE	88	259.822	0.027	9.340	2.953	7.285
I-22	WE	56	191.724	0.034	7.284	3.424	2.517
I-565	WE	32	44.009	0.238	4.485	1.375	1.019
Total		635	1,975.695	0.026	17.046	3.111	8.119

The team paired the crowdsourced traffic data with the traffic crash data to extract the pre-crash traffic flow dynamics. The traffic crash data were obtained from the CARE crash database maintained by the Alabama Department of Transportation (ALDOT) (FHWA, 2020). The crashes that occurred in 2019 on the targeted freeways (I-459, I-59, I-65, I-85, I-10, I-20, I-20/I-59, I-22, I-565) within the state of Alabama were extracted from the CARE crash database. After data cleaning (removing crash records that miss the time or location information), the remaining 9,997

crash records were used to extract the pre-crash traffic information from the HERE database. TABLE 6 shows the descriptive statistics of sampled crashes regarding crash types. In total, the top three crash types - single-vehicle crashes, rear-end crashes, and sideswipe crashes account for 35.6%, 34.7%, and 18.3% of the total crashes, respectively. Spatially, approximately 40% of the crashes occurred on the I-65, followed by 14.4% of the crashes that occurred on the I-20/I-59. The spatial distribution of interstate freeway crash frequency and crash density is shown in FIGURE 6 (a) and (b), respectively.

To reveal the impact of the diverse road environments on crash risk, the team pulled the road infrastructure data from the Highway Performance Monitoring System (HPMS) database. HPMS database provides road environment variables, including annual average daily traffic (AADT), number of lanes, distance to the closest upstream ramp, and land use. As shown in FIGURE 4, the spatial distribution of freeway traffic volume (AADT) is consistent with the spatial distribution of crash density.

TABLE 6 Descriptive statistics for crashes in interstate freeways

Primary Road	Crash Type									
	Rear-End		Sideswipe		Single Vehicle Crash		Other		Total	
	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent
I-459	383	50.4%	142	18.7%	153	20.1%	82	10.8%	760	7.6%
I-59	118	19.9%	77	13.0%	314	53.0%	83	14.0%	592	5.9%
I-65	1,388	35.4%	677	17.2%	1432	36.5%	429	10.9%	3926	39.3%
I-85	360	34.1%	201	19.0%	369	34.9%	126	11.9%	1056	10.6%
I-10	417	45.9%	171	18.8%	246	27.1%	74	8.1%	908	9.1%
I-20	200	22.4%	164	18.4%	400	44.8%	128	14.3%	892	8.9%
I-20/I-59	447	31.1%	331	23.1%	493	34.3%	165	11.5%	1436	14.4%
I-22	40	18.8%	27	12.7%	116	54.5%	30	14.1%	213	2.1%
I-565	112	52.3%	40	18.7%	35	16.4%	27	12.6%	214	2.1%
Total	3465	34.7%	1830	18.3%	3558	35.6%	1144	11.4%	9997	100.0%

4.1.2 ALGO Incident Data

This project acquired traffic incident data from ALGO Traffic which was created by ALDOT to provide live traffic camera feeds, updates on Alabama roads, and access to exclusive ALDOT information such as message sign readouts, incident and construction information, and current road congestion levels. Over the past few years, ALDOT has taken a transportation systems management and operations (TSMO) approach to manage intelligent transportation system (ITS) assets and monitor congestion across their road network. Regional Traffic Management Centers (RTMCs) have been established in four of the five regions to monitor data and information from the ALGO Traffic web interface. FIGURE 8 shows a set of traffic incidents from RTMCs. The University of Alabama (UA) through the Center for Advanced Public Safety (CAPS) was involved in the development of the ALGO Traffic Platform. The team worked with ALDOT and CAPS to obtain the traffic incident reports that cover information including incident start and end time, road type, location, and incident type.

Type	ID	Type	Start Time	End Time	Road Type	Location	County	Event Status
	18757	Abandoned Vehicle	08/24/2017 08:41	08/25/2017 09:41	Interstate	I-20/59 W at Bama Rock Garden	Tuscaloosa	Confirmed
	18756	Bridge Repair / Inspection	08/24/2017 08:00	08/24/2017 14:00	Interstate	I-20/59 W at Holly Springs Road/SR 300 to I-20/59 W at Holly Springs Road/SR 300	Tuscaloosa	Confirmed
	18755	Bridge Repair / Inspection	08/24/2017 07:30	08/24/2017 16:00	Interstate	I-65 S before CR 28; Lake Mitchell Rd to I-65 S before CR 28 Lake Mitchell Rd	Chilton	Confirmed
	18750	Resurfacing or Paving	08/23/2017 08:00	08/24/2017 16:00	U.S. Highway	US 43 N past Shiver De Freeze Rd/Doughty Rd to US 43 N before Fulmer Rd	Tuscaloosa	Confirmed
	18727	Bridge Repair / Inspection	08/21/2017 08:00	08/31/2017 16:00	Interstate	I-20/59 E at University Blvd/US 11/SR 7 to I-20/59 E before Keenes Mill Rd	Tuscaloosa	Confirmed
	18726	Resurfacing or Paving	08/21/2017 10:44	11/30/2018 11:44	Interstate	I-65 S before CR 42 to I-65 S past CR 42	Chilton	Confirmed
	18725	Resurfacing or Paving	08/21/2017 08:00	08/31/2017 16:00	Interstate	I-20/59 E at Skyland Blvd to I-20/59 E at University Blvd/US 11/SR 7	Tuscaloosa	Confirmed
	18724	Resurfacing or Paving	08/21/2017 08:00	01/01/2018 16:00	U.S. Highway	US 31 N before CR 44 to US 31 N before CR 95	Chilton	Confirmed
	18563	Strong Winds	07/29/2017 03:09	07/29/2018 03:09		E at Montgomery	Montgomery	Unconfirmed
	18554	Resurfacing or Paving	07/26/2017 07:00	09/30/2017 17:30	Interstate	I-20/59 W at Exit 68: Joe Mallisham Pkwy to I-20/59 W at Holly Springs Road/SR 300	Tuscaloosa	Confirmed
	18544	Resurfacing or Paving	07/28/2017 06:00	10/31/2017 06:00	Interstate	I-20/59 W at Exit 62 : Holly Springs Rd/SR 300 to I-20/59 W before Exit 52 US 11/CR 231	Tuscaloosa	Confirmed
	18456	Resurfacing or Paving	07/17/2017 07:00	09/01/2017 18:00	U.S. Highway	US 82 EW before SR 5 to US 82 EW past SR 25/Montevallo Rd	Bibb	Confirmed
	18265	Strong Winds	06/21/2017 11:16	06/21/2018 11:16		E at Montgomery	Montgomery	Unconfirmed
	18262	Strong Winds	06/21/2017 09:16	06/21/2018 09:16		E at Montgomery	Montgomery	Unconfirmed
	18164	Strong Winds	05/23/2017 09:18	05/23/2018 09:18		E at Montgomery	Montgomery	Unconfirmed
	18142	Strong Winds	05/22/2017 23:18	05/22/2018 23:18		E at Ft. Payne	DeKalb	Unconfirmed
	18078	Strong Winds	05/01/2017 02:05	05/01/2018 02:05		E at Decatur	Morgan	Unconfirmed
	18069	Strong Winds	04/30/2017 16:05	04/30/2018 16:05		E at Demopolis	Hale	Unconfirmed

FIGURE 8 A sample set of traffic incidents from RTMC

The incident data used in this project are from the Algo traffic incident database maintained by the Alabama Department of Transportation (ALDOT). The Algo traffic incident database stores all the reported incidents that occurred in the State of Alabama. In 2019, there were nearly 50,000 traffic incidents stored in the Algo traffic incident database, of which 39.35% happened on Interstate 65 (I-65). I-65 is the busiest interstate freeway in Alabama, with a total length of 366 miles (as shown in FIGURE 1). The team extracted all reported traffic incidents (N = 9,472) occurring on the Northbound I -65 in 2019.

TABLE 7 shows the frequency and percentage of incidents that occurred on I-65 northbound in 2019. The top three incident types are disabled vehicles, minor crashes, and abandoned vehicles, accounting for 60.30%, 11.92%, and 9.64% of all incidents, respectively. Besides, incident types including moderate crashes, minor crashes, and congestion also stand for a significant percentage (greater than 1%).

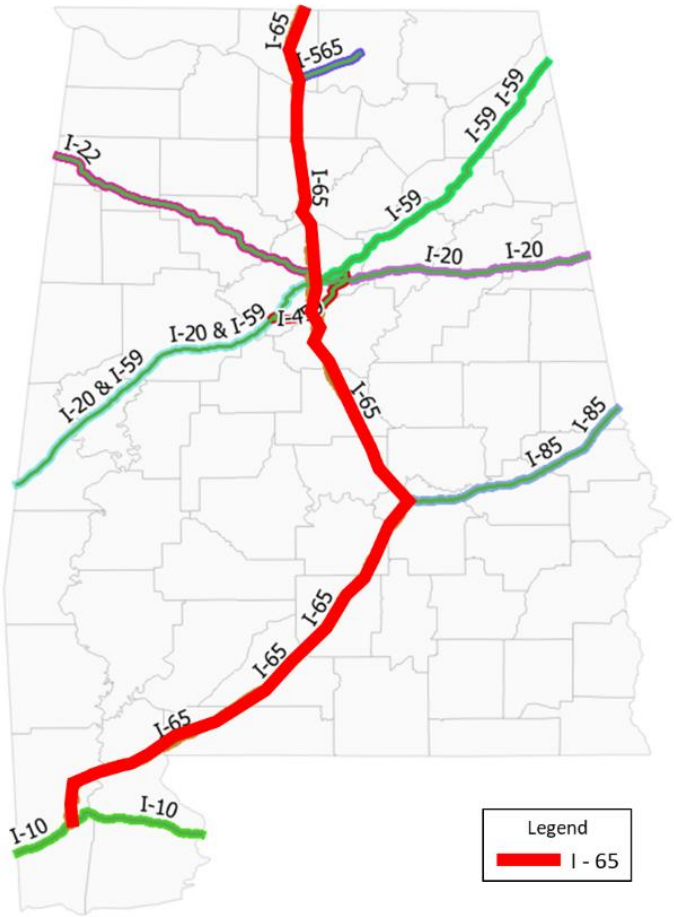


FIGURE 9 Study area – I- 65 in Alabama

TABLE 7 Incident subtype distribution

Incident type	Frequency	Percentage
Abandoned Vehicle	913	9.64%
Congestion	318	3.36%
Debris	349	3.68%
Disabled Vehicle	5,712	60.30%
Grass Fire	23	0.24%
HazMat Spill	5	0.05%
Major Crash	156	1.65%
Medical Emergency	22	0.23%
Minor Crash	1,129	11.92%
Moderate Crash	462	4.88%
Overturned Vehicle	56	0.59%
Police Activity	247	2.61%
Signal Outage	2	0.02%
Smoke	2	0.02%
Structure Fire	1	0.01%
Vehicle Fire	56	0.59%
Wildlife in Roadway	6	0.06%
Null value	13	0.14%
Sum	9,472	100.00%

4.1.3 CARE Crash Data

The Critical Analysis Reporting Environment (CARE) is a data analysis software package originally designed for problem identification and countermeasure development in traffic safety applications. It uses advanced analytical and statistical techniques to generate valuable information directly from data. A screenshot of the CARE online platform is shown in FIGURE 10. Using CARE's step-by-step on-screen menus, it is easy to turn data into enlightening information. The project team worked with researchers and engineers of CAPS at UA to extract traffic crashes on Interstates and other important arterials from CARE. The key crash information for this project includes the location and time of crashes, crash types and injury severities, and other standard crash features.

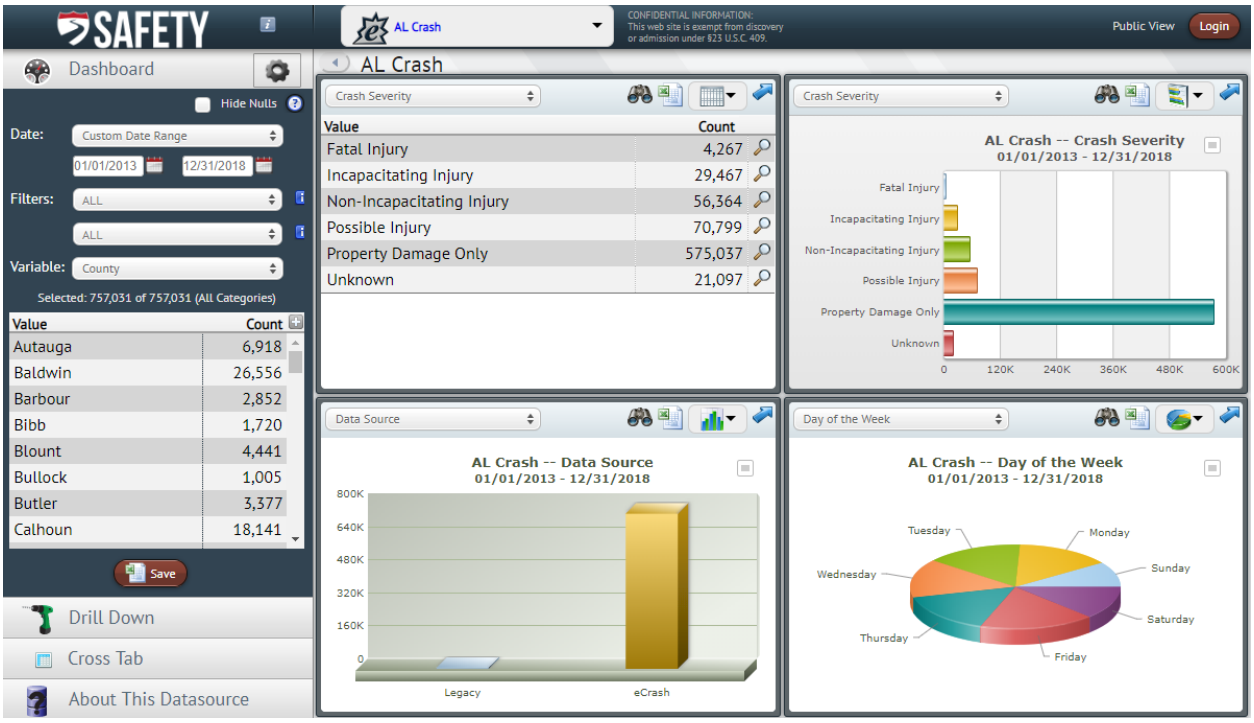


FIGURE 10 CARE online platform

The team paired the crowdsourced traffic data with the traffic crash data to extract the pre-crash traffic flow dynamics. The traffic crash data were obtained from the CARE crash database maintained by the Alabama Department of Transportation (ALDOT) (FHWA, 2020). The crashes that occurred in 2019 on the targeted freeways (I-459, I-59, I-65, I-85, I-10, I-20, I-20/I-59, I-22, I-565) within the state of Alabama were extracted from the CARE crash database. After data cleaning (removing crash records that miss the time or location information), the remaining 9,997 crash records were used to extract the pre-crash traffic information from the HERE database. TABLE 8 shows the descriptive statistics of sampled crashes regarding crash types. In total, the top three crash types - single-vehicle crashes, rear-end crashes, and sideswipe crashes account for 35.6%, 34.7%, and 18.3% of the total crashes, respectively. Spatially, approximately 40% of the crashes occurred on the I-65, followed by 14.4% of the crashes that occurred on the I-20/I-59. The spatial distribution of interstate freeway crash frequency and crash density is shown in FIGURE 3 (a) and (b), respectively.

To reveal the impact of the diverse road environments on crash risk, the team pulled the road infrastructure data from the Highway Performance Monitoring System (HPMS) database. HPMS database provides road environment variables, including annual average daily traffic (AADT), number of lanes, distance to the closest upstream ramp, and land use. As shown in FIGURE 4, the spatial distribution of freeway traffic volume (AADT) is consistent with the spatial distribution of crash density.

TABLE 8 Descriptive statistics for crashes in interstate freeways

Primary Road	Crash Type									
	Rear-End		Sideswipe		Single Vehicle Crash		Other		Total	
	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent
I-459	383	50.4%	142	18.7%	153	20.1%	82	10.8%	760	7.6%
I-59	118	19.9%	77	13.0%	314	53.0%	83	14.0%	592	5.9%
I-65	1,388	35.4%	677	17.2%	1432	36.5%	429	10.9%	3926	39.3%
I-85	360	34.1%	201	19.0%	369	34.9%	126	11.9%	1056	10.6%
I-10	417	45.9%	171	18.8%	246	27.1%	74	8.1%	908	9.1%
I-20	200	22.4%	164	18.4%	400	44.8%	128	14.3%	892	8.9%
I-20/I-59	447	31.1%	331	23.1%	493	34.3%	165	11.5%	1436	14.4%
I-22	40	18.8%	27	12.7%	116	54.5%	30	14.1%	213	2.1%
I-565	112	52.3%	40	18.7%	35	16.4%	27	12.6%	214	2.1%
Total	3465	34.7%	1830	18.3%	3558	35.6%	1144	11.4%	9997	100.0%

4.2 Data Processing

4.2.1 Data Processing for Incident Risk Prediction

Data used in this project are from three sources, including the HERE traffic database, the CARE crash database, and the HPMS highway infrastructure database. The team made a significant effort to link the data features and create variables for crash risk prediction. An automatic traffic characteristic tool was designed in this project to generate the Interstate traffic characteristics related to the incident time and location information. FIGURE 11 shows the framework of data linkage and variable creation. First, the TMCs in the HERE database were linked to the crashes from the CARE database according to the spatial locations of TMCs and crashes. To ensure enough spatial coverage for extracting the traffic data, the team linked crashes to TMCs within 5 miles upstream and downstream of the crash location. As a result, each crash was paired with a 10-mile segment for traffic data extraction from linked HERE TMCs. Second, from paired HERE TMCs the team extracted the traffic speed records before the crash occurrence time. The team considered three different pre-crash time points (10, 15, and 20 minutes before the crash) and one hour after the occurrence of crashes to extract the traffic information. At this point, we obtained the spatial-temporal speed matrix for a specific incident/crash given the unique incident identification number in the ALGO database/crash identification number in the CARE database.

To ensure the modeling results are comparable, the team extracted the traffic data from the same 20-minute intervals. Note, that other time intervals were also attempted, and the 20-min intervals were selected given the model performance. The different pre-crash time points start from the middle of the 20-minute intervals. For instance, if a crash occurred at 8:00 am, the team extracted the traffic data between 7:40 am, and 8:00 am for the 10-minute pre-crash time point, between 7:35 am and 7:55 am for the 15-minute time point, and between 7:30 am and 7:50 am for the 20-min pre-crash time point. Then, for each crash, a spatial-temporal speed matrix was created to

document the speed records extracted from the HERE TMCs. FIGURE 6 shows a schematic diagram of part of the spatial-temporal speed matrix. The speed matrix has a resolution of 0.1 miles per minute. Last, based on the spatial-temporal speed matrix, variables reflecting the traffic characteristics before the occurrence of the incident are calculated which are presented in the later section. Besides the pre-crash traffic information, the road environment attributes from the HPMS database were matched to crashes and TMCs.

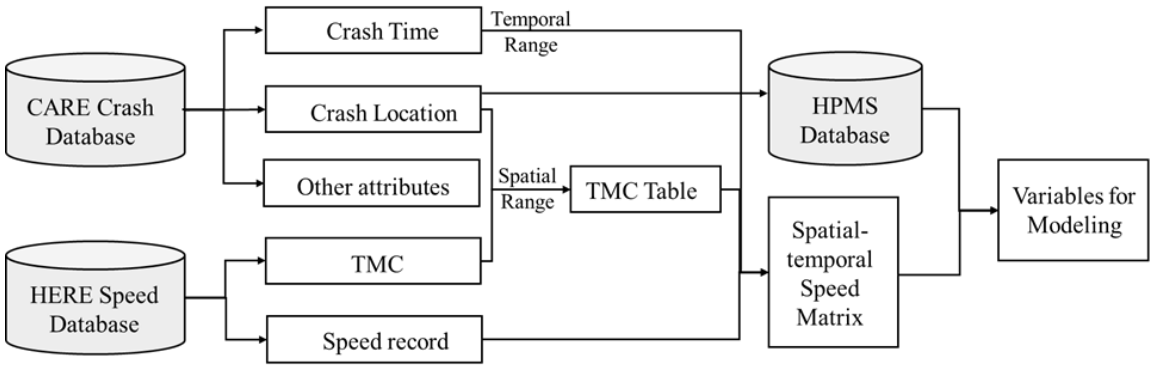
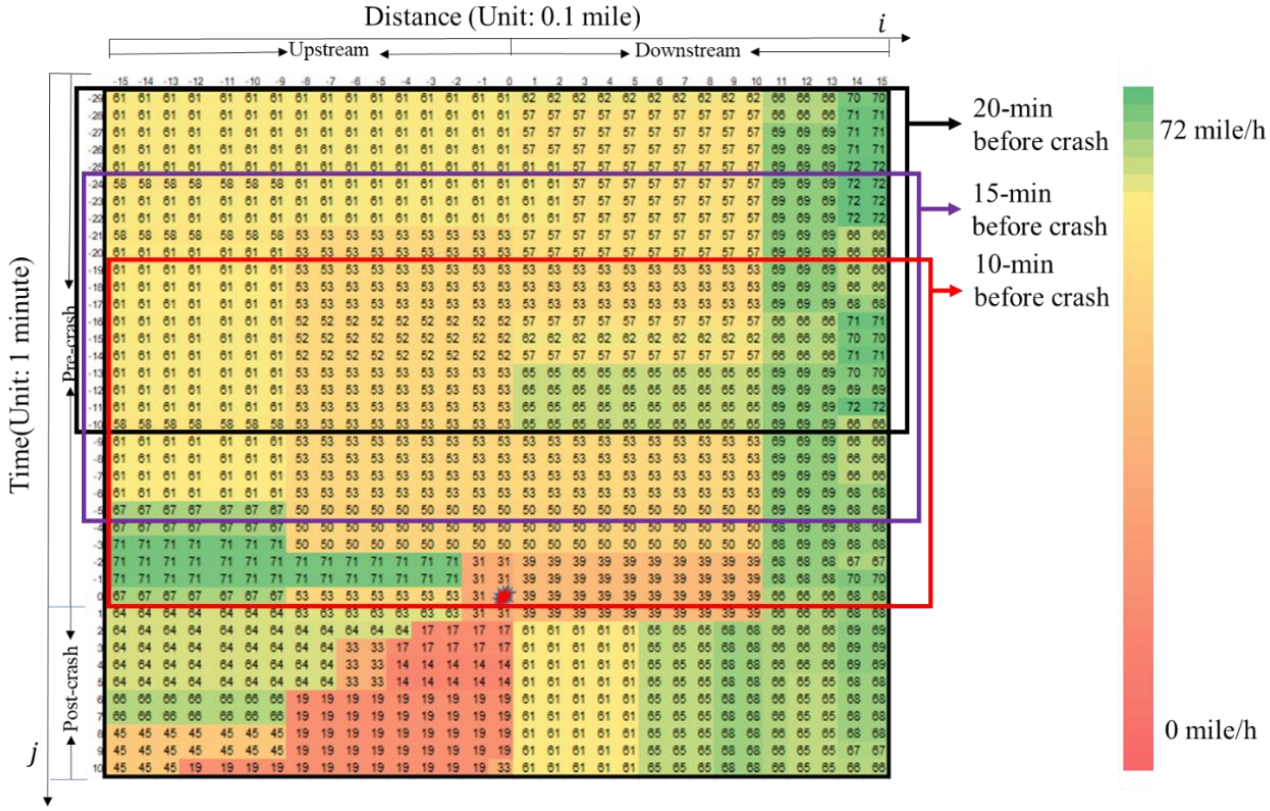


FIGURE 11 The framework of data linkage and processing



Notes: (1) This FIGURE only shows part of the spatial-temporal speed matrix (1.5 miles upstream and 1.5 downstream), the whole spatial-temporal speed matrix covers traffic dynamics in 5 miles upstream and 5 miles downstream. (2) The red point shows the time and location of occurrence of a crash.

FIGURE 12 Spatial-temporal speed matrix

4.2.2 Data Processing for Incident Detection

Similar to 4.2.1 incident risk prediction, one major data preparation step for developing the automatic incident detection model is to link incidents with their corresponding post-incident speed dynamics together. We used the speed information extraction framework shown in FIGURE 12. The main idea behind this framework is to extract the speed information from the HERE database based on the time and spatial location of the incident within predefined temporal and spatial ranges. The outcome of the framework is a spatial-temporal speed matrix with a resolution of 0.1 miles * 1 minute. In other words, the difference of data processing in this section and the previous section is the range of defined spatial dimension and the temporal dimension. FIGURE 13 visualizes the spatial-temporal speed matrix for a major crash. Significant speed reduction can be seen after the occurrence of the crash. The black box shows the detection area, as shown in FIGURE 13. Because the occurrence of an incident seems to have more influence on the upstream traffic than the downstream traffic. Therefore, the spatial extent of the detection area in this project ranges from 2 miles upstream and 1 mile downstream of the incident location. Regarding the temporal range of the spatial-temporal matrix, it covers the traffic dynamics 15 minutes before and 5 minutes after the incident. The speed within the area was extracted and served as the input of the deep learning models later.

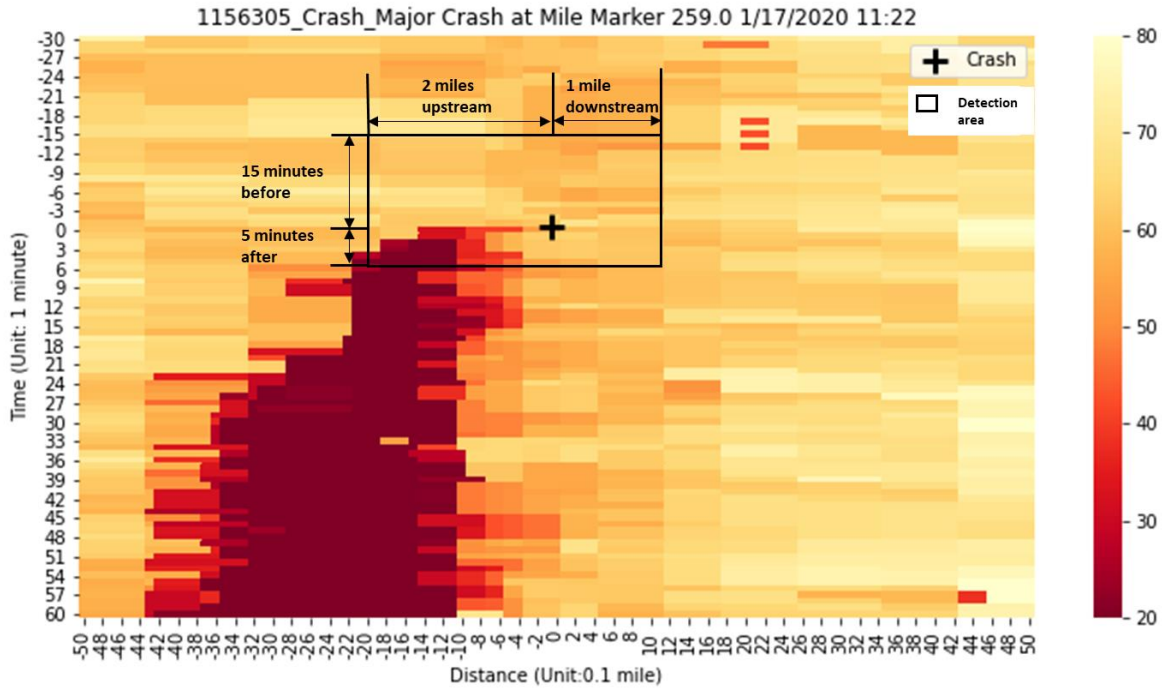


FIGURE 13 Visualization of the spatial-temporal speed matrix

In addition to creating the spatial-temporal speed matrices for incident conditions, the team also created spatial-temporal speed matrices for incident-free conditions paired with the incident conditions. Specifically, for each incident record, one incident-free record was generated at the exact location and the same time of day but on a different day when there was no incident.

4.2.3 Data Processing for Incident Impact Estimation

Pairing the traffic incident data with crowdsourced traffic data using incident time and location information to extract the pre-incident traffic flow dynamics is a crucial step in the data processing

for evaluating the incident/crash impact. From a spatial perspective, the TMCs are used to link the HERE database to the ALGO incident records using location information. As the length of a TMC segment varies across the freeway network, the team ties incidents to TMCs within a 5-mile range upstream and downstream of the incident location. Similarly, from a temporal standpoint, traffic flow speed data or traffic dynamics are retrieved based on the incident's time of occurrence. Therefore, a spatial-temporal speed matrix is generated for each incident to represent the extracted speed records from the HERE database. Lastly, road environment-related variables from the HPMS database are matched to all 8,178 incidents.

5. Methodology

5.1 Machine Learning

5.1.1 Modeling Methods

Logistic Regression

The logistic regression has been widely adopted in crash risk modeling (Hossain et al., 2019). In a crash risk model, the dependent variable Y is a binary variable indicating whether there was a crash at a site during a time period. The model can be expressed in a linear form as:

$$\ln\frac{p}{1-p} = \alpha + \beta X \quad (1)$$

Where p denotes the probability for the crash case (y = 1) and $1-p$ for the crash-free case (y = 0); α is the intercept term, β is the vector of model parameter estimates for independent variables X.

Random Forest

Random Forest (RF) is one of the most popular ensemble methods in machine learning for data classification. The RF is a combination of multiple decision trees, and each tree is a class prediction model (Breiman, 2001). The result of RF is the average prediction of all decision trees. The general idea behind the random forest is shown in FIGURE 7. Compared with the individual tree or class prediction model, the RF is expected to have lower variance and overcome the overfitting problem. In this project, the RF was adopted to model the complex nonlinear relationships between the crash risk and associated factors based on its flexible modeling structure. In this project, two hyperparameters, including the number of decision trees in the forest and the number of features, were considered for each decision tree when splitting a node, and the three-fold cross-validation was performed to optimize the model performance. To determine the number of features at each split, two commonly used methods, including the square root of the total number of variables and base two logarithm of the total number of variables, were tested. For the number of trees, values from 50 to 550 at an interval of 50 were tested. Overall, 20 possible combinations were examined to gain the optimized parameters.

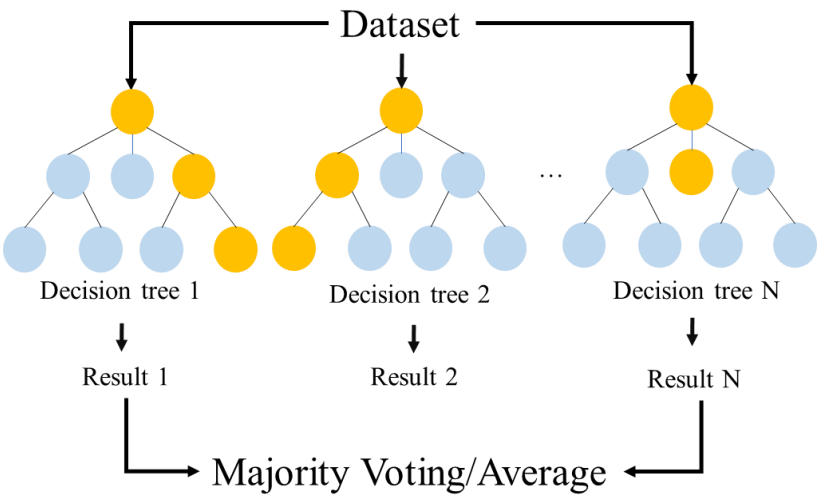


FIGURE 14 Random Forest Model Framework

Extreme Gradient Boosting

The Extreme Gradient Boosting (XGBoost) algorithm is a scalable end-to-end tree boosting system created by Chen & Guestrin (2016). It can be used to solve classification problems and regression problems efficiently. Based on the idea of "boosting," the algorithm combines all the predictions of a group of "weak" learners for developing a "strong" learner through additive training strategies. The XGBoost aims to optimize the value of the objective function and implement the machine learning algorithm within the framework of gradient boosting. The objective function of XGBoost is composed of a loss function and a regularization term:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

where l is a differentiable convex loss function that measures the difference between the predicted value $\hat{y}_i$ and the target value y_i; n is the number of observations in the training dataset; K is the number of trees to be generated; f_k is an independent tree from the ensemble trees. The second term is the regularization term which is used to penalize the complexity of the model. It can be defined as:

$$\Omega(\mathrm{f}) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel \quad (3)$$

where γ is the minimum split loss reduction; λ is the regularization parameter; ω is a vector of weights for leaves. The detailed algorithm and computation procedures of the XGBoost can be found in Chen & Guestrin (2016). In this project, the learning rate, and the number of estimators (trees) were tuned by using a grid search method. The learning rate ranges from 0.05 to 0.3 (with an interval of 0.05) and the number of estimators is from 50 to 550 (with an interval of 50).

Support Vector Machine

The Support Vector Machine (SVM) approach is to find a hyperplane in N-dimensional space (N is the number of features) that the data points can be separated into different classes such as the crash and crash-free observations. The hyperplanes were constructed by the training vector (support vectors) that lie close to the class boundary. The hyperplane can be described using a linear classification function $g(x) = w \cdot x + b$, where w is the normal vector of the hyperplane and b is a variable. The optimal hyperplane is a decision boundary that maximizes the margin distance between the data points and the hyperplane. This optimization problem can be written as:

$$min\emptyset(w) = \frac{1}{2}\|w\|^2 \quad (4)$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 \quad i = 1,2,\ldots,l \quad (5)$$

where $\emptyset(w)$ is an objective function of w. To construct a potentially nonlinear hyperplane between crash and crash-free observations, slack variables ξ and a penalty factor C are integrated into a modified objective function written as:

$$min\emptyset(\omega) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \quad (6)$$

The penalty factor C was used and tuned to control the degree of tolerance of the misclassification. A larger value of C means more penalty for misclassification. To better solve the nonlinear classification problem, kernel functions are often used to transform a nonlinear decision surface to a linear decision surface by mapping the data into high-dimensional spaces (Vapnik, 2013). The dot productions $x_i \cdot x_j$ and $x_i \cdot x$ are transformed by kernel function $K(x_i, x_j) = \left(\emptyset(x_i) \cdot \emptyset(x_j)\right)$ and $K(x_i, x) = \left(\emptyset(x_i) \cdot \emptyset(x)\right)$. After applying the Lagrange theorem, the final decision function can be obtained:

$$f(x) = sign(\sum_{i=1}^{l} a_i\, y_i K(x_i, x) + b) \quad (7)$$

subject to

$$0 \leq\ a_i \leq C, i = 1,2, \dots, l,$$

$$\sum_{i=1}^{l} a_i\, y_i = 0$$

The book by Bishop (2006) provides more mathematical details about the kernel functions. In this project, two types of commonly used kernel functions including the radial basis kernel function (RBF) and polynomial kernel functions (linear kernel, quadratic kernel, and cubic kernel) are adopted along with various combinations of SVM parameters (four values of C: 0.1, 1, 10, 100; three values of gamma: 1, 0.1, 0.01). Models with different kernel functions and parameters were trained based on the splitting ration of 8:2 (80% of the whole dataset as training dataset and 20% of the whole dataset as testing dataset), and the results from models with the best prediction performance are presented in this paper.

The team needed a binary classification model to distinguish the two classes of observations: crash and crash-free. In binary machine learning classification models, the default threshold for distinguishing two classes (labeled as 0 and 1) is 0.5. In this project, when the value of the predicted probability is greater or equal to 0.5, the model predicts a crash case given the traffic dynamics along with other road environment attributes. Otherwise, the model predicts a crash-free case. The choice of this threshold will influence the trade-offs of the positive error and negative error. Therefore, in this project, the ROC curve was used to optimize the threshold after tuning the above-mentioned parameters. The G-Mean was calculated for each threshold based on the X-axis and Y-axis of the ROC curve to balance the true positive rate and false-positive rate.

$$\mathrm{G-Mean}\ =\ sqrt(tpr\ *\ (1 - fpr)) \quad (8)$$

where tpr denotes the true positive rate; fpr denotes the false positive rate.

Categorical Naive Bayes (CNB)

Naive Bayes classifier is a probabilistic machine learning model based on the Bayes theorem that is designed to calculate the probability of each target liable given the feature/variable sets. The stability, reliability, and simplicity of Naive Bayes classifier make it widely applied in text classification and spam filtering. The Categorical Naive Bayes (CNB) is derived from the Naive Bayes and is suitable for classification with discrete features, which assumes categorical distribution for each feature. Give a training dataset X, for each feature i, the probability of category k in feature i given class c could be estimated by the following equation.

$$P(x_i = k|y = c; \alpha) = \frac{N_{i,k,c}+\alpha}{N_c+\alpha.n_i} \quad (9)$$

where, $N_{i,k,c}$ is the frequency of $x_i = k$, which belongs to class c; N_c is the frequency of c, α is the Laplace smoothing parameter to handle zero frequency problem, n_i is the number of all possible categories of feature i. The Maximum A Posterior (MAP) estimation is used to estimate the $P(x_i|y)$ and $P(y)$ from the training dataset and the probability of prediction c given the x_i could be calculated by the following equation.

$$P(y = c|x_1, x_2, x_3, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y = c) \quad (10)$$

$$y = arg\max_y P(y)\prod_{i=1}^{n} P(x_i|y = c) \quad (11)$$

Adaptive Boosting (AdaBoost)

Adaptive Boosting, AdaBoost for short, indicates another method of ensemble learning – boosting. The AdaBoost is a statistical classification meta-algorithm developed by Freund et al. (Freund et al., 1999). The AdaBoost can combine different types of machine learning algorithms (e.g., decision trees) to improve prediction estimations. The training process of AdaBoost follows a sequential process. The model will firstly train a weak learner on the training dataset, and wrong predictions will be highly weighted in the successive models. Following this sequential process, the AdaBoost can converge to a strong learner even if the individual learners can be weak. The team employs multi-class AdaBoosted Decision Trees (Hastie et al., 2009) to train the model, where the maximum number of estimators is set as 100, at which boosting is terminated.

5.1.2 Model Evaluation

To evaluate the performance of different machine learning models, the team used various metrics, including the Area Under the Receiver Operating Characteristics - ROC, Accuracy (ACC), Prediction Rate (PR), and False Alarm Rate (FAR). The selection of these metrics was based on previous studies that modeled the occurrence of traffic incidents (Li et al., 2017; Gakis et al., 2014).

- ROC provides an aggregate measure of the model performance accords all classification thresholds. The ROC value is calculated to show the probability that a randomly chosen positive example (e.g., crash observation) is ranked higher than a randomly chosen negative example (e.g., crash-free observation). A higher AUC value indicates greater model prediction power.

$$AUC = Area\ under\ the\ ROC\ Curve \times 100\% \quad (12)$$

- ACC represents the percentage of all correct classified samples (both crash and crash-free observations) to the number of total training samples.

$$ACC = \frac{Number\ of\ correct\ classified\ samples}{Number\ of\ total\ training\ samples} \times 100\% \quad (13)$$

- PR is defined as the ratio of the number of correctly predicted crash samples to the number of total crash samples.

$$PR = \frac{Number\ of\ correctly\ predicted\ crash\ samples}{Number\ of\ observed\ crash\ samples} \times 100\% \quad (14)$$

- FAR is the proportion of the crash-free samples predicted as crash samples to the number of total crash-free samples.

$$FAR = \frac{Number\ of\ crash-free\ samples\ predicted\ as\ crash}{Number\ of\ observed\ crash-free\ samples} \times 100\% \quad (15)$$

5.1.3 Model Interpretation

Compared with traditional statistical models (e.g., logistic regression), machine learning models are often associated with an improved performance in terms of the model prediction accuracy. However, frequently referred to as a "black box" , machine learning models are usually criticized for lacking interpretability (Zhao et al., 2020). Researchers have undertaken significant efforts to open the "black box" and interpret machine learning models (Molnar, 2020). The team took advantage of two commonly used machine learning tools – Permutation Feature Importance and Partial Dependence Plot to interpret the RF, SVM, and XGBoost modeling results.

Permutation Feature Importance

Permutation Feature Importance measures the importance of a feature or variable by calculating the increase in a model's prediction error after permuting the feature (Breiman, 2001; Fisher et al., 2019). The team used the a Python package called *sklearn.inspection* to compute the importance scores for feature variables (Fabian et al., 2011). The main computation steps include (Molnar, 2020):

1. Estimate the original model error $e^{orig} = L\big(y, f(X)\big)$ (e.g., mean squared error);
2. For each feature ($i = 1, 2, 3, \dots, n$), create a feature matrix X^{perm} by permuting feature i in the data X;
3. Estimate error $e^{perm} = L\big(Y, f(X^{perm})\big)$ based on the predictions of the permuted data;
4. Calculate permutation feature importance $FI^i = e^{perm}/e^{orig}$;
5. Sort features by descending FI.

Partial Dependence Plot

Permutation Feature Importance shows the importance of variables and offers limited information regarding the direction and magnitude of relationships between factors. To reveal the detailed relationship between crash risk and traffic dynamic variables, the team generated Partial Dependence Plots (PDPs) based on modeling results. The PDPs can be used to show how the crash

risk or the probability of having a crash would increase or decrease with the changes in the input feature or independent variable. The same Python package *sklearn.inspection* was used to calculate partial dependence values.

5.2 Deep Learning

In previous research, various statistical models (e.g., ARIMA models, logistic regression models, etc.) and machining learning models (e.g., support vector machine, random forest model etc.) have been used to detect the occurrence of incidents automatically. The input features of these models are usually summarized traffic flow variables (e.g., mean speed, mean volume) created based on detector data. This paper uses real-time speed information provided by probe vehicle data to build AID models to detect the occurrence of incidents. Unlike detector-based data only available with road detectors, probe vehicle data can provide speed information covering a wide spatial range. However, one significant drawback of probe vehicle data is that it cannot provide direct volume information and occupancy information. To avoid information loss, high-resolution spatial-temporal speed matrices are fed into the deep learning models directly without creating summarized speed dynamic variables (e.g., speed mean, speed variance, etc.). Two deep learning models, including Artificial Neural Network (ANN) and Convolution Neural Network (CNN) are used to detect the occurrence of the incident. The occurrence of different incident sub-types may have different impacts on the traffic dynamics. In other words, by classifying different traffic patterns, not only the occurrence of an incident can be detected, but also the incident subtype might be identified. Therefore, two kinds of models are developed in this project, including general AID models and incident subtype classification models.

5.2.1 Artificial Neural Network

Artificial Neural Networks (ANN) are algorithms functioning like the brain and can be used to model complicated patterns and make predictions. A typical ANN comprises three parts in its structure: one input layer, one or more hidden layers, and one output layer. As shown in FIGURE 15, a layer contains several nodes (also called neurons), and layers beside each other are fully connected. In other words, every node in one layer is connected to every other node in the next layer. The key components/technologies in ANN are summarized below.

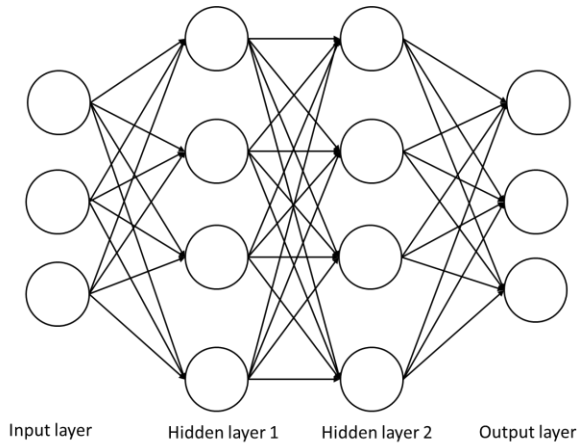


FIGURE 15 Architecture of Artificial Neural Network (ANN)

Node

Each given node in a hidden layer or an output layer can be served as a computation unit. As shown in FIGURE 16, a node first takes the summation of weighted inputs and bias and passes it through a non-linear activation function.

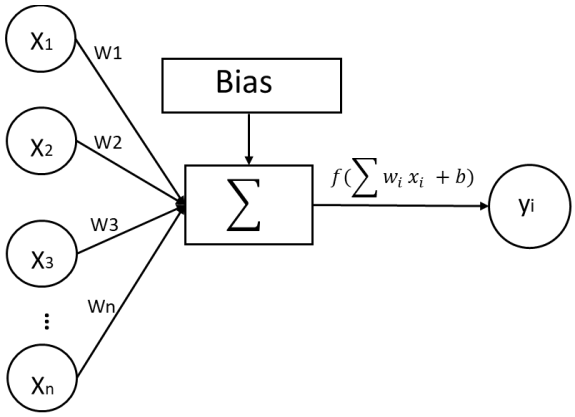


FIGURE 16 A diagram of a node in NN

The mathematic equation for node *h* can be expressed as

$$y_i = f(w_1x_1 + w_2x_2 + \cdots + w_nx_n + b) \tag{16}$$

where, $w_1, w_2, \ldots, w_n$ are weights for input $x_1, x_2, \ldots, x_n$, respectively; *b* is the bias for node *h*; *f()* is the non-linear activation function. The non-linear activation function takes any values as input and outputs a value in the range of 0 to 1. For classification problems, the commonly used activation functions include Sigmoid functions (shown in **Equation 17**), Tanh function (shown in **Equation 18**), Relu function (shown in **Equation 19**) and Exponential Linear Units (ELUs) Function (shown in **Equation 20**).

$$f(x) = \frac{1}{1+e^{-x}} \tag{17}$$

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{18}$$

$$f(x) = \max(0, x) \tag{19}$$

$$f(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^{x}-1) & x < 0 \end{cases} \tag{20}$$

Different activation functions may lead to different prediction results. In practice, the choice of activation function tends to be experimental and needs to be tested. For the last layer of ANN, the choice of activation function should consider the number of labels for classification. In a binary classification problem, the Sigmoid function is often used as the activation function in the last layer. For the multi-class classification problem, the SoftMax function often normalizes the results into probabilities that have the summation of 1. The SoftMax function returns the probability of

each class. As shown in **Equation 21**, the probability of being labeled *i* in a M labels classification problem is:

$$P(y = i) = \frac{\exp(z_i)}{\sum_1^M \exp(z_j)} \quad (21)$$

where, z_i represents the values from the nodes in the output layer.

Loss function

The process of training the neural network is to update the weights to minimize the loss function. Cross-entropy-based loss functions are commonly used in classification problems. In binary classification, the binary cross-entropy is presented as **Equation 22**:

$$L = -(y_i \log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)) \quad (22)$$

where, y_i represents the expected outcome, and $\hat{y}_i$ represents the outcome produced by the model. In a multi-class classification problem, the categorical cross-entropy for M classes is shown in **Equation 23**:

$$L = \sum_{j=1}^{M} y_j \log(\hat{y}_j) \quad (23)$$

Back-Propagation

Back-propagation (Backward propagation of errors) algorithm aims to train artificial neural networks by iteratively updating the weights to minimize the loss function. In other words, the method calculates the gradient of the loss function with respect to all the weights in the network so that the gradient is fed to the gradient descent method, and the weights are updated to minimize the loss function.

5.2.2 Convolutional Neural Network (CNN)

A Convolutional neural network (CNN) architecture consists of three types of layers: convolutional layers, pooling layers, and fully connected (FC) layers, as shown in FIGURE 17 below. The convolutional layer and pooling layers are used to extract features in images, and the FC layers take the output from the previous convolutional layer and predict the class of the image. Compared with the ANN model, CNN has the ability to capture the local connectivity of the image and downscale the image dimension at the same time.

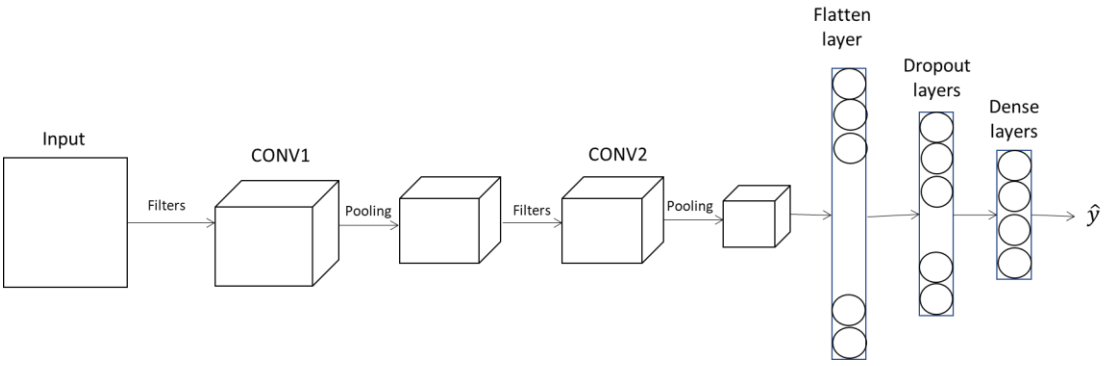


FIGURE 17 A diagram of CNN architecture

Convolutional Layer
The input of this layer is the pixel matrix converted from the raw image. Then, the mathematical operation of convolution is performed between the input matrix and several filters. By sliding the filter over the input matrix, the dot product is taken between the filter and the parts of the input image within the size of the filter. The output is the Feature map that contains the image information, such as the corners and edges.

Pooling Layer
Typically, a Convolutional Layer is followed by a Pooling Layer. The Pooling Layer usually serves as a bridge between the Convolutional Layer and the FC Layer. It aims to decrease the size of the feature map from the previous step to reduce computational costs. There are several types of Pooling operations. Max pooling is the most commonly used one. In Max Pooling, the largest element is taken from the feature map.

Fully Connected Layer/Dense layer
The Fully Connected (FC) /Dense layer is similar to the FC layer in ANN model introduced in the previous section. It consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers. The flattened outcome from previous layers is fed to the FC layers to make the classification.

6. Crash-risk Prediction

6.1 Variable Creation

The team created variables to describe the pre-crash traffic speed characteristics, including the speed reductions and the speed variations within certain times before the crash. The speed-related variables were generated based on the spatial-temporal speed matrices (0.1 miles × 1 minute). Note that the HERE database reports the speeds for each TMC every minute; therefore, the speed could change from one matrix cell to another if the time is different. For a 20-minute time interval, the speed would technically change 20 times. The team created variables to capture the average speed reductions, the variation of speed reductions, and the speed reductions at different percentiles (minimum, 25th, 75th, and maximum). The speed reductions are calculated for the one-minute period and summarized over the 20-minute interval. In other words, in every 20-minute interval, there are 19-speed reduction values to be summarized by their mean, variance, and percentile values. The goal of having these variables is to describe the pre-crash speed curve at or around the crash location. TABLE 9 summarizes the variables created to capture the speed dynamics from various aspects.

It is important to note that the traffic dynamic variables were created for three different pre-crash time points at 10, 15, and 20 minutes (before the event of a crash) separately. In addition, to reflect the variation of data reporting TMCs whose spatial extent could change from time to time, the team created the variables to indicate the mean and variance of the number of dynamic sub-TMCs within a 20-minute time interval at different pre-crash time points.

Besides the traffic dynamics before crashes, the team also extracted the traffic speed information for crash-free conditions paired with the crash conditions. Specifically, for each crash record, one crash-free record was generated at the same location and the same time of day but on a different day when there was no crash. For the crash-free records, the same sets of variables were created through the same methods of data linkage and variable creation. Note that crash-free observations can be created for any time and location where no crashes occurred. The team covers the entire freeway network in Alabama; there would be an extremely large number of crash-free observations if created for any time and location where no crash occurred. Besides, machine learning models often require balanced data; therefore, in this project, for each crash observation, one crash-free observation was created for the same location but at a different time when there was no crash. Further, the way of creating crash-free observations at the same locations could ensure that drivers would face the same roadway environments but different traffic contexts at various times, which supports the aim of the team to explore the relationships between crash risk and traffic dynamics.

TABLE 9 also includes road environment variables extracted from the HPMS road infrastructure database. These variables include the log-transformation of Annual Average Daily Traffic (AADT), the number of through lanes in one direction, and land use type. Hossain & Muromachi (2012) suggested that the conditions near ramp areas are substantially different from basic freeway segments and may affect crash risk differently. The distance to the nearest upstream ramp was obtained based on the road shapefile in the HPMS database. Further, the team pulled the variables of the direction of traffic, the time of day, and the day of the week from the CARE crash database. In total, 17,423 records were generated for modeling, of which 8,688 are crash records (labeled as 1) and 8,736 are crash-free records (labeled as 0). Notably, the number of crash records is not

exactly equal to the number of crash-free records, because some observations with missing information (e.g., HERE data were unavailable at some locations/times) were deleted.

TABLE 9 Variables for the models

Traffic Dynamic Variables (Data source: HERE database)*		
Variables	**Description**	**Equation**
spmean**t**_5m_up	Mean pre-crash traffic speed (over a 20-min interval) for the 5 miles upstream.	$\text{spmean}\mathbf{t}\text{\_5m\_up} = \frac{\sum_{j=-t-10}^{10-t}\sum_{i=-50}^{0} v_{ij}}{1000}$
spvc**t**_5m_up	Variance of pre-crash speed (over a 20-min interval) for the 5 miles upstream.	$\text{spvc}\mathbf{t}\text{\_5m\_up} = \frac{\sum_{j=-t-10}^{10-t}\sum_{i=-50}^{0} (v_{ij} - \text{spmean}\mathbf{t}\text{\_5m\_up})^2}{1000}$
spmean**t**_5m_dn	Mean one-minute pre-crash traffic speed (over a 20-min interval) for the 5 miles downstream.	$\text{spmean}\mathbf{t}\text{\_5m\_dn} = \frac{\sum_{j=-t-10}^{10-t}\sum_{i=0}^{50} v_{ij}}{1000}$
spvc**t**_5m_dn	Variance of pre-crash speed (over a 20-min interval) for the 5 miles downstream.	$\text{spvc}\mathbf{t}\text{\_5m\_dn} = \frac{\sum_{j=-t-10}^{10-t}\sum_{i=0}^{50} (v_{ij} - \text{spmean}\mathbf{t}\text{\_5m\_dn})^2}{1000}$
tmaxspddrop**t**	The pre-crash time point for the max one-minute speed reduction over a 20-min interval at the crash location (from the same column of the spatial-temporal matrix)	$\text{tmaxspddrop}\mathbf{t} = \text{t}_j(\text{ when } \max(v_{ij-1} - v_{ij}))$ $(i = 0, -t - 10 \le j \le 10 - t)$
spmean**t**_5m_df	The difference of mean pre-crash speeds between 5 miles upstream and downstream. It is equal to *spmean**t**_5m_up - spmean**t**_5m_dn.*	$\text{spmean}\mathbf{t}\text{\_5m\_df} = \text{spmean}\mathbf{t}\text{\_5m\_up} - \text{spmean}\mathbf{t}\text{\_5m\_dn}$
spvc**t**_5m_df	The difference of pre-crash speed variances between 5 miles upstream and downstream. It is equal to *spvc**t**_5m_up - spvc**t**_5m_dn*.	$\text{spvc}\mathbf{t}\text{\_5m\_df} = \text{spvc}\mathbf{t}\text{\_5m\_up} - \text{spvc}\mathbf{t}\text{\_5m\_dn}$
spddropmean**t**	Mean one-minute speed reduction (over a 20-min interval) at the crash location (from the same column of the spatial-temporal matrix)	$\text{spddropmean}\mathbf{t} = \frac{\sum_{j=-t-10}^{10-t}(v_{ij-1} - v_{ij})}{20}(i = 0)$
spddropvc**t**	Variance of speed reduction (over a 20-min interval) at the crash location (from the same column of the spatial-temporal matrix)	$\text{spddropvc}\mathbf{t} = \frac{\sum_{j=-t-10}^{10-t}\left((v_{ij-1} - v_{ij}) - \text{spddropmean}\mathbf{t}\right)^2}{20}(i = 0)$
spddropmax**t**	Maximum one-minute speed reduction (over a 20-min interval) at the crash location (from the same column of the spatial-temporal matrix)	$\text{spddropmax}\mathbf{t} = \max(v_{ij-1} - v_{ij})$ $(i = 0, -t - 10 \le j \le 10 - t)$
spddropup**t**	75th percentile speed reduction (over a 20-min interval) at the crash location	$\text{spddropup}\mathbf{t} = \text{Q}_3(v_{ij-1} - v_{ij})$ $(i = 0, -t - 10 \le j \le 10 - t)$

	(from the same column of the spatial-temporal matrix)	
spddroplow**t**	25th percentile speed reduction (over a 20-min interval) at the crash location (from the same column of the spatial-temporal matrix)	$\text{spddroplow}\mathbf{t} = \text{Q}_1\left(v_{ij-1} - v_{ij}\right)$ $(i = 0, -t - 10 \le j \le 10 - t)$
spddropmin**t**	Minimum one-minute speed reduction (over a 20-min interval) at the crash location (from the same column of the spatial-temporal matrix)	$\text{spddropmin}\mathbf{t} = min\left(v_{ij-1} - v_{ij}\right)$ $(i = 0, -t - 10 \le j \le 10 - t)$
Tmcave**t**	Mean of the number of dynamic sub-TMCs /min over a 20-min interval	$\text{Tmcave}\mathbf{t} = \frac{\sum_{j=-t-9}^{10-t} N_j(TMC)}{20}$
Tmcvar**t**	Variance of the number of dynamic sub-TMCs/min over a 20-min interval	$\text{Tmcvar}\mathbf{t} = \frac{\sum_{j=-t-9}^{10-t} (N_j(TMC) - \text{Tmcave}\mathbf{t})^2}{20}$
Other Variables (Data source: CARE database and HPMS database)		
Variables	***Description***	
Logaadt	Logarithm of AADT	
Through_La	Number of through lanes	
NearestDist	Distance to the nearest upstream ramp	
UrbanRural	Land use type (Base: Rural)	
Direction	Northbound/Southbound/Westbound/Eastbound (Base: Eastbound)	
Timeind	a.m. peak (06:00 to 10:00)/midday (10:00 to 16:00)/p.m. peak (16:00 to 20:00)/night (20:00 to 06:00) (Base: a.m. peak)	
Weekday	Weekday/Weekends (Base: Weekends = 0)	

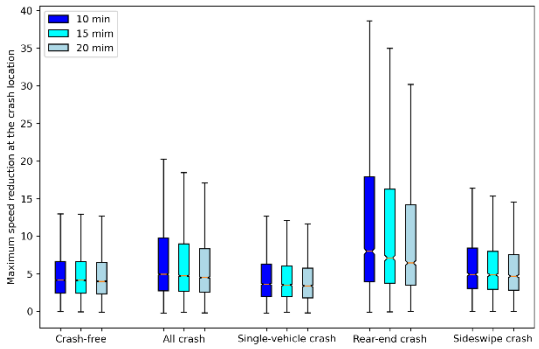
Notes:
(1) * The *Traffic Dynamic Variables* were created for three different pre-crash time points at 10, 15, and 20 minutes (prior to the event of a crash) separately. In the variable names, the letter "**t**" represents the pre-crash time points, and is replaced by the number "10", "15", and "20" respectively in the datasets for modeling. (2) v_{ij} means the speed in location i at pre-crash time point j , and its value equals to the value in cell ij in the spatial-temporal speed matrix. (3) $N_j(TMC)$ denotes the total number of dynamic TMCs both upstream and downstream at pre-crash time point j.

6.2 Results of Statewide Model

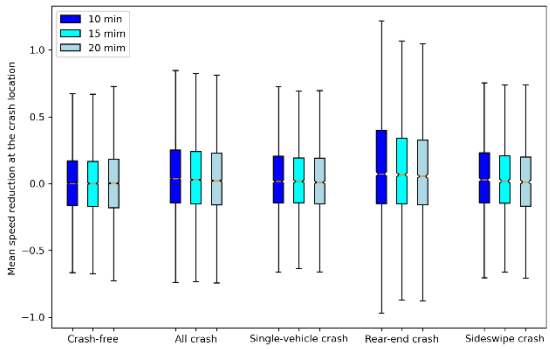
6.2.1 Descriptive Statistics

FIGURE 18 shows the distribution of four selected traffic dynamics independent variables for both crash and crash-free conditions. For crash conditions, separated distributions were shown for all crashes, single-vehicle crashes, sideswipe crashes, and rear-end crashes. In each condition, three boxplots were drawn to show the distribution of selected traffic dynamics variables among three different pre-crash time cases (10, 15, and 20 minutes before the crash). In general, boxplots of rear-end crash groups are comparatively tall in all four plots, especially compared with boxplots of the crash-free group, meaning that rear-end crashes are related to greater speed variations, as demonstrated by wider distributions of speed reductions (as shown in FIGURE 18 (i) and FIGURE 18 (ii)) and speed variances (as shown in FIGURE 18 (iii) and FIGURE 18 (iv)). In addition, the medians of boxplots in the rear-end crash group are greater than the median of boxplots in other groups, indicating that rear-end crashes are more likely to be related to higher speed reduction (as shown in FIGURE 18 (i) and FIGURE 18 (ii)) and speed variances (as shown in FIGURE 18 (iii) and FIGURE 18 (iv)). From a time perspective, wider distributions are shown in times cases closer

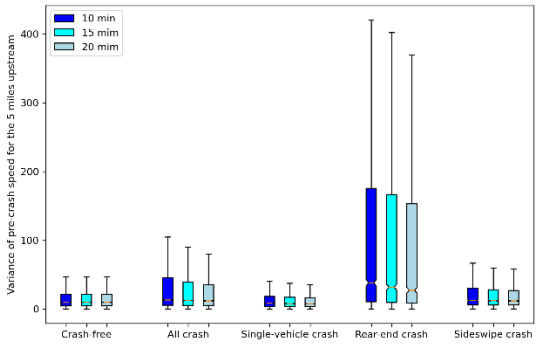
to the occurrence of rear-end crashes, indicating that the speed closer to the event of rear-end crashes changes more significantly.



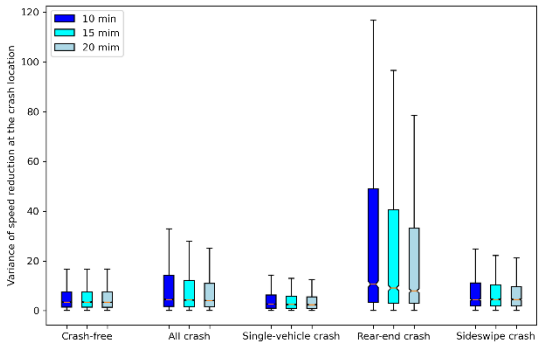
(i) Maximum one-minute speed reduction (over a 20-min interval) at the crash location



(ii) Mean one-minute speed reduction (over a 20-min interval) at the crash location



(iii) Variance of pre-crash speed (over a 20-min interval) for the 5 miles upstream



(iv) Variance of speed reduction (over a 20-min interval) at the crash location

FIGURE 18 Distributions of selected traffic dynamics variables

6.2.2 All Crash Models

TABLE 10 summarizes the performance of all crash models that predict the risk for any type of crash. In general, the performance metrics indicate that all models have limited power for crash risk prediction. The models that took the 10-min pre-crash traffic dynamics as inputs appeared to have a better performance than other models, and the models taking the 15-min pre-crash data were better than the models with the 20-min pre-crash information. The results indicate that the traffic dynamics closer to the event of a crash are more predictive of the crash risk.

The best performance model using the 10-min pre-crash data had an AUC of 61.7% and an accuracy (ACC) of 58.2%. Some previous studies reported crash risk prediction models with an accuracy higher than 80% (Hossain et al., 2019). The low accuracy reported in this project may be explained by the data aspect. Previous studies developed models using traffic data from a limited environment such as one segment or corridor (e.g., I-80 highway in California, I-235 in Des Moines, IA) (Huang et al., 2020; Lin et al., 2020). The traffic data used in this project covered the entire freeway network in Alabama. In addition to the traffic dynamics, many other factors could

also have a relationship with the crash risk. These factors may include the road environment, driver population, and vehicle fleets, and they could vary significantly across geographic areas. The models developed in this project include limited road environment variables and no variables for driver population and vehicle fleets, leading to the issue of unobserved heterogeneity (Abdelrahman et al., 2017). When data used for modeling are collected from a limited environment, the issue of unobserved heterogeneity may be avoided as all observations are from the same environment and share the impacts of unobserved factors on modeling outcomes. Further, the data quality could also affect the modeling performance. The team used the HERE crowdsourced probe vehicle data, which is different from the traffic data collected through loop detectors or other sensors fixed on the road or roadside. Unlike loop detector data, the HERE data generated from probe vehicles does not cover the entire fleet and does not provide information about traffic volumes at the time of or before the event of a crash (Anuar et al., 2015). Models in this project relied on traffic speed records, which are aggregated from probe vehicles' speeds. It is known that not all road segments can be covered by enough probe vehicles due to the high mobility and a limited number of probe vehicles (Naranjo et al., 2012). Historical traffic speeds are reported for these segments (TMC) at this moment. There may be other data issues such as misreporting and delay in data transmission through wireless communication (Zhang et al., 2013). The overall data inaccuracy in the HERE data remains unknown to the authors.

In addition to the data-related reasons, another reason may be that crashes with several types have a different relationship with the traffic dynamics. For example, a rear-end crash occurs when a driver is following too closely to the car in front of him/her, which is likely to happen in congested traffic (Chatterjee & Davis, 2016). Single-vehicle crashes such as run-off-road crashes are more likely to occur in free-flow traffic conditions (Liu & Ye, 2011). Having all crashes in one model may lead to an unclear relationship between crash risk and traffic dynamics, as indicated by the poor model prediction performance.

TABLE 10 Model performance on crash risk prediction for all crash types

Pre-crash time point	Model	Parameter setting	ACC	PR	FAR	AUC
10 min	Logistic model	thresholds = 0.473	57.9%	55.7%	39.9%	60.6%
	RF	max_features='sqrt', n_estimators=494, thresholds = 0.489	58.2%	54.4%	38.0%	61.7%
	XGBoost	learning_rate=0.05, n_estimators=100, thresholds = 0.468	57.1%	55.0%	40.6%	59.8%
	SVM	C=1, gamma=0.01, kernel = "RBF", thresholds = 0.461	58.3%	57.8%	41.1%	60.9%
15 min	Logistic model	thresholds = 0.474	56.0%	54.7%	42.6%	58.5%
	Random Forest model	max_features='sqrt', n_estimators=494, thresholds = 0.491	55.9%	49.8%	37.9%	59.5%

	XGBoost	learning_rate=0.05, n_estimators=100, thresholds = 0.476	55.9%	51.9%	39.8%	58.8%
	SVM	C=0.1, gamma=0.1, kernel = "RBF", thresholds = 0.467	56.0%	52.1%	40.0%	58.0%
20 min	Logistic model	thresholds = 0.476	55.3%	53.0%	42.3%	57.4%
	RF	max_features='sqrt', n_estimators=494, thresholds = 0.489	55.7%	50.8%	39.1%	58.8%
	XGBoost	learning_rate=0.1, n_estimators=100, thresholds = 0.476	55.7%	55.3%	43.8%	58.0%
	SVM	C=100, gamma=0.001, kernel = "RBF", thresholds = 0.469	55.5%	53.0%	42.0%	57.8%

Notes: “n_estimators” denotes the number of decision trees in the forest; “max_features” denotes the number of features is considered by each tree when splitting a node; “thresholds” denotes the thresholds for distinguishing crash observations from crash-free observations; learning_rate denotes the weighting of new trees added to the XGBoost model.

6.2.3 Models for Single-Vehicle, Sideswipe Crash and Rear-End Crashes

Considering that crashes of distinct types may have a different relationship with the traffic dynamics, the team estimated separate models for three major crash types: single-vehicle, sideswipe, and rear-end crashes. As indicated above, the traffic dynamics closer to the event of a crash are more predictive of the crash risk. TABLE 11 presents the models based on the 10-min pre-crash traffic dynamics. In terms of all model evaluation metrics, the models for rear-end crashes have better prediction performance than other models. It implies that rear-end crashes appear to be more predictable than the other two types of crashes if the crash risk models are developed based on the crowdsourced probe vehicle data. For rear-end crashes, in comparison with the logistic regression model, XGBoost, and SVM model, the RF model had a slightly improved prediction performance according to the ROC value (68.4%) and the Prediction Rate (PR) (52.6%).

TABLE 11Model performance on crash risk prediction for different crash types

Crash type	Model	Parameter setting	ACC	PR	FAR	AUC
Single-vehicle crash	Logistic model	thresholds = 0.488	54.1%	55.3%	47.1%	54.3%
	RF	n_estimators=216, thresholds = 0.472	53.9%	61.0%	52.8%	56.0%
	XGBoost	learning_rate=0.05, n_estimators=100, thresholds = 0.487	53.5%	47.0%	40.4%	52.9%
	SVM	C=10, gamma=0.001, thresholds = 0.507	53.8%	44.8%	37.9%	53.8%

Sideswipe crash	Logistic model	thresholds = 0.490	54.2%	51.8%	43.3%	55.8%
	RF	n_estimators=50, threholds = 0.48	47.2%	47.6%	53.2%	45.8%
	XGBoost	learning_rate=0.1, n_estimators=100, thresholds = 0.487	46.4%	57.2%	64.3%	45.3%
	SVM	C=10, gamma=0.001, thresholds = 0.491	55.8%	52.1%	40.4%	55.8%
Rear-end Crash	Logistic model	thresholds = 0.434	64.1%	67.0%	38.7%	67.4%
	RF	max_features='sqrt', n_estimators=494, thresholds = 0.540	65.1%	52.6%	23.0%	68.4%
	XGBoost	learning_rate=0.05, n_estimators=100, thresholds = 0.485	63.1%	57.4%	31.4%	66.3%
	SVM	C=10, gamma=0.01, thresholds = 0.507	64.8%	52.9%	23.9%	65.9%

Notes: “n_estimators” denotes the number of decision trees in the forest; “max_features” denotes the number of features is considered by each tree when splitting a node, and “thresholds” denotes the thresholds for distinguishing crash observations from crash-free observations; learning_rate denotes the weighting of new trees added to the XGBoost model.

6.2.4 Model Interpretation

For rear-end crash models, the team calculated the permutation feature importance for each variable and computed the partial dependence for pre-crash speed-related variables that are top-ranked according to the feature importance. FIGURE 19 shows the ranking of the top 20 variables for the RF model, according to the permutation feature importance. The variable rankings for logistic regression, XGBoost model, and SVM models are also shown in FIGURE 9 as a comparison. Though the rankings are different across the four models, the pre-crash speed-related variables are in general, ranked high in all four models.

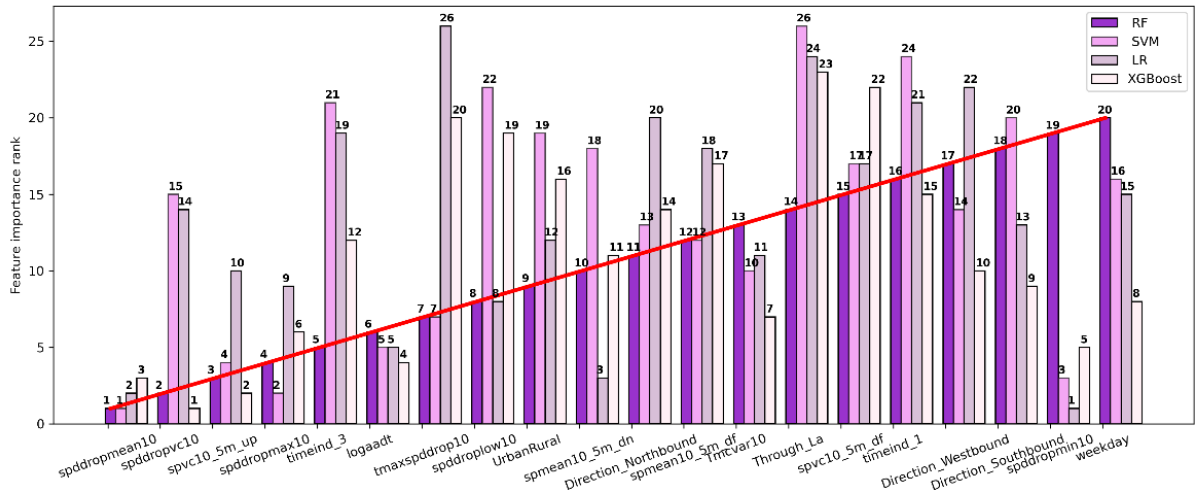
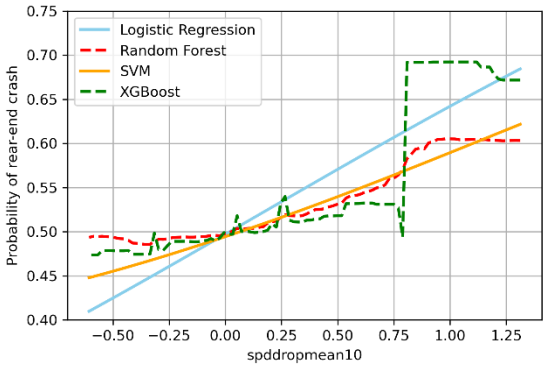


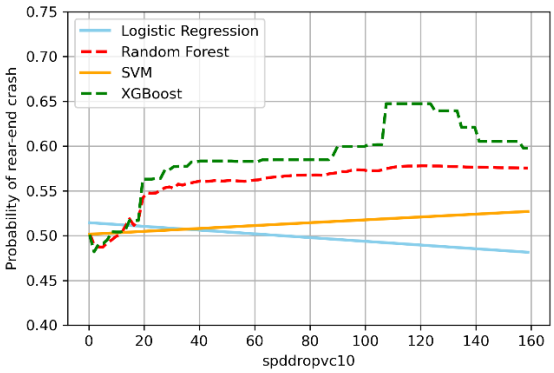
FIGURE 19 Ranking of permutation variable importance for RF, SVM, logistic regression (LR), and XGBoost for rear-end crashes

FIGURE 20 presents the partial dependence plots for the top six pre-crash speed-related variables according to the RF model's permutation feature importance. The plots show forecasted risks for rear-end crashes by varying one particular independent variable (e.g., mean one-minute speed reduction at the crash location) and holding other variables constant. Overall, similar relationships between independent variables and crash risks are found in four modeling approaches - logistic regression, RF, XGBoost, and SVM. Besides the similarities, the differences are also evident. The results of logistic regression and SVM models show nearly linear relationships between crash risk and associated traffic variables, while the tree-based models (RF and XGBoost) uncovered nonlinear relationships.

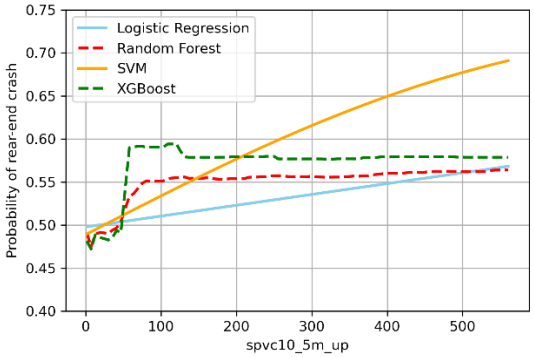
FIGURE 20 (i) shows the relationships between the rear-end crash risk and mean one-minute speed reductions. Greater speed reductions are associated with higher crash risks and a significantly higher crash risk is found for the mean one-minute speed reduction greater than 0.75 mph per minute, according to RF and XGBoost models. FIGURE 20 (iv) shows a similar trend for the maximum one-minute speed reduction. As shown in FIGURE 20 (ii), the variance of one-minute speed reduction also has a significant relationship with rear-end crash risks. Higher crash risks seem to be linked to greater variances. According to RF and XGBoost models, the risks would increase significantly when the variance reaches 20. After this point, the risks appear to be relatively constant. FIGURE 20 (iii) shows the relationship between crash risk and variance of pre-crash speeds for upstream traffic within 5 miles. Similarly, higher crash risks are related to greater speed variances. The non-linear relationships revealed by RF and XGBoost models show the crash risk would increase significantly when the variance reaches around 55 and stay stabilized afterward. FIGURE 20 (v) shows a positive relationship between crash risk and the pre-crash time point for the max 1-min speed reduction over a 20-min interval at the crash location. It means that a higher rear-end crash risk is associated with a closer pre-crash time point for the max 1-min speed reduction over a 20-min interval at the crash location. FIGURE 20 (vi) shows the relationship between the crash risk and the 25th percentile speed reduction. The relationship is relatively stationary across different values of this independent variable, meaning a potentially insignificant relationship.



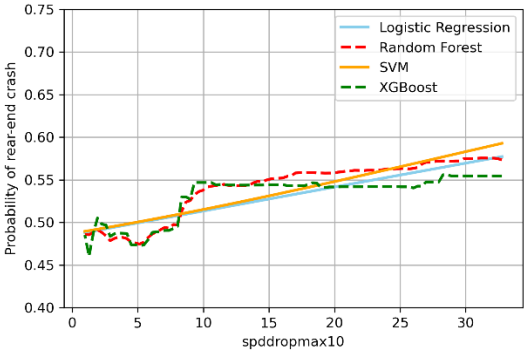
(i) Mean one-minute speed reduction (over 20-min intervals) at the crash location



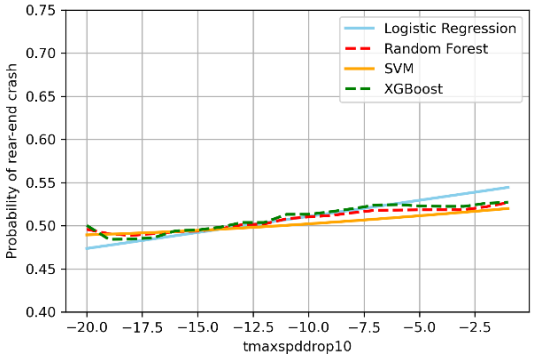
(ii) Variance of one-minute speed reduction (over a 20-min interval) at the crash location



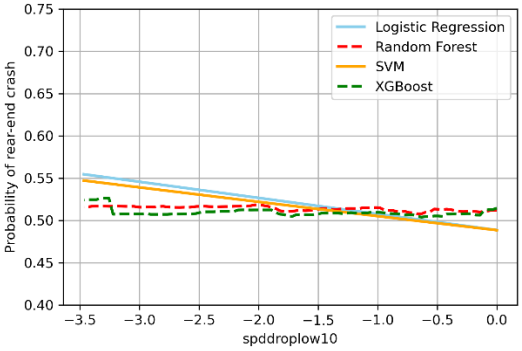
(iii) Variance of pre-crash speed (over a 20-min interval) for the 5 miles upstream.



(iv) Maximum one-minute speed reduction (over a 20-min interval) at the crash location



(v) The pre-crash time point for the max one-minute speed reduction over a 20-min interval at the crash location



(vi) 25th percentile one-minute speed reduction (over a 20-min interval) at the crash location

FIGURE 20 Partial dependence plots of top-six variables in the RF model for rear-end crash

6.3 Results of High Crash Density Freeway Segments Model

6.3.1 High Crash-Density Segments Identification

As shown in TABLE 12, the top three crash subtypes are single-vehicle crashes, rear-end crashes, and sideswipe crashes, accounting for 35.6%, 34.7%, and 18.3% of the total crashes. As demonstrated in our previous study (Zhang et al., 2022), rear-end crashes are more predictable by pre-crash traffic dynamics among other crash subtypes. Therefore, in this weekly report, we built separate models for all types of crashes, single-vehicle crashes, rear-end crashes, and sideswipe crashes under different levels of crash density.

TABLE 12 Frequency of the crash subtypes

Manner of Crash	Frequency	Percentage
Single Vehicle Crash (all types)	3558	35.6%
Rear End (front to rear)	3465	34.7%
Sideswipe - Same Direction	1830	18.3%
Other	421	4.2%
Side Impact (angled)	272	2.7%
Non-Collision	150	1.5%
Angle (front to side) Same Direction	147	1.5%
Head-On (front to front only)	55	0.6%
Side Impact (90 degrees)	37	0.4%
Angle Oncoming (frontal)	24	0.2%
Unknown	13	0.1%
Sideswipe - Opposite Direction	12	0.1%
Angle (front to side) Opposite Direction	10	0.1%
Causal Veh Backing: Rear to Rear	2	0.0%
Causal Veh Backing: Rear to Side	1	0.0%

To identify the high-frequency crash sites, the crashes were aggregated at the freeway segment level based on the recorded mile marker in the CARE crash report. That is, we counted the all types of crash frequency, rear-end crash frequency, sideswipe crash frequency, and single-vehicle crash frequency for every mile of the selected freeway. TABLE 13 shows the distribution of the crash frequencies aggregated at the mile level of the freeway segment. A 1-mile segment can cause up to 52 crashes. The 80th percentile of all-type crash density is 8 and the 50th percentile of all-type crash density is 4. For rear-end crashes, the highest density can be up to 37 crashes/mile. The 80th percentile of rear-end crash density is 3 and the 50th percentile of rear-end crash density is 1. For sideswipe crashes and single-vehicle crashes, the highest segment-level densities are 21 and 27, respectively. FIGURE 1 shows the violin plot of crash frequency by crash types. The shape of violin plots demonstrates that some sites are associated with high crash density, while the crash frequency of the majority of the freeway sites is around the medians.

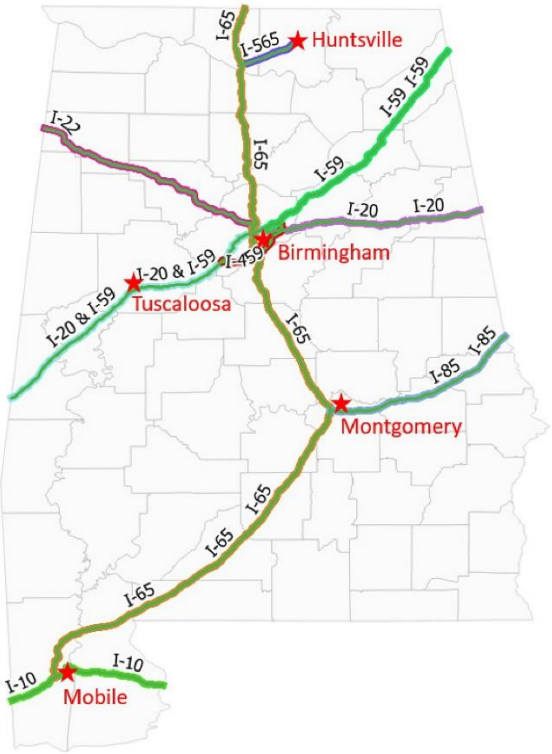


FIGURE 21 Selected freeway within the boundary of the state of Alabama

TABLE 13 Summary statistics for the crash frequencies count at a 1-mile level

Terms	All crash density (crash/mile)	Rear-end crash density (crash/mile)	Sideswipe crash density (crash/mile)	Single vehicle crash density (crash/mile)
Mean	6	2	1	2
Standard deviation	6	3	2	2
Maximum	52	37	21	27
90th percentile	11.2	4	3	4
80th percentile	8	3	2	3
70th percentile	6	2	1	2
60th percentile	5	1	1	2
50th percentile	4	1	1	2
40th percentile	3	1	0	1
30th percentile	2	0	0	1
20th percentile	2	0	0	1
10th percentile	1	0	0	0
Minimum	1	0	0	0

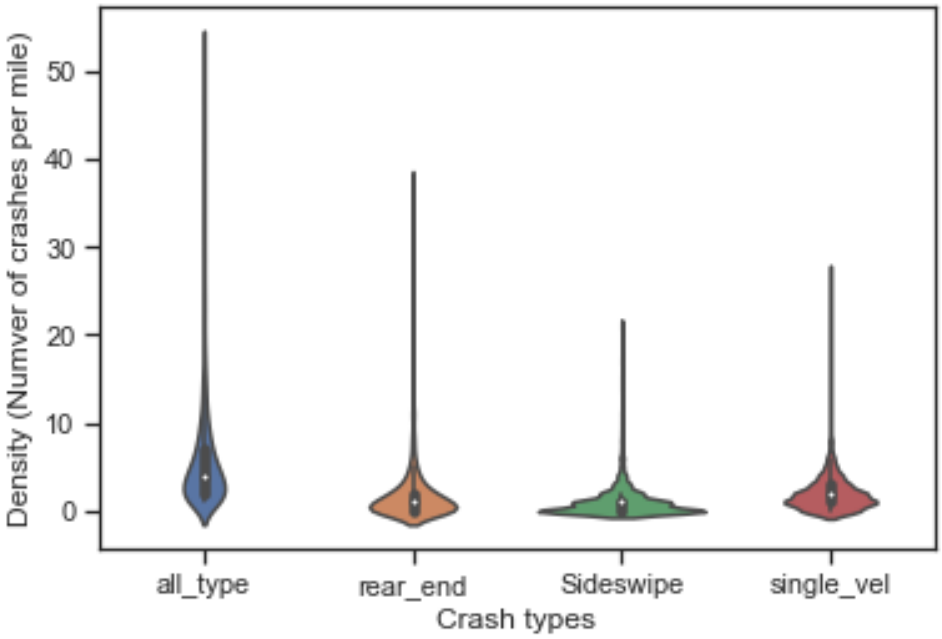


FIGURE 22 Violin plot of crash density

TABLE 14 shows the top 20 crash frequency sites for a selected freeway in Alabama, categorized by all types of crashes, sideswipe crashes, rear-end crashes, and single-vehicle crashes. Note that the 'site_identifier' column in TABLE 14 comprises three parts: road name, road direction, and mile marker. For example, 'I-65_Northbound_259' refers to the number of crashes that occurred between mile markers 259 and 260 on the northbound side of I-65. The range of crash frequencies for the top 20 crash frequency sites is 29-52 for all types of crashes, 7-21 for sideswipe crashes, 19-37 for rear-end crashes, and 8-27 for single-vehicle crashes.

TABLE 14 Top 20 crash frequency sites of selected freeway in Alabama by crash subtypes

(i) All types of crashes

Rank	Site_identifier	Density	Rank	Site_identifier	Density
1	I-65_Northbound_259	52	11	I-459_Southbound_16	33
2	I-20/I-59_Eastbound_76	47	12	I-65_Northbound_4	33
3	I-65_Northbound_260	45	13	I-459_Southbound_17	32
4	I-20/I-59_Westbound_126	44	14	I-20/I-59_Eastbound_124	31
5	I-65_Northbound_252	41	15	I-459_Northbound_28	31
6	I-459_Southbound_19	41	16	I-20/I-59_Westbound_81	30
7	I-65_Southbound_259	40	17	I-10_Westbound_26	30
8	I-20/I-59_Eastbound_123	39	18	I-20/I-59_Eastbound_117	29
9	I-65_Northbound_266	36	19	I-10_Eastbound_25	29
10	I-459_Northbound_16	35	20	I-65_Southbound_260	29

(ii) Sideswipe-vehicle crash

Rank	Site_identifier	Density	Rank	Site_identifier	Density
1	I-65_Northbound_260	21	11	I-20/I-59_Westbound_127	9
2	I-20/I-59_Westbound_72	17	12	I-65_Northbound_259	8
3	I-20/I-59_Westbound_126	15	13	I-65_Southbound_260	8
4	I-20/I-59_Eastbound_76	14	14	I-85_Southbound_2	8
5	I-20/I-59_Eastbound_124	11	15	I-20/I-59_Westbound_128	8
6	I-65_Northbound_261	11	16	I-65_Southbound_173	8

7	I-459_Southbound_19	10	17	I-65_Southbound_261	8
8	I-65_Southbound_259	10	18	I-65_Northbound_252	7
9	I-65_Northbound_5	9	19	I-65_Southbound_5	7
10	I-65_Southbound_242	9	20	I-20/I-59_Eastbound_118	7

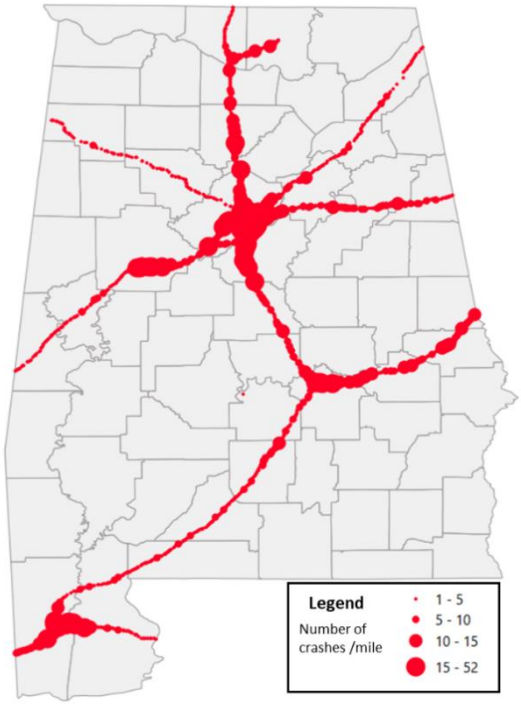
(iii) Rear-end crash density

Rank	Site_identifier	Density	Rank	Site_identifier	Density
1	I-65_Northbound_259	37	11	I-65_Northbound_252	21
2	I-20/I-59_Eastbound_123	30	12	I-459_Northbound_28	21
3	I-20/I-59_Westbound_126	28	13	I-459_Northbound_16	21
4	I-459_Southbound_19	27	14	I-459_Southbound_16	21
5	I-65_Southbound_259	27	15	I-10_Westbound_27	21
6	I-20/I-59_Eastbound_76	26	16	I-65_Northbound_236	21
7	I-459_Northbound_15	24	17	I-65_Northbound_260	20
8	I-459_Southbound_17	24	18	I-65_Northbound_4	19
9	I-65_Southbound_244	24	19	I-10_Eastbound_25	19
10	I-10_Westbound_26	22	20	I-65_Northbound_249	19

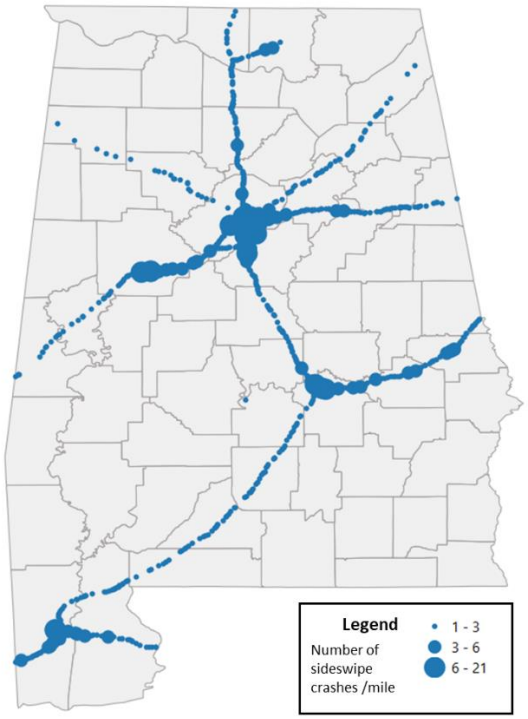
(iv) Single-vehicle crash density

Rank	Site_identifier	Density	Rank	Site_identifier	Density
1	I-65_Northbound_266	27	11	I-59_Southbound_131	10
2	I-20/I-59_Westbound_81	23	12	I-85_Southbound_57	9
3	I-65_Southbound_283	12	13	I-65_Northbound_298	8
4	I-65_Northbound_252	11	14	I-20/I-59_Westbound_115	8
5	I-65_Northbound_265	11	15	I-65_Southbound_270	8
6	I-459_Southbound_29	11	16	I-65_Southbound_199	8
7	I-459_Southbound_31	11	17	I-85_Southbound_22	8
8	I-65_Southbound_269	11	18	I-20/I-59_Westbound_119	8
9	I-20/I-59_Eastbound_115	10	19	I-20_Westbound_137	8
10	I-59_Southbound_130	10	20	I-65_Southbound_271	8

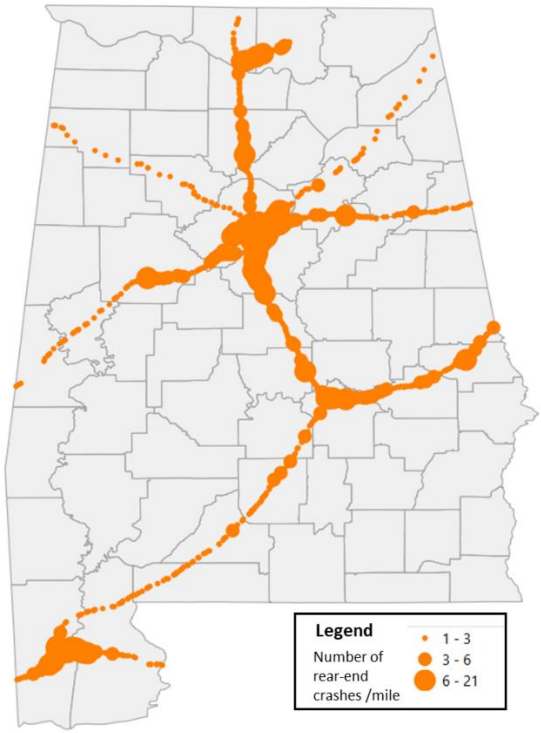
FIGURE 23 shows the mapping of crash frequency by type. The spatial distribution of high crash frequency areas for different crash types is not the same. Most of the high crash frequency sites are located within or near the Birmingham Metropolitan Area and Mobile Metropolitan Area, while I-22 is associated with fewer crashes.



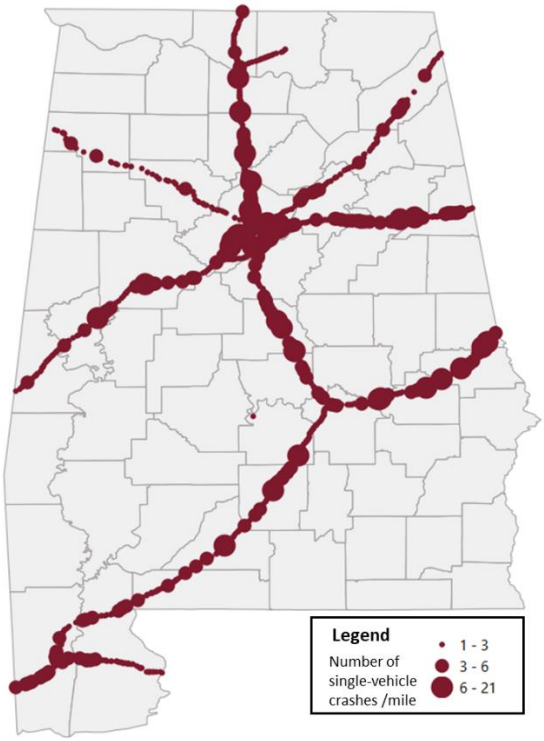
(a) Mapping of the frequency of all-type of crashes at segment level



(b) Mapping of the frequency of sideswipe crashes at segment level



(c) Mapping of the frequency of rear-end crashes at segment level

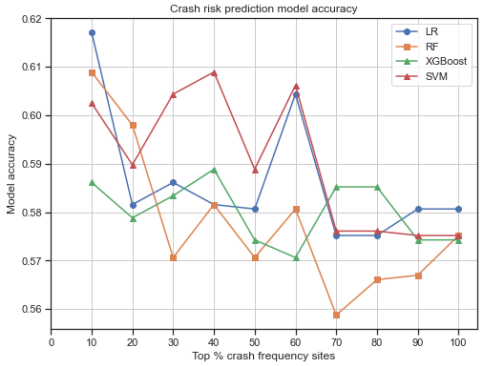


(d) Mapping of the frequency of all-type of crashes at segment level

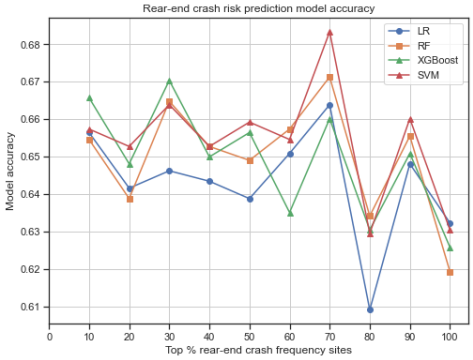
FIGURE 23 Mapping of the spatial distribution of the crash frequency

6.3.2 Separate Crash Risk Prediction Models

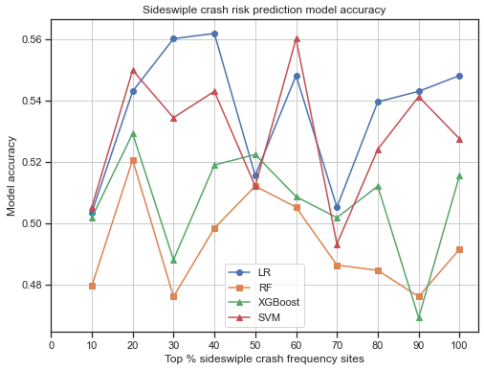
We built crash risk prediction models for crashes located on highway segments with different levels of crash frequency. In particular, we first sorted the 1-mile length freeway segment by crash frequency. Then we extracted the 90 percentile, 80 percentile, 70 percentile, 60 percentile, 50 percentile, 40 percentile, 30 percentile, 20 percentile, and 10 percentile of site-level crash frequency. Then extracted crashes located in freeway segments that have crash frequencies higher than specific percentiles. For example, first, we extracted crashes that occurred at the top 10% crash frequency sites (i.e., 90 percentile of site-level crash frequency), then followed by the 20% crash frequency sites, 30% crash frequency sites, 40% crash frequency sites, 50% crash frequency sites, 60% crash frequency sites, 70% crash frequency sites, 80% crash frequency sites, 90% crash frequency sites, and all sites, and built the models. The model types include logistic regression models (LR) and machine learning models including random forest (RF), support vector machines (SVM), and extreme gradient boosting models (XGBoost). In total, 40 models have been built. The models are all tuned by using the grid search method and 3-folded cross-validation is used to avoid overfitting. The models' accuracy is shown in FIGURE 4. The results show that, for all-type crash prediction models, generally, models for high-density crash freeway segments show better accuracy regardless of the model type. No significant improvement is shown for models separated by type.



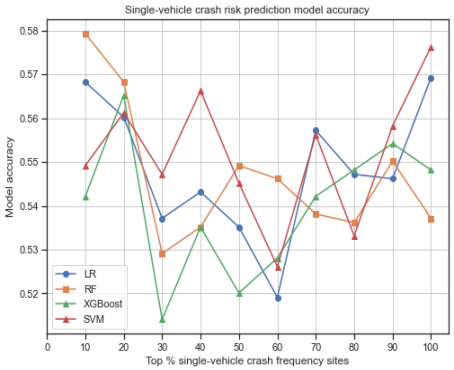
(a) All-type of crashes



(b) Rear-end crash



(c) Sideswipe crashes



(d) Single-vehicle crash

FIGURE 24 Crash prediction model accuracy by crash types and different levels of crash frequency.

6.4 Summary and Conclusion

Real-time crash risk is expected to support proactive traffic incident management by generating critical information for traffic managers to allocate incident response resources to high-risk sites before the occurrence of crashes. The team developed crash risk prediction models by taking advantage of the HERE crowdsourced probe vehicle data from a live database that reports and archives minute-by-minute real-time traffic speeds for freeways. The data is not limited to a specific road segment and covers the entire freeway network in Alabama. Based on the HERE data, the team created a variety of variables to capture the pre-crash traffic dynamics, which are traffic speed characteristics before the event of crashes, measured by mean speed, speed variance, and speed reduction. In addition to the pre-crash traffic dynamics, the team also extracted the traffic dynamics for crash-free conditions. With the data processed for pre-crash and crash-free traffic dynamics, the team developed logistic regression and machine learning models to predict the crash risk on freeways according to traffic dynamics along with static freeway attributes. Three machine learning approaches, including random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost) were tested and compared. Separate models were developed for all crashes, single-vehicle crashes, rear-end crashes, and sideswipe crashes. Modeling results were interpreted by using permutation feature importance and partial dependence plots.

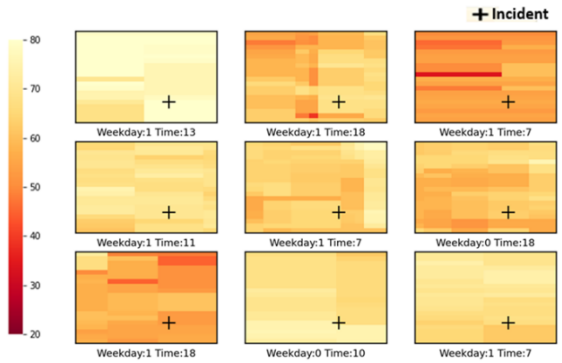
The results indicated that the traffic dynamics closer to the event of a crash are more predictive of the crash risk. Models for rear-end crashes are found to have a greater accuracy than other models, which implies that rear-end crashes have a significant relationship with pre-crash traffic dynamics (especially speed) and are more predictable than other crashes. For rear-end crashes, in comparison with the logistic regression model, XGBoost model, and SVM model, the RF model had a slightly improved prediction performance according to the AUC value (68.4%) and the accuracy (ACC) (65.1%). For rear-end crash models, the team calculated the permutation feature importance for each variable and computed the partial dependence for pre-crash speed-related top-ranked variables according to the feature importance. Though the feature importance rankings are different across models, the pre-crash speed-related variables are generally ranked high in all three models. According to the estimated partial dependence, the rear-end crash risk is positively related to the speed variance and speed reductions. A higher rear-end crash risk is associated with a more significant speed variance upstream, and the risk for a rear-end crash increases when the traffic speed at this location decreases significantly.

This study contributes by using crowdsourced probe vehicle data to develop real-time models to predict crash risk on freeways. Such data have been increasingly used by agencies for traffic monitoring and management at a reasonable cost without installing and maintaining electronic sensors on the road or roadside. Agencies with such data could potentially implement crash risk prediction models to facilitate their traffic incident responses. The team also identified the issues using such data for developing crash risk models. Continuing efforts are needed to examine the accuracy of the crowdsourced probe vehicle data, which is likely to vary across geographic areas and times. Future research will expand the modeling efforts to develop models that account for data issues such as unobserved heterogeneity.

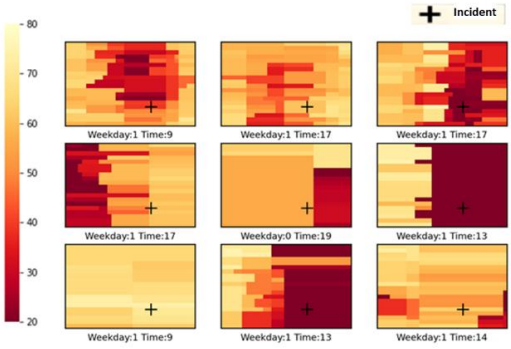
7. Incident Detection

7.1 Data Visualization

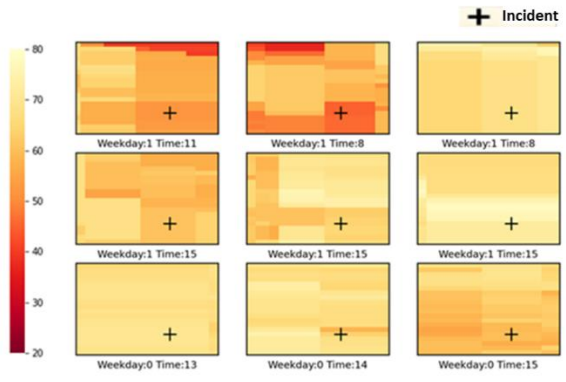
There are seventeen types of sub-incidents recorded in the Algo traffic incident database. However, not all types of sub-incidents have significant impacts on traffic dynamics and can be reflected in the after-incident traffic dynamics captured by probe vehicle data. In this project, detectable incidents are defined as an incident type that has impacts on traffic, and such impact (e.g., speed drop) could be captured by the HERE probe vehicle data. To get an intuitive look at the traffic incident impacts on traffic dynamics, for each incident subtype, 9 randomly selected speed matrices within the detection area for this incident subtype are visualized, as shown in FIGURE 25. As a comparison, the speed matrix within the detection area, when no incident occurs, is also visualized. The visualization results show that the traffic impacts due to the occurrence of abandoned vehicles and disabled vehicles are hard to reflect by the traffic dynamic obtained by HERE probe vehicle data. Conversely, in the majority of cases (based on our randomly selected FIGUREs), the traffic impact resulting from congestion, major crashes, minor crashes, moderate crashes, overturned vehicles, and vehicle fires are significantly captured by the spatial-temporal speed matrix. Based on such results, in this project, we classify congestion, major crashes, minor crashes, moderate crashes, overturned vehicles, and vehicle fire as the detectable incident subtypes. In addition, even for the same kind of incident sub-type, the after-incident traffic impact varies a lot during different times of day as well as the day of the week. Therefore, the team extracted incidents that occurred on weekdays during the daytime for model building later.



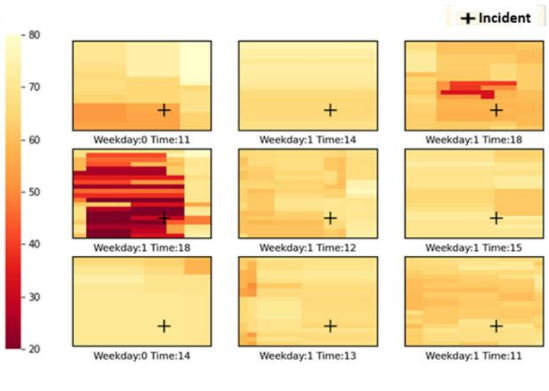
(i) Abandoned Vehicle



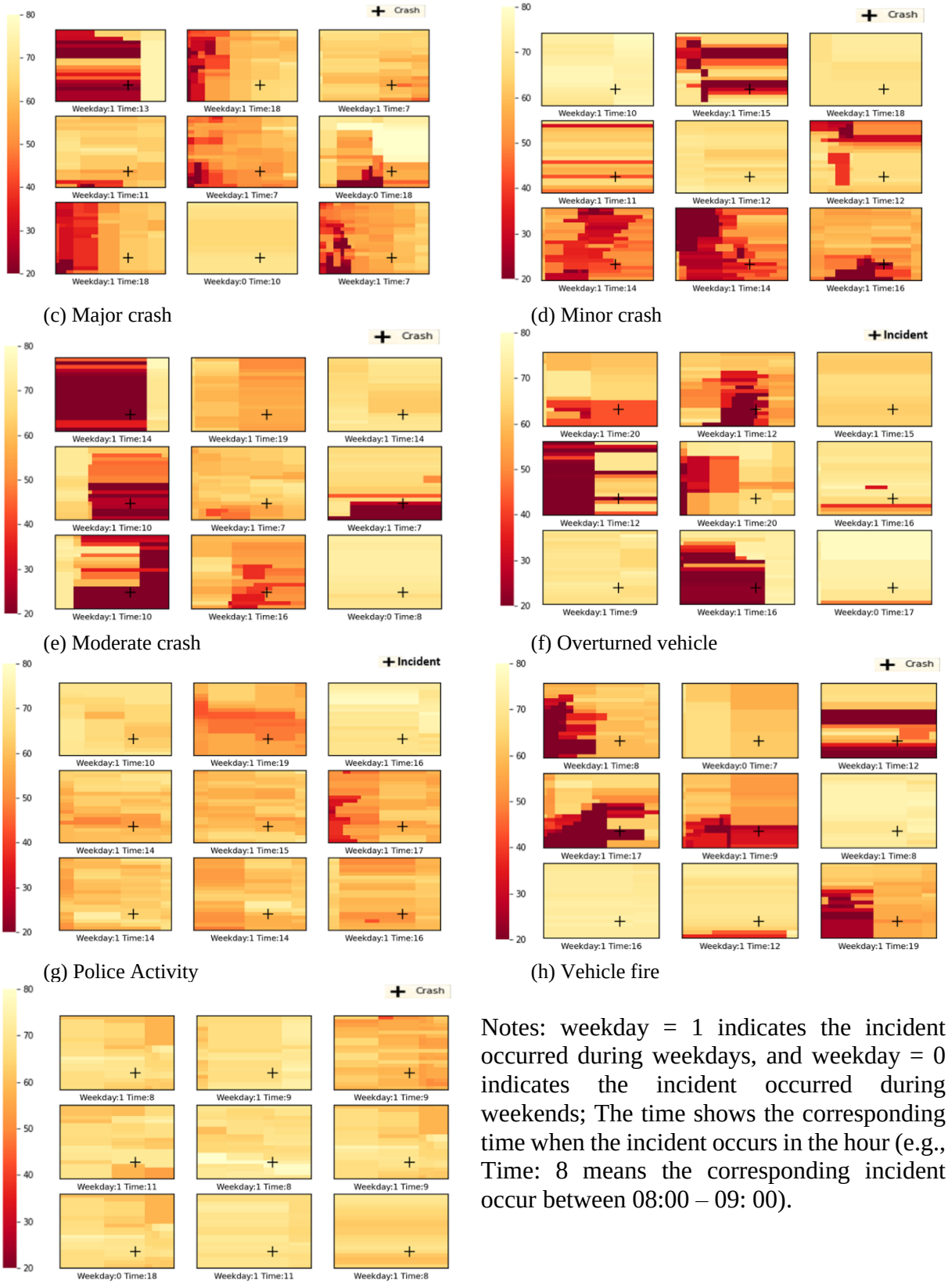
(ii) Congestion



(a) Disabled vehicle



(b) Grass fire



(c) Major crash

(d) Minor crash

(e) Moderate crash

(f) Overturned vehicle

(g) Police Activity

(h) Vehicle fire

(i) Incident free

Notes: weekday = 1 indicates the incident occurred during weekdays, and weekday = 0 indicates the incident occurred during weekends; The time shows the corresponding time when the incident occurs in the hour (e.g., Time: 8 means the corresponding incident occur between 08:00 – 09: 00).

FIGURE 25 Spatial-temporal speed matrix for different incident subtypes

7.2 Results of Statewide Model

7.2.1 Descriptive Statistics

As visualized in FIGURE 25, the occurrence of some specific incident types (e.g., disabled vehicle, abandoned vehicle), though accounting for a majority of the total number of incidents, may have little impact on traffic dynamics captured by HERE probe vehicle data. In other words, through post-incident traffic dynamics provided by HERE vehicle data, some incident types (e.g., major crashes, and moderate crashes) are detectable, and some incident types may not be detectable. To further identify the detachable incident and undetectable incident, the team calculated two mean speed metrics as shown in **Equation 24** (speed difference metrics 1) and **Equation 25** (speed difference metrics 2).

$$Spdif\_mean\_upstream = up15bef\_mean - up5aft\_mean \quad (24)$$

where, $up15bef\_mean$ denotes the mean pre-incident traffic speed (15-min before the incident) for the 2-mile upstream; $up5aft\_mean$ denotes mean after-incident traffic speed (5-min after the incident) for the 2-mile upstream.

$$Spdif\_mean\_updn = dn5aft\_mean - up5aft\_mean \quad (25)$$

where, $dn5aft\_mean$ denotes the mean after-incident traffic speed (5-min after the incident) for the 1 mile downstream; $up5aft\_mean$ has the same meaning in **Equation 9**.

In this project, the speed difference metrics are calculated for each incident sub-type. For comparison purposes, the speed difference metrics for the incident-free condition are also calculated. The incident sub-type with large values of speed difference metrics can be identified in the probe vehicle. Inversely, the incident subtypes with low values of speed difference matrices (or values of speed difference matrices close to that of incident-free condition), cannot be detected in the probe vehicle data. In other words, incident subtypes with higher values of speed difference metrics may be more detectable when HERE probe vehicle data are used to detect incidents. TABLE 15 presents the speed difference measured by speed difference metrics for both incident conditions and incident-free conditions. The results show that the speed difference metrics for the disabled vehicle are 0.099 miles/hour and 0.726 miles/hour, respectively, which are quite like the speed difference metrics for incident-free conditions. Significant speed variations (either speed difference metric 1 or speed difference metric 2 is greater than 3 miles/hour) are found for incident types including minor crashes, moderate crashes, congestion, police activity, major crashes, vehicle fires, overturned vehicles, and grass fire.

TABLE 15 Summarized traffic dynamics variables after the occurrence of different incident subtypes

Term		Frequency	Percentage	Speed difference metric 1	Speed difference metric 2
Incident - free		8,008	100.00%	-0.117	0.453
Incident subtypes	Disabled Vehicle	4,786	59.83%	0.099	0.726
	Minor Crash	982	12.28%	1.779	5.864
	Abandoned Vehicle	762	9.53%	0.106	0.647
	Moderate Crash	402	5.03%	5.845	9.064

	Congestion	314	3.93%	5.643	-3.045
	Debris	310	3.88%	1.096	2.199
	Police Activity	166	2.08%	2.587	4.41
	Major Crash	122	1.53%	7.765	14.346
	Vehicle Fire	51	0.64%	5.883	6.363
	Overturned Vehicle	39	0.49%	10.165	7.247
	Medical Emergency	21	0.26%	0.51	2.705
	Grass Fire	21	0.26%	4.799	2.588
	Null value	11	0.14%	0.408	4.292

Note: this TABLE only shows the incident subtypes which account for at least 0.1% of the total number of incidents.

7.2.2 Model Results

The results of the visualization and speed difference metrics all indicate that the occurrence of some incident sub-types (e.g., abandoned vehicle, disabled vehicle, etc.) may have little impact on traffic flow in terms of speed reduction or the impact of traffic flow may be hardly reflected by the speed information captured by HERE probe vehicle data. Instead of building one single model to detect the occurrence of all types of incidents (i.e., all-type model), the team develops both the all-type model and the model for only detectable incident sub-types. Besides, an additional model is developed to further classify detectable incident types. In this project, detectable incident sub-types are classified as crash, congestion, and other traffic-impacted incidents. The details of model classification labels are shown in TABLE 16. Specifically, Model 1 is used to detect the occurrence of all types of incidents; Model 2 only detects the occurrence of detectable incident subtypes that are identified in the above section; Model 3 further classifies the types of detectable incidents into crash, congestion, and other traffic impacted incident; Model 4 has the same classification labels in Model 3 but with a balanced dataset. Note, for Model 1 to Model 3, one incident observation is paired with one incident-free observation. For Model 4, a balanced dataset is created based on Model 3 using an oversampling method called SMOTE.

TABLE 16 Classification labels of different models

Model	Dataset	Label	Sample size	Description
Model 1	Raw data	0	6673	Incident free
		1	6665	Incident
Model 2	Raw data	0	1416	Incident free
		1	1412	Traffic impact incident (Congestion, Major Crash, Minor Crash, Moderate Crash, Overturned Vehicle, Vehicle Fire)
Model 3	Raw data	0	1416	Incident free
		1	1171	Crash
		2	172	Congestion
		3	69	Other traffic impacted incidents (Overturned Vehicle, Vehicle Fire)
Model 4	Balanced data	0	1416	Incident free
		1	1416	Crash
		2	1416	Congestion
		3	1416	Other traffic impacted incidents (“Overturned Vehicle, Vehicle Fire)

TABLE 17 (i) and (ii) present the modeling performance for Model 1 to Model 4 developed by ANN and CNN, respectively. In terms of incident detection models, the modeling results for both the ANN model and CNN model show that the model detecting traffic impact incidents has better prediction performance than the model detecting all types of incidents. This finding indicates that, for AID models, better prediction performance can be achieved when the model excludes the unpredictable incident sub-types. For incident-type classification models (i.e., model 3 and model 4), consistent with previous studies, the model with balanced data (i.e., Model 4) has a slightly better performance in terms of accuracy than model 3, in both ANN and CNN models. Comparing the two modeling methods, the CNN method outperforms the ANN model in terms of accuracy in Model 1 and Model 3. For Model 2, the two methods all achieve the same accuracy of 0.65. For Model 4, the ANN method achieves better performance with an accuracy of 0.82 than the CNN method with an accuracy of 0.71. This finding indicates that as deep learning models, ANN and CNN can both be used to detect the occurrence of incidents and the method with better prediction power may vary depending on the specific classification problem and whether the data is balanced data or not.

TABLE 17 Model performance

(i) ANN model

Model	Parameter	Label	Precision	recall	f1-score	support
Model 1	activation': 'relu', 'alpha':0.0001, 'hidden_layer_sizes': (150, 100, 50), learning_rate': 'constant', 'max_iter': 300, 'solver': 'adam'	0-Incident free	0.58	0.63	0.60	1,331
		1 - Incident	0.59	0.55	0.57	1,337
		macro avg	0.60	0.60	0.60	2,668
		weighted avg	0.60	0.60	0.60	2,668
		Accuracy			0.59	2,668
Model 2	activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (150, 100, 50), 'learning_rate': 'adaptive', 'max_iter': 300, 'solver': 'adam'	0 - Incident free	0.69	0.84	0.76	277
		1 - traffic impact incident	0.81	0.63	0.71	289
		macro avg	0.75	0.74	0.73	566
		weighted avg	0.75	0.74	0.73	566
		Accuracy			0.74	566
Model 3	activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (150, 100, 50), 'learning_rate': 'adaptive', 'max_iter': 300, 'solver': 'adam'	0	0.67	0.89	0.76	277
		1	0.69	0.45	0.54	242
		2	0.40	0.35	0.38	34
		3	0.29	0.15	0.20	13
		Macro avg	0.51	0.46	0.47	566
		Weighted avg	0.65	0.65	0.63	566
		Accuracy			0.65	566
Model 4	activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (150, 100, 50), 'learning_rate': 'adaptive', 'max_iter': 300, 'solver': 'adam'	0	0.65	0.80	0.72	285
		1	0.80	0.54	0.64	286
		2	0.92	0.99	0.95	275
		3	0.94	0.97	0.95	287
		Macro avg	0.83	0.82	0.82	1,133
		Weighted avg	0.83	0.82	0.82	1,133
		Accuracy			0.82	1,133

(ii) CNN model

Model	Structure	Label	precision	recall	f1-score	support
Model 1		0-Incident free	0.79	0.36	0.49	1,331

	Conv2D + MaxPooling2D + Conv2D + MaxPooling2D + Flatten + Dropout + Dense	1 - Incident	0.59	0.91	0.71	1,337
		macro avg	0.69	0.63	0.60	2,668
		weighted avg	0.69	0.63	0.60	2,668
		Accuracy			0.63	2,668
Model 2	Conv2D + MaxPooling2D + Conv2D + MaxPooling2D + Flatten + Dropout + Dense	0 - Incident free	0.67	0.92	0.78	277
		1 - traffic impact incident	0.88	0.57	0.69	289
		macro avg	0.78	0.75	0.74	566
		weighted avg	0.78	0.74	0.73	566
		Accuracy			0.74	566
Model 3	Conv2D + MaxPooling2D + Conv2D + MaxPooling2D + Flatten + Dropout + Dense	0	0.70	0.87	0.78	277
		1	0.67	0.57	0.61	242
		2	0.38	0.18	0.24	34
		3	--	--	--	13
		Macro avg	0.44	0.40	0.41	566
		Weighted avg	0.65	0.68	0.66	566
		Accuracy			0.68	566
Model 4	Conv2D + MaxPooling2D + Conv2D + MaxPooling2D + Flatten + Dropout + Dense	0	0.67	0.57	0.62	285
		1	0.64	0.46	0.53	286
		2	0.84	0.88	0.86	275
		3	0.68	0.94	0.79	287
		Macro avg	0.71	0.71	0.70	1,133
		Weighted avg	0.71	0.71	0.70	1,133
		Accuracy			0.71	1,133

Note: "--" means there is no correctly classified observation for this class.

7.3 Detectable Incidents in High-Risk Segments

7.3.1 Detectable Incident Subtypes Identification

TABLE 18 shows the distribution of the incident subtypes recorded in the Algo database in the year 2019. The incident subtypes that account for more than 1% of the total incident subtypes including disabled vehicle, abandoned vehicle, minor crash, moderate crash, congestion, debris, police activity, and major crash, account for 54.9%, 11.5%, 10.8%,9.8%, 4.3%, 3.5%, 1.6% and 1.5% of the total incidents.

TABLE 18 Distribution of incident subtypes

Incident Subtype	Frequency	Percentage
Disabled Vehicle	27587	54.9%
Abandoned Vehicle	5796	11.5%
Minor Crash	5440	10.8%
Moderate Crash	4928	9.8%
Congestion	2161	4.3%
Debris	1748	3.5%
Police Activity	781	1.6%
Major Crash	768	1.5%
Vehicle Fire	345	0.7%
Overturned Vehicle	300	0.6%
Grass Fire	116	0.2%
Null value	106	0.2%
Medical Emergency	83	0.2%
Wildlife in Roadway	24	0.0%
HazMat Spill	13	0.0%

Signal Outage	8	0.0%
Smoke	4	0.0%
Structure Fire	1	0.0%

Some incident subtypes may have a minor influence on the traffic dynamic. For example, a disabled vehicle parked on the shoulder of the freeway may have a minor influence on the traffic dynamics. To identify the detectable traffic incident subtypes, we built separate detection models for different incident subtypes. TABLE 19 shows the traffic dynamic variables for the incident detection models.

TABLE 19 Traffic dynamic variables for incident detection

Traffic Dynamic Variables (Data source: HERE database)*	
Variables	**Description**
up15bef_mean	Mean pre-incident traffic speed (15 minutes before the incident) for the 1 mile upstream.
up15bef_var	Variance pre-incident traffic speed (15-min before the incident) for the 1 mile upstream.
dn15bef_mean	Mean pre-incident traffic speed (15 minutes before the incident) for the 1 mile downstream.
dn15bef_var	Variance pre-incident traffic speed (15-min before the incident) for the 1 mile downstream.
df15bef_mean	up15bef_mean - dn15bef_mean
df15bef_var	up15bef_var - dn15bef_var
up4aft_mean	Mean after-incident traffic speed (15 minutes before the incident) for the 1 mile upstream.
up4aft_var	Variance after-incident traffic speed (15 minutes before the incident) for the 1 mile upstream.
dn4aft_mean	Mean after-incident traffic speed (15 minutes before the incident) for the 1 mile downstream.
dn4aft_var	Variance after-incident traffic speed (15 minutes before the incident) for the 1 mile downstream.
df4aft_mean	up4aft_mean - dn4aft_mean
df4aft_var	'up4aft_var' - 'dn4aft_var'
df_up4aft_mean	'up4aft_mean'- up15bef_mean
df_up4aft_var	'up4aft_var'- up15bef_var
df_dn4aft_mean	'dn4aft_mean'- dn15bef_mean
df_dn4aft_var	'dn4aft_var'- dn15bef_var
df_df4aft_mean	'df4aft_mean' - df15bef_mean
df_df4aft_var	'df4aft_var' - df15bef_var
sp4_drop_mean	Mean speed drop at the incident location
sp4_drop_var	Variance of the speed drop at the incident location
logaadt	Log of the annual average daily traffic
logaadt_truck	Log of the annual average daily traffic for trucks
Through_La	Number of the through lanes
UrbanRural	Whether the incident locations are in rural or urban area

TABLE 20 shows the incident detection models for separated incident subtypes. We can find that incident subtypes – minor crashes, moderate crash congestions, and major crashes show an accuracy higher than 60%, which indicates that those crash types are more detecTABLE. In the previous section, we visualized the crash frequencies on the map.

TABLE 20 Incident detection models for separated incident subtypes

Crash subtype	Time period	Detection time	Accuracy	Recall	Precision	F1	FAR	AUC
Disabled Vehicle	ampeak	4	0.514	0.527	0.53	0.529	0.499	0.517
Disabled Vehicle	midday	4	0.513	0.477	0.519	0.497	0.451	0.527
Disabled Vehicle	pmpeak	4	0.524	0.514	0.532	0.523	0.467	0.521
Disabled Vehicle	night	4	0.526	0.451	0.546	0.494	0.395	0.536
Abandoned Vehicle	ampeak	4	0.583	0.547	0.578	0.562	0.382	0.593
Abandoned Vehicle	midday	4	0.55	0.554	0.583	0.568	0.455	0.579
Abandoned Vehicle	pmpeak	4	0.476	0.034	0.714	0.066	0.016	0.509
Abandoned Vehicle	night	4	0.535	0.464	0.472	0.468	0.409	0.488
Minor Crash	ampeak	4	0.622	0.582	0.654	0.616	0.333	0.656
Minor Crash	midday	4	0.672	0.632	0.673	0.652	0.291	0.713
Minor Crash	pmpeak	4	0.566	0.65	0.575	0.61	0.525	0.571
Minor Crash	night	4	0.566	0.377	0.667	0.481	0.217	0.577
Moderate Crash	ampeak	4	0.862	0.793	0.907	0.846	0.074	0.899
Moderate Crash	midday	4	0.731	0.734	0.727	0.731	0.272	0.787
Moderate Crash	pmpeak	4	0.692	0.577	0.727	0.644	0.202	0.739
Moderate Crash	night	4	0.633	0.643	0.643	0.643	0.377	0.686
Congestion	ampeak	4	0.789	0.868	0.759	0.81	0.296	0.785
Congestion	midday	4	0.804	0.851	0.775	0.811	0.241	0.851
Congestion	pmpeak	4	0.773	0.746	0.815	0.779	0.196	0.825
Congestion	night	4	0.833	0.789	0.833	0.811	0.13	0.886
Debris	ampeak	4	0.547	0.528	0.549	0.538	0.434	0.562
Debris	midday	4	0.511	0.622	0.497	0.553	0.594	0.501
Debris	pmpeak	4	0.486	0.472	0.486	0.479	0.5	0.422
Debris	night	4	0.527	0.429	0.667	0.522	0.324	0.572
Police Activity	ampeak	4	0.545	0.429	0.75	0.545	0.25	0.585
Police Activity	midday	4	0.513	0.333	0.652	0.441	0.242	0.494
Police Activity	pmpeak	4	0.517	0.417	0.417	0.417	0.412	0.505
Police Activity	night	4	0.561	0.382	0.568	0.457	0.271	0.496
Major Crash	ampeak	4	0.867	0.81	0.895	0.85	0.083	0.889
Major Crash	midday	4	0.735	0.592	0.935	0.725	0.059	0.77
Major Crash	pmpeak	4	0.871	0.929	0.813	0.867	0.176	0.899
Major Crash	night	4	0.8	0.704	0.792	0.745	0.132	0.798

FIGURE 26 shows the mapping of the spatial distribution of the congestion. It can be seen that the majority of the congestion occurred around the mobile metropolitan area. In addition, the top 20

congestion frequency sites (recorded in Algo) of selected freeways in Alabama are shown in TABLE 9.

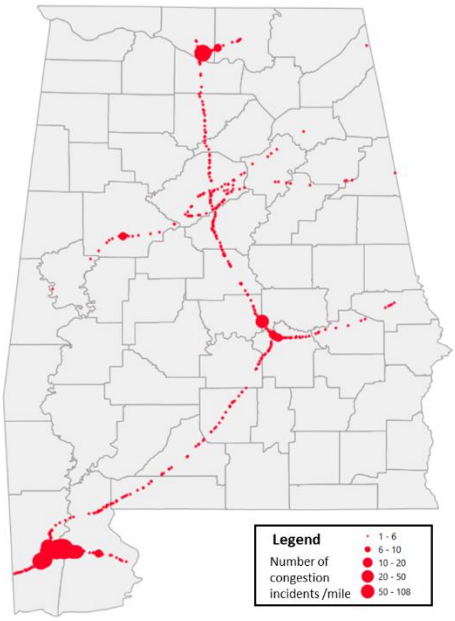


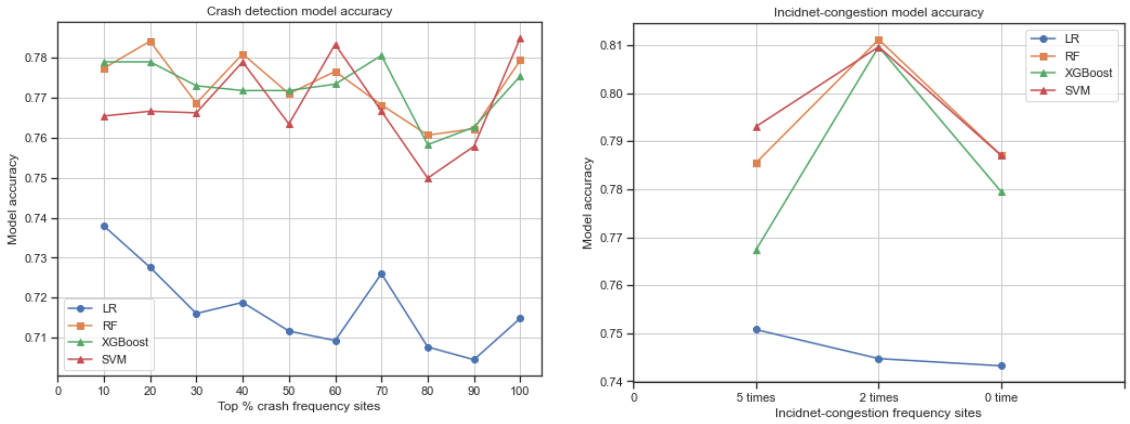
FIGURE 26
Mapping of the spatial distribution of the congestion

Rank	Site_identifier	Density	Rank	Site_identifier	Density
1	I-10_East_30	108	11	I-10_East_17	27
2	I-10_East_26	95	12	I-10_East_24	26
3	I-10_East_25	88	13	I-10_East_28	20
4	I-10_West_27	81	14	I-65_North_181	18
5	I-10_West_28	72	15	I-65_North_3	18
6	I-10_West_30	58	16	I-65_South_1	17
7	I-10_West_29	54	17	I-10_West_31	15
8	I-10_West_15	36	18	I-10_East_32	14
9	I-565_East_1	31	19	I-10_West_32	14
10	I-10_East_29	28	20	I-10_West_36	14

TABLE 21 Top 20 congestion frequency sites (recorded in Algo) of selected freeway in Alabama

7.3.2 Incident Detection Models

FIGURE 27 (a) shows the incident detection models for the crashes across different site-level crash frequencies. Machine learning models including random forest (RF), XGBoost, and support vector machines show higher model accuracy than the logistic regression models. No significant model accuracy can be seen for the three machine-learning models with different levels of incident frequency. But for the logistic regression models, generally, models show higher accuracy in high crash-frequency sites. Since the distribution of congestion is right-tail (i.e., many sites are related to 0 observation of the congestion), we extracted three thresholds for site-level congestion density, including higher than 5 congestions, 2 congestions, and 0 congestion. Similar to the all-type crash detection model, machine learning models show higher detection accuracy, but logistic regression is sensitive to the site-level congestion frequency.



(a) All types of crash (b) Congestion

FIGURE 27 Incident detection model accuracy by detectable incident types and different levels of incident frequency.

7.4 Summary and Conclusion

Real-time automatic incident detection (AID) is expected to support traffic incident management by producing in-time incident information as a backup of other traffic incident detection methods such as 911 calls and police patrol. Developing an AID algorithm is the key to implementing successful incident detection. The team develops three kinds of AID algorithms taking advantage of the HERE probe vehicle database. The HERE database could provide speed information updated every minute. Three AID models are developed in this project: (a) Model 1: AID model to detect the occurrence of all-type incidents; (b) Model 2: AID model to detect traffic incidents that have significant impacts on traffic flow (in terms of speed) and can be captured by HERE probe vehicle data; (c) Model 3: AID model that further classifies the incident sub-types. The inputs of these models are spatial-temporal speed matrix extracted from the HERE database with a resolution of 0.1-mile times by 1 minute. For models with imbalanced datasets (i.e., Model 3), an oversampling method called SMOTE was used to generate a balanced dataset before developing the models (i.e., Model 4).

The results indicated that AID models detecting the occurrence of traffic-impacted incidents have better-predicted accuracy than AID models detecting the occurrence of all types of traffic incidents, which implies that incorporating all types of incidents into the AID model will weaken the prediction performance of the AID model. For AID models that can classify the incident sub-types (crash, congestion, and other traffic-impacted incidents), models with a balanced dataset can achieve higher performance compared to models without balancing. Compared to AID models developed using ANN, AID models developed using CNN achieve better performance accuracy in Model 1 and Model 3 and show similar performance accuracy in Model 2. For incident sub-type classification with balanced data, ANN shows better performance regarding accuracy.

This section of the project contributes to the current state-of-the-art in multifold aspects. First, the team experimented with the possibility of using real-time speed information provided by probe vehicles to detect traffic incidents; Second, unlike previous studies incorporating all types of incidents into the AID model, the team identified traffic-impacted incidents and only included such types of incident into the AID models; Third, deep learning algorithms were used to capture the image-like real-time speed information provided by spatial-temporal speed matrix without

aggregating traffic information into summarized traffic flow variables; Fourth, the AID models developed in this project can classify different incident subtypes, which could provide helpful information for TIM practitioners to take proper countermeasures. Agencies with such data can potentially implement the AID models in this project to detect the occurrence of the incident and classify the incident types. In further studies, more ANN and CNN structures may be tested to achieve improved prediction performance. In addition, road characteristics and temporal information may also need to be considered and incorporated into the model.

8. Incident Impact

8.1 Variable Creation

Regarding the spatial-temporal effects of the incident, the team develops three novel metrics to measure traffic incidents' spatiotemporal impacts: 1) the maximum queue length; 2) time at the maximum queue length; 3) and volume (i.e., the spatiotemporal extent of a queue). Those metrics are based on the traffic dynamics of the HERE data. In addition, the time at the maximum queue length after the occurrence of one incident (one incident could have no queue afterward) is extracted from the HERE database. Notably, the team developed a concept, named "***volume***", representing a queue's spatiotemporal extent calculated in three dimensions: speed reduction, segment length, and time window, as defined by **Equation 26**.

$$Volume = \left\{ \frac{\sum_{m_i}^{m_f} \sum_{t_i}^{t_f} (FFS_i - Speed_{m_t})}{60} \right\} * miles * 3{,}600 \quad (26)$$

where m_i and m_f indicate the start and end milepost over a specific segment, t_i and t_f are the initial and end timestamps over a specific duration, and FFS_i is the free flow speed at the road segment i. In this project, the start and end milepost cover 5-mile segments, the initial and end timestamps cover a 60-minute duration, and a typical free flow speed is 70 mph. **TABLE 23** shows detailed explanations for the abovementioned variables. Moreover, **FIGURE 28** illustrates an example of the maximum queue length and time at the maximum queue length for a major crash incident. **FIGURE 29** demonstrates a 3-D conceptual drawing of the spatiotemporal extent of a queue (***volume***).

TABLE 22 Spatiotemporal Impacts (Bold) and Traffic Dynamics-related Variables

Variables	Descriptions
queue_max	The maximum queue length (up to 5-mile)
time_queue_max	The time reaches the maximum queue length (up to 60-min)
volume	The spatiotemporal extent of a queue (speed reduction * segment length * time window)
up15bef_5m_mean	The mean of pre-incident traffic speed (15-min before the incident) for 5-mile upstream
up15bef_5m_var	The variance of pre-incident traffic speed (15-min before the incident) for 5-mile upstream
up15bef_3m_mean	The mean of pre-incident traffic speed (15-min before the incident) for 3-mile upstream
up15bef_3m_var	The variance of pre-incident traffic speed (15-min before the incident) for 3-mile upstream
up15bef_1m_mean	The mean of pre-incident traffic speed (15-min before the incident) for 1 mile upstream
up15bef_1m_var	The variance of pre-incident traffic speed (15-min before the incident) for 1 mile upstream

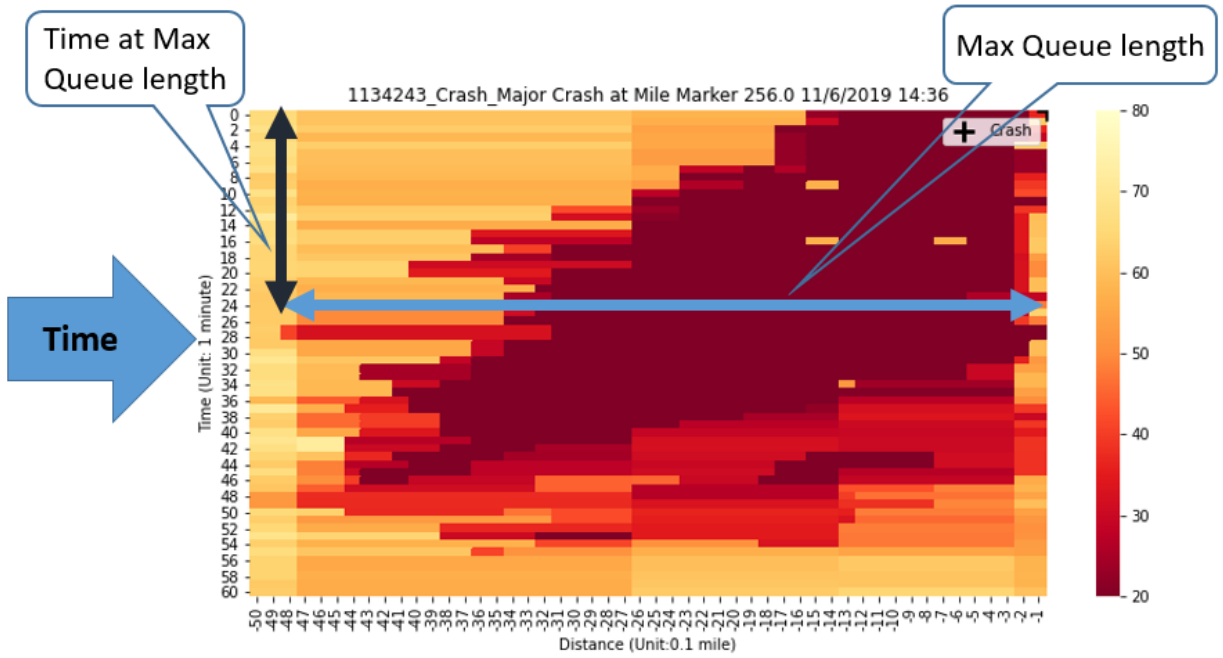


FIGURE 28 Spatiotemporal impacts for an incident (max queue length & time at max queue length)

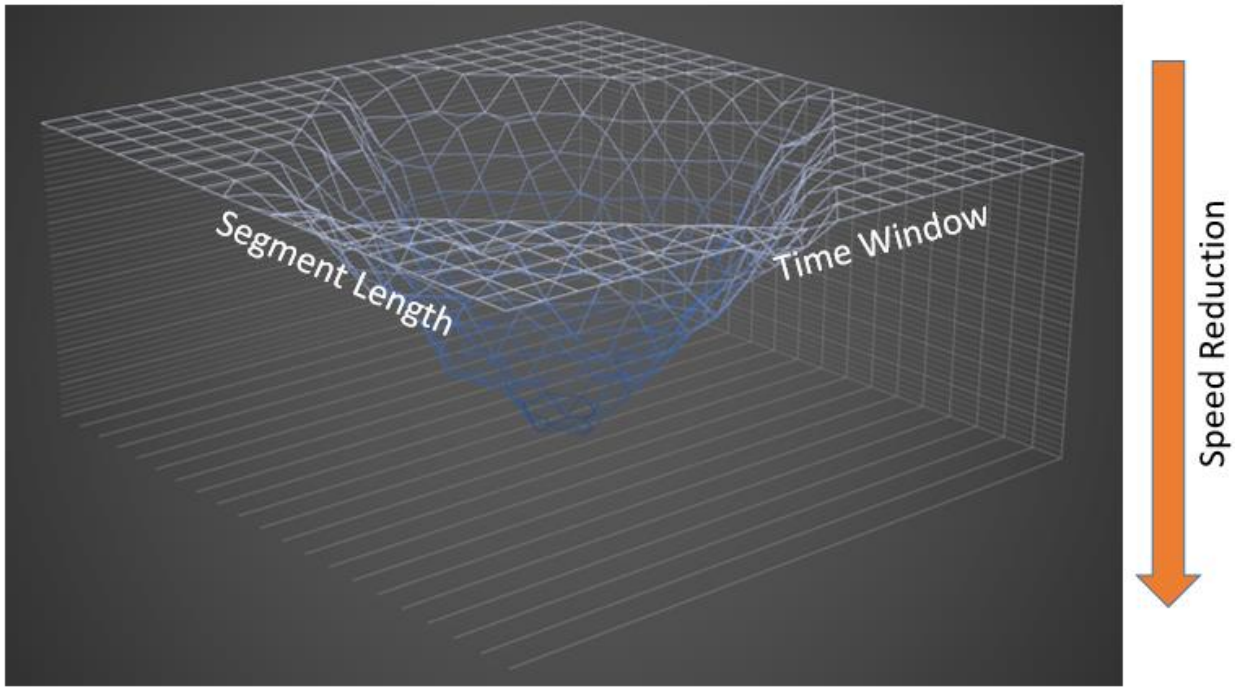


FIGURE 29 Spatiotemporal impacts for an incident (*volume*)

8.2 Descriptive Statistics

TABLE 24 demonstrates the descriptive statistics of variables used for the modeling process. According to **TABLE 24**, the queue length shorter than 1 mile has a higher percentage of data than other groups. Around 27% of queues caused by incidents were reached at their longest queuing distance within 10 minutes. As expected, the ***volume*** or spatiotemporal impacts of traffic incidents of less than 2,000 occupy the largest proportion (28%) of the dataset, indicating that over a quarter of traffic incidents in Alabama have a relatively small impact on road congestion. Similarly, the most considerable portion regarding the mean of pre-incident traffic speed (15 minutes before the incident) for 5 miles upstream of the incident location refers to the speed range between 60 and 80 mph.

From the standpoint of incidents' characteristics, over 85% occurred during weekdays or on freeways in urban areas. The appearance of law enforcement accounts for 30% of total incidents. In terms of incident subtypes, 17% of incidents were minor crashes, and 12% were moderate crashes. It should be noted that disabled vehicles account for around 47% of total incidents, which may not cause serious traffic congestion in most cases. Besides, major crashes were found in 4% of total incidents. Lastly, regarding the effectiveness of Traffic Incident Management (TIM), around 25% of total incidents have less than a 20-minute response time, and over 27% of them can be handled and cleared within 30 minutes.

TABLE 23 Descriptive Statistics of Variables

Variables	Categories	Freq	%	Variables	Categories	Freq	%
queue	(0.0-1.0]	2176	26.61	**Winter (Dec., Jan., or Feb.)**	Yes	1784	21.81
	(1.0-2.0]	2061	25.20		No	6394	78.19
	(2.0-3.0]	1463	17.89	**Weekday**	Yes	7215	88.22
	(3.0-4.0]	886	10.83		No	963	11.78
	(4.0-5.0]	1592	19.47	**AM Peak (6 AM – 9 AM)**	Yes	1452	17.75
time_queue_max	1-10	2266	27.71		No	6726	82.25
	11-20	1564	19.12	**PM Peak (4 PM – 7 PM)**	Yes	2480	30.33
	21-30	1251	15.30		No	5698	69.67
	31-40	1053	12.88	**Police Presence**	Yes	2504	30.62
	41-50	996	12.18		No	5674	69.38
	51-60	1048	12.81	**Towing Presence**	Yes	1292	15.80
volume	(0-2000]	2517	30.78		No	6886	84.20
	(2000-4000]	2365	28.92	**Urban**	Yes	6924	84.67
	(4000-6000]	1308	15.99		No	1254	15.33
	(6000-8000]	708	8.66	**Incident Subtype**	Disabled Vehicle	3906	47.76
	(8000-10000]	522	6.38		***Minor Crash***	1380	16.87
	>10000	758	9.27		***Moderate Crash***	1005	12.29
up15bef_5m_mean	(0-40]	605	7.40		Congestion	920	11.25
	(40-50]	724	8.85		Abandoned Vehicle	638	7.80
	(50-60]	1789	21.88		***Major Crash***	329	4.02
	(60-80]	5060	61.87	**Severity**	Low	5907	72.23
up15bef_5m_var	(0-50]	4263	52.13		Medium	1790	21.89
	(50-100]	1105	13.51		High	481	5.88
	(100-250]	1479	18.09	**Number of Lanes**	3-4 lanes	3224	39.42
	>250	1331	16.28		5-6 lanes	3112	38.05
up15bef_3m_mean	(0-40]	850	10.39		> 6 lanes	1842	22.52
	(40-50]	739	9.04	**AADT**	(0-20000]	569	6.96
	(50-60]	1878	22.96		(20000-40000]	3476	42.50
	(60-80]	4711	57.61		(40000-60000]	1978	24.19
up15bef_3m_var	(0-50]	4593	56.16		(60000-80000]	2155	26.35
	(50-100]	1012	12.37	**Response Time**	(0-10]	1722	21.06

	(100-250]	1437	17.57		(10-20]	399	4.88
	>250	1136	13.89		(20-60]	360	4.40
up15bef_1m_mean	(0-40]	1174	14.36		>60	109	1.33
	(40-50]	784	9.59		NA	5588	68.33
	(50-60]	1710	20.91	**Clearance Time**	(0-10]	1022	12.50
	(60-80]	4510	55.15		(10-20]	721	8.82
up15bef_1m_var	(0-50]	5433	66.43		(20-30]	586	7.17
	(50-100]	880	10.76		(30-60]	1703	20.82
	(100-250]	1180	14.43		(60-90]	834	10.20
	>250	685	8.38		>90	1248	15.26

8.3 Impact Model

8.3.1 Maximum Queue Length Models

The marginal effects of variables on the model of maximum queue length are shown in **TABLE 25**. The TABLE displays the average marginal effects, computed by averaging five machine learning models after removing the maximum and minimum values. The marginal effects of five machine learning models generally have the same signs but differing magnitudes. There are variables with both positive and negative marginal effects, including the mean or variations of pre-incident traffic speed (15 minutes before the incident) for 5 miles upstream, urban area, incident subtype, incident severity, and the number of lanes. The marginal effect magnitudes may imply that these variables have a relatively minor or inconsequential impact on the model of maximum queue length.

In terms of the mean of pre-incident traffic speed (15 minutes before the incident) for 5 miles upstream, negative marginal effects indicate that when upstream traffic flow speed is over 60 mph, it is 25% less likely to have a maximum queue length longer than 4 miles. A similar trend could be found in 3-mile and 1-mile upstream traffic dynamics. From the law enforcement perspective, the presence of police officers could lead to a 4% reduction in shorter than 1-mile maximum queue length. It should be noted that for major crash incidents, it is over 10% less likely to have a maximum queue length shorter than 1 mile but over 8% more likely to have a maximum queue length longer than 4-mile. As expected, similar findings could be found in high-severity traffic incidents. Lastly, with the AADT over 60,000 vehicles, it is on average 3% more likely to cause a maximum queue length longer than 4 miles. More findings regarding how contributing factors affect the maximum queue length can be found in **TABLE 25**.

TABLE 24 Marginal Effects of Variables on Maximum Queue Length

Variables	Categories	Average ME for *queue_max*				
		(0.0-1.0]	(1.0-2.0]	(2.0-3.0]	(3.0-4.0]	(4.0-5.0]
up15bef_5m_mean	(0-40] - Base	-	-	-	-	-
	(40-50]	-1.28%	0.02%	8.15%	1.48%	-6.08%
	(50-60]	5.95%	10.52%	8.13%	-2.23%	-22.02%
	(60-80]	17.02%	11.88%	2.13%	-4.99%	-25.13%
up15bef_5m_var	(0-50] - Base	-	-	-	-	-
	(50-100]	-5.83%	4.03%	2.16%	0.32%	-1.17%
	(100-250]	-12.13%	5.51%	2.75%	2.43%	2.57%
	>250	-15.32%	2.90%	5.59%	3.44%	3.39%
up15bef_3m_mean	(0-40] - Base	-	-	-	-	-
	(40-50]	-1.19%	-1.34%	2.65%	1.17%	0.08%
	(50-60]	3.28%	4.89%	2.59%	0.45%	-8.81%
	(60-80]	10.27%	5.92%	1.61%	-2.34%	-13.11%
up15bef_3m_var	(0-50] - Base	-	-	-	-	-
	(50-100]	-2.43%	-0.23%	0.12%	0.85%	1.33%

	(100-250]	-6.48%	0.15%	1.91%	1.57%	2.10%
	>250	-5.55%	0.08%	3.08%	1.16%	1.77%
up15bef_1m_mean	(0-40] - Base	-	-	-	-	-
	(40-50]	-1.38%	1.85%	0.07%	-0.32%	0.34%
	(50-60]	0.77%	-1.20%	-0.73%	0.74%	-0.28%
	(60-80]	0.14%	-1.01%	0.54%	1.94%	-1.84%
up15bef_1m_var	(0-50] - Base	-	-	-	-	-
	(50-100]	-0.06%	-0.46%	1.14%	-1.17%	0.51%
	(100-250]	-0.75%	-0.45%	-0.01%	-0.35%	1.84%
	>250	-3.56%	-0.91%	1.32%	1.25%	1.96%
Winter (Dec., Jan., or Feb.)	Yes	2.21%	0.14%	-0.92%	-0.66%	-1.05%
Weekday	Yes	0.22%	-0.02%	1.51%	-0.20%	-1.90%
AM Peak (6 AM – 9 AM)	Yes	-1.64%	0.76%	1.05%	0.11%	-0.28%
PM Peak (4 PM – 7 PM)	Yes	-1.27%	0.12%	-0.21%	0.33%	0.36%
Police Presence	Yes	-4.05%	0.30%	1.79%	0.19%	1.00%
Towing Presence	Yes	-0.41%	0.07%	0.51%	-0.79%	0.93%
Urban	Yes	3.65%	1.85%	-0.02%	-0.38%	-4.95%
Incident Subtype	Disabled Veh. - Base	-	-	-	-	-
	Minor Crash	-1.17%	-3.50%	-0.98%	2.04%	2.48%
	Moderate Crash	-4.36%	-3.57%	1.00%	1.54%	4.63%
	Congestion	-3.89%	0.19%	2.47%	-0.55%	0.96%
	Abandoned Vehicle	2.33%	0.66%	-1.48%	-0.15%	-0.80%
	Major Crash	-10.52%	-0.17%	-0.18%	2.64%	8.62%
Severity	Low - Base	-	-	-	-	-
	Medium	-3.21%	-0.79%	2.60%	-0.41%	2.06%
	High	-5.60%	-4.32%	2.64%	2.79%	5.01%
Number of Lanes	3-4 lanes - Base	-	-	-	-	-
	5-6 lanes	0.68%	0.85%	2.17%	-0.02%	-3.38%
	> 6 lanes	6.07%	1.13%	-2.55%	-1.58%	-4.36%
AADT	(0-20000] - Base	-	-	-	-	-
	(20000-40000]	1.46%	-1.37%	-0.51%	-0.03%	0.18%
	(40000-60000]	2.46%	0.90%	-0.83%	-1.02%	-1.53%
	(60000-80000]	-0.49%	-0.62%	-0.05%	-1.95%	3.22%

8.3.2 Time at Maximum Queue Length Models

TABLE 26 demonstrates the marginal effects of variables on the model of time reaches maximum queue length. Similar to the content shown in **TABLE 4**, there are variables with both positive and negative values, for instance, AM peak and PM peak ones. It could be deemed that when incidents occur at AM peak or PM peak, it is more likely to take less than 10 minutes to reach the maximum queue length. On the contrary, it is less likely to have an incident with more than 50 minutes to reach the maximum queue length. Besides, the marginal effects of towing presence reveal that when towing services have been requested for an incident, that incident is 3% less likely to have less than 10 minutes to reach the maximum queue length. From an incident subtype perspective, as expected, minor and moderate crash incidents are more likely to take less than 10 minutes to reach the maximum queue length but less likely to take more than 50 minutes to reach the maximum queue length. Lastly, with the AADT over 60,000 vehicles, it is on average 2% less likely to have an incident with less than 10 minutes to reach the maximum queue length. More information can be found in **TABLE 26** regarding how contributing factors affect the time to reach the maximum queue length.

TABLE 25 Marginal Effects of Variables on Time at Maximum Queue Length

Variables	Categories	Average ME for *time_queue_max*					
		1-10	11-20	21-30	31-40	41-50	51-60
up15bef_5m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	-1.50%	0.50%	0.88%	0.76%	-0.10%	-0.03%
	(50-60]	-4.06%	-0.19%	0.67%	1.06%	1.11%	1.46%
	(60-80]	-8.52%	-0.20%	0.77%	2.15%	1.28%	4.08%
up15bef_5m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	4.18%	0.65%	0.94%	0.05%	-2.04%	-3.85%
	(100-250]	7.76%	2.25%	0.18%	-1.23%	-3.27%	-5.64%
	>250	4.47%	1.75%	0.71%	-0.52%	-1.85%	-5.01%
up15bef_3m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	0.52%	0.32%	0.03%	-0.09%	0.34%	-0.56%
	(50-60]	-1.71%	0.57%	0.37%	0.26%	0.25%	-0.07%
	(60-80]	-3.54%	-0.06%	-0.27%	0.49%	0.83%	1.56%
up15bef_3m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	1.00%	0.05%	-0.21%	0.08%	-0.42%	-0.80%
	(100-250]	4.34%	1.06%	0.09%	-0.93%	-1.63%	-2.52%
	>250	1.64%	0.89%	-0.11%	-0.26%	-1.10%	-0.98%
up15bef_1m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	-0.55%	-0.28%	0.15%	0.55%	-0.10%	0.19%
	(50-60]	0.37%	-0.22%	0.30%	0.13%	-0.65%	0.56%
	(60-80]	-2.00%	-0.13%	-0.07%	0.26%	1.03%	1.37%
up15bef_1m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	3.57%	1.61%	-0.05%	-0.79%	-1.57%	-2.32%
	(100-250]	2.68%	0.36%	-0.10%	0.00%	-1.15%	-1.39%
	>250	3.38%	-0.35%	-0.24%	0.03%	-0.82%	-1.20%
Winter (Dec., Jan., or Feb.)	Yes	-0.43%	0.59%	0.17%	-0.49%	0.22%	-0.16%
Weekday	Yes	-0.07%	-0.59%	0.16%	-0.44%	1.21%	0.68%
AM Peak (6 AM – 9 AM)	Yes	1.34%	-0.04%	0.06%	-0.07%	-1.13%	-0.43%
PM Peak (4 PM – 7 PM)	Yes	1.31%	0.96%	0.22%	0.19%	-0.52%	-1.98%
Police Presence	Yes	-0.14%	0.31%	0.19%	0.40%	0.01%	-0.46%
Towing Presence	Yes	-3.39%	0.53%	0.46%	0.88%	1.33%	-0.07%
Urban	Yes	0.29%	-0.17%	-0.31%	-0.18%	0.10%	-0.02%
Incident Subtype	Disabled Veh. - Base	-	-	-	-	-	-
	Minor Crash	1.96%	0.21%	0.17%	-0.58%	-0.49%	-1.91%
	Moderate Crash	0.21%	2.08%	0.74%	0.38%	-0.35%	-3.81%
	Congestion	2.18%	1.37%	-0.84%	-0.32%	-0.64%	-1.22%
	Abandoned Vehicle	-0.62%	-0.90%	-0.53%	0.13%	0.20%	1.00%
	Major Crash	-0.11%	0.64%	-0.10%	0.01%	0.57%	-1.45%
Severity	Low - Base	-	-	-	-	-	-
	Medium	0.36%	2.44%	0.11%	0.39%	-0.63%	-2.61%
	High	-1.14%	0.70%	-0.01%	0.67%	0.54%	-0.52%
Number of Lanes	3-4 lanes - Base	-	-	-	-	-	-
	5-6 lanes	0.81%	-0.98%	-0.45%	-0.65%	0.47%	1.22%
	> 6 lanes	0.18%	0.02%	-0.12%	0.01%	-0.93%	0.91%
AADT	(0-20000] - Base	-	-	-	-	-	-
	(20000-40000]	-0.65%	0.21%	0.20%	-0.07%	0.17%	0.18%
	(40000-60000]	-0.54%	0.45%	0.06%	-0.43%	0.05%	0.24%
	(60000-80000]	-2.12%	-0.10%	0.01%	0.56%	0.78%	1.06%

8.3.3 Spatiotemporal Impact Model

TABLE 27 illustrates the marginal effects of variables on the model of an incident's spatiotemporal impacts, measured by speed reduction, segment length, and incident time window. The content indicates that it is more likely to cause a low congestion severity when an incident's speed (the mean of pre-incident traffic speed (15 minutes before the incident) for 5 miles upstream of the incident location) is over 60 mph. As expected, the marginal effects regarding the incident

subtype reveal that a major crash incident is more likely to cause a high congestion severity. Further, on average, it is 12% less likely to cause a low congestion severity with the AADT over 60,000 vehicles.

TABLE 26 Marginal Effects of Variables on *Volume* (Spatiotemporal Impacts)

		Average ME for *volume*					
Variables	**Categories**	**(0-2000]**	**(2000-4000]**	**(4000-6000]**	**(6000-8000]**	**(8000-10000]**	**>10000**
up15bef_5m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	0.42%	-0.31%	3.93%	1.53%	-0.43%	-3.40%
	(50-60]	-0.12%	13.35%	9.60%	-0.62%	-4.79%	-15.17%
	(60-80]	21.96%	11.83%	-1.95%	-4.13%	-7.45%	-16.24%
up15bef_5m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	-7.74%	6.11%	1.36%	-0.09%	-0.16%	-0.11%
	(100-250]	-10.34%	3.95%	-0.39%	1.45%	0.47%	1.74%
	>250	-9.63%	0.91%	0.25%	0.96%	1.12%	1.78%
up15bef_3m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	0.07%	-0.20%	1.72%	0.05%	1.68%	0.07%
	(50-60]	-1.44%	5.89%	1.38%	1.50%	-0.05%	-3.88%
	(60-80]	15.94%	0.24%	-0.73%	-1.15%	-1.55%	-5.05%
up15bef_3m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	-2.08%	-0.33%	0.01%	0.18%	0.25%	0.48%
	(100-250]	-2.27%	-0.99%	-0.33%	0.37%	1.27%	1.16%
	>250	-3.01%	0.32%	-0.02%	0.34%	0.96%	0.71%
up15bef_1m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	-0.15%	-0.31%	0.70%	-0.08%	0.01%	-0.04%
	(50-60]	-0.85%	1.95%	-0.55%	-0.05%	0.06%	-0.18%
	(60-80]	3.42%	-1.72%	-0.03%	0.29%	-0.77%	-1.03%
up15bef_1m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	-0.03%	0.00%	-0.45%	0.14%	0.23%	0.20%
	(100-250]	-0.42%	-0.81%	0.18%	-0.03%	0.24%	0.48%
	>250	-3.13%	-0.46%	0.18%	0.46%	1.44%	0.83%
Winter (Dec., Jan., or Feb.)	Yes	0.59%	-0.50%	-0.06%	-0.07%	0.15%	-0.02%
Weekday	Yes	-0.42%	0.14%	0.50%	-0.16%	0.38%	-0.11%
AM Peak (6 AM – 9 AM)	Yes	-0.02%	-0.33%	0.69%	0.00%	-0.15%	-0.16%
PM Peak (4 PM – 7 PM)	Yes	0.23%	-0.69%	0.00%	14.00%	0.07%	0.11%
Police Presence	Yes	-1.63%	-0.28%	0.19%	0.00%	0.30%	0.62%
Towing Presence	Yes	-0.55%	-0.38%	-0.22%	-0.22%	0.22%	1.00%
Urban	Yes	-0.10%	1.08%	0.25%	0.17%	-0.17%	-1.25%
Incident Subtype	Disabled Veh. - Base	-	-	-	-	-	-
	Minor Crash	-0.43%	-0.89%	-0.92%	0.17%	0.16%	1.76%
	Moderate Crash	-2.01%	-0.69%	0.09%	0.07%	0.03%	1.96%
	Congestion	-1.38%	0.09%	1.08%	0.45%	0.57%	-0.45%
	Abandoned Vehicle	1.09%	-0.38%	0.28%	-0.05%	-0.44%	-0.34%
	Major Crash	-9.10%	-0.69%	1.02%	0.84%	1.28%	3.10%
Severity	Low - Base	-	-	-	-	-	-
	Medium	-1.72%	-0.27%	1.23%	-0.16%	0.49%	0.79%
	High	-9.10%	0.58%	2.82%	0.13%	0.40%	3.13%
Number of Lanes	3-4 lanes - Base	-	-	-	-	-	-
	5-6 lanes	1.52%	-0.04%	-0.22%	0.37%	0.16%	-1.22%
	> 6 lanes	1.06%	1.61%	-1.11%	-0.25%	-0.23%	-0.85%
AADT	(0-20000] - Base	-	-	-	-	-	-
	(20000-40000]	0.03%	-0.53%	0.96%	0.13%	0.28%	-0.19%
	(40000-60000]	1.30%	-0.19%	-0.45%	-0.05%	-0.07%	0.32%
	(60000-80000]	-12.07%	8.55%	-0.35%	-0.15%	1.77%	1.53%

8.3.4 Incident Clearance Time Models

Last but not least, **TABLE 28** shows the variables' marginal effects on the incident clearance time model. Incident clearance time refers to the time between the first response to the incident and the time at which the last responder has left the scene. It could be seen from the TABLE that as an incident's spatiotemporal impacts increase, the probability of clearance time of more than 90 minutes goes up dramatically. The presence of police officers or towing services is negatively associated with clearance times shorter than 10 minutes but positively associated with more than 30 minutes of clearance time. The same findings could be found in the marginal effects of crash incidents, including minor, moderate, and major crashes.

TABLE 28 TABLE 27Marginal Effects of Variables on Incident Clearance Time Models

		Average ME for *clearance time*					
Variables	**Categories**	**(0-10]**	**(10-20]**	**(20-30]**	**(30-60]**	**(60-90]**	**>90**
volume	(0-2000] - Base	-	-	-	-	-	-
	(2000-4000]	0.53%	-0.09%	-0.05%	-0.56%	0.07%	0.00%
	(4000-6000]	-0.31%	0.33%	-0.16%	0.54%	0.57%	0.76%
	(6000-8000]	0.44%	0.08%	-0.01%	0.03%	0.16%	3.01%
	(8000-10000]	-0.41%	-0.24%	-0.35%	-2.05%	0.59%	2.89%
	>10000	-0.79%	-0.51%	-0.70%	-2.28%	1.52%	5.00%
up15bef_5m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	-0.40%	-0.47%	-0.09%	-0.41%	0.12%	0.58%
	(50-60]	-0.77%	-0.64%	-0.01%	-0.57%	-0.03%	1.13%
	(60-80]	-1.69%	-1.78%	-0.85%	-0.60%	0.03%	1.37%
up15bef_5m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	1.09%	0.04%	0.22%	0.10%	-0.05%	0.47%
	(100-250]	1.44%	0.83%	0.28%	0.09%	-0.26%	-0.10%
	>250	0.03%	0.43%	0.79%	0.24%	1.24%	0.73%
up15bef_3m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	0.04%	0.08%	0.31%	0.18%	0.58%	-0.22%
	(50-60]	0.45%	0.50%	0.44%	-0.11%	0.99%	0.11%
	(60-80]	0.31%	-0.46%	-0.59%	0.14%	0.57%	0.79%
up15bef_3m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	-0.32%	0.15%	0.33%	-0.04%	-0.01%	0.38%
	(100-250]	0.57%	0.30%	0.25%	1.21%	-0.11%	-0.59%
	>250	0.08%	0.22%	0.10%	0.96%	0.43%	-0.38%
up15bef_1m_mean	(0-40] - Base	-	-	-	-	-	-
	(40-50]	-0.19%	0.13%	0.46%	0.28%	0.26%	-0.20%
	(50-60]	-0.03%	0.02%	-0.09%	-0.12%	-0.06%	0.25%
	(60-80]	0.00%	-0.35%	-0.31%	-0.29%	0.00%	0.66%
up15bef_1m_var	(0-50] - Base	-	-	-	-	-	-
	(50-100]	0.18%	-0.19%	0.17%	1.10%	-0.25%	-0.28%
	(100-250]	-0.59%	0.10%	0.09%	0.47%	0.06%	0.26%
	>250	-0.91%	0.10%	-0.26%	0.18%	0.19%	0.96%
Winter (Dec., Jan., or Feb.)	Yes - Base	-	-	-	-	-	-
	No	0.69%	0.73%	0.15%	-0.28%	-0.15%	0.47%
Weekday	Yes	2.62%	1.86%	0.52%	3.39%	0.70%	-2.03%
AM Peak (6 AM – 9 AM)	Yes	-3.03%	-2.00%	-0.21%	0.50%	-0.44%	-0.52%
PM Peak (4 PM – 7 PM)	Yes	1.50%	1.18%	0.30%	1.65%	0.10%	-1.11%
Police Presence	Yes	-4.36%	0.25%	1.80%	4.48%	4.00%	0.58%
Towing Presence	Yes	-5.80%	0.61%	0.99%	7.30%	5.51%	1.67%
Urban	Yes	0.99%	0.04%	-0.02%	0.84%	-0.31%	-3.46%
Incident Subtype	Disabled Veh. - Base	-	-	-	-	-	-
	Minor Crash	-7.19%	-1.54%	1.10%	8.12%	4.66%	0.60%
	Moderate Crash	-8.82%	-3.94%	-0.77%	3.35%	2.59%	2.99%
	Congestion	-11.86%	-3.53%	-0.95%	0.52%	2.86%	7.82%
	Abandoned Vehicle	4.42%	-2.70%	-2.82%	-6.78%	-1.30%	9.96%
	Major Crash	-4.82%	-3.08%	-1.00%	0.17%	3.07%	5.09%
Severity	Low - Base	-	-	-	-	-	-
	Medium	-1.28%	-1.13%	0.61%	8.42%	2.96%	3.18%

	High	-6.34%	-2.41%	-0.27%	1.98%	2.32%	13.20%
Number of Lanes	3-4 lanes - Base	-	-	-	-	-	-
	5-6 lanes	-1.07%	-0.43%	-0.13%	-1.28%	-0.30%	-0.10%
	> 6 lanes	-1.99%	-0.47%	0.00%	-0.28%	-0.68%	-0.70%
AADT	(0-20000] - Base	-	-	-	-	-	-
	(20000-40000]	0.52%	0.39%	0.27%	0.17%	-0.03%	-1.11%
	(40000-60000]	1.81%	-0.05%	0.33%	2.32%	-0.65%	-4.44%
	(60000-80000]	6.82%	3.38%	1.73%	1.06%	0.13%	-3.98%

8.4 Summary and Conclusion

To support Traffic Incident Management (TIM), research is critical to understand the relationships between the contributing factors and spatiotemporal impacts of traffic incidents. Unlike previous studies that relied on loop detector data from limited sites, the team exploited a large-scale network-wide crowdsourced probe vehicle data to investigate the traffic impacts of freeway incidents. The team created three queueing-related metrics, including the maximum queue length, time at the maximum queue length, and volume (spatiotemporal extent of a queue) to measure traffic incident impacts. The probe vehicle data is linked to traffic incidents and crash data to obtain the characteristics of traffic incidents. Other key factors, such as incident context and road environment, were retrieved from the incident database and HPMS data. To reduce the estimation bias from any single model, the team trained five machine learning models, including Categorical Naive Bayes (CNB), Support vector machine (SVM), Random Forest (RF), AdaBoost (Boost), and Neural network (NN), to interpret the nonlinear relationships between queueing-related metrics and relevant features. In total, 8,178 incidents that occurred on freeways in Alabama were analyzed. In addition, this research computed the marginal effects to quantify the magnitude of independent variables to gain insights into how contributing factors affect the spatiotemporal impacts of traffic incidents.

The modeling results indicate that three queueing-related metrics describing the spatiotemporal impacts of traffic incidents are highly correlated to traffic dynamics on freeways. The incident context and road environment are found to be important contributing factors influencing incident congestion and clearance time. The team extends the understanding of traffic incident impact correlates from a network-wide aspect, providing valuable insights for developing effective traffic and incident management strategies by using emerging crowdsourcing probe vehicle data.

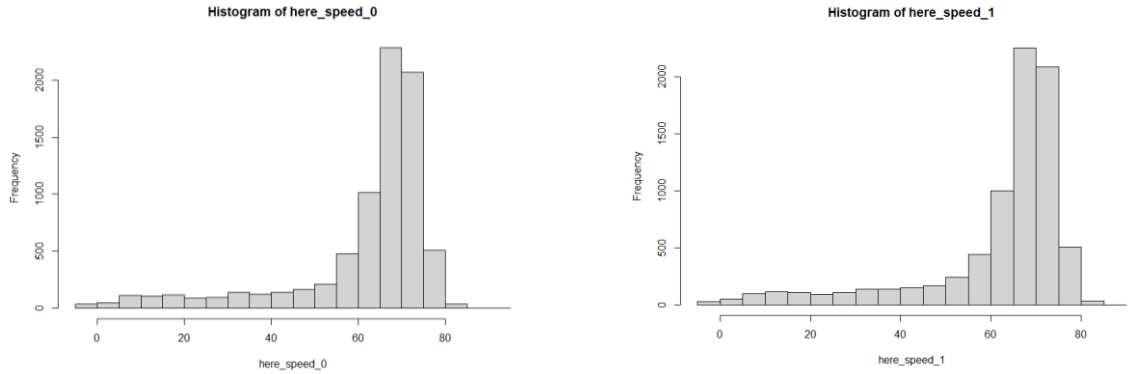
9. Injury Severity Modeling

9.1 Variable Creation

Using the Python speed extraction tool we developed before, we extracted two kinds of speed information from HERE as shown below. The speed information was linked to the corresponding crash records.

- here_speed_0: the speed of the crash location at the recorded crash time
- here_speed_1: the speed of the crash location before the maximum speed drops within 20 minutes before the recorded crash time.

The distributions of *here_speed_0* and *here_speed_1* are shown in FIGURE 30 (i) & (ii), respectively. We coded the data from the CARE crash database and HERE speed database into the categorical variables for inputting them into the Ordered Logistic Modeling. The descriptive statistics of the modeling variables are shown below in TABLE 29.



(i) Histogram of here_speed_0 (ii) Histogram of here_speed_1

FIGURE 30 Histogram of HERE speed variables

TABLE 28 Descriptive statistics of the modeling variables.

(i) Speed-related Variables

Variables		Frequency	Percentage	Variables		Frequency	Percentage
speed_report	00mph	57	0.77%	speed_here_1	00mph	27	0.36%
	01to05mph	96	1.30%		01to05mph	52	0.70%
	06to10mph	76	1.03%		06to10mph	97	1.31%
	11to15mph	68	0.92%		11to15mph	111	1.50%
	16to20mph	95	1.28%		16to20mph	102	1.38%
	21to25mph	63	0.85%		21to25mph	87	1.17%
	26to30mph	100	1.35%		26to30mph	105	1.42%
	31to35mph	99	1.34%		31to35mph	134	1.81%
	36to40mph	138	1.86%		36to40mph	132	1.78%
	41to45mph	188	2.54%		41to45mph	146	1.97%
	46to50mph	202	2.73%		46to50mph	156	2.11%
	51to55mph	248	3.35%		51to55mph	238	3.21%
	56to60mph	447	6.04%		56to60mph	418	5.64%
	61to65mph	747	10.09%		61to65mph	954	12.88%
	66to70mph	2069	27.94%		66to70mph	2143	28.94%
	71to75mph	373	5.04%		71to75mph	1991	26.88%
	75+mph	377	5.09%		75+mph	513	6.93%
	unknown	1963	26.51%	speed_limit	06to10mph	1	0.01%

v2speed_report	00mph	616	8.32%		16to20mph	2	0.03%
	01to05mph	146	1.97%		21to25mph	3	0.04%
	06to10mph	96	1.30%		36to40mph	6	0.08%
	11to15mph	91	1.23%		41to45mph	72	0.97%
	16to20mph	82	1.11%		46to50mph	162	2.19%
	21to25mph	58	0.78%		51to55mph	144	1.94%
	26to30mph	62	0.84%		56to60mph	303	4.09%
	31to35mph	69	0.93%		61to65mph	699	9.44%
	36to40mph	116	1.57%		66to70mph	621	8.39%
	00041to45mph	108	1.46%		71to75mph	5313	71.74%
	46to50mph	156	2.11%		75+mph	14	0.19%
	51to55mph	163	2.20%		unknown	66	0.89%
	56to60mph	288	3.89%				
	61to65mph	452	6.10%				
	66to70mph	824	11.13%				
	71to75mph	85	1.15%				
	no_v2	2777	37.50%				
	unknown	1217	16.43%				
speed_here_0	00mph	30	0.41%				
	01to05mph	47	0.63%				
	06to10mph	106	1.43%				
	11to15mph	100	1.35%				
	16to20mph	114	1.54%				
	21to25mph	85	1.15%				
	26to30mph	85	1.15%				
	31to35mph	138	1.86%				
	36to40mph	119	1.61%				
	41to45mph	135	1.82%				
	46to50mph	158	2.13%				
	51to55mph	209	2.82%				
	56to60mph	451	6.09%				
	61to65mph	961	12.98%				
	66to70mph	2179	29.42%				
	71to75mph	1971	26.61%				
	75+mph	518	6.99%				

(ii) Non-speed-related Variables.

Variables		Frequency	Percentage	Variables		Frequency	Percentage
Injury severity	PDO	6557	88.50%	Lighting condition	Daylight	5070	68.50%
	Possible	288	3.90%		Dawn dusk	285	3.80%
	Minor	401	5.40%		Dark cont. light	114	1.50%
	Serious	125	1.70%		Dark spotlight	399	5.40%
	Fatal	35	0.50%		Dark no light	1508	20.40%
Age	18-25	1873	25.30%		Other	30	0.40%
	26-35	1615	21.80%	Alignment	Straight m level	5453	73.60%
	36-45	1209	16.30%		Straight down	714	9.60%
	46-55	906	12.20%		Straight up	641	8.70%
	56-65	681	9.20%		Curve level	248	3.30%
	65+	452	6.10%		Curve grade	309	4.20%
	Unknown	670	9.00%		Other	41	0.60%
Gender	Male	4332	58.50%	Other vehicle types	Passenger car	1950	26.30%
	Female	2462	33.20%		SUV	1133	15.30%
	Unknown	612	8.30%		Pickup/mini-van	813	11.00%
Primary contributing circumstances	No improper	189	2.55%		Cargo/passenger van	71	1.00%
	DUI	167	2.25%		Truck/tractor	621	8.40%
	Defective equipment	433	5.85%		Other	2818	38.10%
	Distract/Inattention	1021	13.79%	Crash manner	Single vehicle	2653	35.80%
	Close following	1261	17.03%		Sideswipe	2638	35.60%

	Misjudge distance	384	5.18%		Sideswipe	1297	17.50%
	Fast	750	10.13%		Head on/Angle	161	2.20%
	Aggressive	96	1.30%		Side impact	232	3.10%
	Unsee-obstruct	552	7.45%		Other	425	5.70%
	Improper lane change	909	12.27%	Location	On road	5460	73.70%
	Improper steering/Swerved	697	9.41%		Median	624	8.40%
	Other	947	12.79%		Offroad	51	0.70%
Causal unit vehicle type	Passenger car	3582	48.40%		Roadside	707	9.50%
	SUV	1468	19.80%		Shoulder	529	7.10%
	Pickup/Mini-van	1299	17.50%		Other	35	0.50%
	Cargo/passenger van	102	1.40%				
	Truck/tractor	955	12.90%				
Seatbelt	Fully used	6389	86.30%				
	Partially used	13	0.20%				
	Not used	167	2.30%				
	Other	83	1.10%				
	Unknown	754	10.20%				

9.2 Crash Severity Models

We developed models with different speed variables (as shown in TABLE 29) for all crashes and single-vehicle crashes. To see how HERE speed data can be used to benefit modeling injured severity for freeway crashes, we compare the model goodness-of-fit using McFadden R square.

TABLE 29 Description of the speed variables.

Speed variable	Data source	Description
report_speed	CARE crash database	Estimated speed at the impact that is recorded in the CARE database;
here_speed_0	HERE speed database	Extracted HERE speed of the crash location at the recorded crash time;
here_speed_1	HERE speed database	Extracted HERE speed at the crash location before the big speed drop prior to the recorded crash time
speed_var_4	CARE crash database and HERE speed database	Estimated speed at the impact that is recorded in the CARE database with the missing values filled with the *here_speed_0*
speed_var_5	CARE crash database and HERE speed database	Estimated speed at the impact that is recorded in the CARE database with the missing values filled with the *here_speed_1*
speed_var_6	CARE crash database	Estimated speed at the impact that is recorded in the CARE database with the missing values filled with the speed limit

TABLE 30 shows the model we developed and the corresponding reported McFadden R square for all crash types. The results show that Model 5 with speed_var_5 (The estimated speed at the impact that is recorded in the CARE database with the missing values filled with the here_speed_1) achieves the highest McFadden R square. Based on these results, we suggest replacing the missing values in the reported speed with HERE speed at the crash location before the big speed drop before the recorded crash time that achieves the highest accuracy. The model estimation results are shown in TABLE 31.

TABLE 30 Model description from different injured severity models (all crash)

Model number	Speed variable used	Other variables	McFadden R square
1	report_speed	Driver's Age, Gender, Primary Contributing Circumstances, Causal Unit Vehicle Type, Other Vehicle Type, Crash Manner, Seatbelt, Location, Lighting, Alignment	0.1370
2	here_speed_0		0.1206
3	here_speed_1		0.1231
4	speed_var_4		0.1361
5	speed_var_5		0.1372
6	speed_var_6		0.1369

TABLE 31 Estimation Results for Model 5 (all crashes).

Variables		Coef.	Std. Error	p value
Age (Base: 18 - 25)	Age 26-35	0.082	0.108	0.445
	Age 36-45	-0.087	0.125	0.484
	Age 46-55	0.245	0.133	0.065
	Age 56-65	0.118	0.152	0.438
	Age 65+	0.147	0.176	0.401
	Age unknown	**-1.817**	**0.745**	**0.015**
Gender (Base: Male)	**Gender female**	**0.408**	**0.085**	**0.000**
	Gender unknown	**-3.062**	**1.229**	**0.013**
Primary Contributing Circumstances (Base: No improper)	DUI	0.486	0.335	0.147
	Defective Equipment	0.244	0.302	0.419
	distract/inattention	0.385	0.282	0.172
	Close following	-0.373	0.314	0.235
	Misjudge distance	0.217	0.341	0.525
	fast	0.119	0.286	0.677
	aggressive	0.426	0.412	0.301
	Unsee obstruct	**-0.854**	**0.338**	**0.011**
	Improper lane change	0.160	0.326	0.623
	Improper steering/Swerved	0.252	0.285	0.378
	other	**0.566**	**0.285**	**0.047**
Motorist Vehicle Type (Base: Passenger car)	SUV	-0.075	0.102	0.461
	pickup/minivan	**-0.402**	**0.126**	**0.001**
	cargo/passenger van	0.267	0.332	0.421
	truck/tractor	-0.282	0.166	0.090
Other Vehicle Type (Base: Passenger car)	SUV	0.113	0.162	0.488
	pickup/minivan	0.189	0.176	0.281
	cargo/passenger van	0.493	0.467	0.291
	truck/tractor	**1.241**	**0.160**	**0.000**
	Other	0.807	0.468	0.085
Crash Manner (Base: Single-vehicle crash)	sideswipe	0.584	0.338	0.084
	sideswipe	**-0.707**	**0.354**	**0.046**
	Head on/angle	**0.827**	**0.343**	**0.016**
	Side impact	**0.706**	**0.359**	**0.050**
	other	-0.190	0.305	0.533
Seatbelt (Base: Fully used)	Partially used	0.902	0.689	0.191
	Not used	**2.038**	**0.169**	**0.000**
	other	0.803	0.590	0.174

	unknown	**0.847**	**0.193**	**0.000**
Location (Base: on road)	**median**	**0.397**	**0.160**	**0.013**
	Off road	0.644	0.369	0.081
	Other	0.748	0.468	0.110
	roadside	**0.600**	**0.151**	**0.000**
	shoulder	**0.779**	**0.152**	**0.000**
Casual Unit Speed_report with missing information filled with speed_here_1	**speed_report_update00mph**	**1.901**	**0.594**	**0.001**
	speed_report_update06to10mph	-1.232	1.119	0.271
	speed_report_update11to15mph	0.594	0.665	0.372
	speed_report_update16to20mph	-1.073	0.894	0.230
	speed_report_update21to25mph	0.076	0.732	0.918
	speed_report_update26to30mph	-0.014	0.726	0.985
	speed_report_update31to35mph	-0.197	0.724	0.786
	speed_report_update36to40mph	0.970	0.585	0.097
	speed_report_update41to45mph	**1.172**	**0.561**	**0.037**
	speed_report_update46to50mph	**1.175**	**0.552**	**0.033**
	speed_report_update51to55mph	**1.451**	**0.533**	**0.006**
	speed_report_update56to60mph	**1.673**	**0.521**	**0.001**
	speed_report_update61to65mph	**1.552**	**0.512**	**0.002**
	speed_report_update66to70mph	**1.905**	**0.508**	**0.000**
	speed_report_update71to75mph	**1.762**	**0.521**	**0.001**
	speed_report_update75+mph	**2.530**	**0.522**	**0.000**
V2 Speed_report with missing information filled with speed_here_1	v2speed_report_update00mph	-0.233	0.372	0.531
	v2speed_report_update06to10mph	-0.912	0.630	0.148
	v2speed_report_update11to15mph	-0.078	0.495	0.875
	v2speed_report_update16to20mph	0.087	0.545	0.872
	v2speed_report_update21to25mph	0.225	0.532	0.672
	v2speed_report_update26to30mph	-0.344	0.601	0.568
	v2speed_report_update31to35mph	-0.294	0.546	0.590
	v2speed_report_update36to40mph	-0.860	0.491	0.080
	v2speed_report_update41to45mph	-0.421	0.746	0.573
	v2speed_report_update46to50mph	**-1.367**	**0.494**	**0.006**
	v2speed_report_update00041to45mph	-0.426	0.479	0.374
	v2speed_report_update51to55mph	**-0.876**	**0.442**	**0.047**
	v2speed_report_update56to60mph	**-0.974**	**0.411**	**0.018**
	v2speed_report_update61to65mph	**-0.959**	**0.389**	**0.014**
	v2speed_report_update66to70mph	**-0.862**	**0.383**	**0.024**
	v2speed_report_update71to75mph	**-1.181**	**0.469**	**0.012**
	v2speed_report_update75+mph	**-2.490**	**0.846**	**0.003**
	v2speed_report_updateno_v2	-1.032	0.652	0.114
Lighting condition (Base: Daylight)	Dawn dusk	-0.292	0.223	0.191
	Dark cont. light	0.380	0.295	0.198
	Dark spotlight	**0.530**	**0.159**	**0.001**
	Dark no light	0.177	0.098	0.071
	Other	0.635	0.593	0.284
Alignment (Base: straight level)	Straight down	0.031	0.133	0.816
	Straight up	-0.177	0.147	0.229
	Curve level	0.331	0.193	0.085
	Curve grade	0.158	0.178	0.374

	Other	0.271	0.445	0.542
Intercept	**1PDO|2possible**	**4.130**	**0.694**	**0.000**
	2possible|3minor	**4.656**	**0.695**	**0.000**
	3minor|4serious	**6.099**	**0.699**	**0.000**
	4serious|5fatal	**7.713**	**0.714**	**0.000**
Summary Statistics	**Number of observations**			**7406**
	AIC			**6393.015**
	McFadden Pseudo R^2			**0.137**

TABLE 32 shows the model we developed and the corresponding reported McFadden R square for single-vehicle crashes. The results show that Model 11 with speed_var_5 (The estimated speed at the impact that is recorded in the CARE database with the missing values filled with the here_speed_1) achieves the highest McFadden R square. Based on these results, we suggest replacing the missing values in the reported speed with HERE speed at the crash location before the big speed drop before the recorded crash time that achieves the highest accuracy. The model estimation results are shown in TABLE 33.

TABLE 32 Model description from different injured severity models (single vehicle crashes)

Model number	Speed variable used	Other variables	McFadden R square
7	report_speed	Driver's Age, Gender, Primary Contributing Circumstances, Causal Unit Vehicle Type, Other Vehicle Type, Crash Manner, Seatbelt, Location, Lighting, Alignment	0.095
8	here_speed_0		0.089
9	here_speed_1		0.088
10	speed_var_4		0.097
11	speed_var_5		0.097

TABLE 33 Estimation Results for Model 11 (single-vehicle crashes)

Variables		**Coef.**	**Std. Error**	**p value**
Age (base: 18 - 25)	26-35	-0.098	0.156	0.529
	36-45	-0.180	0.173	0.299
	46-55	0.244	0.183	0.181
	56-65	0.220	0.206	0.285
	65+	0.387	0.230	0.092
	Unknown	-1.019	1.149	0.375
Gender (base: male)	Male	0.258	0.118	0.029
	Unknown	-2.589	1.516	0.088
Primary Contributing Circumstances (Base: No improper)	DUI	0.539	0.399	0.177
	Defective Equipment	0.247	0.335	0.461
	Distract/inattention	0.434	0.317	0.171
	Close following	-1.469	1.086	0.176
	Fast	0.119	0.317	0.708
	Aggressive	0.910	0.537	0.090
	Unsee obstruct	**-1.272**	**0.407**	**0.002**
	Improper lane change	0.745	0.575	0.195
	Improper steering/Swerved	0.216	0.315	0.494
	Other	**0.639**	**0.324**	**0.049**
	SUV	0.183	0.139	0.186
	Pickup/minivan	0.036	0.171	0.835

Motorist Vehicle Type (Base: Passenger car)	Cargo/passenger van	0.008	0.522	0.987
	Truck/tractor	0.038	0.230	0.869
Seatbelt (Base: Fully used)	Partially used	0.675	0.896	0.451
	Not used	**1.927**	**0.215**	**0.000**
	Other	-0.453	1.123	0.687
	Unknown	0.498	0.313	0.111
Location (Base: on the road)	Median	0.315	0.184	0.086
	Offroad	0.651	0.382	0.088
	Other	0.529	0.500	0.290
	Roadside	**0.549**	**0.175**	**0.002**
	Shoulder	**0.737**	**0.185**	**0.000**
Speed_report with missing information filled with speed_here_1	Speed_report_update41to45mph	-1.135	1.200	0.345
	Speed_report_update46to50mph	-0.998	0.966	0.301
	Speed_report_update51to55mph	0.503	0.716	0.482
	Speed_report_update56to60mph	0.603	0.668	0.367
	Speed_report_update61to65mph	0.562	0.650	0.387
	Speed_report_update66to70mph	0.939	0.638	0.141
	Speed_report_update71to75mph	0.819	0.654	0.210
	Speed_report_update75+mph	**1.510**	**0.655**	**0.021**
Lighting condition (Base: Daylight)	Dawn dusk	-0.587	0.328	0.073
	Dark cont. light	0.283	0.459	0.538
	Dark spotlight	0.443	0.245	0.071
	Dark no light	0.076	0.130	0.557
	Other	0.992	0.757	0.190
Alignment (Base: straight level)	Straight down	0.060	0.176	0.733
	Straight up	-0.394	0.219	0.073
	Curve level	**0.545**	**0.221**	**0.014**
	Curve grade	0.112	0.226	0.620
	Other	0.872	0.658	0.185
Intercept	**PDO | possible**	**3.291**	**0.702**	**0.000**
	Possible | minor	**3.837**	**0.703**	**0.000**
	Minor | serious	**5.284**	**0.711**	**0.000**
	Serious | fatal	**7.120**	**0.749**	**0.000**
Summary statistics	**Number of observations**			**2653**
	AIC			**3224.563**
	McFadden Pseudo R^2			**0.097**

9.3 Summary and Conclusion

This section replaces the missing reported estimated speed at impact with the speed information extracted from HERE speed and compares the modeling results regarding the goodness-of-fit. In comparison, we also replace the missing reported estimated speed at impact with the speed limit documented in the CARE crash report. The modeling results show that replacing the missing reported estimated speed at impact with the speed information extracted from HERE speed can improve the goodness of fit of injured severity models.

10. Summary and Recommendations

This report documents the work conducted by the team to take advantage of the existing state-wide large-scale database to develop a series of tools and models to assist proactive traffic incident management in Alabama. The major tasks completed in this project are as follows: 1) review of related work of state practice and scholarly research, 2) data collection and data processing, 3) development of crash risk prediction model， 5) development of incident detection model, 6) development of incident impact model, 7) development of injury severity model. Additionally, the real-time spatial-temporal speed extraction tool is also provided in the report.

Specifically, the team makes use of the ALGO traffic incident database, the CARE crash database, the HERE traffic database, and the publicly accessible HPMS database. A real-time spatial-temporal speed extraction tool was created in Python which enables extraction and export of the spatial-temporal speed matrix in 0.1 miles * 1-minute resolution given incident/crash unique ID in the ALGO/CARE database. Based on the spatial-temporal speed matrix generated for each incident/crash, the team created various traffic characteristics variables (e.g., average speed, speed variances, upstream and downstream speed differences, etc.) and spatial-temporal speed matrix given the predefined time range and spatial coverage. The team also extracted static environmental information from the HPMS database to provide supplementary road geometry and land use information.

Regarding crash risk prediction, machine learning models are developed to predict the occurrence of crashes based on the pre-crash traffic characteristics. We built models for all types of crashes and subtypes of crashes. The results show that rear-end crashes are more predictable using the traffic characteristics provided by the probe vehicle-based HERE data. The partial dependent plots identified the non-linearity between the precrash traffic variable and the crash risk. Specifically，according to the estimated partial dependence, the rear-end crash risk is positively related to the speed variance and speed reductions. A higher rear-end crash risk is associated with a more significant speed variance upstream, and the risk for a rear-end crash increases when the traffic speed at this location decreases significantly. Additional work was conducted to map the high-crashes frequency sites. The results show that, for all-type crash prediction models, generally, models for high-density crash freeway segments show better accuracy regardless of the model type. No significant improvement is shown for models separated by type.

For automatic incident detection (AID), taking advantage of the spatial and temporal coverage of the speed information provided by the HERE data, this project develops the CNN model to detect the occurrence of incidents at the state level. The results indicated that AID models detecting the occurrence of traffic-impacted incidents have better-predicted accuracy than AID models detecting the occurrence of all types of traffic incidents, which implies that incorporating all types of incidents into the AID model will weaken the prediction performance of the AID model. For AID models that can classify the incident sub-types (crash, congestion, and other traffic-impacted incidents), models with a balanced dataset can achieve higher performance compared to models without balancing. The mapping results of the incident locations show the spatial uneven distribution of the incident location. High-frequency incident segments are identified and provided in the report.

This report also documents the team's effort to evaluate the impact of the traffic incident by exploring the influence of the occurrence of the incident on the maximum queue length, time of the maximum queue, and the drop of the traffic volume within the defined time and space ranges. We identified the factors associated with these three metrics. The modeling results indicate that three queueing-related metrics describing the spatiotemporal impacts of traffic incidents are highly correlated to traffic dynamics on freeways. The incident context and road environment are found to be important contributing factors influencing incident congestion and clearance time. The team extends the understanding of traffic incident impact correlates from a network-wide aspect, providing valuable insights for developing effective traffic and incident management strategies by using emerging crowdsourcing probe vehicle data.

Lastly, the team also explored the potential of using the HERE speed information to improve the crash severity modeling. The team replaced the missing reported estimated speed at impact with the speed information extracted from HERE speed and compared the modeling results regarding the goodness of fit. In comparison, we also replace the missing reported estimated speed at impact with the speed limit documented in the CARE crash report. The modeling results show that replacing the missing reported estimated speed at impact with the speed information extracted from HERE speed can improve the goodness of fit of injured severity models.

In the future, more connected vehicle data will become available with higher vehicle concentration and wider coverage, and there is more potential to use real-time traffic information to facilitate proactive traffic incident management practice. Besides, with the fast development of the artificial intelligence model and computation ability, more model types and structures can be tested to explore the possibility of facility the traffic incident management practice to prevent the occurrence of incidents, speed up the detection of the incidents, and shorten the clearance time of the incident.

Appendix A

Speed Extraction Tool in Python code

```
##Input crash

import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
os.chdir(r"F:\PHD Project\TIM\Data process\Crash_speed_extract\I-20")

#Event_I-65_Incident_January_North.csv
#Event_I-65_Crash_January_North.csv
eventrec = pd.read_csv("2019-2020 ALGO Crash_I20_I59.csv")
options = ['High', 'Medium']

# selecting rows based on condition
eventrec = eventrec.loc[eventrec['Last Severity'].isin(options)]

##Filter road and direction first
event = eventrec[["Event ID","DateTime","Mile Marker",'Direction','Primary Road','Incident
Type','Incident Subtype']]
options = ['Major Crash', 'Overturned Vehicle','Moderate Crash']
event = event.loc[event['Incident Subtype'].isin(options)]
##Notice：Northbound： From < To; Southbound: To < From
##Notice：Eastbound： From < To; Westbound: To < From
options = ['East']
event = event.loc[event['Direction'].isin(options)]
event.head()

###Connect to HERE database
import pyodbc
[x for x in pyodbc.drivers()]
driver = "{ODBC Driver 17 for SQL Server}"
server = "XXX"  # server name
username = "XXX"
password = "XXX!"
database = "XXX"  # database name
conn =
pyodbc.connect('DRIVER={};SERVER={};DATABASE={};UID={};PWD={}'.format(drive
r, server, database, username, password))

"""Function get_record: Input: item in gp; Output speed record in one minute with (-a miles, b
miles)"""

def get_speed_matrix(speed_TABLE,event_location,a,b):
```

```
    gp = speed_TABLE.groupby(by = 'tstamp')
    speed_matrix = pd.DataFrame()

    def get_record(minute):
        minute = list(minute)[1]
        minute.reset_index(inplace=True)
        tmrange = End - Start + 1
        record = list(np.zeros(tmrange))
        for i in range(len(minute)):
            startp = int(minute.loc[i,"MF"]-Start)
            endp =  int(minute.loc[i,"MT"]-Start)
            speed = minute.loc[i,"speed"]
            for j in range(startp,endp):
                record[j] = speed
        event_L = int(round(event_location,1)*10)
        start_L = event_L - a * 10 - Start
        end_L = event_L + b * 10 - Start
        record = record[start_L:end_L+1]
        return record

    for minute in gp:
        record = get_record(minute)
        record = pd.DataFrame([record])
        speed_matrix = pd.concat([speed_matrix,record],axis = 0)

    return speed_matrix

"""
If the speed_matrix is 3 miles before incident - IncidentLc = 3 + 0.1
Bef, Aft - the spatial range
"""
def HDSM(IncidentLc,a,b,ff, speed_matrix):
    IncidentLc = IncidentLc * 10
    Bef = int(IncidentLc - a * 10)
    Aft = int(IncidentLc + b * 10)
    for i in range(len(speed_matrix)):
        score = 0
        for j in range(a, b):
            score = score + (ff - speed_matrix.iloc[i,j])
        speed_matrix.loc[i,"score"] = score
    return speed_matrix


os.chdir(r"F:\PHD Project\TIM\Data process\Crash_speed_extract")
#Query TMC - Upstream a miles; Downstream b miles;
# search and change: a = 10,b = 5
```

```
##Notice：Northbound： From < To; Southbound: To < From
##Notice：Eastbound： From < To; Westbound: To < From
#i = i
for i in range(len(event)):
    record = event.iloc[i]
    event_ID = record["Event ID"]
    Incident_type = record["Incident Type"]
    Incident_Subtype = record["Incident Subtype"]
    event_location = record["Mile Marker"]
    a = 10   #Upper bound - 10 miles before
    b = 5   #Lower bound - 5 miles after
    Upmm = str(round(record["Mile Marker"]-a,2))
    Dnmm = str(round(record['Mile Marker']+b,2))

    if float(Upmm) < 0:
        continue
    sql = "select
Id,ROAD_NUM ,ROAD_DIR,GeometryLength,MeasureAscending,MeasureFrom ,MeasureT
o from InterstateRouteGeometry where ROAD_NUM='I-65' and ROAD_DIR = 'Northbound'"
    TMC =  pd.read_sql_query(sql, conn)
    Maxmm = TMC["MeasureTo"].max()
    if float(Dnmm) > Maxmm:
        continue

    sql = "select
Id,ROAD_NUM ,ROAD_DIR,GeometryLength,MeasureAscending,MeasureFrom ,MeasureT
o from InterstateRouteGeometry where ROAD_NUM='I-65' and ROAD_DIR = 'Northbound'
and MeasureFrom <= "+ Upmm + " and MeasureTo >= "+ Upmm + "order by MeasureTo"
    UpTMC = pd.read_sql_query(sql, conn)
    sql = "select
Id,ROAD_NUM ,ROAD_DIR,GeometryLength,MeasureAscending,MeasureFrom ,MeasureT
o from InterstateRouteGeometry where ROAD_NUM='I-65' and ROAD_DIR = 'Northbound'
and MeasureFrom <= "+ Dnmm + " and MeasureTo >= "+ Dnmm + "order by MeasureTo"
    DnTMC = pd.read_sql_query(sql, conn)
    Startmm = str(round(UpTMC.loc[0,"MeasureFrom"]-0.01,2))
    Endmm = str(round(DnTMC.loc[0,"MeasureFrom"]+0.01,2))

    ##Query TMC incuding a miles upstream and b miles downstream
    sql = "select
Id,ROAD_NUM ,ROAD_DIR,GeometryLength,MeasureAscending,MeasureFrom ,MeasureT
o,_TMC from InterstateRouteGeometry where ROAD_NUM='I-65' and ROAD_DIR =
'Northbound' and MeasureFrom >= "+ Startmm + " and MeasureFrom <= "+ Endmm + " order
by MeasureTo"
    TMC = pd.read_sql_query(sql, conn)
    TMC["Index"]= [item+1 for item in range(len(TMC))]
    print(TMC["_TMC"])
```

```
    TMC_query = TMC.rename(columns={"_TMC":"TMC"})
    TMC_query = TMC_query[["MeasureFrom","MeasureTo","TMC","Index"]]

    ##Query speed including
    #Spatical range
    TMCall = "'" + TMC.loc[0,"_TMC"] + "'"
    for j in range(1,len(TMC)):
        tem = "'" + TMC.loc[j,"_TMC"] + "'"
        TMCall = TMCall + ',' + tem
    print(TMCall)
    #Temporal range
    from datetime import datetime
    TimeStr = record["DateTime"]   #####
    Time = datetime.strptime(TimeStr, "%m/%d/%Y %H:%M")
    #1)Upper bound
    c = 60  # c minutes before accident
    d = 60  # d minutes after accident


    import datetime
    TimeUp = Time - datetime.timedelta(minutes = c)
    from datetime import datetime
    TimeUp = "'" + datetime.strftime(TimeUp,"%Y-%m-%d %H:%M:%S") + "'"
    #2)Lower bound
    import datetime
    TimeDb = Time + datetime.timedelta(minutes = d)
    from datetime import datetime
    TimeDb = "'"+ datetime.strftime(TimeDb,"%Y-%m-%d %H:%M:%S") + "'"

    sql = "select * from speeds where TMC in (" + TMCall + ") and tstamp >=" + TimeUp + "
and tstamp <= " + TimeDb + "order by tstamp,TMC"
    speed_Query = pd.read_sql_query(sql, conn)
    if len(speed_Query) == 0:
        continue

    speed_TABLE = speed_Query.merge(TMC_query,on="TMC",how = 'left')
    speed_TABLE.sort_values(by=['tstamp','Index'],inplace=True)
    ##Update MeasureFrom，Measure To consider dynamic sub-segment
    speed_TABLE.fillna(0,inplace=True)
    speed_TABLE["MeasureFrom"] = speed_TABLE["MeasureFrom"] +
speed_TABLE["offset"]

    ##when sub_len != 0 --> Dynamic sub-segment
    speed_TABLE.loc[speed_TABLE["sub_len"]!= 0,"MeasureTo"] =
speed_TABLE.loc[speed_TABLE["sub_len"]!= 0,"MeasureFrom"] +
speed_TABLE.loc[speed_TABLE["sub_len"]!= 0,"sub_len"]
```

```
    ##Set 0.1 mile as slice
    speed_TABLE["MF"]=round(speed_TABLE["MeasureFrom"],1)*10
    speed_TABLE["MT"]=round(speed_TABLE["MeasureTo"],1)*10


    ##Speed matrix
    ##Initialization
    Start = int(speed_TABLE["MF"].min())
    End = int(speed_TABLE["MT"].max())

    ##Get speed_matrix
    speed_matrix = get_speed_matrix(speed_TABLE,event_location,a = 10,b = 5)

    ##Visulization
    speed_matrix.index = list(range(-c,len(speed_matrix)-c))
    speed_matrix.columns = list(range(-a*10,len(-speed_matrix.columns)-a*10))
    fig,ax=plt.subplots(1,1,figsize=(12,6))
    Ftitle = str(event_ID) + '_' + Incident_type + '_' + Incident_Subtype + ' at Mile Marker ' +
str(event_location) + ' ' + TimeStr
    plt.title(Ftitle) #FIGURE titile
    sns.heatmap(speed_matrix,cmap='YlOrRd_r', vmin=20, vmax=80)
    plt.xlabel("Distance (Unit:0.1 mile)")
    plt.ylabel("Time (Unit: 1 minute)")
    plt.scatter(x = 50, y = 30, marker="+",color = 'k',s = 180,label = "Crash",linewidths = 3)
    plt.legend()
    filename = Incident_type + '_' + Incident_Subtype + '_' + 'Speed_matrix_for_Event_'+
str(event_ID) +'.png'
    fig.savefig(filename)
    plt.clf() # to save memory

    ##Calculate SDMH score
    speed_matrix.reset_index(inplace=True)
    speed_matrix = HDSM(IncidentLc=10.1,a = 10,b = 5,ff=70,speed_matrix = speed_matrix)

    ##Save to csv
    file_name = Incident_type + '_' + Incident_Subtype + '_' + 'Speed_matrix_for_event'+
str(event_ID) + '.csv'
    speed_matrix.to_csv(file_name)
```

References

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention, 32(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abdelrahman, A., Abu-Ali, N., & Hassanein, H. S. (2017). Driver Behavior Classification in Crash and Near-Crash Events Using 100-CAR Naturalistic Data Set. 2017 IEEE Global Communications Conference, GLOBECOM 2017 - Proceedings, 2018-January 1–6. https://doi.org/10.1109/GLOCOM.2017.8253921

Ahmed, F., &Hawas,Y. E. (2012).Athreshold-based real-time incident detection system for urban traffic networks. Procedia—Social and Behavioral Sciences, 48, 1713–1722.

Ahmed, M. and Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. Transportation Research Part C: Emerging Technologies, 26, pp.203-213.

Ahmed, S. A., & Cook, A. R. (1982). Application of time-series analysis techniques to freeway incident detection. Transportation Research Record, 841(3), 19-21.

Alabama Department of Transportation (ALDOT), 2014. Alabama Traffic Incident Management Guidelines. Online at: https://algotraffic.com/Content/documents/alabama_TIM_Guidelines_v12_091516.pdf

Alabama Traffic Incident Management, 2020. Traffic Incident Management (TIM). Online at: https://alabamatim.org/.

Anuar, K., Habtemichael, F., & Cetin, M. (2015). Estimating Traffic Flow Rate on Freeways from Probe Vehicle Data and Fundamental Diagram. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2015-October 2921–2926. https://doi.org/10.1109/ITSC.2015.468

Bae, B., Liu, Y., Han, L.D. and Bozdogan, H., 2019. Spatio-temporal traffic queue detection for uninterrupted flows. Transportation Research Part B: Methodological, 129, pp.20-34.

Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. Transportation Research Part C: Emerging Technologies, 86, 202–219. https://doi.org/10.1016/J.TRC.2017.11.014

Bishop, C. M. (2006). Pattern Recognition and Machine Learning.

Breiman, L. (2001). Random Forests. Machine Learning 2001 45:1, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J. and Wu, Y., 2020. Real-time crash prediction on expressways using deep generative models. Transportation research part C: emerging technologies, 117, p.102697.

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. Transportation Research Part C: Emerging Technologies, 117(June), 102697. https://doi.org/10.1016/j.trc.2020.102697

CDC. (2020). Global Road Safety | CDC. https://www.cdc.gov/injury/features/global-road-safety/index.html

Chatterjee, I., & Davis, G. A. (2016). Analysis of Rear-End Events on Congested Freeways by Using Video-Recorded Shock Waves: Https://Doi.Org/10.3141/2583-14, 2583, 110–118. https://doi.org/10.3141/2583-14

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Liu, X.C. and Zhang, G., 2016. Non-recurrent congestion analysis using data-driven spatiotemporal approach for information construction. Transportation Research Part C: Emerging Technologies, 71, pp.19-31.

Chung, Y. and Recker, W.W., 2013. Spatiotemporal analysis of traffic congestion caused by rubbernecking at freeway accidents. IEEE Transactions on Intelligent Transportation Systems, 14(3), pp.1416-1422.

Chung, Y., 2013. Identifying primary and secondary crashes from spatiotemporal crash impact analysis. Transportation research record, 2386(1), pp.62-71.

D'Andrea, E. and Marcelloni, F., 2017. Detection of traffic congestion and incidents from GPS trace analysis. Expert Systems with Applications, 73, pp.43-56.

Delaware Valley Regional Planning Commission, 2020a. Interactive Detour Route Mapping (IDRuM). Online at: https://www.dvrpc.org/Transportation/TSMO/IDRuM/#googtrans/en/zh.

Delaware Valley Regional Planning Commission, 2020b. Regional Integrated Multi-Modal Information Sharing (RIMIS) Project. Online at: dvrpc.org/Transportation/TSMO/RIMIS/#googtrans/en/zh.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. Advances in neural information processing systems, 9, 155-161.

Drucker, H., Burges, C.J., Kaufman, L., Smola, A. and Vapnik, V., 1996. Support vector regression machines. Advances in neural information processing systems, 9.

Esri. (2020). HERE data layers—ArcGIS StreetMap Premium | Document. https://doc.arcgis.com/zh-cn/streetmap-premium/get-started/dd-here-data.htm

Fabian, P., Michel Vincent, Olivier, G., Mathieu, B., Peter, P., Ron, W., Vanderplas Jake, & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. In Journal of Machine Learning Research (Vol. 12, Issue 85). https://scikit-learn.org/sTABLE/inspection.html

Federal Highway Administration (FHWA), 2010. Traffic incident management handbook (No. FHWA-HOP-10-013). U.S. Department of Transportation, Federal Highway Administration, Office of Transportation Operations.

FHWA (2010). Best Practices in Traffic Incident Management. Available online at: https://ops.fhwa.dot.gov/publications/fhwahop10050/ch2.htm

FHWA, 2020. Traffic Incident Management. Online at: https://ops.fhwa.dot.gov/eto_tim_pse/about/tim.htm.

FHWA. (2017). Traffic Control Systems Handbook: Chapter 6 . Detectors. 1–12. https://ops.fhwa.dot.gov/publications/fhwahop06006/chapter_6.htm

FHWA. (2018). Traffic Incident Management (TIM) Performance Measurement: On the Road to Success (Issue FHWA-HOP-10-009). https://ops.fhwa.dot.gov/publications/fhwahop10009/tim_fsi.htm

FHWA. (2020) Federal Highway Administration Focus States Initiative: Traffic Incident Management Performance Measures Final Report. Available online at: https://ops.fhwa.dot.gov/publications/fhwahop10010/presentation.htm

FHWA. (2020). 2020 Alabama Highway Safety Improvement Program.

FHWA. (2021). EDC-5: Crowdsourcing for Operations | Federal Highway Administration. https://www.fhwa.dot.gov/innovation/everydaycounts/edc_5/crowdsourcing.cfm

FHWA. 2020. Traffic Incident Management. Available online at: https://ops.fhwa.dot.gov/plan4ops/traffic_incident.htm#:~:text=Traffic%20incident%20management%20(TIM)%20consists,safely%20and%20quickly%20as%20possible.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research, 20, 1–81. http://jmlr.org/papers/v20/18-760.html.

Florida TIM Responder, 2020a. Traffic Incident Management Efficient Traffic Rerouting and Agency Coordination. Online at: http://www.floridatim.com/TIM%20Newsletter/Florida%20TIM%20Responder%20-%20June%202020.pdf.

Florida TIM Responder, 2020b. Florida TIM Responder - September 2020. Online at: http://www.floridatim.com/TIM%20Newsletter/Florida%20TIM%20Responder%20-%20September%202020.pdf.

Freund, Y., Schapire, R. and Abe, N., 1999. A short introduction to boosting. Journal-Japanese Society for Artificial Intelligence, 14(771-780), p.1612.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society for Artificial Intelligence, 14(771-780), 1612.

Gakis, E., Kehagias, D. and Tzovaras, D., 2014, October. Mining traffic data for road incidents detection. In 17th International IEEE Conference on Intelligent Transportation Systems (ITSC) (pp. 930-935). IEEE.

Gakis, E., Kehagias, D., & Tzovaras, D. (2014). Mining traffic data for road incidents detection. 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014, 2, 930–935. https://doi.org/10.1109/ITSC.2014.6957808

Ghosh, B., & Smith, D. P. (2014). Customization of Automatic Incident Detection Algorithms for Signalized Urban Arterials. Journal of Intelligent Transportation Systems, 18(4), 426–441. https://doi.org/10.1080/15472450.2013.806843

Green, E. R., Pigman, J. G., Walton, J. R., & McCormack, S. (2012). Identification of Secondary Crashes and Recommended Countermeasures to Ensure More Accurate Documentation.

GS&P, 2014. Alabama Traffic Incident Management Guidelines. Online at: https://algotraffic.com/Content/documents/alabama_TIM_Guidelines_v12_091516.pdf.

Gu, Y., Qian, Z.S. and Chen, F., 2016. From Twitter to detector: Real-time traffic incident detection using social media data. Transportation research part C: emerging technologies, 67, pp.321-342.

Hall, Fred L., Shi, Yong, Atala, George, 1993. On-line testing of the mcmaster incident detection algorithm under recurrent congestion. Transport. Res. Rec. 1394, 1–7

Hastie, T., Rosset, S., Zhu, J. and Zou, H., 2009. Multi-class adaboost. Statistics and its Interface, 2(3), pp.349-360.

HERE, 2021. Road Traffic Analytics & Location Intelligence | HERE. https://www.here.com/platform/traffic-solutions/real-time-traffic-information. Accessed June 1, 2022.

HERE. (2021). Road Traffic Analytics & Location Intelligence | HERE. https://www.here.com/platform/traffic-solutions/real-time-traffic-information

HERE. (2021). Road Traffic Analytics & Location Intelligence | HERE. https://www.here.com/platform/traffic-solutions/real-time-traffic-information.

Hojati, A. T., Ferreira, L., Washington, S., & Charles, P. (2013). Hazard based models for freeway traffic incident duration. Accident Analysis and Prevention, 52, 171–181. https://doi.org/10.1016/j.aap.2012.12.037

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. https://doi.org/10.1016/J.AAP.2011.08.004

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. Accident Analysis and Prevention, 124(July 2018), 66–84. https://doi.org/10.1016/j.aap.2018.12.022

Huang, T., Wang, S. and Sharma, A., 2020. Highway crash detection and risk estimation using deep learning. Accident Analysis & Prevention, 135, p.105392.

Huang, T., Wang, S., & Sharma, A. (2020). Highway crash detection and risk estimation using deep learning. Accident Analysis & Prevention, 135, 105392.

Huang, T., Wang, S., & Sharma, A. (2020). Highway crash detection and risk estimation using deep learning. Accident Analysis and Prevention, 135(December 2019), 105392. https://doi.org/10.1016/j.aap.2019.105392

Huang, Z., Arian, A., Yuan, Y., and Chiu, Y.C., 2020. Using conditional generative adversarial nets and heat maps with simulation-accelerated training to predict the spatiotemporal impacts of highway incidents. Transportation research record, 2674(8), pp.836-849.

INRIX, Inc., 2019. INRIX 2018 Global Traffic Scorecard. Online at: http://inrix.com/scorecard/

Islam, N., Hainen, A., Burdette, S., Jones, S., & Smith, R. (2021). An analytical assessment of freeway service patrol on incident clearance times. Advances in Transportation Studies, 54, 75–88. https://doi.org/10.53136/97912599405446

Karim, Asim, Adeli, Hojjat, 2002a. Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm. J. Transport. Eng. 128 (1), 21–30.)

Khattak, A. J., Wang, X., & Zhang, H. (2010). Spatial analysis and modeling of traffic incidents for proactive incident management and strategic planning. Transportation Research Record, 2178, 128–137. https://doi.org/10.3141/2178-14

Kitali, A.E., Alluri, P., Sando, T. and Lentz, R., 2019. Impact of primary incident spatiotemporal influence thresholds on the detection of secondary crashes. Transportation research record, 2673(10), pp.271-283.

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. Current Biology, 29(7), R231-R236.

Li, L., Lin, Y., Du, B., Yang, F., & Ran, B. (2020). Real-time traffic incident detection based on a hybrid deep learning model. Transportmetrica A: Transport Science, 1-21.

Li, L., Qu, X., Zhang, J. and Ran, B., 2017. Traffic incident detection based on extreme machine learning. Journal of Applied Science and Engineering, 20(4), pp.409-416.

Li, L., Qu, X., Zhang, J., & Ran, B. (2017). Traffic incident detection based on extreme machine learning. Journal of Applied Science and Engineering, 20(4), 409–416. https://doi.org/10.6180/jase.2017.20.4.01

Li, X., Liu, J., Khattak, A., & Nambisan, S. (2020). Sequential Prediction for Large-Scale Traffic Incident Duration: Application and Comparison of Survival Models: Https://Doi.Org/10.1177/0361198119899041, 2674(1), 79–93. https://doi.org/10.1177/0361198119899041

Lin, Y., Li, L., Jing, H., Ran, B. and Sun, D., 2020. Automated traffic incident detection with a smaller dataset based on generative adversarial networks. Accident Analysis & Prevention, 144, p.105628.

Lin, Y., Li, L., Jing, H., Ran, B., & Sun, D. (2020). Automated traffic incident detection with a smaller dataset based on generative adversarial networks. Accident Analysis and Prevention, 144(September 2019), 105628. https://doi.org/10.1016/j.aap.2020.105628

Liu, C., & Ye, T. J. (2011). Run-Off-Road Crashes: An On-Scene Perspective. www.ntis.gov

Liu, C., Zhang, S., Wu, H. and Fu, Q., 2017. A dynamic spatiotemporal analysis model for traffic incident influence prediction on urban road networks. ISPRS International Journal of Geo-Information, 6(11), p.362.

Liu, J., Khattak, A., & Zhang, M. (2016). What Role Do Precrash Driver Actions Play in Work Zone Crashes? Application of Hierarchical Models to Crash Data: Https://Doi.Org/10.3141/2555-01, 2555, 1–11. https://doi.org/10.3141/2555-01

Lu, W., Liu, J., Fu, X., Yang, J. and Jones, S., 2022. Integrating machine learning into path analysis for quantifying behavioral pathways in bicycle-motor vehicle crashes. Accident Analysis & Prevention, 168, p.106622.

Mannino, N. (2021). What Causes Traffic Jams? https://extramile.thehartford.com/auto/what-causes-traffic-jams/

Miller, M. and Gupta, C., 2012, August. Mining traffic incidents to forecast impact. In Proceedings of the ACM SIGKDD international workshop on urban computing (pp. 33-40).

Molnar, C. (2020). InterpreTABLE machine learning. Lulu.com.

Motamed, M., & Machemehl, R. (2014). Real time freeway incident detection (No. SWUTC/14/600451-00083-1). Texas A&M Transportation Institute.

Motamed, M., & Machemehl, R. (2014). Real time freeway incident detection. 7(2), 42.

Naranjo, J. E., Jiménez, F., Serradilla, F. J., & Zato, J. G. (2012). Floating car data augmentation based on infrastructure sensors and neural networks. IEEE Transactions on Intelligent Transportation Systems, 13(1), 107–114. https://doi.org/10.1109/TITS.2011.2180377

Naranjo, J. E., Jiménez, F., Serradilla, F. J., & Zato, J. G. (2012). Floating car data augmentation based on infrastructure sensors and neural networks. IEEE Transactions on Intelligent Transportation Systems, 13(1), 107–114. https://doi.org/10.1109/TITS.2011.2180377

NCHRPTIMPM, 2020. Overview of TIM Program. Online at: http://nchrptimpm.timnetwork.org/?page_id=83.

Nevada TIM Coalition, 2020a. About Us. Online at: http://nvtim.com/about-us/.

Nevada TIM Coalition, 2020b. TIM INITIATIVES. Online at: http://nvtim.com/tim-initiatives/.

NYC Gov, 2020. Citywide Incident Management System. Online at: https://www1.nyc.gov/site/em/about/citywide-incident-management-system.page.

Ou, J., Xia, J., Wang, Y., Wang, C. and Lu, Z., 2020. A data-driven approach to determining freeway incident impact areas with fuzzy and graph theory-based clustering. Computer-Aided Civil and Infrastructure Engineering, 35(2), pp.178-199.

Owens, N., Armstrong, A., Sullivan, P., Mitchell, C., Newton, D., Brewster, R., & Trego, T. (2010). Traffic incident management handbook.

Pan, B., Demiryurek, U., Gupta, C. and Shahabi, C., 2015. Forecasting spatiotemporal impact of traffic incidents for next-generation navigation systems. Knowledge and Information Systems, 45(1), pp.75-104.

Park, H., & Haghani, A. (2016). Real-time prediction of secondary incident occurrences using vehicle probe data. Transportation Research Part C: Emerging Technologies, 70, 69–85. https://doi.org/10.1016/J.TRC.2015.03.018

Proper, A. T. (1999). Intelligent transportation systems benefits: 1999 update (No. FHWA-OP-99-012). United States. Joint Program Office for Intelligent Transportation Systems.

Qu, X., Wang, W., Wang, W., & Liu, P. (2011). Real-time freeway sideswipe crash prediction by support vector machine; Real-time freeway sideswipe crash prediction by support vector machine. IET Intell. Transp. Syst, 7(4), 445. https://doi.org/10.1049/iet-its.2011.0230

Qu, X., Wang, W., Wang, W., Liu, P., & Noyce, D. A. (2012). Real-Time Prediction of Freeway Rear-End Crash Potential by Support Vector Machine.

Ren, J., Chen, Y., Xin, L., Shi, J., Li, B. and Liu, Y., 2016. Detecting and positioning of traffic incidents via video-based analysis of traffic states in a road segment. IET Intelligent Transport Systems, 10(6), pp.428-437.

Samant, A., Adeli, H., 2000. Feature extraction for traffic incident detection using wavelet transform and linear discriminant analysis. Comput. -Aided Civil Infrastructure Eng. 15 (4), 241–250

Sun, L., Lin, Z., Li, W. and Xiang, Y., 2019. Freeway incident detection based on set theory and short-range communication. Transportation letters, 11(10), pp.558-569.

Texas Transportation Institute (TTI), 2020. Urban mobility report and appendices. http://mobility.tamu.edu/ums/report/. Accessed June 1, 2022.

Vapnik, V. (2013). The Nature of Statistical Learning Theory. In Springer science & business media. https://doi.org/10.1080/00401706.1996.10484565

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C. and Halkias, B.M., 2010. Freeway operations, spatiotemporal-incident characteristics, and secondary-crash occurrence. Transportation research record, 2178(1), pp.1-9.

Wali, B., Khattak, A. J., & Liu, J. (2021). Heterogeneity assessment in incident duration modelling: Implications for development of practical strategies for small- & large-scale incidents. https://doi.org/10.1080/15472450.2021.1944135

Wang, L., Abdel-Aty, M., Shi, Q., & Park, J. (2015). Real-time crash prediction for expressway weaving segments. Transportation Research Part C: Emerging Technologies, 61, 1–10. https://doi.org/10.1016/J.TRC.2015.10.008

Wang, Z., Qi, X. and Jiang, H., 2018. Estimating the spatiotemporal impact of traffic incidents: An integer programming approach consistent with the propagation of shockwaves. Transportation Research Part B: Methodological, 111, pp.356-369.

WisDOT TIME, 2017. WisDOT's Traffic Incident Management Enhancement (TIME) Program. Online at: https://wisdottribaltaskforce.org/wp-content/uploads/2015/06/TIME-Presentation.pdf.

Wu, Y., Abdel-Aty, M., & Lee, J. (2018a). Crash risk analysis during fog conditions using real-time traffic data. Accident Analysis and Prevention, 114, 4–11. https://doi.org/10.1016/j.aap.2017.05.004

Wu, Y., Abdel-Aty, M., Cai, Q., Lee, J., & Park, J. (2018b). Developing an algorithm to assess the rear-end collision risk under fog conditions using real-time data. Transportation Research Part C, 87, 11–25. https://doi.org/10.1016/j.trc.2017.12.012

Xiao, J. and Liu, Y., 2012. Traffic incident detection using multiple-kernel support vector machine. Transportation research record, 2324(1), pp.44-52.

Xiao, J., & Liu, Y. (2012). Traffic incident detection using multiple-kernel support vector machine. Transportation research record, 2324(1), 44-52.

Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. Accident Analysis and Prevention, 57, 30–39. https://doi.org/10.1016/j.aap.2013.03.035

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis and Prevention, 51, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Yu, W., Park, S., Kim, D. S., & Ko, S. S. (2015). An Arterial Incident Detection Procedure Utilizing Real-Time Vehicle Reidentification Travel Time Data. Http://Dx.Doi.Org/10.1080/15472450.2014.972762, 19(4), 370–384. https://doi.org/10.1080/15472450.2014.972762

Yuan, F., & Cheu, R. L. (2003). Incident detection using support vector machines. Transportation Research Part C: Emerging Technologies, 11(3-4), 309-328.

Zhang, H. and Khattak, A., 2011. Spatiotemporal patterns of primary and secondary incidents on urban freeways. Transportation research record, 2229(1), pp.19-27.

Zhang, Z., Liu, J., Li, X., & Khattak, A. J. (2021). Do Larger Sample Sizes Increase the Reliability of Traffic Incident Duration Models? A Case Study of East Tennessee Incidents: Https://Doi.Org/10.1177/0361198121992063, 036119812199206. https://doi.org/10.1177/0361198121992063

Zhang, Z., Nie, Q., Liu, J., Hainen, A., & Yang, C. (2022). Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data. Journal of intelligent transportation systems. Under publication.

Zhang, Z., Yang, D., Zhang, T., He, Q., & Lian, X. (2013). A study on the method for cleaning and repairing the probe vehicle data. IEEE Transactions on Intelligent Transportation Systems, 14(1), 419–427. https://doi.org/10.1109/TITS.2012.2217378

Zhang, Z., Yang, D., Zhang, T., He, Q., & Lian, X. (2013). A study on the method for cleaning and repairing the probe vehicle data. IEEE Transactions on Intelligent Transportation Systems, 14(1), 419–427. https://doi.org/10.1109/TITS.2012.2217378

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. Travel Behaviour and Society, 20(August 2019), 22–35. https://doi.org/10.1016/j.tbs.2020.02.003

Zheng, Z., Qi, X., Wang, Z. and Ran, B., 2021. Incorporating multiple congestion levels into spatiotemporal analysis for the impact of a traffic incident. Accident Analysis & Prevention, 159, p.106255.