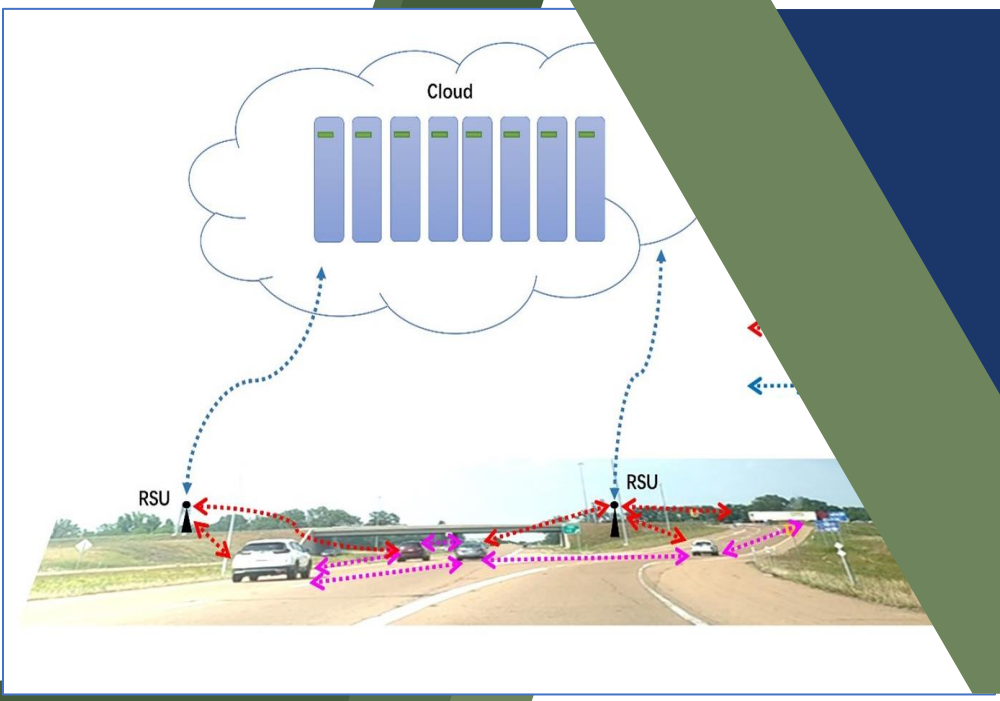

FINAL REPORT

PROJECT F4

AUGUST 2022

Automatic Safety Diagnosis in a Connected Vehicle Environment

Shuang Z. Tu, Ph.D., Jackson State University
Robert W. Whalin, Ph.D., Jackson State University
Di Wu, Doctoral Candidate, Jackson State University

STRIDE | Southeastern Transportation Research, Innovation, Development and Education Center

UF | Transportation Institute UNIVERSITY of FLORIDA

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. Project F4	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Automatic Safety Diagnosis in a Connected Vehicle Environment		**5. Report Date** 8/10/2022 **6. Performing Organization Code**
7. Author(s) Shuang Z. Tu, Ph.D., Jackson State University Robert W. Whalin, Ph.D., Jackson State University Di Wu, Ph.D. Candidate, Jackson State University		**8. Performing Organization Report No.** STRIDE Project F4
9. Performing Organization Name and Address University of Florida, Occupational Therapy, PO Box 100164, Gainesville, FL 32611 University of Alabama at Birmingham, Department of Civil, Construction, & Environmental Engineering, Hoehn Engineering Bldg., Rm 140, 1075 13th Street, Birmingham, AL 35294		**10. Work Unit No.** **11. Contract or Grant No.** Funding Agreement Number 69A3551747104
12. Sponsoring Agency Name and Address *University of Florida Transportation Institute/* Southeastern Transportation Research, Innovation, Development and Education Center (STRIDE) 365 Weil Hall, P.O. Box 116580 Gainesville, FL 32611 *U.S Department of Transportation/Office of Research, Development & Tech* 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** 3/1/2020 to 8/10/2022 **14. Sponsoring Agency Code**

15. Supplementary Notes N/A

16. Abstract - Previous researchers found that the most important accident causation factor was the driver's abnormal driving status, which was associated with driving volatility. And the driving volatility can be traced from the trajectories of the vehicles that were embedded in the BSMs. Based on these findings, we developed an automatic safety diagnosis system for the connected vehicle environment (ASDSCE), a real-time near crash warning tool with a multi-dimensional cloud-based driving anomaly detection (DAD) model and a conflict identification model (CIM) on the individual level specifically configured for BSMs. The architecture of the proposed system is composed of two components: one is in the cloud who collects and stores BSMs of the CVs and determines in batch mode the thresholds of each vehicle; the other is in the in-vehicle subsystem which determines the driving anomalies and detect conflicts. A near crash will be warranted when the traffic situation satisfies both of the following two conditions: (a) a conflict is identified and, (b) at least one of the drivers that is involved in the conflict is in abnormal driving status. The ASDSCE contains the following features: focusing on detecting abnormal drivers instead of normal drivers; using the trajectory data embedded in the BSM to study driving volatility; implementing on the individual drivers instead of the aggregate level; and reducing the model training time in order to leave sufficient time to the involved drivers to perform successful evasive actions. The presented computational pipeline of ASDSCE includes raw data collection, data preprocessing, data analysis, data communication and warning message generation. ASDSCE is built with Python on Visual Studio 2019 using the BSMs from the CV pilot studies and evaluated using the SHRP2 naturalistic driving study crash data.

17. Key Words driving status, abnormal detection, BSM, conflict, safety	**18. Distribution Statement** No restrictions

19. Security Classif. (of this report) N/A	**20. Security Classif. (of this page)** N/A	**21. No. of Pages** 62 pages	**22. Price** N/A

DISCLAIMER

ACKNOWLEDGEMENT OF SPONSORSHIP AND STAKEHOLDERS

This work was sponsored by a contract from the Southeastern Transportation Research, Innovation, Development and Education Center (STRIDE), a Regional University Transportation Center sponsored by a grant from the U.S. Department of Transportation's University Transportation Centers Program

Funding Agreement Number - 69A3551747104

LIST OF AUTHORS

Lead PI:

Shuang Z. Tu, Ph.D.
Jackson State University
shuang.z.tu@jsums.edu
https://orcid.org/0000-0002-4506-6447

Co-PI:

Robert W. Whalin, Ph.D.
Jackson State University
Robert.w.whalin@jsums.edu
https://orcid.org/0000-0002-8712-9434

Additional Researchers:

Di Wu, Ph.D. Candidate
Jackson State University
di.wu@students.jsums.edu
https://orcid.org/0000-0003-3169-3041

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Previous researchers found that the most important accident causation factor was the driver's abnormal driving status, which was associated with driving volatility. And the driving volatility can be traced from the trajectories of the vehicles that were embedded in the BSMs. Based on these findings, we developed an automatic safety diagnosis system for the connected vehicle environment (ASDSCE), a real-time near crash warning tool with a multi-dimensional cloud-based driving anomaly detection (DAD) model and a conflict identification model (CIM) on the individual level specifically configured for BSMs. The architecture of the proposed system is composed of two components: one is in the cloud who collects and stores BSMs of the CVs and determines in batch mode the thresholds of each vehicle; the other is in the in-vehicle subsystem which determines the driving anomalies and detect conflicts. A near crash will be warranted when the traffic situation satisfies both of the following two conditions: (a) a conflict is identified and, (b) at least one of the drivers that is involved in the conflict is in abnormal driving status.

The ASDSCE contains the following features: focusing on detecting abnormal drivers instead of normal drivers; using the trajectory data embedded in the BSM to study driving volatility; implementing on the individual drivers instead of the aggregate level; and reducing the model training time in order to leave sufficient time to the involved drivers to perform successful evasive actions. The presented computational pipeline of ASDSCE includes raw data collection, data preprocessing, data analysis, data communication and warning message generation. ASDSCE is built with Python on Visual Studio 2019 using the BSMs from the CV pilot studies and evaluated using the SHRP2 naturalistic driving study crash data.

Keywords:

driving status, abnormal detection, BSM, conflict, safety

EXECUTIVE SUMMARY

The purpose of this project is to construct a computational pipeline to identify near-crash events using basic safety messages (BSMs) in the connected vehicle (CV) environment and generate near-crash warnings to the driver.

We define near crash as a situation that satisfies both of the following two conditions: (a) a conflict is identified and, (b) at least one of the drivers involved in the conflict is in abnormal driving status. We built an automatic safety diagnosis system in the connected vehicle environment with Python on Visual Studio 2019 using the BSM data from the CV pilot studies and evaluated with the SHRP2 naturalistic driving study crash data. Our system is composed of a multi-dimensional driving anomaly detection model and a conflict identification model on the individual level using only the BSM data.

Our system can be used as a real-time near crash warning tool in the CV environment. The significance of our system lies in its special data source. Because the data source is solely the BSMs, our system can serve as an additional collision warning tool which may supplement the current popular advanced driver assistance systems that rely on the data collected by the sensors on the ego vehicle. With our system, traffic safety can be hoped to be significantly improved because the collision warning can be triggered from another vehicle other than the ego vehicle itself.

In addition, our system provides a way to reuse the BSMs. Due to the tremendous volume and complexity, it is not realistic to store all the BSMs generated in the CV environment into the data center. This research built a practical way to extract from BSMs the thresholds of the key performance indicators of each vehicle and only store these thresholds and the BSMs over a short period of time for traffic safety analyses. Therefore, the BSM storing problem can be mitigated.

This study combines the CV, traffic conflict technology and big data technology together. The designed system can be used as a real-time near-crash warning tool in the CV environment. It can help to improve the safety of connected vehicles in driving and increase the market penetration of connected and autonomous vehicles (CAVs).

Future work includes pilot studies to generate more data sets and further validate the current system and upgrading the model from sequential processing to parallel processing to reliably ensure real-time safety analysis and processing.

1.0 INTRODUCTION

Crashes are a major cause of traffic congestion and reducing crashes is a prominent task for congestion mitigation. As human factors contribute to more than 90% traffic crashes, abnormal driving behavior has been intensively studied to improve traffic safety. In the connected vehicle (CV) environment, Basic Safety Messages (BSMs) transmitted between CVs. A driver's behavior can be reflected by the vehicle's trajectories which are embedded in the BSMs. If a driver's abnormal driving behavior can be somehow detected, a potential crash can be avoided. Based on this reasoning, this project aims to utilize the BSMs to construct a near-crash warning system.

1.1 Objective

The objective of this project is to construct a computational pipeline to identify near-crash events using basic safety messages (BSMs) in the connected vehicle (CV) environment and generate near-crash warnings to the driver.

1.2 Scope

The computational pipeline is an automatic safety diagnosis system in the CV environment (ASDSCE). The ASDSCE consists of the traffic management center (TMC), all the CVs under its surveillance, and the datapath between them. The concept of the ASDSCE is illustrated in Figure 1.

The system in the cloud stores the historical BSMs of all the CVs under its surveillance, operates a continuous threshold calculation using the historical BSMs, and maintains a flag list of the current abnormal CVs. The historical BSMs are of a certain time period, say a month, calibrated according to the local conditions. The in-vehicle subsystem is equipped in all the CVs. It is composed of an on-board unit (OBU) and a computer. The two-way datapath is composed of the CV environment with the BSMs, including V2I and V2V, and the backhaul system. It transmits the information between the CVs, the roadside units (RDUs) and the TMC. The whole system configuration is illustrated in Figure 2.

The process of the in-vehicle subsystem is illustrated in Figure 3. Once the engine of a CV starts, the OBU starts to receive streams of information from the cloud and the nearby CVs. This information will be passed to the in-vehicle computer. The computer runs the abnormal driving status detection model using the ego BSMs and the thresholds received from the cloud to determine the status of the ego vehicle. If the ego vehicle is detected abnormal, the ego vehicle will be flagged, and the flag information will be uploaded to the cloud through the datapath. The joined efforts between the cloud and the CV complete the so-called task: Driving Anomaly Detection (DAD). If any abnormal driving status is found in the ego and other nearby CVs, the in-vehicle computer will run the Conflict Identification Model (CIM) of the corresponding conflict

scenario to identify conflicts. If any conflict is identified, a collision warning will be issued.

The datapath is not within the study scope of this project. The datapath involves the vehicle cloud which is an open research problem and is one of the major challenges of the CV.

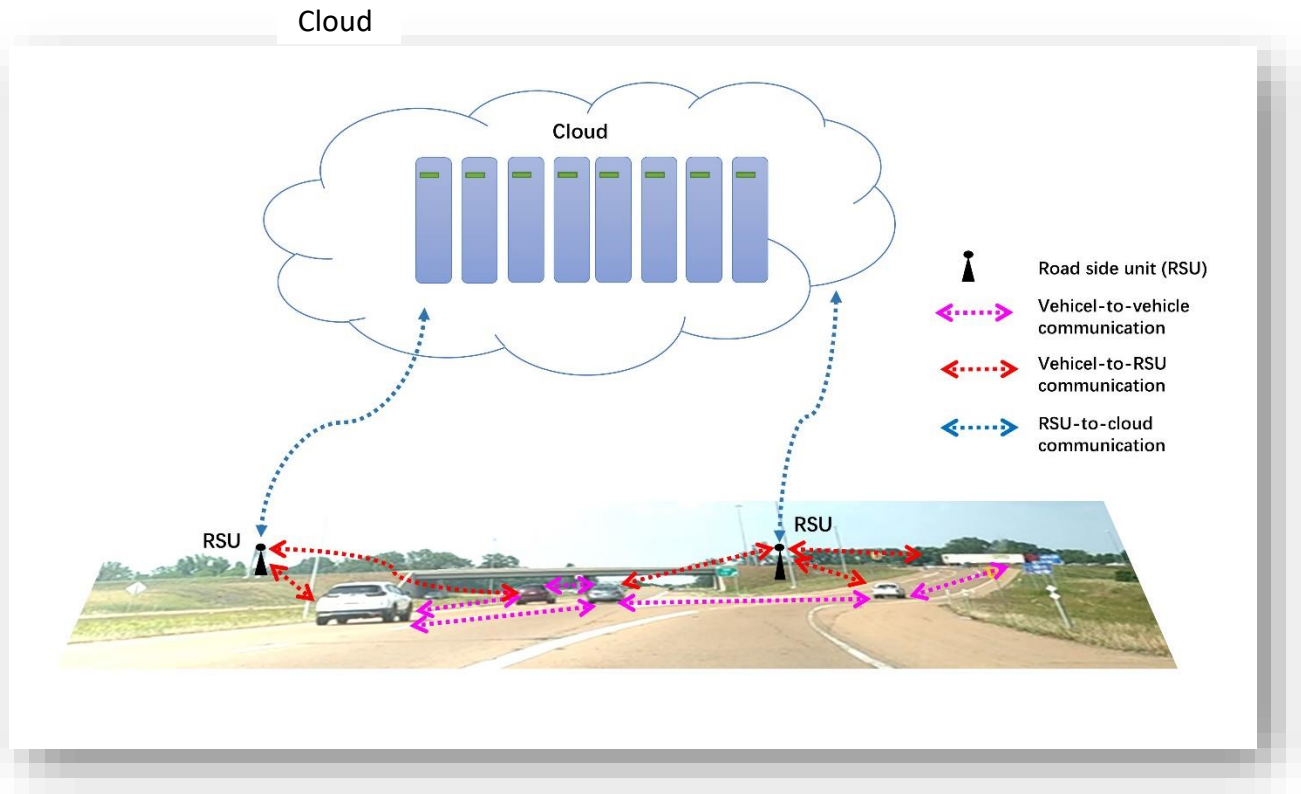

FIGURE 1. THE CONCEPT OF THE ASDSCE.

The proposed system uses solely raw BSMs in the CV environment, determines if the driver is in the abnormal driving status, and generate warnings when a conflict is identified. Here we define the near crash in a new way. A near crash needs to meet both of the following two conditions: first, at least one of the vehicles in a driver-vehicle unit (DVU) pair is in abnormal driving condition, and second a conflict is present. This project focuses on two tasks: task one is to perform Driving Anomaly Detection (DAD) with joint efforts from the cloud and the in-vehicle subsystem, in which the CVs with abnormal status are identified; task two is to perform Conflict Identification (CI) which is carried out the in-vehicle subsystem.

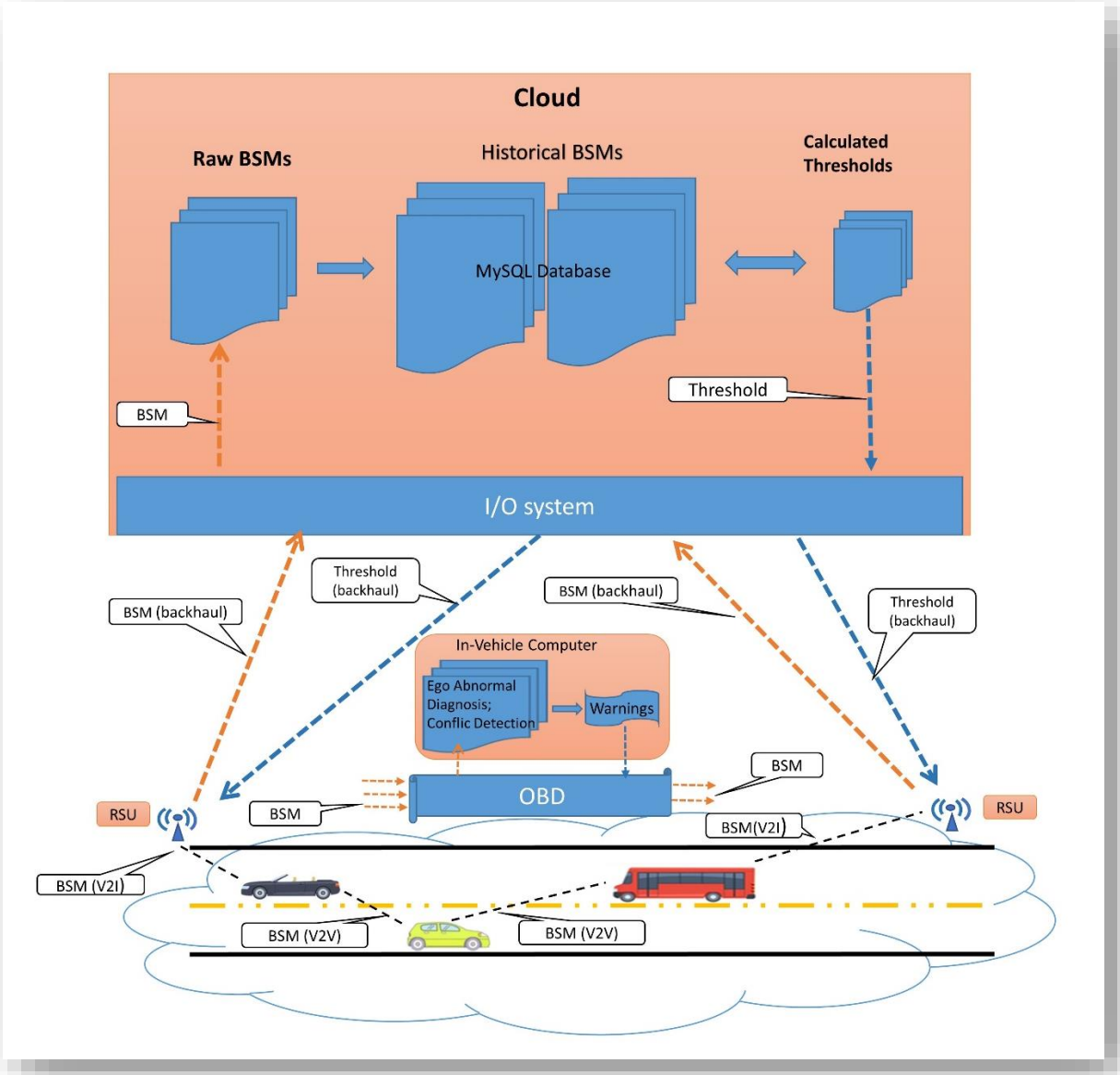

FIGURE 2. THE PROCESS OF ASDSCE.

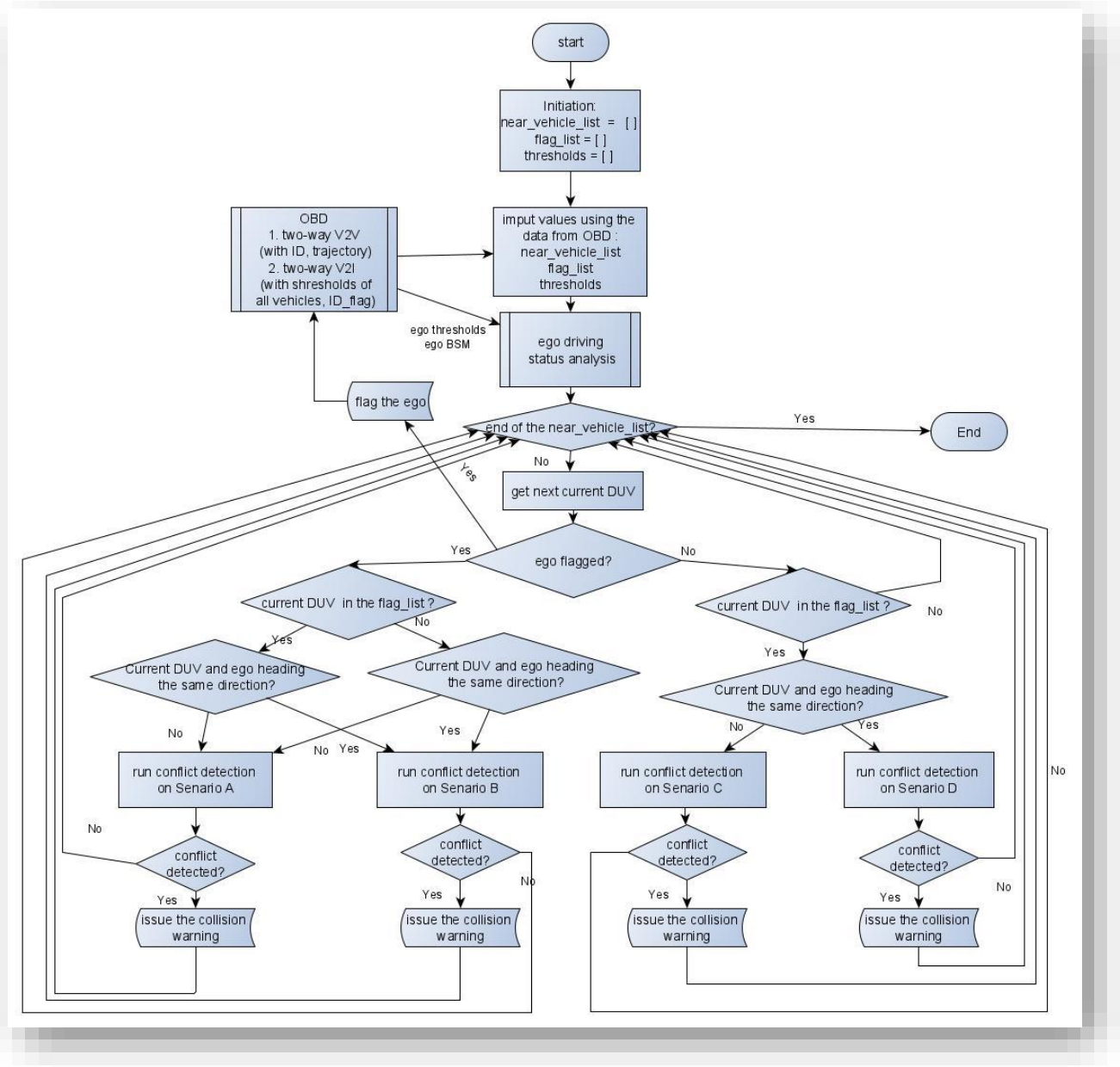

FIGURE 3. THE IN-VEHICLE SUBSYSTEM.

2.0 LITERATURE REVIEW

More than six million crashes resulting in more than 30 thousand fatalities and two million injuries have been reported annually on U.S. highways and streets (NHTSA, 2021; NCS, 2021). Due to enormous societal impact, highway safety has long been intensively studied. The mainstream traffic safety studies are crash record based, whose major tools are statistical models, and the direct measures are crash frequency and crash severity (Tarko, 2018). However, the further advancing of this approach is encumbered because of the following situations: (a) the necessity of waiting for crashes to happen, which is the most undesirable defect of this approach (Tarko, 2005); (b) crash data were not strictly accurate: crashes tend to be under reported and the rules of reporting vary, which may bias data sampling and mislead the statistic models (Wang, 2010; Han, 2009); (c) incomplete information: information in the circumstances preceding the recorded crash is seldom available; (d) problems in transfer: the related statistic models are site specific and local calibration is needed for transfer and some models might not be transferable (Wasconcelos, 2014). Nevertheless, the traffic conflict technique (TCT), a simulation-based approach, was proposed in late 1960s to measure the crash potential from the traffic kinematic characteristics instead of crash records. Having withstood considerate studies on its reliability and validity, TCT was gradually accepted by the safety community as a surrogate method of proactive safety analysis (Zheng, 2014; Chin,1997).

The key concept of TCT is conflict, which was first proposed by Perkins and Harris as “any potential accident situation” including evasive actions of drivers and traffic violations (Per-kins1968traffic). The definition of conflict experienced many years of discussion and settled down as “an observable situation in which two or more road users approach each other in space and time to such an extent that there is a risk of collision if their movements remained unchanged” (Amundsen, 1977). The TCT overcomes the aforementioned disadvantages of the crash data-based safety study approaches and was utilized in numerous traffic studies. The earliest attempts can be traced back to 1976 in Cooper's simulation study of a T-junction (Cooper, 1976), followed by a good number of studies investigating the traffic conflict profiles, involving the total number of conflicts and the number of vehicles that encountered conflicts during simulation, on various road configurations (Archer,2005; Huguenin, 2005; Saccomanno, 2008). Meanwhile a good number of safety performance indicators of TCT, referred to as surrogate safety measures (SSMs), were developed. The SSMs can be categorized into temporal based, such as TTC and post-encroachment time (PET); distance based, such as proportion of stopping distance (PSD); deceleration based, such as deceleration rate to avoid a crash (DRAC); and other indicators, such as crash index (CI) and margin to collision (MTC) etc. Review of the SSMs can be found in the literature (Zheng,2014; Mahmud,2017; Chin, 1997).

However, there are also issues in TCT. For example, the definition of conflict was complained to be too simple and unrealistic to describe traffic behaviors (Saunier, 2006; Hidas, 2005 modelling). It was also complained about ambiguity because the cutoff boundary between a conflict and a non-conflict situation is indistinct (Mahmud, 2017). Besides, many surrogate measures were developed but no consensus has been reached on what is the most preferable measure (tageld-in2017comparison). Moreover, the number of conflicts to an equivalent collision was found in-consistent and contextual with very high variation (Fazio, 1993; Hidas, 2005). These issues caused difficulty in implementing SSMs as indirect measures for the safety study in practice.

Nevertheless, SSMs found their uses as a post-processor for safety evaluation in microscopic simulation models. In 2008 the Federal Highway Administration (FHWA) released the surrogate safety assessment model (SSAM) using the simulation data, including the trajectories of the simulated vehicles. SSAM uses combined TTC and PET to detect conflicts according to specific configuration of traffic. For example, in a car-following scenario, a conflict would not be war-ranted when a TTC reached the threshold while the PET did not due to evasive maneuver; but in a head-on scenario, TTC reaching the threshold alone can warrant a conflict. The process of conflict determination of SSAM indicated that conflict was not a stationary term to be defined, but a user defined situation. In SSAM, the threshold of TTC was user defined but recommended to be 1.5 seconds for the values above it was not generally considered “severe” in a traditional field conflict study (Sayed, 1994; Gettman, 2008; Das, 2020). PET was defined as the time differential between the time the leading vehicle occupied a location and when the trailing vehicle arrived. The threshold value of PET was also user defined and needed calibration. PET threshold determination for heterogeneous traffic scenario is an open research problem (paul2020post). SSAM can be used as a safety evaluation add-on module of the microscopic simulation models such as VISSIM (Fellendorf, 2010) Paramics (Cameron, 1996) and CORSIM (Halati, 1997). Many studies used SSAM to identify traffic conflicts, as SSAM is based on the SSMs of TTC and PET. TTC and PET became the most used SSMs (Alrajie, 2015).

Although SSAM is powerful in determining many safety features, using SSAM directly in the real world of the CV environment might cause some problems. For example, SSAM has its own module for driving abnormal detection by calculating the probability of collision using the trajectories, which needs at least five seconds to collect the data on the scene and train the model be-fore analyzing the driving status. Whereas the CV environment cannot afford the five seconds to perform this task of training the model on the scene. As in the CV environment, the effective range of BSMs is as short as 300 meters, in the case that two CVs are 300 meters apart and both are running at the speed of 50mph (22.352 m/s) in the opposite direction, TTC is 6.7 seconds. If counting from the moment of receiving the first BSM from the vehicle of 300 meters away, after five seconds consumed by the model training process, there would be only 1.7 seconds

left, but the driver needs 2.5 seconds to perceive the danger and takes actions. Similar problems also exist in the car-following scenario when the leading vehicle brakes sharply, which happens when the driver is under abnormal driving status. Therefore, there is a need to reduce the model training time in order to leave sufficient time to the involved vehicles to perform successful evasive actions.

Other than in stimulation models, TCT was also widely utilized in automobile industry on Advanced Driver Assistance Systems (ADAS), such as adaptive cruise control (ACC). While ACC is expected to reduce rear-end collisions caused by driver's error, it cannot completely replace driver's braking. Even equipped with ACC, a vehicle still needs the advanced real-time safety warnings (Bose, 2003). Therefore, collision warning/avoidance, such as lane departure warning (LDW), forward collision warning (FCW), pedestrian detection (PD), and automatic emergency braking (AEB) are installed in high-end cars and AVs and will be installed in low-end vehicles as well in the near future (Hu, 2020; Wang, 2011). In addition, emergency steering assistance (ESA) is receiving increasing research attention for it can help evade from a collision that is unavoidable by AEB alone (Eckert, 2011; He, 2019). For ADAS, TTC was the most widely used SSM because of its simplicity and applicability (Van, 1993; Farah, 2009; Qu, 2014; Qu, 2014; Meng, 2012; Jin, 2011; Li, 2017). In the area of ADAS, the concept of conflict is not popular because conflict is a concept that is too relax. A typical equivalent conflict - to-crash rate is at million level (Fazio, 1993). If the collision warnings were based on conflicts, there would be too many false alarms. Therefore, a concept of near crash was used in ADAS to describe a situation of potential crash when a warning is needed.

Near crash, also called near miss, was first formalized by McFarland and Moseley as the "emergency situation or critical incidents which could easily have led to a crash" (Williams, 1981). Hanowski et al., defined near crash as any circumstance that requires a rapid, evasive maneuver by the subject vehicle, or any other vehicle, pedestrian, cyclist, or animal to avoid a crash. And a rapid, evasive maneuver is defined as a steering, braking, accelerating, or any combination of control inputs that approach the limits of the vehicle capabilities (Hanowski, 2006). To be more descriptive, the measure of scale for criticality assessment (SCA) was developed to clarify how close a near crash is to a collision. Based on TTCs, SCA classifies the traffic situations to groups of imperceptible, harmless, unpleasant, dangerous, and uncontrollable (Sieber, 2016).

The data for ADAS are mostly collected from the in-vehicle sensors such as radar, Linda, camera, speed sensor, and throttle position sensor etc. During the process, data fusion and image processing techniques were utilized to extract the trajectories of the nearby vehicles from the collected images. Currently data collection of the of nearby vehicles relies on the in-vehicle sensors. This posts a safety issue in the cases when the sensors all break down and therefore additional information from other channels are necessary. And there are other channels ready in deployment. As known by all, autonomous

vehicle (AV) is expected to be the ultimate solution to future transportation (Wang, 2020). And "The full benefits of vehicle automation can be achieved only through connectivity" (USDOT, 2020). The ITS Joint Program Office of United States Department of Transportation (US DOT) is already moving forward with research on joining connected vehicle (CV) to AV.

BSM is a class of SPMD data, which is a part of the connected vehicle (CV) program. BSM is the basic application known as the "Here I Am" data message. The format of BSM is defined by Society of Automotive Engineers J2735: The Dedicated Short-Range Communications (DSRC) Message Set Dictionary. BSMs are broadcast from the in-vehicle device at the dedicated bound of 5.9 GHz spectrum at the query of 10 Hz to surrounding (maximum 300 miters) vehicles (Henclewood, 2014). A BSM is composed of two parts: part one is the main part of the message, which includes information such as the vehicle ID, epoch time, GPS location, speed, acceleration, yaw rate, and associated accuracy measurements; part two provide supplementary information. BSMs are overall regarded as snapshot for safety data and not be reused or stored. Unique research was found to try to reuse BSMs for an In-vehicle computing system (Benaissa, 2020).

As a major initiative of US DOT, the CV technology enables safe, interoperable networked wireless communications among vehicles, the infrastructure, and passengers' personal communications devices. The BSM data generated in the CV operation are massive in amount and become an innovative data source for traffic safety community and thus opened the door for the research topics on traffic safety that should have been carried out yet had not been done because of lack of data, in which a fundamental one is the driving behavior regime analysis. While the availability of big data provides opportunities for data driven modeling and analysis on safety data, it also calls for the effective data collecting, storage and quarrying. Once the CV project is brought about, up to 200GB/second data will be typically generated for a traffic management center (TMC). Due to the massive volume, it is not realistic to store all the BSMs, but only the most relevant information that is extracted from the raw data will be stored. This research is proposed to explore what information of BSMs need to be kept, how to extract it and how to process it for the real-time safety diagnosing.

Another issue of ADAS is related to the focus of the warning criteria. From the systematical viewpoint the driving system is composed of drivers, vehicles, roads and environment. A crash is the result of a serial of malfunctions of the driving components as shown in Figure 4. Research shows that human factors contribute more than 90% crashes, as shown in Figure 5 (Treat, 1979; Singh, 2015; Dingus, 2016). As most crashes are due to the drivers who are in abnormal driving status, the safety study should focus on the abnormal driving status instead of the normal drivers. But the reality is, in developing the criteria for ADAS conflict detection, although many safety features were addressed, such as the types of vehicles involved, friction, and lighting conditions, no

driving status of the driver was considered. The driver responses differently when he/she is under abnormal driving status. For example, the perception and reaction time of drivers under influence (DUI) is longer than those who are under normal status, and the action of a DUI is unpredictable. If focus is put on the normal or average drivers, the warning criteria would not be able to represent the abnormal drivers. Therefore, abnormal driving status needs to be a focus of ADAS and additional conflict or near crash detection criteria needs to be established especially for abnormal drivers. And real-time driving anomaly detection (DAD) in the CV environment to issue real-time near crash waring is imperative and an active research problem.

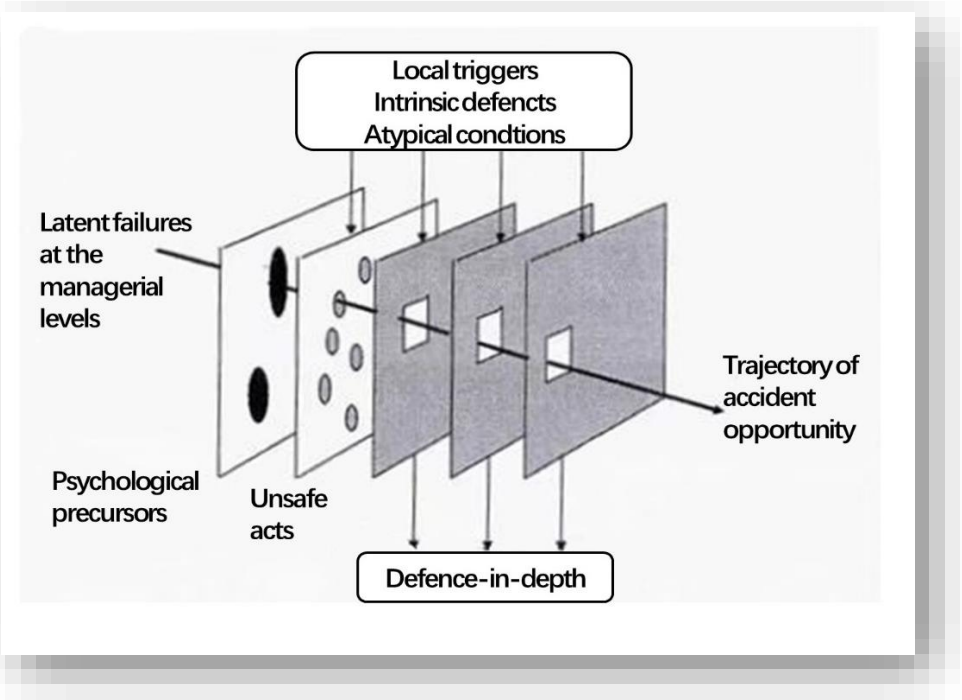

FIGURE 4.THE DYNAMICS OF CRASH CAUSATION (REASON1990).

To define driving anomaly (DA) is an open research problem. Currently, there are three approaches: first, from the common sense, abnormal driving behaviors include driving under the influence (DUI), driving with distraction, aggressiveness, and drowsiness because these behaviors will likely cause crashes. So traditionally DA was defined as a situation in which the driver is not concentrating on driving (Miyaji, 2008); Second, from statistics of the majority drivers: as crashes are rare events, complying with the majority in driving maneuver is considered safe and normal. Hence DA can be defined as

deviating from the statistical majority; Third, from the statistics of the individual driving behavior: as driving behaviors differ substantially between individuals and everyone has one's own driving patterns such as the way hitting the gas and brake pedals, wheel steering, and in the distance they keep when following a vehicle (Fancher, 1998; Igarashi, 2004), a driver might drive years to have a crash, so not complying with one's own driving pattern can also be considered as DA. In the occasions of DA, which indicates that the driver is not in the best mode and cannot judge if the driving status is normal, a safety alarm will be helpful to avoid a penitential crash.

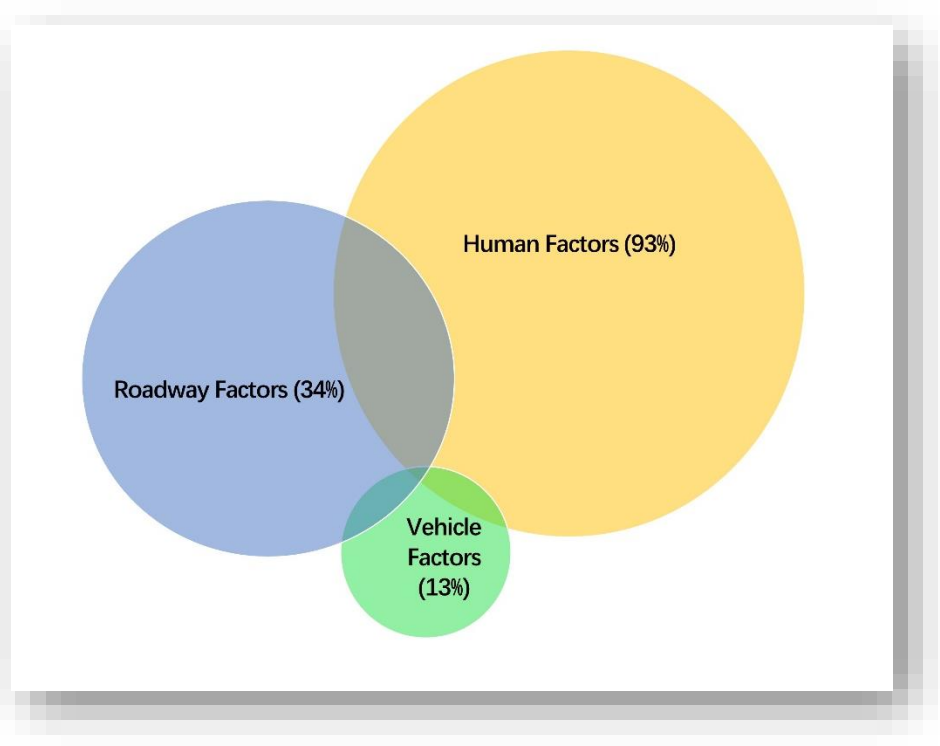

FIGURE 5. THE PRESENCE OF ACCIDENT CAUSATION FACTORS (TREAT1979).

According to the definitions of DAD, the direct approach is to monitor the driver's exhalation, facial and body movements using in-vehicle alcohol sensors and/or cameras and analyze the images using computer vision for DAD. The drawbacks of this approach are the cost of computation in deep learning and high-end cameras, the privacy issue, and its limitation from being freely broadcast at real-time (Janai, 2017).

An indirect approach is to use social economic data to categorize risky drivers or use trajectory data which are the results from the driving maneuver. Social-economic factors are assumed to have impacts on driving behavior in a psychological way (Boyle,

2007). Age, gender and income level etc. were widely used because they are found statistically correlated with the number of crashes and these measurements are easily available. This method was widely utilized by automobile manufacturers and insurance companies to identify risky drivers since 1968 (Ayuso, 2019). For example, pay-as-you-drive insurance systems calculate premiums according to how risky the insured driver is. The riskiness factors used include driven miles, time of day, speed, and how accident-prone the often-visited places are. "Aggressive driving" is a term used by the National Highway Traffic Safety Administration (NHTSA) to classify "driving actions that markedly exceed the norms of safe driving behavior and that directly affect other road users by placing them in unnecessary danger" (Richard, 2018). However, a theoretical definition for aggressive driving "has been proven challenging to arrive at a consensus" (Richard, 2018). In case of the non-administrative research, "driving volatility" was proposed to replace "driving aggressive" as a measure of the instantaneous driving decisions (Wang, 2015). The transition of the nomenclature opened the horizon of approaches of using only the data of vehicle trajectory to describe the driver's behavior.

In traffic safety, vehicular trajectory data were studied as the footprints of driving behaviors to identify DA and give warnings when abnormal events warrant a warning flag. The vehicular trajectory data contain detailed information on microscopic phenomena. Embedded in the trajectories, the speed, acceleration (Lajunen, 1997), jerk (Ericsson, 2000) were mostly selected as key performance indicators (KPIs) to measure driving volatility. Speeding is an aggressive behavior and very common among drivers but directly using speed as a KPI for DAD is naïve because speed is contextual to speed limits (Ellison, 2010). A simple solution is to use higher maximum speeds, which are associated with drivers who have more accident records (Lajunen, 1997). Another solution is to use acceleration which is also found associated with risky drivers. The change of acceleration with respect to the speed (Langari, 2005) and with respect to the time (Murphey et al., 2009) (also called the vehicle jerk) were also used to classify drivers driving behavior. The cut-off values for abnormal acceleration were studied, for example, 1.47 m/s^2 as the threshold for aggressive acceleration and 2.28 m/s^2 for extremely aggressive (Kim, 2013), and another study set the range of 0.85 to 1.10 m/s^2 as aggressive acceleration (De,2000). So far, no consensus threshold has been reached because it is contextual sensitive (Wang, 2015). Meanwhile, accelerations were found to vary with speeds and accelerations on different directions cannot change together, the thresholds for longitudinal and lateral accelerations of various speed bins were set to be the multivariate KPIs (Liu, 2014; Liu, 2016). This rule-based method has the advantages of simplicity and efficiency (Martinez, 2017) while its disadvantage is it cannot address the different driving patterns of individuals.

The next improvement would be analyzing BSMs at the individual-level, which is the approach we employ in this project. This research is aimed to construct such a system

using the trajectory data embedded in BSMs to identity near crash to generate potential collision warning.

3.0 PROBLEM STATEMENT

In retrospection on the literature, traffic safety is facing a leapfrogging development. As the transportation system is evaluating toward ITS and CAV, the traditional safety statistical models can no longer bear the requirement to be the mainstream method. Until the full automation is achieved, as abnormal human behavior is a major causation factor of crashes, abnormal driving behavior will still be a focus of traffic safety. Each driver has the won driving pattern; today's computation capability allows modeling the driving behavior at individual level. From the technology of ADAS, synthetizing TCT to the CV environment to form digital twins is a promising approach.

However, there are salient shortcomings of adopting TCT and its key concept-- conflict: firstly, conflict is a loose measure, and the warnings can be triggered too often and result in too many false alarms; secondly, as the current ADASs use the data collected on the scene, the identification of abnormal human behavior is not prompt enough.

From the geniture of the crash, psychological precursors of abnormal status already exist before the crash scene. It is possible to identify the abnormal status of the driver before the scene. Research have shown that the driving anomaly can be traced from the vehicle trajectories, which are embedded in the BSMs of the CV environment.

With the growth of the market penetration rate, there will be massive BSMs, and it would be unpractical to store all of them in data centers. It is imperative to find a way to extract and store the valuable and storable information from the BSMs before they perish.

Based on the background, we propose an automatic safety diagnosis system in the CV environment (ASDSCE). The ASDSCE contains the following features:

a) Focus on detecting abnormal drivers instead of normal drivers,
b) Use the trajectory data embedded in BSM to study driving volatility,
c) On the individual driver level instead of the aggregate level, and
d) Reduce the model training time in order to leave sufficient time to the involved drivers to perform successful evasive actions.

4.0 DATA DESCRIPTION

4.1 BSM Data

The basic safety message (BSM) data were the working data of our project. BSM is a basic application of CV program known as the "Here I Am" data message. BSMs are generated in the on-board-devices (OBDs) that were specifically designed for CVs. In the air, the BSMs are broadcasted at the dedicated bound of 5.9 GHz spectrum at the frequency of 10 Hz (Henclewood, 2014) and can be received by the nearby CVs and roadside units (RSU). The effective transmitting distances of BSMs are ranged from 300 meters to 1000 meters. The format of a BSM is defined by the Society of Automotive Engineers J2735: The Dedicated Short-Range Communications (DSRC) Message Set Dictionary. A typical BSM is composed of two parts: part one is the main part of the message, including the vehicle ID, epoch time, GPS location, speed, acceleration, yaw rate, and associated accuracy measurements; part two provides supplementary information. BSMs were considered disposable and not reused.

The Safety Pilot Model Deployment (SPMD) project is a part of the CV program. It was a research initiative on CVs and collected and stored the BSM data during the tests. The SPMD data are available on the Intelligent Transportation System (ITS) DataHub (its.dot.gov/data/). The working data used in this project are the field BSM data from a SPMD test conducted in Ann Arbor, Michigan, in October 2012. A Comma Separated Values (CSV) BsmP1 file of a size of 67GB stores all the BSMs generated by the 1527 test vehicles in the test. The original downloaded data file had 19 attributes and over 500 million records. During our data pre-processing, the irrelevant attributes were filtered out and the resulted data file has 11 attributes including $DevID$ for the vehicle ID, $EpochT$ for timestamp and attributes for latitude, longitude, accelerations, heading and yaw-rate. The descriptions of the attributes are shown in Table 1.

TABLE 1 ATTRIBUTE LIST OF THE BSM DATA

Attributes Name	Type	Units	Description
DevID	Integer	None	Test vehicle ID assigned by the CV program
EpochT	Integer	seconds	Epoch time, the number of seconds since the January 1 of 1970 Greenwich Mean Time (GMT)
Latitude	Float	Degrees	Current latitude of the test vehicle
Longitude	Float	Degrees	Current longitude of the test vehicle
Elevation	Float	Meters	Current elevation of test vehicle according to GPS
Speed	Real	m/sec	Test vehicle speed
Heading	Real	Degrees	Test vehicle heading/direction
Ax	Real	m/sec^2	Longitudinal acceleration
Ay	Real	m/sec^2	Lateral acceleration
Az	Real	m/sec^2	Vertical acceleration
Yawrate	Real	Deg/sec	Vehicle yaw rate

4.2 SHRP2 Data

The crash data from Naturalistic Driving Study (NDS) for the second Strategic Highway Research Program (SHRP 2) were our model evaluation data. NDS is a research program to address the impact of driver performance and behavior in traffic safety. The Virginia Tech Transportation Institute (VTTI) serves as the technical coordination and study design contractor for the NDS and maintains the InSight Data Access Website (Jafari2017).

In the InSight Data Access Website, the Event Detail Table section there lists 41,530 records of crashes and near crashes. Each record is posted with detailed information of the event including a video of up to 25 seconds before the event, event detail data and the final narrative. There are readily fetched data sets that had been used by previous studies and can be obtained by other institutes with no cost. However, although our required data can be retrieved from the crashes there was no used data set could meet our requirement.

For acquiring NDS data, a data use license with VTTI and the proof of Institutional Review Board (IRB) approval are required. The data users also need to take the VTTI training in the protection of human subjects. We contacted VTTI with a data description and initiated the data purchase process. The data description is shown as the following:

- Participant driver is at fault;
- Police reportable or most severe crash severity;
- Event nature of conflict with another vehicle (e.g. exclude run off road or tire strikes);
- Incident type excludes backing or rear-end struck scenarios;
- Crash occurred at least 60 days after participant's entry into study;
- Lowest volume and traffic movement roadways excluded (i.e., functional class 5).

We screened all the crashes of SHRP2 and hand-picked 47 crash events which meet the data description, and the driver was under abnormal driving status, such as driving under influence, driver fatigue, or driving while texting on the phone. We settled down on purchasing the data set that includes 47 crashes and 60-day time series data before the crash day. According to data requirement, from the SHRP 2 data collection system, a total of 12500 trips from 46 events were retrieved (one of the crash events did not have trip time information). As crashes are rare invent and the instrumented vehicles are limited in number, they do not have crashes in which both of the vehicles involved in a crash were instrumented. The crashes happened between an instrumented vehicle and a stationary object, such as a tree, fence or roadway curbs.

In order to protect potentially identifying information (PII), however, there were some restrictions. For example, the GPS coordinates for crash trips, the exact time, and any information can trace the driver's identification cannot be released. The attributes of the time series data include vtti_timestamp, vtti.file_id, vtti.accel_x, vtti.accel_y, vtti.heading_gps, vtti.speed_gps vtti.speed_network,, x_position, y_position, as shown in Table 2.

TABLE 2. ATTRIBUTE LIST OF THE SHRP2 DATA

Attributes Name	Type	Units	Description
vtti_timestamp	Integer	millisecond	The time steps from the released point of the trip.
vtti.file_id	Integer	/	The identification number of the trip file.
vtti.accel_x	float	g	Vehicle acceleration in the longitudinal direction versus time.
vtti.accel_y	float	g	Vehicle acceleration in the lateral direction versus time.
vtti.heading_gps	float	g	Compass heading of vehicle from GPS.
vtti.speed_gps	float	km/h	Vehicle speed from GPS.
vtti.speed_network	float	km/h	Vehicle speed indicated on speedometer collected from network.
x_position	float	meter	The relative X coordinate to an point (fixed in one trip).
y_position	float	meter	The relative Y coordinate to an point (fixed in one trip).

In order to use the SHRP2 data in the DAD model, the number of days prior to crash need to be identified. Although the query to fetch the data used the drivers that included to the program 60 days before the crash, the maximum days

5.0 TASK 1: DRIVING ANOMALY DETECTION (DAD) MODEL

5.1 Introduction of the DAD Model

We propose a multi-dimensional driving anomaly detection (DAD) system on the individual level specifically configured for BSMs. This DAD model is a crucial component of our automatic safety diagnosing system in the CV environment (ASDSCE).

Anomaly detection is an interdisciplinary problem, and it has been applied in many domains such as finance for credit card fraud detection, healthcare for magnetic resonance imaging (MRI) diagnosis on malignant tumors (Wilson, 1934; Sundt, 1974), astronomy for damage detection on space craft, and cybersecurity for intrusion detection (Chandola, 2009), signal intrusion in the CV environment (Rajbahadur, 2018), but no application was found in DAD using the BSM data.

Traditionally most highway safety studies have relied on historical crash data and statistical models. Yet crash data possess the notorious deficiency in availability and quality because crashes are rare events. As an alternative, the traffic conflict technique (TCT) that measures the crash potential -- conflict, which is defined as "an observable situation in which two or more road users approach each other in space and time to such an extent that there is a risk of collision if their movements remained unchanged" -- without having to wait for crashes to happen emerged in late 1960s. The non-crash

data approach based TCT was widely used because it makes safety analysis much less expensive, can be well connected to traffic simulation models and has good performance in countermeasure analysis. However, the disputes on TCT's qualification for surrogate measures for crash data have never been resolved. Especially when the modern transportation development calls for advanced safety diagnosis, TCT appears not readily tuned for the new challenge. This research is proposed to integrate TCT measures into the near-crash identification process, and to demonstrate that TCT is a useful tool for safety diagnosis.

Although the idea of autonomous vehicles (AVs) has been around for more than a century, and it has reached a point where it can begin to be offered to the public, AV is still not largely accepted. Safety is the major issue especially in obstacle detection because the information merely from the ego vehicle cannot guarantee 100 percent safety. However, this problem can be readily solved in the environment of CV through exchanging BSMs. The driving automation cannot be achieved without CV. While the private sector is moving quickly in the AV space, the USDOT will play a significant role in the deployment of AVs. By integrating CV with AV, we can improve the safety of our roads, expand our transportation capabilities, and greatly extend mobility options to everyone.

Since abnormal driving behaviors are present in more than 90% of crashes, a conflict together with at least one driver in abnormal driving status can be a closer stage to a crash. The hypothesis of this research is: each driver has his/her own driving patterns of normal and aggressive/abnormal status, which can be identified by the patterns of the driver's BSMs, such as abnormal acceleration rates. The patterns found in the historical BSMs will be recorded and updated by the traffic management centers. The real-time safety diagnosis is running non-stop comparing the incoming BSMs and the stored patterns of BSMs. Once the real-time/new pattern of a driver is categorized to be aggressive, and a conflict is also identified, a near-crash event will be identified. While real-time safety diagnosing is still an open area for research, this research is proposed to superimpose the missing modules and integrating BSM and TCT as part of the pipeline for automate safety diagnosing.

Data science is a mash-up of different disciplines which can help decision makers shift from ad hoc analysis to an ongoing conversation with data, and its intuition-based essence helps finding out the hidden patterns of the data. The outstanding achievement of data companies such as Google, Amazon, Facebook, Twitter, and LinkedIn manifestos the magic power of data science and makes other companies even academia interested in using data science for breakthroughs in their problems that traditional methods won't offer. However, in the merging of data science into an established discipline, especially the stringent transportation engineering field, rejections and obstacles arise, which is not uncommon in the evolution history of the

human knowledge body. So, for the data scientists, it is important to be precautious of the natural tendency of overlooking the fundamentals of the field; meanwhile for the established discipline, it would be wise to keep an open mind while scrutinizing the data-driven applications coming into its field. Benchmarked with the CV and AV, modern intelligent transportation system (ITS) transportation is moving toward fully automation. Instrumented with digital devices producing big data, ITS brings both challenges and opportunities to technological development and application innovation. For the traffic safety community, the challenge lies in keeping up with the modern driving environment and the tools from fast developing areas such as artificial intelligence, and the opportunity in reshuffling the safety analysis methods and guidelines.

The proposed DAD system is a component of our computational pipeline to identify near-crash events. The designed functionality of the DAD is to take recent historical big BSM data to learn the thresholds to differentiate the normal and abnormal driving status, and with the thresholds to identify the anomalies using the real-time BSM data.

5.2 Methodology of the DAD Model

The methodology of our DAD is determined by the nature of the working data, the nature of the anomaly, the availability of the labels and the constraints and requirements of the traffic safety domain.

The proposed DAD system is divided to two parts: in the cloud and in the in-vehicle subsystem. In the TMC, the system collects and stores BSMs of the vehicles it covers for a certain period of time, say a month, and determines in batch mode the thresholds of the selected key performance indicators (KPIs) representing the normal status for each vehicle, and broadcast the thresholds through BSMs; in the in-vehicle device, as new BSMs streaming in, the device compares the new values of each KPI with the received thresholds and determines if it is an outlier. The outliers will be analyzed to determine if the outliers combined deserve an anomaly event against that vehicle. Finally, the system will determine the impact factors to update thresholds according to the significances of the outliers. The whole process has five modules as follows (also illustrated in Figure 6): Module 1. Data Preprocessing and Selecting KPIs; Module 2: Learning What Is Normal; Module 3: Detecting Outliers; Module 4: Determine Abnormal Driving Event; Module 5: System Updating.

5.2.1 Module 1: Selecting Key Performance Indicators (KPIs)

The nature of the input data is key to DAD because it determines the techniques. By structure BSMs are discontinuous time series (TS) data, which is a type of sequence data where data instances are linearly ordered but have lots of not available (NA) records. BSMs are also spatial data because coordinates are included. TS data typically consist of two components: contextual attributes, which are used to determine the context for that instance, such as timestamps

and coordinates; and behavior attributes, such as speed and accelerations. Here if we treat timestamps as the contextual attribute and coordinates as behavior attribute, then BSM has high (almost infinitive) cardinality; If alternatively, we treat coordinates as the contextual attribute and time as the behavior attributes, then BSM will also has high cardinality; If we treat both time and coordinates as contextual attributes, then the number of contexts will be infinite. To make the problem solvable, we propose to treat time and space in different stages – as coordinates shows the environments, such as road conditions and traffic congestion etc. Hence in this project we focus on determining the driving status and ignore spatial coordinates. The environment impact will be handled by the Surrogate Safety Assessment Model (SSAM) to analyze the conflict.

Having excluded the spatial component, the KPIs were determined based on the goals of making full use of the rest of BSM attributes. As driving behavior is complicated, we included multiple varieties since each KPI might have its own pattern. All we can have now are speed, acceleration, jerk, and yaw-rate. Conventionally the first step of TS analysis is to decompose TS data according to the context, in our case the context is the time. We did visualization of all the selected KPIs with respect to time, and no identifiable periodicity was observed. We also performed autocorrelation process, with no seasonality identified either. We also visualized the KPIs with respect to speed, and found they change with the speed, which is consistent with the literature findings. Figure 7 shows the relationship between acceleration and speed, which is in line with the previous studies that the accelerations have some special patterns with respect to speeds (Liu, 2014; Liu, 2016).

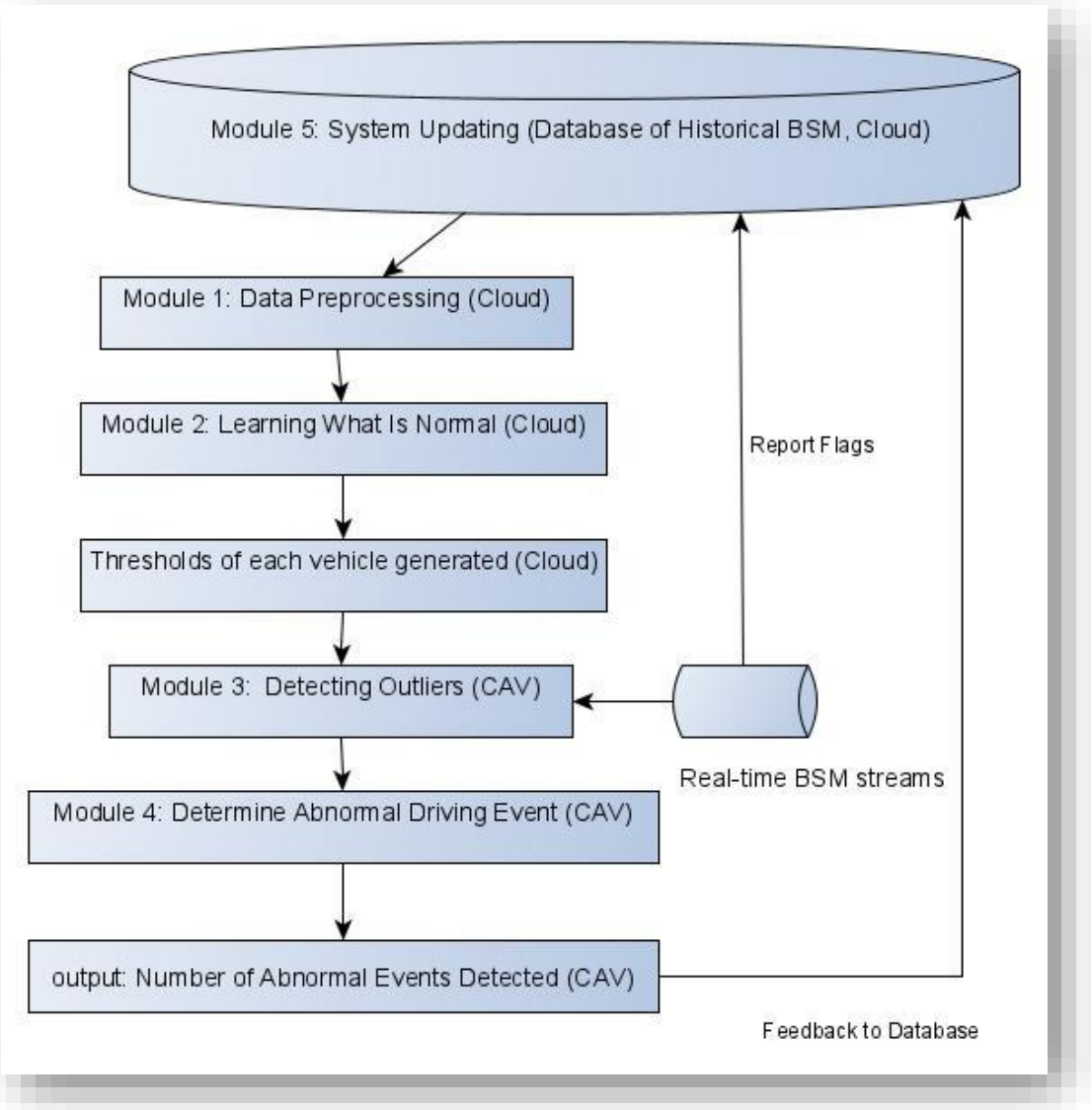

FIGURE 6. PROCESS OF THE PROPOSED DAD.

As the KPIs of A$celeration - longitudinal, acceleration - lateral, jerk - longitudinal, jerk - lateral$ are found to co-exist with abnormal driving status (Lajunen, 1997; Ericsson, 2000; Langari, 2005; Murphey, 2009), we used speed as a context variable instead of time. As the yaw rate describes the rate of change of the heading angle and is directly related to the lateral acceleration, we did not include the heading and raw rate as a KPI. Since the cut-off values of the thresholds for normal vs. abnormal are contextually sensitive, and no consensus thresholds have been reached (Wang, 2015), we set up the thresholds as variables.

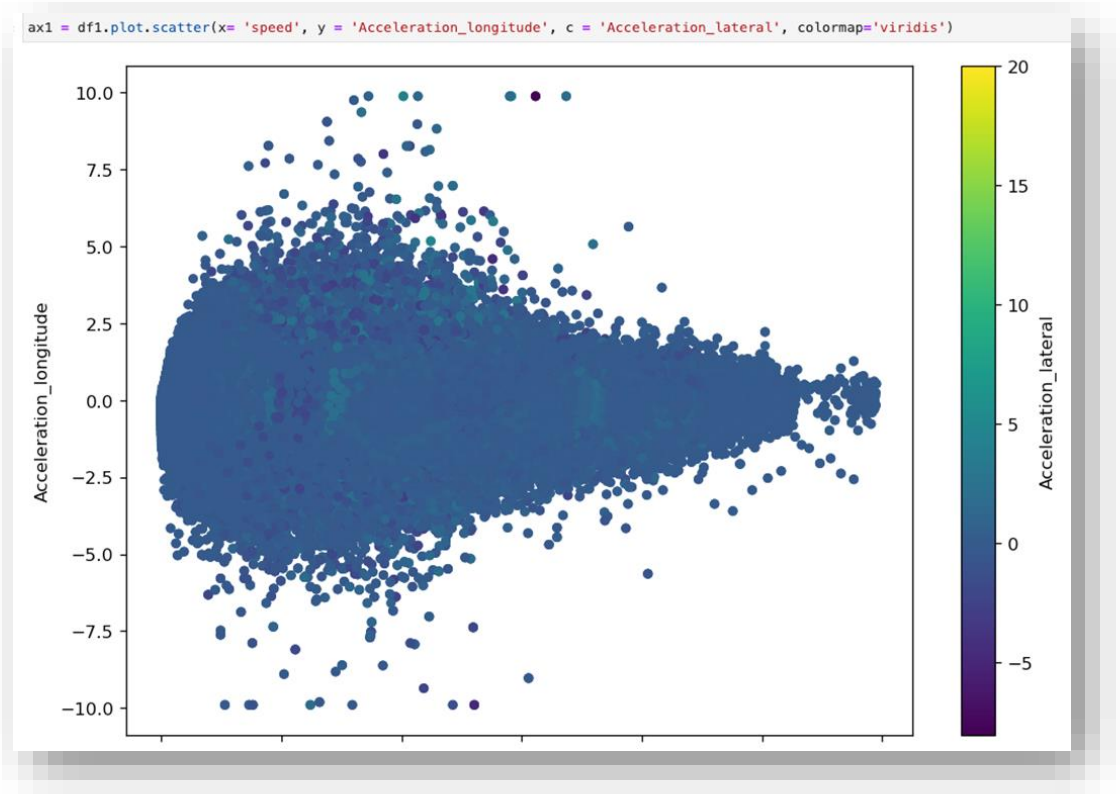

FIGURE 7.THE SCATTER PLOT OF ACCELERATIONS TO SPEEDS.

The processing of the BSMs is illustrated as follows: first, the csv file is converted to parquet format using the $Dask$ package in Python. This step reduces about 40% query time. Then the file is split into smaller data files by the vehicle ID: $DevID$. Subsets are created by $DevID$ and saved to 1527 small csv files. Since we are going to work on the individual level and we will always query by $DevID$, splitting big data into smaller ones drops query time from minutes to seconds. As the BSMs are generated at a frequency of $10\ Hz$, for each second there are 10 instances. We group the instances by seconds and take the average and derive average BSMs for the second. The next step is to select KPIs from the attributes. After initial investigation and according to the literature review, we selected KPIs as: $acceleration - longitudina$l, $acceleration - lateral$, $jerk - longitudinal$, $jerk - lateral$. Jerk is calculated using the moving average of accelerations by second using the Pandas module of Python. The formulas for calculating jerks at longitudinal and lateral in the i^{th} row of the Pandas $DataFrame$ are given in Equation (3-1) and (3-2).

$$jerk_{longitudianal[i]} = \frac{\text{acceleration}_{longitudinal[i]} - acceleration_{\text{longitudinal}[i-1]}}{\text{timestamp[i]} - timestamp[i-1]} \quad (3-1)$$

$$jerk_{lateral[i]} = \frac{acceleration_{lateral[i]} - acceleration_{lateral[i-1]}}{timestamp[i] - timestamp[i-1]} \quad (3-2)$$

Furthermore, we divide the KPIs to be positive and negative groups, e.g. $Acceleration - longitudinal\_possitive$ and $Acceleration - longitudinal\_negative$, because they represent different movements of the driver stepping on the gas or the brake and might have different patterns. Therefore, we have eight KPIs in total.

5.2.2 Module 2: Learning What Is Normal

Before detecting the anomaly, the system needs to learn what the "normal behaviors" look like. This is because the instances of crash are rare and obtaining labeled data of driving anomaly is prohibitively difficult while getting labels for normal behavior is much easier and less expensive. Nowadays in both research and practice, average thresholds at aggregate level are used. In our study we calculated the threshold at the individual level, resulting in a panel of thresholds for each driver.

For each KPI, the values of "normal behavior" are determined in this module. As we have resampled the individual BSMs by taking the average of each parameter, we group the rows by speed bins. The speed bins of size of 1 mph are set up and instances by seconds are redistributed to the bins. As discussed previously, the driver who is complying with its historical driving pattern is considered normal and we assume the values of each sample bin are normally distributed. The mean and standard deviations of each KPI are calculated for each speed bin. Thus, the panel of what is normal for an individual driver is generated, which is the information that needs to be abstracted from historical BSMs. In our study we use one-month data of 1527 vehicles and the calculation for this module costs three hours when we use only one process thread on our workstation. In practice, when the number of vehicles gets huge, parallel computing needs to be applied. The duration of the data for storage is a trade-off between the cost and the accuracy. The more the data stored the better the accuracy, but more expensive in storage and computation cost. The data panel of normal status of an individual vehicle is shown in Table . The rows are speed bins, the columns are the mean and standard deviation of the averages of positive and negative of $acceleration\_longitudinal$, $acceleration\_lateral$, $jerk\_longitudinal$ and $jerk\_lateral$.

TABLE 3. DATA PANEL EXTRACTED FROM AN VEHICLE (PARTIAL).

	speed_bin	0	1	2	3	4	5	6	7	8	9	10
acc_lon_p	mean	0.002436	0.148376	0.56307	0.787194	0.983954	1.192235	1.337538	1.32516	1.398614	1.323334	1.300234
acc_lon_p	std	0.002225	0.18192	0.392187	0.606541	0.733267	0.806321	0.839056	0.804345	0.804976	0.800801	0.76969
acc_lon_n	mean	-0.01405	-0.08628	-0.73162	-0.86491	-0.99819	-1.04187	-1.14423	-1.1853	-1.20188	-1.21563	-1.23855
acc_lon_n	std	0.00694	0.195623	0.507102	0.596489	0.661933	0.753567	0.74688	0.771699	0.786363	0.808835	0.804062
acc_lat_p	mean	20.01	2.196286	0.268705	0.05936	0.057242	0.069047	0.085095	0.096859	0.120503	0.212377	0.3169
acc_lat_p	std		6.06406	1.801026	0.396841	0.45105	0.431809	0.325246	0.154866	0.276925	0.371819	0.735151
acc_lat_n	mean	-0.01	-0.01008	-0.01336	-0.01803	-0.02216	-0.02688	-0.03648	-0.05113	-0.06153	-0.08261	-0.10866
acc_lat_n	std	4.88E-20	0.001113	0.009681	0.019348	0.028027	0.040362	0.07236	0.132858	0.170901	0.212122	0.300253
jerk_lon_p	mean	0.026028	0.223262	0.568748	0.709553	0.732469	0.824624	0.802729	0.692113	0.62276	0.542077	0.510498
jerk_lon_p	std	0.094186	0.37151	0.466118	0.591015	0.624505	0.696375	0.680028	0.612652	0.605413	0.552651	0.499198
jerk_lon_n	mean	-0.0088	-0.03473	-0.28252	-0.3217	-0.36004	-0.42201	-0.46223	-0.39244	-0.40045	-0.42351	-0.38723
jerk_lon_n	std	0.008	0.083935	0.315748	0.3446	0.344509	0.433422	0.487027	0.401976	0.395484	0.414006	0.360067
jerk_lat_p	mean	8.94E-06	0.00149	0.01969	0.023769	0.029042	0.035219	0.050722	0.043935	0.054583	0.092956	0.145396
jerk_lat_p	std	0.000325	0.069469	0.359987	0.157544	0.265913	0.286575	0.237766	0.083867	0.110184	0.174958	0.510506
jerk_lat_n	mean	-0.00174	-0.04154	-0.03556	-0.02783	-0.02013	-0.05251	-0.03598	-0.04478	-0.05237	-0.09907	-0.10388
jerk_lat_n	std	0.002408	0.442005	0.372267	0.268677	0.032055	0.582464	0.064078	0.084626	0.126077	0.458761	0.282969

Speed bin		**5**	**6**	**7**	**8**
KPI	Measure				
Acc_lon_p	Mean	1.192235	1.337538	1.32516	1.398614
	Std	0.806321	0.839056	0.804345	0.804976
Acc_lon_n	Mean	-1.04187	-1.14423	-1.1853	-1.20188
	Std	0.753567	0.74688	0.771699	0.786363
Acc_lat_p	Mean	0.069047	0.085095	0.096859	0.120503
	Std	0.431809	0.085095	0.096859	0.120503
Acc_lat_n	Mean	-0.02688	-0.03648	-0.05113	-0.06153
	Std	0.040362	0.07236	0.132858	0.170901
Jerk_lon_p	Mean	0.824624	0.802729	0.692773	0.62276
	Std	0.696375	0.680028	0.612652	0.605413
Jerk_lon_n	Mean	-0.42201	-0.46223	-0.39244	-0.40045
	Std	0.433433	0.487027	0.401976	0.395484
Jerk_lat_p	Mean	0.035219	0.050722	0.043935	0.054583
	Std	0.286576	0.237766	0.083867	0.110184
Jerk_lat_n	Mean	-0.05251	-0.03598	-0.04478	-0.05237
	Std	0.582464	0.064078	0.084626	0.126077

5.2.3 Module 3: Detecting Outliers

Outliers or anomalies are data points that do not meet the condition of what is normal in Module 2. As we assume the BSMs are normally distributed, the data located in the 95% probability regions are considered normal and the other 5% as outliers. From statistics, the cut-off value of 95% is two times of standard deviation away from the mean. Question might arise on our assumption: are the KPIs normally distributed? If a data set is normally distributed, the residual needs

to be random. The answer is no and but approximately yes. As shown in Figure 8, the $acceleration\_longitudinal$ of ID 6010 is not strictly normally distributed but close enough. Other scholars found it can be simulated with Negative Binomial distribution (Liu 2016), but we decide to take an approximation of normal distribution because we are solving an engineering problem, all we need to know is whether the vehicle has the potential to cause a crash, and we can use engineering alternatives to replace difficult mathematical problems. This philosophy is similar to our leaving the coordinates (environmental) impact from BSM to SSAM.

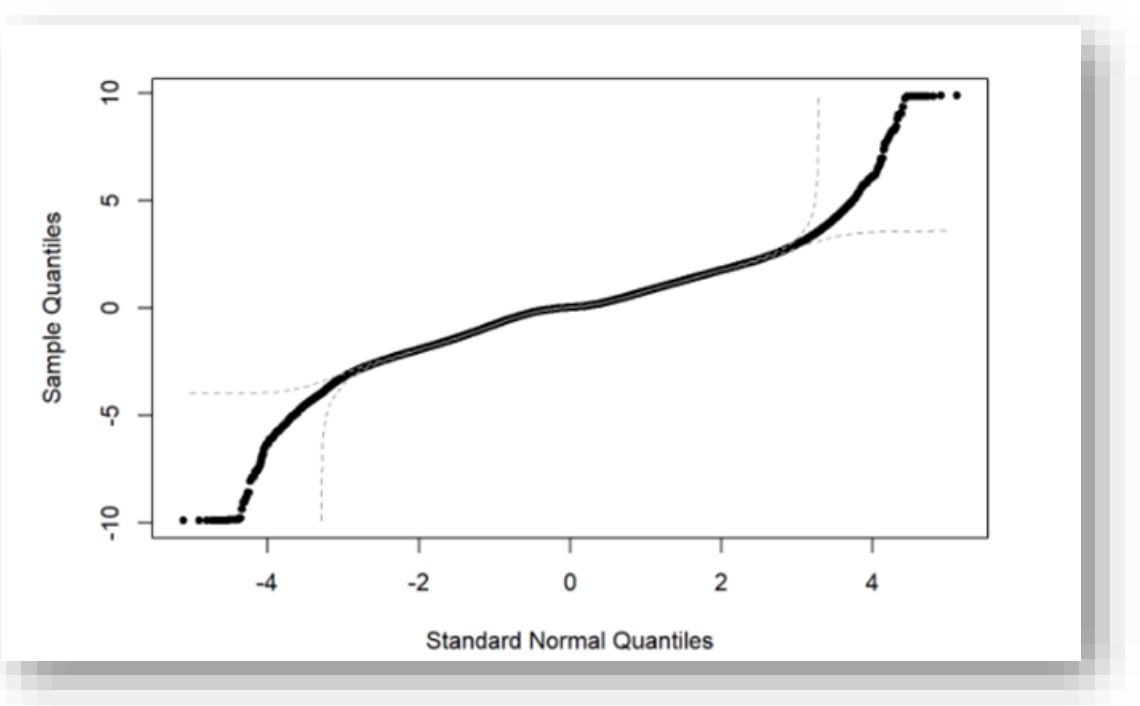

FIGURE 8. Q-Q PLOT OF LONGITUDINAL ACCELERATION OF A SAMPLE VEHICLE.

We set the regions of low probability to be 5 percentiles, with the values falling onto the range of $(mean - 2 * standard\ deviation, mean + 2 * standard\ deviation)$. The ranges of all KPIs at all the speed bins compose the thresholds. In our test runs there are many outliers detected. Since our goal is to minimize false-alarms, we are going to reduce the number of false-positives with the subsequent modules. In our DAD system, this outlier detection module is processed in the in-vehicle device. The in-vehicle computer stores the up-to-date thresholds received from the TMC. When the vehicle starts, and the in-vehicle computer will start generating BSMs. The new BSMs will be processed to detect outliers.

5.2.4 Module 4: Determine Abnormal Driving Event

A single outlier might not mean abnormal driving status, but multiple outliers in a short period of time signals anomaly: either the driver or the vehicle is not in good condition. In a DAD system, the outliers can be scored by the magnitude of deviation from the norm and/or duration of the outliers. In cases of comparing the impacts from different KPIs, a machine learning (ML) technique of normalization is commonly used. Considering it might introduce unnecessary uncertainties as we are not clear about the relationship between the comparative impacts of different KPIs, we decide not to use the magnitude of deviations. In the literature, five successive abnormal events of accelerations together warrant a safety alarm (Liu, 2016). If five successive outliers happen in a row of the same KPI, or more than two KPIs are outliers in the same second warrants an event of abnormal driving status.

More research and calibration need to be done to make the initial state more reasonable. As our system is dealing with multiple stages of uncertainties on the driving status, normal driving status and abnormal driving status, and the output is an alarm for a possible crash, which is also uncertain, we are aiming at minimizing false alarms, but the system needs to be calibrated when used. Here we build the system and leave the users to adjust the values of the parameters according to the local conditions.

5.2.5 Module 5: System Updating

So far, our system is designed to update all the thresholds in batch mode periodically. When the data of the next month are collected, they will be processed by the system in the same way as described in this section. We mentioned the period to be one month simply because the only data we have covers a month's period. However, a more advanced way of system updating is to apply auto-tuning. In DAD, auto-tuning adjusts the thresholds to provide an accurate baseline. After an anomaly is detected, the system needs to decide whether to use it to update what is normal. A driver might be in abnormal status today, for example DUI, and will be normal again the next day. In this case, no updating is needed. But there are cases that the driver changes his/her driving habits. For example, a near-sighted driver who starts to wear glasses and can see clearer than before might use a higher acceleration/deceleration rate. In the beginning, the system might treat it as abnormal, but if the anomaly persists then the system needs to gradually accept it and update to the new state. The tool to control how the anomalies are treated is the learning rate, which can adjust the trade-off between how fast the system learns and how adaptive it is. For example, the learning rate can be defined as 0.001 for the first day when the abnormal status is detected, and the corresponding threshold will change by

0.001 times, or almost no change happens since the rate is very small; if similar anomaly persists for 5 days in a row, the learning rate will be assigned a much bigger value, say 0.1. The ML algorithm of K-means is utilized to classify the abnormal status of the days. K-means is based on similarity of multidimensional variables. We will discuss auto-tuning in the Results and Discussion section.

5.3 Evaluation of the DAD Model

Evaluation on unsupervised anomaly detection is a constant challenge, so is the application of machine learning to practical engineering problems. Nevertheless, modeling driving status using BSMs of connected vehicles at the individual level is an unavoidable task for the automation of our traffic safety diagnosis system. In the model evaluation, the measurement of average precision is utilized, and the model is validated. Our model is a combination of machine learning (ML) and engineering modeling. As a data science technique, ML has become a hot topic in many domains since the deluge of big data, but when it comes to engineering, where accuracy and proving is emphasized, ML is not as successfully applied. One of the reasons is that ML is known as a "black box", neither convincing nor easily assortative with domain knowledge. Through this research, we found it a practical way to build the ML model for engineering problem, basing on the domain knowledge, and use measures in the information system to validate the model using underlying assumptions.

As our goal of building the DAD model is not to compete with other algorithms but to apply it in the real-world traffic safety engineering, in model evaluation, we evaluate the reasonableness of the model — whether the model can function as proposed. The research is based on the assumption that the driver is under abnormal driving status in an accident. So, the trajectory of the accident trip should have more outliers than a normal one. In our previous study, we found that the KPI data did not follow any statistical distribution strictly, but somehow close to the normal distribution. So, we assume they are normally distributed. In the model building, we set the inliers to be the ones that fall in the range of two times the standard deviation around the mean of each interval and consider the cases that is out of that range to be the outliers. Therefore, the 95% of the instances should be normal, if the driving status is normal, but in the trip when the driver is under abnormal driving status, the outliers of the trajectory should be more than 5%. Hence, we set up the test as follows: select the accident trip files of all the drivers, if the number of anomaly cases detected is more than 5%, then the DAD model is valid. In the test, we calculated the average precision of each driver's testing file.

The measure precision is the fraction of relevant instances among the retrieved instances and the average precision is a measure that combines recall and precision for

ranked retrieval results. The average precision is the mean of the precision scores after each relevant record is retrieved. Mathematically, the average precision is written as Equation (3-3) where r is the rank of each relevant document, R is the total number of relevant records, and $P@r$ is the precision of the top-r retrieved records (Zhang2009).

$$AveragePrecesion = \frac{\sum_r p@r}{R} \qquad (3-3)$$

The results of the test show that all the 42 drivers have been found to exhibit more than 5% abnormal of the accident trips, which proves that the DAD model is valid, as shown in Figure 9.

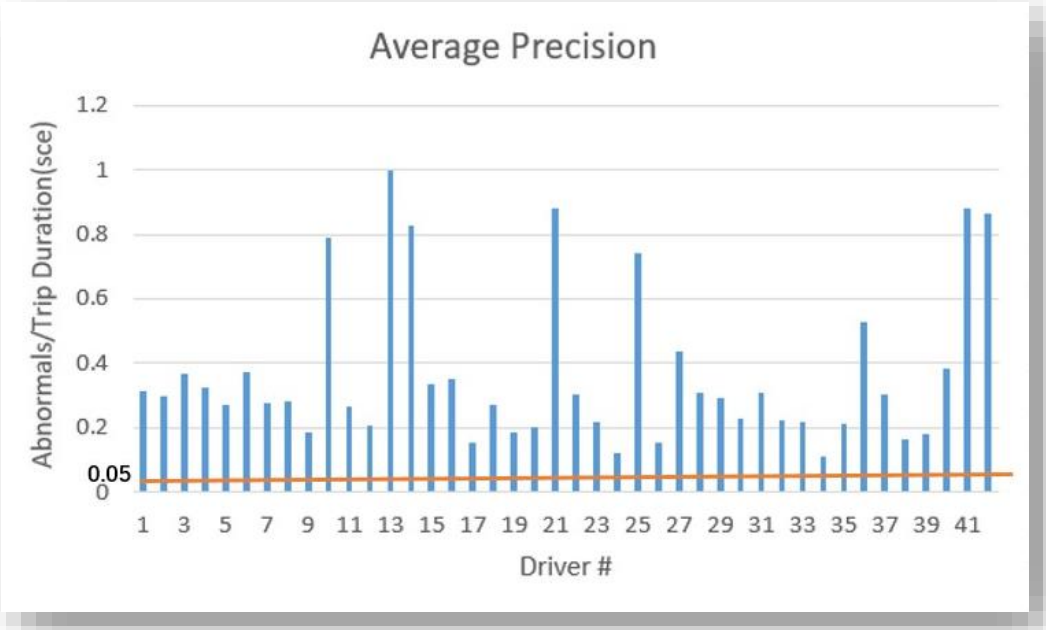

FIGURE 9. EVALUATION OF THE OUTLIER DETECTION MODEL.

5.4 Sensitivity Analysis of the DAD Model

In the domain of engineering, sensitivity analysis is a widely used tool in model evaluation. Parameter sensitivity analysis is usually performed in which a series of tests on the model with different parameter values to observe the dynamic behavior of the model responses to the parameter changes. And proper parameter values can be recommended through analyzing the patterns of the results.

Again, as our goal of building the DAD model is for real world application, the number of safety alarms need to be reasonable. In our tests, many abnormal instances were detected in a test file, from a few to thousands depending on the driver and trip duration. Too frequent alarms might annoy the driver and a single outlier might not mean abnormal driving status, but multiple outliers over a short period of time do signal anomaly. Therefore, we added Module five in our DAD model to cut down the occurrences of alarms. Sensitivity analysis is performed to generate the reasonable numbers of alarms.

In our model, some parameters are defined as follows:

N_v – $the\ number\ of\ KPIs\ being\ identified\ as\ outliers\ in\ the\ same\ second$;

N_s – $the\ number\ of\ successive\ seconds$;

N_{std} – $the\ number\ of\ times\ of\ standard\ deviation\ away\ from\ the\ mean\ to\ calculate\ the\ thresholds$;

N_d – $the\ number\ of\ days\ prior\ to\ the\ crash\ to\ calculate\ the\ threshold$.

An abnormal event (triggering an alarm) will be warranted if any of the following conditions is met:

1. The number of KPIs being identified as outliers in the same second is larger or equal to Nv;

2. Within Ns more than one KPI are identified as an outlier in a row.

In the sensitivity analysis, we test with various values of Nv and N_s to determine the best value by observing the model response. As aforementioned, we treat the threshold as a variable, here we use sensitivity analysis to investigate the proper value range. We are interested in how the model responds to the days prior to crash to calculate the thresholds. We also treat the number of days prior to the crash as a testing parameter. The parameter settings for sensitivity analysis are shown in Table .

TABLE 4. PARAMETER SETTING FOR SENSITIVITY ANALYSIS.

Parameter	Test Value	Initial Value
N_v	1, 2, 3, 4, 5, 6, 7, 8	2
N_s	3, 5, 10, 15, 20, 30	5
N_{std}	2, 2. 25, 2. 5, 2. 75, 3	2
N_d	15, 30, 45, 60	30

5.4.1 Sensitivity Analysis on N_v

In our model, there are 8 KPIs: acceleration-longitudinal, acceleration-lateral, jerk-longitudinal, jerk-lateral, each of which has positive and negative items. In sensitivity tests, we tested the Nv value from 1 through 8. The model responses are shown in Figure 10 . With Nv set to be 1 or 2, more than 15% of the instances of the trip seconds will be identified as alarms, which will result in too many alarms. Meanwhile the detailed recorded data show that in many cases, the acceleration and jerk at the same direction were identified as outliers at the same time, which indicated that the pair are correlated to some extent. So, we eliminated values 1 and 2. Figure 10 also shows that when Nv is more than 3, the curve becomes flat, which means the number of abnormal cases identified are very close. Therefore Nv was determined to be 3.

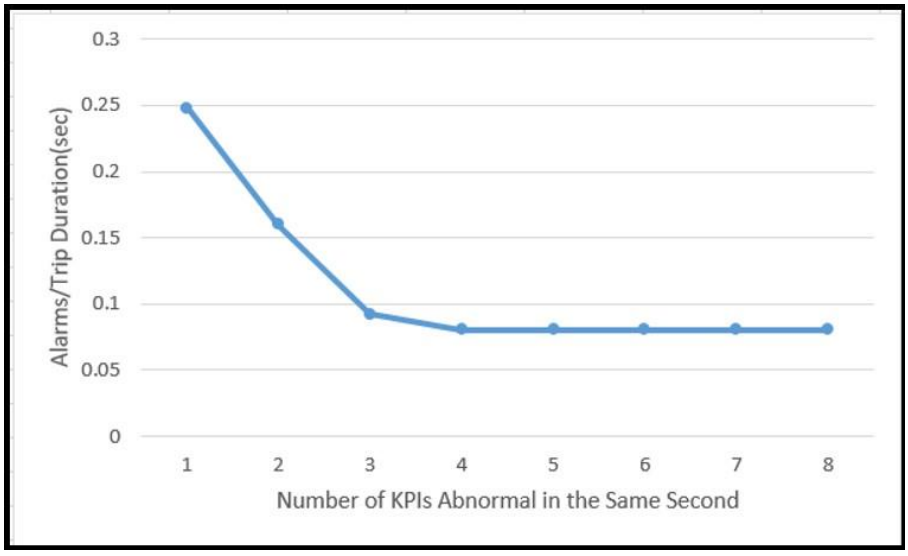

FIGURE 10. THE SYSTEM RESPONDING TO A STEP INCREASE IN THE NUMBER OF KPIS IS DETECTED AS ABNORMAL IN THE SAME SECOND.

5.4.2. Sensitivity Analysis on N_s

Figure 11 shows how the system responds to the various values of the number of successive seconds when a single KPI is found to be abnormal in a row. The value of 10 seconds is selected for N_s because it is where the curve changes the slope at that point and the value of the ratio is close to 5%.

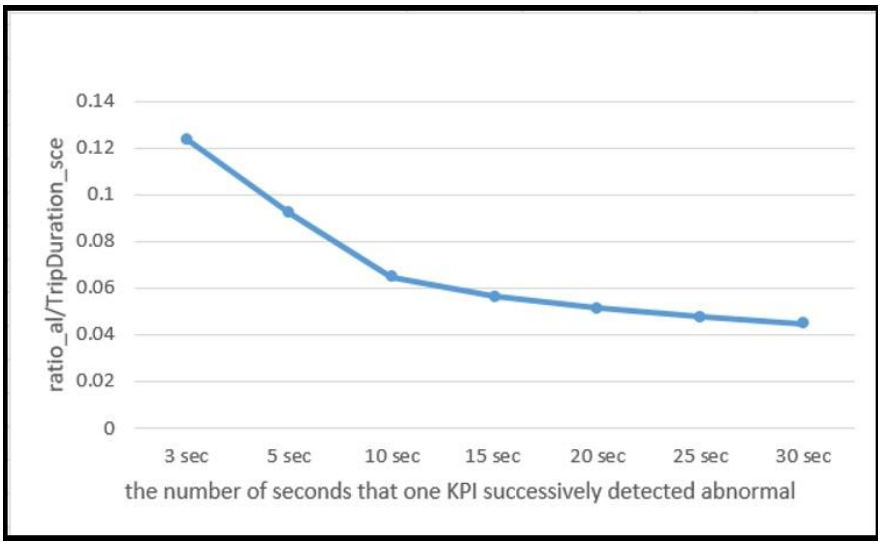

FIGURE 11. THE SYSTEM RESPONDING TO A STEP INCREASE IN THE NUMBER OF SECONDS THAT ONE KPI SUCCESSIVELY DETECTED ABNORMAL.

5.4.3. Sensitivity Analysis on N_{std}

Figure 12 shows how the system responds to the different settings (the number of times of standard deviation away from the mean) to calculate the thresholds. It shows that the value of 2 and 2.5 did not result in significant changes. And from Figure 13, no alarms are generated for the testing file after the value of 2.25, which violates the purpose of the model, which is to detect abnormal for all the potential crashes. We decided to use 2 because this is the widely used value and 2.25 did not make significant difference.

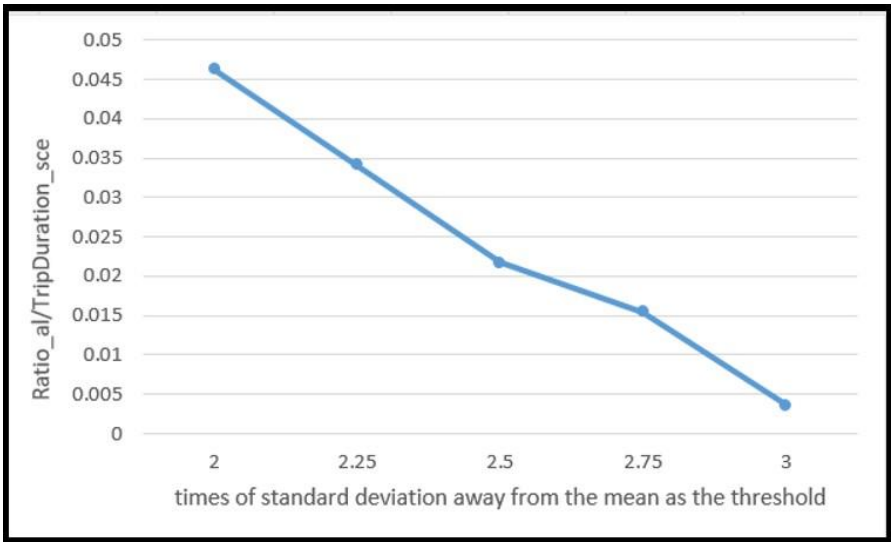

FIGURE 12. THE SYSTEM RESPONDING TO A STEP INCREASE IN THE TIMES OF STANDARD DEVIATION AWAY FROM THE MEAN.

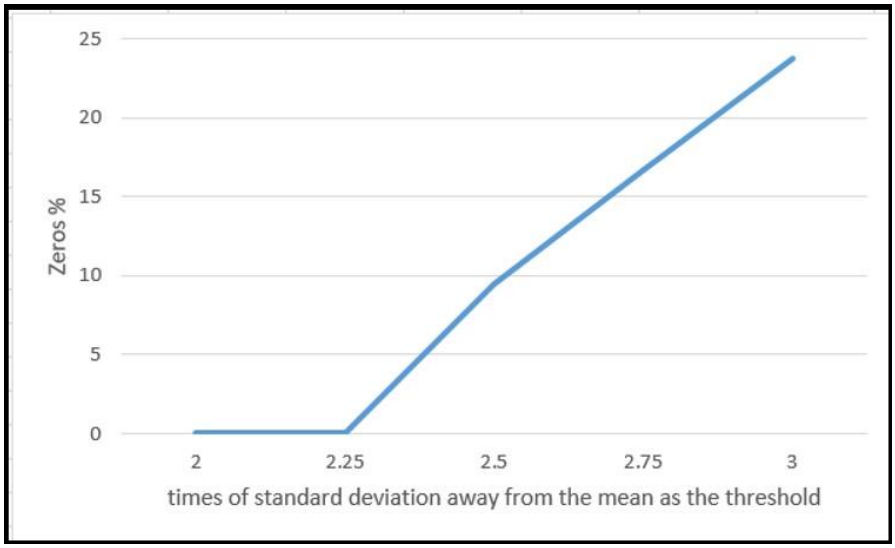

FIGURE 13. THE SYSTEM RESPONDING TO A STEP INCREASE IN THE TIMES OF STANDARD DEVIATION AWAY FROM THE MEAN.

5.4.4. Sensitivity Analysis on N_d

Figure 14 shows how the system responds to how many days the cloud saves the raw BSMs to calculate the thresholds. We assume that the cloud uses the batch mode to calculate the thresholds. The curve changes values within a small range, which means that system is not highly sensitive to the change ofN_d. We selected 30 days because it identifies the most anomaly events. In practice, this parameter is better to be determined by the number of vehicles covered by the cloud and the computational capacity of the server. Furthermore, the auto-tuning is expected to replace the batch mode, then this parameter will no longer exist.

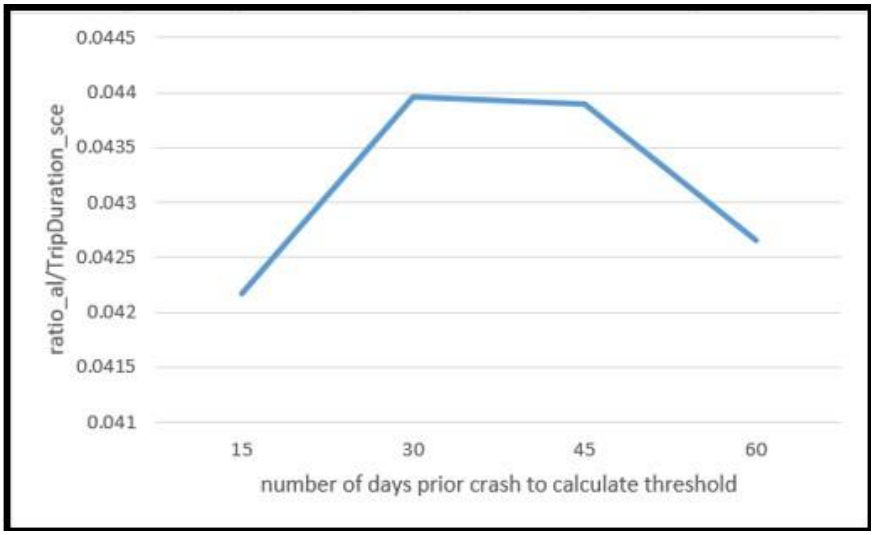

FIGURE 14. THE SYSTEM RESPONDING TO A STEP INCREASE IN NUMBER OF DAYS PRIOR CRASH TO CALCULATE THRESHOLD.

5.5 Results and Discussion of the DAD Model

The results from the proposed driving anomaly detection (DAD) model are the threshold panel of what is normal for an individual vehicle, which is the information that needs to be extracted from historical BSMs and stored in the TMC. In the threshold panel, the selected KPIs include $acceleration - longitudinal, acceleration - lateral, jerk - longitudinal,\ and\ jerk - lateral$. The mean and standard deviation of each KPI in each speed bin of a size of 1 mph are the major contents of the panel.

Before model implementation, as our working data is a TS data, we performed standard TS data analysis with no periodicity identified. This is reasonable because the driving behavior is substantially complicated, and the majority of TS data do not have periodicity any way. Therefore, we conclude that the models for the TS data do not apply to BSMs. We implement our DAD model from scratch instead of using the existing machine learning models.

After implementation of DAD, we performed model evaluation using the Average Precision method to evaluate our DAD model. The evaluation results show that our DAD model is valid. Then, sensitivity analysis was carried out to determine the recommended values for some model parameters. Table summarizes the results of the sensitivity analysis.

TABLE 5. DETERMINED PARAMETER VALUES.

Parameter	Determined Value
Nv	3
N_s	10
N_std	2
N_d	30

There are several limitations. First of all, human behavior is complicated, and the attempt of determining the behavior status based on the footprint of a vehicle can be inaccurate. Second, in scoring the outliers, we are not clear about the relationship of the comparative impacts of different KPIs. We keep whatever KPI that might have some impact instead of running the statistical testing to exclude those not statistically related KPIs. This is again because of the complication of human behavior, and we do not have the luxury of plenty of data and understanding of human mental processing. And finally, we did not run the auto-tuning due to lack of data and the changing of driving habits might need time longer than one month.

In this section, we described a DAD system that determines if the driver is in abnormal driving status according to the driving volatility using solely the BSM data. We explained the theoretical foundation, the mathematical model of the proposed DAD model and performed model implementation. The resulted threshold panels are what need to be extracted from the BSMs and need to be stored in the cloud for traffic safety analysis. The proposed DAD passed the model evaluation and through sensitivity analysis the recommended values of certain model parameters were obtained based on the working data.

6.0 TASK 2: CONFLICT IDENTIFICATION MODEL (CIM)

6.1 Introduction of CIM

A major obstacle to the prevalence of AV is safety. Currently, the safety of AVs relies largely on the surveillance systems and motion detection in the ego vehicle. The real-world detection is affected by many factors such as weather, interference, and sensibilities. This safety issue can be mitigated by sending near crash warnings to the drivers in the CV environment, through analyzing the trajectories of the vehicles embedded in the BSMs.

In the literature, the research using trajectories to detect the potential crashes utilized the trajectory data that were collected on the scene. However, in the case of using BSMs, as the effective transmission distance of the V2V BSMs is limited, there might be no sufficient time to perform a chain of tasks to avoid a crash after the vehicles come into the effective V2V range, including collecting the data, training the model, analyzing

the data, broadcasting the alarm and for the driver to perceive the alarm and take actions. In order to leave adequate time to the drivers, we separate the process into two steps: step one is the driving anomaly detection including the threshold values determined in the cloud using historical BSMs and the driving abnormal detection (DAD) in the in-vehicle subsystem using real time BSMs; step two is to detect the conflicts by the in-vehicle system using the real time BSMs and the results from the first step. This section focuses on the second step in which we define a conflict as the condition when attention is needed when a vehicle is under abnormal state. This section also describes a conflict detection model (CIM) using the profiles of speed and distance to identify conflicts.

6.2 Methodology of CIM

6.2.1 Conflict Scenarios

The speed distance profile (SDP) of a vehicle is a sequence of time-stamped measurements of the vehicle's position and speed, often recorded by the odometer or the Global Positioning System (GPS) (Andrieu, 2013). The SDP here is the time-stamped sequence of the coordinates and speeds of a driver-vehicle units (DVUs) pair, in which at least one DVU is under abnormal status. The proposed CIM is to identify conflicts between the DVU pair using the SDP extracted from their BSMs. As the purpose of the ASDSCE is to generate warnings against potential crashes, the safest way to define the conflict scenario is to embrace the worst cases. Four of such scenarios are defined when DVU_a and DVU_b are close to each other (within the effective BSM V2V range) and at least one of them is in abnormal driving status (with a flag), as shown in Figure 15.

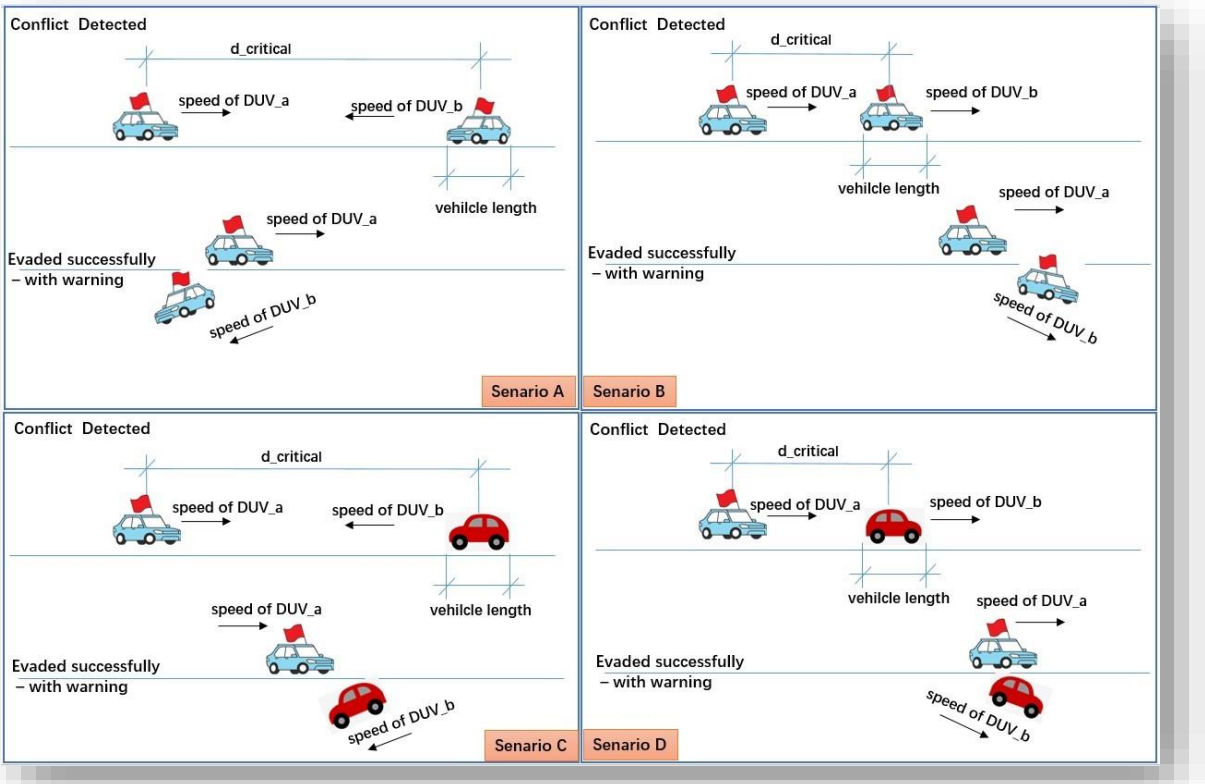

FIGURE 15. CONFLICT SCENARIOS UNDER ABNORMAL DRIVING STATUS.

- Scenario A: Head-on scenario in which both of DVU_a and DVU_b are in abnormal driving status (flagged). DVU_a is heading toward DVU_b at the maximum possible speed while DVU_b is trying to steer away;
- Scenario B: Car-following scenario in which both of DVU_a and DVU_b are flagged. DVU_a is heading toward DVU_b at the maximum possible speed while DVU_b is trying to steer away;
- Scenario C: Head-on scenario in which only DVU_a is flagged. DVU_a is heading toward DVU_b at the maximum possible speed while DVU_b is trying to steer away;
- Scenario D: Car-following scenario in which only DVU_a is flagged. DVU_a is heading toward DVU_b at the maximum possible speed while DVU_b is trying steer away.

In Figure 15, d_critical denotes the critical distance which is the distance between the DVU pair when a conflict is detected.

6.2.2 Mathematical Model for the Speed Distance Profile (SDP)

To illustrate our SDPs, as shown in Figure 15, we introduce the time to evade (TTE), which is defined as the time interval that DVU_b needs to perform a chain

of actions including hearing and understanding the warning, checking the surrounding and taking actions and steers away from the location of potential collision. During the TTE, the relative distance between the DVU pair decreases from the critical distance to zero. In order to determine the value of TTE, the perception-reaction time (PRT) was reviewed. In the transportation safety community, PRT is defined as the time for a driver to perceive and respond appropriately to an impending hazard (Bates, 1995). The American Association of State Highway and Transportation Officials (AASHTO) recommended a PRT of 2.5 seconds, including 1.5 seconds for visual perception time (VPT) and 1.0 second for reaction time, as the design standard for calculating the stopping sight distance (SSD). Although this PRT was determined through extensive studies and it passed some extreme test cases, "surprise intrusion" tests and in full-scale road tests, however, certain other tests showed that 2.5 seconds were not safe enough and some researchers called for increasing the value to 3.0 seconds or more (Sens, 1989; Grime, 1952). As drivers under influence typically have longer PRT, our understanding is that AASHTO's standard was set for average normal drivers and a higher value of PRT is more appropriate for abnormal drivers. Our TTE is similar to the PRT but with one difference – PRT was defined to perceive and react to the same object – the conflicting vehicle, while TTE is used to perspective two objects – the warning and the conflicting vehicle. So, an additional audio perception time (APT) to hear the warning, and pass it to the brain, and for the brain to comprehend the warning and instruct the eyes to look for the conflicting vehicle needs to be added. As APT was tested shorter than VPT (Jain, 2015), we take the value of 1.0 second for the APT. Therefore, in our case, the TTE is set up to be 3.5 seconds for the normal DVUs, and 4.0 seconds for the abnormal DVUs. From the results of the related studies (Bokare, 2017; Kusano, 2011), the maximum acceleration for the vehicle was selected as $2.87m/s^2$ for the critical distance calculation, and a deceleration rate of $0.52m/s^2$ was selected for the braking distance calculation.

Using the SDP data extracted from the BSMs, we constructed the math model to detect the potential conflicts. Given that DVU_a and DVU_b are in the V2V effective range (which means they can exchange BSMs) and at least one of DVU_a and DVU_b is under abnormal status. The in-vehicle subsystem will check the headings – $ABS * (heading_a - heading_b)$ to determine the scenario type and the distance d between DVU_a and DVU_b, as shown in Eq. (4-1). If d is not greater than the critical distance d_{crit}, as shown in Eq. (4-2), or in Eq. (4-3), then a conflict is identified. In the equations, the location of DVU_a and DVU_b are denoted as $P_a\ (x_a, y_a)$, $P_b\ (x_b, y_b)$, respectively, l denotes the length of a DUV, and $v_{a,}$ and v_b denote the speeds of DVU_A and DVU_B, respectively. $TTE =$

$3.5seconds/4.0seconds$, and $a_{max} = 2.87m/s^2$ denotes the maximum acceleration of DVU_a.

$$d = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \quad (4-1)$$

$$d_{crit-head-on} = TTE * (v_a - v_b) + 0.5 * a_{max} * TTE^2 + l \quad (4-2)$$

$$d_{crit-car-floowing} = TTE * (v_a - v_b) + 0.5 * a_{max} * TTE^2 \quad (4-3)$$

Thus, the conflict in our system is defined as the situation when the actual distance between DVU_a and DVU_b is not greater than the critical distance when abnormal driving status is present in at least one of DVU_a and DVU_b.

6.3 Case Study of CIM

We implemented the CIM in Python for all of our working data which are in the comma-separated values (CSV) format and Python is powerful in manipulating tabular data. In the CIM, we loaded the CSV files of the DUV pair that are under investigation to different Pandas $DataFrames$.

6.3.1. Data Description

The SHRP2 data set has two sets of speeds: the network speed and the GPS speed. The network speeds are recorded at the frequency of $10Hz$ and the GPS speed at $1Hz$. The network speed is generated by the vehicle's speedometer through multipart tools such as dive cable, speed cup, hairspring, and pointer needle on the dial panel. It includes many errors because of the long-chain process and different manufacturers might have different standards of error tolerances. The network speed is shown on the driver's dash panel to give the driver some idea of the driving speeds. The network speeds usually are higher than the actual speeds, with acceptable error tolerance of 10 percent. On the other hand, the GPS speed is more accurate as it is from the satellite. We use network speeds to test the driver's driving status for the driver is directly influenced by them. The speed bin was set up as 1 mph, and the instances are aggregated to 1 sec with means.

In order to simulate the BSMs, which were generated at a frequency of 10 Hz, we paused the program for 0.1 second after each time it reads one record of the data. Although the proposed CIM is straightforward in theory, to demonstrate it is difficult because of lack of data. We did not find any crash record that both of the involved vehicles are CVs as crashes are rare events and the number of CVs in the CV pilot studies were limited. However, in the Naturalistic Driving Study (NDS) $InSight$ Data, there are some crashes recorded between one equipped vehicle and a stationary object, such as a

fence or a roadway curb. For our case study, a total of 23 such crashes are selected, in which the drivers were reported under abnormal driving status, such as DUI, driving while texting on the phone or being tired after long driving etc. The driving status information were provided by the Strategic Highway Research Program (SHRP II) NDS $InSight$ Data Access (SHRP2, 2020). Under the principle of the privacy protection, the coordinates of the trajectories of the 23 cases were revised to be relative to certain points that were unpublished. As the stationary object can be used to represent the normal DUV with the speed of zero, all the 23 cases fall into scenario C as described in Subsection 2.4. We created the data set of DVU_a as staying on the crash point from each trajectory files of the 23 crashes. In the CIM testing, when a conflict is firstly identified, the $timestamp$, the speed of DVU_b and the distance between DVU_a and DVU_b, are recorded.

6.3.2. CIM Algorithm and Running Results

Figure 16 shows that in all of the 23 cases, the first conflict was identified at least 19 seconds before the crash. This indicates that DVU_a should have enough time to take evasive actions and demonstrates that the CIM is functioning as proposed. Table records the details of the tests runs.

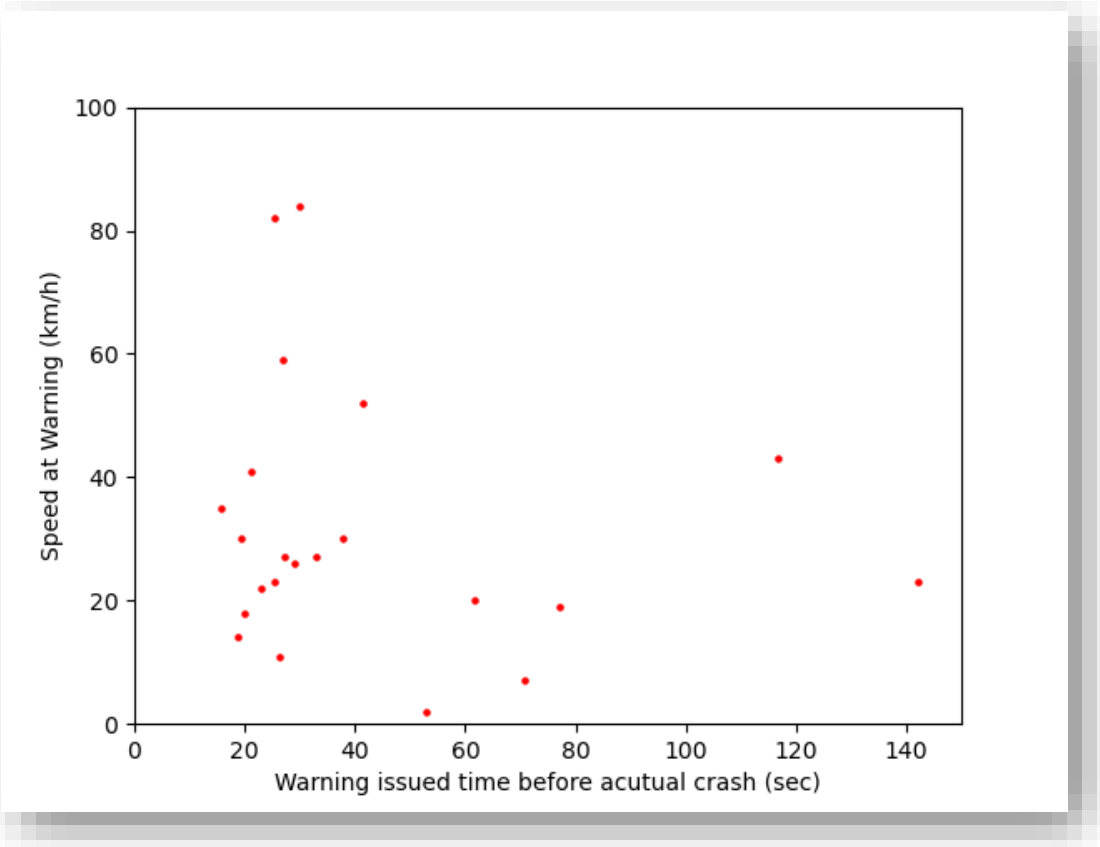

FIGURE 16. SPEED AND TIME REMAINING OF CONFLICTS IDENTIFIED FIRST TIME IN TEST RUNS

Table 6. Conflict identification test records.

Case Number	Speed	Distance	Confict_timestamp	Crash_timestamp	Remain_time
Unit	km/h	m	sec	sec	sec
1	82	117	23	49	26
2	35	65	207	223	16
3	18	46	9802	9822	20
4	19	47	1000	1077	77
5	20	48	1053	1115	62
6	23	51	363	505	142
7	22	50	10594	10617	23
8	26	55	406	435	29
9	84	119	1349	1379	30
10	41	49	2428	2449	21
11	14	19	2397	2415	18
12	55	87	1335	2320	985
13	52	84	1649	1690	41
14	11	38	518	545	27
15	27	56	1189	1222	33
16	2	18	3	56	53
17	7	7	4	75	71
18	23	51	1349	1375	26
19	30	59	64	83	19
20	43	74	3633	3750	117
21	59	91	2493	2520	27
22	27	56	3110	3137	27
23	30	59	522	560	38

6.4 Results and Discussion of CIM

In this section we constructed a conflict detection model (CIM) using the speed distance profile (SDP) to detect conflicts between a driver-vehicle unit (DVU) pair under abnormal driving status. In our system, a conflict is defined as a traffic situation involving a DVU pair which satisfies the following two conditions: (1) at least one of them is under abnormal driving status; (2) the actual distance between them is less or equal to the critical distance, which is calculated by the time to evade (TTE) with the maximum possible approaching speed. The model was tested on the SHAPII crash data. The results show that the conflict identification model can function as expected.

The contribution of this CIM lies in that it creatively introduced a collision warning tool using the data sources not from the traditional in-vehicle sensors but from the BSMs of the CV environment. Collision warnings from outside the ego vehicle can greatly enhance safety especially in the circumstances of unexpected malfunctioning of the ego vehicle. This section also emphasizes the importance of abnormal driving status from a systematic viewpoint. Abnormal driving status is a major collision causation factor and deserves more attention of the ADAS. The authors call for putting focus on abnormal driving status instead of the normal drivers. Substantial future work is expected to

specialize in many aspects of the CIM, including but not limited to, implementing the thresholds and the flag list in the datapath, upgrading the model from sequential programming to parallel programming, specifying vehicle type, vehicle length, acceleration rate, deceleration rate, and improving the sophistication of the conflict scenarios.

The CIM is a component of the in-vehicle subsystem of ASDSCE, in which a near crash warning will be generated if a conflict is detected between the pair of ego CV and a nearby CV when any CV in the pair is under abnormal driving status. The functionality of CIM is to generate collision warnings solely using BSMs.

This section reports our work on several issues: a) redefining the conflict. Conflict is a key concept of surrogate safety analysis, which was originally designed for traffic simulation data. We tailor it to fit our system; b) developing the mathematical algorithms to identify the conflicts; c) implementing the algorithms in the in-vehicle subsystem. The algorithm is tested on the SHRP2 crash data which contain similar features of the BSMs.

7.0 CONCLUSIONS

In this project, we built an automatic safety diagnosis system in the connected vehicle environment (ASDSCE). It is a real-time near crash warning tool on the individual level specifically configured for BSMs. The architecture of the proposed composed of two components: one is the driving anomaly detection (DAD) model, which collects and stores historical BSMs in the cloud and determines in batch mode the thresholds of each vehicle and identify the abnormal driving behavior from the real-time BSMs; the other is a conflict identification model (CIM) which is in the in-vehicle subsystem which detects conflicts. A near crash warning will be warranted when the traffic situation satisfies both of the following two conditions: (a) a conflict is identified and, (b) at least one of the drivers that is involved in the conflict is in abnormal driving status.

Using solely the BSM data, the DAD system determines if the driver is in abnormal driving status according to the driving volatility. The DAD contains two parts: one is in the cloud where the threshold panels defending what is normal of each CV are generated using the historical BSMs; the other is in the in-vehicle computer where the current BSMs of the ego vehicle are compared with the thresholds that are being broadcasted from the cloud. We explained the theoretical foundation and the mathematical algorithm for the proposed DAD model and implemented the model. To answer the initial project target problem, the content of what need to be extracted from the BSMs and can to be stored in the cloud for traffic safety analysis are the threshold panels of all the individual CVs. The proposed DAD model passed the model evaluation. Through sensitivity analysis the recommended values of certain model parameters were established based on the working dataset.

The CIM is a component of the in-vehicle subsystem of ASDSCE, in which a near crash warning will be generated if a conflict is detected between the pair of the ego CV and a nearby CV when any CV in the pair is under abnormal driving status. The functionality of CIM is to generate collision warnings solely using BSMs. In building the CIM, we redefined conflict to fit our system, developed the mathematical algorithm and implemented the algorithm with Python. The algorithm is tested on the SHRP2 crash data.

The ASDSCE contains the following features: focusing on detecting abnormal drivers instead of normal drivers; using the trajectory data embedded in the BSM to study driving volatility; implementing on the individual level instead of the aggregate level; and reducing the model training time to leave sufficient time to the involved drivers to perform successful evasive actions. The present computational pipeline of ASDSCE includes raw data collection, data preprocessing, data analysis, data communication and warning message generation. ASDSCE is built with Python on Visual Studio 2019 using the BSMs from the CV pilot studies and evaluated using the SHRP2 naturalistic driving study crash data.

The ASDSCE system can be used as a real-time near-crash warning tool in the CV environment. This project can help to improve the safety of CVs in driving and It open a new approach for the safety of CAV operations.

8.0 RECOMMENDATIONS

The contribution of this project lies in that it creatively introduces a collision warning tool using the data sources not from the traditional in-vehicle sensors but from the BSMs generated in the CV environment. Collision warnings triggered from outside the ego vehicle can greatly enhance safety especially in the circumstances of unexpected malfunctioning of the ego vehicle. As the automatic safety diagnosis system in the connected vehicle environment (ASDSCE) utilizes solely the BSM data, the ASDSCE can serve as an additional collision warning tool supplementing the current tools that rely on the data collected by the sensors on the ego vehicle. However, the system is based on many assumptions due to lack of data and because of its nature of complexity, the system still needs fine tuning on many aspects.

There are several limitations of the DAD model. First of all, human behaviors are complicated. The attempt to determine the behavior status through the footprint of a vehicle can be inaccurate. Second, in scoring the outliers, we are not clear of the relative impacts of different KPIs. We keep whatever KPI that might have some impact instead of running the statistical testing to exclude the not statistically related KPIs. This is again because of the complication of human behavior, and we do not have the luxury of plenty of data and understanding of human mental processing. And finally, we did not run the auto-tuning due to lack of data and changing driving habits for a driver might need time longer than one month.

This project also emphasizes the importance of abnormal driving status from a systematic viewpoint. Abnormal driving status is a major collision causation factor and deserves more attention of the ADAS. The authors call for putting focus on abnormal driving status instead of the normal drivers. Substantial future work is expected to investigate many aspects of the CIM, including but not limited to, implementing the thresholds and the flag list in the datapath, upgrading the model from sequential processing to parallel processing, specifying vehicle type, vehicle length, acceleration rate, deceleration rate, and improving the sophistication of the conflict scenarios.

The datapath is not within the study scope of this project. The datapath involves the vehicle cloud which is an open research problem and is one of the major challenges of the CV. Future efforts are expected on the datapath research.

The computational models of this project are coded in sequential manner. As in the real word of CVs, the BSMs and the number of CVs will be overwhelming. Real-time safety analysis demands parallel computing to speed up data processing. Therefore, upgrading the current project by employing parallel computing technologies is inevitable.

9.0 REFERENCE LIST

1. AlRajie, H. O. Investigation of using microscopic traffic simulation tools to predict traffic conflicts between right-turning vehicles and through cyclists at signalized intersections. PhD thesis. Carleton University, 2015.
2. Amundsen, F. H. and Hyden, C. Proceedings of first workshop on traffic conflicts. In: Oslo, TTI, Oslo, Norway and LTH Lund, Sweden (1977), p. 78.
3. Andrieu, C., Pierre, G. S. and Bressaud, X. A functional analysis of speed pro les: smoothing using derivative information, curve registration, and functional boxplot. In: arXiv preprint arXiv:1312.2252 (2013).
4. Archer, J. Indicators for traffic safety assessment and prediction and their application in microsimulation modelling: A study of urban and suburban intersections. PhD thesis. KTH, 2005.
5. Bates, J. T. Perception-reaction time. ITE journal. 1995;65 (2) :35-6.
6. Bokare, P. S., and Maurya, A. K. Acceleration-deceleration behaviour of various vehicle types. Transportation research procedia. 2017 Jan pp: 4733-49.
7. Bose, A., and Ioannou, P. A. Analysis of traffic flow with mixed manual and semiautomated vehicles. IEEE Transactions on Intelligent Transportation Systems. 2003 Dec;4(4):173-88.
8. Boyle, J. M., Lampkin, C., and Schulman, R. 2007 motor vehicle occupant safety survey. Volume 2, Seat belt report. United States. National Highway Traffic Safety Administration; 2008 Jul 1.
9. Cameron, G. D., and Duncan, G. I. PARAMICS—Parallel microscopic simulation of road traffic. The Journal of Supercomputing. 1996 Mar;10(1):25-53.
10. Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. ACM computing surveys (CSUR). 2009 Jul 30;41(3):1-58.
11. Chin, H.C., and Quek, S.T. Measurement of traffic conflicts. Safety Science. 1997 Aug 1;26(3):169-85.
12. Cooper, D.F., and Ferguson, N. Traffic studies at T-Junctions. 2. A conflict simulation Record. Traffic Engineering & Control. 1976 Jul;17(Analytic).
13. Das, S., and Maurya, A.K. Defining Time-to-Collision Thresholds by the Type of Lead Vehicle in Non-Lane-Based Traffic Environments. IEEE Transactions on Intelligent Transportation Systems. 2019 Oct 14;21(12):4972-82.
14. De Vlieger, I., De Keukeleere, D., and Kretzschmar, J. G. Environmental effects of driving behaviour and congestion related to passenger cars. Atmospheric Environment. 2000 Jan 1;34(27):4649-55.
15. Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., and Hankey, J. Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences. 2016 Mar 8;113(10):2636-41.
16. Eckert, A., Hartmann, B., Sevenich, M., and Rieth, P. Emergency steer & brake assist: a systematic approach for system integration of two complementary driver assistance systems. In22nd International Technical Conference on the Enhanced Safety of Vehicles (ESV) 2011 Jun 13 (pp. 13-16).
17. Ellison, A. B., and Greaves, S. Driver characteristics and speeding behaviour. In Proceedings of the 33rd Australasian Transport Research Forum (ATRF'10) 2010 Sep.
18. Ericsson, E. Variability in urban driving patterns. Transportation Research Part D: Transport and Environment. 2000 Sep 1;5(5):337-54.
19. Fancher, P. Intelligent cruise control field operational test. Final report. Volume I: Technical report. 1998.
20. Farah, H., Bekhor, S., Polus, A. Risk evaluation by modeling of passing behavior on two-lane rural highways. Accident Analysis & Prevention. 2009 Jul 1;41(4):887-94.

21. Fazio, J., Holden, J., and Rouphail NM. Use of freeway conflict rates as an alternative to crash rates in weaving section safety analyses. Transportation Research Record. 1993;1(401):61.
22. Fellendorf, M., and Vortisch, P. Microscopic traffic flow simulator VISSIM. In Fundamentals of traffic simulation 2010 (pp. 63-93). Springer, New York, NY.
23. Gettman, D., Pu, L., Sayed, T., Shelby, S. G., and Energy S. Surrogate safety assessment model and validation. Turner-Fairbank Highway Research Center; 2008 Jun 1.
24. Grime, G. Traffic and road safety research at the Road Research Laboratory, England. InHighway Research Board Proceedings 1952 (Vol. 31).
25. Halati, A., Lieu, H., and Walker, S. CORSIM-corridor traffic simulation model. In Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and Opportunities Urban Transportation Division, ASCE; Highway Division, ASCE; Federal Highway Administration, USDOT; and National Highway Traffic Safety Administration, USDOT. 1997.
26. Han, I., and Yang, K. S. Characteristic analysis for cognition of dangerous driving using automobile black boxes. International journal of automotive technology. 2009 Oct;10(5):597-605.
27. Hanowsk, R. J., Hickman, J. S., Wierwille, W. W., and Keisler, A. A descriptive analysis of light vehicle–heavy vehicle interactions using in situ driving data. Accident Analysis & Prevention. 2007 Jan 1;39(1):169-79.
28. He, X., Liu, Y., Lv, C., Ji, X., and Liu, Y. Emergency steering control of autonomous vehicle for collision avoidance and stabilization. Vehicle system dynamics. 2019 Aug 3;57(8):1163-87.
29. Henclewood, D., Abramovich, M., and Yelchuru, B. Safety pilot model deployment–one-day sample data environment data handbook. Research and Technology Innovation Administration. Research and Technology Innovation Administration, US Department of Transportation, McLean, VA. 2014 Jul.
30. Hidas, P. Modelling vehicle interactions in microscopic simulation of merging and weaving. Transportation Research Part C: Emerging Technologies. 2005 Feb 1;13(1):37-62.
31. Hu, L., Ou, J., Huang, J., Chen, Y., and Cao, D. A review of research on traffic conflicts based on intelligent vehicles. IEEE Access. 2020 Jan 29; 8:24471-83.
32. Huguenin, F., Torday, A., and Dumont, A. Evaluation of traffic safety using microsimulation. InProceedings of the 5th Swiss Transport Research Conference–STRC, Ascona, Swiss 2005 Mar 9.
33. Igarashi, K., Miyajima, C., Itou, K., Takeda, K., Itakura, F., and Abu,t H. Biometric identification using driving behavioral signals. In2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763) 2004 Jun 27 (Vol. 1, pp. 65-68). IEEE.
34. Jafari, M. Traffic safety measures using multiple streams real time data. Rutgers University. Center for Advanced Infrastructure & Transportation; 2017 Jan 4.
35. Jain, A., Bansal, R., Kumar, A., and Singh, K.D. A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. International Journal of Applied and Basic Medical Research. 2015 May;5(2):124.
36. Janai, J., Güney, F., Behl, A., Geiger A. Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision. 2020 Jul 5;12(1–3):1-308.
37. Jin, S., Qu, X., and Wang, D. Assessment of expressway traffic safety using Gaussian mixture model based on time to collision. International Journal of Computational Intelligence Systems. 2011 Dec 1;4(6):1122-30.
38. Kim, E., and Choi, E. Estimates of critical values of aggressive acceleration from a viewpoint of fuel consumption and emissions. 2013.
39. Kusano, K. D., Gabler, H., Method for estimating time to collision at braking in real-world, lead vehicle stopped rear-end crashes for use in pre-crash system design. SAE International Journal of Passenger Cars-Mechanical Systems. 2011 Apr 12;4(2011-01-0576):435-43.

40. Lajunen, T., Karola, J., and Summala, H. Speed and acceleration as measures of driving style in young male drivers. Perceptual and motor skills. 1997 Aug;85(1):3-16.
41. Langari, R., and Won, J. S. Intelligent energy management agent for a parallel hybrid vehicle-part I: system architecture and design of the driving situation identification process. IEEE transactions on vehicular technology. 2005 May 23;54(3):925-34.
42. Li, Y., Li, Z., Wang, H., Wang, W., and Xing, L. Evaluating the safety impact of adaptive cruise control in traffic oscillations on freeways. Accident Analysis & Prevention. 2017 Jul 1; 104:137-45.
43. Liu, J., and Khattak, A. J. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. Transportation research part C: emerging technologies. 2016 Jul 1; 68:83-100.
44. Liu, J., Wang, X., and Khattak, A. Generating real-time driving volatility information. In2014 World Congress on Intelligent Transport Systems 2014.
45. Mahmud, S. S., Ferreira, L., Hoque, M.S., and Tavassoli, A. Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS research. 2017 Dec 1;41(4):153-63.
46. Martinez, C. M., Heucke, M., Wang, F. Y., Gao, B., and Cao, D. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. IEEE Transactions on Intelligent Transportation Systems. 2017 Aug 4;19(3):666-76.
47. Meng, Q., and Qu, X. Estimation of rear-end vehicle crash frequencies in urban road tunnels. Accident Analysis & Prevention. 2012 Sep 1; 48:254-63.
48. Miyaji, M., Danno, M., and Oguri, K. Analysis of driver behavior based on traffic incidents for driver monitor systems. In2008 IEEE Intelligent Vehicles Symposium 2008 Jun 4 (pp. 930-935). IEEE.
49. Murphey, Y. L., Milton, R., and Kiliaris, L. Driver's style classification using jerk analysis. In2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems 2009 Mar 30 (pp. 23-28). IEEE.
50. National Highway Traffic Safety Administration. Early estimate of motor vehicle traffic fatalities in 2020 (Crash Stats Brief Statistical Summary. Report No. DOT HS 813 115). Tech. rep. National Highway Traffic Safety Administration, May 2020.
51. Paul, M., and Ghosh, I. Post encroachment time threshold identification for right-turn related crashes at unsignalized intersections on intercity highways under mixed traffic. International journal of injury control and safety promotion. 2020 Apr 2;27(2):121-35.
52. Perkins, S. R., Harris, J. L. Traffic conflict characteristics-accident potential at intersections. Highway Research Record. 1968(225).
53. Qu, X., Yang, Y., Liu, Z., Jin, S., and Weng, J. Potential crash risks of expressway on-ramps and off-ramps: a case study in Beijing, China. Safety science. 2014 Dec 1; 70:58-62.
54. Reason, J. Human error. Cambridge university press; 1990 Oct 26.
55. Richard, C. et al. Countermeasures That Work: A Highway Safety Countermeasure Guide for State Highway Safety Offices, 2017. Tech. rep. United States. Department of Transportation. National Highway Traffic Safety, 2018.
56. Saccomanno, F. F., Cunto, F., Guido, G., and Vitale, A. Comparing safety at signalized intersections and roundabouts using simulated rear-end conflicts. Transportation Research Record. 2008 Jan;2078(1):90-5.
57. Saunier, N., Sayed, T. A feature-based tracking algorithm for vehicles in intersections. In The 3rd Canadian Conference on Computer and Robot Vision (CRV'06) 2006 Jun 7 (pp. 59-59). IEEE.
58. Sayed, T., Brown, G., Navin, F. Simulation of traffic conflicts at unsignalized intersections with TSC-Sim. Accident Analysis & Prevention. 1994 Oct 1;26(5):593-607
59. Sens, M.J., Cheng, P. H., Wiechel, J. F., and Guenther, D. A. Perception/reaction time values for accident reconstruction. SAE Technical Paper; 1989 Feb 1.

60. Shin, D., and Yi, K. Human factor considered risk assessment of automated vehicle using vehicle to vehicle wireless communication. International Journal of Automotive Engineering. 2018;9(2):56-63.
61. Singh, D., and Reddy, C. K. A survey on platforms for big data analytics. Journal of big data. 2015 Dec;2(1):1-20.
62. Sundt, T. M., Sharbrough, F. W., Anderson, R. E., and Michenfelder, J. D. Cerebral blood flow measurements and electroencephalograms during carotid endarterectomy. Journal of neurosurgery. 1974 Sep 1;41(3):310-20.
63. Tageldin, A., Sayed, T.,and Shaaban, K. Comparison of time-proximity and evasive action conflict measures: Case studies from five cities. Transportation research record. 2017;2661(1):19-29.
64. Tarko, A. P. Surrogate measures of safety. In safe mobility: challenges, methodology and solutions 2018 Apr 18. Emerald Publishing Limited.
65. Tarko, A. P., and Songchitruksa, P. Estimating the frequency of crashes as extreme traffic events. In84th Annual Meeting of the Transportation Research Board 2005.
66. Treat, J. R. A study of precrash factors involved in traffic accidents. HSRI Research review. 1980 May.
67. Van Der Horst, R., and Hogema J. Time-to-collision and collision avoidance systems.
68. Vasconcelos, L., Neto, L., Seco, Á. M., and Silva, A. B. Validation of the surrogate safety assessment model for assessment of intersection safety. Transportation Research Record. 2014 Jan;2432(1):1-9.
69. Wang, J., Zhang, L., Huang, Y., and Zhao, J. Safety of autonomous vehicles. Journal of advanced transportation. 2020 Oct 6;2020.
70. Wang, W. H., Cao, Q., Ikeuchi, K., and Bubb, H. Reliability and safety analysis methodology for identification of drivers' erroneous actions. International Journal of Automotive Technology. 2010 Dec;11(6):873-81.
71. Wang, W., Mao, Y., Jin, J., Wang, X., Guo, H., Ren, X., and Ikeuch,i K. Driver's various information process and multi-ruled decision-making mechanism: a fundamental of intelligent driving shaping model. International Journal of Computational Intelligence Systems. 2011 May 1;4(3):297-305.
72. Wang, X., Khattak, A. J., Liu, J., Masghati-Amoli, G., and Son, S. What is the level of volatility in instantaneous driving decisions? Transportation Research Part C: Emerging Technologies. 2015 Sep 1; 58:413-27.
73. Williams, M. J. Validity of the traffic conflicts technique. Accident Analysis & Prevention. 1981 Jun 1;13(2):133-45.
74. Wilson, F. N., Johnston, F. D., Macleod, A. G., and Barker, P.S. Electrocardiograms that represent the potential variations of a single electrode. American Heart Journal. 1934 Apr 1;9(4):447-58.
75. Zhang, E. and Zhang, Y. Average Precision. In: Encyclopedia of Database Systems. Ed. by Ling Liu and M. Tamer ÖZSU. Boston, MA: Springer US, 2009, pp. 192 193. isbn: 978-0-387-39940-9. doi:10.1007/978-0-387-39940-9 482. url: https://doi.org/10.1007/978-0-387-39940-9 482.
76. Zheng, L., Ismail, K., and Meng, X. Traffic conflict techniques for road safety analysis: open questions and some insights. Canadian journal of civil engineering. 2014;41(7):633-41.

10.0 APPENDICES

10.1 Appendix A – Acronyms, abbreviations, etc.

AASHTO -- American Association of State Highway and Transportation Officials

ACC -- adaptive cruise control

ADAS -- advanced driver assistance systems

AEB -- automatic emergency braking

APT -- audio perception time

AV -- autonomous vehicle

AV -- autonomous vehicle

BSM -- basic safety message

CI -- crash index

CIM -- conflict detection model

CSV -- comma-separated values

CV -- connected vehicle

DA -- driving anomaly

DVU -- driver-vehicle unit

ESA -- emergency steering assistance

FCW -- forward collision warning

FHWA -- Federal Highway Administration

GPS -- Global Positioning System

ITS -- intelligent transportation system

ITS -- intelligent transportation system

KPI -- key performance indicator

LDW -- lane departure warning

MTC -- margin to collision

NDS -- Naturalistic Driving Study

NHTSA -- National Highway Traffic Safety Administration

NHTSA -- National Highway Traffic Safety Administration

OBU -- on-board unit

PAID -- Pay-as-you-drive

PD -- pedestrian detection

PET -- post-encroachment time

PRT -- perception-reaction time

PSD -- proportion of stopping distance

SCA -- scale for criticality assessment

SHRP II -- the Strategic Highway Research Program

SPMD -- Safety Pilot Model Deployment (SPMD)

SSAM -- surrogate safety assessment model

SSM -- surrogate safety measure

TCT -- traffic conflict technique

TMC -- traffic management center

TMC -- traffic management center

TS -- time series

TTE -- time to evade

US DOT -- United States Department of Transportation

V2V -- vehicle-to-vehicle

VPT -- visual perception time

10.2 Appendix B – Associated websites, data, etc., produced

https://insight.shrp2nds.us/login/auth

https://www.its.dot.gov/pilots/

10.3 Appendix C – Summary of Accomplishments

Date	**Type of Accomplishment** *(select from drop down list)*	**Detailed Description** *Provide name of person, name of event, name of award, title of presentation, location and any links to announcements if available Please attach any abstracts, summaries, high quality photos, or additional details as an appendix.*
09/07/2020	Conference Paper	We submitted the abstract of a paper titled "Anomaly Detection on Driving Status Using Basic Safety Messages in Connected Vehicle Environment" to the International Conference on Transportation and Development (ICTD) 2021.
11/31/2020	Conference Paper	We submitted the full paper titled "Anomaly Detection on Driving Status Using Basic Safety Messages in Connected Vehicle Environment" to the International Conference on Transportation and Development (ICTD) 2021.
6/8/2021	Conference Presentation	We presented online our paper titled "Anomaly Detection on Driving Status Using Basic Safety Messages in Connected Vehicle Environment "on the International Conference on Transportation and Development (ICTD) 2021.
10/26/2021	Publication	We submitted the abstract of a paper titled "Evaluation and Sensitivity Analysis of Unsupervised Driving Anomaly Detection" to vehicles of Multidisciplinary Digital Publishing Institute
11/9/2021	Conference Paper	We submitted the abstract of a paper titled "Conflict Identification Using Speed Distance Profile on Basic Safety Messages" to the International Conference on Transportation and Development (ICTD) 2022.
01/24/2022	Conference Paper	We submitted the full paper titled "Conflict Identification Using Speed Distance Profile on Basic Safety Messages" to the International Conference on Transportation and Development (ICTD) 2022. The conference accepted our paper in the conference program of ICTD 2022.

Abstract of the paper submitted to the International Conference on Transportation and Development (ICTD) 2021. https://www.asce-ictd.org/

Anomaly Detection on Driving Status Using Basic Safety Messages in Connected Vehicle Environment

Di Wu, P.E.[1], Shuang Z. Tu, Ph.D.[2], and Robert W. Whalin, Ph.D., P.E., D.CE[3]

1Computational and Data Enabled Science and Engineering Program, Jackson State University, Email: dwzoon@gmail.com
2Department of Electrical and Computer Engineering and Computer Science, Jackson State University, Email: shuang.z.tu@jsums.edu
3Department of Civil and Environmental Engineering and Industrial Systems and Technology, Jackson State University, Email: robert.w.whalin@jsums.edu

ABSTRACT

As human factors contribute to more than 90% crashes, abnormal driving behavior has been intensively studied to improve traffic safety. Basic Safety Messages (BSMs) transmitted between connected vehicles (CVs) are time series data with high-cardinality. Real-time anomaly detection on driving status using BSMs is important but neglected. This paper is to explore what information imbedded in BSMs needs to be stored, how to extract and process it for real-time safety diagnosis. We propose a real-time multi-dimensional driving anomaly detection (DAD) system on individual level specifically configured for BSMs. The architecture of the proposed system is composed of two parts: in cloudthe system collects and stores BSMs of the vehicles it covers for a short period of time, and determines in batch mode the thresholds of the selected key performance indicators (KPIs) representing normal status for each vehicle, and broadcast the thresholds through the BSMs; in the in-vehicle device, as new BSMs streaming in, the device compares them with the received thresholds and determines the outliers; if detected, the outliers will be analyzed and determines if the vehicle warrants an anomaly flag; finally, the system will determine the impact factors to update thresholds according to the significances of the outliers. This system of AD is a crucial component of our pipeline for the automatic safety diagnosing system in the CV environment. This research is sponsored by the Southeastern Transportation Research, Innovation, Development and Education Center (STRIDE).

Keywords: anomaly detection, BSM, connected vehicle, safety, algorithm

Abstract of the paper submitted to Vehicles of Multidisciplinary Digital Publishing Institute.
https://www.mdpi.com/journal/vehicles

Evaluation and Sensitivity Analysis of Unsupervised Driving Anomaly Detection

Di Wu 1, † , ‡ , Shuang Tu 2, † , ‡ * and Robert Whalin 3, † , ‡

1 Computational and Data Enabled Science and Engineering Program, Jackson State University; dwzoon@gmail.com
2 Department of Electrical and Computer Engineering and Computer Science, Jackson State University; shuang.z.tu@jsums.edu
3 Department of Civil Engineering, Jackson State University; robert.w.whalin@jsums.edu
* Correspondence: shuang.z.tu@jsums.edu;
First Note: Current address:1400 John R. Lynch St, Jackson, MS 39217
Second Note: These authors contributed equally to this work.

Abstract: Evaluation on unsupervised anomaly detection is a constant challenge, so is the application of machine learning to practical engineering problems. Nevertheless, modeling driving status using basic safety messages (BSMs) of connected vehicles at the individual level is an unavoidable task for the automation of our traffic safety diagnosis system. This paper records our efforts in model evaluation and sensitivity analysis in building the driving anomaly detection system. In the model evaluation, the measurement of average precision was utilized and the model was validated. In the sensitivity analysis, a number of key performance indicators (KPIs) were set up based on the model responding to the changing values of the KPIs. Our model is a combination of machine learning (ML) and engineering modeling. As a data science technique ML has become a hot topic in many domains since the deluge of big data, but when it comes to engineering, where accuracy and proving is emphasized, ML is not as successfully applied. One of the reasons is that ML is known as a "black box", neither convincing nor easily assortative with domain knowledge. Through this research, we found it a practical way to build the ML model for engineering problem, basing on domain knowledge, and use measures in information system to validate the model using underling assumptions. Given the lack of information on the distributions followed by the model parameters, sensitivity analysis is a practical tool to set up parameters at reasonable ranges. This paper is a part of our ongoing project of Anomaly Detection on Driving Status Using Basic Safety Messages in the Connected Vehicle (CV) Environment.

Keywords: model evaluation; unsupervised; machine learning; driving status; outlier detection; sensitivity analysis; traffic safety

Abstract of the paper submitted to the International Conference on Transportation and Development (ICTD) 2022. https://www.asce-ictd.org/

Conflict Identification Using Speed Distance Profile on Basic Safety Messages

Di Wu [1], Li Zhang[2], Robert Whalin[3] and Shuang Tu[4]*

1Computational and Data Enabled Science and Engineering Program, Jackson State University; e-mail: di.wu@students.jsums.edu
2Department of Civil and Environmental Engineering, Mississippi State University; e-mail: li.zhang@ngsim.com
3Department of Civil and Environmental Engineering and Industrial Systems and Technology, Jackson State University; e-mail: robert.w.whalin@jsums.edu
4 Department of Electrical and Computer Engineering and Computer Science, Jackson State University; e-mail: shuang.z.tu@jsums.edu
*corresponding author

ABSTRACT

A major obstacle to the prevalence of autonomous vehicles (AVs) is safety. The safety of AVs relies largely on the surveillance systems and motion detection in the ego vehicle. The real-world detection is affected by many factors such as weather, interference and sensibilities. This safety issue can be mitigated by sending near crash warnings to the drivers in the connected vehicle (CV) environment, through analyzing the trajectories of the vehicles embedded in the basic safety massages (BSMs). In the literature, the research using trajectories to detect the potential crashes utilized the trajectory data that were collected on the scene. However, in the case of using BSMs, as the effective transmission distance of the vehicle-to-vehicle (V2V) BSMs is limited, there might be no sufficient time to perform a chain of tasks to avoid a crash after the vehicles come into the effective V2V range, including collecting the data, training the model, analyzing the data, broadcasting the alarm and for the driver to perceive the alarm and take actions. In order to leave adequate time to the drivers, we separate the process into two steps: step one is the driving anomaly detection including the threshold generation in the cloudusing historical BSMs and the driving abnormal detection (DAD) in the in-vehicle subsystem using real time BSMs; step two is to detect the conflicts using the real time BSMs and the results from the first step. In our system, a near crash warning will be generated if a conflict is detected between the pair of ego CV and a nearby CV when any CV in the pair is under abnormal driving status. This paper focuses on the second step, in which several issues are solved: a) redefining the conflict. Conflict is a key concept of surrogate safety analysis, which was originally designed for traffic simulation data. We tailor it to fit our system; b) developing the mathematical algorithms to identify the conflicts; c) implementing the algorithms in the in-vehicle subsystem. The algorithm is tested on the SHRP2 crash data which contain similar features of the BSMs. This paper is part of our ongoing project of Anomaly Detection on Driving Status Using Basic Safety Messages in the Connected Vehicle (CV) Environment (ASDSCE).