ASCE

Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques

Amin Ariannezhad, Ph.D.[1]; Abolfazl Karimpour[2]; Xiao Qin, Ph.D., P.E.[3]; Yao-Jan Wu, Ph.D., P.E.[4]; and Yasamin Salmani, Ph.D.[5]

Abstract: With a growing number of intelligent transportation system sensors and the networkwide deployment of those across the nation's roadway facilities, current research and practices should concentrate on more proactive safety strategies. In recent years, real-time traffic data collected from ITS sensors have been utilized to develop crash prediction models. Real-time crash prediction models can be used to identify hazardous traffic conditions that might cause a crash. This study aims to examine how employing data mining techniques that account for imbalanced data could improve the predictive capability of real-time crash prediction models. The term *imbalanced data* refers to a condition where the number of observations in each class is not equally distributed among the data set (noncrash cases outnumber crash cases). To decrease the within-class variation of imbalanced data, the data were split into two traffic-state data sets: free-flow speed (FFS) and congestion. Three models, including logistic regression as the baseline, random forest (RF) with random undersampling, and Adaptive Boosting (AdaBoost), were estimated with each data set. The results were compared with the models that were estimated using the complete set of data. Model comparisons indicated that all three models achieved significantly better predictive results with the congested and FFS data sets as opposed to the data set containing all crashes and that, while in some cases the results of the undersampled RF model were slightly better than those of AdaBoost, both models outperformed the logistic regression model. The results of this study demonstrated that using models to deal with imbalanced data and lowering the variation of imbalanced data could substantially improve crash prediction accuracy. The findings could help traffic agencies to practically implement and deploy crash prediction models for real-time applications and develop crash prevention strategies accordingly. **DOI: 10.1061/JTEPBS.0000499.**

Author keywords: Real-time crash prediction; Imbalanced data; Traffic conditions; Logistic regression; Adaptive boosting; Undersampling; Random forest (RF).

Introduction

Rapid growth in population and car ownership per capita significantly increases the need for safety studies. According to the National Highway Traffic Safety Administration (NHTSA), in 2019, over 36,120 people were killed in roadway crashes in the United States. In recent years, many studies have focused on microlevel safety analysis, such as crash severity (Ariannezhad et al. 2014; Ariannezhad and Wu 2018; Razi-Ardakani et al. 2014; Uddin and Huynh 2017) and crash prediction analysis (Mansourkhaki et al. 2017a, b; Eftekharzadeh and Khodabakhshi 2014), as well as macrolevel safety research (Ariannezhad et al. 2020; Lee et al. 2018), with the goal of improving roadway safety. With a growing number of intelligent transportation system (ITS) sensors and networkwide deployment of those across the nation's roadway facilities, current research and practices should concentrate on more proactive safety strategies. With traffic agencies collecting and archiving real-time traffic data, it is now feasible to develop real-time crash prediction models to identify crash precursors. Real-time crash prediction models can be used to identify hazardous traffic conditions that might cause a crash. These hazardous traffic conditions are identified by comparing the traffic flow characteristics of crash and noncrash situations. Based on these models, crash prevention strategies can be developed to help alleviate crash risks or possibly avoid a crash. While developing crash risk prediction models, three questions will be addressed: (1) What type of traffic variables should be included in the predictive model?; (2) What is the relation between crash type and severity when estimating crash risk?; and (3) What type of predictive models is appropriate for handling both real-time traffic and imbalanced crash data?

In recent years, significant efforts have been made to test different variables in crash risk prediction models. Most studies used only traffic-related variables (Abdel-Aty et al. 2008; Hossain and Muromachi 2012; Pande et al. 2011; Yu and Abdel-Aty 2013). These variables included average, standard deviation (SD), and COV of speed, volume, and occupancy before crashes. In addition to these variables, other factors, such as weather conditions (Chen et al. 2017; Lin et al. 2015; Xu et al. 2013a; You et al. 2017), visibility (Lin et al. 2015), and geometric data (Xu et al. 2013a), have also been included in predictor variables.

[1]Research Assistant, Dept. of Civil and Architectural Engineering and Mechanics, Univ. of Arizona, Tucson, AZ 85721 (corresponding author). ORCID: https://orcid.org/0000-0001-6679-7428. Email: ariannezhad@email.arizona.edu

[2]Graduate Research Assistant, Dept. of Civil and Architectural Engineering and Mechanics, Univ. of Arizona, Tucson, AZ 85721. ORCID: https://orcid.org/0000-0002-8707-6408. Email: karimpour@email.arizona.edu

[3]Professor, Dept. of Civil and Environmental Engineering, Univ. of Wisconsin-Milwaukee, Milwaukee, WI 53211. Email: qinx@uwm.edu

[4]Associate Professor, Dept. of Civil and Architectural Engineering and Mechanics, Univ. of Arizona, Tucson, AZ 85721. ORCID: https://orcid.org/0000-0002-0456-7915. Email: yaojan@email.arizona.edu

[5]Assistant Professor of Project and Operations Management, Dept. of Management, College of Business, Bryant Univ., Smithfield, RI 02917. ORCID: https://orcid.org/0000-0002-7691-7240. Email: ysalmani@bryant.edu

Note. This manuscript was submitted on October 25, 2019; approved on October 27, 2020; published online on December 29, 2020. Discussion period open until May 29, 2021; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering, Part A: Systems***, © ASCE, ISSN 2473-2907.

The majority of crash prediction models use these predictive variables to predict all types of crashes in general (Abdel-Aty et al. 2008; Ahmed et al. 2012a; Hossain and Muromachi 2012; Lin et al. 2015; Xu et al. 2013a; Yu and Abdel-Aty 2013). However, some researchers (Ahmed et al. 2012b; Pande et al. 2011) have analyzed rear-end crashes as the most common crashes on freeways. Based on the findings in Ahmed et al. (2012b), the rear-end crash model achieves an accuracy of 72%, while the generic all-crash model only identified 69% of crashes correctly. You et al. (2017) included both rear-end and side-swipe crashes in their data set to develop their crash prediction model. Chen et al. (2017) selected only lane-change-related crashes to investigate how lane-specific real-time traffic factors were associated with these crashes. Xu et al. (2013a) focused on crash risk with consideration of crash severity to link the likelihood of crash occurrence with various traffic flow characteristics at different severity levels. They considered three levels of severity based on their crash data—fatal/incapacitating injury, nonincapacitating/possible injury, and property-damage-only—and concluded that traffic flow variables contributing to crash risk were different at various severity levels.

To develop real-time crash prediction models, various parametric and nonparametric models have been used. Matched case-control logistic regression, a traditional statistical model, is one of the most commonly used techniques for comparing crash and noncrash cases (Abdel-Aty et al. 2004; Chen et al. 2017; Lee et al. 2006). Bayesian logistic regression (Ahmed et al. 2012a) and Bayesian updating approach (Ahmed et al. 2012b) are also commonly used Bayesian inference techniques in crash risk evaluation. In addition to these models, several artificial intelligence (AI) techniques have been applied. Neural network (NN)-based classification (Abdel-Aty et al. 2008; Pande et al. 2011; Pande and Abdel-Aty 2006b), support vector machine (SVM) (Yu and Abdel-Aty 2013), and genetic programming (Xu et al. 2013b) are the most commonly used AI techniques for predicting real-time crash risk. Pande et al. (2011) and Abdel-Aty et al. (2008) developed a multilayer perceptron neural network (MLPNN) modeling framework with the main goal of investigating the transferability of models developed for one corridor onto another. They concluded that the same crash risk prediction model may only work for corridors with very similar traffic conditions. Yu and Abdel-Aty (2013) criticized the linear functional form in traditional logistic regression models as well as overfitting problems in NN models, and they developed several SVM models with different kernel functions. They evaluated the effect of sample size on SVM models' predictive capability and found that such models' classification accuracy would increase if smaller sample sizes were used.

Although previously estimated real-time crash prediction models including traditional and AI models are capable of predicting crashes, several issues need to be addressed to improve the predictive capability of these models. First, in general, the data used for crash prediction are imbalanced. The term *imbalanced data* refers to a condition where the number of observations in each class is not equally distributed among the data set. In reality, the number of crashes severely outnumber noncrash instances. When data are imbalanced, depending on the severity of the class imbalance, a model's performance will be affected as a result of ignoring the minority class instances (crash cases) in favor of the majority class instances (noncrash cases). Several studies utilized imbalanced data sets to estimate crash prediction models (Basso et al. 2018; Theofilatos et al. 2018). These two studies used a linear logistic regression model and SVM model with a radial kernel to predict crash risk. One drawback of the logistic regression model is that it assumes a linear functional form between independent and dependent variables. Therefore, inefficient and biased estimations might be produced by these models when the assumption is violated (Mussone et al. 1999; Yu and Abdel-Aty 2013). In addition, dealing with an imbalanced data set, the data instances in one class greatly outnumber the instances of the other class, and classification algorithms such as SVM and NN usually tend to overpredict the majority class (Seiffert et al. 2010).

Second, most studies estimate only one model for the entire day. The findings in Xu et al. (2012) indicate that the contributing factors leading to crashes are different, in every traffic state. The average values of traffic variables, including speed, volume, and occupancy, differ significantly in different traffic states. This fluctuation in the data not only substantially increases the variation within the two classes but also increases the similarities between the classes. Therefore, one generic model for an entire day cannot accurately capture the underlying trend of the data and would not fit the data well. Moreover, few studies have considered various traffic states while developing their real-time crash prediction models. Xu et al. (2014) developed a crash risk prediction model that included various traffic conditions across multiple traffic states. The authors considered four traffic states and incorporated various crash mechanisms across these traffic states. In another study, Abdel-Aty et al. (2005) split their crash prediction model into high-speed and low-speed traffic conditions.

Though a large body of literature exists on real-time crash prediction, to the authors' best knowledge, no studies developed a real-time crash prediction model considering both imbalanced data and various traffic states in tandem. Even the existing literature that accounts for different traffic states in the prediction step (Abdel-Aty et al. 2005; Xu et al. 2014, 2012) usually ignores the fact that the crash data set is imbalanced and only assumes a linear functional form between independent and dependent variables. Therefore, the objectives of this study are, first, to investigate crash prediction accuracy using boosting and sampling techniques and compare it with the accuracy of traditional models in the context of an imbalanced data set and, second, to further examine how different traffic states (congestion versus free-flow speed) could improve models' predictive performance. The results of this study could help in the development of more accurate crash prediction models, where they could practically be used in real time. Furthermore, proactive traffic management strategies could be developed to help mitigate crash risk and improve freeway safety.

The rest of this paper is organized as follows. The next section thoroughly investigates the existing background on various crash risk prediction approaches and the major variables used in the prediction model. The next section describes the study segment and data used in this study. Then the methodology and a short description of various models and variables used in this study are discussed. Finally, model results, discussion, and conclusions are then presented.

Data Collection and Preparation

A 48-km (30-mi) long segment of the Interstate 10 (I-10) freeway in the Phoenix metropolitan area in Arizona was selected as the study corridor. This corridor is one of the major corridors carrying inbound and outbound traffic for the Phoenix Central Business District. What follows are descriptions of three data sets obtained from the Arizona Department of Transportation (ADOT):

1. Loop detector traffic data: The ADOT Traffic Operation Center (TOC) archives 20-s data from all dual loop detectors located along freeways in the Phoenix area in real time. Average speed, volume, and occupancy values at the 20-s aggregation interval were obtained from the detectors in each lane. In the study

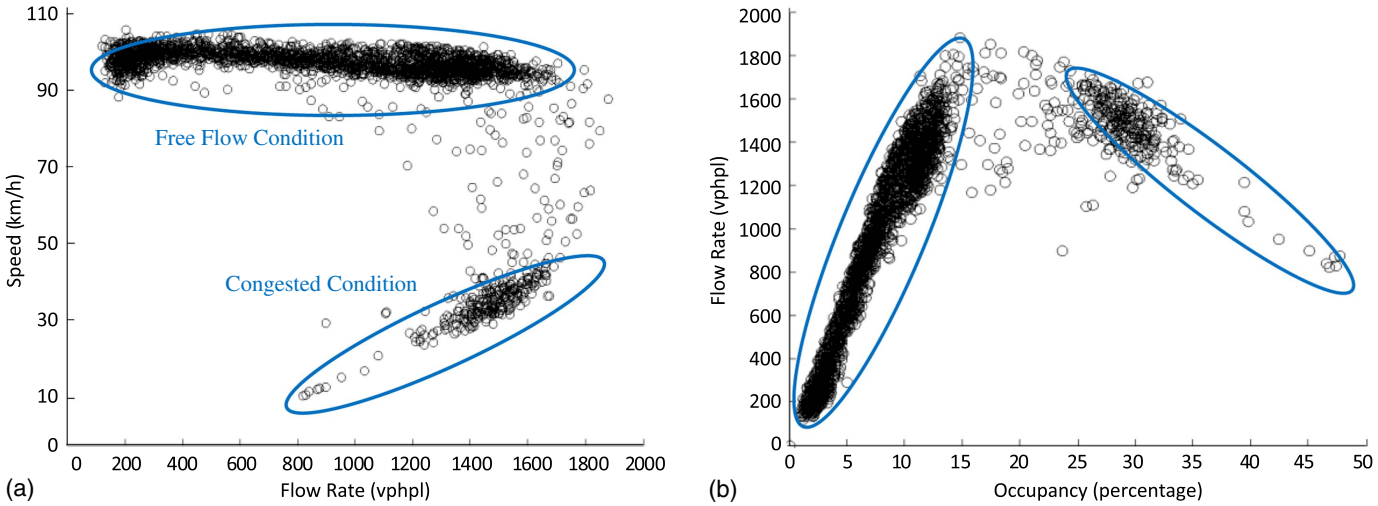

Fig. 1. Traffic flow analysis of sample data on study corridor: (a) speed-flow diagram; and (b) flow-occupancy diagram.

corridor, loop detectors are spaced 1.61 km (1-mi) apart on average in each direction.

2. Traffic crash data: All reported crashes in the state of Arizona are stored in the Accident Location Identification Surveillance System (ALISS). The ALISS database includes key information, required for this study, about crash events, including date, time, and location. The primary source of data for the ALISS database is the State Highway Log (SHL) system in Arizona. The incident reports are logged by both police officers and ADOT personnel.

The analysis done for this study included 1,984 crashes that occurred on weekdays on the study corridor in 2015. Usually, loop detectors traffic data are not perfectly clean so that they can be directly used for analytics. Therefore, generally, invalid and missing loop data need to be detected and appropriate data imputation techniques need to be applied (Ariannezhad and Wu 2020; Karimpour et al. 2019). In this study, after data preprocessing and data imputation of the erroneous and missing traffic data records, the loop detector traffic data corresponding to each crash were extracted from the traffic data set.

To examine how decreasing the variation in imbalanced data sets would affect the predictive performance of models, the final data were divided into two sets with congested and free-flow speed (FFS) traffic states before crash occurrence. It is important to note that investigation of the factors that contribute to crashes in different traffic states was not the purpose of this paper but was thoroughly discussed elsewhere (Xu et al. 2012). In this study, the fundamental traffic flow concept is used to understand various traffic states. Based on the empirical traffic data collected in the study segment, traffic flow speed and traffic flow occupancy diagrams are developed (Fig. 1). Based on a fundamental observational traffic diagram, FFS and congested conditions occur when the speed is approximately 88 km/h (55 mi/h) and less than 48 km/h (30 mi/h), respectively. The values of occupancy corresponding to these states are approximately 13% and 25%, respectively. After splitting the data, the final data sets of FFS and congested states included 941 and 532 crashes, respectively.

To develop an imbalanced data set, for each crash case approximately 200 noncrash cases were randomly selected at different times of day, on different days of the week, and on different road segments. Note that data with a ratio of 1:200 reflect the characteristics of an imbalanced data set, so an increase in the imbalance severity would not affect the distribution of the data, which makes the models computationally more expensive. This value was selected based on a suggestion from previous studies. Most of the previous literature and case studies recommended a ratio of 1:100 (Liu et al. 2006), 1:50 (Akbani et al. 2004), 1:20 (Tang and He 2017), 1:10 (Sun and Sun 2016), and 1:4 (You et al. 2017) for an imbalanced data condition. In the real world, the ratio of crash to noncrash cases is higher; however, collecting all data for every road segment is impractical, especially when the nontransferability of crash prediction models creates the need to train a model for all road corridors.

As shown in Table 1, previous studies used different numbers of loop detector stations to collect relevant traffic data for each crash. Depending on the distance between sensors and data availability, traffic sensors at crash locations or upstream and downstream of the crash location were considered for data collection. In this study, owing to the relatively large distance between loop detectors, one loop detector downstream of each crash location and one upstream were used.

Tclassify crash and noncrash data, precrash and normal conditions were defined first. Previous studies considered different definitions for precrash conditions. Some defined it as a time period starting from 5 or 10 min before a crash to crash time (Oh et al. 2005), while many studies used a time period between 5 and 10–20 min before a crash (Lin et al. 2015; Pande et al. 2011;

Table 1. Comparison of loop sensor locations used in previous studies

Reference	Upstream	Crash location	Downstream
Pande et al. (2011)	3 detectors	—	3 detectors
Hossain and Muromachi (2012)	2 detectors	1 detector	2 detectors
Yu and Abdel-Aty (2013)	1 detector	1 detector	1 detector
Xu et al. (2013a)	1 detector	—	1 detector
Lin et al. (2015) and You et al. (2017)	—	1 detector	—

Yu and Abdel-Aty 2013). In this study, 5–15 min of loop data prior to the reported crash time were extracted to avoid confusing pre- and postcrash conditions (Yu and Abdel-Aty 2013).

The polling interval of traffic data from the loop detector is 20 s. The 20-s raw data might have random noise and are usually challenging to include in the model framework (Abdel-Aty and Pemmanaboina 2006; Ahmed and Abdel-Aty 2012; Yu et al. 2018). Therefore, previous studies suggested aggregating traffic data into 5-min time intervals (Abdel-Aty and Pemmanaboina 2006; Ahmed and Abdel-Aty 2012; Liu and Chen 2017; Xu et al. 2014). It is worth mentioning that because of the high correlation between traffic data collected from adjacent lanes, aggregation was done over all lanes. Therefore, a time period of 5–15 min prior to crash time was divided into two time slices: 10–15 min and 5–10 min before a crash. The average and standard deviation of 20-s speed, volume, and occupancy were calculated for the two time slices and aggregated over all lanes. The final data set was prepared in the following format. For a crash that occurred on Thursday (March 16, 2016) at 9:35 a.m., loop detector–related variables were finalized into two time periods: 9:20 to 9:25 a.m. and 9:25 to 9:30 a.m.

A total of 24 variables for each crash and noncrash were prepared for the classification models. These variables included average and SD of speed, volume, and occupancy for two time slices and two loop stations. In summary, the final data included 1,984 crashes that occurred in the study corridor and 396,800 noncrashes. The congested and FFS data sets included 532 and 941 crashes with corresponding 106,400 and 188,200 noncrashes, respectively.

Methodology

Binary classifiers are usually judged by their accuracy or misclassification error (Mease et al. 2007). Therefore, data mining algorithms are aimed at increasing prediction accuracy by classifying the maximum number of data instances in their correct classes. A model's accuracy could be challenging when data are imbalanced as traditional algorithms tend to overpredict the majority class (Seiffert et al. 2010). This situation is prevalent in many real-world classification problems, as well as in crash prediction. Therefore, it is important to use a technique that ensures the correct identification of the instances of rare classes. A rare class is crash cases, and it is the focus of this study.

Boosting and data sampling are two methods that have been introduced for dealing with imbalanced data (Weiss 2004). One of the most commonly used boosting algorithms is Adaptive Boosting (AdaBoost) (Seiffert et al. 2010). It is one of the top 10 data mining algorithms based on its predictive capability (Wu et al. 2008). AdaBoost has demonstrated acceptable performance in the classification of imbalanced data problems. Data sampling has also proven capable of improving classification results by balancing the distribution of classes in a data set (Galar et al. 2012).

To evaluate the effect of an imbalanced data set on crash prediction model performance, three models were estimated in this study:

- Binary logistic regression: used as the benchmark model as one of the most commonly used methods in crash prediction studies. The data for this study were retrieved from a fairly homogeneous segment, in terms of the geometric conditions and number of lanes; thus, the model is not affected by confounding variables. In addition, noncrash cases were selected randomly; therefore, the binary logit model was thought to be an appropriate model compared to a conditional binary logit model with matched case control;
- AdaBoost: used to investigate boosting solutions to the imbalanced data issue; and
- Random forest with undersampling: used to investigate the sampling technique solution to the imbalanced data issue.

The results of boosting and sampling techniques will be compared with each other as well as with the benchmark model to understand the predictability of each classifier. The following subsections describe AdaBoost and RF with undersampling models. To conserve space, the well-known binary logistic regression model is not described in this section. Refer to Train (2009) for details of the model.

Random Forest with Undersampling

Data sampling techniques are data-level approaches to handling imbalanced data. These techniques are independent of the underlying classifiers and are only used for balancing the distribution of classes in the data set. It was empirically proven that using data sampling techniques to change the distribution of classes would be a positive solution to deal with imbalanced data sets (Galar et al. 2012). Random undersampling is one of the sampling techniques that creates a balanced subset of data by eliminating instances from the majority class.

While undersampling balances the distribution of classes, one of its disadvantages is losing data, especially when the imbalance between classes is severe. The training data in this situation would not be truly representative of the population. To account for this issue, the undersampling technique was integrated with the RF model. In this approach, for each tree in the RF, the majority class is undersampled to the same size as the minority class. The instances of the majority class that are not included in a given iteration would have the chance to be included in further iterations. Therefore, a large number of iterations in RF allows one to largely overcome the drawback of undersampling, which is the loss of information (Seiffert et al. 2010). Random forest, proposed by Breiman (2001), is an ensemble method that uses classification and regression tree (CART) as the base learner. This model generates different bootstrap samples of the original data set and constructs decision trees with each sample. RF changes the structure of the classification trees during the modeling process by splitting each node in the trees using a random subset of variables instead of all the variables. This strategy makes RF robust against overfitting, and the model performs very well when compared with other classifiers such as SVMs, NN, and discriminant analysis (Liaw and Wiener 2002). The steps of the RF algorithm with undersampling are described as follows:

1. Draw n (number of trees) bootstrap samples from the original data set; in each sample, the number of instances from each class equals the size of the minority class (crash cases) in the training set.
2. Grow an unpruned tree (always grown to maximal depth) with each bootstrap sample. At each node of a tree, m random predictor variables are selected to split each node of the tree. Two-thirds of the cases are used to grow the tree, and the remaining cases, called out-of-bag (OOB) data, are retained to validate the tree.
3. Predict the test data by aggregating the predictions obtained by the n trees.

The error rate in the training data is estimated as follows:

1. After each tree is grown with the two-thirds of the bootstrap sample, the OOB data are predicted by the grown tree.
2. The predictions of the OOB data from each tree are aggregated, and the error rate, called the OOB error, is calculated.

Adaptive Boosting

AdaBoost is an ensemble supervised learning algorithm proposed by Freund and Schapire (1995) that uses a whole data set to train a weak learner through several iterations. During each iteration, the strategy of the model is to focus on those instances that were misclassified during the current iteration, with the goal of correctly classifying them in the next iteration. The weights of all data instances are equal at first. After each iteration, the weights of the misclassified instances are increased while the weights of correctly classified ones are decreased. Some of the notions in the AdaBoost algorithm are listed below:

- Sequence of N labeled data instances: $(x_i, y_i), \ldots, (x_N, y_N)$, in which x_i is the vector of 24 traffic-related explanatory variables and y_i is the class label (crash or noncrash) for the data instance i;
- Distribution D over N instances, set to be uniform: $D(i) = 1/N$;
- Weak learning algorithm, CART in this study; and
- Number of iterations, T.

Given the following inputs for the model, the algorithm is described by Steps 1–3 (Freund and Schapire 1995):

Step 1: Initialize the weight vector as $W_i^1 = D(i)$ for $i = 1, \ldots, N$.

Step 2: Repeat the following steps for $t = 1, 2, \ldots, T$:

a. Compute the distribution P^t by normalizing the weights as

$$P^t = \frac{w^t}{\sum_{i=1}^{N} w_i^t} \tag{1}$$

b. Call the decision tree, provide it with the distribution P^t, and generate a hypothesis $h_t: X \rightarrow [0, 1]$, in which 1 and 0 are used for crash and noncrash cases, respectively.

c. Calculate the error of the decision tree (h_t) as

$$\varepsilon_t = \sum_{i=1}^{N} P_i^t |h_t(x_i) - y_i| \tag{2}$$

d. To update the weight vector, set the parameter β_i as a function of the error: $\beta_i = \varepsilon_t/(1 - \varepsilon_t)$.

e. Update the weight vector and set the new weight vector as

$$w_i^{t+1} = w_i^t \beta_t^{1-|h_t(x_i)-y_i|} \tag{3}$$

Step 3: Calculate the final hypothesis after T iterations. The final hypothesis uses a weighted majority vote to combine the outputs of the T decision trees, and it is calculated as

$$h_{f(x)} = \begin{cases} 1 & if \ \sum_{t=1}^{T} \left(\frac{\log 1}{\beta_t}\right) h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \log 1/\beta_t \\ 0 & otherwise \end{cases} \tag{4}$$

Variable Selection

Prior to estimating the crash prediction models, the most significant variables for each model were identified to be further used to train the models. CART is one of the most frequently used variable selection models in crash prediction (Pande and Abdel-Aty 2006a, b; Yu and Abdel-Aty 2013). However, owing to the instability of a single decision tree, RF with the Gini importance measure was alternatively used in multiple studies (Ahmed and Abdel-Aty 2012; Hassan and Abdel-Aty 2013; Pande et al. 2011; You et al. 2017). Permutation-based importance is another available variable importance measure in the RF. Both of these measures have some drawbacks. The Gini index results in biased estimation of variable ranking when explanatory variables vary in their scale of measurement or number of categories (Strobl et al. 2007), and the permutation-based measure could be affected by imbalanced data since it evaluates the importance of variables based on the error rate. Therefore, the permutation importance based on area under curve (AUC) proposed by Janitza et al. (2013) was used in this study. For imbalanced data sets, this measure shows much better performance than the regular permutation importance, while both are more reliable than the Gini importance measure (Janitza et al. 2013).

To train the undersampled RF model, the default number of variables (square root of the total number of input features) was used to split each node of the tree. Also, the number of trees was selected to be 1,000 in the train model. Most of the previous studies pointed out that traffic conditions leading to crashes are well predicted by traffic-related variables, such as average and standard deviation of speed, volume, and occupancy. Based on the most recent studies conducted on real-time crash prediction models, from 2001 to 2020, over 80% of these studies have treated traffic-related variables collected from the loop detector as predictor variables. In this study, based on the traffic data collected upstream and downstream of crash and noncrash cases, it was seen that the average and standard deviation of speed and occupancy are different in the upstream compared to the downstream. In addition, it is known that crashes are more likely to happen in high-volume traffic conditions. Furthermore, during free-flow traffic conditions, drivers are more encouraged to frequently change lane, which will result in speed variation.

After performing significant variable selection using AUC-based permutation, the results of variable selection showed a similar outcome. The SD of the volume was only significant in the congested model, speed variance (SD-speed) was a significant variable in the FFS model, and average speed and occupancy were significant in both models. These findings are also consistent with previous studies (Parsa et al. 2019; Roshandel et al. 2015; You et al. 2017). It is worth mentioning that the SD is the most commonly used measure of dispersion. Therefore, in this study, instead of variance, the SD was used. Table 2 illustrates the most important variables, with importance above average, in the congested and FFS models. Table 2 illustrates the most important variables, with importance above average, in the congested and FFS models for the RF model. In addition, the coefficient and z-value of the corresponding variables in the logistic regression model are also demonstrated in this table.

The results indicated that the most important variable for both models was the average speed at the upstream loop station 5–10 min before the crash. Other high-importance variables in the congested model were found to be different from those in the FFS model. All the variables with high importance in the congested model were related to the traffic condition of the location upstream of the crash, while in the FFS model, average traffic speed at the downstream loop station was found to be of high importance. The implication is that traffic flow characteristics leading to high crash risk are different in congestion and FFS conditions.

Model Comparison and Discussion

To evaluate the performance of models with the imbalanced data set of each traffic state, their prediction results were compared with the single model of all crashes. Three models, logistic regression, RF with undersampling, and AdaBoost, were estimated with each data set. To conduct a fair comparison between the three models, the same imbalanced data sets were used as the inputs for the logistic

Table 2. Important variables in RF model and their corresponding coefficients in logistic regression model

Important variables	Congested model			FFS model		
	RF important variable	Logistic regression coefficient	z-value	RF important variable	Logistic regression coefficient	z-value
Average speed of upstream, 5–10 min before	X	−0.04	−5.77	X	0.09	7.84
Average speed of upstream, 10–15 min before	X	−0.03	−4.01	X	−0.03	−3.79
SD-volume of upstream, 10–15 min before	X	0.11	3.75	—	—	—
SD-occupancy of upstream, 10–15 min before	X	−0.02	−3.61	—	—	—
SD-occupancy of upstream, 5–10 min before	X	0.004	1.99	—	—	—
Average occupancy of upstream, 5–10 min before	—	—	—	X	−0.08	−5.53
Average speed of downstream, 10–15 min before	—	—	—	X	−0.04	−5.00
Average speed of downstream, 5–10 min before	—	—	—	X	0.02	2.78

Note: X indicates variable is important to corresponding model.

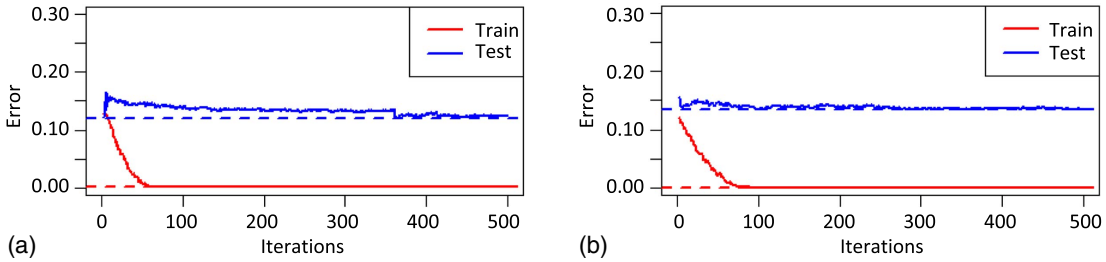

Fig. 2. AdaBoost models' train and test errors versus iteration number: (a) congested model; and (b) FFS model.

regression, RF with undersampling, and AdaBoost models. Logistic regression and AdaBoost ran on the entire data set, while RF with undersampling performed resampling at each iteration. Undersampling is an internal feature of RF that handles imbalanced data. It is worth mentioning that at each iteration, data resampling is performed on a data set similar to that of the other models. To train the models, 70% of the final data sets were randomly selected and used as the training data. The remaining 30% of the data in each data set was used to test the predictability of the models. To train the AdaBoost model, the number of iterations needs to be tuned. Fig. 2 shows the relationship between the number of iterations and the train and test errors for congested and FFS models. As the number of iterations increases, the training error decreases until it reaches zero at Iterations 60 and 80 for the congested and FFS models, respectively. Zero train error after some iterations could cause a model to overfit and affect the classification accuracy of the test data. However, the trained AdaBoost models did not overfit as the error rate for the test data did not increase after the training error reached zero. For the congestion and FFS, although the test error curve converted into a straight line after almost 360 and 260 iterations, respectively, the training and testing process was continued to 500 iterations to illustrate the model's resistance to overfitting. Such resistance to overfitting is one of the properties of AdaBoost that make it such a robust machine learning algorithm (Schapire 2013). Fig. 2 shows that the test errors did not change significantly after some iterations, and they converged to approximately 12% and 14% for congested and FFS models, respectively. Therefore, 500 iterations were considered to train all three models.

When using imbalanced data, model accuracy could be a misleading measure to evaluate prediction performance because it cannot distinguish between the number of correctly classified instances of each class (Galar et al. 2012). For instance, even if the proposed model classified all the data points as noncrashes, the accuracy of the model would be approximately 99% since 99.5% of data points are noncrashes. Therefore, to evaluate the performance of classifiers in the imbalanced data sets, generally the following measures are more appropriate than accuracy:

1. True positive rate (TP_{rate} or sensitivity): percentage of crashes correctly classified as crashes.
2. False positive rate, (FP_{rate} or false alarm): percentage of noncrashes misclassified as crashes.
3. True negative rate (TN_{rate} or specificity): percentage of noncrashes correctly classified as noncrashes.
4. False negative rate (FN_{rate}): number of crashes misclassified as noncrashes.

Generally, gaining low misclassification errors for both classes is the best achievement for a classifier. None of the four mentioned measures alone is able to completely evaluate the quality of classification for both classes. The receiver operating characteristic (ROC) curve is a criterion that combines these four measures (Bradley 1997). This curve shows the trade-off between TP_{rate} (benefit) and FP_{rate} (cost) of a model based on different probability cutoffs. ROC curves for the three models with three data sets are illustrated in Fig. 3. Figs. 3(a, c, and e) illustrate the ROC curves for the logistic regression, RF with undersampling, and AdaBoost models for the congested, FFS, and all data models, respectively. It is concluded that achieving a larger TP_{rate} in a model by changing the probability threshold would cause a higher cost by increasing FP_{rate}. However, TP_{rate} and FP_{rate} do not change by the same amount while changing the threshold. To understand the effect of threshold on both TP_{rate} and FP_{rate} at the same time, the sum of the sensitivity and specificity of the models for different thresholds is shown in Figs. 3(b, d, and f). These figures indicate that for some thresholds, the sum of specificity and sensitivity is larger. Technically, the perfect model would be one with both sensitivity and specificity equal to one. Therefore, the optimal threshold, as suggested in Youden (1950), is the one that results in the largest sum of sensitivity and specificity (the point on a ROC curve with the maximum distance to the diagonal line).

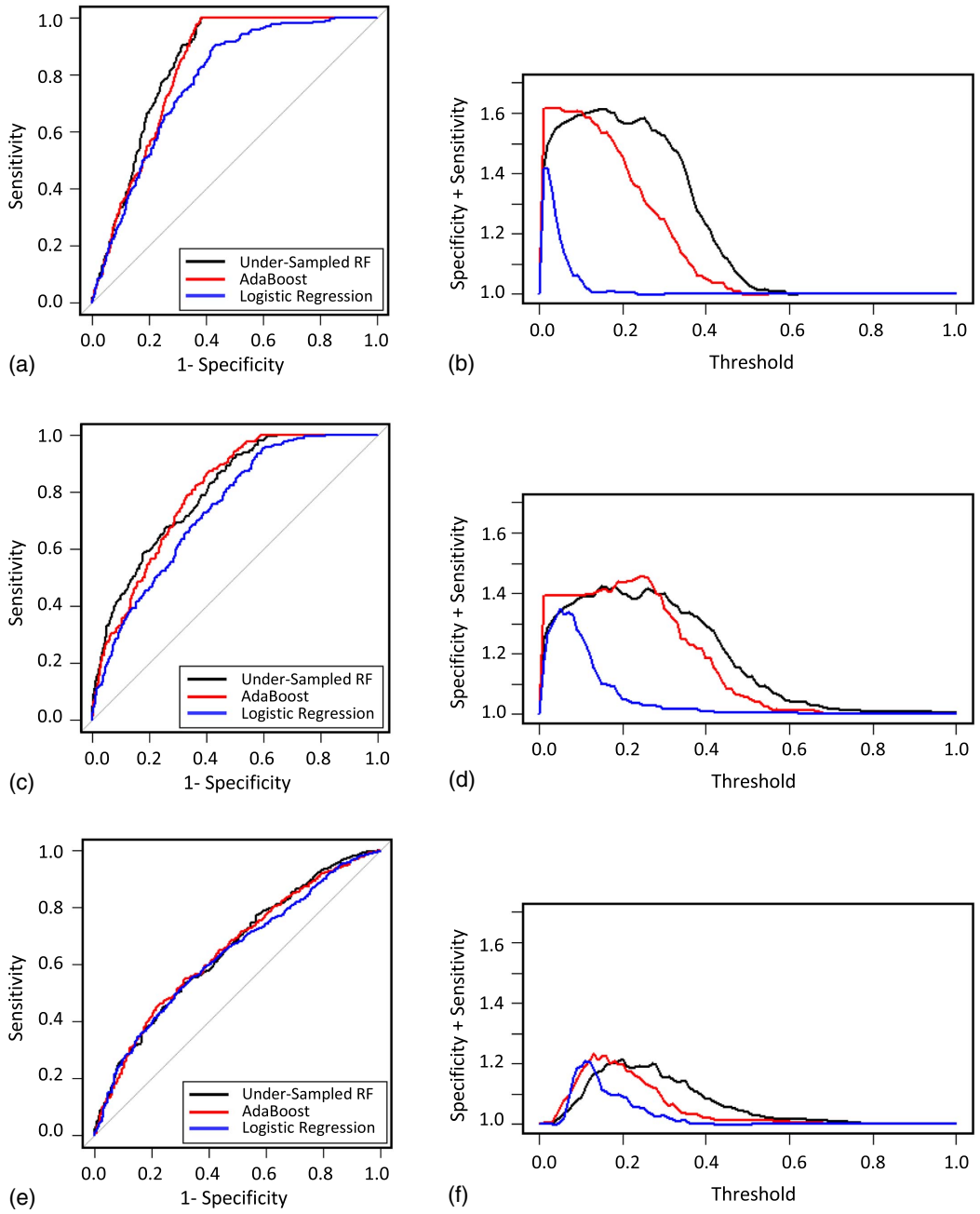

Fig. 3. (a) ROC curves for congested models; (b) summation of specificity and sensitivity versus probability threshold: congested models; (c) ROC curves for FFS models; (d) summation of specificity and sensitivity versus probability threshold: FFS models; (e) ROC curves for all data models; and (f) summation of specificity and sensitivity versus probability threshold: all data models.

The AUC is a measure of a classifier's performance and is used to determine which model is better on average (Galar et al. 2012). AUC values for all the models are shown in Table 3.

The following observations may be made by evaluating and comparing the models' prediction results:

1. Both undersampled RF and AdaBoost models outperformed the traditional logistic regression model.
2. Although the undersampled RF and AdaBoost models performed similarly with the FFS data set, undersampled RF was slightly better with congestion data, indicating that boosting could be an alternative to undersampling in the case of imbalanced data.
3. The prediction results for all three models were significantly more accurate with the congested and FFS data sets, as opposed to all the data together, implying that, when using imbalanced data, regardless of the model type, lowering the within-class variation in the data enhances a model's predictive capability.
4. Unlike the superiority of the undersampled RF and AdaBoost models in the FFS and congested data sets, these models performed only slightly better with the data set of all crashes, implying that when an imbalanced data set has a high fluctuation, regardless of the model's type, the model was barely able to identify the underlying trend in the data.
5. All three models performed slightly better with the congested data set as compared to the FFS data set. This finding is consistent with the finding in the previous study (Xu et al. 2012). The reason for this might be because during nonpeak hours, it is mainly non-traffic-related factors, such as lane change and

Table 3. Comparison of results of models based on AUC and different probability cutoffs

Traffic state	Model	Arbitrary probability cutoff		Optimal probability cutoff		AUC
		FP_{rate} (%)	TP_{rate} (%)	FP_{rate} (%)	TP_{rate} (%)	
Congested	Logistic regression	30	71	43	90	0.776
	Undersampled RF		87	39	100	0.836
	AdaBoost		82	38	100	0.819
Free-flow speed	Logistic regression	30	60	58	93	0.741
	Undersampled RF		69	50	92	0.800
	AdaBoost		73	41	87	0.796
All data	Logistic regression	30	51	34	55	0.637
	Undersampled RF		51	32	53	0.653
	AdaBoost		52	22	45	0.650

Note: AUC = area under ROC curve; FP_{rate} = false positive rate; and TP_{rate} = true positive rate.

speed variance, that contribute to crashes (Chang and Xiang 2003). Therefore, compared to the congested data set, observations with a hard-to-detect trend would be higher in the FFS data set.

The overall results revealed the importance of using techniques that account for the bias in the prediction of instances from two classes while working with an imbalanced data set. For this study, sufficient crash cases were collected for the sample data. However, because crashes are rare events, the number of crashes in a given segment could be small. Therefore, not only is the ratio of crash to noncrash cases in the sample imbalanced, but the number of minority cases is small. In such cases, selecting the right model becomes even more critical because this issue could severely affect the correct identification of rare crash cases. Like the findings in a previous study (Xu et al. 2012), it was also found in the variable selection section that the most important variables for predicting crashes in congested and FFS conditions were different. Crash prediction models should be developed separately for different traffic states. Otherwise, using all the data together would severely affect model prediction performance.

Table 3 also shows the TP_{rate} and FP_{rate} of the models for the optimal threshold. The results showed that logistic regression, undersampled RF, and AdaBoost models with FFS data sets could respectively predict 93% of crashes with a 58% false alarm rate, 92% of crashes with a 50% false alarm rate, and 87% of crashes with a 41% false alarm rate. In other words, these were the best combination of sensitivity and specificity that these models could achieve. However, based on the practical applications of the models, the FP_{rate} could be set to any value to obtain the corresponding crash prediction accuracy. To compare the relative performance of the models, sensitivity values for a sample FP_{rate} of 30% were calculated for all the models (Table 3). The best prediction results were achieved by those models developed with the congested data set. With this FP_{rate}, undersampled RF, AdaBoost, and logistic regression models could correctly predict 87%, 82%, and 71% of crashes, respectively.

In this study, along with previous studies, it was shown that using only a single model for all traffic conditions might produce bias and error in the prediction model. For the best results, dividing data into either congested and free-flow conditions or four traffic states (free flow, bunched flow, bunched congested, and standing congested traffic) is the most commonly used approach in developing real-time crash risk models. Better accuracy is achieved in such cases because the mechanisms of crash occurrence could be better captured under various traffic states. However, this could increase the cost of model development for traffic agencies. In addition, having more models to estimate could make the model development process computationally complex. Therefore, from an application point of view, one would use separate models to predict crash risk in different traffic states, which considers the estimation process cost.

From a modeling evaluation point of view, both parametric models, such as logistic regression, and nonparametric models, such as RF, have been commonly used as classifiers in the literature. While nonparametric models could result in better prediction accuracy, these models require more data to accurately estimate the mapping function between independent and response variables and are usually computationally more expensive. Real-time crash prediction models could be used by transportation agencies as a mechanism to notify drivers of hazardous traffic conditions that could lead to crashes. In this case, a model's accuracy and a low false alarm rate would be the most important factors for model selection because drivers might start ignoring these crash risk warnings if they are provided with a high number of false alarms. Therefore, nonparametric models, such as AdaBoost and undersampling RF, are the recommended models, despite their possible higher cost.

Overall, to achieve the best practical results, it is recommended not only that a model's prediction performance be considered but also that the application of the model, data limitations, and system restrictions be determined in the model selection process. Therefore, based on the results of this study, if an entity prefers to implement only one model for all crashes, as a result of either an insufficient number of crashes in each traffic state or limited budget and time, then a logistic regression model could be selected over machine learning models. The reason for this would be the simplicity, computational efficiency, and interpretability of logistic regression models, as well as their comparable results with machine learning models when estimated with all crash data.

Conclusions

Real-time crash prediction models can be used to identify hazardous traffic conditions that might result in a crash. Based on these models, crash prevention strategies can be developed to help mitigate crash risks or even avoid crashes. In this study, three models, binary logistic regression, AdaBoost, and RF with undersampling, were used to develop crash prediction models for congested and FFS traffic conditions.

The findings suggested that crash risk prediction models must be developed for each traffic condition separately. Otherwise, the high variation in traffic flow characteristics in the data could compromise the model prediction results. In addition, the results suggest that the probability threshold should be selected based on the application of the model, the importance of the correct classification of the instances of each class, and the cost associated with

misclassifying the instances of each class. In the case of real-time crash prediction, the sum of sensitivity and specificity as well as the trade-off between FP_{rate} and TP_{rate} should be considered at the same time by traffic agencies, based on practical requirements.

The outcomes of this study provide valuable insights for agencies who plan to develop and deploy real-time crash prediction models. Traffic agencies could develop proactive crash prevention strategies, such as variable speed limit (VSL), ramp metering, and dissemination of warning messages to drivers through connected vehicle technologies or variable message signs. Therefore, improving correct identification of potential crash risks could help reduce actual crashes and increase confidence in user experience. Before any countermeasures are implemented by agencies, the trade-off between crash identification rate and false alarm rate should be carefully considered based on practical application. The choices of crash prediction accuracy with the corresponding false alarm rate and their optimal combination can be aided by ROC curve analysis. In the future, this study could be extended by evaluating the transferability of the developed models to other freeway corridors. Also, to improve the accuracy of the models, other data sources, such as weather and roadway geometric data, could be integrated into the current data set. In addition, it is suggested that a more in-depth study be conducted to concentrate on evaluating the impacts of different data aggregation levels on the accuracy of crash prediction models.

Data Availability Statement

All data, models, or code generated or used during the study are confidential in nature. All these items are part of a project with the Arizona DOT, so they are not allowed to be shared.

Acknowledgments

The authors would like to thank the Arizona DOT for funding and data support. Special thanks go to Vahid Goftar and Brent Cain for their support of innovative research. The authors wish to extend their thanks to Mr. Adrian Cottam for valuable comments and proofreading.

References

Abdel-Aty, M., A. Pande, A. Das, and W. Knibbe. 2008. "Assessing safety on dutch freeways with data from infrastructure-based intelligent transportation systems." *Transp. Res. Rec.* 2083 (1): 153–161. https://doi.org/10.3141/2083-18.

Abdel-Aty, M., N. Uddin, and A. Pande. 2005. "Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways." *Transp. Res. Rec.* 1908 (1): 51–58. https://doi.org/10.1177/0361198105190800107.

Abdel-Aty, M., N. Uddin, A. Pande, F. Abdalla, and L. Hsia. 2004. "Predicting freeway crashes from loop detector data by matched case-control logistic regression." *Transp. Res. Rec.* 1897 (1): 88–95. https://doi.org/10.3141/1897-12.

Abdel-Aty, M. A., and R. Pemmanaboina. 2006. "Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data." *IEEE Trans. Intell. Transp. Syst.* 7 (2): 167–174. https://doi.org/10.1109/TITS.2006.874710.

Ahmed, M., M. Abdel-Aty, and R. Yu. 2012a. "Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data." *Transp. Res. Rec.* 2280 (1): 51–59. https://doi.org/10.3141/2280-06.

Ahmed, M., M. Abdel-Aty, and R. Yu. 2012b. "Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data." *Transp. Res. Rec.* 2280 (1): 60–67. https://doi.org/10.3141/2280-07.

Ahmed, M. M., and M. A. Abdel-Aty. 2012. "The viability of using automatic vehicle identification data for real-time crash prediction." *IEEE Trans. Intell. Transp. Syst.* 13 (2): 459–468. https://doi.org/10.1109/TITS.2011.2171052.

Akbani, R., S. Kwek, and N. Japkowicz. 2004. "Applying support vector machines to imbalanced datasets." In *Proc., European Conf. on Machine Learning*, 39–50. Berlin: Springer.

Ariannezhad, A., A. Karimpour, and Y.-J. Wu. 2020. "Incorporating mode choices into safety analysis at the macroscopic level." *J. Transp. Eng., Part A: Syst.* 146 (4): 04020022. https://doi.org/10.1061/JTEPBS.0000337.

Ariannezhad, A., H. Razi-Ardakani, and M. Kermanshah. 2014. "Exploring factors contributing to crash severity of motorcycles at suburban roads." In *Proc., 93rd Annual Meeting of the Transportation Research Board*. Washington, DC: Transportation Research Board.

Ariannezhad, A., and Y.-J. Wu. 2018. "Effects of heavy rainfall in different light conditions on crash severity during Arizona's monsoon season." *J. Transp. Saf. Secur.* 11 (6): 579–594. https://doi.org/10.1080/19439962.2018.1454561.

Ariannezhad, A., and Y.-J. Wu. 2020. "Large-scale loop detector troubleshooting using clustering and association rule mining." *J. Transp. Eng., Part A: Syst.* 146 (7): 04020064. https://doi.org/10.1061/JTEPBS.0000387.

Basso, F., L. J. Basso, F. Bravo, and R. Pezoa. 2018. "Real-time crash prediction in an urban expressway using disaggregated data." *Transp. Res. Part C: Emerging Technol.* 86 (Jul): 202–219. https://doi.org/10.1016/j.trc.2017.11.014.

Bradley, A. P. 1997. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognit.* 30 (7): 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Chang, G.-L., and H. Xiang. 2003. *The relationship between congestion levels and accidents*. Rep. No. MD-03-SP 208B46. College Park, MD: Univ. of Maryland.

Chen, Z., X. Qin, and M. R. R. Shaon. 2017. "Modeling lane-change-related crashes with lane-specific real-time traffic and weather data." *J. Intell. Transp. Syst.* 22 (4): 291300. https://doi.org/10.1080/15472450.2017.1309529.

Eftekharzadeh, S. F., and A. Khodabakhshi. 2014. "Safety evaluation of highway geometric design criteria in horizontal curves at downgrades." *Int. J. Civ. Eng.* 12 (3): 326–332.

Freund, Y., and R. E. Schapire. 1995. *A desicion-theoretic generalization of on-line learning and an application to boosting*, 23–37. Berlin: Springer.

Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. 2012. "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches." *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 42 (4): 463–484. https://doi.org/10.1109/TSMCC.2011.2161285.

Hassan, H. M., and M. A. Abdel-Aty. 2013. "Predicting reduced visibility related crashes on freeways using real-time traffic flow data." *J. Saf. Res.* 45 (Jun): 29–36. https://doi.org/10.1016/j.jsr.2012.12.004.

Hossain, M., and Y. Muromachi. 2012. "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways." *Accid. Anal. Prev.* 45 (Mar): 373–381. https://doi.org/10.1016/j.aap.2011.08.004.

Janitza, S., C. Strobl, and A.-L. Boulesteix. 2013. "An AUC-based permutation variable importance measure for random forests." *BMC Bioinf.* 14 (1): 119. https://doi.org/10.1186/1471-2105-14-119.

Karimpour, A., A. Ariannezhad, and Y.-J. Wu. 2019. "Hybrid data-driven approach for truck travel time imputation." *IET Intel. Transport Syst.* 13 (10): 1518–1524. https://doi.org/10.1049/iet-its.2018.5469.

Lee, C., M. Abdel-Aty, and L. Hsia. 2006. "Potential real-time indicators of sideswipe crashes on freeways." *Transp. Res. Rec.* 1953 (1): 41–49. https://doi.org/10.1177/0361198106195300105.

Lee, J., S. Yasmin, N. Eluru, M. Abdel-Aty, and Q. Cai. 2018. "Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed

fractional split multinomial logit modeling approach with spatial effects." *Accid. Anal. Prev.* 111 (Feb): 12–22. https://doi.org/10.1016/j.aap.2017.11.017.

Liaw, A., and M. Wiener. 2002. "Classification and regression by random forest." *R News* 2 (3): 18–22.

Lin, L., Q. Wang, and A. W. Sadek. 2015. "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction." *Transp. Res. Part C: Emerging Technol.* 55 (Jun): 444–459. https://doi.org/10.1016/j.trc.2015.03.015.

Liu, M., and Y. Chen. 2017. "Predicting real-time crash risk for urban expressways in China." *Math. Probl. Eng.* 2017: 1–10. https://doi.org/10.1155/2017/6263726.

Liu, Y., A. An, and X. Huang. 2006. "Boosting prediction accuracy on imbalanced datasets with SVM ensembles." In *Advances in knowledge discovery and data mining*, edited by W.-K. Ng, M. Kitsuregawa, J. Li, and K. Chang, 107–118. Berlin: Springer.

Mansourkhaki, A., A. Karimpour, and H. Sadoghi Yazdi. 2017a. "Non-stationary concept of accident prediction." In Vol. 170 of *Proc., Institution of Civil Engineers-Transport*, 140–151. London: Thomas Telford. https://doi.org/10.1680/jtran.15.00053.

Mansourkhaki, A., A. Karimpour, and H. S. Yazdi. 2017b. "Introducing prior knowledge for a hybrid accident prediction model." *KSCE J. Civ. Eng.* 21 (5): 1912–1918.

Mease, D., A. J. Wyner, and A. Buja. 2007. "Boosted classification trees and class probability/quantile estimation." *J. Mach. Learn. Res.* 8 (Mar): 409–439.

Mussone, L., A. Ferrari, and M. Oneta. 1999. "An analysis of urban collisions using an artificial intelligence model." *Accid. Anal. Prev.* 31 (6): 705–718. https://doi.org/10.1016/S0001-4575(99)00031-7.

Oh, C., J.-S. Oh, and S. G. Ritchie. 2005. "Real-time hazardous traffic condition warning system: Framework and evaluation." *IEEE Trans. Intell. Transp. Syst.* 6 (3): 265–272. https://doi.org/10.1109/TITS.2005.853693.

Pande, A., and M. Abdel-Aty. 2006a. "Assessment of freeway traffic parameters leading to lane-change related collisions." *Accid. Anal. Prev.* 38 (5): 936–948. https://doi.org/10.1016/j.aap.2006.03.004.

Pande, A., and M. Abdel-Aty. 2006b. "Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways." *Transp. Res. Rec.* 1953 (1): 31–40. https://doi.org/10.1177/0361198106195300104.

Pande, A., A. Das, M. Abdel-Aty, and H. Hassan. 2011. "Estimation of real-time crash risk." *Transp. Res. Rec.* 2237 (1): 60–66. https://doi.org/10.3141/2237-07.

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

Razi-Ardakani, H., A. Ariannezhad, and M. Vaziri. 2014. "Identifying factors affecting severity of urban and rural bus crashes." In *Proc., 93rd Annual Meeting of the Transportation Research Board*. Washington, DC: Transportation Research Board.

Roshandel, S., Z. Zheng, and S. Washington. 2015. "Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis." *Accid. Anal. Prev.* 79 (Jun): 198–211. https://doi.org/10.1016/j.aap.2015.03.013.

Schapire, R. E. 2013. "Explaining AdaBoost." In *Empirical inference*, 37–52. Berlin: Springer.

Seiffert, C., T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. 2010. "RUSBoost: A hybrid approach to alleviating class imbalance." *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* 40 (1): 185–197. https://doi.org/10.1109/TSMCA.2009.2029559.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC Bioinf.* 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Sun, J., and J. Sun. 2016. "Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model." *IET Intell. Transport Syst.* 10 (5): 331–337. https://doi.org/10.1049/iet-its.2014.0288.

Tang, B., and H. He. 2017. "GIR-based ensemble sampling approaches for imbalanced learning." *Pattern Recognit.* 71 (Nov): 306–319. https://doi.org/10.1016/j.patcog.2017.06.019.

Theofilatos, A., G. Yannis, P. Kopelias, and F. Papadimitriou. 2018. "Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events." *Accid. Anal. Prev.* 130 (Sep): 151–159. https://doi.org/10.1016/J.AAP.2017.12.018.

Train, K. E. 2009. *Discrete choice methods with simulation*. Cambridge, UK: Cambridge University Press.

Uddin, M., and N. Huynh. 2017. "Truck-involved crashes injury severity analysis for different lighting conditions on rural and urban roadways." *Accid. Anal. Prev.* 108 (Nov): 44–55. https://doi.org/10.1016/j.aap.2017.08.009.

Weiss, G. M. 2004. "Mining with rarity: A unifying framework." *ACM SIGKDD Explor. Newsl.* 6 (1): 7. https://doi.org/10.1145/1007730.1007734.

Wu, X., et al. 2008. "Top 10 algorithms in data mining." *Knowl. Inf. Syst.* 14 (1): 1–37. https://doi.org/10.1007/s10115-007-0114-2.

Xu, C., P. Liu, W. Wang, and Z. Li. 2012. "Evaluation of the impacts of traffic states on crash risks on freeways." *Accid. Anal. Prev.* 47 (Jul): 162–171. https://doi.org/10.1016/j.aap.2012.01.020.

Xu, C., A. P. Tarko, W. Wang, and P. Liu. 2013a. "Predicting crash likelihood and severity on freeways with real-time loop detector data." *Accid. Anal. Prev.* 57 (Aug): 30–39. https://doi.org/10.1016/j.aap.2013.03.035.

Xu, C., W. Wang, and P. Liu. 2013b. "A genetic programming model for real-time crash prediction on freeways." *IEEE Trans. Intell. Transp. Syst.* 14 (2): 574–586. https://doi.org/10.1109/TITS.2012.2226240.

Xu, C., W. Wang, P. Liu, R. Guo, and Z. Li. 2014. "Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models." *Transp. Res. Part C: Emerging Technol.* 38 (Jan): 167–176. https://doi.org/10.1016/j.trc.2013.11.020.

You, J., J. Wang, and J. Guo. 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.

Youden, W. J. 1950. "Index for rating diagnostic tests." *Cancer* 3 (1): 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Yu, R., and M. Abdel-Aty. 2013. "Utilizing support vector machine in real-time crash risk evaluation." *Accid. Anal. Prev.* 51 (Mar): 252–259. https://doi.org/10.1016/j.aap.2012.11.027.

Yu, R., M. Quddus, X. Wang, and K. Yang. 2018. "Impact of data aggregation approaches on the relationships between operating speed and traffic safety." *Accid. Anal. Prev.* 120 (Nov): 304–310. https://doi.org/10.1016/j.aap.2018.06.007.