Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 7833555, 19 pages
https://doi.org/10.1155/2023/7833555

Research Article

Time and Distance Gaps of Primary-Secondary Crashes Prediction and Analysis Using Random Forests and SHAP Model

Xinyuan Liu,[1] **Jinjun Tang**,[1] **Fan Gao**,[1,2] **and Xizhi Ding**[1]

[1]*Smart Transportation Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China*
[2]*Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*

Correspondence should be addressed to Jinjun Tang; jinjuntang@csu.edu.cn

Received 18 September 2022; Revised 12 December 2022; Accepted 18 March 2023; Published 14 April 2023

Academic Editor: Wen Liu

Secondary crashes (SCs) are typically defined as the crash that occurs within the spatiotemporal boundaries of the impact area of the primary crashes (PCs), which will intensify traffic congestion and induce a series of road safety issues. Predicting and analyzing the time and distance gaps between the SCs and PCs will help to prevent the occurrence of SCs. In this paper, a combined data-driven method of static and dynamic approaches is applied to identify SCs. Then, the random forests (RF) method is implemented to predict the two gaps using temporal, primary crash, roadway, and real-time traffic characteristics data collected from 2016 to 2019 at California interstate freeways. Subsequently, the SHapley Additive explanation (SHAP) approach is employed to interpret the RF outputs. The results show that the traffic volume, speed, lighting, and population are considered the most significant factors in both gaps. Furthermore, the main and interaction effects of factors are also quantified. High volume possibly promotes the time and distance gaps, while low volume inhibits them. And volume affects the distance gap inconsiderably when it falls between 300 and 400 veh/5 min. Traffic conditions with high speed and low volume are strongly associated with short-time and short-distance gaps. Darker surroundings probably accelerate the occurrence of SCs. Moreover, crashes involving the violation categories of improper turns or unsafe lane changes likely result in long time and distance gaps. These results have important implications for proposing traffic management and improving road safety.

1. Introduction

Road traffic crashes pose a threat to normal traffic operations and safety and can cause property damage or even serious injuries. According to the world health organization [1], approximately 1.3 million people die each year as a result of road traffic crashes. Between 20 and 50 million more people suffer nonfatal injuries, with many incurring a disability. Furthermore, road traffic crashes cost most countries 3% of their gross domestic product [1]. SCs, happening in the spatiotemporal impact area of primary crashes (PCs), commonly result in an additional impact on traffic and extra personal injury [2, 3]. According to [4], SCs can account for 20% of all crashes and 18% of all fatalities on freeways in the United States. In this context, SC prevention has become a major consideration in the traffic safety field.

In the past decades, a large body of literature has been devoted to investigating the identification of SCs and modeling the risk of SC occurrence [5–13]. Various statistical and machine learning (ML) methods were applied to explore these two aspects of SCs [9–12]. However, the time gap (i.e., the time difference) and distance gap (i.e., the spatial separation) between an SC and the corresponding PC have received less attention, which might hinder a better understanding of the possible time and location of SCs. Among the few methods applied to study these two gaps, statistical approaches subjected themselves to the possibility of predicting infinitely large gaps [14, 15], while ML methods

failed to provide satisfactory prediction performance on the distance gap [16]. Moreover, the black-box models need more explanation to discuss the effects of contributing factors in detail [16]. Therefore, some promising methods and data experiments are required.

To better capture the characteristics of SCs, we first developed a hybrid method (i.e., static spatiotemporal threshold-based and speed contour map-based methods) to identify SCs and obtain the time and distance gaps. Subsequently, random forests (RFs) were used to predict the time and distance gaps, which have high prediction performance and diversity. And an interpretation technique, namely the SHapley Additive explanation (SHAP) approach, was applied to examine the model outputs and estimate the global and local effects of the influencing factors. Understanding time and distance gaps and their influencing factors can provide management strategies for transportation agencies and improve traffic operations and road safety.

2. Literature Review

2.1. Secondary Crash Identification. Overall, two types of methods, static and dynamic methods, were widely used to identify SCs. Static methods identify SCs by setting the fixed spatiotemporal thresholds, which means crashes are identified as SCs if they fall within the spatiotemporal thresholds of another crash [17]. First introduced this method and defined the thresholds equivalent to one mile upstream of a PC and 15 minutes after clearance time. Following this study, further research associated with static methods has been explored [5–7, 18]. For example, some studies proposed a spatial threshold of 2 miles and time thresholds of 2, 1, and 2 hours, respectively, to identify California secondary crashes [19–21]. SCs can be selected quickly and effectively from massive crashes according to spatiotemporal thresholds [2, 16]. However, static methods have the problem of subjective judgment: overestimation or underestimation of the thresholds [2, 22]. As an improvement [7], we introduced three sets of spatiotemporal thresholds to identify SCs on Florida interstates. The spatial thresholds for all three sets were 2 miles, and the time thresholds were 2 h, 15 minutes, and 30 minutes after the PCs' clearance time. Their results confirmed that the identification ratio of SCs varied for different sets.

With the support of various sensor technologies, dynamic methods are becoming increasingly popular and used because of an improvement in the misclassification of SCs [22]. There are three main dynamic methods: (a) queuing theory-based method [23, 24]; and (b) shockwave-based approaches [25, 26]; (c) speed contour map-based method [11, 13, 18]. In practical application, due to the data quality and quantity requirements of methods (a) and (b), the models are often simplified and set assumptions, failing to reflect the actual condition in the real world. Nevertheless, the speed contour map-based method has performed well without any simplification or assumptions since it can accurately capture the impact area of PCs [13, 27, 28]. For example, [18] compared the crash state speed with the historical average speed to brighten the impact area. Likewise [11, 13], we applied this method to identify SCs and considered recurrent congestion.

In summary, static methods are easy to implement and quickly obtain identification results, while dynamic methods achieve better performance but consume a lot of computational time. Combining these two methods for SC identification can improve efficiency and accuracy [16, 25]. This paper proposes a two-stage strategy to identify SCs by incorporating the fixed spatiotemporal threshold-based and speed contour map-based methods.

2.2. Secondary Crash Risk Modeling and Predicting. Several statistical and ML models have been applied to explore the relationship between SC occurrences and contributing factors [9–12]. For example, [10] proposed a logit model to predict SC likelihood, and their results revealed that rear-end crashes could increase the SC likelihood [11] developed a random effects logit model to link the probability of SCs with real-time traffic volume conditions, primary crash characteristics, environmental conditions, and geometric characteristics. Similarly, [29] used the Bayesian complementary log-log model to predict the likelihood of SCs and examine their relationship with several variables.

However, previous studies focused less on the time and distance gaps between the SCs and PCs. Several studies have made attempts using regression approaches. For example, [14] selected the ordinary least-squares (OLS) regression to model the two gaps separately. Their results showed that time and distance gaps were closely associated with collision type and the duration of the primary crash. Likewise, [15] applied OLS regression to evaluate the relationship between the time and distance gaps concerning individual crash characteristics. They found that the number of lanes, total vehicles involved in the crash, morning time, and AADT were the most significant factors affecting time and distance gaps. Although most independent variables had a high significance, traditional statistical models usually made more prior assumptions for input variables, and they were unable to predict the possibility of massive gaps. Moreover, [14, 15] built an independent regression model for the time and distance gaps, ignoring the potential correlation of the two gaps because they happen at the same time. Therefore, it is necessary to consider an alternative model to investigate gaps simultaneously.

By contrast, ML methods have become increasingly attractive and have gained more attention due to their high prediction power and low limitation on data [30]. Multiple ML methods have been employed in traffic safety studies [8, 13, 16, 29], such as neural network models, genetic algorithms, random forests, XGBoost, etc. In a small number of studies on the time and distance gaps [16], the authors utilized a linear regression model and two ML algorithms, including a back-propagation neural network (BPNN) and the least-squares support vector machine (LSSVM), to build three prediction models. The results indicated that the BPNN and LSSVM models outperformed the linear regression model, but these two ML models also failed to

provide adequate performance on distance gap prediction. Regarding ML models, many other promising approaches, such as ensemble algorithms, combine several base learners to enhance the prediction performance [31–33].

Besides, relatively fewer studies have focused on SC prevention. As [2] summarized, available data and high costs have limited relevant investigations, so continued endeavors are still needed. The main objective of this study is to develop a reliable model to predict the time and distance gaps and analyze associated influencing factors, which can help with proactive prevention and improve safety. Several existing research gaps and insufficiencies were mitigated and supplemented in this study.

3. Data Preparation

In this study, crash data were collected from the Statewide Integrated Traffic Records System (SWITRS), which records detailed description of crash-related information, such as the unique case identifier, location (state route, postmile, latitude and longitude), collision year and time, collision severity and type, lighting, weather, etc. A total of 24643 crashes were collected from freeways I-10, I-5, US-101, I-210, and I-110 in Los Angeles County of California over four years, from June 2016 to December 2019 [34]. Through a detailed examination, we removed the issues of redundant attributes and missing values from the crash data.

In order to combine real-time traffic data into the analysis of crashes, volume, and speed were extracted from the caltrans performance measurement system [35]. In PeMS, data were gathered from a set of loop detectors on the road and transmitted to the management center for storage. And the configuration information of the detector was integrated, including the location and unique identification number. A two-step matching strategy is devised to obtain traffic volume and average speed for each crash. The first step matches the nearest detector upstream for every crash based on the latitude and longitude of the crashes and the loop detectors. The second step is extracting the volume and speed for 5 minutes before the crashes.

Referring to the previous studies on SCs [14, 16], 17 variables were selected from 4 dimensions. Specifically, temporal characteristics consist of 5 variables, namely, peak, weekend, weather, lighting, and population, which reflect the environment's state. Population density has a relationship with vehicle trips [36, 37]. Primary crash factors include 8 variables: collision severity, collision type, violation category, part count, etc. These variables demonstrate all the information associated with a crash. Road condition and surface reflect the roadway characteristics, including whether the pavement is a maintenance area or free from abnormal conditions or whether the pavement is dry/wet. Traffic volume (veh/5 min) and speed (mile/h) report the traffic characteristics. Detailed descriptions and statistical information are expressed in Table 1. Additionally, the Pearson correlation coefficients (PCCs) were applied to examine the multicollinearity between the 17 variables. Figure 1 demonstrates the computed results. As shown, all the absolute values of PPC are less than 0.8, indicating a low linear correlation between variables.

4. Methodology

4.1. SC Identification. The identification of SCs is the basis for conducting SC modeling and analysis. The static spatiotemporal threshold-based estimation is the first stage to identify SCs roughly, and it can be defined in the following equation:

$$\mathrm{SC} = \begin{cases} 1, & if\left[t_B \in \left(t_A, t_A + t_{\text{threshold}}\right)\right] \cup \left[S_B \in \left(S_A, S_A + S_{\text{threshold}}\right)\right], \\ 0, & \text{others}, \end{cases} \tag{1}$$

where (t_A, S_A) denotes the location and occurrence time of the crash A, (t_B, S_B) denotes the location and occurrence time of another crash B that needs to be examined, $(t_{\text{threshold}}, S_{\text{threshold}})$ denotes the defined time threshold and spatial threshold, and the value of 1 means that crash B is identified as a secondary crash corresponding to crash A and 0 otherwise.

Speed contour map-based method estimates the impact area of the PC based on the change in traffic speed, and a SC is identified when it is discovered in this area. The speed contour map comprises grid cells split by defined time intervals and the milepost of sensor stations [2]. The impact area can be ascertained by checking the speed of each cell near the crash. In general, it can be written as the following equation:

$$V^b_{(t,S)} = \begin{cases} 1, & \text{if } V_{(t,S)} < V^r_{(t,S)}, \\ 0, & \text{others}, \end{cases} \tag{2}$$

where $V_{(t,S)}$ and $V^r_{(t,S)}$ denote the current and the reference speed of one cell; $V^b_{(t,S)} = 1$ denotes that the cell is affected; and $V^b_{(t,S)} = 0$ denotes that the cell is not affected. The size of the impact area was determined by the reference speed $V^r_{(t,S)}$. The detailed procedures of the identification method are as follows:

(i) Apply the fixed spatiotemporal thresholds to identify the candidate SCs. Referring to previous studies on SC analysis in California [19–21], 2 miles and 2 hours were selected as the thresholds in this study. The initial identification on 24,643 crashes has yielded 563 possible SCs.

(ii) Extract the 5-min speed data to develop a speed contour map for a potential PC. More specifically, given the fixed spatiotemporal thresholds that have been determined, the time period for extracting speed data is between 2 hours before and 2 hours

Table 1: Description of variables used in crash analysis.

Variables	Types	Description	Count	Percent	Mean	Std
Temporal characteristics						
Peak	Binary	0 = no	261	70.9	—	—
		1 = yes (7 : 00–9:00 or 17 : 00–19 : 00)	107	29.1		
Weekend	Binary	0 = no	259	70.4	—	—
		1 = yes	109	29.6		
Weather	Categorical	0 = clear	306	83.2	—	—
		1 = cloudy	46	12.5		
		2 = rainy	16	4.3		
Lighting	Categorical	0 = daylight	217	59.0	—	—
		1 = dusk-dawn	17	4.6		
		2 = dark-streetlights	92	25.0		
		3 = dark-no streetlights	42	11.4		
Population	Categorical	0 = incorporated (less than 25,000)	10	2.7	—	—
		1 = incorporated (25,000–100,000)	93	25.3		
		2 = incorporated (100,000–250,000)	67	18.2		
		3 = incorporated (over 250,000)	188	51.1		
		4 = unincorporated (rural)	10	2.7		
Primary crash characteristic						
Collision severity	Categorical	0 = fatal	3	0.8	—	—
		1 = severe injury	98	26.6		
		2 = other visible injury	15	4.1		
		3 = complaint of pain	252	68.5		
Collision type	Categorical	0 = head on	2	0.5	—	—
		1 = sideswipe	53	14.4		
		2 = rear-end	242	65.8		
		3 = broadside	11	3.0		
		4 = hit object	45	12.2		
		5 = overturned	12	3.3		
		6 = vehicle/pedestrian	3	0.8		
Violation category	Categorical	0 = alcohol or drug	22	6.0	—	—
		1 = unsafe speed	247	67.1		
		2 = following too closely	4	1.1		
		3 = unsafe lane change	44	12.0		
		4 = improper turning	38	10.3		
		5 = other	13	3.5		
Party count	Discrete	Counting total parties in the collision	—	—	—	—
		0 = 1 party	44	12.0		
		1 = 2 parties	202	54.9		
		2 = 3 parties	87	23.6		
		3 = 4 parties	29	7.9		
		4 = 5 parties	4	1.1		
		5 = 6 parties	2	0.5		
Tow away	Binary	0 = no	128	34.8	—	—
		1 = yes	240	65.2		
Truck involved	Binary	0 = no	343	93.2	—	—
		1 = yes	25	6.8		
Hit and run	Categorical	0 = felony	29	7.9	—	—
		1 = misdemeanor	14	3.8		
		2 = no hit and run	325	88.3		
Alcohol involved	Binary	0 = no	333	90.5	—	—
		1 = yes	35	9.5		
Roadway characteristic						
Road condition	Binary	0 = construction or repair zone	23	6.2	—	—
		1 = no unusual condition	345	93.8		
Road surface	Binary	0 = dry	335	91.0	—	—
		1 = wet	33	9.0		
Traffic characteristics						
Volume (veh/5 min)	Continuous	Vehicle counts over the 5 minutes period preceding PCs	—	—	369.73	158.82
Speed (mile/h)	Continuous	Vehicle speed over the 5 minutes period preceding PCs	—	—	48.22	17.16

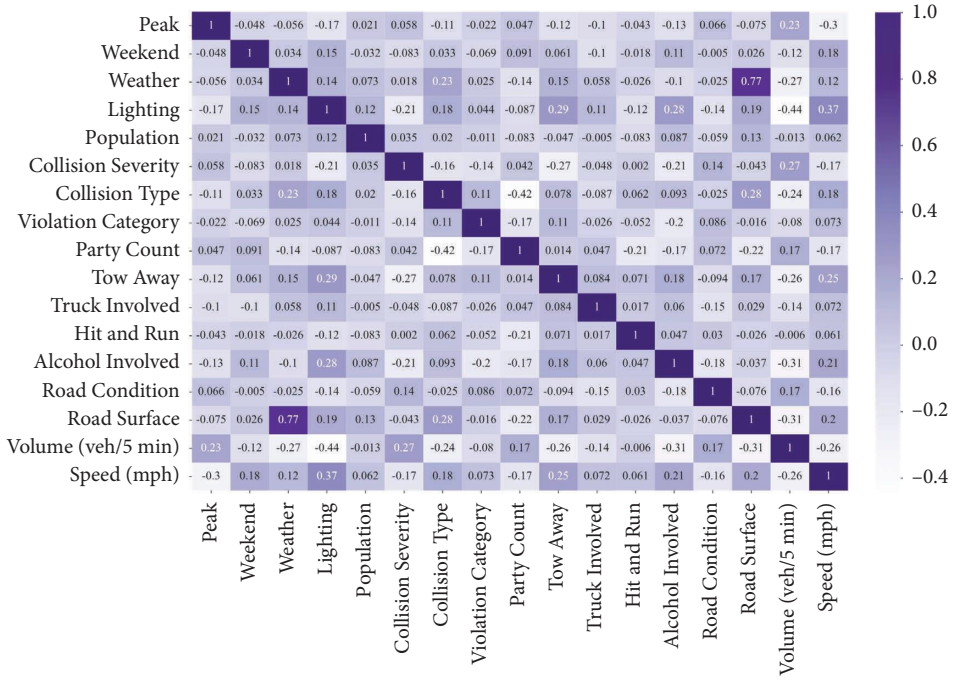

Figure 1: Pearson correlation coefficients of variables.

after the PC, and the spatial period is 2 miles upstream and 2 miles downstream of the PC location. To eliminate the effects of recurrent congestion, the historical average speed was calculated by collecting speed data from the PC-free days in a year [13, 18].

(iii) Estimate the impact area of a potential PC using equation (2). The crashes that occur in the impact area of PC are identified as SCs.

Following the two-stage identification method, 368 SCs are identified in this study. The ratio of the number of SCs to the number of all crashes is 1.49%, which is consistent with the findings of the references in this area that this ratio is around 1–1.6% [11–13, 18, 25, 38–40].

4.2. Random Forests. This study used RF to predict the time and distance gaps, which has been widely used in the transportation field [41–46]. RF uses a bootstrap sampling method to change the training set to build an integration of regression trees [47]. Such a mechanism expresses the following advantages: gaining higher performance. Furthermore, RF can perform multiple output modeling [48, 49], which is suitable for simultaneously predicting the time and distance gaps.

The input vectors for the RF model are represented as $\{\boldsymbol{x} = [x_{i1}, x_{i2}, \ldots, x_{i,M}], \boldsymbol{y} = [y_{i1}, y_{i2}]\}, i = 1, 2, \ldots, N$. M and N are the number of features and samples, y_{i1} and y_{i2} indicate the time gap and the distance gap of sample i, respectively. Figure 2 expresses the structural framework of RF, which consists of the following three parts: (1) Sample set selection: using the resampling method p times on the original dataset to generate a sample set. In other words, some samples are likely to be chosen multiple times, while others will not be selected once. After k rounds of extraction, k new sample sets are obtained. (2) Decision trees generation: training k decision trees using k sample sets of data. During each round of generating trees, m variables from M ($m < M$) features are selected for training. The randomness of the training data and variable combinations improves the prediction performance of the model and essentially prevents overfitting. (3) Result combination. Since the decision trees generated are independent, they have the same contribution to the predicted result. Therefore, the final result is obtained by averaging the k predicted results. For multioutput problems, the following changes are required in the decision trees: First is to store several output values instead of 1. Then use splitting criteria that calculate the average reduction across all outputs.

4.3. SHAP Method. ML methods commonly demonstrate an outstanding prediction performance, while their abilities are limited due to their low interpretability. Although the RF model can obtain global explanations (i.e., the relative importance), it cannot quantify local explanations for individual predictions. Nevertheless, local explanations provide more detailed information than global ones [50, 51]. Shapley additive explanations (SHAP) technology is a representative local interpretation method that can explain the main local effects and interaction effects of independent variables on dependent variables, as proposed by [52]. Furthermore, [53] improved SHAP to better and faster explain tree-based ML models, such as random forests and gradient boosted trees.

SHAP value is the core of the method which is computed based on the game-theoretic approach, and it represents the average marginal contributions of one variable on a single prediction. SHAP value is defined as the following equation:

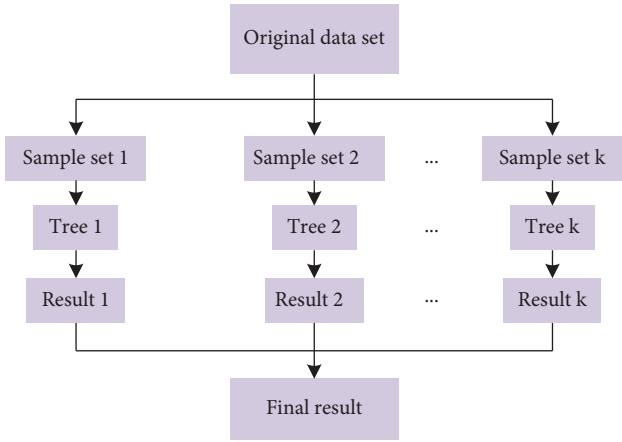

Figure 2: Structural framework of RF.

$$\phi_i(f,x) = \sum_{R\in\mathscr{R}} \frac{1}{M!}\left[f_x\left(P_i^R \cup i\right) - f_x\left(P_i^R\right)\right], \tag{3}$$

where $\mathscr{R}$ indicates the set of all variable orderings, P_i^R represents the set of all variables that rank before the variable i in the ordering R, M is the number of variables, x means the values of explanatory variables, and f_x refers to the single prediction, which can be written by the following equation:

$$f(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f,x), \tag{4}$$

where $\phi_0(f)$ means the base value, i.e., the average value of overall predictions.

Additionally, the global importance of variables is the sum of the contribution of one variable on all predictions, which is calculated by averaging absolute SHAP values as shown in the following equation:

$$I_i = \frac{1}{n}\sum_{j=1}^{n}\left|\phi_i^{(j)}\right|, \tag{5}$$

where I_i represents the importance of variable i, $\phi_i^{(j)}$ indicates the SHAP value for variable i in the single prediction j, and n is the number of all predictions.

The proposed RF model and SHAP method were mainly implemented in Python (3.8.8) using scikit-learn (0.24.1) and shap (0.40.0). The SHAP package contains three applications: force plot, summary plot and dependence plot. In this study, we apply the summary plot to describe the importance of each variable and the dependence plot to reflect the main effects and the interaction effects of all variables.

5. Results and Discussion

5.1. Results. In this study, the grid-search with 5-fold cross-validation techniques (i.e., GridSearchCV) was used to determine the core parameters of the RF model. Table 2 reports the optimal values of the parameters. In the application, the proposed RF model is compared with two traditional multivariate models: the K-nearest neighbor (KNN) model and the multilayer perceptron regression (MPR) model. All the models were trained and validated by applying the same dataset to guarantee the reliability of the comparison results. Specifically, at a ratio of 7 : 3, the raw samples were split into a training set and a testing set for training and testing model. Two classical regression evaluation measures, namely, mean absolute error (MAE) and mean squared error (MSE), were used to assess model performance. The final evaluation results are presented in Table 3. As shown, the RF model mostly outperformed the other two models on both the training and testing sets in terms of predicting the time and distance gaps.

Table 2: Optimal values of parameters of the RF model.

Parameters	Values
n_estimators	110
max_depth	10
max_features	"auto"
min_samples_split	2
min_samples_leaf	1

5.2. Global Importance of Variables. Figure 3 visualizes the global importance of variables on the time gap. In the left part, variables are sorted in descending order according to their global importance, computed by averaging their absolute SHAP values per variable. The left x-axis indicates the mean (|SHAP value|). As shown, lighting is the most dominant variable on the time gap, and its average effect on the predicted value is 0.11, followed closely by volume and speed, which change the predicted value by 0.093 and 0.056, respectively, on average. It suggested that the traffic characteristics significantly affect the time gap. This finding is not surprising; Traffic characteristics are the direct response of the traffic state, which largely influences the travel surroundings and driver status. As [11] indicated, more than geometric characteristics and primary crash characteristics, traffic characteristics could significantly affect the SC likelihood. Subsequently, population has a greater contribution than party count and collision severity, indicating that the temporal characteristic of population impacts the time gap more than the primary crash characteristic. By contrast, the roadway characteristics of road surface and condition have a substantially minor effect on the time gap, with the mean (|SHAP value|) less than 0.005.

In the right part, the diagram consists of points representing the samples, and the color visually reveals the magnitude of variables (red means a high value, while blue means a low value). The right x-axis indicates the SHAP value, which refers to the effects of all variables on a single model output (i.e., the local effect). This diagram roughly illustrates the variation of effects with the change of either variable. Taking lighting as an example, its left side of the vertical axis is covered with red points (indicate dark) and its right side is stacked with blue points (refer to daylight). This demonstrates that night may decrease the time gap, while the daytime probably promotes the time gap. In addition, high volume (red points) mostly has a positive SHAP value and

Table 3: Results of several models.

	Time gap				Distance gap			
	MAE		MSE		MAE		MSE	
	Training set	Testing set	Training set	Testing set	Training set	Testing set	Training set	Testing set
RF	**0.22**	**0.46**	**0.07**	**0.31**	**0.45**	**0.45**	0.33	**0.31**
KNN	0.44	0.47	0.28	0.32	0.49	0.47	0.36	0.36
MPR	0.45	0.46	0.30	0.32	0.46	0.47	**0.31**	0.32

Bold values refer to the maximum prediction performance in each circumstance.

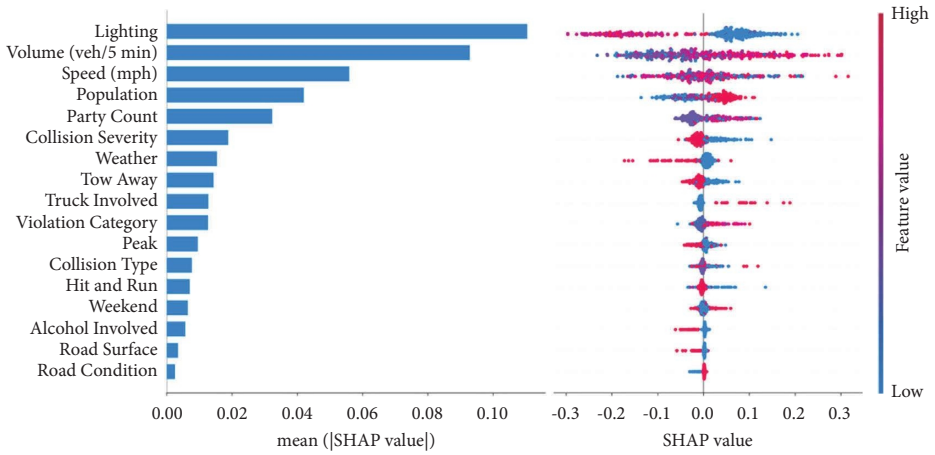

Figure 3: Global importance of independent variables and a summary of local explanations for the time gap.

low volume (blue points) mainly has a negative one, revealing that high volume promotes the time gap while low volume inhibits it.

Figure 4 represents the global importance of variables on the distance gap. As shown in the left part, volume is the most significant contributor and has an overwhelming effect on the distance gap, changing the predicted value by 0.136. Definitely, volume size directly influences the length of the vehicle queue and, thus, the distance gap between the PC and SC. Lighting, speed, and population also rank at the top of the importance list. Road surface and condition are in the bottom third and second places. Generally, the importance ranking of variables for the two gaps is different, but there are overall similarities. Traffic features are always the most important. Crash and temporal characteristics are commonly distributed throughout the importance list. And road traits contribute relatively small to both time and distance gaps. Regarding the right part, it shows that high volume, daylight, enormous speed, and a dense population have a positive SHAP value, possibly increasing the distance gap.

5.3. Local Effects of Variables. In previous studies, the local effects of a particular variable on the predicted outcome are often observed assuming that other variables are constant. The drawback is that this way does not consider the issue that the changes of specific variable likely cause variations in other variables (rather than assuming that all other variables are constant). The local dependence plot obtained based on the SHAP method can quantify the variables' effects while avoiding the abovementioned disadvantage. The main effects were calculated for each variable. In addition, considering the nontrivial effects of traffic characteristics on the time and distance gaps (see Figures 3 and 4 in the previous section), their interaction effects with the rest of the variables were also estimated. In this section, we select variables with strong effects for analysis.

Figure 5 shows the local dependence plots for volume on the time and distance gaps. Specifically, the first two plots reveal the main effects of volume, and the last two reflect the interaction effects between volume and speed. Moreover, the left column is for the time gap, while the right column is for the distance gap. In each plot, every point corresponds to a sample. The x-axis represents the volume value; the left y-axis indicates the SHAP value (i.e., the local effect); the right y-axis and the different colored points in the last two plots describe the speed value. As shown in Figures 5(a) and 5(b), plots for volume reveal an overall upward trend. When volume is around 100 veh/5 min in the two plots, its local effects remain at the negative highest level, suggesting that low volume may lead to a sharp decline in the time and distance gaps. One possible explanation is that low volume allows for such long distances between vehicles that drivers tend to relax their vigilance generally. When faced with a sudden crash, they are likely to react slowly and are unable to stop timely at high speed (as shown in the lower-left corner of Figures 5(c) and 5(d), the corresponding speed is around 65 mph). Another reasonable interpretation is that low volume does not contribute to long queue length formation, thus creating a short-distance gap. As volume grows to 500 veh/5 min, its local effects remain at the positive highest level, indicating that high volume is likely to rapidly

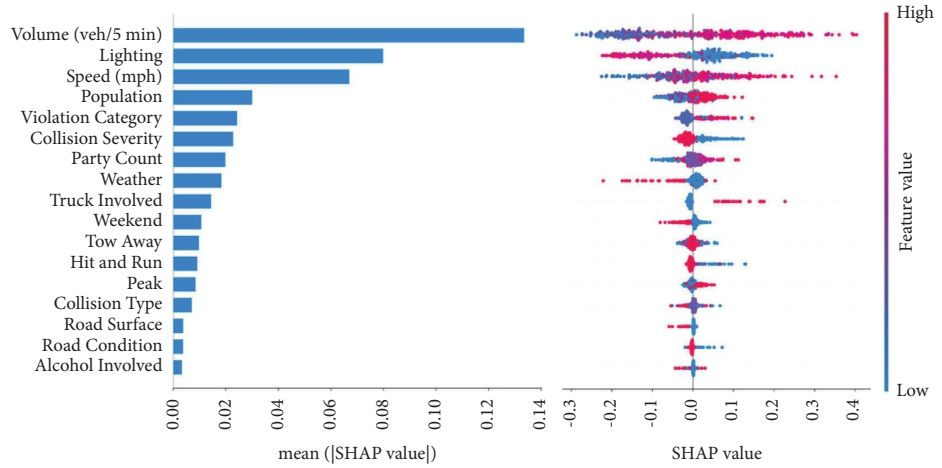

Figure 4: Global importance of independent variables and a summary of local explanations for the distance gap.

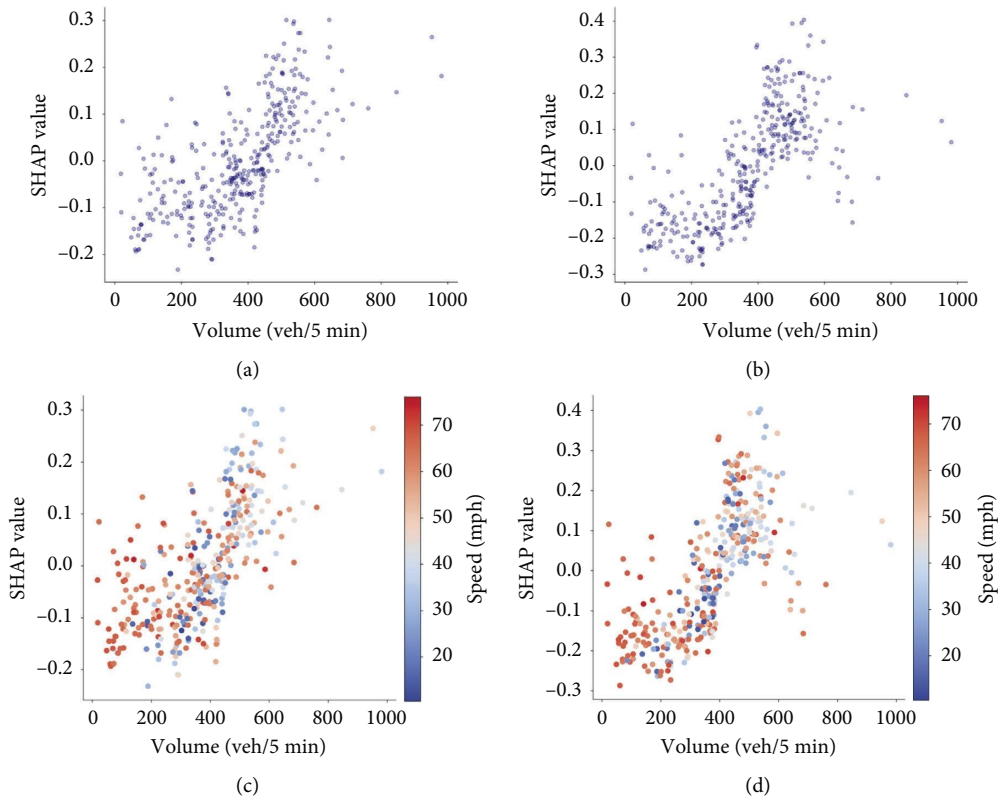

Figure 5: SHAP local dependence plots of volume. (a) Main effects of volume on the time gap. (b) Main effects of volume on the distance gap. (c) Interaction effects between volume and speed on the time gap. (d) Interaction effects between volume and speed on the distance gap.

increase the time gap and distance gap. This finding is consistent with existing works [15]. The reason might be that high volume makes the traffic situation entirely stressful, and drivers have developed a cautious driving style under this circumstance. When a PC occurs, drivers in the immediate vicinity upstream will not feel large shock, so SC does not occur as quickly. Moreover, high volume can prolong queue length and thus increase the distance gap. When volume is around 500 veh/5 min, its corresponding speed falls in an extensive range of 24–76 mph.

Figures 6(a) and 6(b) show the main local effects of speed on the time and distance gaps, respectively. The trends in the two plots are similar in general (down then up), but the inflection points correspond to different speed values. In Figure 6(a), as speed ranges between 0 and 50 mph, its local effects on the time gap decline to negative from positive as it increases. When speed falls 50–75 mph, its local effects show a steep upward trend. As for Figure 6(b), when speed increased from 0 to 30 mph, its local effects decline from 0.05 to −0.22, indicating that this value range of speed inhibits the

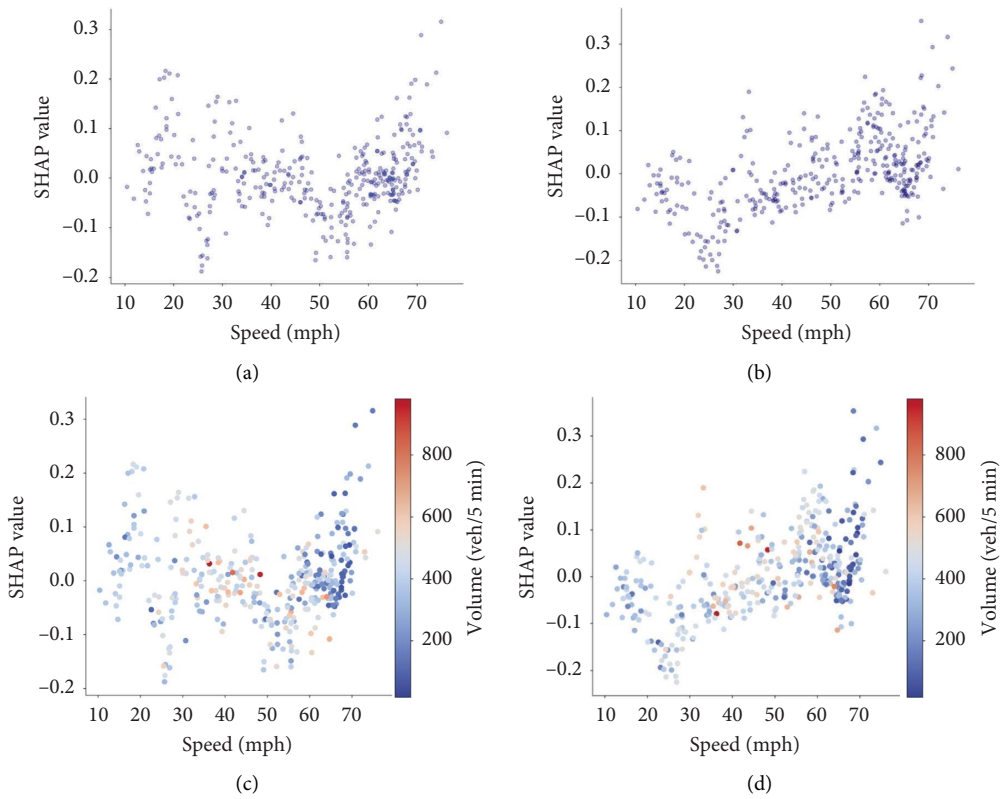

Figure 6: SHAP local dependence plots of speed. (a) Main effects of speed on the time gap. (b) Main effects of speed on the distance gap. (c) Interaction effects between speed and volume on the time gap. (d) Interaction effects between speed and volume on the distance gap.

distance gap. As the speed continues to increase, the local effects grow to be positive. Moreover, we found that when the speed ranges between 60 and 75 mph (the average volume for this speed range is 281 veh/5 min), the corresponding effects for both time and distance gaps are stable around value 0, as observed from Figures 6(c) and 6(d). Such a finding demonstrates that this traffic state has minor promotion/inhibition on both gaps.

Figures 7(a) and 7(b) demonstrate the main effects of lighting on the time and distance gaps; the two plots reveal an approximate concave trend. As shown in Figure 7(a), the local effects of daylight and dawn (i.e., lighting = 0 and 1) on the time gap fall in the range of 0–0.20, while streetlights and no streetlights (i.e., lighting = 2 and 3) have the most negative effects. Such variations in local effects indicate that a darker environment will accelerate the occurrence of SCs. Probably because the driver's sight distance in dark situations depends on the space illuminated by the streetlights and headlights, leading to a lack of timely and clear perception of the current road condition, resulting in insufficient avoidance of PCs and thus reducing the time gap. Figures 7(c) and 7(d) display the interaction effects between lighting and volume. As observed, all points are approximately divided by their color into upper-right and lower-left parts, with most of the pale and dark blue points (i.e., representing daylight and dawn) being above the horizontal axis where the local effect is −0.1, the red and orange points (i.e., denoting streetlights and no streetlights) being below it. In other words, a bright environment has a larger volume and positive local effects, while a dark condition has a relatively smaller volume and negative local effects. It makes sense that the vehicle trips are more during the day than at night. Likewise, it is reasonable to consider that high volume likely prolongs queue length and therefore increases the distance gap.

Figures 8(a) and 8(b) represent the main effects of violation category on the two gaps. As observed, improper turns (i.e., violation category = 4) have the maximum SHAP value. Specifically, its local effects on the time and instance gaps roughly fall between 0–0.10 and 0–0.15, respectively; such ranges indicate that this violation category promotes the time and distance gaps to a varying degree. The reason might be that the crashes in which the violation category is improper turns probably block turn lanes (usually on a one-way road), thus affecting the vehicles behind and causing a long queue length. Followed by another violation category of unsafe lane changes (i.e., violation category = 3), which shows positive correlations with both gaps. Likewise, crashes caused by unsafe lane changes likely block multiple lanes and involve several vehicles, thus decreasing the road capacity significantly and extending the queue length. Besides, this type of crash is more visible. That means drivers behind can catch the crash information at a distance and drive more carefully, increasing the time and distance gaps. By contrast, the other four violation categories have more negligible local effects. As shown in Figures 8(c) and 8(d), it is the

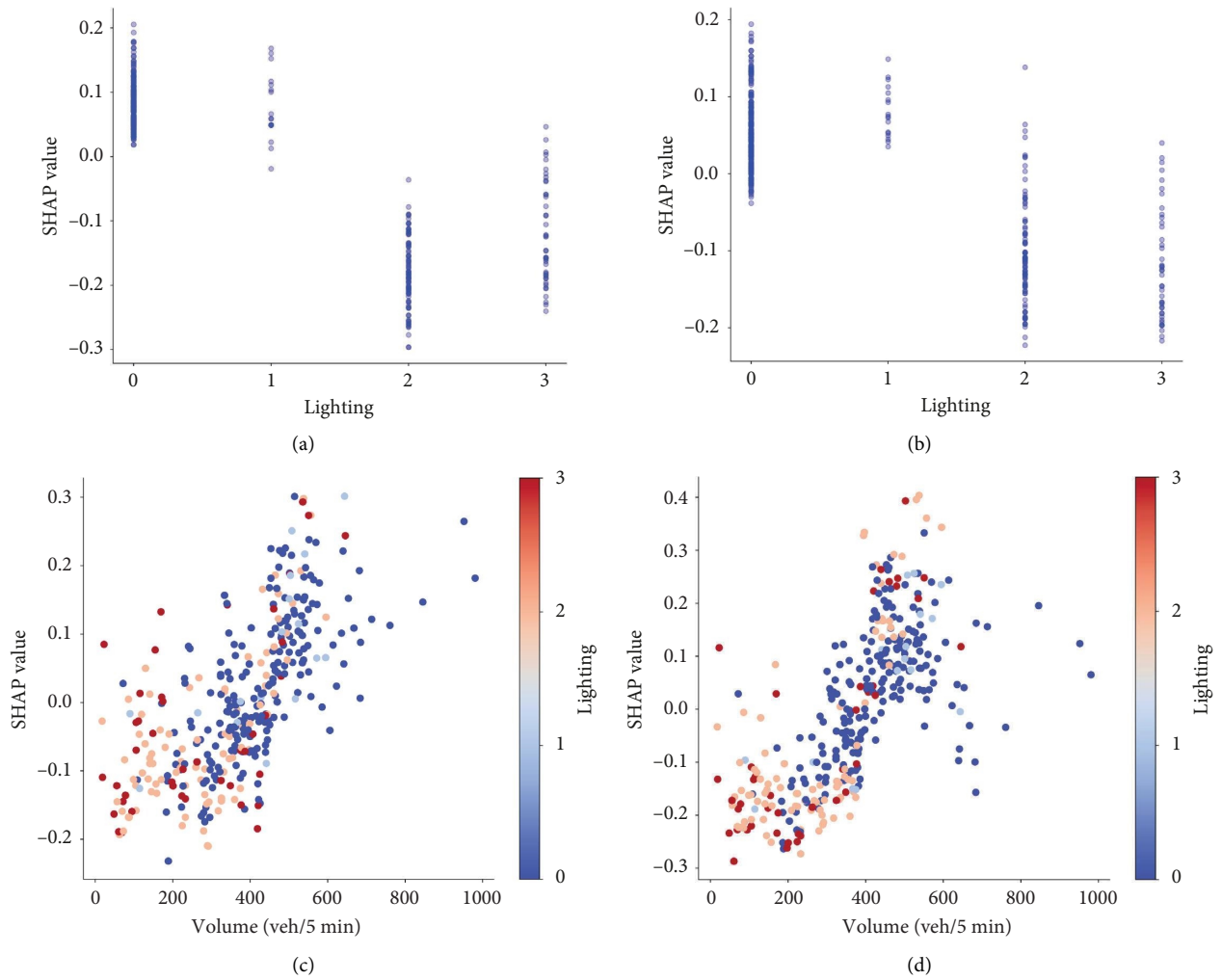

Figure 7: SHAP local dependence plots of lighting. (a) Main effects of lighting on the time gap. (b) Main effects of lighting on the distance gap. (c) Interaction effects between volume and lighting on the time gap. (d) Interaction effects between volume and lighting on the distance gap.

interaction effects between violation category and speed. We find a strong association between crashes involving alcohol (i.e., violation category = 0) and high speed, because points are red on the first vertical column. Another interesting finding is that the red points in the fifth vertical column (i.e., violation category = 4) are concentrated at the bottom, illustrating that those crashes, which occurred due to unsafe lane changes at high speeds, reduce the time and distance gaps.

Figures 9(a) and 9(b) represent the main effects of collision severity. The fatal crashes, severe injury crashes, and light injury crashes (i.e., collision severity = 0, 1, and 2) have a promotion on the time and distance gaps, while only complaining crashes (i.e., collision severity = 3) mainly have inhibition on the two gaps. One possible reason is that serious crashes attract more attention, such as rapid rescue and intervention by traffic police, so that SCs do not occur at a close time and distance. Figures 9(c) and 9(d) show the interaction effects between collision severity and speed. As observed, most of the blue points (represent the sample of fatal crashes) occur in the speed range of 60–70 mph, suggesting that serious crashes frequently occur at high speeds.

The main and interaction effects of other variables are presented in Figures 10 and 11. As shown, plots of population reveal a broadly upward trend, varying from negative to positive. A dense population (i.e., Population = 3 and 4, indicating the population is more than 250000) promotes the time gap and distance gap. One possible explanation is that car ownership and travel trips may be relatively high in these densely populated areas, leading to long queuing times and length. The local effects of most weekdays (i.e., weekend = 0) and peak periods (i.e., peak = 1) on the distance gap are greater than the value 0. It makes sense that weekdays and peak periods have many commuter trips, resulting in high volume on the road. The plot for collision type shows a v trend on the time gap while a downward trend on the distance gap. The effects of clear days are around the value 0, while the effects of most cloudy days are less than the value 0. Such a comparison indicates that cloudy days will inhibit both gaps, i.e., SCs will occur sooner and closer on cloudy

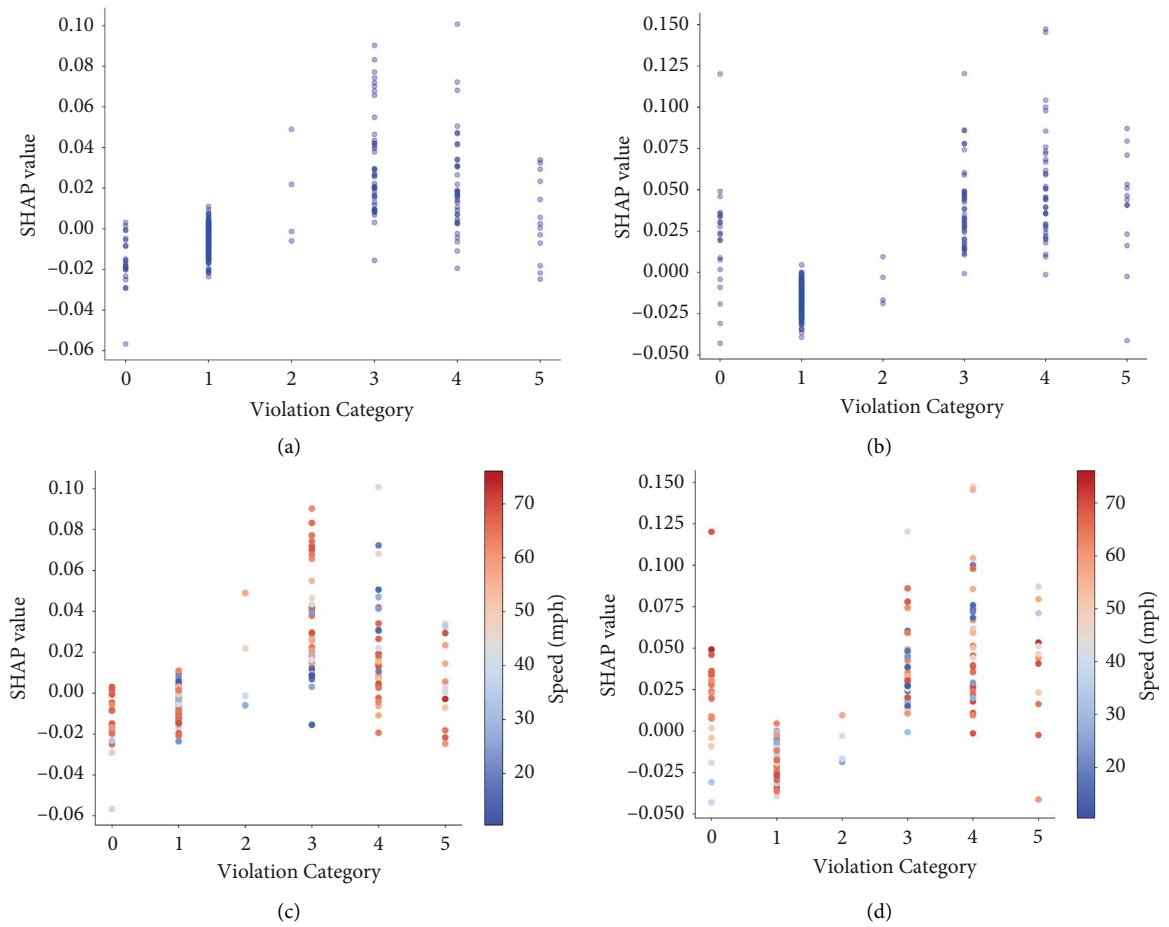

Figure 8: SHAP local dependence plots of violation category. (a) Main effects of the violation category on the time gap. (b) Main effects of violation category on the distance gap. (c) Interaction effects between violation category and speed on the time gap. (d) Interaction effects between violation category and speed on the distance gap.

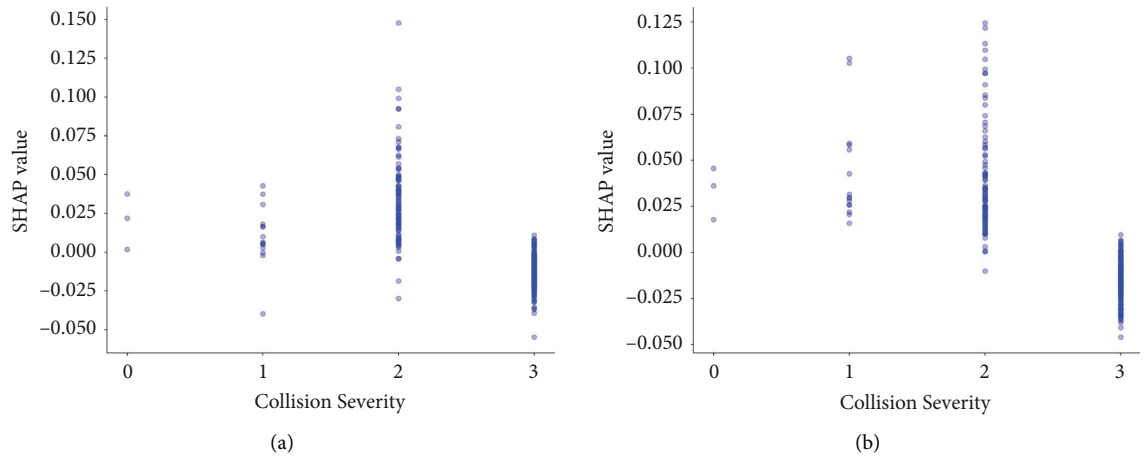

Figure 9: Continued.

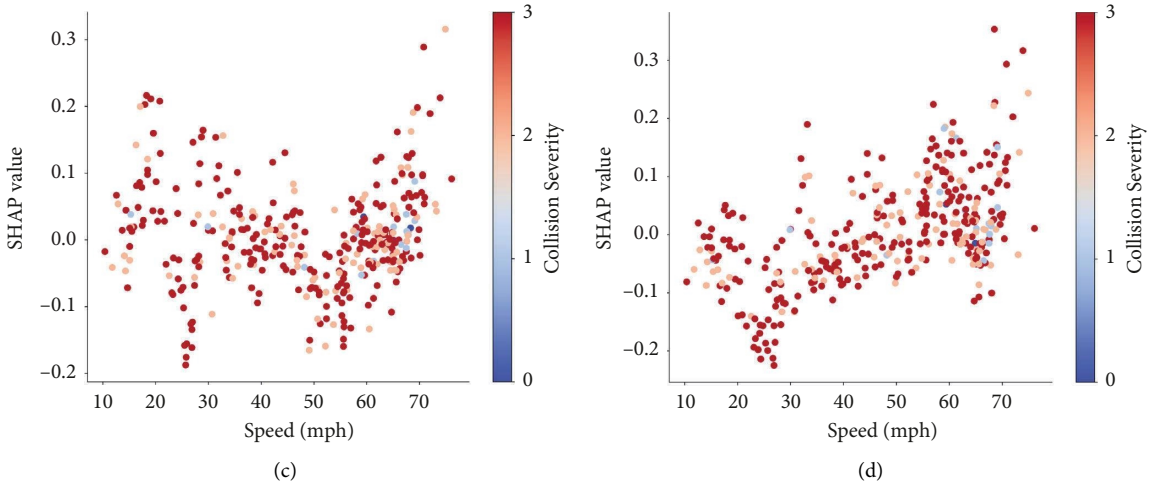

Figure 9: SHAP local dependence plots of collision severity. (a) Main effects of collision severity on the time gap. (b) Main effects of collision severity on the distance gap. (c) Interaction effects between collision severity and speed on the time gap. (d) Interaction effects between collision severity and speed on the distance gap.

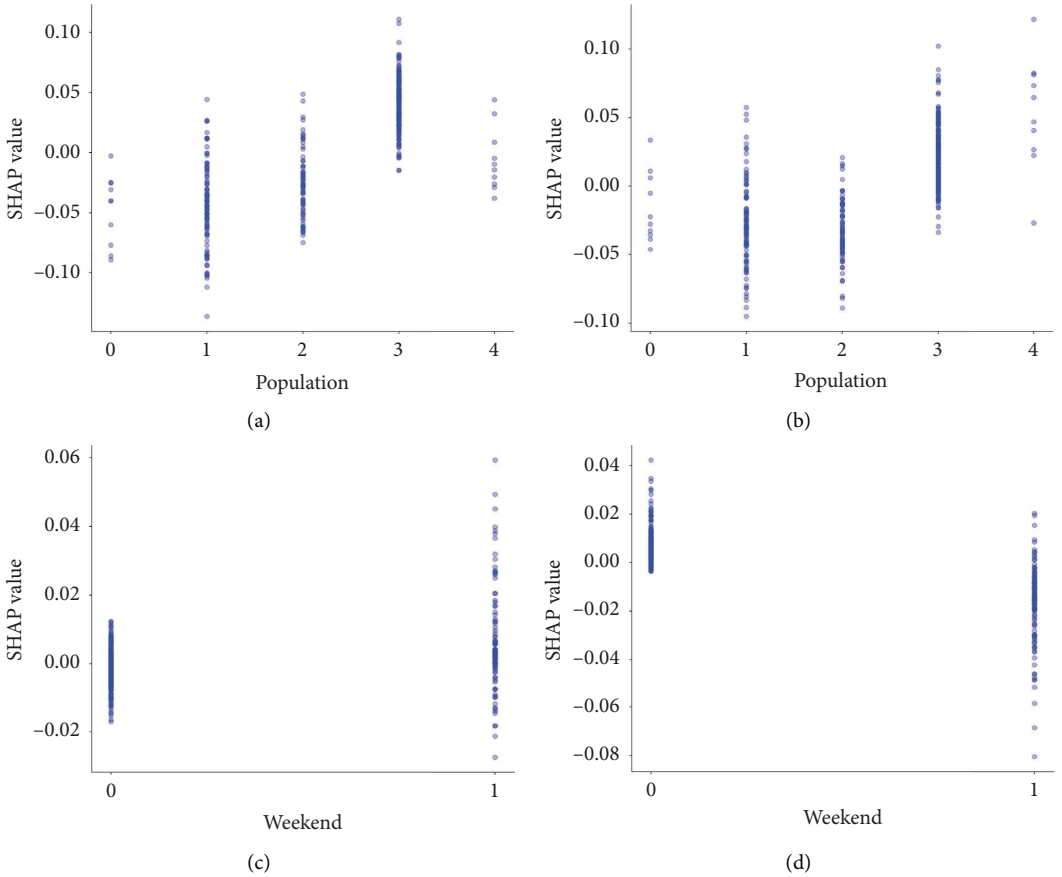

Figure 10: Continued.

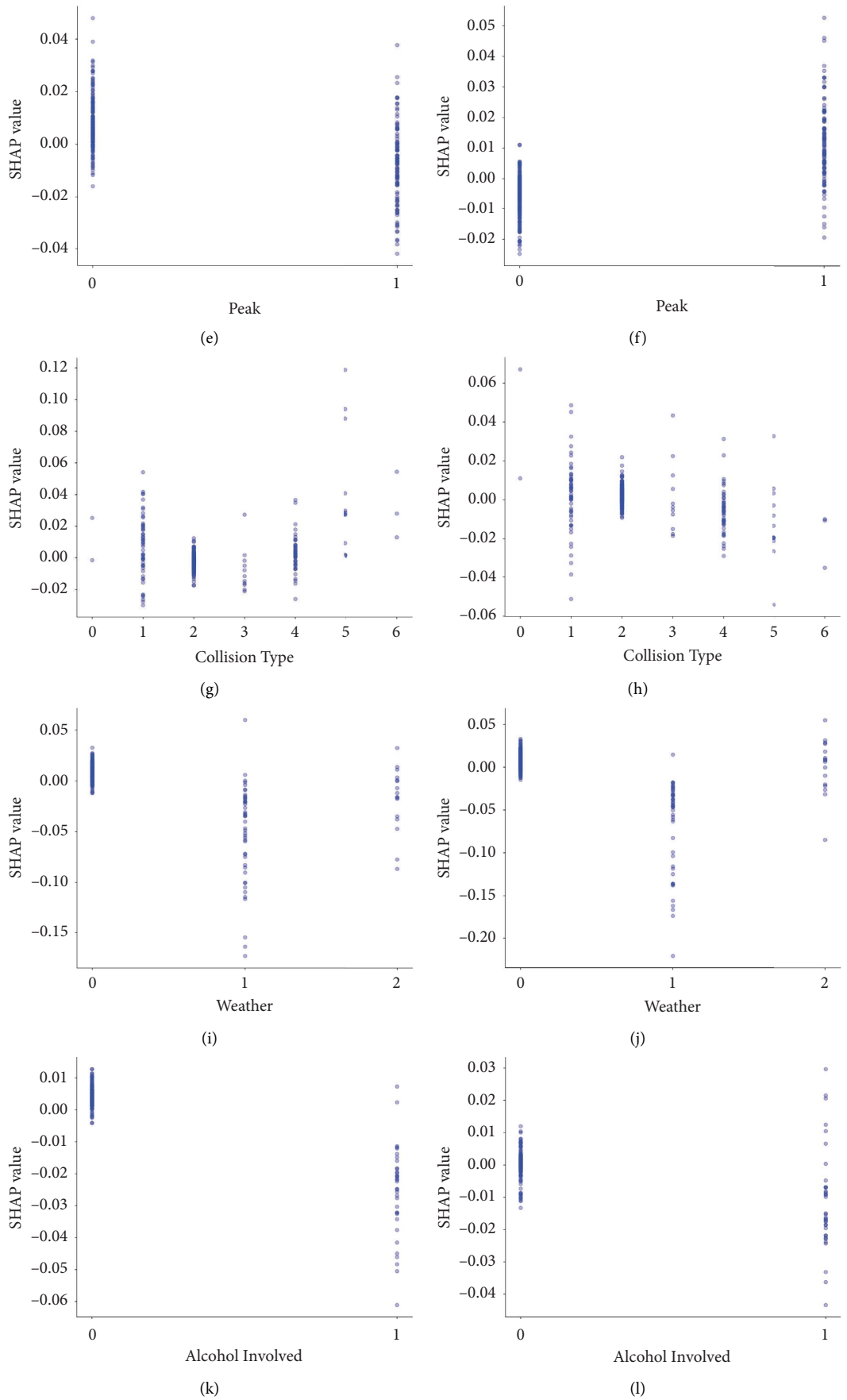

Figure 10: Continued.

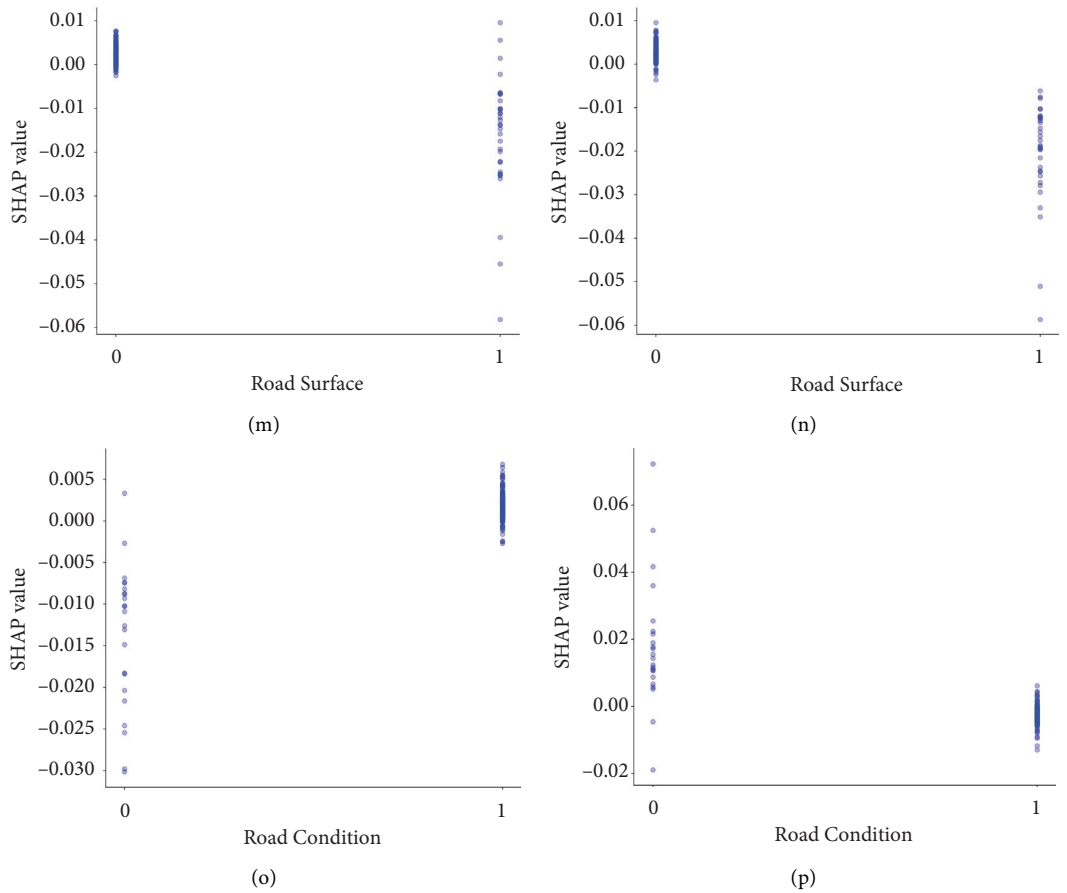

Figure 10: SHAP main effects of variables on the time gap and the distance gap. (a) Population on the time gap. (b) Population on the distance gap. (c) Weekend on the time gap. (d) Weekend on the distance gap. (e) Peak on the time gap. (f) Peak on the distance gap. (g) Collision type on the time gap. (h) Collision type on the distance gap. (i) Weather on the time gap. (j) Weather on the distance gap. (k) Alcohol involved on the time gap. (l) Alcohol involved on the distance gap. (m) Road surface on the time gap. (n) Road surface on the distance gap. (o) Road condition on the time gap. (p) Road condition on the distance gap.

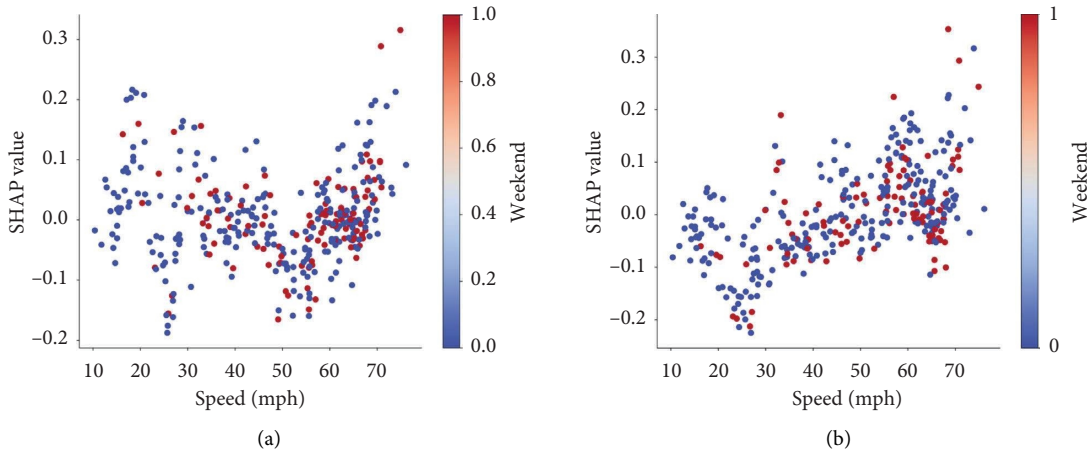

Figure 11: Continued.

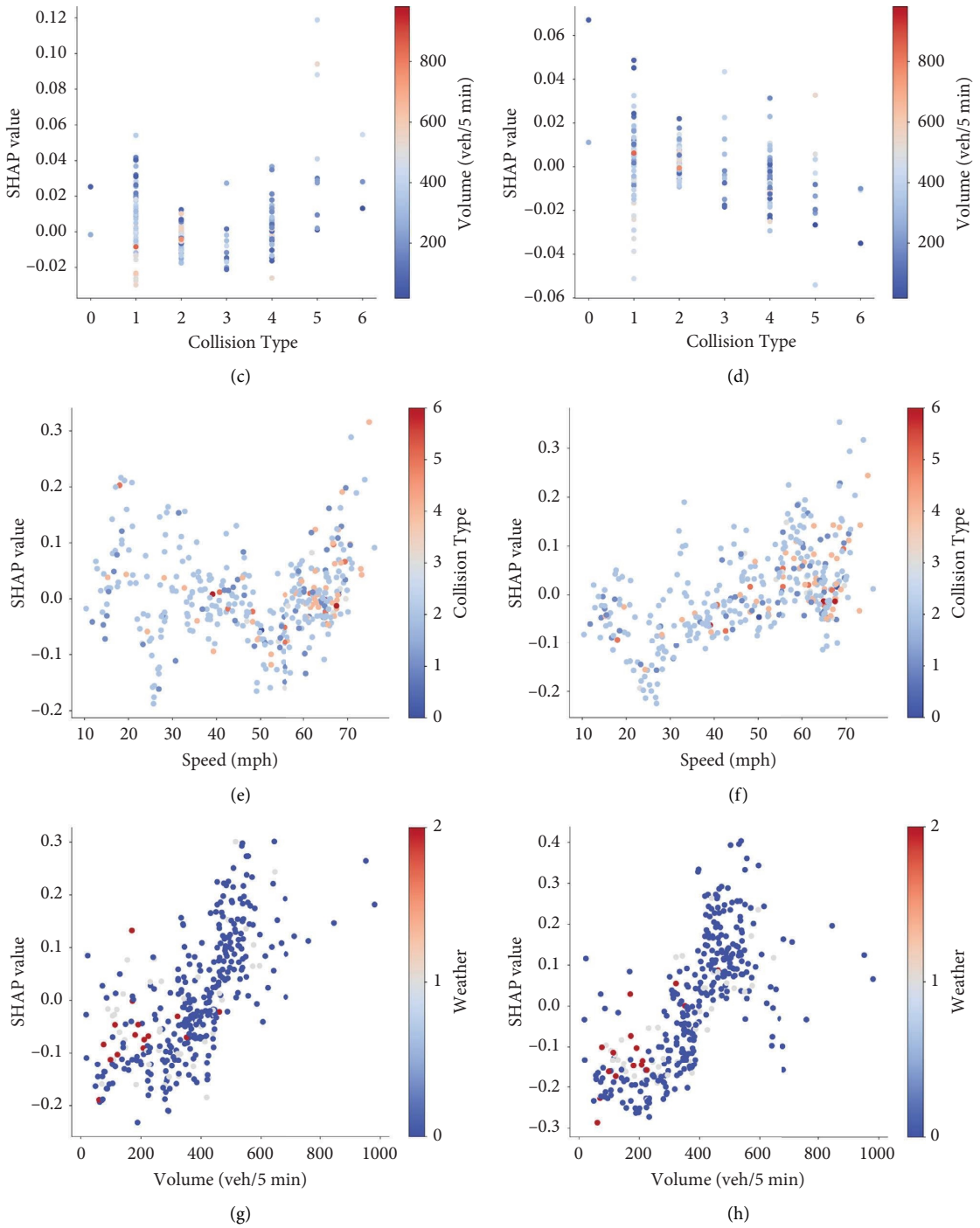

Figure 11: Continued.

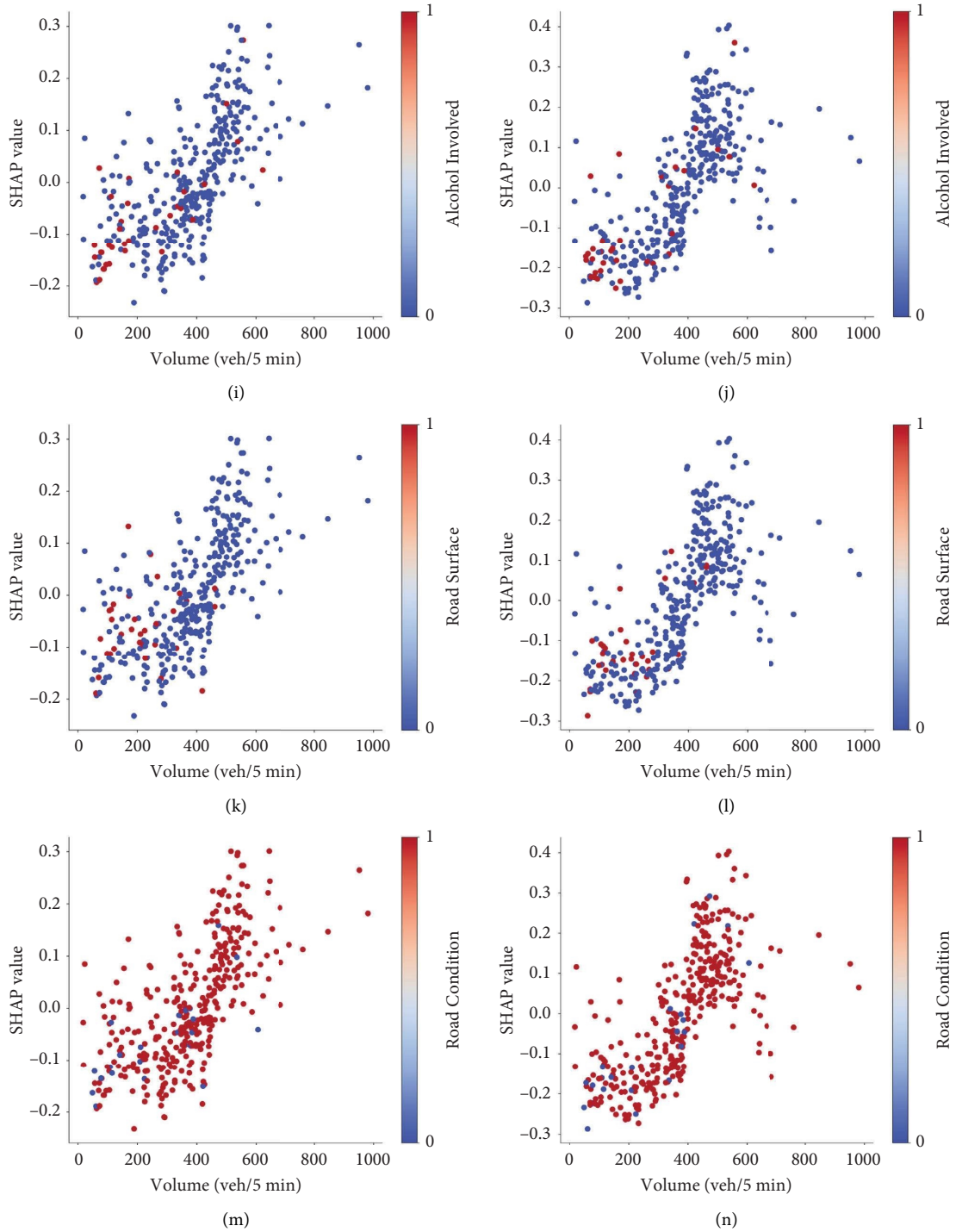

Figure 11: SHAP interaction effect plots among variables on the time gap and the distance gap. (a) Speed and weekend on the time gap. (b) Speed and weekend on the distance gap. (c) Collision type and volume on the time gap. (d) Collision type and volume on the distance gap. (e) Speed and collision type on the time gap. (f) Speed and collision type on the distance gap. (g) Volume and weather on the time gap. (h) Volume and weather on the distance gap. (i) Volume and alcohol involved on the time gap. (j) Volume and alcohol involved on the distance gap. (k) Volume and road surface on the time gap. (l) Volume and road surface on the distance gap. (m) Volume and road condition on the time gap. (n) Volume and road condition on the distance gap.

days. Drinking (i.e., alcohol involved = 1) mostly has negative local effects, meaning that drinking will reduce the time and distance gap. This is consistent with reality. The wet surface (i.e., road surface = 1) inhibits both gaps, which is consistent with existing knowledge [16]. It makes sense that a wet road surface harms the vehicle's stability, such as a brake failure, thus accelerating the occurrence of SCs.

6. Conclusions

This study aimed at predicting the time and distance gaps between SCs and PCs on highways and to analyze how the influencing factors contribute to the gaps comprehensively. First, a data-driven identification method combining the fixed spatiotemporal thresholds-based method and the speed contour map-based method was developed to identify SCs. A total of 368 SCs were sought out from the total number of 24643 crashes. Then, the RF model was applied to predict the two gaps. The data samples were split into training and testing sets at a ratio of 7 : 3. The results showed that the RF model performed better than KNN and MPR. Additionally, the SHAP method was conducted to explain the outputs of the RF model. Based on this local interpretation method, we revealed variables' global importance and main and interaction effects on the time and distance gaps.

We found that traffic volume and speed are the important contributors to the time and distance gaps; monitoring traffic conditions helps implement timely and effective management to prevent SCs. Several temporal characteristics, such as lighting and population, contribute more to both gaps than primary crash features and road factors. Compared with road factors, the primary crash characteristics of violation category, party count, and collision severity demonstrate more significant effects. With these findings about factor priorities, traffic managers and policymakers can develop prevention plans and allocate resources more efficiently.

The local dependence plots quantify the effects of variables. Plots for the continuous variables, i.e., volume and speed, reveal developing trends and several inflection points. For example, the local effects of volume increase monotonically from −0.3 to 0.4 as the volume grows. Such variation indicates that low volume sharply inhibits the time and distance gaps, while high volume boosts them significantly. Additionally, the local effects on the distance gap are around value 0 when volume falls between 300 and 400 veh/5 min, suggesting that the traffic state in this volume affects the gap inconsiderably. The plot for the main effects of speed on the distance gap shows an obvious inflection point. Such critical information above is considerable for traffic safety managers. As for plots about the discrete variables, demonstrate the local effects and corresponding characteristics of different categories of variables. Take lighting as an example: the effects of daylight and dawn are positive, while those of streetlights and no streetlights are mostly having negative effects. That is to say, a darker environment probably accelerates the occurrence of SCs. Where the economic condition allows, it is advantageous to increase the intensity of the lighting. Moreover, crashes involving the violation categories of improper turns or unsafe lane changes possibly cause long time and large distance gaps.

The contributions of this study can be summarized in the following three aspects: (1) proposing a two-stage SC identification method, which combined the static and dynamic approaches. And the identification results on the test data are consistent with existing works, providing a reliable basis for SC analysis. (2) Applying random forest to simultaneously predict the time and distance gaps, which facilitated understanding the relationship between the dependent and independent variables. 17 independent variables selected from temporal, primary crash, roadway, and traffic characteristics and two dependent variables, namely time gap and distance gap, were used as inputs to train and test the random forest model. The results achieved better performance compared with other models. (3) Using a brand-new interpretation technique SHAP to explain the RF model from global and local ways. We made several significant findings which will be definitely helpful for traffic decision makers to formulate strategies.

This research also raises issues in need of further explorations in the future. First, 368 crashes were used in the model training. Although we applied ML models that are advantageous for handling sparse data, small sample sizes may reduce the performance of the models. More data are expected to be required to improve the model performance. Second, 17 variables were used, and future work will cover more types of factors. This study focused on temporal characteristics, primary crash factors, roadway conditions, and real-time traffic parameters. Other factors, such as shoulder width and truck proportion which have shown correlations with the time gap and distance gap of SCs, will be considered in future research. The SC factors are also worthy of being discussed. In the future, it is a potential idea to combine the PC and SC characteristics to explore the time and distance gaps.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded in part by the National Natural Science Foundation of China (grant no. 52172310), Humanities and Social Sciences Foundation of the Ministry of Education (grant no. 21YJCZH147), Innovation-Driven Project of Central South University (grant no. 2020CX041), and the Fundamental Research Funds for the Central Universities of Central South University (grant no. 2022ZZTS0717).

References

[1] World health organization, "Road traffic injuries," 2023, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2] H. Yang, Z. Wang, K. Xie, K. Ozbay, and M. Imprialou, "Methodological evolution and frontiers of identifying, modeling and preventing secondary crashes on highways," *Accident Analysis and Prevention*, vol. 117, pp. 40–54, 2018.

[3] S. A. Tedesco, V. Alexiadis, W. R. Loudon, R. Margiotta, and D. Skinner, "Development of a 40 model to assess the safety impacts of implementing IVHS user services, moving toward deployment," in *Proceedings of the IVHS America Annual Meeting*, pp. 343–352, Atlanta GA, USA, March 1994.

[4] N. Owens, A. Armstrong, P. Sullivan et al., *Traffic Incident Management Handbook*, Federal Highway Administration, Washington, DC, USA, 2010.

[5] J. G. Pigman, J. R. Walton, and E. R. Green, "Identification of secondary crashes and recommended countermeasures," *Crash Severity*, Transport Research International Documentation, Washington, DC, USA, 2011.

[6] M. Jalayer, F. Baratian-Ghorghi, and H. Zhou, "Identifying and characterizing secondary crashes on the Alabama state highway systems," *Advances in Transportation Studies*, vol. 37, pp. 129–140, 2015.

[7] Y. Tian, H. Chen, and D. Truong, "A case study to identify secondary crashes on interstate highways in Florida by using geographic information systems (gis)," *Advances in Transportation Studies*, vol. 2, pp. 103–112, 2016.

[8] H. Yang, Z. Wang, and K. Xie, "Impact of connected vehicles on mitigating secondary crash risk," *International Journal of Transportation Science and Technology*, vol. 6, no. 3, pp. 196–207, 2017a.

[9] C. Zhan, L. Shen, M. A. Hadi, and A. Gan, "Understanding the characteristics of secondary crashes on freeways," in *Proceedings of the Transportation Research Board 87th Annual Meeting*, Washington, DC, USA, January 2008.

[10] H. Yang, K. Ozbay, and K. Xie, "Assessing the risk of secondary crashes on highways," *Journal of Safety Research*, vol. 49, no. 143, pp. 143.e1–149, 2014.

[11] C. Xu, P. Liu, B. Yang, and W. Wang, "Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 406–418, 2016.

[12] H. Park and A. Haghani, "Real-time prediction of secondary incident occurrences using vehicle probe data," *Transportation Research Part C: Emerging Technologies*, vol. 70, pp. 69–85, 2016.

[13] P. Li and M. Abdel-Aty, "A hybrid machine learning model for predicting Real-Time secondary crash likelihood," *Accident Analysis and Prevention*, vol. 165, Article ID 106504, 2022.

[14] H. B. Zhang and A. Khattak, "Spatiotemporal patterns of primary and secondary incidents on urban freeways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2229, no. 1, pp. 19–27, 2011.

[15] D. Chimba and B. Kutela, "Scanning secondary derived crashes from disabled and abandoned vehicle incidents on uninterrupted flow highways," *Journal of Safety Research*, vol. 50, no. 5, pp. 109–116, 2014.

[16] J. Wang, B. Liu, T. Fu, S. Liu, and J. Stipancic, "Modeling when and where a secondary accident occurs," *Accident Analysis and Prevention*, vol. 130, pp. 160–166, 2019.

[17] R. Raub, "Occurrence of secondary crashes on urban arterial roadways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1581, no. 1, pp. 53–58, 1997.

[18] B. Yang, Y. Guo, and C. Xu, "Analysis of freeway secondary crashes with a two-step method by loop detector data," *IEEE Access*, vol. 7, pp. 22884–22890, 2019.

[19] J. E. Moore, G. Giuliano, and S. Cho, "Secondary accident rates on Los Angeles freeways," *Journal of Transportation Engineering*, vol. 130, no. 3, pp. 280–285, 2004.

[20] W. Hirunyanitiwattana and S. P. Mattingly, "Identifying Secondary Crash Characteristics for California Highway System," in *Proceedings of the Transportation Research Board Meeting*, Washington, DC, USA, January 2006.

[21] L. Kopitch and J.-D. M. Saphores, "Assessing effectiveness of changeable message signs on secondary crashes," in *Proceedings of the Transportation Research Board 90th Annual Meeting*, Washington, DC, USA, January 2011.

[22] H. Yang, Z. Wang, K. Xie, and D. Dai, "Use of ubiquitous probe vehicle data for identifying secondary crashes," *Transportation Research Part C: Emerging Technologies*, vol. 82, pp. 138–160, 2017.

[23] E. I. Vlahogianni, M. G. Karlaftis, and F. P. Orfanou, "Modeling the effects of weather and traffic on the risk of secondary incidents," *Journal of Intelligent Transportation Systems*, vol. 16, no. 3, pp. 109–117, 2012.

[24] M.-I. M. Imprialou, F. P. Orfanou, E. I. Vlahogianni, and M. G. Karlaftis, "Methods for defining spatiotemporal influence areas and secondary incident detection in freeways," *Journal of Transportation Engineering*, vol. 140, no. 1, pp. 70–80, 2014.

[25] W. Junhua, L. Boya, Z. Lanfang, and D. R. Ragland, "Modeling secondary accidents identified by traffic shock waves," *Accident Analysis and Prevention*, vol. 87, pp. 141–147, 2016.

[26] A. A. Sarker, R. Paleti, S. Mishra, M. M. Golias, and P. B. Freeze, "Prediction of secondary crash frequency on highway networks," *Accident Analysis and Prevention*, vol. 98, pp. 108–117, 2017.

[27] N. J. Goodall, "Probability of secondary crash occurrence on freeways with the use of private-sector speed data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2635, no. 1, pp. 11–18, 2017.

[28] H. Park, S. Gao, and A. Haghani, "Sequential interpretation and prediction of secondary incident probability in real time," in *Proceedings of the Transportation Research Board 96th Annual Meeting*, Washington, DC, USA, January 2017.

[29] A. E. Kitali, P. Alluri, T. Sando, H. Haule, E. Kidando, and R. Lentz, "Likelihood estimation of secondary crashes using Bayesian complementary log-log model," *Accident Analysis and Prevention*, vol. 119, pp. 58–67, 2018.

[30] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Systems with Applications*, vol. 91, pp. 452–463, 2018.

[31] X. M. Chen, S. Zhang, and L. Li, "Multi-model ensemble for short-term traffic flow prediction under normal and abnormal conditions," *IET Intelligent Transport Systems*, vol. 13, no. 2, pp. 260–268, 2018.

[32] A. Jamal, M. Zahid, M. Tauhidur Rahman et al., "Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study," *International Journal of Injury Control and Safety Promotion*, vol. 28, no. 4, pp. 408–427, 2021.

[33] G. Asencio-Cortés, E. Florido, A. Troncoso, and F. Martínez-Álvarez, "A novel methodology to predict urban traffic

congestion with ensemble learning," *Soft Computing*, vol. 20, no. 11, pp. 4205–4216, 2016.
[34] SWITRS, "Statewide Integrated Traffic Records System (SWITRS)," 2022, https://iswitrs.chp.ca.gov/Reports/jsp/index.jsp.
[35] Pems, "Caltrans PeMS," 2022, https://pems.dot.ca.gov/.
[36] H. T. Yang, G. C. Zhai, L. C. Yang, and K. Xie, "How does the suspension of ride-sourcing affect the transportation system and environment?" *Transportation Research Part D: Transport and Environment*, vol. 102, Article ID 103131, 2022.
[37] H. T. Yang, J. H. Huo, R. B. Pan, K. Xie, W. J. Zhang, and X. J. Luo, "Exploring built environment factors that influence the market share of ridesourcing service," *Applied Geography*, vol. 142, Article ID 102699, 2022.
[38] A. A. Sarker, A. Naimi, S. Mishra, M. M. Golias, and P. B. Freeze, "Development of a secondary crash identification algorithm and occurrence pattern determination in large scale multi-facility transportation network," *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 142–160, 2015.
[39] S. Mishra, M. Golias, A. Sarker, and A. Naimi, *Effect of Primary and Secondary Crashes: Identification, Visualization, and Prediction Research Report No. CFIRE 09-05*, University of Wisconsin-Madison, Madison, WI, USA, 2016.
[40] J. Wang, W. Xie, B. Liu, S. Fang, and D. R. Ragland, "Identification of freeway secondary accidents with traffic shock wave detected by loop detectors," *Safety Science*, vol. 87, pp. 195–201, 2016.
[41] D. Yao, J. Yang, and X. Zhan, "A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines," *Journal of Computers*, vol. 8, no. 1, pp. 73-74, 2013.
[42] K. Miller, F. Huettmann, B. Norcross, and M. Lorenz, "Multivariate random forest models of estuarine-associated fish and invertebrate communities," *Marine Ecology Progress Series*, vol. 500, pp. 159–174, 2014.
[43] H. Hong, H. R. Pourghasemi, and Z. S. Pourtaghi, "Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models," *Geomorphology*, vol. 259, pp. 105–118, 2016.
[44] Y. Li, C. Zou, M. Berecibar et al., "Random forest regression for online capacity estimation of lithium-ion batteries," *Applied Energy*, vol. 232, pp. 197–210, 2018.
[45] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, "Crash injury severity analysis using a two-layer stacking framework," *Accident Analysis and Prevention*, vol. 122, pp. 226–238, 2019.
[46] G. Xu, M. Liu, Z. Jiang, D. Söffker, and W. Shen, "Bearing Fault diagnosis method based on deep convolutional neural network and random forest ensemble learning," *Sensors*, vol. 19, no. 5, p. 1088, 2019.
[47] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[48] G. De'ath, "Multivariate regression trees: a new technique for modeling species-environment relationships," *Ecology*, vol. 83, no. 4, pp. 1105–1117, 2002.
[49] M. Segal and Y. Xiao, "Multivariate random forests," *WIREs Data Mining and Knowledge Discovery*, vol. 1, pp. 80–87, 2011.
[50] X. Wen, Y. Xie, L. Wu, and L. Jiang, "Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP," *Accident Analysis and Prevention*, vol. 159, Article ID 106261, 2021.
[51] L. Xiao, S. Lo, J. Liu, J. Zhou, and Q. Li, "Nonlinear and synergistic effects of TOD on urban vibrancy: applying local explanations for gradient boosting decision tree," *Sustainable Cities and Society*, vol. 72, Article ID 103063, 2021.
[52] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
[53] S. M. Lundberg, G. Erion, H. Chen et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.