Accident Analysis and Prevention 146 (2020) 105711

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles

Yanqi Lian, Guoqing Zhang, Jaeyoung Lee*, Helai Huang

School of Traffic & Transportation Engineering, Central South University, Changsha, Hunan, People's Republic of China

ARTICLE INFO

Keywords:
Big Data
Data mining
Machine learning
Traffic safety
Intelligent transportation systems
Connected and automated vehicles

ABSTRACT

The era of Big Data has arrived. Recently, under the environment of intelligent transportation systems (ITS) and connected/automated vehicles (CAV), Big Data has been applied in various fields in transportation including traffic safety. In this study, we review recent research studies that employed Big Data to analyze traffic safety under the environment of ITS and CAV. The particular topics include crash detection or prediction, discovery of contributing factors to crashes, driving behavior analysis, crash hotspot identification, etc. From the reviewed studies, employing advanced analytics for Big Data has a great potential for understanding and enhancing traffic safety. Big Data application in traffic safety integrates and processes massive multi-source data, breaks through the limitations of the traditional data analytics, and discovers and solves the problems, which cannot be solved by the traditional safety analytics. Lastly, suggestions are provided for future Big Data safety analytics under the environment of ITS and CAV.

1. Introduction

In the recent decade, the volume, the variety, and the availability of data keep growing rapidly, which is the result of the fast advancement of various technologies. According to Sagiroglu and Sinanc (2013), humanity had created only five Exabytes (5×10^6 Terabytes) of data until 2003. Surprisingly, the same amount of data was generated within only two days in 2012 (McAfee et al., 2012). International Data Corporation (IDC) predicted that the amount of the global datasphere will increase from 33 Zettabytes (33×10^9 Terabytes) in 2018 to 175 Zettabytes (175×10^9 Terabytes) by 2025 (Reinsel et al., 2018). The rapid growth of data attracts a growing body of researchers to apply Big Data to investigate an array of studies, which offers opportunities to explore the knowledge that was not able to be achieved in the past (Chen et al., 2014).

Most research studies on traffic safety have been conducted using certain data because outcomes verified by data are more convincing. Previously, traffic safety analysis mostly used manually collected data such as police crash report data (hand written but not digitalized) along with estimated aggregated static traffic volume (e.g., AADT). Some safety analyses were not able to be conducted due to the unavailability of data, such as dynamic traffic flow data and driving behavior data. Recently, extended installations of detectors, sensors, and other sources have led to the rapid popularization of ITS (intelligent transportation systems) and the emergence of CAV (connected/automated vehicles), accompanied by the skyrocketed amount of data regarding human, vehicles, roads, and environments. Such advancements of data collection have resulted in significant changes in safety analyses. The recent safety studies have used more dynamic (or real-time) and/or high-resolution data owing to the Big Data. Nevertheless, traditional analytics are not suitable for such Big Data while advanced Big Data analytics are because they have a strong data processing capability. In traffic safety, Big Data analytics have multiple advantages particularly in some perspectives. First, researchers analyze when and where a crash would occur combining crash cases and non-crash cases. Meanwhile, a traffic crash occurs due to numerous factors including human, environment, design, vehicle, socio-economic characteristics. Therefore, data used in safety analyses are large, various and heterogeneous, and the underlying relationships of these data can be uncovered by Big Data analytics. Second, the availability of a large amount of data has led to the revolution of ITS from conventional technology-driven system to data-driven intelligent transportation system (Zhang et al., 2011). Under the environment of ITS, to provide a proactive countermeasure for preventing a traffic crash, real-time analyses and real-time traffic management of massive traffic data are essential and Big Data analytics are the main way to achieve it. Third, Big Data are produced since vehicle-to-everything (V2X) communication-based CAV monitor drivers, other road users, and environments in real-time. And Big data analytics will

* Corresponding author.
E-mail address: jaeyoung@knights.ucf.edu (J. Lee).

https://doi.org/10.1016/j.aap.2020.105711
Received 15 April 2020; Received in revised form 25 July 2020; Accepted 29 July 2020
Available online 04 September 2020

have considerable benefits in analyzing traffic safety for CAV due to its rapid processing capability. Therefore, Big Data safety analytics in ITS and CAV research have attracted a remarkable amount of research interest, which include but not be limited to: traffic crash prediction, real-time safety analysis, damage severity modeling and so on.

Although many Big Data safety analytics for ITS and CAV have been applied, some challenges still remain. There have been very few systematic reviews of Big Data analytics specifically addressed safety under the environment of ITS and CAV. Therefore, we aim to suggest a systematic review by summarizing data, models and techniques, topics, and discussing the challenges and the future of Big Data safety applications in ITS and CAV research.

2. Literature selection

2.1. Criteria

The criteria for material collection are shown below:

1 Before searching for papers, we identified keywords related to our research topic. The keywords can be divided into two classes: keywords related to Big Data analytics including "Big Data, machine learning, deep learning, and data mining" and keywords related to traffic safety including "safety, crash, accident, collision, incident, and severity". The paper was selected if any possible combination between two types of keywords appears in the title or keywords of the paper.
2 We searched for papers in scientific journals and conferences in 2013 and later.
3 We focused on papers that conducted quantitative analyses of traffic safety with Big Data. Papers about qualitative analyses of Big Data applications in safety and papers with pure mathematical modeling but no data analysis were not considered.
4 Papers whose citations are at least five were of our interest.

We searched for studies based on the above criteria via the article databases including Google Scholar and Web of Science.

2.2. Safety-related Big Data

Big Data was defined from three aspects including volume, velocity, and variety as early as 2001 (Laney, 2001). As data increases, the definition of Big Data changes. In 2011, IDC presented one new characteristic of Big Data—value (Gantz and Reinsel, 2012). Then, White (2012) put forward another new aspect of Big Data—veracity, which changes the definition of Big Data from "4V" to "5V". For accident investigation and analysis, Huang et al. (2018) defined safety-related Big Data from five aspects: volume (large magnitude); velocity (fast speed of data generation); variety (structural heterogeneity); veracity (unknown accuracy); value (safety information contained in safety-related Big Data).

Big Data contains rich information and a complicated structure, which results in a long time to clean and process the data. For the purpose of the current study, the data used in the papers are considered as Big Data if the data, which contain multiple data types are combined and data preprocessing is conducted via a huge proportion of computation. Moreover, according to Ghofrani et al. (2018), the data sets with over 50,000 variables or 50,000 observations are considered as Big Data. Therefore, after collecting the papers based on the criteria in Section 2.1, the papers with data cannot meet the requirements are to be deleted. In the end, we got a total of 57 articles.

3. Big Data safety analytics in ITS and CAV research

The widespread use of various sensors and detectors has led to the fast advancement of ITS and CAV, accompanied by the advent of vast and multiple data. With continuously increasing data, researchers have shifted their attention to analyzing traffic safety under the environment of ITS and CAV by using Big Data analytics. We classified the 57 studies of the topic into two categories: Big Data safety analytics in ITS research (53), and Big Data safety analytics in CAV research (4). The numerical values in parentheses represent the number of papers investigating that category.

3.1. Big Data safety analytics in ITS research

With the rapid popularization of ITS, Big Data in the field of traffic are continuously collected from multiple sources over vast geographic scale. To make the transportation system safer, more efficient and reliable, these data are leveraged to conduct various research in the field of traffic safety and provide new insights to safety analysis. Big Data safety analytics in ITS research are further classified into four categories, including crash detection or prediction, discovery of contributing factors to traffic crashes, driving behavior analysis, and crash hotspot identification. Moreover, it should be mentioned that many papers focus on more than one aspect.

3.1.1. Crash detection or prediction

In ITS research, considerable efforts have been devoted to predicting a traffic crash prior to the occurrence or detecting a crash after the occurrence. The former enables us to prevent the crash or decrease crash severity by providing appropriate intelligent real-time countermeasures while the latter can assist in taking the necessary precautions as early as possible. Theofilatos and Yannis (2014) reviewed the effects of traffic and weather characteristics on road safety by dividing the papers into three classes: crash occurrence likelihood, crash frequency, and crash injury severity. The same categories were adopted by us for traffic crash prediction.

Considerable efforts have been devoted to developing a model or system to detect a crash after the occurrence. Ozbayoglu et al. (2016) proposed a real-time crash detection system to monitor real-time traffic flow and detect the occurrence of a crash by using three separate computational intelligence techniques including Nearest Neighbor, Regression Tree, and Feedforward Neural Network. To detect an incident in large urban networks, the spatial correlation is a significant factor to be considered, Zhu et al. (2018) presented a network-level incident detection model based on Convolutional Neural Network (CNN) and solved high spatial connections of urban network with a connectivity matrix. The authors found that the spatial correlations may exist between neighboring corridors and CNN can learn the complex correlations correctly. Zhang et al. (2018d) employed a deep learning approach for detecting a traffic crash with social media data and validated the credibility of crash-related massages as well as the time and location effectiveness via crash log and traffic data. The authors found that the social network messages are more probable to be a viable supplement rather than a replacement to the existing detection method because such messages can capture only few events that seldom arouse public attention. By comparing environmental conditions and vehicle kinematics signatures of near-crash events with data collected from the Second Strategic Highway Research Program (SHRP2), Ali et al. (2019) used both parametric and non-parametric methods to detect near-crashes on freeways with a time chunking technique monitoring changes in vehicle kinematics on a timescale.

Crash occurrence likelihood has been investigated by a surge of papers. Ahmed and Abdel-Aty (2013) proposed a real-time risk assessment framework to distinguish crash and no-crash and compared the prediction accuracy of traffic data collected from different sources at the same location. By calculating the entry and departure time to and from each segment of 2 miles, Al Najada and Mahgoub (2016a) predicted the crashes early in a spatiotemporal scenario with considering primary and secondary crashes as well as the availability of infrastructure. To predict highway traffic crash, Park et al. (2016) proposed a distributed Hadoop framework with a synthetic minority over-sampling technique which was used to resolve the problem of data imbalance. With the same way to tackle imbalanced dataset issues as the last study, You et al. (2017) estimated a novel approach based on Support Vector Machine (SVM) to predict crash likelihood with traffic data as well as weather data and found the model with weather data outperforms that without weather data. Yuan et al. (2017) explored effective techniques to better predict whether a crash will occur for each road segment in each hour. In this study, eigen-analysis of the road network was leveraged to incorporate SpatialGraph features to address the spatial heterogeneity challenge, which could significantly enhance the performance of all the models.

Several studies developed models to predict crash frequency at given time and locations. Considering the effect of nearby facilities naturally, Zhang et al. (2018b) presented an auto-searching algorithm to predict annual crashes of any facility within the square-shaped region and provided a built-in tool that allows user map displays of annual crashes. To predict crash number across crash severities, Dong et al. (2018) proposed an innovative deep learning technique which consists of two modules including an unsupervised feature learning module to extract feature representations and a supervised fine-tuning module to predict crashes. Several studies investigated crash frequency prediction in real-time or at high resolution (Ren et al., 2018; Yuan et al., 2018; Cai et al., 2019). Ren et al. (2018) proposed a deep learning model based on Long Short-Term Memory (LSTM) to predict the citywide traffic crash frequency with temporal resolution of one hour and spatial resolution of 1 km × 1 km uniform grids. The findings revealed that the traffic crash frequency has strong periodical temporal patterns and regional spatial correlation. Yuan et al. (2018) proposed a heterogeneity convolutional LSTM framework which tackled the spatiotemporal heterogeneity challenge with time dimension of day and spatial dimension of 5 km × 5 km grid. The findings indicated that daily average is generally good at predicting long-term traffic crashes with low average error. Cai et al. (2019) introduced a deep learning architecture based on CNN to predict crashes for transportation safety planning with high-resolution data. The authors divided each grid into 10,000 (100 × 100) cells and both the width and height of each cell are 0.03 mile. Then, the proposed method using high-resolution data based on cells was compared with three conventional models with low-resolution data based on grids. The results demonstrated that the local interactions between cells could be hierarchically captured by the proposed model and the deep learning method with high-resolution data could significantly improve the crash prediction accuracy.

A host of studies classified crashes into several levels based on severity to predict crash injury severities with the consideration of the features related to crash. For instance, to predict crash severity, various machine learning methods were implemented, compared, and evaluated (Effati et al., 2015; Babič and Zuskáčová, 2016). Similarly, the performances of various machine learning methods and statistical methods with distinct modeling logic used to predict crash severities were compared (Delen et al., 2017; Iranitalab and Khattak, 2017; Taamneh et al., 2017; Zhang et al., 2018a). The results demonstrated that the predicting accuracy of the machine learning methods is higher than that of the statistical methods. In contrast, Castro and Kim (2016) found the accuracy of the statistical method is higher than machine learning methods. Meanwhile, novel traffic crash's severity prediction methods were established and compared with statistical models and machine learning models (Alkheder et al., 2017; Zheng et al., 2019). The findings of each study showed that the proposed model outperforms the baseline models. In addition, a deep learning framework, named DeepScooter was developed to predict motorcycle-involved crash severities by analyzing only at-fault motorcycle rider crashes (Das et al., 2018).

Some articles worked on multiple aspects simultaneously. In some literature (Chen et al., 2016; Ren et al., 2017; Bao et al., 2019), the sum of severity level of all crashes in a specific grid was defined as the crash risk of that grid and data was aggregated into given time and space to tackle spatiotemporal heterogeneity. In detail, by meshing location into approximate 0.5 km × 0.5 km grids with one hour as the time interval to analyze spatial and temporal varied data, Chen et al. (2016) estimated traffic crash risk in a city or national scale with a deep learning approach. Ren et al. (2017) collected big and heterogeneous data related to traffic crashes and predicted short-term traffic crash risks based on the same method and spatiotemporal resolution with another study of them (Ren et al., 2018). The authors revealed that the predictive power of the daily periodic feature of traffic pattern is better compared to that of the weekly feature and the short-term feature for traffic crash risk prediction. Bao et al. (2019) developed a deep learning model based on LSTM network with incorporating multi-source datasets to predict citywide short-term crash risk by conducting a total of nine prediction tasks, including weekly, daily and hourly models with 8 × 3,15 × 5 and 30 × 10 grids, respectively. The authors indicated that the increase of the spatiotemporal resolution of prediction task would worsen the prediction performance of the proposed model and the proposed models incorporating multiple source data achieve the best performance over baseline methods.

3.1.2. *Discovery of contributing factors to traffic crashes*

Crashes under different circumstances are often regarded as random events which are determined by numerous factors including drivers, vehicles, roadway, traffic and environment condition, etc. To prevent potential crashes, decrease the severe outcomes from crashes, and save lives and cost in advance, an increasing number of researchers have attempted to uncover the root causes of traffic crashes. The categories used by Theofilatos and Yannis (2014) were also adopted in this section, including factors for the crash occurrence, factors for crash frequency, and factors for crash injury severity.

Investigating the contributing factors to crash occurrence harnessing Big Data has attracted a remarkable amount of research interest. Shi and Abdel-Aty (2015) leveraged both Random Forests (RF) and Bayesian statistics method to investigate real-time safety with one-minute interval traffic data and verified that congestion has a significant impact on rear-end crashes. The crash data was first clustered and association rule mining was further conducted to identify various circumstances related to the occurrence of a crash (Kumar and Toshniwal, 2015a, b; Kumar and Toshniwal, 2016). To prevent traffic casualties and congestion, Al Najada and Mahgoub (2016b) explored the main causes of traffic crashes and eventually found age band of drivers plays an important role in causing traffic crashes. You et al. (2017) employed RF to select the contributing factors and utilized the mean importance value to evaluate the relative effects of the factors. The results showed that the flow of the corresponding segment as well as the weather condition can affect traffic safety. Prati et al. (2017) leveraged Decision tree and Bayesian network analysis to uncover the factors related to severities of bicycle crashes and demonstrated that road type, crash type, and type of opponent vehicle are identified as important predictors by two models. Besides, Chen et al. (2018b) used unbalanced panel data mixed logit model to investigate the contributing factors to crash likelihood. The findings of the study confirmed that both time-varying factors (e.g., road surface condition and hourly traffic volume) and site-varying factors (e.g., curvature and speed limit) may have a great influence on the crash likelihood. Das et al. (2019) conducted association rule mining to discover the relationships between the crash-related factors and the crash occurrence in the rainy weather with only considering at-fault driver information for the final analysis.

With respect to crash frequency influencing factors, Simandl et al. (2016) aimed at evaluating the impact of selective enforcement on the numbers of citations and crashes before and during selective enforcement and found that citations increased significantly at all identified selective enforcement locations while the number of crashes decreased significantly. With dividing the study area into 1113 Traffic Analysis

Zones (TAZs), Jiang et al. (2016) studied the macro-level factors leading to crash frequency of different injury levels, crash types, and collision types. The results suggested that the two most important factors proactively alleviating traffic safety issues are the distribution of road networks and socioeconomics. Similarly, by aggregating the collected data into 896 TAZs, Bao et al. (2017) classified the twitter-based human activities into seven categories to explore the effects of various contributing factors on crashes. The authors demonstrated that the numbers of eating, recreation, and education related activities significantly affect the pedestrian-involved crashes, while the numbers of eating, shopping, and social related activities significantly affect the vehicle-to-vehicle crashes. Dong et al. (2018) investigated the relationship between the examined factors and the traffic crashes and found that traffic, geometric, pavement and environmental factors have direct effects on traffic crashes across injury severities. El Mazouri et al. (2019) conducted association rule mining to uncover all possible correlations and relationships between victims and traffic crashes. The authors indicated that the excess speed and carelessness of drivers are the main causes of the crashes and the pedestrians are the first vulnerable victims of these crashes.

In terms of the factors related to crash injury severity, Kashani et al. (2014) explored the factors related to the crash severity of motorcycle pillion passengers and found area type, land use, and injured part of the body are the most influential factors. Appling association rule mining approach, Moradkhani et al. (2014) found spatial features and infrastructure play a major role in the crash. Using the same method as the last study, Babič and Zuskáčová (2016) found crashes that happen on Sunday have fatal consequences with the highest probability, despite the count of crashes is least on Sunday. Castro and Kim (2016) focused on the effects of various road-related factors on crash severity and discovered that road type, vehicle maneuver, and light condition have significant effects on crash severity, whereas the age of vehicle and weather conditions have no significant effects. Moreover, Taamneh et al. (2017) extracted the rules generated by the decision tree and the rules induction to understand the main factors related to crash severity and found that age, gender, nationality, year of the crash, casualty status, and collision type have the largest effects on fatal severity. Li et al. (2017) investigated the factors contributing to fatal rates and discovered the effects of the environmental factors (e.g. roadway surface, weather, light condition) are not strong while the effects of human factors (e.g. drunk or not) and collision type are stronger. To explore the influential factors and their marginal effects on truck crash severity, Zheng et al. (2018) developed a Gradient boosting model. The authors found trucking company as well as driver characteristics affect truck crash injury severity significantly and fatal crashes are likely to happen under good weather or good road surface. What's more, the authors also indicated that even though most of the identified contributing factors have significant effects on all four levels of crash severity, their relative importance and marginal effect are all different. In addition, various methods were conducted and compared to predict crash severity and further sensitivity analysis were applied to analyzing variable importance (Effati et al., 2015; Delen et al., 2017; Zhang et al., 2018a).

Some papers focused on the crash-related factors considering other aspect or several aspects. For instance, Wang et al. (2015) explored the relationship among driving risk, driver/vehicle characteristics, and road environment with naturalistic driving data. The authors indicated that the most important factors related to the driving-risk level are brake velocity, triggering factors, potential object type, and potential crash type. Xie et al. (2017) investigated the factors contributing to crash cost which accounts for both crash frequency as well as severity and found many factors (e.g., vehicle-miles-traveled, truck ratio, subway ridership, bus stop density, taxi trip, ratio of commercial area) are positively associated with the cost of the crash involving pedestrian.

3.1.3. Driving behavior analysis

According to Pakgohar et al. (2011), 97.5 percent of crashes are caused by human factors. Especially, aggressive driver behavior, which is broadly defined as any willful driving behaviors with no intention of harming another road user but the disregard of their safety, is a leading cause of traffic crash (Tasca, 2000). It was found that 56 percent of all fatal crashes were occurred due to aggressive driving behaviors, which include but not be limited to: speeding, failure to yield right of way, careless driving, improper turning and so on (American Automobile Association, 2009). Therefore, to improve traffic safety, considerable efforts have been devoted to analyzing driving behavior with the availability of driving behavior data.

St-Aubin et al. (2015) proposed a theoretical and practical framework of a large-scale automated video-based proactive road safety analysis system and applied the system to the cross-sectional analysis of driver behavior with roundabout video data. Al Najada and Mahgoub (2016b) investigated the impact of driver's behavior on causing traffic crashes and found the driver age, gender as well as the crash severity could be obtained from the driver's behavior mining. Risky driving behavior is a leading factor in traffic crashes. With the goal to improve traffic safety, considerable efforts have been devoted to analyzing how driving detection system is used to identify risky driving behaviors. For instance, using vehicle dynamics data collected from a static driving simulator, Tango and Botta (2013) conducted several machine learning techniques to detect driver's visual distraction and pointed out the personalization analysis, with one specific model for each participant. Tran et al. (2018) proposed a driver distraction detection system to detect ten normal as well as distracted driving behaviors and presented a conversational warning system to alert the distracted drivers in real-time by using data from a driving simulator. The results indicated that the similarities of different behaviors such as "hair and makeup" and "talking on the phone-left" would result in misclassifications. Wang et al. (2016) used RF model to detect drowsy behavior with data from the driving simulator and found 20s-size datasets using parameter combination of lateral and longitudinal accelerations are the best inputs for drivers' drowsy behavior detection.

3.1.4. Crash hotspot identification

A crash hotspot is defined as the area where traffic crashes happen with a high probability. To apply the precautions effectively, a key step towards improving traffic safety is crash hotspot identification which is generally conducted based on road safety experts' opinions or via analysis of historical crash data. With massive data available, crash hotspot identification can be implemented with more methods and higher resolution. For reviewed papers on crash hotspot identification, some papers identified crash hotspots with crash data while some papers identified crash hotspots based on other data (e.g. social media data), and crash data were only used to verify the accuracy of identification.

For instance, there are some papers used crash data to identify crash hotspots. Jiang et al. (2016) identified hot zones at macro-level which was defined as the TAZ level and examined the appropriateness of alternative crash risk measures. The authors found that the optimum macro-level crash hot zone identification accuracy could be achieved by measuring crash risk with crashes per square mile as compared to crashes per mile and crashes per million vehicle-miles-traveled. Based on an existing immuno-inspired mechanism, namely SeleSup, Triguero et al. (2017) introduced a Big Data approach for road incident hotspot identification using Apache Spark. Xie et al. (2017) identified crash hotspot involving pedestrians based on the crash cost that accounts for both crash frequency as well as severity and the potential for safety improvement (PSI) that could be obtained by using the actual crash cost minus the cost of "similar" sites. The results indicated that the hotspot identification approach based on crash cost would neglect sites with relatively low crash costs while the method based on PSI has the potential to find such sites.

Meanwhile, several articles used other data to conduct crash hotspot identification and used crash data to verify accuracy. Sinnott and Yin

(2015) proposed a Cloud-based software system to predict and verify crash black spots with social media (Twitter) data and benchmarked the identification accuracy of Twitter with historic crash hotspot information as well as official tweet information. By establishing a visual interaction system, Itoh et al. (2015) explored caution spots with large-scale vehicle recorder data and provided a 3D spatiotemporal visualization space to show caution spots. Moreover, by dividing the area into disjoint cells with the size of 60 m × 60 m, Zhang et al. (2018c) developed a multi-view learning approach for risky traffic location identification in a city and tackled data availability as well as location accuracy by jointly exploring the social and remote sensing data.

3.2. Big Data safety analytics in CAV research

Integrating advanced sensing, communications, and electronic technologies, Connected Vehicles (CVs) that communicate with everything along with Automated Vehicles (AVs) that drive without human input are rapidly emerging. CVs can integrate neighbouring vehicles and facilities into the transportation system and support large-scale ITS applications while AVs can liberate drivers and allow drivers to do other things. The emergence of CAV has the potential to revolutionize the transportation system by increasing capacity, improving safety and reducing congestion, energy consumption, and pollution. With such promising advantages, more and more researchers started to investigate CAV to realize the benefits of them. Meanwhile, Big Data analytics have been applied to safety analysis under the environment of CAV since not only data generated from CAV but also data used to investigate CAV is extensive.

For instance, some papers focused on crash prediction and detection under the environment of CAV (Chen et al., 2018a; Dogru and Subasi, 2018). By exploring the correlation between important influential factors of crashes and the occurring probability of crashes on the Internet of Vehicles, a genetic algorithm optimized Neural Network was established to predict rear-end crashes with training data generated from VISSIM (Chen et al., 2018a). Then, the authors compared three communication methods between two nodes, i.e., unicast, broadcast and multicast, and demonstrated that broadcasting is the only appropriate method of information exchange for the discussed case. Collecting traffic data from sensors installed on roads and sending alert messages to vehicles can prevent or decrease crashes, but unfortunately not all roads are equipped with such traffic sensors. To solve this problem, Dogru and Subasi (2018) considered that the low number of vehicles with vehicle-to-vehicle (V2V) communication devices broadcast some of their microscopic vehicle parameters so that other vehicles which are able to receive the information can use machine learning methods to detect crashes by analyzing the behaviors of vehicles in last a few seconds. With the V2V multisource Big Data collected from the platform established by Peng and Shao, the driving behaviors of 20 professional drivers were analyzed in the field driving test (Peng and Shao, 2018). Moreover, the authors developed a Neural Network-Bayesian filter identification model to identify risky driving behavior including overclose car-following and lane departure by analyzing the lane line distance, vehicle following distance, relative speed, as well as TTC values within a time window. The findings of this study revealed that the identification rate of the proposed model is 8% higher compared to that of a pure Neural Network model and the best identification time window for risky driving behavior is 1.5 s. With the Basic Safety Message data transmitted between more than 2800 connected vehicles and historical crash as well as road inventory data, Arvin et al. (2019) explored the relationship between the frequencies of rear-end, sideswipe, angle as well as head-on crashes at intersections and instantaneous driving behaviors which were regarded as the variations in longitudinal and lateral vehicular control.

4. Discussion

After reviewing the literature for Big Data safety analytics under the environment of ITS and CAV, we summarized all papers from five aspects: 1) data; 2) topics; 3) types of analytics; 4) models and techniques; 5) real-time and high resolution. In terms of data, we elaborated data sources, data types, and data preprocessing.

4.1. Data

4.1.1. Data sources

We divided the data sources into two categories: 1) simulation-generated data sources; 2) pre-existing data sources. The brief description of these data sources is listed below.

1) Simulation-generated data sources: data collected from this type of data source is generated by the driving simulator or traffic simulation such as VISSIM, Simulation of Urban Mobility.
2) Pre-existing data sources: including social media, detector, and various databases such as the latest National Highway-Rail Crossing Inventory Database.

The numbers of the two data sources used in the reviewed papers were counted and shown in Fig. 1. Among the 57 reviewed papers, the studies analyzing traffic safety with data from the pre-existing data sources are more than 90 %, which demonstrates the high accessibility and convenience of collecting data from this type of data source. Meanwhile, the reason for only five papers using data from simulation-generated data sources may be the difficulty and the disadvantage of collecting data using this way. Collecting data by traffic simulation need to create a road segment or road network and set up relative parameters while collecting data using a driving simulator need to find professional drivers to conduct massive experiments, which are complicated and time-consuming. Moreover, the driver's response to the surrounding environment in a driving simulator is not exactly the same as that in reality so that the reliability of data is suspect.

The numbers of various equipment used to collect data were counted and depicted in Fig. 2. The majority of papers did not present equipment, so we only counted equipment that the authors pointed out. Equipment presented in the reviewed papers fall into two categories: traffic flow detectors and weather station. For traffic flow detectors, Martin et al. (2003) presented that detection techniques can be divided into three classes: in-roadway detectors, over-roadway detectors, and off-roadway technologies. Inductive loop detector, one of the most widely applied in-roadway detectors, can detect volume, speed, presence, occupancy and so on (Park et al., 2016; Al Najada and Mahgoub, 2016a; You et al., 2017; Zhang et al., 2018d; Zhu et al., 2018; Chen et al., 2018b). One of the disadvantages of the inductive loop detector is that it will disrupt traffic during installation and maintenance, which can be improved by over-roadway detectors. Over-roadway detectors, commonly known as non-intrusive detectors, are mounted above the roadway or alongside it. Camera (Wang et al., 2015; St-Aubin et al., 2015; Al Najada and Mahgoub, 2016a; Yuan et al., 2018; Peng and Shao, 2018; Ali et al., 2019) and microwave radar sensor (Ahmed and Abdel-Aty, 2013; Shi and Abdel-Aty, 2015; Ozbayoglu et al., 2016) are over-roadway detectors presented in the reviewed papers. Off-roadway technologies, emerging techniques, are becoming more popular in the transportation field. Off-roadway technologies shown in the reviewed papers include global positioning system (Chen et al., 2016; Simandl et al., 2016; Xie et al., 2017; Ren et al., 2017; Bao et al., 2019), automatic vehicle identification (Ahmed and Abdel-Aty, 2013) and driving detector used to collect vehicle running data (Wang et al., 2015; Itoh et al., 2015).

4.1.2. Data types

We also focused on the type of data used to investigate traffic safety

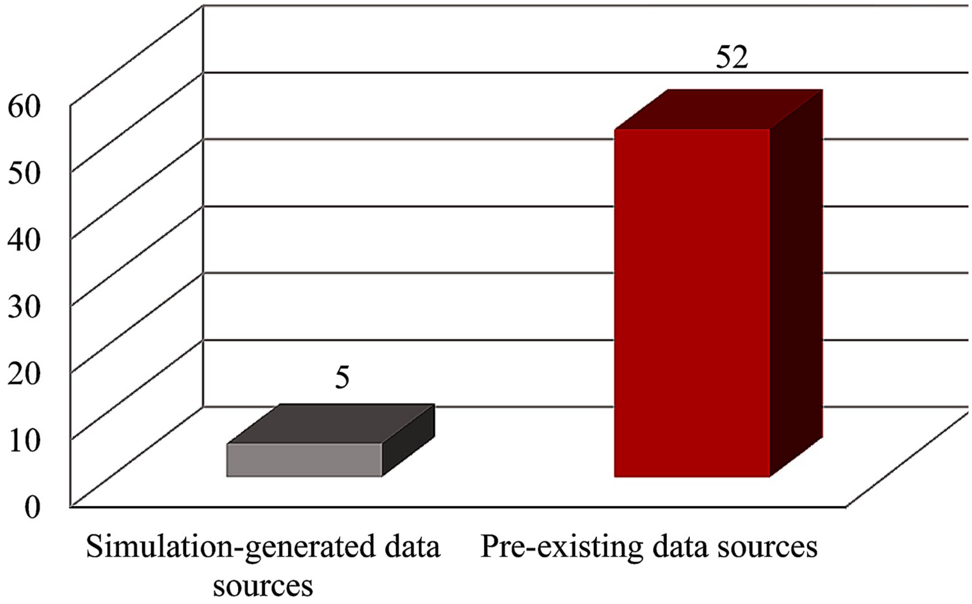

Fig. 1. Frequency of data sources used in the reviewed papers.

under the environment of ITS and CAV. Crash data in some papers are crash record data, and we grouped attributes of crash record data into eight categories according to the characteristic of the data. The numbers of data type and crash data attributes were depicted in Figs. 3 and 4, respectively. As shown in Fig. 3, consistent with our thought, crash data is the most common data type in traffic safety analyses. Traffic data used in the reviewed papers contain two classes: traffic flow data (10 out of 14 traffic data) and traditional traffic exposure variables (e.g. AADT). Social media data refers to data collected from social networks where users share individual social media activity. The low utilization of social media data may track back to the challenge of processing the data. It would be a complicated and huge task for researchers since they need to identify the keywords and extract useful information from social media data. GPS data presented in the reviewed papers mainly include taxi GPS data (Xie et al., 2017; Ren et al., 2017; Bao et al., 2019), human mobility data (Chen et al., 2016) and officer patrol route GPS data (Simandl et al., 2016). Naturalistic Driving Studies (NDS), emerged in the early 2000s, aim at collecting naturalistic driving data which aid in understanding what happened before, during, and after crash and near-crash. Especially, the largest naturalistic driving data was generated by the Second Strategic Highway Research Program Naturalistic Safety Study (SHRP2 NDS), which completed in 2015 in the United States. SHRP2 NDS data (Ali et al., 2019) and smaller scale NDS data were used to examine research issues related to human factors and traffic in the reviewed papers. It is worth noting that the utilization of GPS data and NDS data in Big Data safety analytics is low, which is beyond our expectation.

We considered the remaining data in Fig. 3 as the attributes and analyzed them in combination with the related attributes in Fig. 4. We found the numbers of the attributes related to road, light, and weather are approximate and such data are the most popular data except crash data, which indicates that the attributes related to road, light, and weather are considered most frequently as the contributing factors to traffic crashes. Attributes related to vehicle and driver are used at a similar frequency and some papers combined vehicle dynamic data with driver behavior data because some behaviors of the driver will influence the running status of the vehicle. Socio-demographic and land use, which are considered generally as the macroscopic factors to

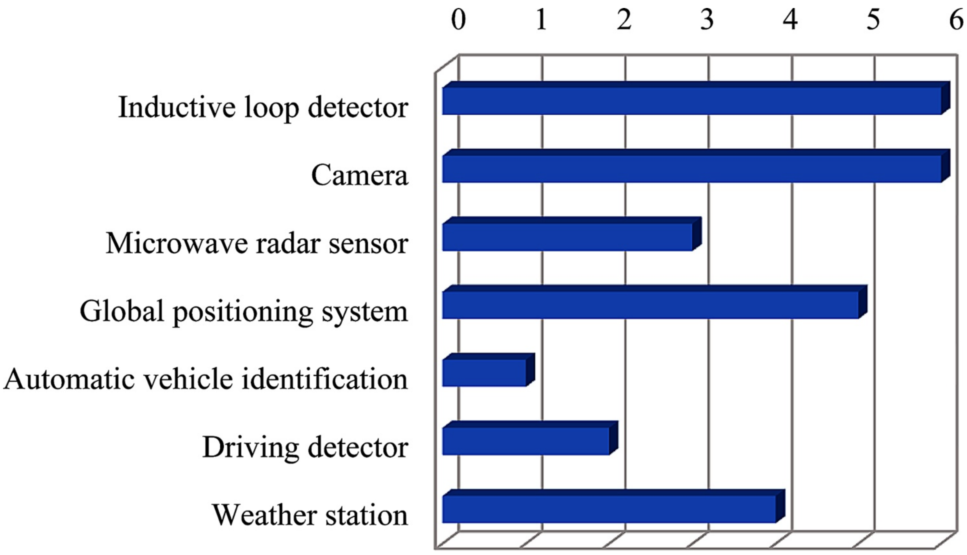

Fig. 2. Frequency of equipment used to collect data in the reviewed papers.

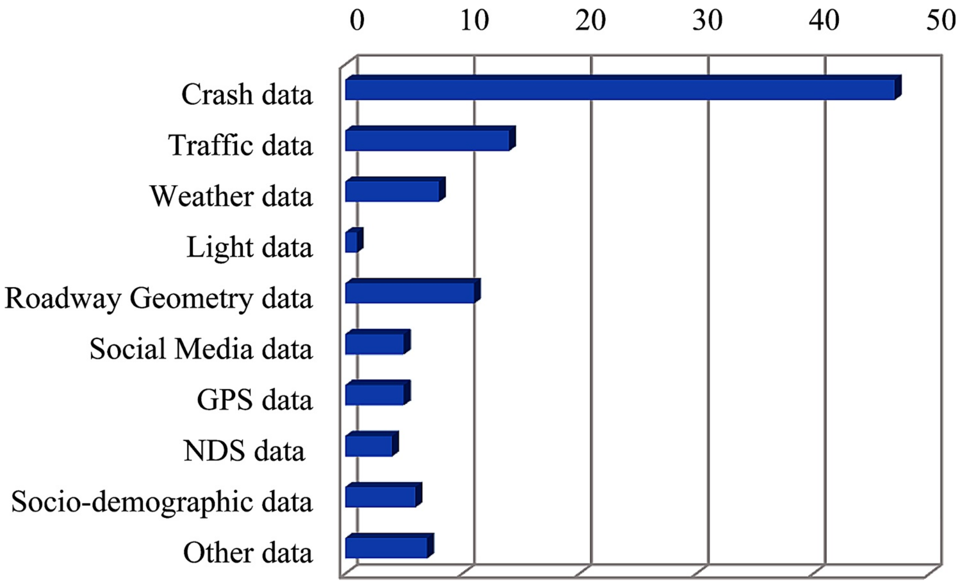

Fig. 3. Frequency of data types used in the reviewed papers.

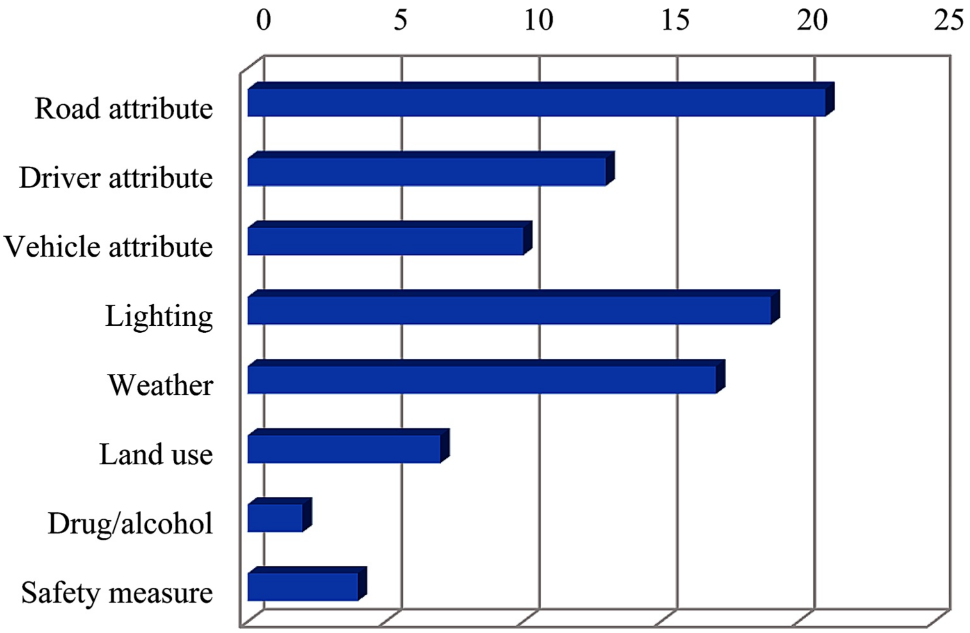

Fig. 4. Frequency of attributes contained in crash record data.

predict traffic crashes, are used infrequently. Therefore, we can speculate that the macroscopic factors related to crash are less considered when researchers analyze traffic safety using Big Data. Safety measure refers to the use of security systems such as airbag, seat belt, safety helmet. Whether the drivers use safety measures and whether they drink alcohol or use drugs are also factors that researchers consider when analyzing traffic safety, even if they are less considered in the reviewed papers.

4.1.3. Data preprocessing

The use of Big Data enables us to explore the problem which cannot be solved by the traditional data analysis. However, Big Data brings us not only opportunities but also challenges including noise, data heterogeneity, data imbalance, high dimensionality, etc. In detail, massive missing values and unrealistic values in data are regarded as noise. Data heterogeneity is a major challenge which is mainly caused by the temporal or spatial variation of data in the reviewed papers. For example, some data is spatially varied such as land use data, demographic data, and road network data while some data is temporally varied such as weather data. Meanwhile, some data is both varied spatially and temporally such as crash data, taxi trip data. Data imbalance refers to the minority of data points fall in the positive class while the majority of data points fall in the negative class, which will cause high false negatives. High dimensionality means that the number of attributes is so large that some of the attributes are useless for traffic safety analysis.

To tackle these problems, the essential procedure before analyzing traffic safety with Big Data is data preprocessing including 1) data cleaning; 2) data integration; 3) data transformation; and 4) data reduction. Data cleaning is done by removing or filling in missing value, cleaning up unrealistic data and repeated data, smoothing noise, and correcting errors. Data integration is used to integrate multi-source data into a unified dataset. Data transformation is done by data aggregation, data normalization and so on to transform the data into a suitable form for learning. Data reduction is used to reduce the size of data while the integrity of the original data is maintained. The above data preprocessing procedures were conducted in almost all studies. Meanwhile, data heterogeneity was tackled by aggregating data into given time and location while high dimensionality was tackled by feature selection generally. Especially, some studies implemented specific methods for data preprocessing. For instance, the missing values was imputed by inverse distance weighting (Yuan et al., 2017). Data heterogeneity was addressed by the matched case-control method (Shi and Abdel-Aty, 2015; You et al., 2017; Das et al., 2019) and eigenanalysis of the road network (Yuan et al., 2017). Data imbalance problems were tackled by some specific methods, including under sampling (Delen et al., 2017), over sampling (Park et al., 2016; Prati et al., 2017; You et al., 2017; Zheng et al., 2019), informative negative sampling approach (Yuan et al., 2017) and bagging based approach (Al Najada and Mahgoub, 2016b). Besides, z-score normalization method (Zheng et al., 2019) was used to normalize data, and kernel density function was used to show the spatial distribution of data (Xie et al., 2017).

4.2. Topics

In total, we reviewed 57 articles using Big Data to study traffic safety and grouped these studies into two classes including Big Data safety analytics under the environment of ITS and Big Data safety analytics under the environment of CAV. Meanwhile, these topics were further grouped into four classes: crash detection or prediction, discovery of contributing factors to traffic crashes, driving behavior analysis, and crash hotspot identification. It should be noted that some studies concentrated on more than one perspective for traffic safety analysis.

As indicated in Fig. 5, most of the reviewed studies focused on the investigation of crash detection or prediction which is always considered to be one of the most important problems in the traffic safety area. Benefiting from the increasingly comprehensive data, the frequency of the papers on crash detection or prediction has increased over years and is almost the highest every year, which indicates the effectiveness of traffic crash detection or prediction research. Crash detection or prediction provides a promising solution to prevent the occurrences of traffic crashes and secondary crashes and reduce the costs of crashes in a proactive way. The second largest number of studies using Big Data, a total number of 24 articles, is the discovery of contributing factors to crashes. Investigating what causes traffic crash is crucial to develop various strategies for preventing crashes and providing a safer road. Even though there are less literature about crash hotspot identification and driving behavior analysis, they are investigated almost every year, which demonstrates that crash hotspot identification and driving behavior analysis also attract attention from researchers. CAV is a research hotspot now but safety analysis for CAV using Big Data is lacking. In addition, the reviewed papers for CAV focused on crash detection or prediction, the discovery contributing factors to crashes, and driving behavior analysis. Crash hotspot identification under the environment of CAV has not been investigated so far.

4.3. Types of analytics

We classified the reviewed papers into three categories: descriptive, predictive, and prescriptive analysis. Descriptive analyses are conducted for describing what happened in the past while predictive analyses for the future and prescriptive analyses for decision making (Delen and Demirkan, 2013). This classification is widely used in Big Data analytics (Delen and Demirkan, 2013; Duan and Xiong, 2015; Nguyen et al., 2018; Ghofrani et al., 2018).

Fig. 6 depicts the frequency of each type of analytics in each year. For the total number of analysis types, we found the applications of Big Data analytics in traffic safety have mainly concentrated on prescriptive analyses (26 out of 65 papers) and descriptive analyses (23 out of 65 papers). On the other hand, predictive analyses have been relatively less conducted. For the number of analysis types in each year, we found that the majority of papers had focused on descriptive analyses in the early stage (2013–2016), and descriptive analyses have begun to decline year by year since 2016. Predictive analyses have increased since 2014 and dominated in 2017. Prescriptive analyses had the largest increase from 2017 to 2018 and dominated from 2018.

Table 1 presents the types of analytics for each application of traffic safety. The values in parentheses are the number of studies related to CAV. This table shows that prescriptive and predictive analyses

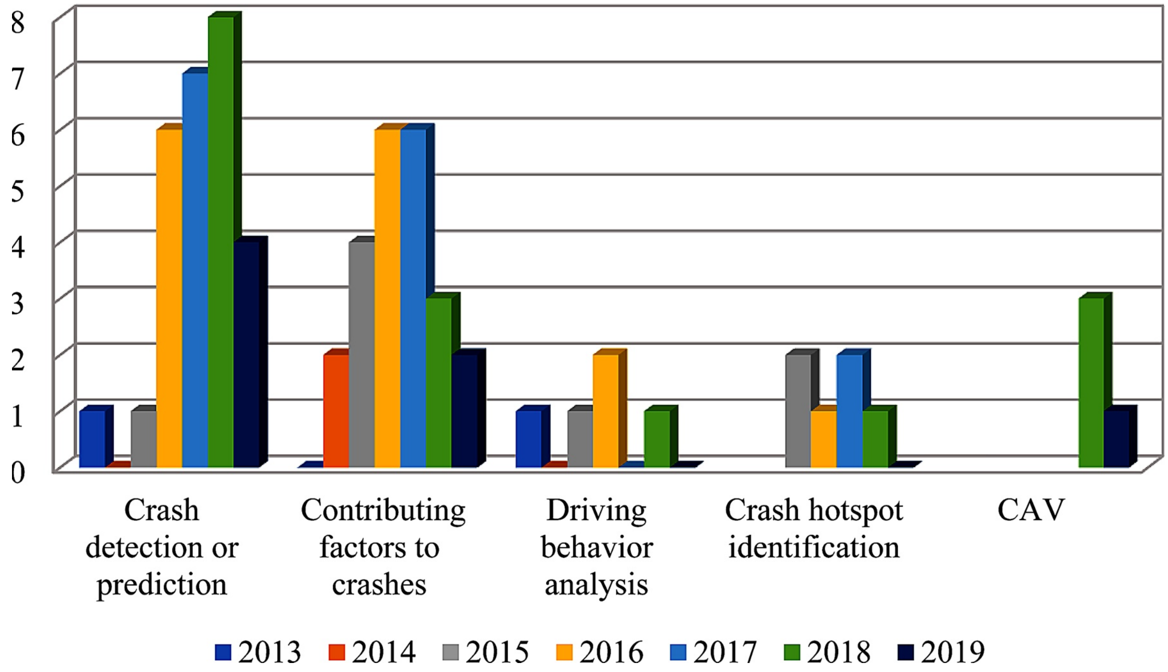

Fig. 5. Frequency of Big Data applications in the reviewed papers.

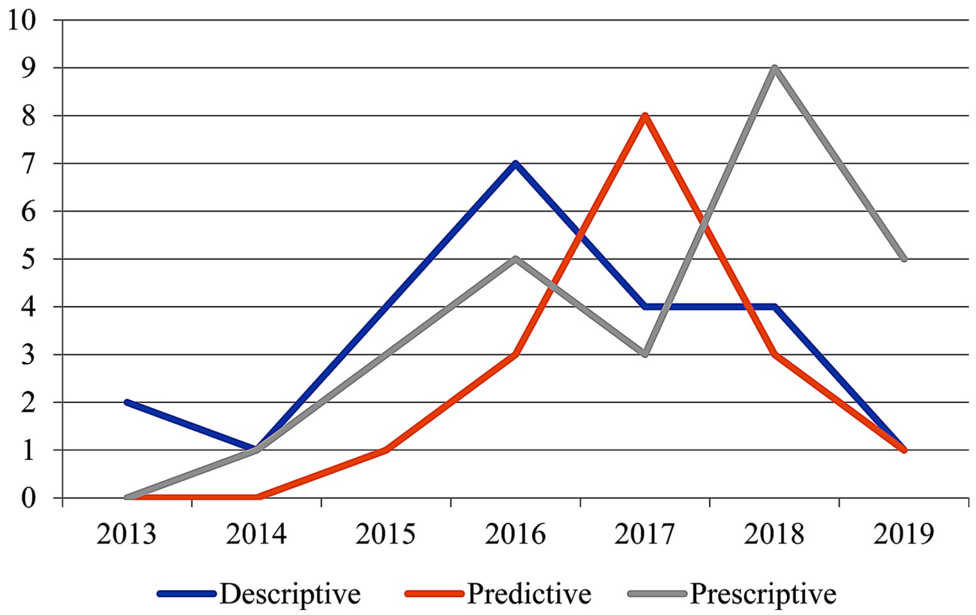

Fig. 6. Distribution of types of analytics in the reviewed papers.

Table 1
Distribution of types of analytics in Big Data safety analytics.

Types of analytics→ ↓Topics	Descriptive	Predictive	Prescriptive	Total
Crash detection or prediction	3(1)	11	13(1)	29
Discovery of contributing factors	13	5	5(1)	24
Driving behavior analysis	3	0	2(1)	6
Crash hotspot identification	3	0	3	6
Total	23	16	26	65

dominated in crash detection or prediction while descriptive analyses dominated in the discovery of contributing factors to traffic crashes. The importance of prescriptive and predictive analyses in reducing crash costs and saving lives is the main reason why these two analyses dominate crash detection or prediction. Finally, the review shows that predictive analyses have not been conducted in driving behavior analysis and crash hotspot identification.

4.4. Models and techniques

The most widely used models in Big Data analytics include classification, clustering, association and so on (Erl et al., 2016). We classified models used in the reviewed papers and summarized the number of models used in each application. Each model includes a number of techniques, which were also summarized. The values in parentheses are the numbers of models or techniques used to analyze traffic safety under the environment of CAV. It is worth mentioning that some papers used more than one model, such as the clustering model first and then the classification model. In this case, we counted all models used in the paper (Table 2). In terms of techniques, it should be noted that some papers used comparative analysis, and some papers used baseline techniques to prove the advantages of the used technique. All techniques used in the papers were counted by us (Table 3).

Classification models are used to identify the category of input data, including binary classification models and multi-class classification models. Classification models, usually used to conduct predictive analyses, are the most widely used models in crash detection or prediction. In detail, binary classification models are generally used to predict crash likelihood while multi-class classification models are used to predict crash injury severity. Another application that often uses classification models is the discovery of contributing factors to traffic crashes. Tree-based techniques (e.g., Decision Tree, Random Forest) are the most applied techniques in classification models. For the discovery of contributing factors to traffic crashes, eight out of eleven classification models used tree-based techniques.

Table 2
Distribution of models used in Big Data safety analytics.

Topics → ↓ Models	Crash prediction	Contributing factors	Behavior analysis	Hotspot identification	Total
Classification	19(1)	11	3(1)	2	37
Regression	9	5(1)		1	16
Association	1	8			9
Clustering	4	4		2	10
Image processing	1		2		3
Simulation	0(1)				1
Visualization	1			1	2
Optimization-based model	0(1)			1	2
Total	38	29	6	7	80

Regression models are used to find trends in data. The regression models are the second most frequently used models, especially for the discovery of contributing factors to traffic crashes and crash detection or prediction. For crash detection or prediction, regression models are generally used to predict crash count or risk and LSTM is the most widely used regression technique (Ren et al., 2017, 2018; Yuan et al., 2018; Bao et al., 2019). Statistical analysis techniques are popular regression techniques in the discovery of contributing factors to traffic crashes (Shi and Abdel-Aty, 2015; Xie et al., 2017; Bao et al., 2017; Chen et al., 2018b; Dong et al., 2018). In addition, NN and statistical analysis techniques are often used as baseline techniques for regression models.

Association models, usually used to implement descriptive analyses, are applied to uncovering the strength of the relationship between the items. Association rule mining is a popular technique used in association models, and Apriori algorithm is most widely applied to finding association rules (Agrawal et al., 1993). All association models in the reviewed papers were investigated by association rule mining and Apriori algorithm (Moradkhani et al., 2014; Babič and Zuskáčová,

Table 3
Distribution of techniques used in Big Data safety analytics.

Models→ ↓ Techniques	Classification	Regression	Association	Clustering	Image processing	Simulation	Visualization	Optimization-based model	N/A	Total
NN	22(2)	11			3	0(1)				39
SVM	10(1)	4								15
Associate rule			9							8
Tree-based	34(1)	7								42
Clustering				10						10
Statistical analysis	19	13(4)								36
Heuristic methods							1	1(1)		3
N/A	6	3			1		1		1	12
Total	95	42	8	10	4	1	2	2	1	165

2016; Kumar and Toshniwal, 2015a, b; Kumar and Toshniwal, 2016; Li et al., 2017; Zhang et al., 2018d; Das et al., 2019; El Mazouri et al., 2019).

Clustering models are a series of models that divide data into several clusters in such a way that objects in the same cluster are more alike compared to that in different clusters. Except for one paper on crash hotspot identification only used clustering model (Sinnott and Yin, 2015), other papers used clustering models first, and then conducted classification models (Wang et al., 2015; Park et al., 2016; Yuan et al., 2017; Alkheder et al., 2017; Iranitalab and Khattak, 2017) or association models (Kumar and Toshniwal, 2015a, b; Kumar and Toshniwal, 2016). Using clustering models before classification or association models can provide high accuracy in the process. For example, Alkheder et al. (2017) firstly split crash data into three clusters and then used ANN classifier to predict crashes. The results after clustering indicated significant improvement in the prediction accuracy. Techniques used in the reviewed papers for clustering include K-means clustering (Park et al., 2016; Alkheder et al., 2017; Wang et al., 2015; Kumar and Toshniwal, 2016), DBSCAN (Sinnott and Yin, 2015), Hierarchical clustering (Itoh et al., 2015), and Spectral clustering (Yuan et al., 2017). In addition, Iranitalab and Khattak (2017) compared the effects of K-means clustering and Latent Class clustering on the performance of different classification models.

The numbers of image processing, simulation, visualization, and optimization-based models are fewer. CNN is the most implemented technique in the image processing models (Tran et al., 2018; Cai et al., 2019). Exact and heuristic algorithms are the main techniques for optimization-based models. Exact algorithms are the ones that always solve an optimization problem to optimality while heuristic algorithms refer to algorithms of solving problems through inductive reasoning of past experience to find the sub-optimal solution of the problem or its optimal solution with a certain probability. For reviewed papers, heuristic algorithms were applied to solving optimization-based models (Triguero et al., 2017; Chen et al., 2018a).

4.5. Real-time and high resolution

The recent safety studies have used more dynamic (or real-time) and/or high-resolution data because of the recent advancement of traffic surveillance. Real-time analysis could address the issue of aggregate analysis—the ignorance of important within-period variations. Meanwhile, the accuracy of traffic safety analysis with high resolution data may be higher since the spatial heterogeneity could be better considered. We conducted further analysis on articles that are real-time or high-resolution for crash detection or prediction. Stylianou et al. (2019) considered papers that used traffic variables aggregated in intervals smaller than six min as real-time analysis. Moreover, according to Cai et al. (2019), data are low-resolution if the data are aggregated based on the zone. The high-resolution explanatory data are aggregated at smaller units in a zone and a zone should have multiple smaller units. The same definitions of real-time and high resolution were applied to this study.

After reviewing real-time Big Data safety analytics, we found the objective of real-time safety analysis is to predict or detect crash likelihood with short time interval traffic data such as 1-min interval traffic data (Ahmed and Abdel-Aty, 2013; Al Najada and Mahgoub, 2016a; Ozbayoglu et al., 2016; Park et al., 2016; You et al., 2017; Zhu et al., 2018; Ali et al., 2019). Papers that used high resolution data to predict crashes usually spatially attached data to the corresponding grids (Chen et al., 2016; Ren et al., 2017, 2018; Yuan et al., 2018; Cai et al., 2019; Bao et al., 2019) or road segments (Al Najada and Mahgoub, 2016a; Yuan et al., 2017). However, the numbers of papers in real-time or at high resolution are very small: only seven papers in real-time and eight papers at high resolution for the topic of crash detection or prediction. Aggregate analysis at low resolution could introduce biases and misleading for crash detection or prediction. With the advancement of intelligent transportation and the enrichment of data, real-time Big Data safety analytics at high resolution should attract more research concerns.

5. Challenges and suggestions

Although Big Data safety analytics under the environment of ITS and CAV have made great achievements, there are still several challenges that need to be investigated and solved. This section introduces these challenges and recommends some suggestions for them.

It is worth thinking about what can be defined as Big Data in transportation. Considering the quantity and diversity of data in some papers, they were not sufficient for being Big Data but the authors still claimed that their studies are Big Data analytics. For example, Bharti et al. (2016) predicted traffic crashes by using only a set of 300 data from the questionnaire. Moreover, how much data are enough for various aspects of safety analysis is a question worth studying. Insufficient data may lead to inaccurate models while excessive data lead to waste resources. Therefore, to better leverage Big Data, some studies should be devoted to embodying and quantifying the standard of Big Data.

Concerning the data source in the reviewed papers, the low utilization of social media data is a disadvantage for Big Data safety analytics, which deserves more attention. As a platform where users share their status at a specific moment and place, social media can provide data about human mobility, traffic crashes, etc. in a timely manner. By applying social media, we can predict traffic crashes in time before any official crash notification arrives at the scene, which assists in triggering a quick crash response. Besides, the proof of crashes measured by traffic flow and occupancy is indirect while the report of crashes from social media data is direct (Zhang et al., 2018d). With the advancement of the traffic surveillance system, NDS data and GPS data are accessible and contain rich information, but the data are not being used adequately. And mobile phone based crowdsourced data (e.g. Strava) is absent from the reviewed literature, which should also arise attention from researchers. Moreover, the review indicates that there is not a good understanding of the linkage between small crash data and other Big Data. Therefore, integrating information from multiple data sources should get more attempts to realize multi-source information fusion, which assists in providing valuable and comprehensive insights into safety analysis.

Regarding the methods used in Big Data safety analytics under the environment of ITS and CAV, we find that no model can accurately predict crashes and study intrinsic causation comprehensively at the same time from reviewed papers, which is a limitation to traffic safety analysis with Big Data. The effective framework to trade off the performance of prediction and the ability of exploring causation should be investigated to assist in analyzing traffic safety.

With respect to Big Data safety applications under the environment of ITS and CAV, firstly, crash hotspot identification under the environment of CAV should draw the attention from researchers, which has not been investigated so far. Secondly, studies on policies for traffic safety with Big Data are lacking. More studies about this topic should be investigated using Big Data, such as the impact of policies on traffic safety and different countermeasures adopted to decrease crashes in different conditions. These studies can provide valuable insights to develop effective policies to prevent crashes. Thirdly, more than half of all road traffic deaths are among vulnerable road users including pedestrians, bicyclists, and motorcyclists (WHO, 2019). However, there are few papers with Big Data to investigate traffic safety for vulnerable road users under the environment of ITS and CAV, which is a problem worthy of researchers' attention. Lastly, the number of studies on Big Data safety analytics at real-time or in high resolution is still small.

Besides, the studies on CAV are one of the most important topics in these days. On the basis of equipped sensors of CAV and road-side units (e.g. video), Big Data can be generated and collected for deeply understanding traffic crash mechanisms and improving the safety of CAV as well as manual vehicles (MV). However, the studies focused on Big Data safety analytics under the environment of CAV are lacking. Connected vehicles can provide awareness of the surrounding situation to reduce blind zone for drivers via V2X communications while automated vehicles of different automation levels can reduce the possibility of crashes caused by driver error, which assists in improving traffic safety. In contrast, CAV may impose new crash risks since a range of driver issues including over-reliance on technology, reduced situation awareness, and loss of driving skills. Therefore, to promote the development of CAV, extensive studies should be contributed to improving traffic safety for CAV with Big Data. First, it is worth investigating the influence that CAV bring to drivers' behavior of not only CAV but also MV by using Big Data since driving behavior is one of the main factors related to crashes. Meanwhile, when and how to take over control by drivers of AVs should be explored to prevent crashes caused by takeover behavior. Second, it is important to investigate the impacts of drivers' behaviors of CAV on safety and new traffic risks raised by CAV in the mixed environment where CAV and MV coexist. Meanwhile, the impacts of environment and market penetration rates of CAV on such risks and how to reduce such risks are worth exploring. Last but not least, Big Data from both CAV and road-side units can be processed in real-time and delivered to relevant road users. Therefore, to enhance safety, considerable computing performance and reliable methods based on cloud-computing which can process Big Data quickly will be required for real-time processing.

6. Conclusions

The advancement of the recent technologies has led to the emergence of massive amounts of data regarding human, vehicles, roads, and environments. Compared with the traditional data, multi-source Big Data can provide more comprehensive information, which enables us to analyze traffic safety more accurately and timely. Moreover, effective analyses of complex Big Data are conducive to identifying new patterns and trends. Due to the benefits of Big Data, a growing body of research has applied Big Data analytics to investigate traffic safety under the environment of ITS and CAV. By summarizing Big Data applications in safety analysis and uncovering the challenges in front of Big Data safety analytics under the environment of ITS and CAV, we reviewed the papers which analyzed traffic safety with Big Data.

To achieve the goal of the paper, we first grouped the reviewed papers into four categories and made a brief description of them. Then, we summarized the data from data sources, data types to data preprocessing. Meanwhile, Big Data applications in safety analysis and models as well as techniques used to investigate traffic safety were discussed. Last but not least, the challenges surrounding Big Data safety analytics under the environment of ITS and CAV were discussed and several suggestions were recommended for these challenges. First, what can be regarded as Big Data analytics and how much data are enough should be embodied and quantified. Second, what should be paid attention to are multi-source information fusion and the low utilization of social media data, NDS data, and mobile phone based crowdsourced data. Third, Big Data analytics on policies for traffic safety, crash hotspot identification for CAV, and the safety of vulnerable road users (e.g., bicyclists, pedestrians, motorcyclists) should attract more attention from researchers. Fourth, to analyze traffic safety more effectively, it is important to explore the effective framework, which can trade-off the performance of prediction and the ability of exploring causation. Finally, the future directions of CAV studies mainly for traffic safety were discussed. Big Data from both CAV and road-side units can be applied to deeply explore traffic crash mechanism, which is helpful for enhancing safety. To realize such promising advantages, more research should be focused on the impacts of CAV on drivers' behaviors of both CAV and MV and new traffic risks raised by CAV. Meanwhile, to process Big Data from CAV and road-side units in a timely manner, considerable computing performance and reliable methods based on cloud-computing are worth investigating. By summarizing data, models, and applications of Big Data in safety research of ITS and CAV, this study will be helpful for researchers to discover the data and methods that have not attracted their attention before and provide new insights to the future study directions for Big Data safety analytics under the environment of ITS and CAV.

CRediT authorship contribution statement

Yanqi Lian: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Guoqing Zhang:** Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Jaeyoung Lee:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing. **Helai Huang:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was funded by Innovation-Driven Project of Central South University (No. 2020CX013).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105711.

References

Agrawal, R., Imielinski, T., Swami, A.N., 1993. Data mining: a performance perspective. IEEE Trans. Knowl. Data Eng. 5 (6), 914–925. https://doi.org/10.1109/69.250074.

Ahmed, M., Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. Transp. Res. Part C Emerg. Technol. 26, 203–213. https://doi.org/10.1016/j.trc.2012.09.002.

Al Najada, H., Mahgoub, I., 2016a. Anticipation and alert system of congestion and accidents in VANET using Big Data analysis for Intelligent Transportation Systems. Paper Presented at the 2016 IEEE Symposium Series on Computational Intelligence (SSCI).

Al Najada, H., Mahgoub, I., 2016b. Big vehicular traffic data mining: towards accident and congestion prevention. Paper Presented at the 2016 International Wireless Communications and Mobile Computing Conference (IWCMC).

Ali, E.M., Ahmed, M.M., Wulff, S.S., 2019. Detection of critical safety events on freeways in clear and rainy weather using SHRP2 naturalistic driving data: parametric and non-parametric techniques. Saf. Sci. 119, 141–149.

Alkheder, S., Taamneh, M., Taamneh, S., 2017. Severity prediction of traffic accident using an artificial neural network. J. Forecast. 36 (1), 100–108. https://doi.org/10.1002/for.2425.

American Automobile Association, 2009. Aggressive Driving: Research Update. American Automobile Association Foundation for Traffic Safety, pp. 5–6.

Arvin, R., Kamrani, M., Khattak, A.J., 2019. How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data. Accid. Anal. Prev. 127, 118–133. https://doi.org/10.1016/j.aap.2019.01.014.

Babič, F., Zuskáčová, K., 2016. Descriptive and predictive mining on road accidents data. In: 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE. pp. 87–92.

Bao, J., Liu, P., Yu, H., Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accid. Anal. Prev. 106, 358–369. https://doi.org/10.1016/j.aap.2017.06.012.

Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 122, 239–254.

Bharti, S., Katiyar, V.K., Kranti, K., 2016. Traffic accident prediction model using support vector machines with Gaussian Kernel. Proceedings of Fifth International Conference on Soft Computing for Problem Solving 1–10.

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., Yuan, J., 2019. Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. Transp. Res. Part A Policy Pract. 127, 71–85. https://doi.org/10.1016/j.tra.2019.07.010.

Castro, Y., Kim, Y.J., 2016. Data mining on road safety: factor assessment on vehicle accidents using classification models. Int. J. Crashworthiness 21 (2), 104–111.

Chen, M., Mao, S., Liu, Y., 2014. Big data: a survey. Mob. Netw. Appl. 19 (2), 171–209.

Chen, Q., Song, X., Yamada, H., Shibasaki, R., 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. Paper Presented at the Thirtieth AAAI Conference on Artificial Intelligence.

Chen, C., Xiang, H., Qiu, T., Wang, C., Zhou, Y., Chang, V., 2018a. A rear-end collision prediction scheme based on deep learning in the Internet of Vehicles. J. Parallel Distrib. Comput. 117, 192–204. https://doi.org/10.1016/j.jpdc.2017.08.014.

Chen, F., Chen, S., Ma, X., 2018b. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. J. Safety Res. 65, 153–159. https://doi.org/10.1016/j.jsr.2018.02.010.

Das, S., Dutta, A., Dixon, K., Minjares-Kyle, L., Gillette, G., 2018. Using deep learning in severity analysis of at-fault motorcycle rider crashes. Transp. Res. Rec. 2672 (34), 122–134.

Das, S., Dutta, A., Sun, X., 2019. Patterns of rainy weather crashes: applying rules mining. J. Transp. Saf. Secur. 1–23. https://doi.org/10.1080/19439962.2019.1572681.

Delen, D., Demirkan, H., 2013. Data, information and analytics as services. Decis. Support Syst. 55 (1), 359–363.

Delen, D., Tomak, L., Topuz, K., Eryarsoy, E., 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. J. Transp. Health 4, 118–131.

Dogru, N., Subasi, A., 2018. Traffic accident detection using random forest classifier. Paper Presented at the 2018 15th Learning and Technology Conference (L&T).

Dong, C., Shao, C., Li, J., Xiong, Z., 2018. An improved deep learning model for traffic crash prediction. J. Adv. Transp. 2018.

Duan, L., Xiong, Y., 2015. Big data analytics and business analytics. J. Manag. Anal. 2 (1), 1–21.

Effati, M., Thill, J.C., Shabani, S., 2015. Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor. J. Geogr. Syst. 17 (2), 107–135.

El Mazouri, F.Z., Abounaima, M.C., Zenkouar, K., 2019. Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France. J. Big Data 6 (1), 5.

Erl, T., Khattak, W., Buhler, P., 2016. Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall Press.

Gantz, J., Reinsel, D., 2012. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Anal. Future 2007 (2012), 1–16.

Ghofrani, F., He, Q., Goverde, R.M., Liu, X., 2018. Recent applications of big data analytics in railway transportation systems: a survey. Transp. Res. Part C Emerg. Technol. 90, 226–246.

Huang, L., Wu, C., Wang, B., Ouyang, Q., 2018. A new paradigm for accident investigation and analysis in the era of big data. Process. Saf. Prog. 37 (1), 42–48.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.

Itoh, M., Yokoyama, D., Toyoda, M., Kitsuregawa, M., 2015. Visual interface for exploring caution spots from vehicle recorder big data. Paper Presented at the 2015 IEEE International Conference on Big Data (Big Data).

Jiang, X., Abdel-Aty, M., Hu, J., Lee, J., 2016. Investigating macro-level hotzone identification and variable importance using big data: a random forest models approach. Neurocomputing 181, 53–63.

Kashani, A.T., Rabieyan, R., Besharati, M.M., 2014. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. J. Safety Res. 51, 93–98.

Kumar, S., Toshniwal, D., 2015a. A data mining framework to analyze road accident data. J. Big Data 2 (1), 26.

Kumar, S., Toshniwal, D., 2015b. Analysing road accident data using association rule mining. In: 2015 International Conference on Computing, Communication and Security (ICCCS). IEEE. pp. 1–6.

Kumar, S., Toshniwal, D., 2016. A data mining approach to characterize road accident locations. J. Mod. Transp. 24 (1), 62–72.

Laney, D., 2001. 3-d Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 2001.

Li, L., Shrestha, S., Hu, G., 2017. Analysis of road traffic fatal accidents using data mining techniques. In: 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA). IEEE. pp. 363–370.

Martin, P.T., Feng, Y., Wang, X., 2003. Detector Technology Evaluation. Mountain-Plains Consortium.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D., 2012. Big data: the management revolution. Harv. Bus. Rev. 90 (10), 60–68.

Moradkhani, F., Ebrahimkhani, S., Sadeghi Begham, B., 2014. Road accident data analysis: a data mining approach. Indian J. Sci. Res. 3 (3), 437–443.

Nguyen, T., Li, Z.H.O.U., Spiegler, V., Ieromonachou, P., Lin, Y., 2018. Big data analytics in supply chain management: a state-of-the-art literature review. Comput. Oper. Res. 98, 254–264.

Ozbayoglu, M., Kucukayan, G., Dogdu, E., 2016. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. Paper Presented at the 2016 IEEE International Conference on Big Data (Big Data).

Pakgohar, A., Tabrizi, R.S., Khalili, M., Esmaeili, A., 2011. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Comput. Sci. 3, 764–769.

Park, S.-h., Kim, S.-m., Ha, Y.-g., 2016. Highway traffic accident prediction using VDS big data analysis. J. Supercomput. 72 (7), 2815–2831. https://doi.org/10.1007/s11227-016-1624-z.

Peng, J., Shao, Y., 2018. Intelligent method for identifying driving risk based on V2V multisource big data. Complexity 2018, 1–9. https://doi.org/10.1155/2018/1801273.

Prati, G., Pietrantoni, L., Fraboni, F., 2017. Using data mining techniques to predict the severity of bicycle crashes. Accid. Anal. Prev. 101, 44–54.

Reinsel, D., Gantz, J., Rydning, J., 2018. The Digitization of the World From Edge to Core. IDC White Paper. .

Ren, H., Song, Y., Liu, J., Hu, Y., Lei, J., 2017. A deep learning approach to the prediction of short-term traffic accident risk. arXiv preprint arXiv:1710.09543.

Ren, H., Song, Y., Wang, J., Hu, Y., Lei, J., 2018. A deep learning approach to the citywide traffic accident risk prediction. Paper Presented at the 2018 21st International Conference on Intelligent Transportation Systems (ITSC).

Sagiroglu, S., Sinanc, D., 2013. Big data: a review. Paper Presented at the 2013 International Conference on Collaboration Technologies and Systems (CTS).

Shi, Q., Abdel-Aty, M., 2015. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C Emerg. Technol. 58, 380–394. https://doi.org/10.1016/j.trc.2015.02.022.

Simandl, J.K., Graettinger, A.J., Smith, R.K., Jones, S., Barnett, T.E., 2016. Making use of big data to evaluate the effectiveness of selective law enforcement in reducing crashes. Transp. Res. Record: J. Transp. Res. Board 2584 (1), 8–15. https://doi.org/10.3141/2584-02.

Sinnott, R.O., Yin, S., 2015. Accident Black spot identification and verification through social media. Paper Presented at the 2015 IEEE International Conference on Data Science and Data Intensive Systems.

St-Aubin, P., Saunier, N., Miranda-Moreno, L., 2015. Large-scale automated proactive road safety analysis using video data. Transp. Res. Part C Emerg. Technol. 58, 363–379. https://doi.org/10.1016/j.trc.2015.04.007.

Stylianou, K., Dimitriou, L., Abdel-Aty, M., 2019. Big data and road safety: a comprehensive review. Mobility Patterns, Big Data and Transport Analytics. Elsevier, pp. 297–343.

Taamneh, M., Alkheder, S., Taamneh, S., 2017. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. J. Transp. Saf. Secur. 9 (2), 146–166.

Tango, F., Botta, M., 2013. Real-time detection system of driver distraction using machine learning. IEEE Trans. Intell. Transp. Syst. 14 (2), 894–905. https://doi.org/10.1109/tits.2013.2247760.

Tasca, L., 2000. A Review of the Literature on Aggressive Driving Research. Ontario Advisory Group on Safe Driving Secretariat, Road User Safety Branch, Ontario Ministry of Transportation, Ontario, Canada.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256.

Tran, D., Do, H.M., Sheng, W., Bai, H., Chowdhary, G., 2018. Real-time detection of distracted driving based on deep learning. IET Intell. Transp. Syst. 12 (10), 1210–1219.

Triguero, I., Figueredo, G.P., Mesgarpour, M., Garibaldi, J.M., John, R.I., 2017. Vehicle incident hot spots identification: an approach for big data. Paper Presented at the 2017 IEEE Trustcom/BigDataSE/ICESS.

Wang, J., Zheng, Y., Li, X., Yu, C., Kodaka, K., Li, K., 2015. Driving risk assessment using near-crash database through data mining of tree-based model. Accid. Anal. Prev. 84, 54–64.

Wang, M.S., Jeong, N.T., Kim, K.S., Choi, S.B., Yang, S.M., You, S.H., ... Suh, M.W., 2016. Drowsy behavior detection based on driving information. Int. J. Automot. Technol. 17 (1), 165–173.

White, M., 2012. Digital workplaces: vision and reality. Bus. Inf. Rev. 29 (4), 205–214.

World Health Organization, 2019. Global Status Report on Road Safety (2018). WHO, Geneva, Switzerland.

Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of traffic crashes involving pedestrians using big data: investigation of contributing factors and identification of hotspots. Risk Anal. 37 (8), 1459–1476. https://doi.org/10.1111/risa.12785.

You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. J. Mod. Transp. 25 (2), 116–123. https://doi.org/10.1007/s40534-017-0129-7.

Yuan, Z., Zhou, X., Yang, T., Tamerius, J., Mantilla, R., 2017. Predicting traffic accidents through heterogeneous urban data: a case study. In: Paper Presented at the Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017). Halifax, NS, Canada.

Yuan, Z., Zhou, X., Yang, T., 2018. Hetero-ConvLSTM. Paper Presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C., 2011. Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. 12 (4), 1624–1639.

Zhang, J., Li, Z., Pu, Z., Xu, C., 2018a. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access 6, 60079–60087. https://doi.org/10.1109/access.2018.2874979.

Zhang, W., Xiao, L., Wang, Y., Kelarestaghi, K., 2018b. Big data approach of crash prediction. Paper Presented at the Transportation Research Board 97th Annual Meeting.

Zhang, Y., Lu, Y., Zhang, D., Shang, L., Wang, D., 2018c. Risksens: a multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing. Paper Presented at the 2018 IEEE International Conference on Big Data (Big Data).

Zhang, Z., He, Q., Gao, J., Ni, M., 2018d. A deep learning approach for detecting traffic accidents from social media data. Transp. Res. Part C Emerg. Technol. 86, 580–596. https://doi.org/10.1016/j.trc.2017.11.027.

Zheng, Z., Lu, P., Lantz, B., 2018. Commercial truck crash injury severity analysis using gradient boosting data mining model. J. Safety Res. 65, 115–124.

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Wang, Z., 2019. Traffic accident's severity prediction: a deep-learning approach-based CNN network. IEEE Access 7, 39897–39910. https://doi.org/10.1109/access.2019.2903319.

Zhu, L., Guo, F., Krishnan, R., Polak, J.W., 2018. A deep learning approach for traffic incident detection in urban networks. Paper Presented at the 2018 21st International Conference on Intelligent Transportation Systems (ITSC).