Accident Analysis and Prevention 136 (2020) 105405

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis

Amir Bahador Parsa[a,*], Ali Movahedi[a], Homa Taghipour[a], Sybil Derrible[b], Abolfazl (Kouros) Mohammadian[a]

[a] *Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W Taylor St, 2095 ERF, Chicago, IL 60607, United States*
[b] *Department of Civil and Materials Engineering, Institute for Environmental Science and Policy, University of Illinois at Chicago, 842 W Taylor St, 2095 ERF, Chicago, IL 60607, United States*

ARTICLE INFO

Keywords:
Accident detection
XGBoost
SHAP
Real time data
Machine learning

ABSTRACT

Detecting traffic accidents as rapidly as possible is essential for traffic safety. In this study, we use eXtreme Gradient Boosting (XGBoost)—a Machine Learning (ML) technique—to detect the occurrence of accidents using a set of real time data comprised of traffic, network, demographic, land use, and weather features. The data used from the Chicago metropolitan expressways was collected between December 2016 and December 2017, and it includes 244 traffic accidents and 6073 non-accident cases. In addition, SHAP (SHapley Additive exPlanation) is employed to interpret the results and analyze the importance of individual features. The results show that XGBoost can detect accidents robustly with an accuracy, detection rate, and a false alarm rate of 99 %, 79 %, and 0.16 %, respectively. Several traffic related features, especially difference of speed between 5 min before and 5 min after an accident, are found to have relatively more impact on the occurrence of accidents. Furthermore, a feature dependency analysis is conducted for three pairs of features. First, average daily traffic and speed after accidents/non-accidents time at the upstream location are interpreted jointly. Then, distance to Central Business District and residential density are analyzed. Finally, speed after accidents/non-accidents time at upstream location and speed after accidents/non-accidents time at downstream location are evaluated with respect to the model's output.

1. Introduction

The occurrence of traffic accidents is a major concern in countries worldwide. With a rapid increase in the number of highways and motorized vehicles in most countries (Global status report on road safety, 2015), the total number of accidents has increased substantially in the world, and an annual report from the National Highway Traffic Safety Administration (NHTSA) reported that around 5,000,000 traffic accidents occur in the United States (US) each year (Traffic Safety Facts, 2013). In fact, traffic accidents have become the second main cause of death for young people and the third reason of death for people who are between 30 and 44 in the US (Traffic safety facts, 2013). Moreover, traffic accidents resulting in death or injury increased by 3 % and 2 %, respectively, between 2011 and 2012 (Traffic safety facts, 2013). In the world, the World Health Organization reported that 1.25 million people die in traffic accidents every year (Global status report on road safety, 2015). Beyond the human health impacts, traffic accidents also generate negative impacts on traffic, especially on highways (Azimi et al., 2019), at intersections (Arvin et al., 2019a), and in work zones (Mokhtarimousavi et al., 2019), often resulting in increased congestion and emission and the worsening of other factors if accidents are not detected properly and rapidly (Traffic Incident Management, 2013).

The transport community has been increasingly making use of novel computational techniques and new data sources (Sharifi et al., 2019; Golshani et al., 2018; Parsa et al., 2019a; Nasr Esfahani et al., 2019; Razi-Ardakani et al., 2018; Ahangari et al., 2019), which have also facilitated the prediction (Mansourkhaki et al., 2016, 2017), detection (Parsa et al., 2019b) and estimation of severity (Azimi et al., 2020; Arvin et al., 2019b) of accidents. For example, smartphones are equipped with several sensors such as an accelerometer, a magnetometer, and a gyroscope that can provide data to detect accidents (Alwan et al., 2016; Fernandes et al., 2016). In fact, traffic information from smartphones has been integrated with other sources of data such as NetLogo simulated data (Thomas and Vidal, 2017) and airbag

[*] Corresponding author.
E-mail addresses: aparsa2@uic.edu (A.B. Parsa), amovah2@uic.edu (A. Movahedi), htaghi2@uic.edu (H. Taghipour), derrible@uic.edu (S. Derrible), kouros@uic.edu (A.K. Mohammadian).

https://doi.org/10.1016/j.aap.2019.105405
Received 9 August 2019; Received in revised form 24 October 2019; Accepted 15 December 2019
Available online 20 December 2019

triggers data (Zaldivar et al., 2011) to detect accidents. These types of data rarely available, however, and they tend to be expensive to acquire (White et al., 2011). Other potential data sources include visual data such as photo and video. Several studies have utilized these data sources along with different techniques including matrix approximation (Xia et al., 2015), statistic heuristic method (Maaloul et al., 2017), hybrid support vector machine with extended Kalman filter (Vishnu and Nedunchezhian, 2018), and extreme learning machine (Chen et al., 2016b) to detect accidents. Vision-based accident detection requires a large amount of information provided by photos and videos, however, and it requires large storage capacity (Chen et al., 2016b). In addition, the accuracy of vision-based accident detection models is significantly affected by weather condition and picture/frame resolution (Xia et al., 2015). Social media is another new data source that can be used to detect accidents for instance through crawling, processing, and filtering tweets (Gu et al., 2016). As an example, Deep Belief Network (DBN) and Long Short-Term Memory (LSTM) are two deep learning models that have been used to detect accidents from Twitter content (Zhang et al., 2018). In these models, it is preferable to fuse social media with traffic data to achieve higher performance (Zhang and He, 2016). In fact, social media data are mostly considered as a supplement than a main source of data for accident detection since they tend to be less available than traffic data (Schulz et al., 2013). Moreover, the precision of the location of accidents reported in social media is generally poor (Zhang and He, 2016).

In the end, traditional traffic data offers a rich and relatively available source of data that can be used for accident detection (Amin and Jalil, 2012; Xu et al., 2016a). In particular, loop detector data is available for most highways and expressways in the US. Many techniques have been used to detect accidents using traffic data (e.g., conditional logistic regression (Kwak and Kho, 2016)) as well as dynamic methods that are able to detect shock waves caused by an accident to detect secondary accidents (Wang et al., 2016). Moreover, many machine learning models have been used to detect accidents, including k-nearest neighbor (Ozbayoglu et al., 2017), regression tree (Ozbayoglu et al., 2017), feed-forward neural network (Ozbayoglu et al., 2017), support vector machine (Dong et al., 2015), probabilistic neural network (Parsa et al., 2019c), dynamic Bayesian network (Sun and Sun, 2015), and deep learning (Chen et al., 2016a; Parsa et al., 2019a). To this end, traffic data is often considering to be best suited to detect and predict the occurrence of accidents (Xu et al., 2016b).

eXtreme Gradient Boosting (XGBoost) is a relatively new method, initially proposed by Chen and Guestrin in 2016 (Chen and Guestrin, 2016), that generally generates high accuracy and fast processing time while being computationally less costly and less complex (Chen and Guestrin, 2016; Hamilton et al., 2019). Meng (2018) has already leveraged XGBoost to predict the occurrence and the duration of accidents using several data sources including road geometric design, historical accident data, as well as traffic and weather data. Furthermore, two studies (Hamilton et al., 2019; Schlögl et al., 2019) showed that XGBoost performed better than several other machine learning techniques to predict the likelihood of an accident including Logistic Regression, Bayesian Regularized Neural Network, Pegasos SVM, Bagging Average Neural Networks, Deep Neural Network, and Gradient Boosting. Shan et al. (2018) also employed artificial neural network to integrate multiple XGBoost models together to predict accident duration. Finally, XGBoost has also been used to predict the severity of traffic accidents, and it achieved high performance, especially when using spatial data (Mokoatle et al., 2019).

In terms of model interpretation—which is especially important when using ML models that are often difficult to interpret—several studies have started to take advantage of SHAP (SHapley Additive exPlanation) (Ribeiro et al., 2016; Štrumbelj and Kononenko, 2014). SHAP was initially proposed by Shapley in 1953 and it is based on game theory (Shapley, 1953). It offers a powerful and insightful measure of the importance of a feature in a model. In 2017, Lundberg and Lee developed a practical package in Python that is able to calculate SHAP for different techniques including LightGBM, GBoost, CatBoost, XGBoost, and Scikit-learn tree models (Lundberg and Lee, 2017). Having known advantages of SHAP, researchers started to utilize this technique (Movahedi and Derrible, 2020). In traffic safety, Mihaita et al. (Mihaita et al., 2019) employed SHAP in 2019 to analyze the impact of different features on accident duration.

The main objective of this study is both to assess the performance of XGBoost to detect the occurrence of accidents in real time and to analyze the importance of individual features for accident detection using SHAP. This work includes traffic, network, demographic, land use, and weather data sources. The data pre-processing procedure adopted in this article is similar to Parsa et al. (2019b)—from the authors of this article—where we used Synthetic Minority Oversampling Technique (SMOTE) to prepare the dataset, and support vector machine and probabilistic neural network for accident detection modeling. The knowledge gained from the study can help urban planners to evaluate (Kashani et al., 2019) and inform policy decisions.

Semantically, we note that the terms *variable* and *feature* are identical. The former tends to be used in statistics and the latter tends to be used in computer science. In this article, we will use *feature* in line with most works that use XGBoost and SHAP.

The rest of this article is organized as follows. First, the data sources used in this study and the feature extraction procedure are presented. Then, XGBoost and SHAP are reviewed in depth in the methodology section. Finally, the performance of the accident detection model is analyzed and discussed, a comprehensive features interpretation is provided through SHAP.

2. Data analysis and feature extraction

The accident data used in this study were collected and archived by the Illinois Department of Transportation (IDOT). The dataset includes 244 accident cases that occurred on the Chicago metropolitan expressways between December 2016 and December 2017. Fig. 1 displays the location of these accidents. In addition, 6073 non-accident cases are selected randomly from the same time period. Each accident/non-accident case occurred on a section that has a loop detector located at the beginning and end of the section in such a way that the length of each section is roughly one kilometer (i.e., 0.6 mile).

By nature, accident data is imbalanced, and datasets include many more instances of non-accident occurrence. In general, oversampling and undersampling techniques have been applied to cope with imbalanced data. Undersampling techniques tend to lead to the loss of a significant portion of data, which can result in a decrease in model accuracy (Han et al., 2005). Oversampling is therefore generally preferred. In this work, we use SMOTE developed by Chawla et al. (2002) that synthesize new data points from members of a minority class through a convex combination of adjacent members. Specifically, SMOTE uses each data point of a minority class and generates new members along the line joining them to their *k* nearest neighbors. In the literature, SMOTE has been used extensively thanks to its ability to generate larger and less specific decision regions (Han et al., 2005) and because it can deal with noisy (Kaur and Gosain, 2018), large, and sparse datasets (Vanhoeyveld and Martens, 2018). In traffic safety, Parsa et al. (2019c) tested several variants of SMOTE on a highly imbalanced dataset similar to that of the current study.

In this study, we used Python 3.6 and employed regular SMOTE to balance the data.

2.1. Traffic data

The traffic data consist of volume, occupancy, and speed collected every 20 s by every loop detector that are present in almost all the expressways of Chicago. After applying some cleaning techniques, the data are then aggregated to 5-min intervals, which tend to offer the

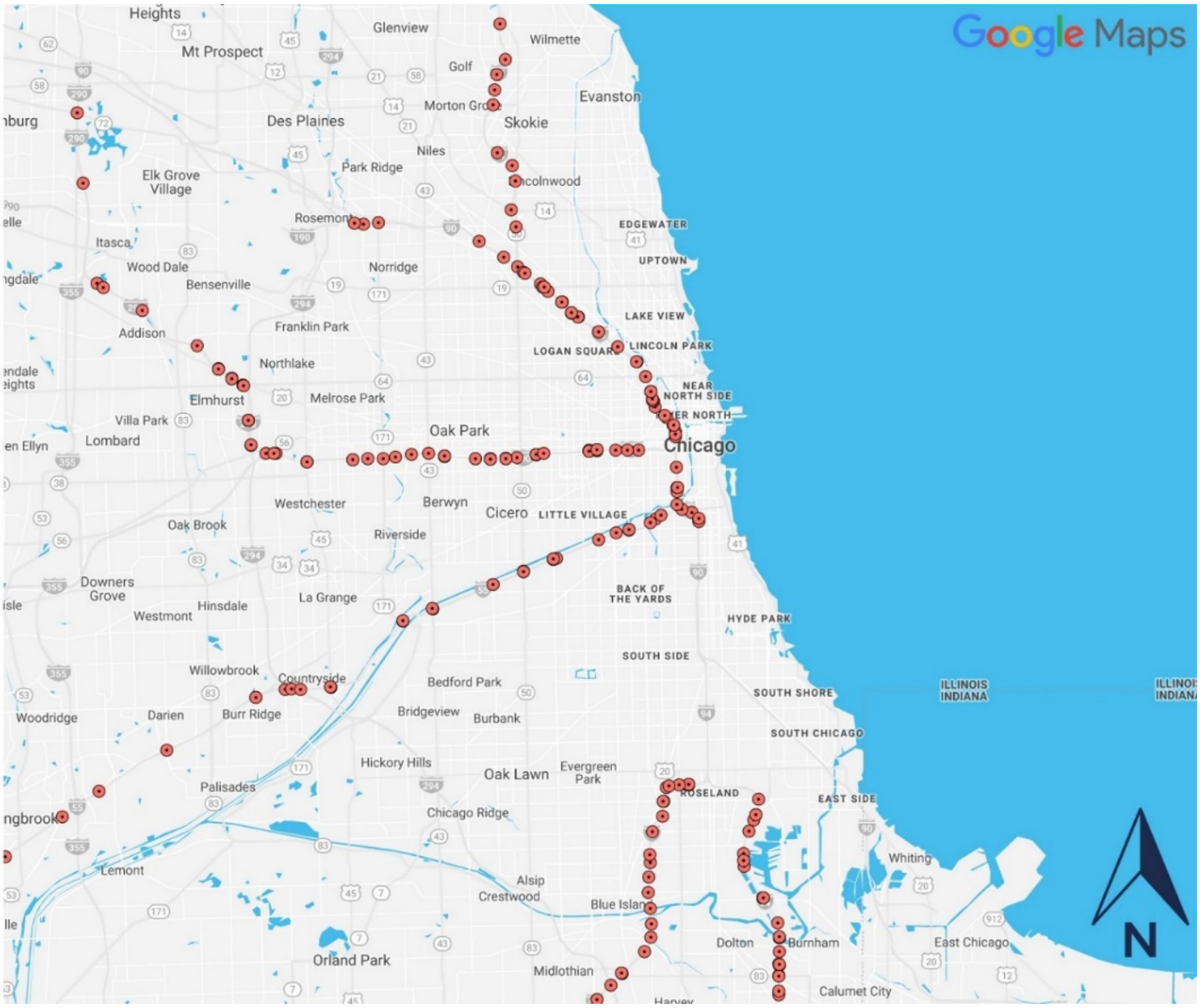


Fig. 1. Location of 244 accidents.

optimal time interval for accident detection from our previous study (Parsa et al., 2019a). Essentially, for a time *t*, each data point contains the traffic conditions from both loop detectors (i.e., the one located upstream and the one located downstream) 5 min before and 5 min after that time. From these data, the difference in traffic conditions between before and after time *t*, and between the upstream and downstream loop detectors are calculated. In addition, we extracted the Average Daily Traffic (ADT) feature for each link of the network from the Highway Performance Monitoring System (HPMS) data.

2.2. Network data

In terms of network data (i.e., the physical transport system), the number of lanes were found to be important and were collected and used in the model. Specifically, two features are extracted: (1) distance of the link to the Central Business District (CBD), and (2) connectivity that is calculated by adding up the number of links connected to the beginning and end of each link and by dividing the result by the length of the link.

2.3. Demographic data

In the network, each link passes through census blocks for which several demographic features are available from the Smart Location Database provided by the US Environmental protection agency (EPA). In this study, the residential density of the block groups around a link was found pertinent and used in the accident detection model. Specifically, a weighted average of the residential density of block groups around a link is calculated with respect to proportional length of the link located in each block group.

2.4. Land use data

Different land use features are extracted from (Clark, 2016) provided by the Chicago Metropolitan Agency for Planning (CMAP). To do so, first, a buffer area is created around each link in the traffic network. Then, the percent of area occupied by different land use types are calculated and assigned to that link. Among all land use types, commercial land use—that includes shopping malls, regional and community retail centers, single large-site retails, offices, cultural or entertainment places, hotels, and motels—is selected to be used in the model.

2.5. Weather condition data

Weather conditions are usually collected from airport weather stations. For this study, two weather stations are available, one at Chicago O'Hare International Airport and one at Chicago Midway International Airport. The distance of each link to the two airports is calculated, and the weather conditions from the closest airport are assigned to the link. Moreover, similar to (Parsa et al., 2019b), in this study, weather conditions are aggregated into one ordinal feature that varies from 1 for

Table 1
Description of explanatory features.

Features	Description	Mean
Traffic		
SpeedA_up	Aggregated speed of 5 min after accident/non-accident, at upstream	52.22 mph
SpeedA_down	Aggregated speed of 5 min after accident/non-accident, at downstream	50.96 mph
DiffSpdBA_U	Difference of speed between before and after of accident/non-accident, at upstream	2.77 mph
VolA_up	Aggregated volume of 5 min after accident/non-accident, at upstream	18.79 veh
DiffVolUD_A	Difference of volume between upstream and downstream, after accident/non-accident time	0.701 veh
ADT	Average daily traffic of link	45346 veh
Network		
Nlanes	Number of lanes	3.48
DistCBD	Distance from centroid of link to the centroid of CBD	10.22 mi
Connectivity	Role of link in making connection between links of network	0.031
Demographic		
ResDensity	Gross residential density (house unit/acre) on unprotected land	4.55 hu/ac
Land Use		
Commercial	Area percentage of buffer zone around link covered by commercial land use	0.47 %
Weather Condition		
Weather	Ordinal feature from 1 for sunny to 4 for stormy weather conditions	1.17

Number of observations: 6317.

sunny to 4 for stormy and harsh weather conditions.

All features generated and used in this work and their descriptions are displayed in Table 1.

3. Methodology

In this study, XGBoost is used to model accident detection. XGBoost is an efficient implementation of gradient boosted decision trees. A Decision Tree (DT) has a structure similar to a tree with a root node (topmost node), internal nodes, and leaf nodes (end nodes). DT algorithms generally use simple rules to start from the root node and branch out, going through internal nodes, to finally end up in the leaves. In contrast, gradient boosted decision tree is an ensemble learning technique that uses a sequence of decision trees, where each decision tree learns from the previous tree and affects the next tree to improve the model and build a strong learner (Friedman, 2001). In the next section, we explain the equations behind XGBoost; interested readers are referred to a study conducted by Chen and Guestrin (Chen and Guestrin, 2016) for more details.

In this work, Python 3.6 was systematically used, from applying SMOTE to model testing. For model training, we used the XGBoost package and the Scikit-learn library.

3.1. Extreme gradient boosting – XGBoost

Given a dataset with n samples, there are independent variables x_i, and each of these variables has m features therefore $x_i \in \mathbb{R}^m$. For each of these variables, there are corresponding dependent variables y_i, $y_i \in \mathbb{R}$. A tree ensemble model predicts the dependent variable such $y\hat{}_i$ using the independent variables and K additive functions:

$$y\hat{}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

Here, f_k represents an independent tree structure with leaf scores and F is the space of trees. The goal is to minimize Eq. (2):

$$\mathscr{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

where l is a loss function, Ω is a term for penalizing the complexity of the model, and:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega_i\|^2 \tag{3}$$

In Eq. (3), T is the number of leaves, and ω_i is the score of the i^{th} leaf. By solving Eqs. (1)–(3), the optimal values for ω_j^*, and the corresponding values are:

$$\omega_j^* = -\frac{\sum_{i\in I_j} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})}{\sum_{i\in I_j} \partial^2{}_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} \tag{4}$$

$$\mathscr{L}^{\sim t}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})\right)^2}{\sum_{i\in I_j} \partial^2{}_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} + \gamma T \tag{5}$$

Since, in practice, it is difficult to calculate this value for all possible tree structures, instead, the following formula is used:

$$\begin{aligned}\mathscr{L}_{split} = \frac{1}{2}\Bigg[&\frac{\left(\sum_{i\in I_L} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})\right)^2}{\sum_{i\in I_L} \partial^2{}_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} + \frac{\left(\sum_{i\in I_R} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})\right)^2}{\sum_{i\in I_R} \partial^2{}_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda}\\ &- \frac{\left(\sum_{i\in I} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})\right)^2}{\sum_{i\in I} \partial^2{}_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda}\Bigg] - \gamma\end{aligned} \tag{6}$$

where:

$$I = I_L \cup I_R \tag{7}$$

Here I_L are the instances sets of left nodes after the split and I_R are the instances sets of right nodes after the split (Chen and Guestrin, 2016).

As an added benefit, as a decision tree algorithm, XGBoost is not impacted by multicollinearity (Badr, 2019). Therefore, even if two variables capture the same phenomenon in a system, both can be kept, which is particularly desirable here since we perform a significant feature analysis through SHAP (detailed below).

3.2. Model parameters

In XGBoost, several parameters need to be selected to maximize model performance. Parameter tuning is important for XGBoost to prevent overfitting and too much complexity, although it can be difficult since XGBoost uses multiple parameters. Overfitting happens when a model starts to learn noises and random fluctuations and finally considers them as meaningful facts or concepts.

The number of iterations is the number of trees that are fitted in the model. The maximum depth of the tree represents the maximum number of splits; increasing the maximum depth can cause overfitting.

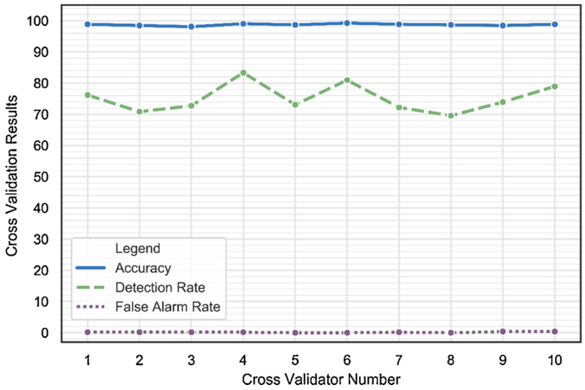


Fig. 2. Model cross validation.

The fraction of observations which randomly selected for the training instances are defined by "subsample" which prevents overfitting. At each step, the learning rate is used to shrink the weights and change the impact of each individual tree and make the model more robust (Chen and Guestrin, 2016). The next parameter that can help prevent overfitting through subsampling the columns is "colsample_bytree". The parameters "lambda" and "alpha" are L2 and L1 regularization terms on weights, respectively, and their increment makes the model more conservative. The optimal XGBoost hyper-parameters values selected after cross-validation are: number of iterations: 500, max depth: 6, subsample: 0.8, colsample bytree: 0.4, lambda: 1.5, alpha: 0.2, and learning rate: 0.01.

3.3. Model evaluation

In this study, accuracy, detection rate (a.k.a. *sensitivity*), and false alarm rate are selected to evaluate model performance; they are defined in Eqs. (8)–(10). The goal is to develop a model with a high accuracy and detection rate, and with a low false alarm rate.

$$Accuracy = \frac{Numberoftruereports}{Totalnumberofcases} \times 100 \tag{8}$$

$$DetectionRate = \frac{Numberoftrueaccidentreports}{Totalnumberofaccidents} \times 100 \tag{9}$$

$$FalseAlarmRate = \frac{Numberoffalseaccidentreports}{Totalnumberofcases} \times 100 \tag{10}$$

It is worth noting that an alternative to false alarm rate for showing performance of model in detecting non-accident cases (i.e., majority class members) is *specificity* which is proportion of non-accident reports detected correctly by the model.

3.4. Model interpretation

SHAP (Shapley Additive exPlanations), proposed by Lundberg and Lee (Lundberg and Lee, 2017), is used to interpret the output of the model. SHAP is based on game theory (Štrumbelj and Kononenko, 2014) and local explanations (Ribeiro et al., 2016), and it offers a means to estimate the contribution of each feature. Assume an XGBoost model where a group N (with n features) used to predict an output (N). In SHAP, the contribution of each feature (ϕ_i is contribution of feature i) on the model output $v(N)$ is allocated based on their marginal contribution (Shapley, 1953). Based on several axioms to help fairly allocate the contribution of each feature, shapely values are determined through:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S \cup \{i\}) - v(S)] \tag{11}$$

A linear function of binary features g is defined based on the following additive feature attribution method:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{12}$$

where $z' \in \{0, 1\}^M$, equals to 1 when a feature is observed, otherwise it equals to 0, and M is the number of input features (Lundberg and Lee, 2017).

4. Results and discussion

4.1. XGBoost model

In this work, XGBoost is trained on 65 % of the data that is selected randomly, and the remaining 35 % is used to test the model. In addition, a 10-fold cross-validation process is employed on the training data to test the stability of the model performance—that is, the training data is divided into ten subsamples randomly, and ten models are trained in such a way that each time nine subsamples are used to train a model and one subsample is used to test a model.

Fig. 2 displays the results of the 10-fold cross-validation for the three performance measures: accuracy, detection rate, and false alarm rate. We can see that the accuracy is close to 100 % and the false alarm rate are less than 0.4 % for all 10 models. Moreover, the detection rate varies between 70 % and 83 %. Overall, these indicators are significantly high and suggest that XGBoost can be used to model accident detection. XGBoost is finally retrained on the full 65 % training set and tested on the 35 % testing set. The final performance indicators achieved are as follows: accuracy of 99 %, detection rate (a.k.a. sensitivity) of 79 %, and false alarm rate of 0.16 % (specificity is equal to 0.998), which is again indicative of a significantly high performance. The steady low false alarm rate (high specificity) of XGBoost was further tested and validated in other studies as well (Soleimani et al., 2019).

To interested readers, we also compared the performance of XGBoost with four other popular machine learning techniques, and the results are reported in Supplementary materials. From the five techniques tested, XGBoost performed best. Supplementary materials also contain a short discussion of the benefits of using SMOTE to improve the accuracy of the model.

4.2. Feature analysis

Fig. 3 shows the SHAP summary plot that orders features based on their importance to detect accidents. We can see that traffic-related features are the most important features in the model. Specifically, the difference in speed between before and after an accident at the upstream location (i.e., DiffSpdBA_U) has the greatest impact on the model. This observation is likely related to the creation of a shockwave

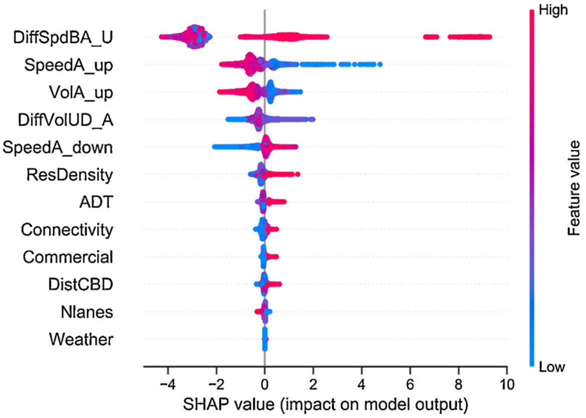


Fig. 3. SHAP summary plot.

that moves rapidly toward the traffic at the upstream location and that manifests itself by a sudden drop of traffic speed (Wang et al., 2016). Moreover, higher values of this feature result in higher SHAP values, which correspond to a higher probability that an accident has occurred. Speed and volume after accident at the upstream location are the next two most important features, respectively, and lower values of these features (i.e., lower speeds and volumes) correspond to a higher chance of accident occurrence. In fact, after an accident, the congestion created at the upstream location slows down the traffic and reduces the volume. In contrast, higher values of speed after an accident at the downstream location correspond to a higher chance of accident occurrence. In fact, an analysis of driving behavior showed that when drivers pass an accident scene, they tend to increase their speed to or even slightly higher than the speed limit (Yishui et al., 2015).

These observations are reasonable since an accident is more likely represented by a low upstream speed and a high downstream speed than by both low/ high upstream and downstream speeds. A similar observation can be made with volume, and therefore we expect that a higher difference in upstream and downstream volume (i.e., higher DiffVolUD_A) translates to a higher probability that an accident has occurred.

In terms of traffic characteristics, Fig. 3 also shows that higher average daily traffic results in a higher probability of an accident. This might stem from the fact that as more vehicles pass through a specific link, the probability of having an accident increases.

Subsequently, residential density, as a demographic feature, is found to have a significant impact—in fact, its impact is higher than average daily traffic. The results suggest that links located in block groups with a higher residential density are more likely to have accidents. Although the reason is not definite, we might hypothesize that areas with higher building density have a mixed traffic composition that impacts the probability of accident occurrence. This is in line with the results from previous studies on traffic mixture of residential area (Boulange et al., 2017; Lachapelle, 2015). In addition, although their SHAP values are significantly lower than the features already mentioned, both connectivity and distance to CBD also have a positive impact on accident occurrence. That is, when a link makes more connections in the network, it is more likely to pass through more traffic and the chance of accident occurrence increases (similar to average daily traffic). Moreover, links located further away from the CBD are more likely to have accidents. Once again, the reason is not definite, but we might hypothesize that the average speed increase as we move away from the CBD reduces traffic safety thus leading to a higher probability of accidents (the relationship is actually more complex as we will see in the next section). To test the relationship, a correlation analysis between average speed and distance to CBD is performed and found to be positive with a correlation coefficient of 0.252.

In terms of land use, commercial land use also has a positive impact on occurrence of accident, although here as well, the SHAP value is significantly low. Similar to building density, commercial areas may have more mixed traffic that may increase the probability of accident occurrence. The number of lanes slightly negatively affects accident occurrence, which means accidents are more likely to occur on expressways with fewer lanes.

Finally, weather condition is the least significant feature, which is interesting and unexpected. Although we should note that a high number of accidents in the dataset (i.e., 88 %) have happened during sunny and fine weather conditions. That being said, Fig. 3 shows that weather has a slight positive impact on accident occurrence, meaning that harsh weather conditions increase the probability of accident occurrence, which is expected.

4.3. Feature dependency analysis

In Fig. 4, we plot the value of a feature on the *x*-axis and the SHAP value of it on the *y*-axis by changing a specific feature in the model. In Fig. 4-a, we select ADT as the feature to determine its impact when SpeedA_up increases from 10 mph to 80 mph. The red points represent higher values of SpeedA_up, and the blue points represent lower ones. When ADT is low, SHAP values for high SpeedA_up are above zero, which suggests that increasing ADT increases the probability of accident occurrence. That is, increasing traffic while speeds are high and traffic volume is low is resulting in higher chance of accident. In contrast, SHAP values for low SpeedA_up are below zero, which suggests that increasing ADT while speeds and traffic volumes are low reduces the probability of accident occurrence which can be because of less maneuverability in this condition.

Ignoring the color of the figure and focusing on how changing ADT impacts the model outputs reveals that for low ADT values, overall SHAP values are positive up to a point around 25,000. Then, SHAP values are negative, which means that by increasing ADT, the probability of accident occurrence decreases. After that, when ADT is higher than 60,000, SHAP values become positive again, which means increasing ADT will increase the probability of accident occurrence.

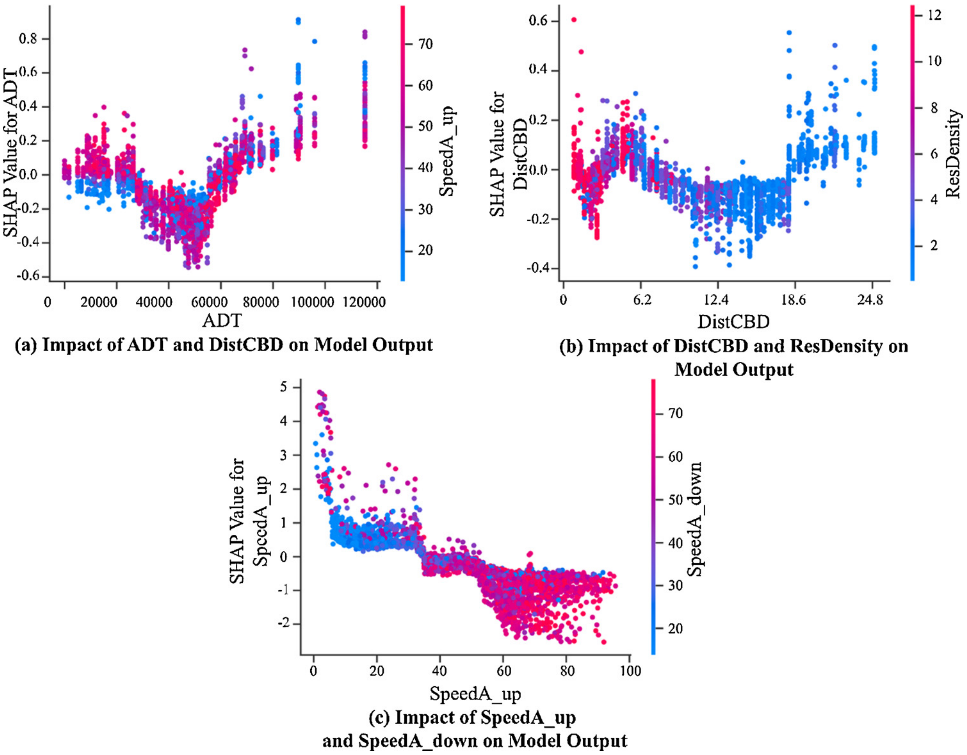
Fig. 4-b shows the impact of DistCBD and ResDensity on accident detection. Despite some noises, SHAP values are mostly negative until a DistCBD value of 3 miles. Then SHAP values increase and become positive in the range of 3–9 miles, and again a reduction happens in SHAP values and they become negative up to DistCBD of 18.5 miles. Finally, after a DistCBD of 18.5 miles, SHAP values become positive again. The highly nonlinear impact of DistCBD is interesting, and it demonstrates that looking at one single parameter (e.g., whether it is positive or negative) is often not sufficient. Essentially, the SHAP values suggest that the impact of DistCBD starts negatively; that is, increasing DistCBD when it is below 3 miles, results in fewer accidents. Between 3–9 miles, the relationship becomes positive, and increasing in DistCBD causes more accidents. Once again, the impact of DistCBD becomes negative between 9 and 18.5 miles, after which it finally stays positive suggesting that increasing in DistCBD results in more accidents. Furthermore, the blue points mostly appear for higher values of DistCBD; essentially telling us the residential density tends to be lower away from the CDB.

Finally, Fig. 4-c displays the impact of SpeedA_up and SpeedA_down in the accident detection model. Points with low SpeedA_down values (i.e., blue points) are mostly on the left-hand side of the figure, where values of SpeedA_up are also low. The red points located on the left-hand side (i.e., with low SpeedA_up and high SpeedA_down) likely correspond to the accidents present in the dataset. SHAP values are positive for points with SpeedA_up below 37 mph, which states increasing SpeedA_up when it is lower than 37 mph also increases the probability of accident occurrence. In contrast, SHAP values become negative for points with SpeedA_up above 37 mph, which shows the negative correlation between SpeedA_up and accident occurrence.

5. Summary and conclusion

In this study, XGBoost is trained to model accident detection using a set of real-time data extracted and generated from different data sources. In total, 244 accident cases and 6073 non-accident cases are used to train the model that achieved an accuracy of 99 %, detection rate of 79 %, and false alarm rate of 0.16 %. Feature importance analysis is applied to the final model using SHAP, and traffic related features (especially speed) is found to have a substantial impact on the probability of accident occurrence in the model. Furthermore, a SHAP dependency analysis is performed, and the impacts of three pairs of features on the model are captured and described. After traffic related features, demographic, network, land use, and weather condition features are found to have a high impact on the probability of accident occurrence.

The results from the model can also inform policy decisions. As an example, since residential density and speed both have positive impacts on the probability of accident occurrence, one can suggest



(a) Impact of ADT and DistCBD on Model Output

(b) Impact of DistCBD and ResDensity on Model Output

(c) Impact of SpeedA_up and SpeedA_down on Model Output

Fig. 4. SHAP dependency analysis.

implementing more traffic calming measures in neighborhoods with relatively higher residential density.

The high performance of XGBoost supports its ability to detect accidents. Moreover, steady low values of false alarm rates in model training (i.e., 10-fold cross validation) confirms the robust behavior of this technique in detecting accidents as well.

Furthermore, SHAP offers an insightful means to interpret the results from a complex algorithm such as XGBoost. The technique is not only capable of evaluating the importance and direction of the impacts of a feature on the output of model, it can also extract complex and nonlinear joint impacts of features on the output of a model. In particular, in this work, the fluctuation impact of distance to CBD provided interesting information that is not captured by most other techniques.

As a limiting factor, to be able to detect accidents more rapidly, it would be desirable to lower the aggregation interval below 5 min as is currently performed to maximize performance accuracy. A possible future avenue is to fuse loop detector data with other data sources such as Telematics data, although these sources are rarely made available. Moreover, data on pavement condition and geometric design of expressways could improve the accuracy of the model but were not available for this study.

Finally, for proper application of the method developed here, we note that real-time loop detectors should be maintained regularly as any malfunction in the detectors would affect the accuracy of the model prediction.

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.

Acknowledgments

The research leading to these findings has received funding from the Illinois Department of Transportation (IDOT) and from the National Science Foundation (NSF) CAREER award #1551731. The authors would also like to thank IDOT for collecting and archiving the loop detectors data used in this study.

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.2019.105405.

References

Ahangari, S., Jeihani, M., Dehzangi, A., 2019. A Machine Learning Distracted Driving Prediction Model. International Symposium of Intelligent Unmanned Systems on Artificial Intelligence.

Alwan, Z.S., Muhammed, H., Alshaibani, A., 2016. Car accident detection and notification system using smartphone. Int. J. Comput. Sci. Mob. Comput.(January).

Amin, S., Jalil, J., 2012. Accident detection and reporting system using GPS, GPRS and GSM technology. Int. Conf. Informatics, Electron. Vis 640–643. https://doi.org/10.1109/ICIEV.2012.6317382.

Arvin, R., Kamrani, M., Khattak, A.J., 2019a. How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data. Accid. Anal. Prev. 127, 118–133.

Arvin, R., Kamrani, M., Khattak, A.J., 2019b. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. Accident Analysis & Prevention 132.

Azimi, G., Asgari, H., Rahimi, A., Xia, J., 2019. Investigation of heterogeneity in severity analysis for large truck crashes. 98th Annu. Meet. Transp. Res. Board.

Azimi, G., Rahimi, A., Asgari, H., Jin, X., 2020. Severity analysis for large truck rollover crashes using a random parameter ordered logit model. Accident Analysis & Prevention 135, 105355.

Badr, W., 2019. Why feature correlation matters…. A lot! Towar. Data Sci.

Boulange, C., Gunn, L., Giles-corti, B., Mavoa, S., Pettit, C., Badland, H., 2017. Examining associations between urban design attributes and transport mode choice for walking, cycling, public transport and private motor vehicle trips. J. Transp. Health 6 (July), 155–166.

Chawla, N., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chen, Q., Song, X., Yamada, H., Shibasaki, R., 2016a. Learning deep representation from big and heterogeneous data for traffic accident inference. 30th AAAI Conf. Artif. Intell. 338–344.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. Proc. 22nd acm sigkdd Int. Conf. Knowl. Discov. Data Min. 785–794.

Chen, Y., Yu, Y., Li, T., 2016b. A vision based traffic accident detection method using extreme learning machine. Int. Conf. Adv. Robot. Mechatronics 567–572. https://doi.org/10.1109/ICARM.2016.7606983.

Clark, D., 2016. Chicago Metropolitan Agency for Planning's 2013 Land Use Inventory for Northeastern Illinois, Version 1.0. Chicago Metrop. Agency Plan.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. Accid. Anal. Prev. 82, 192–198. https://doi.org/10.1016/j.aap.2015.05.018.

Fernandes, B., Alam, M., Gomes, V., Ferreira, J., Oliveira, A., 2016. Automatic accident detection with multi-modal alert system implementation for ITS. Veh. Commun. 3, 1–11. https://doi.org/10.1016/j.vehcom.2015.11.001.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Global status report on road safety 2015, 2015. World Heal. Organ.

Golshani, N., Shabanpour, R., Mahmoudifard, S.M., Derrible, S., Mohammadian, A., 2018. Modeling travel mode and timing decisions: comparison of artificial neural networks and copula-based joint model. Travel Behav. Soc. (September 2017), 21–32. https://doi.org/10.1016/j.tbs.2017.09.003.

Gu, Y., Sean, Z., Chen, F., 2016. From Twitter to detector: real-time traffic incident detection using social media data. Transp. Res. Part C 67, 321–342. https://doi.org/10.1016/j.trc.2016.02.011.

Hamilton, B.A., Bakhit, P.R., Ishak, S., 2019. An eXtreme gradient boosting method for identifying the factors contributing to crash/near-crash events: a naturalistic driving study. Can. J. Civ. Eng. 1–32.

Han, H., Wang, W., Mao, B., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Int. Conf. Intell. Comput. 878–887.

Kashani, H., Movahedi, A., Morshedi, M.A., 2019. An agent-based simulation model to evaluate the response to seismic retrofit promotion policies. International Journal of Disaster Risk Reduction 33, 181–195. https://doi.org/10.1016/j.ijdrr.2018.10.004.

Kaur, P., Gosain, A., 2018. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. ICT Based Innov. 23–30.

Kwak, H., Kho, S., 2016. Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. Accid. Anal. Prev. 88, 9–19. https://doi.org/10.1016/j.aap.2015.12.004.

Lachapelle, U., 2015. Walk, bicycle and transit trips of transit dependent and choice riders in the NHTS 2009. J. Phys. Act. Health (April), 1139–1147. https://doi.org/10.1123/jpah.2014-0052.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 4765–4774.

Maaloul, B., Taleb-ahmed, A., Niar, S., Harb, N., Valderrama, C., 2017. Adaptive video-based algorithm for accident detection on highways. 12th IEEE Int. Symp. Ind. Embed. Syst. 1–6. https://doi.org/10.1109/SIES.2017.7993382.

Mansourkhaki, A., Karimpour, A., Sadoghi Yazdi, H., 2016. Non-stationary concept of accident prediction. Proceedings of the Institution of Civil Engineers-Transport 170, 140–151.

Mansourkhaki, A., Karimpour, A., Sadoghi Yazdi, H., 2017. Introducing prior knowledge for a hybrid accident prediction model. KSCE Journal of Civil Engineering 1912–1918.

Meng, H., 2018. Expressway crash prediction based on traffic big data. 2018 Int. Conf. Signal Process. Mach. Learn. 1–6.

Mihaita, A.-S., Liu, Z., Cai, C., Rizoiu, M.A., 2019. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. arXiv Prepr. arXiv1905.12254.

Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved support vector machine models for work zone crash injury severity prediction and analysis. Transp. Res. Rec. https://doi.org/10.1177/0361198119845899.

Mokoatle, M., Marivate, V., Esiefarienrhe, M., 2019. Predicting road traffic accident severity using accident report data in South Africa. 20th Annu. Int. Conf. Digit. Gov. Res. 11–17.

Movahedi, A., Derrible, S., 2020. Interrelated Patterns of Electricity, Gas, and Water Consumption in Large-Scale Buildings. (under review). Sustainable Cities and Society.

Nasr Esfahani, H., Arvin, R., Song, Z., Sze, N.N., 2019. Prevalence of cell phone use while driving and its impact on driving performance, focusing on near-crash risk: A survey study in tehran. Journal of Transportation Safety & Security. https://doi.org/10.1080/19439962.2019.1701166.

Ozbayoglu, M., Kucukayan, G., Dogdu, E., 2017. A Real-Time Autonomous Highway Accident Detection Model Based on Big Data Processing and Computational Intelligence. IEEE, pp. 1807–1813.

Parsa, A.B., Chauhan, R.S., Taghipour, H., Derrible, S., Mohammadian, A. (Kouros), 2019a. Applying Deep Learning to Detect Traffic Accidents in Real Time Using Spatiotemporal Sequential Data. arXiv Preprint arXiv 1912.06991.

Parsa, A.B., Kamal, K., Taghipour, H., Mohammadian, A.K., 2019b. Does security of neighborhoods affect non-mandatory trips? A copula-based joint multinomial-ordinal model of mode and trip distance choices. Transportation Research Board 98th Annual Meeting.

Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019c. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129 (January), 202–210. https://doi.org/10.1016/j.aap.2019.05.014.

Razi-Ardakani, H., Mahmoudzadeh, A., Kermanshah, M., 2018. A Nested Logit analysis of the influence of distraction on types of vehicle crashes. European Transport Research Review 44.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should i trust you?: Explaining the predictions of any classifier. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.

Schlögl, M., Stütz, R., Laaha, G., Melcher, M., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Accid. Anal. Prev. 127 (February), 134–149. https://doi.org/10.1016/j.aap.2019.02.008.

Schulz, A., Ristoski, P., Paulheim, H., 2013. I see a car crash: real-time detection of small scale incidents in microblogs. Ext. Semant. Web Conf.

Shan, L., Yang, Z., Zhang, H., Shi, R., Kuang, L., 2018. Predicting duration of traffic accidents based on ensemble learning. Int. Conf. Collab. Comput. Networking, Appl. Work 252–266.

Shapley, L.S., 1953. A value for n-person games. Contrib. to Theory Games. pp. 307–317.

Sharifi, M.S., Song, Z., Nasr Esfahani, H., Christensen, K., 2019. Exploring heterogeneous pedestrian stream characteristics at walking facilities with different angle intersections. Physica A: Statistical Mechanics and its Applications.

Soleimani, S., Mousa, S.R., Codjoe, J., Leitner, M., 2019. A comprehensive railroad-highway grade crossing consolidation model: a machine learning approach. Accid. Anal. Prev. 128 (April), 65–77. https://doi.org/10.1016/j.aap.2019.04.002.

Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 647–665.

Sun, Jie, Sun, Jian, 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transp. Res. Part C 54, 176–186. https://doi.org/10.1016/j.trc.2015.03.006.

Thomas, R.W., Vidal, J.M., 2017. Toward detecting accidents with already available passive traffic information. IEEE 7th Annu. Comput. Commun. Work. Conf. 1–4. https://doi.org/10.1109/CCWC.2017.7868428.

Traffic Incident Management, 2013. Fed. Highw. Adm. Traffic safety facts 2012: Young drivers, 2014. Natl. Highw. Traffic Saf. Adm.

Traffic Safety Facts FARS, 2013. Natl. Highw. Traffic Saf. Adm. (NHTSA), GES Annu. Rep.Natl. Highw. Traffic Saf. Adm. (NHTSA), GES Annu. Rep.

Vanhoeyveld, J., Martens, D., 2018. Imbalanced classification in sparse and large behaviour datasets. Data Min. Knowl. Discov. 32 (1), 25–82.

Vishnu, V.C.M., Nedunchezhian, M.R.R., 2018. Intelligent traffic video surveillance and accident detection system with dynamic traffic signal control. Cluster Comput. 135–147. https://doi.org/10.1007/s10586-017-0974-5.

Wang, J., Xie, W., Liu, B., Ragland, D.R., 2016. Identification of freeway secondary accidents with traffic shock wave detected by loop detectors. Saf. Sci. 87, 195–201. https://doi.org/10.1016/j.ssci.2016.04.015.

White, J., Thompson, C., Turner, H., Dougherty, B., Schmidt, D.C., 2011. WreckWatch: automatic traffic accident detection and notification with smartphones. Mob. Netw. Appl. 285–303. https://doi.org/10.1007/s11036-011-0304-8.

Xia, S., Xiong, J., Liu, Y., Li, G., 2015. Vision-based traffic accident detection using matrix approximation. 10th Asian Control Conf 1–5. https://doi.org/10.1109/ASCC.2015.7244586.

Xu, B., Barkley, T., Lewis, A., Macfarlane, J., Pietrobon, D., Stroila, M., 2016a. Real-time detection and classification of traffic jams from probe data. Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.

Xu, C., Assistant, P.D., Liu, P., Ph, D., Yang, B., Candidate, P.D., Wang, W., Ph, D., 2016b. Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. Transp. Res. Part C 71, 406–418. https://doi.org/10.1016/j.trc.2016.08.015.

Yishui, S., Wei, C., Hongjiang, Z., 2015. Research of highway bottlenecks based on catastrophe theory. 2015 Int. Conf. Transp. Inf. Saf. 138–142. https://doi.org/10.1109/ICTIS.2015.7232066.

Zaldivar, J., Calafate, C.T., Cano, J.C., Manzoni, P., 2011. Providing accident detection in vehicular networks through OBD-II devices and Android-based Smartphones. In: IEEE 36th Conf. Local Comput. Networks. IEEE. pp. 813–819.

Zhang, Z., He, Q., 2016. On-site traffic accident detection with both social media and traffic data. Proc. 9th Trienn. Symp. Transp. Anal. (TRISTAN) 3.

Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting tra ffi c accidents from social media data. Transp. Res. Part C Emerg. Technol. 86 (November 2017), 580–596. https://doi.org/10.1016/j.trc.2017.11.027s.