Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 8820402, 13 pages
https://doi.org/10.1155/2021/8820402

Research Article

Crash Prediction on Expressway Incorporating Traffic Flow Continuity Parameters Based on Machine Learning Approach

Tian Lei[1], **Jia Peng**[2], **Xingliang Liu**[3] and **Qin Luo**[1]

[1]*College of Urban Transportation and Logistics, Shenzhen Technology University and Guangdong Rail Transit Intelligent Operation and Maintenance Technology Development Center, Shenzhen 518118, Guangdong, China*
[2]*Highway School, Chang'an University, Xi'an 710064, Shaanxi, China*
[3]*College of Traffic & Transportation, Chongqing Jiaotong University, Chongqing 400074, China*

Correspondence should be addressed to Jia Peng; pengjia0916@outlook.com

Received 19 September 2020; Revised 13 February 2021; Accepted 19 March 2021; Published 30 March 2021

Academic Editor: Mei Chen

Real-time crash prediction helps identify and prevent the occurrence of traffic crash. For years, various real-time crash prediction models have been investigated to provide effective information for proactive traffic management. When building real-time crash prediction model, a suitable variable space together with a specific time interval for traffic data aggregation and an appropriate modelling algorithm should be applied. Regarding the intercorrelation problem with variable space, comprehensive real-time crash prediction model considering available traffic data characteristics in applicable circumstances needs to be explored. Taking Xi'an G3001 Expressway as study area, real road traffic and accident data during the period from January 2014 to January 2019 on this expressway are applied for real-time crash prediction. To better capture traffic flow characteristics on expressway and improve the practicality of real-time crash prediction model, two new variables (segment difference coefficient and lane difference coefficient) describing the smoothness and continuity of traffic flow in spatial dimension are developed and incorporated in building the crash prediction model to solve the intercorrelation problem with variable space. Random forest (RF) is then adopted to specify the quantitative relationship between specific variable and crash risk. Real-time crash prediction model based on support vector machine (SVM) using new composed variable space is built. The results show that simplified variable space could contribute to the same classification power in currently used real-time crash prediction models compared with traditional variable space. Moreover, the prediction model based on SVM reaches an accuracy level of 0.9, which performs better than other currently used prediction models.

1. Introduction

Expressway safety has remained as a major concern in traffic system management. With higher operating speed, crashes on expressway are more likely to lead to huge life and property loss compared with other types of road [1]. According to preliminary estimates from National Highway Traffic Safety Administration (NHTSA), 36120 people died in motor vehicle crashes on highway in USA in 2019 [2]. The increasing need to reduce traffic fatalities and injuries has prompted research on proactive traffic management strategies for crash prevention. With the advancement of transportation information systems and traffic sensing technology, real-time crash prediction on expressway receives much attention from transportation professionals as it is regarded as a promising solution to road safety issues. Through predicting the time and location of possible crash occurrences in real time, proactive traffic management strategies can be applied to prevent crashes in time and improve traffic safety. Moreover, with the rapid development of autonomous vehicle techniques, it is important to accurately identify unsafe traffic condition to ensure the fast reaction of these new techniques and improve the proactive safety control of traffic systems [3].

For years, a wide array of attempts have been made for real-time crash prediction making use of multisource data.

Prediction models using traditional statistical methods such as logistic model [4] and log linear model [5] or machine learning methods such as support vector machine [6] and random forest [7] have been explored. Although statistical methods provide better interpretation for contributing factors, machine learning methods are proved to have higher prediction accuracy [8]. While developing crash prediction model, one critical step is to identify contributing factors related to the occurrence of crash. In previous studies, variables including road geometric alignment factors, traffic condition-related factors, and environment related factors were considered [9–13]. Although these variable systems cover most of factors that may contribute to crash occurrence, there still exists intercorrelation or redundant variables problem with variable space when building real-time crash prediction model. Moreover, within traffic condition-related factors, variables such as vehicle count, occupancy, and velocity are commonly adopted [14–16], while variables describing traffic flow continuity characteristics are less considered [6].

This paper therefore aims to investigate the performance of real-time crash prediction model using machine learning methods on freeways incorporating traffic flow continuity parameters considering the intercorrelation problem with variable space. Traffic and crash data collected on Xi'an G3001 Expressway from January 2014 to January 2019 are applied. Through a comprehensive analysis towards existing variables adopted in related work, two new variables' traffic flow continuity characteristics were introduced to address the previous mentioned problems. Support vector machine (SVM) is applied to build the crash prediction model considering its better performance in low data volume circumstance. Then variable selection results and crash risk prediction results are discussed and prediction results of the model built in the present work are compared with currently used prediction models by ROC curves.

2. Literature Review

In recent years, with large development in real-time road network supervision system, adequate traffic data could be used in crash prediction, and various crash prediction models have been built to provide effective information for proactive traffic management. In related works, variable space is commonly used as the prerequisite of building real-time crash risk prediction model, which is defined as a set composed of predisposing factors of traffic accidents such as vehicle counts, velocity, and weather conditions [17]. Variables adopted in related studies could be roughly divided into three categories. The first category includes average velocity, vehicle count, velocity variation, velocity standard deviation, and occupancy. These data are usually obtained from both upstream and downstream detectors of crash position [14–16], representing real-time traffic condition in a certain time interval and its dynamic characters. The second category includes differences of vehicle counts, occupancy, and velocity between adjacent lanes and so forth, which describe vehicles' lateral movements [5, 10, 18, 19]. Variables in both the first and second categories are considered as critical parameters in predicting real-time crash risk, and the importance of these variables has been proved. However, compared with variables in the first category, variables in the second category are less used due to the fact that traffic data from different lanes cannot always be specified by detectors. The third category includes vehicles' safe stopping distance [20], traffic state of free or congested flow [10, 15, 18], highway geometric alignment such as segment length and surface width [11, 21], and weather conditions [12, 22], which can be summarized as environmental or special factors that may influence traffic safety. Compared with variables in the first two categories, variables in this category are proved to be less important in forecasting real-time crash risk.

As can be concluded from related researches, variables used in real-time crash prediction models are basically selected following the comprehensive principle, and a systematical variable space does show some superiority in real-time crash risk forecasting accuracy. However, a large variable space may also lead to inefficiency in data analysis. Redundant variables may cause overfitting issues and increase the computational complexity [15]. Besides, some variables are intercorrelated, for instance, velocity, vehicle count, and occupancy, which may affect the accuracy of prediction results. Moreover, among variables adopted in existing works, variables describing traffic flow continuity are proved to be critical and significant in forecasting real-time crash risk [9, 10, 16, 19, 21, 23, 24]. Therefore, if a systematic but well-specified variable space (with inner independent variables and no redundant variables) including appropriate traffic flow continuity variables can be built, higher efficiency and practicality for real-time crash prediction can be obtained.

To forecast the real-time crash risk, traffic data collected will be aggregated using specific time interval. Together with the variable space and appropriate modelling algorithm, crash-prone traffic conditions can be successfully distinguished from normal traffic conditions [17]. The length of time interval adopted in related works varies in a wide range. Although some researchers applied highly disaggregated time intervals such as 30 seconds and 1 min, it is commonly accepted that time intervals of 5–10 minutes before the crash perform better than these highly disaggregated ones [12, 14, 15, 21]. Besides, 5–10-minute time interval is also sufficient for the traffic management center to analyze, react to, and announce warning information to the drivers [22]. However, in many countries, or expressway out of urban area, traffic data aggregated in less than 10 minutes may not be available, owing to the shortage and inferiority of traffic supervision facilities. Moreover, the reported time and location of a certain crash depend on the subjective volition of the policeman reaching the crash site [23]. As a result, there might be a larger error between reported crash time and the actual time when using a relatively smaller time interval. From another aspect, dealing with traffic data in abovementioned precision level will lead to high cost and redundant work in some circumstances. For instance, predicting crash risk in small time interval may cause frequent warning, and it is not realistic to send policemen to every potential accident point. In fact, if a larger time interval can be

applied with a reasonable prediction accuracy, the application scope of real-time crash prediction model can be broadened. Some researchers applied longer time traffic data such as 20–40 mins [15, 25] or 1 hour [26–28] before crash and also obtained reasonable prediction results. That is to say, if proper classification algorithm can be adopted when a longer time interval is applied, prediction accuracy can also be guaranteed.

On the basis of variable space, appropriate method should be applied to select significant variables when building real-time crash risk prediction model. Traditional variable selection methods can be summarized as two types: approaches based on engineering practice and statistical approaches such as logistic regression [29, 30]. With the rapid development of artificial intelligent methods, more robust and intuitive approaches such as classification tree and random forest are introduced to rank the importance of each variable [6, 24]. Though artificial intelligent methods have some natural advantages, traditional methods have not been replaced as the key point of variable selection is to make sure variables selected are independent and significant.

Once variable space is decided, real-time crash risk prediction model should be built based on appropriate algorithm. Modelling methods employed in recent decades can be divided into two categories: traditional algorithms based on mathematical statistics and modern approaches represented by artificial intelligence and data mining techniques. Typical statistical methods found in related studies mainly include matched case-control logistic regression [4, 10, 16, 18], aggregate log linear model [5], and Bayesian statistics [3, 9, 14, 19]. Algorithms based on neural networks [31, 32], fuzzy logic method [20], classification trees [33], machine learning [6, 9, 34], and deep learning [8, 35–37] are encompassed in modern methods. Regarding the intercorrelation problem of traffic variables, statistical approaches usually delete the intercorrelated variables during modelling process [14]. As a contrast, modern approaches perform better in accommodating correlated variables. However, some of the modern approaches such as neural network based modelling method have higher demand in data sources which may not always be available. Therefore, an appropriate modelling method considering the intercorrelation problem, volume of available data, calculation complexity, and predicting accuracy comprehensively is expected.

Based on all these considerations, this study attempts to explore the performance of real-time crash prediction model on freeways using machine learning methods when considering traffic flow continuity characteristics. This attempt would also address the existing intercorrelation problem with variable space and the circumstance that only longer time interval can be applied for data aggregation. This paper is organized as follows: Section 3 describes the study area and the collected data. Section 4 presents the variable selection process through a comprehensive analysis towards existing variables adopted in related work and proposing new variables. Section 5 introduces the methodology applied for building real-time crash prediction model. Section 6 summarized the model performance and modelling results. Finally, Section 7 summarizes and concludes this paper.

3. Study Area and Data Collection

In this article, Xi'an G3001 Expressway was selected as targeted road, which steps over several districts as depicted in Figure 1. As can be seen, G3001 is divided into 11 basic segments by intersected highways. Traffic detectors are deployed at middle positions of these basic segments. Detectors used for data collection are video cameras. Therefore, G3001 was divided according to positions of detectors, using adjacent detectors as the start and end points of a specific segment. In this way, segment length varied in certain range, and it would be reasonably considered as a variable. Traffic and crash data on G3001 from January 2014 to January 2019 were collected by Shaanxi Transportation Department. Data obtained from Shaanxi Transportation Department are well aggregated by vehicle type. Traffic data were given in 1 h time interval, including vehicle count in each lane, vehicle type, and average velocity. During this time period, 575 crash cases were obtained, and 110 special cases (35 drunk driving cases, 7 drug driving cases, 39 fatigue driving cases, and 29 vehicle broke down cases) considered being unpredictable were excluded. Among these data, 350 cases were randomly selected as training data and 115 cases were settled as verification data. In each of the abovementioned cases, corresponding upstream and downstream traffic data in one hour prior to the crash were extracted as crash dataset.

Besides crash dataset, noncrash dataset should also be prepared for building the crash prediction model. Noncrash dataset refers to extracted upstream and downstream traffic data in normal traffic conditions in the same road segments and also the same time periods to specific crash cases. Based on crash dataset, noncrash dataset was selected according to case-control sampling design. In this design method, the best ratio between crash data volume and noncrash data volume is 0.2 [15, 29, 38]. Therefore, for each crash dataset, five corresponding noncrash cases were randomly selected in the same segment and during time in the same month, where no crash occurred within one hour of the original crash time.

According to related works, public real-time traffic data from government department was the main data resource used in building real-time crash prediction. In the present work, traffic data and accident data obtained from Shaanxi Transportation Department and traffic data were given in 1 hour. Due to this data limitation, the raw traffic data were aggregated to 1-hour interval to obtain averages, standard deviations, and coefficient of variations prior to accident occurrence. Such intervals may be too large to capture short-term variations [26]; however, to the best of our knowledge, this is one of the first attempts to utilize traffic flow continuity description variables instead of traditional traffic condition-related variables (vehicle count, occupancy, and velocity) when building real-time crash prediction model. Traffic flow continuity characteristics in spatial dimension may not be well captured in a too-short-term situation. Besides, it can be concluded from previous studies that reasonable prediction results can be obtained even when a time interval of 1 hour is applied [26–28].

Figure 1: Research objective G3001 and detectors distribution.

4. Variable Selection

Variables used in building real-time crash risk estimation models were mainly selected based on comprehensive and practical principles. These variables should be independent from each other and comprehensive enough to accommodate all aspects related to occurrence of traffic crash. Besides, data collected from real roads vary from different resources, which were depicted as detectors. Therefore, selected variables should also be suitable for collected data. As can be concluded from previous studies, variables related to traffic condition, geometric alignment, and environment areas were always considered. Specific variables considered in related works [10, 16, 19, 21] are summarized in Table 1. Further explanation about the importance of each variable is provided as follows.

In the abovementioned studies, crash risk was analyzed using algorithms including Genetic Programming (GP) method, Binary Logit (BL) model, Multinomial Logit (MNL), model and Bayes model. The importance of certain variable towards crash risk can be obtained through the values of Gini indexes, which provides a solid basis in building variable space in this study. On the basis of comprehensive variable coverage selected from related research, variables' importance represented by Gini indexes can help further exclude the variables with no or low impact on crash risk. In the present work, 4 qualitative description indicators were applied to analyze the variables' importance in different variable space in a unified system. In the proposed evaluation system, "VI = very important" means that the variable belongs to the ranking range of 0–25%. Similarly, "IM = important" and "CO = common" represent ranking ranges of 25–50% and 50–75%, respectively. Those in ranking range of 75–100% and not selected were attributed to "NC = not chosen." While the importance of the variable in different variable space cannot be compared directly, the rank of a variable's Gini index in a certain variable space is horizontally comparable, which represents the importance of the variable. For instance, in Xu's work in 2013 [16], 28 candidate variables were considered, 12 of them were used in uncongested traffic situation, while 8 of them were applied in congested traffic situation when building estimation model cooperating Genetic Programming (GP) method. Therefore, the importance level of each variable is summarized in Table 1 by different situations and different methods.

Considering variables related to traffic conditions, occupancy and velocity were significantly important; in particular, OCC_{up}, OCC_{do}, V_{up}, V_{do}, and Std. V_{do} have the highest selection rate. Further, other variables belonging to occupancy and velocity were more or less considered in different models, except Dif. OCC_{up}. In vehicle count, only VC_{up} and VC_{do} were selected, showing less importance compared with occupancy and velocity related ones. Among geometric alignment related variables, SL was recognized as the most important one. MA and DA were also considered as relatively important. Regarding environment related variables, WC was treated as important one.

Analysis in Table 1 and in the above paragraph provides a baseline of variable selection in this research. There are two more concerns that need to be addressed when deciding the most important variables. First, as indicated in related research [16, 19, 21], the variable space needs to be simplified, since the complexity of variable space leads to heavy work load and less practicality. Moreover, while most of these works concentrated on simplifying the variables using

Table 1: Variable coverages in related works.

Major categories	Minor categories	Specific variables	Abbreviation	In GP model (uncongested)	In GP model (congested)	In BL model	In MNL model	In Bayes model
Traffic conditions	Vehicle count	Upstream vehicle count	VC_{up}	NC	**IM**	NC	**IM**	NC
		Std. dev. of upstream vehicle count	Std. VC_{up}	NC	NC	NC	NC	NC
		Difference in upstream vehicle counts between adjacent lanes	Dif. VC_{up}	NC	NC	NC	NC	NC
		Downstream vehicle count	VC_{do}	NC	NC	NC	**IM**	NC
		Std. dev. of downstream vehicle count	Std. VC_{do}	NC	NC	NC	NC	NC
		Difference in downstream vehicle counts between adjacent lanes	Dif. VC_{do}	NC	NC	NC	NC	NC
		Difference in vehicle counts between upstream and downstream	Dif. $VC_{up\text{-}do}$	NC	NC	NC	NC	NC
	Occupancy	Upstream occupancy	OCC_{up}	**IM**	NC	NC	**IM**	**IM**
		Std. dev. of upstream occupancy	Std. OCC_{up}	NC	**IM**		NC	NC
		Difference in upstream occupancy between adjacent lanes	Dif. OCC_{up}	NC	NC	NC	NC	NC
		Downstream occupancy	OCC_{do}	**IM**	**VI**	NC	**IM**	NC
		Std. dev. of downstream occupancy	Std. OCC_{do}	NC	**VI**	NC	NC	NC
		Difference in downstream occupancy between adjacent lanes	Dif. OCC_{do}	**VI**	NC	NC	NC	NC
		Difference in occupancy between upstream and downstream	Dif. $OCC_{up\text{-}do}$	**IM**	NC	NC	NC	NC
	Velocity	Upstream velocity	V_{up}	**IM**	**VI**	**CO**	**CO**	NC
		Std. dev. of upstream velocity	Std. V_{up}	NC	NC	**IM**	**IM**	NC
		Difference in upstream velocity between adjacent lanes	Dif. V_{up}	**IM**	NC	NC	NC	NC
		Downstream velocity	V_{do}	**VI**	**VI**	**CO**	**CO**	**IM**
		Std. dev. of downstream velocity	Std. V_{do}	**IM**	NC	**IM**	**IM**	**IM**
		Difference in downstream velocity between adjacent lanes	Dif. V_{do}	**IM**	NC	NC	NC	NC
		Difference in velocity between upstream and downstream	Dif. $V_{up\text{-}do}$	**IM**	NC	NC	NC	**CO**
Proportion of trucks		PT	NC	NC	NC	**VI**	NC	
Geometric alignment	Mainline	Segment length	SL	**IM**	**IM**	**VI**	NC	**CO**
		Number of lanes	NL	NC	NC	**IM**	NC	NC
		Road surface width	RSW	NC	NC	NC	NC	**CO**
		Lane width	LW	NC	NC	NC	NC	NC
		Inner shoulder width	ISW	NC	NC	NC	NC	NC
		Outer shoulder width	OSW	NC	NC	**CO**	NC	NC
		Median width	MW	NC	NC	**IM**	NC	NC
	Ramp	Merge area	MA	NC	NC	**IM**	NC	**VI**
		Diverge area	DA	NC	NC	**VI**	NC	**VI**
		Distance to merging ramp	DMR	NC	NC	**CO**	NC	NC

Table 1: Continued.

Major categories	Minor categories	Specific variables	Abbreviation	In GP model (uncongested)	In GP model (congested)	In BL model	In MNL model	In Bayes model
Environment		Weather conditions	WC	**IM**	**IM**	NC	NC	NC
		Time of day	TD	NC	NC	NC	**IM**	NC
		Peak period	PP	NC	NC	NC	NC	NC
		Velocity limit	VL	NC	NC	**IM**	NC	NC

VI = very important; IM = important; CO = common; NC = not chosen.

mathematical methods, the innercorrelations among variables were less considered. Regarding such situation, designing new parameters using existing variables can provide a reliable solution to these problems. To address these issues, variable occupancy could be used, as calculated in the following equation [39]:

$$\mathrm{OCC} = \frac{1}{L}\sum_{i=1}^{n} l_i \propto k = \frac{\mathrm{VC}}{V}. \quad (1)$$

In equation (1), l_i refers to the length of a certain vehicle in a specific road segment and L refers to the length of this segment. From this equation, definition of occupancy (OCC) follows the same baseline of density (k) definition, where k is proportional to vehicle count (VC) and inversely proportional to velocity (V). As a result, the connection among OCC, VC, and V is found, which could provide accordance when designing new variables. Usually, variables OCC, VC, and V are used to describe the real-time traffic condition (free flow or congestion) of a certain road segment [21]. As can be seen from Table 1, the standard deviations of OCC, VC, and V for upstream and downstream occupancy as well as the difference of OCC, VC, and V between upstream and downstream show great importance. Variables including Std. $\mathrm{VC_{up}}$, Dif. $\mathrm{VC_{up}}$, Std. $\mathrm{OCC_{up}}$, and Dif. $\mathrm{OCC_{up}}$ represent the traffic flow stability in a certain road segment or a cross section [6]. To be more specific, these variables mainly reflect the stability of traffic flow in spatial dimension. In spatial dimension, traffic flow stability is reflected in two directions, movements between adjacent lanes and movements between upstream and downstream in segment at certain length. According to the abovementioned contents, two new variables describing the stability of traffic flow in spatial dimension are developed taking the connection among OCC, VC, and V as reference. The first variable is segment difference coefficient (Dif. $\mathrm{DE_{up\text{-}do}}$), which describes traffic density variation along a certain segment through aggregating $\mathrm{VC_{do}}$, V_{do}, $\mathrm{VC_{up}}$, V_{up}, and SL into one variable. The other variable is lane difference coefficient (Dif. $\mathrm{DE_{do}}$), which describes traffic density variation among different lanes through aggregating VC, V, and NL into one variable. The definitions of these two variables are provided in the two following equations:

$$\mathrm{Dif.DE_{up-do}} = \left|\frac{\mathrm{VC_{do}}/V_{\mathrm{do}} - \mathrm{VC_{up}}/V_{\mathrm{up}}}{\mathrm{SL}}\right|, \quad (2)$$

$$\mathrm{Dif.DE_{do}} = \frac{1}{\mathrm{NL}} \cdot \frac{\sum_{i=1}^{\mathrm{NL}} \left|\mathrm{VC}_i/V_i - \mathrm{VC}_{i-1}/V_{i-1}\right|}{\sum_{i=1}^{n} \mathrm{VC}_i/V_i}. \quad (3)$$

In equation (3), i represents the number of a specific lane, $i \in [1, \mathrm{NL}]$. With these two variables, traffic stability can be more or less depicted, and SL is comprehensively considered. Although traffic density difference has been adopted to predict crash risk in a previous study [6], it is considered in temporal dimension. That is, traffic density difference at a certain location is calculated every five minutes. Meanwhile, in the present work, Dif. $\mathrm{DE_{up\text{-}do}}$ and Dif. $\mathrm{DE_{do}}$ are developed to describe traffic density variation in spatial dimension. To the best of our knowledge, this is one of the first attempts to consider density variation in spatial dimension when building real-time crash prediction model. According to the definitions of these two variables, Dif. $\mathrm{DE_{up\text{-}do}}$ represents the traffic condition variation along the road segment, while Dif. DEdo considers the traffic condition variation along the cross section of specific downstream section. Therefore, the two constructed variables are independent.

Once DEup-do and Dif. $\mathrm{DE_{do}}$ are applied, traffic condition variables $\mathrm{VC_{do}}$, V_{do}, $\mathrm{VC_{up}}$, V_{up}, NL, and SL used to construct new variables could be replaced. For other traffic condition variables depicting traffic flow stability applied in traditional studies [6, 21], such as Dif. $\mathrm{VC_{do}}$, Dif. $\mathrm{VC_{up\text{-}do}}$, Dif. $\mathrm{OCC_{do}}$, Dif. $\mathrm{OCC_{up\text{-}do}}$, Dif. V_{do}, and Dif. $V_{\mathrm{up\text{-}do}}$, if new designed variables $\mathrm{DE_{up\text{-}do}}$ and Dif. $\mathrm{DE_{do}}$ can be proved to have better performance in predicting crash risk than these traditional variables, it is reasonable to replace these traffic condition variables with the new designed variables, and intercorrelation problem can be addressed. Besides these two new designed variables, PT is considered additionally in the present work as it is proved to have great significance in our previous work [40] and it is not correlated to density variation variables. Apart from traffic condition-related variables, MA and DA attributed to geometric alignment and WC related to travel environment are also considered based on the significant level summarized from previous studies, as shown in Table 1. The final variable system used in this paper is depicted in Table 2.

5. Methodology

Based on available traffic data obtained on G3001, considering the variable space adopted in existing researches, a comprehensive variable system considering the importance of traffic flow continuity characteristics was built in the previous part. To build an effective crash prediction model, quantitative relationship between specific variable and crash risk needs to be specified. In the present work, RF is applied to specify the relationship mentioned above as it is commonly used in ranking the importance of each variable by Gini indexes, which is intuitive [6, 12, 21, 32]. It also has better antioverfitting ability and operational stability compared with traditional engineering practice based or statistical variable selection methods. Then real-time crash risk prediction can be interpreted as judgement upon whether traffic accident will happen in certain segment, which actually becomes a binary classification problem. When building the crash prediction model, appropriate modelling method considering the intercorrelation problem, volume of available data, calculation complexity, and predicting accuracy should be selected. As mentioned in the Data Preparation part, traffic data aggregated in small time interval are not available on G3001, which means that only lower traffic data volume can be applied to build the prediction model. To address the intercorrelation problem and at the same time considering the available data characteristics, support vector machine (SVM) is adopted to build the crash risk prediction model.

Table 2: Variable system.

Selected variables	Quantization	Abbreviation
Segment difference coefficient	Equation (2)	Dif. $DE_{up\text{-}do}$
Lane difference coefficient	Equation (3)	Dif. DE_{do}
Proportion of trucks	%	PT
Merge area	Ratio to segment length (%)	MA
Diverge area	Ratio to segment length (%)	DA
Weather conditions	Snow = 3; fog = 2; rain = 1; others = 0	WC

5.1. Random Forest. Once the variable system used for building the prediction model is specified, relationship between specific variable and crash risk will be quantified through RF. RF is a widely used machine learning method for classification and regression. Usually, in process of building classification model, variables' quantized importance will be intuitively obtained. Consider a database θ containing N records depicted as $[x_1, x_2, \ldots, x_N]$. Every record consists of an explanatory variable set $V = [V_1, \ldots, V_n]$ and a response variable V_r. To successfully predict V_r, classification tree $\widehat{f}$ known as CART was proposed. The prediction error $R(\widehat{f}, \overline{\theta})$ based on validation subset $\overline{\theta}$ is given in the two following equations:

$$R(\widehat{f}, \overline{\theta}) = \frac{1}{|\overline{\theta}|} \sum_{i \in \overline{\theta}} I\left(\widehat{f}(V_i) = V_{\text{ir}}\right), \tag{4}$$

$$I(e) = \begin{cases} 1, & \text{if } e \text{ is true,} \\ 0, & \text{if } e \text{ is false.} \end{cases} \tag{5}$$

Among the abovementioned equations, V_{ir} represents the observed value of variable V_r, corresponding to the ith record. CART mentioned above has shortage of inaccurate prediction led by small turbulence in training sample. To overcome this problem, RF was introduced [41]. In RF algorithm, the trees are formed depending on n_{RF} bootstrap samples $\overline{\theta}^1, \overline{\theta}^2, \ldots, \overline{\theta}^{n_{\text{RF}}}$ of database θ. To a specific tree, a subset of variables n_{var} is randomly chosen for splitting rule in each node. Every tree is completely grown until all of the nodes are pure, and the trees are not pruned. The resulting learning rule is the aggregation of all the tree-based estimators denoted by $f_1, f_2, \ldots, f_{n_{\text{RF}}}$ [42]. The class with the maximum number of votes among the n_{RF} trees in the forest is the predicted class of an observation.

As depicted previously, variables' quantized importance will be intuitively obtained by using RF, which is represented by Gini indexes. Split with lowest impurity at each node is selected based on Gini criterion. While forming the forest, reduction of Gini node impurity is recorded for variable $V_i \in [V_1, \ldots, V_n]$. Average of all the reductions in Gini impurity in the forest where V_i forms the split is its Gini variable importance. At last, the variables can be ranked according to the Gini variable importance measure [43].

5.2. Support Vector Machine. As mentioned at the beginning of this section, the essence of real-time crash prediction is a binary classification problem, and SVM has been proved to be effective in solving such issues [9, 24]. Previous studies integrated all variables (i.e., all traffic, geometric, sociodemographic, and trip generation variables) in SVM, which was considered deficient due to overfitting problem [9]. Meanwhile, in the present work, this issue can be tackled through using simplified and derived variables.

Real-time crash prediction using SVM can be taken as a linear separable problem, and the training set T can be defined as

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\} \in (X * Y)^l. \tag{6}$$

In the above contents, $x_i \in X = R^n$, $y_i \in Y = \{+1, -1\}$, $i = 1, \ldots, l$, and hyperplane given below could be found in n-dimension Euclidean space R^n.

$$\{x \in R^n \mid (w \cdot x) + b = 0\}, \quad w \in R^n, b \in R. \tag{7}$$

Further, the parameters w and b are proposed in equation (8), while the decision function is provided in equation (9).

$$y_i = \operatorname{sgn}((w \cdot x_i) + b), \quad i = 1, \ldots, l, \tag{8}$$

$$f(x) = \operatorname{sgn}((w \cdot x) + b). \tag{9}$$

Based on the form, along the necessary and sufficient condition of a standard hyperplane, the previously mentioned problem can be converted into an optimization issue combined with maximum spacing principle, shown in the two following equations:

$$\min_{w,b} \tau(w) = \frac{1}{2}|w|^2, \tag{10}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geq 1, \quad i = 1, \ldots, l. \tag{11}$$

To approximate the linear separable problem, relaxation variable $\xi_i \geq 0 (i = 1, \ldots, l)$ and penalty parameter $C > 0$ can be introduced to soften the restrictions and to provide a penalty to the scenario when the value of ξ_i is too large in the objective function.

Thus, equations (10) and (11) can be converted as follows:

$$\min_{w,b,\xi} \frac{1}{2}|w|^2 + C \sum_{i=1}^{l} \xi_i, \tag{12}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \; i = 1, \ldots, l. \tag{13}$$

Based on dual theory, optimization problem can be converted as a dual problem, shown in the two following equations:

$$\min_{a} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \left(x_i \cdot x_j\right) - \sum_{j=1}^{l} \alpha_j, \tag{14}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C,\ i = 1, \ldots, l. \tag{15}$$

Assuming that $\alpha^* = (\alpha_1^*, \ldots, \alpha_i^*)^T$ is an arbitrary solution to dual problem and that arbitrary solution (w^*, b^*, ξ^*) of equation (10) can be obtained using linear support vector machine, a kernel function $K(x_i, x_j)$ that can be used to create the linear mapping was introduced in equation (14). A different kernel function (linear kernel function, polynomial kernel function, radial basis kernel function, as well as sigmoid kernel function) will create a different support vector machine. The general form of the decision function is described as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i^* y_i K\left(x_i, x_j\right) + b^* \right). \tag{16}$$

In previous studies, the efficiencies of different kernel functions have been investigated [6, 9]. But unified conclusion has not been obtained yet. Therefore, all of the four kernel functions will be applied to predict the real-time crash risk, and the classification result of each kernel function will be contrasted to find the optimal algorithm.

6. Results and Discussion

In this section, significant variables' importance which is represented by Gini indexes will be obtained using RF, and quantitative relationship between specific variable and crash risk will also be verified. Real-time crash prediction models will be built through SVM using MATLAB LibSVM toolbox, based on both simplified variable space and traditional variable space. The accuracy and practicality will be further proved using comparison analysis.

6.1. Variable Significance Identification Based on RF. In the Variable Selection section, newly designed parameters Dif. $DE_{up\text{-}do}$ and Dif. DE_{do} representing traffic dynamic in crash risk prediction are provided. To further verify the influence of these two new variables on crash occurrence and specify the quantitative relationship between specific variable and crash risk, RF model is applied. Variables shown in Table 2 were taken as part of the input. Others were considered according to Table 1, including VC_{up}, V_{up}, VC_{do}, V_{do}, SL, and density in upstream (DE_{up}) and downstream (DE_{do}) representing OCC_{up}/OCC_{do}. To conduct the RF method, MATLAB RF toolbox was applied, and the results of output Gini indexes were shown in Figure 2.

As shown in Figure 2, Dif. $DE_{up\text{-}do}$ and Dif. DE_{do} ranked the first (0.97) and second (0.93), respectively, among all variables used, which is well aligned with the assumption mentioned above. Parameters V_{do} and V_{up} also show the importance of 0.89 and 0.72, which met the results in related works (16). The Gini indexes of DA, MA, and WC were 0.69, 0.65, and 0.64, respectively, which rank in the middle among all variables. For the proportion of trucks (PT), the results show that PTup (upstream) and PTdo (downstream) have the same index of 0.61, and the importance level is close to the three abovementioned variables. This result differs from the evidence found in related study, which claimed that proportion of trucks has no effect on crash occurrence [28]. Moreover, a recent study showed that consideration of vehicle type will increase the prediction power [44], which supports the decision of adopting PT in the present work. Except for DE_{do}, other traditionally used parameters were ranked in relatively lower positions, which is well aligned with the results in related works [21]. Though V_{do}, DE_{do}, and V_{up} have anterior positions, they can be represented by Dif. $DE_{up\text{-}do}$ and Dif. DE_{do}. As can be summarized from this result, the abovementioned assumption, Dif. $DE_{up\text{-}do}$ and Dif. DE_{do} are strongly related to crash occurrence, could be proved, and variables shown in Table 2 have the potential to be further used in building the SVM prediction model.

6.2. Crash Prediction Model Based on SVM. As mentioned in Section 2, 350 pieces of crash data and 1750 pieces of corresponding normal traffic data were selected as training group. Testing group consisted of 115 pieces of crash data and 575 pieces of corresponding normal traffic data were used to analyze the model accuracy. According to Section 5.1, variables shown in Table 2 compose a simplified variable space in building SVM prediction model. To further prove model performance when adopting this simplified variable space, a traditional variable space including variables shown in Figure 2 (except Dif. $DE_{up\text{-}do}$ and Dif. DE_{do}) is also applied to build SVM prediction model as a comparison. After SVM training, correct rate, which refers to the ratio between correct prediction volume and all prediction volumes in testing group, is used to describe the accuracy of prediction model. In SVM model, four kernel functions, linear kernel function (LKF), polynomial kernel function (PKF), radial basis kernel function (RBKF), and sigmoid kernel function (SKF), were used. The performance of each kernel function should also be verified [6]. Thus, correct rate of two SVM models based on each kernel function was provided in Figure 3.

As can be seen from Figure 3, the gap between correct rates of simplified variable space based SVM prediction model and traditional variable space based SVM prediction model lies in 0–5%. Therefore, the accuracies of these two SVM prediction models could be considered in the same level. Compared to traditional variable space, variables volume reduces more than 50% in simplified variable space, owing to the usage of newly designed variables (Dif. $DE_{up\text{-}do}$ and Dif. DE_{do}). The intercorrelation problem is also solved in simplified variable space, as introduced in Section 3. Moreover, the performances of LKF and PKF are basically the same, which are better than those of RBKF and SKF.

Though the accuracy of simplified variable space based SVM prediction model has been proved, it could not be concluded that this accuracy is mainly contributed by newly designed variables, Dif. $DE_{up\text{-}do}$ and Dif. DE_{do}. Therefore, we

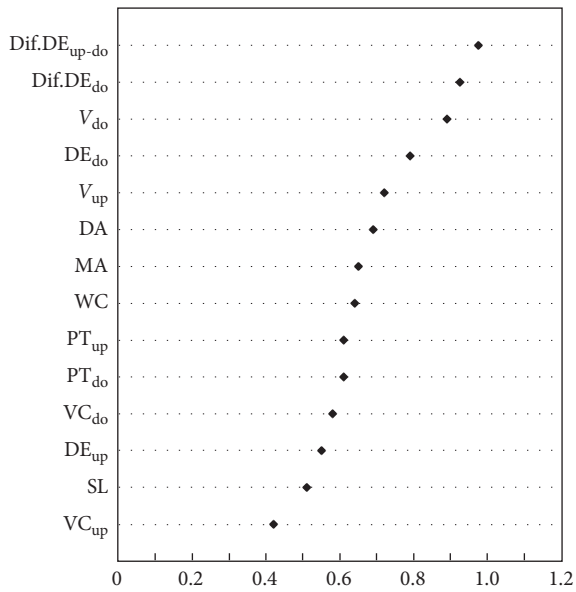

Figure 2: Output Gini indexes of selected variables.

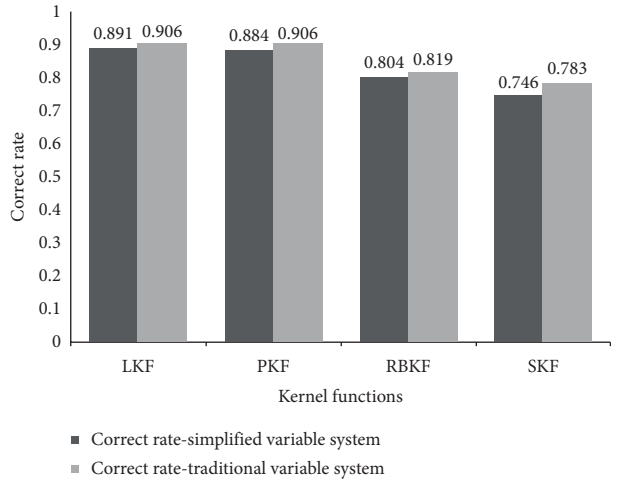

Figure 3: Correct rate of two SVM models based on each kernel function.

also choose correct rate as objective, and, using simplified variable space and the abovementioned four kernel functions based SVM, in each prediction we remove one significant variable, and corresponding correct rate could be obtained, as shown in Figure 4.

As can be seen in Figure 4, in each kernel function based SVM prediction model, Dif. DE$_{up-do}$ and Dif. DE$_{do}$ in simplified variable space have much more obvious impact on correct rate, which proved that these two newly designed variables mainly contribute to the accuracy of simplified variable space based SVM prediction model. Four other variables in simplified variable space also affect correct rate, which show basically the same importance. Results in Figures 3 and 4 proved the feasibility of simplified variable space based SVM prediction model with correct rate level of 0.90, and its classification power compared to other prediction models should also be studied.

Traditionally, AUC (the area under the curve) value of ROC (receiver operating characteristic) curve is used to describe the classification power of prediction models. In typical ROC curve, true positive rate (TPR) represents the probability of correct prediction in positive samples. False positive rate (FPR) represents the probability of mistake prediction in negative samples. ROC curve describes the relationship between TPR and FPR in specific prediction model, and AUC value is understood to better when it is closer to 1. In this research, AUC values of train group and test group were selected to analyze the classification power of simplified variable space based SVM prediction model. To make the results more reliable, previously used traditional variable space is also adopted as a comparison. Furthermore, six currently used real-time crash prediction models were chosen to specify the superiority of simplified variable space based SVM prediction model: Binary Logit (BL) model and

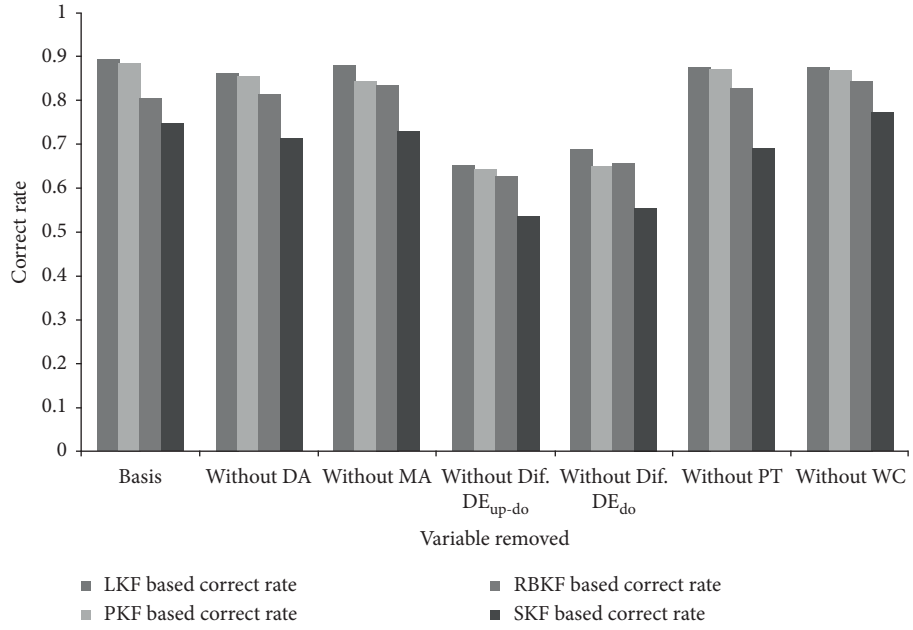

Figure 4: Impact of each significant variable on SVM prediction accuracy.

Table 3: Prediction performances of real-time crash prediction models.

Real-time crash prediction models	AUC values	
	Simplified variable space	Traditional variable space
SVM (train data)	0.975	0.977
SVM (test data)	0.952	0.961
BL (uncongested)	0.771	0.770
BL (congested)	0.712	0.705
GP (uncongested)	0.820	0.792
GP (congested)	0.754	0.773
UCC	0.897	0.903
MCC	0.722	0.755

Genetic Programming (GP) model in uncongested and congested traffic flow [21] and unmatched/matched case-control models (UCC/MCC) [19]. All AUC values in the abovementioned models are depicted in Table 3.

Based on results in Table 3, we could draw the two following conclusions. First, compared with the prediction results when traditional variable space is applied, AUC value of each prediction model remains of the same level when simplified variable space is applied (error lower than 4.37%). That is, with 50% reduction in variables volume, simplified variable space could contribute the same classification power in currently used real-time crash prediction models compared with traditional variable space. The generality and efficiency of the proposed variable space can be verified. Moreover, the intercorrelation problem can be better addressed when applying simplified variable space. Second, AUC values of SVM predictions (0.952–0.977) are larger than those of other models (0.705–0.903), which proves that the real-time crash prediction method of simplified variable space based SVM has a stronger classification power than those of other currently used models. These two conclusions both support the importance and significance of newly designed variables, segment difference coefficient and lane difference coefficient.

7. Conclusions

In this article, a real-time crash risk prediction model based on SVM was built considering the importance of traffic flow continuity parameters. To build the prediction model, data groups with one-hour time interval were selected using real road traffic and accident data from January 2014 to January 2019 on Xi'an G3001 Expressway. Based on a comprehensive analysis of previously applied variables, six important variables were selected including two newly designed variables, Dif. DE_{up-do} and Dif. DE_{do}. Method of random forest was adopted to specify the quantitative relationship between specific variable and crash risk. Based on the result, the significance of Dif. DE_{up-do} and Dif. DE_{do} was verified with high Gini indexes. The real-time crash risk prediction model was then built based on SVM LKF. The result showed that the prediction model built in the present work obtained the accuracy level of 0.9, and its feasibility and practicality were verified

by ROC curves, which showed better performance compared to other currently used prediction models. It should be noticed that the two newly designed variables proposed in the current work are applied under the circumstances of a longer time interval, and further studies should concentrate more on exploring comprehensive variables for other circumstances.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Tian Lei and Jia Peng conceptualized and designed the study; Xingliang Liu collected data; Tian Lei, Jia Peng, and Xingliang Liu analyzed and interpreted the results; Tian Lei, Xingliang Liu, Jia Peng, and Qin Luo prepared the draft manuscript. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

This research was partly supported by the Guangdong Basic and Applied Basic Research Foundation (no. 2020A1515111001), Ordinary University Engineering Technology Development Center Project of Guangdong Province under Grant no. 2019GCZX006, and Research Project of Natural Science Foundation of Guangdong Province under Grant no. 2018A030313119. The authors wish to acknowledge the Department of Transport of Shaanxi Province (DTSP), China, for sharing the traffic and accident data on Xi'an G3001 Expressway.

References

[1] J. Wang, T. Luo, and T. Fu, "Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach," *Accident Analysis & Prevention*, vol. 133, Article ID 105320, 2019.

[2] Facts + statistics: highway safety, 2020, https://www.iii.org/fact-statistic/facts-statistics-highway-safety#Motor%20vehicle%20crashes.

[3] C. Katrakazas, M. Quddus, and W. H. Chen, "A simulation study of predicting real-time conflict-prone traffic conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 10, pp. 3196–3207, 2017.

[4] A. Pande and M. Abdel-Aty, "Multiple-model framework for assessment of real-time crash risk," *Transportation Research Record*, vol. 2019, no. 1, pp. 99–107, 2019.

[5] C. Lee, B. Hellinga, and F. Saccomanno, "Real-time crash prediction model for application to crash prevention in freeway traffic," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1840, no. 1, pp. 67–77, 2003.

[6] F. Basso, L. J. Basso, F. Bravo, and R. Pezoa, "Real-time crash prediction in an urban expressway using disaggregated data," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 202–219, 2018.

[7] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 444–459, 2015.

[8] Q. Cai, M. Abdel-Aty, J. Yuan, J. Lee, and Y. Wu, "Real-time crash prediction on expressways using deep generative models," *Transportation Research Part C: Emerging Technologies*, vol. 117, Article ID 102697, 2020.

[9] L. Wang, M. Abdel-Aty, J. Lee, and Q. Shi, "Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors," *Accident Analysis & Prevention*, vol. 122, pp. 378–384, 2019.

[10] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a real-time crash risk prediction model incorporating the various crash mechanisms across different traffic states," *Traffic Injury Prevention*, vol. 16, no. 1, pp. 28–35, 2015.

[11] A. Pande and M. Abdel-Aty, "Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1953, no. 1, pp. 31–40, 2006.

[12] M. M. Ahmed and M. A. Abdel-Aty, "The viability of using automatic vehicle identification data for real-time crash prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 459–468, 2012.

[13] L. Dimitriou, K. Stylianou, and M. A. Abdel-Aty, "Assessing rear-end crash potential in urban locations based on vehicle-by-vehicle interactions, geometric characteristics and operational conditions," *Accident Analysis & Prevention*, vol. 118, pp. 221–235, 2018.

[14] M. Hossain and Y. Muromachi, "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways," *Accident Analysis & Prevention*, vol. 45, no. 2012, pp. 373–381, 2012.

[15] J. Sun and J. Sun, "A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data," *Transportation Research Part C: Emerging Technologies*, vol. 54, no. 2015, pp. 176–186, 2015.

[16] S. Yasmin, N. Eluru, L. Wang, and M. A. Abdel-Aty, "A joint framework for static and real-time crash risk analysis," *Analytic Methods in Accident Research*, vol. 18, pp. 45–56, 2018.

[17] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *Journal of Safety Research*, vol. 36, no. 1, pp. 97–108, 2005.

[18] C. Lee, M. Abdel-Aty, and L. Hsia, "Potential real-time indicators of sideswipe crashes on freeways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1953, no. 1, pp. 41–49, 2006.

[19] C. Xu, P. Liu, and W. Wang, "Evaluation of the predictability of real-time crash risk models," *Accident Analysis & Prevention*, vol. 94, pp. 207–215, 2016.

[20] C. Oh, S. Park, and S. G. Ritchie, "A method for identifying rear-end collision risks using inductive loop detectors," *Accident Analysis & Prevention*, vol. 38, pp. 295–301, 2006.

[21] C. Xu, W. Wang, and P. Liu, "A genetic programming model for real-time crash prediction on freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 574–586, 2013.

[22] L. Wang, M. Abdel-Aty, Q. Shi, and J. Park, "Real-time crash prediction for expressway weaving segments," *Transportation Research Part C: Emerging Technologies*, vol. 61, pp. 1–10, 2015.

[23] M.-I. Imprialou, "Developing accident-speed relationships using a new modelling approach," Ph.D. dissertation, Loughborough University, Loughborough, UK, 2015.
[24] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007.
[25] W. Xie, J. Wang, and D. R. Ragland, "Utilizing the eigenvectors of freeway loop data spatiotemporal schematic for real time crash prediction," *Accident Analysis & Prevention*, vol. 94, no. 2016, pp. 59–64, 2016.
[26] A. Theofilatos, "Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials," *Journal of Safety Research*, vol. 61, pp. 9–21, 2017.
[27] A. Theofilatos, G. Yannis, C. Antoniou, A. Chaziris, and D. Sermpis, "Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: a combined approach," *Journal of Transportation Safety & Security*, vol. 10, no. 5, pp. 471–490, 2018.
[28] A. Theofilatos, G. Yannis, P. Kopelias, and F. Papadimitriou, "Predicting road accidents: a rare-events modeling approach," *Transportation Research Procedia*, vol. 14, pp. 3399–3405, 2016.
[29] M. Abdel-Aty, N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia, "Predicting freeway crashes from loop detector data by matched case-control logistic regression," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1897, no. 1, pp. 88–95, 2004.
[30] A. Pande and M. Abdel-Aty, "A freeway safety strategy for advanced proactive traffic management," *Journal of Intelligent Transportation Systems*, vol. 9, no. 3, pp. 145–158, 2005.
[31] C. Oh, J.-S. Oh, S. G. Ritchie, and M. Chang, "Real-time hazardous traffic condition warning system: framework and evaluation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 3, pp. 265–272, 2005.
[32] M. Abdel-Aty, A. Pande, A. Das, and W. J. Knibbe, "Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2083, no. 1, pp. 153–161, 2008.
[33] A. Pande and M. Abdel-Aty, "Assessment of freeway traffic parameters leading to lane-change related collisions," *Accident Analysis & Prevention*, vol. 38, no. 2006, pp. 936–948, 2006.
[34] R. Yu and M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation," *Accident Analysis & Prevention*, vol. 51, no. 2013, pp. 252–259, 2013.
[35] A. Theofilatos, C. Chen, and C. Antoniou, "Comparing machine learning and deep learning methods for real-time crash prediction," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 8, pp. 169–178, 2019.
[36] C. Dong, C. Shao, J. Li, and Z. Xiong, "An improved deep learning model for traffic crash prediction," *Journal of Advanced Transportation*, vol. 2018, Article ID 3869106, 13 pages, 2018.
[37] J. Yuan, M. Abdel-Aty, Y. Gong, and Q. Cai, "Real-time crash risk prediction using long short-term memory recurrent neural network," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 4, pp. 314–326, 2019.
[38] M. A. Abdel-Aty, H. M. Hassan, M. Ahmed, and A. S. Al-Ghamdi, "Real-time prediction of visibility related crashes," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 288–298, 2012.
[39] D. C. Gazis, "The origins of traffic theory," *Operations Research*, vol. 50, no. 1, pp. 69–77, 2002.
[40] X. L. Liu, J. L. Xu, M. H. Li, L. Y. Wei, and H. Ru, "General-logistic-based speed-density relationship model incorporating the effect of heavy vehicles," *Mathematical Problems in Engineering*, vol. 2019, Article ID 6039846, 10 pages, 2019.
[41] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[42] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2013.
[43] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.
[44] F. Basso, L. J. Basso, and R. Pezoa, "The importance of flow composition in real-time crash prediction," *Accident Analysis & Prevention*, vol. 137, Article ID 105436, 2020.