Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 5789573, 18 pages
https://doi.org/10.1155/2023/5789573

Research Article

Using Multidimensional Data to Analyze Freeway Real-Time Traffic Crash Precursors Based on XGBoost-SHAP Algorithm

Jie Li,[1,2] Yang Yang,[3,4] Yanran Hu,[5] Xinyuan Zhu,[1,6] Naixuan Ma,[1,6] and Xiaojing Yuan[5]

[1]*Shandong Key Laboratory of Highway Technology and Safety Assessment, Jinan, Shandong 250101, China*
[2]*Shandong Hi-Speed Information Group Co., Ltd., Jinan, Shandong 250000, China*
[3]*School of Transportation Science and Engineering, Beihang University, Beijing 100191, China*
[4]*Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Beihang University, Beijing 100191, China*
[5]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[6]*Shandong Hi-Speed Engineering Test Co., Ltd., Jinan, Shandong 250002, China*

Correspondence should be addressed to Yang Yang; yangphd@buaa.edu.cn

Received 12 October 2022; Revised 28 November 2022; Accepted 23 December 2022; Published 4 April 2023

Academic Editor: Wen Liu

The traditional freeway safety studies with "poststatic" thinking basically use cross-sectional data or panel data, which find it difficult to figure out real-time traffic crash risk factors. With the development of information collection technology, it is possible to obtain high-resolution traffic flow data currently, which provide a data basis for the dynamic traffic safety research towards freeways. This research aims at accurately identifying the real-time traffic crash precursors on freeways and addressing the shortcomings of conventional dynamic traffic safety research with the thinking of limited factor dimensions. In this research, dimensional data were applied as input model variables, the input dataset includes traffic crash data and the matched dynamic traffic flow data, and weather information and road characteristics were also considered to figure out the interaction effects between these dimensional factors. The XGBoost (eXtreme Gradient Boosting) was carried out to identify the dynamic crash-prone variables and the SHAP (SHapley Additive exPlanations) interpreter was introduced to interpret the XGBoost model, as well as the visualization of the influence of each eigenvalue on the traffic crash was realized. The results indicate that, in addition to traffic flow variables, road, weather, and temporal characteristics also have an impact on the traffic crash risk, and there is an interaction between each feature. The results of this research can provide the theoretical basis for freeway real-time traffic crash prediction and safety control.

1. Introduction

Over the past decades, road traffic safety has always been a hot topic in the field of traffic engineering. Due to the special driving characteristics of freeways, the quantity and severity of traffic crashes on freeways are often higher than those on ordinary roads [1], and about 1.25 million people die from traffic crashes every year worldwide, which is the most important cause of death among youths aged 15–29 [2]; the safety of freeways has attracted many scholars' attention. Nowadays, the collection of real-time traffic flow data becomes possible [3]; we can realize the dynamic control for road traffic safety through predicting the real-time freeway safety status, quantifying the temporal-spatial impact of the crash, and issuing early warning or alert information to drivers or managers via crash data and real-time traffic flow data, thus reducing the incidence and severity of freeway crashes, safeguarding people's lives, and reducing property losses.

Traditional freeway crash risk feature identification basically uses cross-sectional static data collected after a traffic crash, such as the variables including driver age, gender, whether to use a seat belt or if drink alcohol applicable, road characteristics, and weather characteristics, to

analyze its correlation with crash severity or frequency of occurrence, with the view to find ways to mitigate crash severity. Carson and Mannering [4] investigated the effectiveness of icing warning signs in reducing the crash frequency and crash severity in Washington State using statistical analysis. Kassu and Hasan [5] analyzed the key factors affecting fatal, nonfatal injury, and property damage crashes on four-lane and six-lane interstate freeway segments using a negative binomial regression approach for the three types of crashes and provided a meaningful statistical interpretation of the developed model estimates based on crash rate ratios. Yang et al. [1, 6] developed a novel data mining algorithm to explore the traffic crash occurrence mechanism for cross-area freeways using cross-section data, and the results show that there are significant differences in the causal factors of freeway crashes by regional type. Zhang et al. [7] proposed a spatial multinomial logit (SMNL) model for possible spatial correlations in freeway traffic crash injury data, considered spatial effects in the multinomial logit (MNL) model framework, and used a Gaussian conditional autoregressive prior to capture spatial correlations. Zeng et al. [8] developed a Bayesian spatial generalized ordered logit model based on 1424 crash records of a Chinese freeway in 2014 and 2015 to simulate the severity of crashes using hourly wind speed, temperature, precipitation, visibility, and humidity and other observations, and the severity was classified into three categories: mild, moderate, and severe. Wen et al. [9] studied the effect of road conditions and weather conditions and their interaction on crash incidence. A Bayesian spatiotemporal model is proposed to measure the relationship between the frequency of traffic crashes and possible risk factors, including traffic composition, the presence of curves and slopes, weather conditions, and their interactions. Malin et al. [10] analyzed the risk of crash under different road and weather conditions. The results showed that, for precipitation types, the highest relative crash risk was for snowfall; for road and weather conditions, the highest relative crash risk was for muddy road conditions; for road types, the relative crash risk was generally higher for highways compared to two-lane and multilane roads; for crash types, the relative crash risk corresponding to single-vehicle crash was generally higher than that of multivehicle crash.

However, these studies are mainly applying static data and do not consider the impact of dynamic traffic flow characteristics on crash risk. With the improvement of real-time information collection technology, research on identifying freeway traffic crash risk characteristics from a mesoscopic perspective based on real-time traffic flow data is gradually becoming popular. Golob and Recker [11] used a combination of nonlinear correlation and cluster analysis to identify traffic flow states under different crash types, and the case study used data from over 1000 crashes in Southern California to identify 21 traffic flow states under three different environmental conditions: dry roads during the day (8), dry roads at night (6), and wet (7), with each crash type corresponding to a different traffic flow characteristic. Zheng et al. [12] studied the effect of traffic oscillations (repeated deceleration followed by acceleration and stop-and-go driving) on freeway traffic safety. A conditional logistic regression model was developed using a case-control paired design with the traffic flow data before the crash as the case sample and the traffic flow data of the same period in the date without the crash as the control. Xu et al. [13] used traffic flow data and crash data from the northbound section of I-880 freeway in California, USA, performed K-means cluster analysis to classify the traffic flow into five different states, then used a conditional logistic regression model to study the relationship between traffic crash risk and traffic state, and compared the traffic flow characteristics under different traffic states to establish a model of traffic crash risk under different traffic states. Yu et al. [14] explored the effects of weather and traffic conditions on crash frequency under different scenarios based on crash data, traffic flow data, and real-time weather for the Colorado I-70 mountain freeway from August 2010–August 2011. This research uses a Poisson model and two random effects' models with a Bayesian inference approach for comparison, using the deviance information criterion (DIC) as a comparison factor. Later, Yu and Abdel-Aty [15] also studied mountainous freeway sections and classified traffic crashes on mountainous freeways according to their severity into serious (injury and fatal) and nonserious (property damage only) crashes and established a traffic crash injury severity analysis model. Sun et al. [16] proposed a new crash risk assessment method based on traffic safety state classification to explore the freeway crash risk under different traffic conditions. Yang et al. [17] considered the differences in traffic flow states to identify the dynamic crash risk for cross-area freeways. Table 1 summarizes the real-time dynamic factors affecting the crash risk obtained in these studies.

The identification for freeway traffic crash risk factors is an important premise for freeway safety improvement, as well as the theoretical basis for predicting the frequency and severity of traffic crashes. In particular, the recognition of crash precursors based on dynamic traffic flow is the main work of freeway real-time safety operation management. However, few studies have included weather, road, and time characteristics variables simultaneously among the explanatory variables and considered the interaction between traffic flow, weather, road, and time characteristics in the identification of real-time traffic crash risk on freeways. Based on crash data, matched traffic sensor data, weather data, and road characteristics, this research analyzes the relationship between the influence of each factor on freeway traffic crash, adds weather, road, and time characteristics to the explanatory variables for real-time freeway crash risk prediction, and analyzes the influence of each variable on crash risk and the interaction between them using the SHAP (SHapley Additive exPlanations) interpreter.

2. Data Process

Real-time traffic crash-prone identification research based on high-resolution traffic flow basically requires a huge number of traffic crash data and the matched traffic flow data. The accuracy of traffic flow data in this kind of research is usually 5 minutes in the temporal dimension [17], and the

Table 1: Real-time traffic crash precursors obtained from some research results.

Number	Author (year)	Factors affecting the risk of the crash
1	Golob and Recker [11] (2004)	Environmental conditions and traffic flow status
2	Lee et al. [18] (2009)	Average volume, flow variance coefficient, and flow ratio
3	Xu et al. [13] (2010)	Speed standard deviation
4	Christoforou et al. [19] (2011)	Average flow, lane to lane speed variance, average speed, lane-to-lane flow variance, and average density
6	Yu and Abdel-Aty [15] (2014)	Snow season indicators, slope indicators, speed standard deviation, and temperature
7	Wang et al. [20] (2015)	Mainline speed at the beginning of the interweaving section, speed difference at the beginning and end of the interweaving section, and logarithm of the traffic volume
8	Yang et al. [21] (2018)	Upstream average flow, crash section average flow, crash section average speed, and crash section speed standard deviation
9	Yin [22] (2021)	Downstream speed, upstream occupancy, downstream speed coefficient of variation, upstream speed standard deviation, traffic flow status, and time conditions

step size of traffic flow which is larger than 5 minutes may affect the accuracy of the model negatively.

2.1. Data Description. Freeway traffic crash is the result of multiple factors, which needs to consider traffic flow, weather, environment, and other factors at the same time. In this research, the Beijing section of the Beijing-Harbin freeway was selected as the research object, and the crash data, traffic flow data, weather and road characteristics of the study section, and period were collected, and these data were preprocessed and matched, after which suitable alternative feature parameters were selected to establish the data base for the analysis of the impact factors of freeway traffic crash.

In this research, the section of Beijing-Harbin freeway with milepost number k0-k39 of total length about 39 km is taken as the specific research section: traffic crash data (including crash type, severity, time, and location), traffic flow data (including massive high-precision traffic flow, speed, occupancy, and 85% speed in 1-minute sets), weather data (including visibility, rainfall, dew point, temperature, and wind speed), and roadway characteristics data (including upstream and downstream detector spacing, roadway width, shoulder width, and number of ramps). The characteristics of each data source are described.

2.1.1. Traffic Crash Data. A total of 198 traffic crash data were collected from January 2013 to September 2014 for the inner section of the study road, with data fields including crash time, stake number, direction, crash type, and crash description. Traffic crashes caused by vehicle fire and damage were deleted, and the crash data were numbered and operated in the order of occurrence to facilitate the next step. After data preprocessing, a total of 164 traffic crashes were extracted from the studied road sections and time periods.

2.1.2. Traffic Flow Data. Traffic flow data can usually be acquired using induction loops, microwave, or video detection. The traffic flow data used in this research were collected by microwave radar traffic information detectors. Microwave radar traffic information detectors are usually placed at the side of a freeway or above a lane and are capable of detecting traffic flow parameters in multiple lanes simultaneously. The microwave radar detector used for traffic flow data collection in this research allows simultaneous detection of traffic flow parameters for one cross section, i.e., six lanes in both directions.

A total of 20 sets of microwave radar traffic information detectors are deployed in the upstream and downstream directions of the road section to be studied, and the collected raw traffic information data mainly include speed, volume, and lane occupancy. The distance between the upstream and downstream microwave detectors varies widely, with the farthest distance reaching 6.18 km, the shortest distance 0.8 km, and the average distance about 1.9 km. The data are saved as a "csv" file. The main data collected by the microwave detector include direction, volume of each lane, occupancy, and speed. The collection time interval is 1 minute, and some of the data are shown in Table 2. Taking the data in the first row of Table 2 as an example, it means that, during the period 2013/04/29 19:29:00-2013/04/29 19:29:59, in direction 2 (outbound direction to Beijing), the detector number 523050003 detected a traffic flow of 22 in lane 1 with a speed of 59 and an occupancy rate of 16%.

2.1.3. Weather Data. The road section to be studied passes through Chaoyang District and Tongzhou District of Beijing, China, and meteorological data for these two districts are available for 2013-2014, with the data in 1-hour sets. The fields mainly include time, rainfall, visibility, wind speed, temperature, and dew point. These weather features were clustered and the clustering results are shown in Table 3.

2.1.4. Road Features. The road feature data used in this research include information on the number of lanes, road width, shoulder width, the number of on-ramps and off-ramps between the upstream and downstream detectors, and whether they are curves, as shown in Table 4.

Table 2: Examples of partial traffic flow data.

Detection time	Detector number	Lane number	Direction	Volume	Speed	Occupancy rate
04/11/2013 19:29	523050003	1	2	22	59	16
04/11/2013 19:29	523050003	2	2	35	66	24
04/11/2013 19:29	523050003	3	2	30	70	23
04/11/2013 19:29	523050003	4	1	19	72	10
04/11/2013 19:29	523050003	5	1	27	70	19
04/11/2013 19:29	523050003	6	1	17	58	15
04/11/2013 19:30	523050003	1	2	18	62	14
04/11/2013 19:30	523050003	2	2	14	71	9

Table 3: Weather features and their values.

Weather features	Scope	Variable values
Rainfall (mm)	0–5	1
	5~10	2
	>10	3
Visibility (km)	<1	1
	1–5	2
	5–10	3
	>10	4
Dew point (°C)	<10	1
	10–15	2
	15–20	3
	>20	4
Temperature (°C)	<10	1
	10–20	2
	20–30	3
	>30	4
Wind speed (km/h)	0–9	1
	9–18	2
	18–27	3
	>27	4

Table 4: Examples of road features and their values.

Number of on-ramps	Number of off-ramps	Curve	Upstream and downstream distance (m)	Number of lanes	Road width (m)
2	1	0	1067	3	19
0	1	1	6180	3	11
1	1	0	2733	3	11
1	0	0	1020	3	11
0	0	1	3933	3	11
2	0	0	1200	3	11

2.2. Sample Structure Design

2.2.1. Unpaired Case-Control Sample Structure Design. The modeling work of highway traffic crash risk influence factor analysis and highway crash risk real-time prediction requires not only the traffic flow data before the crash but also the inclusion of traffic flow data in noncrash conditions, which is clearly a nonequilibrium binary classification problem. This type of problem can be solved by under-sampling, where a small number of samples are extracted from the traffic flow data samples in the noncrash state according to a certain extraction method, and the extracted samples can represent the traffic flow features in the noncrash state. Commonly used methods are paired case-control and unpaired case-control methods [23].

The paired case-control method is typically used to control for confounding factors such as weather, season, and road features and to study the effect of traffic flow variables on traffic crash risk by selecting traffic flow data in the noncrash conditions with crash time and location as paired variables. Therefore, the data samples constructed using the paired case-control method do not contain types of data such as weather, road, and time features [17] and are not suitable for studying the interaction between weather, road, time, and traffic flow features. The unpaired case-control method, on the other hand, is a random sampling of

noncrash data, which can include variables such as weather, road, and time of day, and is more appropriate for this research, so the unpaired case-control method was used in this research.

It has been shown that the larger the ratio of the number of control samples to the number of case samples, the more accurate the results obtained, but this ratio exceeds 4 : 1 will have minimal improvement in the accuracy of the predictions [24]. Therefore, in this research, the ratio of crash sample and noncrash sample is set to 1 : 4 to construct the data sample:

(1) Selection of sample data for cases (crash).

Microwave detector selection: Existing study shows that both upstream and downstream traffic flow features may have an impact on freeway traffic crashes [25], and in this research, the two upstream and downstream microwave detectors closest to the crash site were selected as the collectors of traffic flow data, and the locations of the microwave detectors are shown in Figure 1.

Selection of temporal period: Since the intention of the research is to provide traffic management with early warning before traffic crashes and then take active measures to avoid or mitigate the risk of crashes, it is necessary to set aside a certain amount of time for prejudgment and processing. In this research, the traffic flow data from 10 minutes to 5 minutes before the occurrence of a freeway traffic crash are used as the case sample data. For example, if a freeway traffic crash occurs at 15:35, then the traffic flow data in the period of 15:25–15:30 will be screened to characterize the traffic flow before this crash.

(2) Selection of control (noncrash) sample data: In this research, noncrash data are extracted according to the ratio of 1:4, and the crash data used in this research are 164, then it is necessary to make a random sampling method to extract 656 noncrash data, and the extraction steps are divided into the following three steps:

(i) First, we use the pandas and datetime toolkit in Python to generate 183745 time series with "2013-01-01 00:00:00" as the time starting point and "2014-10-01 00:00:00" as the time check extracted time series, delete the data that are too close to the crash data, and then conduct random sampling until 656 time series of noncrash data are generated to meet the requirements, and number them 1–656.

(ii) The ratio of direction 1 (entering Beijing) and direction 2 (leaving Beijing) in the crash data is 1:1, and the direction of travel in the noncrash data is generated according to the same ratio. From the 656 noncrash data, 328 data were randomly selected and the direction of travel was set to 1, and the direction of travel of the rest noncrash data was set to 2.

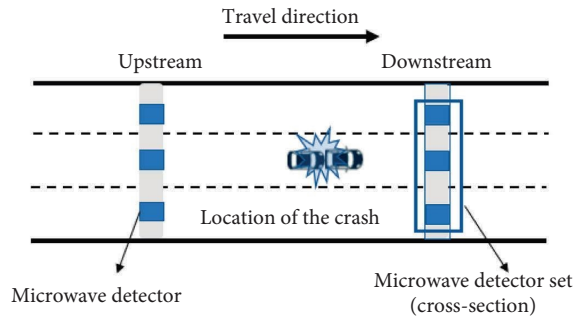

Figure 1: Location of traffic information detectors in relation to the location of traffic crash.

(iii) For each noncrash data, two adjacent detectors are randomly selected from 20 microwave detectors as the upstream and downstream detectors of the noncrash data, and the numbers of the upstream and downstream detectors are determined according to the direction of travel of the noncrash data.

2.2.2. Data Match. Various types of datasets are used in this research, including crash data, noncrash data, traffic flow, roads, and weather, which need to be matched according to certain fields to generate the datasets needed for the research. The data-matching work is divided into matching of traffic flow data, matching of road features, and matching of weather features.

(1) Traffic flow data match.

Matching of crash data and traffic flow data: First, locate the nearest upstream and downstream detector numbers according to the mile post and travel direction of the crash data; then find the corresponding traffic flow data of 10–5 minutes before the crash according to the actual time of the crash and the upstream and downstream detector numbers. The matching process is shown in Figure 2.

Matching of noncrash data and traffic flow data: Since noncrash data have already completed the matching work with upstream and downstream detector numbers and time in the process of data sampling, it is straightforward to search in the traffic flow data collected by detectors based on noncrash data times.

As an example, the traffic flow data matching process is specified for crash No. 5:

(i) The crash location of crash No. 5 is found from the crash data as milepost K25, the direction of traffic corresponding to the crash is out of Beijing, and the time of the crash is 2013/1/10 20: 58. According to the location data of each detector can be matched to the No. 5, the crash location is located between No. 16 detector (K24 + 680) and No. 17 detector (K25 + 700), and

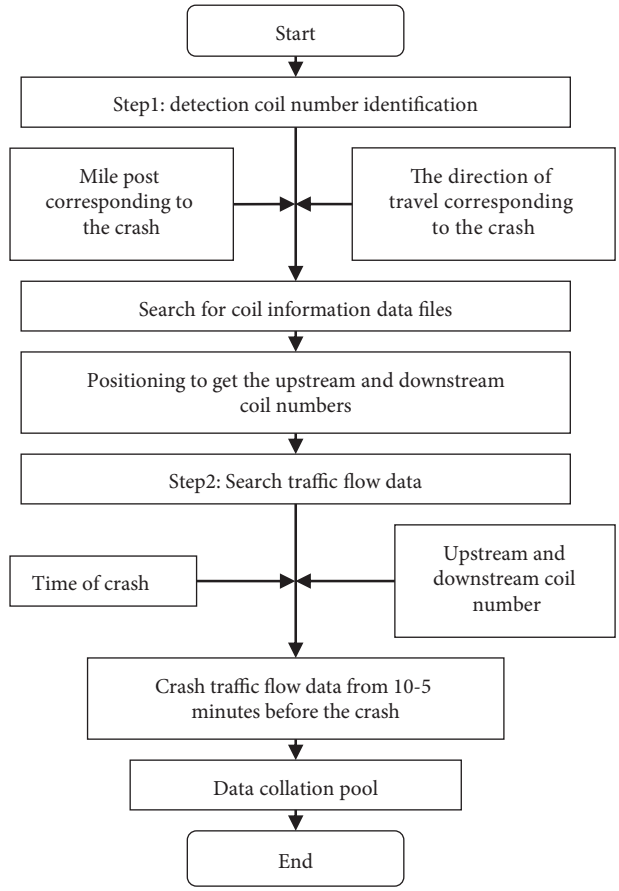

Figure 2: Matching flowchart of the crash data and traffic flow data.

then according to the direction of travel corresponding to the crash we can determine the upstream detector is No. 16 detector and the downstream detector is No. 17 detector.

(ii) According to the crash time: 2013/1/10 20:58, the traffic flow data including flow, occupancy, and speed in the period of 2013/1/10 20:48–20:53 can be filtered for 10–5 minutes before the crash. Using the collection time within the period of 2013/1/10 20:48–20:53 as the filtering condition, the eligible traffic flow data were filtered out from the traffic flow data collected by detectors No. 16 and No. 17 as the traffic flow data corresponding to crash No. 5.

(2) Weather and road feature match.

The nearest weather station can be matched according to the location of the crash, and then the weather data corresponding to the time of the crash can be found before the crash, including temperature, dew point, wind speed, rainfall, and visibility. Since the meteorological data are counted in 1-hour sets, the meteorological data matched to the crash data are the meteorological data for the nearest 1 hour to the time of the crash. The matching method for noncrash and meteorological data is the same.

In the matching of crash data and traffic flow data, each crash data has been matched to the corresponding upstream and downstream detector number, and the corresponding road information can be matched according to the location of upstream and downstream detectors and the direction of crash traffic, including the number of lanes, road width, road shoulder width, distance between upstream and downstream detectors, the number of on and off ramps between upstream and downstream detectors, and whether it is a curve. The matching method for noncrash data and road information is the same.

2.3. Selection of Candidate Feature Variables

2.3.1. Traffic Flow Parameters' Selection. The raw traffic flow data collected by the microwave detector are a 1-minute set meter, which may have a large noise level and the results may be biased. Most existing studies use traffic flow data from 5-minute set meters, which are less noisy and can characterize the traffic flow more accurately [26]. Therefore, in this research, the original 1-minute set count traffic flow data were processed into 5-minute set count traffic flow data. In addition, due to the strong correlation of traffic flow data between lanes, it is also necessary to set and count traffic flow

data in one cross section in spatial dimension. This results in six basic variables of traffic flow: upstream detector average flow, speed, and occupancy and downstream detector average flow, speed, and occupancy. These six basic traffic flow variables can roughly reflect the operation status of traffic flow, and these six variables are used as alternative traffic flow parameters. In addition, the difference in flow, occupancy, and speed between adjacent lanes can characterize to some extent the lane changing behavior of vehicles in the traffic flow [17], and frequent lane-changing behavior is an important cause of freeway traffic crashes. Therefore, the differences in flow, occupancy, and speed between adjacent lanes are considered as alternative traffic flow parameters. The difference between upstream and downstream flows, occupancy, and speed can reflect the features of traffic flow changes between the upstream and downstream, reflecting the change of traffic flow status from upstream to downstream. In general, a change in traffic flow state from congestion to traffic flow or from traffic flow to congestion can significantly increase the risk of crashes [27, 28]. Therefore, the difference between upstream and downstream flows, occupancy, and speed are also included in the traffic flow alternative parameters.

In summary, in this research, the difference in volume, speed, and occupancy between adjacent lanes of upstream and downstream detectors, as well as the difference in volume, speed, and occupancy of upstream and downstream detectors were calculated and used as candidate parameters for traffic flow [17]. The variable names and variable descriptions are shown in Table 5.

2.3.2. Parameters' Selection of Weather, Road, and Time Feature. Many studies have indicated that weather features significantly affect the risk of freeway crashes [10, 18, 29, 30]. For example, rainfall may make the road slippery, vehicle braking performance would be reduced because of the road friction coefficient and become worse, the vehicle is easy to skid. For other examples, the low visibility of the outside environment will make it more difficult for drivers to drive and increase the risk of crash. Therefore, it is necessary to include weather features in the analysis of traffic crash risk.

The percentages of each weather feature in the crash and noncrash data are shown in Table 6. As can be seen from Table 6, there are significant differences in the percentages of crashe and noncrash for the three weather features of rainfall, visibility, and dew point. For example, the ratio of rainfall between 5 and 10 mm and the ratio of rainfall greater than 10 mm in the noncrash group were 5.95% and 2.44%, respectively, while the ratio of rainfall between 5 and 10 mm and the ratio of rainfall greater than 10 mm in the crash group were 12.20% and 7.93%, which were significantly higher than those in the noncrash group. However, in the two weather features of temperature and wind speed, the difference between the proportion of crash and noncrash is so small that it is almost negligible. Therefore, air temperature, dew point temperature, and visibility are included in the candidate feature parameters to be considered, and the two weather features, air temperature and wind speed, are not considered.

Because the research sections are all two-way 6 lanes and their shoulder widths do not vary much, these two road feature factors are not considered. Roadway features were selected as alternative feature variables, such as roadway width, distance between upstream and downstream coils, number of up-ramps and number of down-ramps between upstream and downstream coils, and whether they are curves.

Yin's research has shown that the crash traffic flow features on weekdays and weekends are significantly different, with weekend crashes and weekday crashes occurring under different traffic conditions [22], and in addition, different time periods are often seen as influencing factors in freeway crashes. Therefore, in this research, whether it is a working day and the period in which the crash data and noncrash data are located are included as temporal feature parameters for consideration.

In summary, the input variables considered for modeling in this research include four dimensions including traffic flow, weather, road, and time feature variables, as shown in Table 7.

3. Modeling

The XGBoost model requires little input feature variables and no feature filtering, and it can also solve the feature covariance problem well. The SHAP interpreter can be used to obtain the importance ranking of each feature and to determine whether each feature has a positive or negative effect on the occurrence of traffic crash, in addition to visualizing the interaction between each feature [30]. To analyze the influence of features such as weather, road, and time on traffic crash risk, as well as the interaction between each feature, this research introduces the SHAP interpreter to explain the freeway traffic crash risk prediction model built based on XGBoost.

3.1. Principle of XGBoost Model. Boosting algorithm is a supervised machine learning algorithm that focuses on reducing bias. Each base evaluator in the Boosting algorithm is not independent of each other but related; it is built in a certain order. The principle of the Boosting algorithm is shown in Figure 3. The working mechanism is to first construct a base evaluator based on the initial weights of the initial training set and update the weights of each sample in the training set based on the performance of this base evaluator to increase the weights of error-correcting samples so that they receive more attention in the next training. Afterwards, the training set with updated sample weights is used to construct the next base evaluator, which is repeated several times, where it is set in advance to obtain a base evaluator, and then, these base evaluators are integrated according to certain weights to form the final integrated evaluator. AdaBoost and gradient boost decision tree (GBDT) are both relatively typical Boosting algorithms. XGBoost is an improvement on GBDT, which is also a Boosting algorithm [31].

Table 5: Candidate variables of traffic flow.

Variable name	Variable description
up_q	Average upstream detector volume 10–5 minutes before the crash
up_v	Average upstream detector speed 10–5 minutes before the crash
up_o	Average upstream detector occupancy 10–5 minutes before the crash
up_dif_q	The average of the absolute value of the difference in traffic flow in the adjacent lane of the upstream detector 10–5 minutes before the crash
up_dif_v	Average of the absolute value of the speed difference between adjacent lanes of the upstream detector 10–5 minutes before the crash
up_dif_o	Average of the absolute value of the difference in adjacent lane occupancy of upstream detectors 10–5 minutes before the crash
down_q	Average downstream detector volume 10–5 minutes before the crash
down_v	Average downstream detector speed 10–5 minutes before the crash
down_o	Average downstream detector occupancy 10–5 minutes before the crash
down_dif_q	The average of the absolute value of the difference in the volume of the adjacent lanes of the downstream detector 10–5 minutes before the crash
down_dif_v	Average of the absolute value of the speed difference between adjacent lanes of the downstream detector 10–5 minutes before the crash
down_dif_o	Average of the absolute value of the difference in adjacent lane occupancy of downstream detectors 10–5 minutes before the crash
dif_q	Absolute value of the difference in volume between the upstream and downstream detectors 10–5 minutes before the crash
dif_v	Absolute value of the speed difference between upstream and downstream detectors 10–5 minutes before the crash
dif_o	Absolute value of the difference between upstream and downstream detector occupancy 10–5 minutes before the crash

Table 6: Distribution of weather features in crash and noncrash data.

Weather features		Crash		Noncrash	
		Quantity	Percentage (%)	Quantity	Percentage (%)
Rainfall	0–4	131	79.88	601	91.62
	4–10	20	12.20	39	5.95
	>10	13	7.93	16	2.44
	Sum	164	100	656	100
Visibility	<1	1	0.61	2	0.30
	1–5	67	40.85	188	28.66
	5–10	83	50.61	377	57.47
	>10	13	7.93	89	13.57
	Sum	164	100	656	100
Dew point	<10	18	11.69	36	5.49
	10–15	130	84.42	573	87.35
	15–20	15	9.74	45	6.86
	>20	1	0.65	2	0.30
	Sum	164	100	656	100
Temperature	<10	63	38.41	254	38.72
	10–20	36	21.95	141	21.49
	20–30	64	39.02	257	39.18
	>30	1	0.61	4	0.61
	Sum	164	100	656	100
Wind speed	0–9	79	48.17	319	48.63
	9–18	74	45.12	295	44.97
	18–27	9	5.49	35	5.34
	>27	2	1.22	7	1.07
	Sum	164	100	656	100

XGBoost is the abbreviation for eXtreme gradient boosting, which is a new gradient boosting algorithm proposed in 2015 [32]. Because of its high computational speed as well as prediction accuracy, XGBoost is highly preferred in machine learning competitions and is widely used in various fields.

Table 7: Input variables.

Variable category	Variable name	Variable description
Traffic flow variables	*up_q*	Average upstream detector volume 10–5 minutes before the crash
	up_v	Average upstream detector speed 10–5 minutes before the crash
	up_o	Average upstream detector occupancy 10–5 minutes before the crash
	up_dif_q	Average of the absolute value of the difference in traffic flow in the adjacent lane of the upstream detector 10–5 minutes before the crash
	up_dif_v	Average of the absolute value of the speed difference between adjacent lanes of the upstream detector 10–5 minutes before the crash
	up_dif_o	Average of the absolute value of the difference in adjacent lane occupancy of upstream detectors 10–5 minutes before the crash
	down_q	Average downstream detector volume 10–5 minutes before the crash
	down_v	Average downstream detector speed 10–5 minutes before the crash
	down_o	Average downstream detector occupancy 10–5 minutes before the crash
	down_dif_q	Average of the absolute value of the difference in traffic flow between adjacent lanes of downstream detectors 10–5 minutes before the crash
	down_dif_v	Average of the absolute value of the speed difference between adjacent lanes of the downstream detector 10–5 minutes before the crash
	down_dif_o	Average of the absolute value of the difference in adjacent lane occupancy of downstream detectors 10–5 minutes before the crash
	dif_q	Absolute value of flow difference between upstream and downstream detectors 10–5 minutes before the crash
	dif_v	Absolute value of speed difference between upstream and downstream detectors 10–5 minutes before the crash
	dif_o	Absolute value of upstream and downstream detector occupancy difference 10–5 minutes before the crash
Weather variables	*dewp*	Dew point (°C), take the value 1 (≤10); 2 (10–15); 3 (15–20); 4 (≥20)
	vis	Visibility (km), taking values 1 (<1); 2 (1–5); 3 (5–10); 4 (≥10)
	prep	Rainfall (mm), taking values 1 (0–5); 2 (5–10); 3 (≥10)
Road features	*width*	Road width (m)
	length	Upstream and downstream detector distance (km)
	curve	Whether it is a curve segment, takes the value 1 (curve segment); 0 (not a curve segment)
	ramp_up	Number of on-ramps between upstream and downstream detectors
	ramp_down	Number of down-ramps between upstream and downstream detectors
Time features	*weenkend*	Whether it is a weekday, takes the value 1 (weekend); 0 (weekday)
	hour	The time corresponding to the traffic flow data, taking values of 0 (0–1); 1 (1–2); 2 (2–3); 3 (3–4); 4 (4–5); 5 (5–6); 6 (6–7); . . .; 22 (22–23); 23 (23–24)

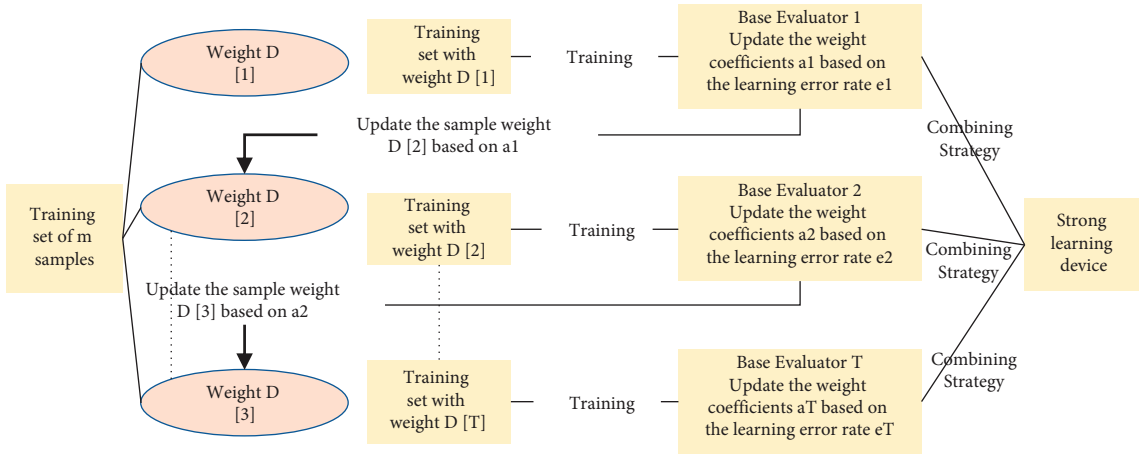

Figure 3: Schematic diagram of boosting.

The objective function of XGboost consists of two components, the loss function and the regularization term, with the following equation:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{t} \Omega(f_j), \tag{1}$$

where $l(y_i, \hat{y}_i)$ denotes the loss function of XGboost, y_i is the actual value of the i^{th} sample, and $\hat{y}_i$ denotes the predicted value of the XGboost model for the i th sample. The loss function is chosen according to the actual situation and is used to measure the difference between the actual and predicted values. $\Omega(f_j)$ is the complexity of a single tree and $\sum_{j=1}^{t}\Omega(f_j)$ is the overall complexity of the model, i.e., the sum of the complexity of all t trees, which is used as the canonical term of the objective function.

In constructing the t^{th} tree, the parameters of the first $t-1$ trees can be considered as constant terms, and thus, the objective function can be transformed into the following form:

$$\begin{aligned}\text{Obj}^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{j=1}^{t} \Omega\left(f_j\right) \\ &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{j=1}^{t} \Omega\left(f_j\right) \\ &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + \text{constant}.\end{aligned} \tag{2}$$

According to the Taylor expansion,

$$f(x+\Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2. \tag{3}$$

The approximation of the objective function (equations (2) and (3)) can be obtained as follows:

$$\begin{aligned}\text{Obj}^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] \\ + \Omega(f_t) + \text{constant},\end{aligned} \tag{4}$$

where g_i is the first-order partial derivative of the loss function and h_i is its second-order partial derivative; g_i and h_i can be written as follows:

$$\begin{aligned} g_i &= \partial_{y(t-1)} l\left(y_i, \hat{y}^{t-1}\right), \\ h_i &= \partial^2_{y(t-1)} l\left(y_i, \hat{y}^{t-1}\right). \end{aligned} \tag{5}$$

Since $l(y_i, \hat{y}^{(t-1)})$ and the constant term constant will not have an effect on the training of the t^{th} tree, the objective function can be written as follows:

$$\text{Obj}^{(t)} \simeq \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t). \tag{6}$$

The next further refinement f_t can be expressed as follows:

$$f_t(x) = w_q(x), \tag{7}$$

where w is the weight occupied by the leaf node and q represents the mapping relationship from the sample to the leaf node. According to the above equation, the canonical term of the objective function, which is the complexity of the model $\Omega(f_t)$, can be further refined as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_j^2, \tag{8}$$

where T is the number of leaf nodes of the t^{th} tree.

Combining equations (7) and (8), we can obtain

$$\begin{aligned}\text{Obj}^{(t)} &\simeq \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{i=1}^{n}\left[g_i w_q(x_i) + \frac{1}{2}h_i w_q^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{i=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T.\end{aligned} \tag{9}$$

We define the following equation:

$$G_j = \sum_{i \in I_j} g_i,$$
$$H_j = \sum_{i \in I_j} h_i. \tag{10}$$

Then, the objective function can be reduced to the following equation:

$$\mathrm{Obj}^{(t)} = \left[\mathrm{G}_j \mathrm{w}_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2 \right] + \gamma \mathrm{T}. \tag{11}$$

This is the final objective function of XGBoost.

XGBoost applies a second-order Taylor expansion to the deformation of the objective function, and the first-order partial derivatives and second-order partial derivatives of the loss function are also used in this process, which makes the gradient descent faster and more accurate. At the same time, this feature also allows XGBoost to optimize the model and select parameters based only on the values of the input samples without specifying the loss function in advance, which greatly improves the applicability of the model. In addition, the inclusion of the regular term in the objective function can control the complexity of the model, which can effectively model to prevent overfitting and reduce the variance of the model.

3.2. SHAP-Based Model Interpretation Method. Except for linear models and some simple models, such as decision trees and plainly Bayes, most machine learning models are black-box models and cannot be interpreted. Explainable models usually have insufficient accuracy, while some black-box models tend to have high predictive accuracy [33]. Scholars have explored how to improve the prediction accuracy of models while making them interpretable. Lundberg and Lee proposed the SHAP method in their paper "a unified approach to interpreting model predictions" published in 2017, which allows the interpretation of model predictions [34]. The SHAP method mainly relies on the Shapley value, which was proposed by Shapley in 1953, and the core idea is cooperative game theory. Later, Lundberg et al. proposed TreeSHAP in 2018 in the paper "consistent individualized feature attribution for tree ensembles" [35]. TreeSHAP is a variation of the SHAP method, and the application objects are mainly tree models, such as random forests and XGboost. TreeSHAP is much faster and can handle the correlation of features well [35–38].

The model makes predictions based on the features of each sample, and each sample corresponds to a prediction result, and the value assigned to each feature that contributes to the prediction result is called the Shapley value. The Shapley value has additivity, i.e.,

$$y_i = \overline{y} + \sum_{j=1}^{k} f\left(x_i, j\right), \tag{12}$$

where x_i is the i^{th} data sample, each data sample has k features, y_i is the prediction value corresponding to x_i, $\overline{y}$ is the baseline of the model, and $f(x_i, j)$ is the Shapley value of the j^{th} feature of x_i. It is possible to determine whether the feature has a positive or negative effect on the prediction result based on the positivity or negativity of $f(x_i, j)$. Taking the binary classification problem as an example, if $f(x_i, j)$ is positive, it means that this feature will increase the likelihood of positive classes appearing, and if $f(x_i, j)$ is negative, it means that this feature will have a reverse effect on the model, increasing the likelihood of negative classes appearing.

The SHAP interpreter can evaluate the importance of features, unlike feature_importance in models such as XGboost, which is mainly based on the degradation of model performance, while the SHAP interpreter is based on the magnitude of feature attribution. The SHAP interpreter ranks the importance of a feature based on the mean of the absolute value of the degree of influence of this feature on the prediction result, i.e., the mean of the Shapley absolute values, as in equations (3–13). The larger the mean value is, the higher the importance of the feature and the greater the impact on the prediction result are:

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right|, \tag{13}$$

where I_j is the feature importance of the j^{th} feature and $\phi_j^{(i)}$ is the Shapley value of the j th feature of the i^{th} sample.

The SHAP interpreter can also calculate the interactions between features. The interaction is an additional combined feature effect after considering the individual feature effects. The Shapley interaction index is defined in the following form:

$$\phi_{i,j} = \sum_{S \subseteq \{i,j\}} \frac{|S|!\,(M = |S| - 2)!}{2\,(M-1)!} \delta_{ij}(S). \tag{14}$$

When $i \neq j$,

$$\delta_{ij}(S) = f_x(S \cup \{i, j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S). \tag{15}$$

The above equation yields the pure interaction effect after removing the main effect of the features and then averaging the values of all possible feature coalitions S. In calculating the SHAP interaction values for all features, a matrix of dimension $M \times M$ is obtained for each sample, where M is the number of features.

3.3. XGBoost Parameter Selection. In this research, the value of area under curve (AUC) was used as the model evaluation index. Accuracy (the ratio of predicted samples to all samples) is the most commonly used model evaluation index in classification algorithms, but this index is not applicable to nonequilibrium data sets. For example, in this research, the crash data used the following: noncrash data = 1:4 in the data set; if the model judges all labels as 0, that is, noncrash, the accuracy rate of the model can reach 80%, which can neither identify the crash data nor reflect the prediction effect of the model, and is obviously meaningless. Therefore, for non-equilibrium data sets, it is necessary not only to ensure a high accuracy of the model, but also to ensure that the model has

a good performance in the prediction of the majority and minority classes. Based on the above considerations and existing studies, this research takes the area under the curve (AUC) of receiver operating characteristic (ROC) as the evaluation index of model prediction performance.

As shown in Figure 4, the closer the ROC curve is to the upper left corner, the better the predictive performance of the classifier is. Sometimes the ROC curves of two models may cross, at which time the AUC value of the area under the ROC curve is needed to evaluate which is better. The closer the AUC is to 1, the better the predictive performance of the model is.

The parameters "*n_estimatores*," "*subsample*," "*max_depth,*" and "*learning_rate*" are selected as the parameters that have a great impact on the classification performance of the model, where *n_estimatores* is the number of weak classifiers; if this value is too large, the model will be over-fitted, but if it is too small, it may be under-fitted, and its general value is 100–300; *subsample is the ratio of the data used in training to the total training set and generally takes the value of* 0.5–1; *learning_rate* is the learning rate; when updating the feature weights, the learning rate shrinkage can prevent overfitting when updating the feature weights and make the boosting process more conservative; its general value is 0.01–0.2; *max_depth* is the maximum depth of the tree, increasing *max_depth* will increase the complexity of the tree model and make the model fall into overfitting, but if the tree depth is not enough, it will also appear underfitting, its general value is 1–11; *scale_pos_weight* is the weight of positive samples, and when the sample dataset is a non-equilibrium dataset, increasing the weight of positive samples will improve the recall of the model, i.e., the ability to capture the incident data, so it is possible to improve the overall prediction performance of the model. Since this dataset is an unbalanced dataset and the ratio of noncrash samples: incident samples is 4:1, the value of *scale_pos_weight* is set to 1–10.

The explanatory variables of the model are as shown in Table 7; the category label is whether a crash occurred, the sample label for a crash occurred is noted as 1, and the sample label for no crash occurred is noted as 0. The XGBClassifier is used to build the initial XGboost classification model, and then, the GridSearchCV of the sklearn machine learning package in Python is used to implement the grid search, and *n_estimatores* are searched in the range of 100–300 in steps of 10; the search range of *learning_rate* is 0.01–0.2, using the linspace function in numpy to generate 50 random data in the range of 0.01–0.2 as the actual search range; the search range of *max_depth* is 1–11, with 1 as the step; the search range of *subsample* is 0.5–1 in steps of 0.1; the search range of *scale_pos_weight* is 1–10 in steps of 1. The optimal combination of parameters obtained from the grid search is $n\_estimatores = 200$, $max\_depth = 9$, $learning\_rate = 0.02$, $subsample = 0.7$, and $scale\_pos\_weight = 1$.

Using the optimal combination of parameters, 70% of the data samples are used as the training set in the model, the remaining 30% of the data samples are used as the test set, and the XGBoost classification model is rebuilt; then, the SHAP interpreter is built based on this model to interpret the model and visualize the interpretation results.

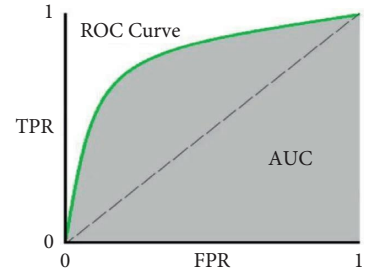

Figure 4: ROC curve.

4. Results and Discussion

4.1. Extraction and Analysis of Major Traffic Crash Precursors. The SHAP interpreter ranks the importance of each feature based on the mean value of the absolute value of each feature's influence on a crash occurrence, with the greater the mean value, the higher the importance of the feature, i.e., the greater the influence on whether a crash will occur. Figure 5 shows the ranking of the importance of each feature based on the SHAP interpreter, and it can be found that the feature distance between upstream and downstream detectors (*length*) has the highest importance, which means that this feature contributes the most to whether a crash occurs or not, and far exceeds the other variables. In addition, the absolute value of the difference in speed between upstream and downstream detectors (*dif_v*), the average speed of downstream detectors (*down_v*), the average speed of upstream detectors (*up_v*), and the average value of the absolute value of the difference in flow between adjacent lanes of upstream detectors (*up_dif_q*) in the traffic flow variables contribute significantly to whether a crash occurs. Among the weather characteristics, rain (*rain*) has the highest contribution to whether a crash occurs or not. In addition, the hour (*hour*) corresponding to the traffic flow data are also an important factor influencing the occurrence of crashes.

The SHAP summary plot (Figure 6) combines feature importance with feature effects. Each point on the summary plot corresponds to the Shapley value of a feature of a sample. The position on the *y*-axis is determined by the importance of the features, which is ordered according to their importance, consistent with Figure 4, and the position on the *x*-axis is determined by the Shapley value, with the Shapley value equal to 0 as the central axis, and the Shapley value to the right of the central axis is greater than 0; at the left of the central axis, the Shapley value is less than 0. The color represents the size of the feature value; the redder the color, the larger the corresponding feature value, and the bluer the color, the smaller the corresponding feature value. The overlap point is jittered upwards on the *y*-axis. Therefore, the SHAP summary plot gives an idea of the distribution of the Shapley values for each feature. The following conclusions can be drawn from Figure 5:

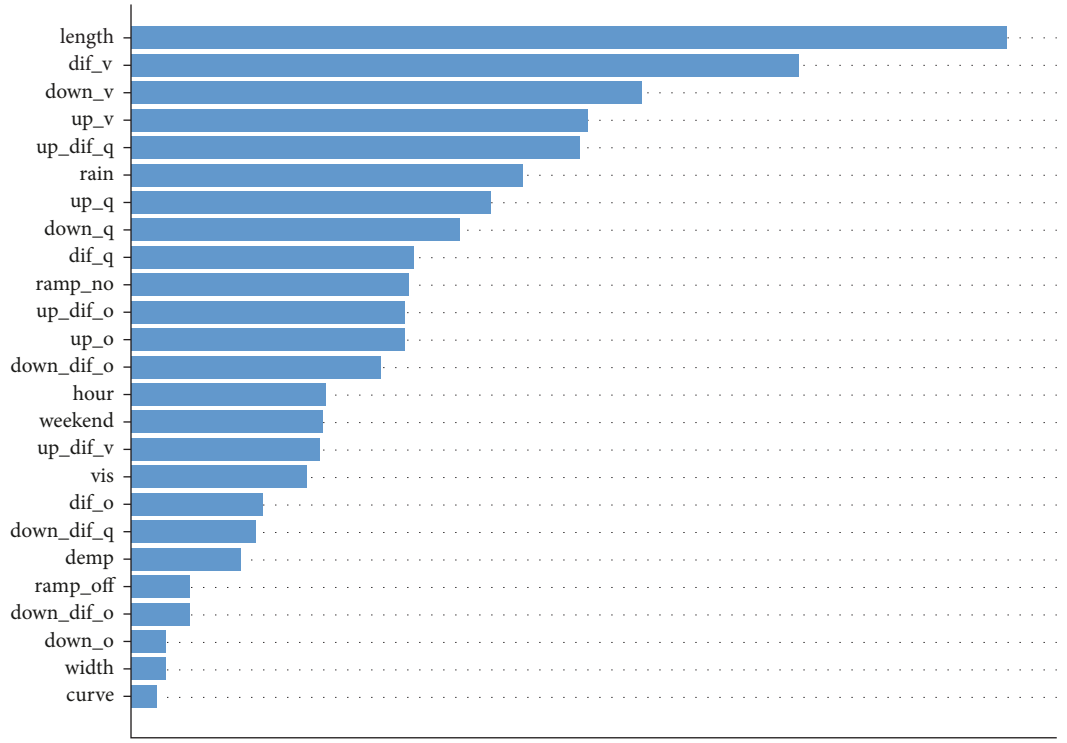

Figure 5: Ranking chart of variables' importance.

(1) It is obvious that the red points of the feature of the distance between upstream and downstream detectors (*length*) are basically located on the right side of the Shapley value equal to 0, and the blue points are basically located on the left side of the Shapley value equal to 0. This means that the smaller the distance between upstream and downstream detectors, the lower the risk of traffic crash, and the larger the distance between upstream and downstream detectors, the higher the risk of traffic crash. The reason for this situation may be because the greater the distance between the upstream and downstream detectors, the longer the length of the section between them and the greater the likelihood that the crash point will fall on this section of the freeway, in which traffic crash is more likely to occur.

(2) For the absolute value of the upstream and downstream detector velocity difference (*dif_v*), the red points are mostly located to the right of the Shapley value equal to 0, and the blue points are basically located to the left of the Shapley value equal to 0. This indicates that the absolute value of the speed difference between the upstream and downstream detectors is positively correlated with the crash risk, and the larger it is, the larger the corresponding Shapley value is and then the greater the likelihood of a crash is. This may occur because the sudden change in speed increases the probability of traffic crash when the freeway goes from congested to free flow or from free to congested flow.

(3) For the downstream detector average velocity (*down_v*) and the downstream detector flow average (*down_q*), the red points are mainly concentrated to the left of the Shapley value equal to 0, and the blue points are concentrated to the right of the Shapley value equal to 0. This indicates that the average speed downstream is negatively related to the crash risk and has a negative effect on the occurrence of crash; the smaller it is, the higher the traffic crash risk is.

(4) The average upstream detector velocity (*up_v*) has no significant positive or negative correlation with the occurrence of crashes, but its high values, i.e., the red points are mostly distributed to the left of the Shapley value equal to 0, indicating that crashes are more likely to occur when the average upstream velocity is lower. The low values, i.e., the blue points, are more scattered, and the corresponding Shapley values are both greater than 0 and less than 0. There is no obvious linear relationship with the occurrence of crashes.

(5) For the average value of the absolute value of the difference in flow between adjacent lanes of upstream detectors (*up_dif_q*), most of the red points are distributed to the right of the Shapley value equal to 0; that is, the larger the absolute value of the difference in flow between adjacent lanes of upstream detectors, the more likely it is that a traffic crash will occur. This may be due to the fact that vehicle lane changing behavior is more likely to occur when there is a large difference in traffic flow

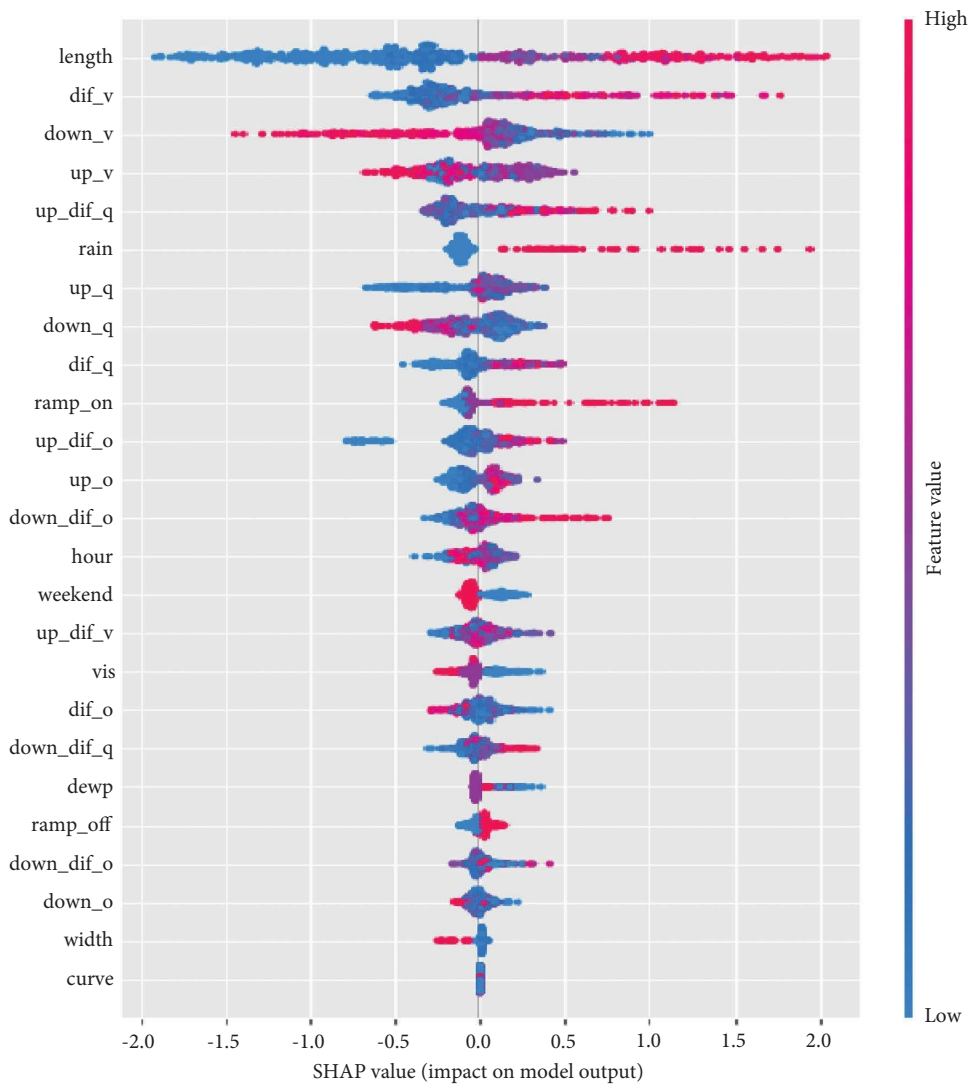

Figure 6: SHAP summary plot.

between adjacent lanes upstream, and frequent lane changing behavior can lead to an increased risk of crash.

(6) Among the weather features, rainfall (*rain*) has a greater influence on the occurrence of traffic crashes, while visibility (*vis*) and dew point (*dewp*) have less influence on the occurrence of traffic crashes. The amount of rainfall is positively correlated with the risk of traffic crash, and the increase of rainfall will increase the risk of traffic crashes. Visibility is negatively correlated with traffic crash risk; the lower the visibility, the higher the risk of a traffic crash.

(7) Average value of the upstream detector flow (*up_q*), absolute value of the upstream and downstream detector flow difference (*dif_q*), absolute value of the upstream detector adjacent lane occupancy difference (*up_dif_o*), absolute value of downstream detector adjacent lane speed difference (*down_dif_v*), and absolute value of downstream detector adjacent lane flow difference (*down_dif_q*), most of the high values of these features are distributed to the right of the Shapley value equal to 0, and the low values are to the left of the Shapley value equal to 0. All of them are positively correlated with traffic crash risk. As the values of these features increases, the likelihood of traffic crash increases.

(8) There is no obvious positive or negative relationship between the absolute value of the upstream detector adjacent lane speed difference (*up_dif_v*) and the time corresponding to the traffic flow data (*hour*) and the traffic crash risk, but its low value is mostly near the left side of the Shapley value equal to 0, indicating that the traffic crash risk is lower when the absolute value of the upstream adjacent lane speed difference is smaller or the time corresponding to the traffic flow is in the early morning.

4.2. Analysis of the Impact of Eigenvalues on Traffic Crash. To understand the exact form of the impact of each eigenvalue on the traffic crash risk, it is necessary to see the SHAP dependence plot. The x-axis of the SHAP dependence plot corresponds to the feature values and the y-axis to the corresponding Shapley values. The SHAP dependence plot can also represent the interaction effect between features.

The SHAP dependence plots of the distance between the upstream and downstream detectors, the absolute value of the speed difference between the upstream and downstream detectors, the downstream speed, the upstream speed, visibility, and the absolute value of the flow difference between the adjacent lanes of the upstream detectors, as shown in Figure 7, are used as examples to analyze the relationship between these features and the traffic crash risk and the interaction that exists among each feature.

As can be seen from Figure 7(a), the greater the distance between the upstream and downstream detectors, the higher the Shapley value, and the higher the likelihood of a crash. With 2 km as the dividing point, when the distance between upstream and downstream detectors is greater than 2 km, the Shapley value is basically positive and has a positive driving effect on the occurrence of a crash; when the distance between upstream and downstream detectors is less than 2 km, the Shapley value is basically negative and has a negative driving effect on the occurrence of a crash. The dispersion in the vertical direction on the right side of Figure 7(a) shows that the distance between the upstream and downstream detectors interacts most strongly with the mean downstream velocity. For the distance between upstream and downstream detectors less than 2 km, a higher downstream velocity increases the risk of crash, but for the distance between upstream and downstream detectors greater than 2 km, a higher downstream velocity decreases the risk of crash.

From Figure 7(b), it can be seen that the absolute value of upstream and downstream detector speed difference and crash risk generally show the relationship that the greater the value, the greater the risk of traffic crash, but when the absolute value of upstream and downstream detector speed difference is greater than 20 and less than 40, the distribution of Shapley value spans a larger range, and when the absolute value of upstream and downstream detector speed difference is greater than 40, the Shapley value shows a decreasing trend and the risk of traffic crash decreases. The dispersion in the vertical direction on the right side of Figure 7(b) shows the strongest interaction between the absolute value of upstream and downstream detector velocity difference and the distance between upstream and downstream detectors. When the absolute value of upstream and downstream detector velocity difference is less than 20, a larger upstream and downstream spacing will have a positive driving effect on the occurrence of crash, but when the absolute value of upstream and downstream detector velocity difference is greater than 20, a smaller upstream and downstream distance will increase the risk of crash.

The overall trend in Figure 7(c) is decreasing, indicating that the higher the downstream velocity, the higher the possibility of a crash. With 90 as the dividing point, when the downstream velocity is greater than 90, the Shapley value is basically negative, which has a negative impact on the occurrence of crashes; when the downstream velocity is less than 90, the Shapley value is basically positive, which has a positive impact on the occurrence of crash. The dispersion in the vertical direction on the right side of Figure 7(c) shows that the downstream velocity interacts most strongly with whether or not it is a curve, with a curve decreasing the likelihood of a crash at downstream velocities greater than 90 and increasing the likelihood of a crash at downstream velocities less than 90.

From Figure 7(d), it can be seen that the traffic crash risk increases with the increase of upstream speed when the upstream speed is less than 85, while it decreases with the increase of upstream speed when the upstream speed is greater than 85. The dispersion in the vertical direction on the right side of Figure 7(d) shows the strongest interaction between the upstream speed and the distance between the upstream and downstream detectors. At upstream speeds less than 85, a larger upstream and downstream detector spacing increases the likelihood of a traffic crash, while at upstream speeds greater than 85, a larger upstream and downstream detector spacing decreases the likelihood of a traffic crash.

As can be seen from Figure 7(e), the Shapley values are mostly positive at low visibility, indicating that the low visibility increases the crash risk; at higher visibility, the Shapley values are mostly negative, indicating that the crash risk decreases when the visibility is better. The dispersion in the vertical direction on the right side of Figure 7(e) shows the strongest interaction between visibility and curves, where curves increase the likelihood of traffic crash when visibility is low but decrease the likelihood of crash when visibility is high instead.

From Figure 7(f), it can be seen that, as the absolute value of the upstream detector adjacent lane flow difference increases, the corresponding Shapley value decreases, then increases, then decreases, and then increases again, resembling a "W" shape, indicating that the upstream adjacent lane flow difference and traffic crash risk are not simply positively or negatively correlated. The dispersion in the vertical direction on the right side of Figure 7(f) shows the strongest interaction between the difference in upstream adjacent lane flow and the number of on-ramps, where a low number of on-ramps reduces the risk of traffic crash when the difference in upstream adjacent lane flow is small, and the presence of on-ramps increases the risk of traffic crash when the difference in upstream adjacent lane flow is large, indicating that the lane changing behavior is more likely to cause traffic crash near the merging area.

Based on the above analysis, it can be found that not only traffic flow variables but also the features of road, weather, and time all have an impact on traffic crash risk, and there is an interaction between each feature. Since SHAP-based model interpretation can only visualize second-order feature interactions, higher-order feature interactions cannot be demonstrated, but the existence of higher-order feature interactions is possible.

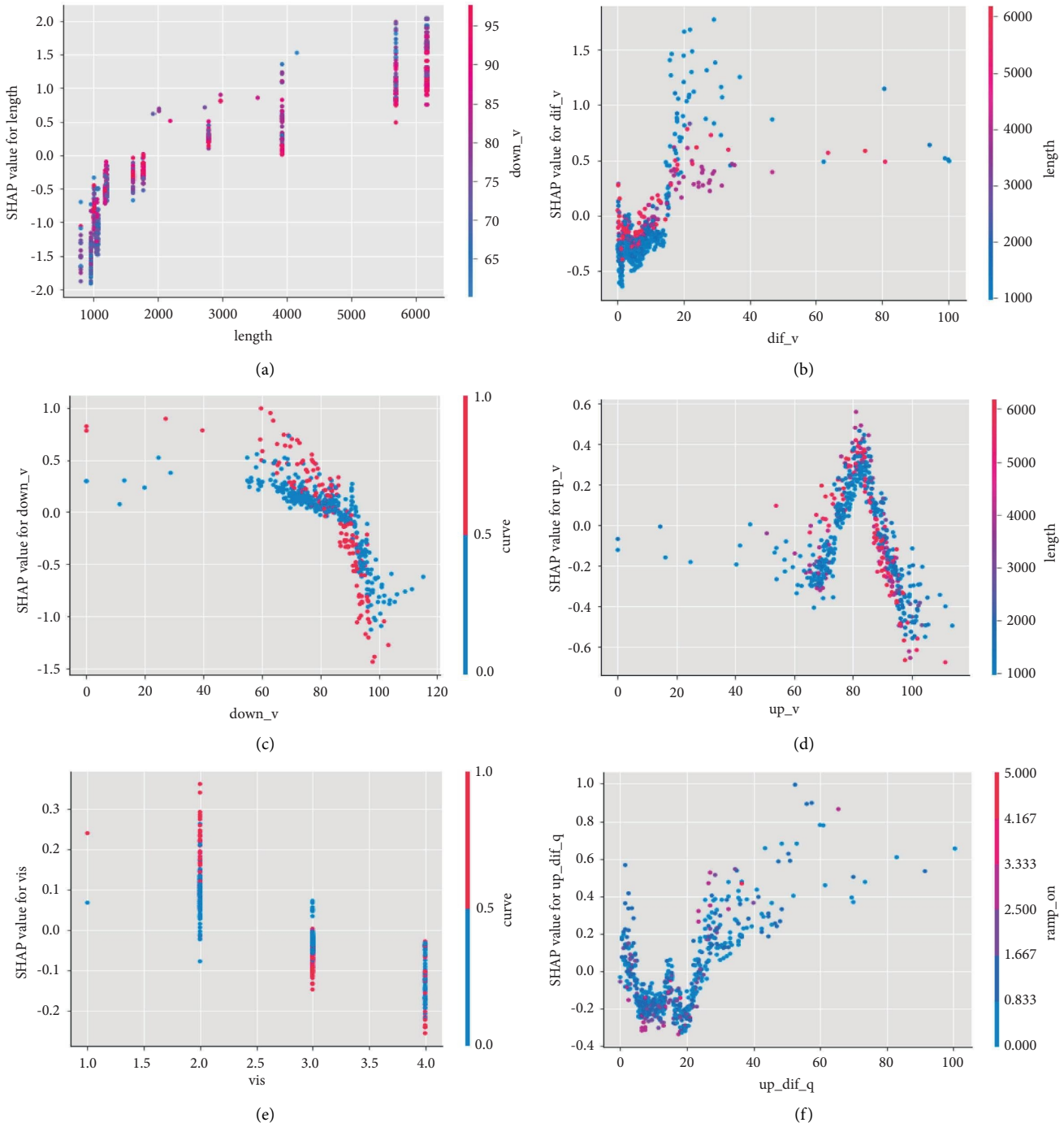

Figure 7: SHAP dependence plot.

5. Conclusions

In this research, based on the dataset of the Beijing section of the Beijing-Harbin freeway in China, XGboost was applied to model, and the model was interpreted via the SHAP interpreter to obtain the ranking of the importance of each feature on the impact of crash risk, analyze the specific relationship between each feature and the traffic crash risk, and further explore and visualize the two-dimensional interaction effect between each feature. The main research findings are as follows:

(1) The distance between upstream and downstream detectors in road features is the most important for crash risk, which is far more than other variables. The absolute values of upstream and downstream speed difference, downstream average speed, upstream average speed, and upstream adjacent lane flow difference in the traffic flow variables also contribute much to the traffic crash occurrence. Among the weather features, rainfall has the highest contribution to the traffic crash occurrence. Among the time features, the time corresponding to the traffic flow data is also an important factor influencing the occurrence of crashes.

(2) The variables of distance between the upstream and downstream detectors, absolute value of upstream and downstream speed difference, absolute value of upstream adjacent lane flow difference, rainfall, upstream flow average, absolute value of upstream and downstream flow difference, absolute value of upstream adjacent lane occupancy difference, and absolute value of downstream adjacent lane speed difference are positively correlated with traffic crash risk; the variables of downstream average speed, downstream flow average, and visibility are negatively correlated with traffic crash risk.

(3) The effects of traffic flow variables, road characteristics, weather, and temporal features on traffic crash are not independent of each other, but there is a relatively complex interaction effect. Second, the order characteristic interactions between these factors exist, and there's a possibility that higher-order characteristic interactions also exist.

(4) The approach proposed in this research can deeply excavate the mechanism of dynamic traffic crash occurrence on freeways, and the research results can be applied to real-time traffic risk monitoring on freeways, so as to provide the theoretical basis for the traffic crash prediction work and warning technology. Simultaneously, it is possible to develop targeted road management measures based on indicators such as real-time road traffic flow operating conditions, weather, and road features, to identify the risk of crashes in a timely manner, and to improve the safety and operational efficiency of freeways.

(5) The dynamic traffic flow data applied in this research are based on the microwave radar traffic information detector collection, and the obtained traffic flow features are not comprehensive. Subsequently, we may obtain the indicator parameters such as lane changing behavior and following behavior of vehicles through video detectors or collect parameters such as acceleration and headway time distance of vehicles based on real driving data to improve the required traffic flow information for modeling.

Data Availability

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Open Project of Shandong Key Laboratory of Highway Technology and Safety Assessment (SH202105).

References

[1] Y. Yang, Z. Yuan, and R. Meng, "Exploring traffic crash occurrence mechanism toward cross-area freeways via an improved data mining approach[J]," *Journal of Transportation Engineering Part A Systems*, vol. 148, no. 9, Article ID 04022052, 2022.

[2] Who, *Global Status Report on Road Safety[R]*, World Health Organization, Geneva, Switzerland, 2018.

[3] H. Wumaier, J. Gao, and J. Zhou, "Short-term forecasting method for dynamic traffic flow based on stochastic forest algorithm[J]," *Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology*, vol. 39, no. 2, pp. 1501–1513, 2020.

[4] J. Carson and F. Mannering, "The effect of ice warning signs on ice-accident frequencies and severities," *Accident Analysis and Prevention*, vol. 33, no. 1, pp. 99–109, 2001.

[5] A. Kassu and M. Hasan, "Factors associated with traffic crashes on urban freeways," *Transport Engineer*, vol. 2, Article ID 100014, 2020.

[6] Y. Yang, N. Tian, Y. Wang, and Z. Yuan, "A parallel FP-growth mining algorithm with load balancing constraints for traffic crash data[J]," *International Journal of Computers, Communications and Control*, vol. 17, no. 4, p. 4806, 2022.

[7] X. Zhang, H. Wen, T. Yamamoto, and Q. Zeng, "Investigating hazardous factors affecting freeway crash injury severity incorporating real-time weather data: using a Bayesian multinomial logit model with conditional autoregressive priors," *Journal of Safety Research*, vol. 76, pp. 248–255, 2021.

[8] Q. Zeng, W. Hao, J. Lee, and F. Chen, "Investigating the impacts of real-time weather conditions on freeway crash severity: a Bayesian spatial analysis[J]," *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, pp. 27681/1–15, 2020.

[9] H. Wen, X. Zhang, Q. Zeng, and N. N. Sze, "Bayesian spatial-temporal model for the main and interaction effects of roadway and weather characteristics on freeway crash

incidence[J]," *Accident Analysis and Prevention*, vol. 132, Article ID 105249, pp. 105249.1–105249.6, Nov. 2019.
[10] F. Malin, I. Norros, and S. Innamaa, "Accident risk of road and weather conditions on different road types," *Accident Analysis and Prevention*, vol. 122, pp. 181–188, JAN. 2019.
[11] T. F. Golob and W. W. Recker, "A method for relating type of crash to traffic flow characteristics on urban freeways," *Transportation Research Part A: Policy and Practice*, vol. 38, no. 1, pp. 53–80, 2004.
[12] Z. Zheng, S. Ahn, and C. M. Monsere, "Impact of traffic oscillations on freeway crash occurrences," *Accident Analysis and Prevention*, vol. 42, no. 2, pp. 626–636, 2010.
[13] C. Xu, P. Liu, W. Wang, and Z. Li, "Evaluation of the impacts of traffic states on crash risks on freeways," *Accident Analysis and Prevention*, vol. 47, pp. 162–171, 2012.
[14] R. Yu, M. Abdel-Aty, and M. Ahmed, "Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors," *Accident Analysis and Prevention*, vol. 50, pp. 371–376, 2013.
[15] R. Yu and M. Abdel-Aty, "Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data," *Safety Science*, vol. 63, pp. 50–56, 2014.
[16] D. Sun, Y. Ai, Y. Sun, and L. Zhao, "A highway crash risk assessment method based on traffic safety state division," *PLoS One*, vol. 15, no. 1, Article ID e0227609, 2020.
[17] Y. Yang, K. He, Y. P. Wang, Z. Z. Yuan, Y. H. Yin, and M. Z. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," *Physica A: Statistical Mechanics and Its Applications*, vol. 595, Article ID 127083, 2022.
[18] C. Lee, Y. J. Park, and M. A. Abdelaty, "Effects of lane-change and car-following-related traffic flow parameters on crash occurrence by lane[C]," *Transportation Research Board Meeting*, 2009,https://trid.trb.org/view/881429.
[19] Z. Christoforou, S. Cohen, and M. G. Karlaftis, "Identifying crash type propensity using real-time traffic data on freeways," *Journal of Safety Research*, vol. 42, no. 1, pp. 43–50, 2011.
[20] 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.
[21] K. Yang, X. Wang, and R. Yu, "A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation," *Transportation Research Part C: Emerging Technologies*, vol. 96, pp. 192–207, 2018.
[22] Y. Yin, *Freeway Real-Time Crash Risk Analysis and Prediction Considering the Characteristics of Traffic Flow under Different Time*, Beijing Jiaotong University, Haidian, Beijing, 2021.
[23] V. E. Miller, T. S. Carey, B. N. Gaynes et al., "Innovations in suicide prevention research (INSPIRE): a protocol for a population-based case–control study[J]," *Injury Prevention*, vol. 330, pp. 1–8, 2022.
[24] M. D. T. Hitchings, A. L. Joseph, E. D. Natalie, A. I. Ko, O. T. Ranzani, and J. R. Andrews, "Use of recently vaccinated individuals to detect bias in test-negative case–control studies of COVID-19 vaccine effectiveness[J]," *Epidemiology*, vol. 33, no. 4, 2022.
[25] H. Hassan and M. A. Abdelaty, "Exploring visibility-related crashes on freeways based on real-time traffic flow data[C]," *Transportation Research Board Meeting*, 2011,https://trid.trb.org/view/1091668.
[26] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Analysis and Prevention*, vol. 45, no. 2, pp. 478–486, 2012.
[27] C. Lee, F. Saccomanno, and B. Hellinga, "Analysis of crash precursors on instrumented freeways," *Transportation Research Record*, vol. 1784, no. 1, pp. 1–8, 2002.
[28] M. Hossain and Y. Muromachi, "Understanding crash mechanisms and selecting interventions to mitigate real-time hazards on urban expressways," *Transportation Research Record*, vol. 2213, no. 1, pp. 53–62, 2011.
[29] M. Abdelaty, A. Pande, and N. Uddin, "Relating crash occurrence to freeway loop detectors data, weather conditions and geometric factors[J]," *Geometric Design*, 2005,https://trid.trb.org/view/768065.
[30] Y. Yang, K. Wang, Z. Yuan, and D. Liu, "Predicting freeway traffic crash severity using XGBoost-bayesian network model with consideration of features interaction[J]," *Journal of Advanced Transportation*, vol. 2022, Article ID 4257865, 16 pages, 2022.
[31] S. Rath, G. S. Priyanka, N. Nagappan, and T. Tiju, "Discovery of direct band gap perovskites for light harvesting by using machine learning[J]," *Computational Materials Science*, vol. 210, Article ID 111476, 2022.
[32] A. Dadj, A. Lbdc, B. Jobda, A. N. Rodolfo, and G. Marcelo, "Automatic method for classifying COVID-19 patients based on chest X-ray images, using deep features and PSO-optimized XGBoost - ScienceDirect[J]," *Expert Systems with Applications*, vol. 183, Article ID 115452.
[33] A. Herold, "Application of machine learning for the Higgs boson mass reconstruction using ATLAS data," CERN Accelerating science, Geneva, Switzerland, 2022.
[34] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions[C]," in *Proceedings of the Conference on NeuralInformation Processing Systems*, pp. 4765–4774, Long Beach, CA, USA, December 2017.
[35] S. M. Lundberg, G. G. Erion, and S. I. Lee, "Consistent individualized feature attribution for tree ensembles[J]," 2018, https://arxiv.org/abs/1802.03888.
[36] X. Chen, S. Wu, C. Shi et al., "Sensing data supported traffic flow prediction via denoising schemes and ann: a comparison," *IEEE Sensors Journal*, vol. 20, no. 23, pp. 14317–14328, 2020.
[37] X. Chen, J. Ling, S. Wang, Y. Yang, L. Luo, and Y. Yan, "Ship detection from coastal surveillance videos via an ensemble Canny-Gaussian-morphology framework," *Journal of Navigation*, vol. 74, no. 6, pp. 1252–1266, 2021.
[38] T. Cardoso, P. Ballester, F. P. Moreira et al., "Identifying nonlinear patterns of 5-year suicide risk incidence in youth: a gradient tree boosting and SHAP study," *Biological Psychiatry*, vol. 89, no. 9, p. S283, 2021.