Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 9123399, 16 pages
https://doi.org/10.1155/2022/9123399

Research Article

Risk Measurement Model for Vehicle Group Based on Temporal and Spatial Similarities

Huiying Wen,[1,2] **Xiaohua Chen**,[1] **and Sheng Zhao**[1]

[1]*School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China*
[2]*Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Southeast University, Nanjing 211189, China*

Correspondence should be addressed to Sheng Zhao; ctszhao@scut.edu.cn

Received 5 January 2022; Revised 23 June 2022; Accepted 15 July 2022; Published 16 August 2022

Academic Editor: Arkatkar Shriniwas

Vehicle rear-end collisions are primarily caused by tight car following in a continuous traffic flow, as well as a driver's incorrect perception of the traffic environment ahead and delayed response. To facilitate an investigation pertaining to rear-end collision mechanisms and accurately measure the risk, the concept of a vehicle group is introduced herein. A risk measurement model for a vehicle group (RMVG) based on temporal and spatial similarities is proposed. First, vehicles are categorized based on their temporal and spatial similarities. Risk measurement metrics are defined based on the traffic composition, movement state, and conflict extent. Subsequently, vehicle group risk identification and risk measurement models based on an isolation forest are established. the rear-end collision risk of the vehicle groups is analyzed both qualitatively and quantitatively. Finally, the RMVG is tested using the vehicle trajectory data set of Longpan South Road, Nanjing City, Jiangsu Province, China, and the results are compared with those of a support vector machine and local outlier factor. The results show that the accuracy of the RMVG is higher than those of other models: its accuracy rate and specificity are 95.68% and 88.89%, respectively, whereas its false alarm rate is only 3.47%.

1. Introduction

Owing to the continuously increasing demand for travel and car ownership, traffic collisions have become more prevalent, thus posing severe safety risks to road users. According to the Global Status Report on Road Safety, road traffic crashes, which is the eighth leading cause of death, caused 1.3 million fatalities in 2016 worldwide [1]. A significant proportion of road-traffic crashes involve rear-end collisions. According to the National Highway Traffic Safety Administration, rear-end collisions in the United States constituted 32.5% of all crashes in 2019 [2]. In Shanghai, China, approximately 20% of all road crashes were rear-end collisions: 49% are elevated expressway collisions, and 67% are tunnel collisions [3]. Therefore, rear-end collisions must be prevented to improve road traffic safety. However, most existing studies focus on the macro-evaluation of traffic flow [4, 5] or the microrecognition of individual vehicles [6, 7]. These methods cannot simultaneously consider the effects of individual behaviors and interactions among surrounding vehicles on driving safety. Owing to the significant increase in traffic volume, the close car-following phenomenon during driving and the group response phenomenon of some traffic flows to random interference have become more evident. These factors contribute significantly to vehicle rear-end collisions and pile-ups and cannot be considered as merely a microscopic or macroscopic traffic flow. Therefore, comprehensive investigations must be performed from a new perspective. Herein, a risk measurement model for a vehicle group (RMVG) based on temporal and spatial similarities is proposed. This model considers a vehicle group as an object for the identification and quantification of rear-end collision risk.

2. Literature Review

2.1. Conventional Traffic Safety Analysis Methods. Traffic safety has been investigated extensively, and the results obtained have been effectively applied to solve practical

engineering problems [8, 9]. Conventional methods primarily use crash data and statistical techniques for traffic safety analysis. Generally, the crash rate or death rate is combined with possible influencing factors, including speed [10], population [11], traffic volume [12], and land use [13], to analyze traffic safety. For instance, the Smeed model uses regression analysis to combine population, motor vehicle ownership, and crash fatalities to analyze traffic safety [14]. Similarly, Ng used regression analysis to predict traffic collisions and identify high risks [15]. Rabbani established a time-series model to predict collision rates based on seasonality in historical collision data [16]. Dong used a mixed logit model to analyze the effects of traffic flow and road environment on single-vehicle and multivehicle collisions [17].

2.2. Traffic Safety Analysis Method Based on Surrogate Safety Measures (SSMs). Many studies involving traffic safety analysis have been conducted based on historical collision data. However, the use of collision data poses several problems, including difficult access, data scarcity, and a long data acquisition period [8]. Therefore, SSMs were proposed for safety evaluation. Hayward introduced the time-to-collision (TTC) concept [18]. Similarly, YDEN proposed the concept of postencroachment time [19]. Because the TTC concept does not apply to cases with a speed difference of 0 km/h, Balas presented the inverse time-to-collision (TTC^{-1}) concept [20, 21]. Tarko investigated the causal relationship between conflict and collisions, where the results proved that SSMs can be used as a basis for safety evaluation [22, 23]. SSMs based on temporal logic are widely used in investigations pertaining to rear-end collisions [24, 25] and lane-changing safety [26, 27]. Meanwhile, some SSMs are based on distance logic, in which the safety distance is used as a risk evaluation factor. Wu used the stopping sight distance (SSD) to establish a rear-end collision risk index and identified the risk of vehicles in a foggy environment [6]. Hema evaluated road traffic safety by comparing the SSDs of preceding and following vehicles [28].

2.3. Modern Traffic Safety Analysis Methods. Owing to the continuous development of big data technology and machine learning, several researchers have used a significant amount of microtraffic data combined with intelligent learning and SSMs to identify and quantify traffic safety hazards. Zhanyong used the support vector machine (SVM) method based on the maximum classification interval to train and optimize a complex traffic accident black spot model [29]. Torok used a one-class SVM method to detect human-related emergencies during driving; this system can assist self-driving cars in generating risk warnings [30]. Djenouri used the local outlier factor (LOF) algorithm to analyze the effects of events, particular weather conditions, or planning decisions on traffic flow in an urban area [31]. Elassad established a real-time collision-prediction fusion framework that integrates Bayesian learners, k-nearest neighbors, an SVM, and a multilayer perceptron to predict traffic collisions [32]. Shen combined Bayesian deep learning and Gaussian mixture clustering based on SSMs to predict the risk of road traffic collisions [33]. In addition, back propagation neural networks [34, 35], generative adversarial networks [36, 37], convolutional neural networks [38], XGBoost [39], long short-term memory [40], and random forests [41] have been widely used to measure driving risks.

In summary, traffic safety analysis has received significant attention from researchers and the industry. A series of important results have been obtained through basic theoretical research and technical applications. In existing analysis methods, the data sources used primarily include historical collision and conflict data based on SSMs. Although methods based on collision data are reliable, the data acquisition process involved is difficult and time-consuming. By contrast, indirect evaluation methods based on SSMs can be used more widely for traffic safety analyses. Generally, these methods exhibit high accuracy, flexibility, and stability, among others. However, they depend significantly on field data to derive various conflict indicators. The research scope for these methods primarily includes macroanalysis based on traffic flow and microanalysis based on individual vehicles. The analysis methods based on traffic flow provide average results for the entire traffic scenario without considering the effect of individual vehicle behavior on safety. Analyses based on individual vehicles primarily consider front and rear vehicles and disregard the effects of interactions among surrounding vehicles on driving safety. Therefore, driving safety has been analyzed from the perspective of vehicle groups [42], focusing on the correlation between the main vehicle and its surrounding vehicles. This correlation considers the preceding and following cars in the same lane as well as other surrounding vehicles. Most existing safety analysis methods employ statistical regression or machine learning to analyze traffic or driving risks [43]. In the absence of collision data, data attributes must be manually annotated when using supervised learning methods—a process that is highly dependent on experience. However, unsupervised learning methods do not require advance data labeling.

2.4. Contributions and Framework. The contributions of this study are as follows:

(1) A vehicle group categorization rule is proposed to categorize vehicles based on temporal and spatial characteristics, through which continuous vehicles with mutual influence can be separated.

(2) An RMVG was proposed. The RMVG comprehensively reflects individual behaviors and group effects during the driving process. Using this model, the source of risk can be considered more comprehensively while the scope of risk investigation is reduced.

(3) A rear-end collision risk quantification method that considers the possibility and severity of collisions is established. Conflict probability and severity represent the risk levels in different dimensions. A comprehensive consideration of these two factors

can improve the effectiveness of risk measurements. The overall framework of the RMVG is presented in Figure 1.

The remainder of this paper is organized as follows: Section 3 introduces the data sources used in this study. Section 4 describes the vehicle group categorization method based on temporal and spatial similarities. Section 5 describes a risk evaluation index system and the proposed RMVG. Section 6 presents a validation of the proposed model using a real trajectory data set. Finally, Section 7 concludes the paper.

3. Data Preparation

The vehicle trajectory data used to analyze the rear-end collision risk of the vehicle groups were provided by the Southeast University Intelligent Traffic System (ITS) laboratory. These data were obtained from an elevated section of the expressway on Longpan South Road, Nanjing, China. The acquisition began at 7:30 a.m. on April 23, 2018 (Monday), and the weather was clear at that time. The study section was 427 m long and located in the East-West direction; the east and west sections were two-way eight-lane and six-lane roads, respectively, as shown in Figure 2. The data were continuously obtained for 4 min and 15 s using a DJI Mavic 2 drone at an altitude of 310 m with a frame rate of 24 frames per second, including 498,266 trajectory data points from 921 vehicles. The ITS researchers extracted the complete vehicle trajectory data from the video and manually verified them to ensure that the public data can realize complete vehicle identification and tracking. This data set provides trajectory information with a time accuracy of 0.1 s, including the speed, acceleration, lane, and driving distance of each vehicle. The data formats are listed in Table 1. The data set used in this study was smoothed via Kalman filtering to eliminate any possible noise.

4. Vehicle Group Categorization Rule Based on Temporal and Spatial Similarities

Most rear-end collisions are caused by vehicles trailing extremely closely; in such cases, accurate traffic information cannot be obtained timely. To measure the risk of rear-end collisions more conveniently, a categorization rule was proposed. Vehicles that trail closely and indicate group responses to random interference factors are categorized into the same vehicle group. Rear-end collisions and pile-ups can be expressed as a process in which the stable and close car-following state of the vehicle group is discontinued because of random interference factors. After the categorization, the group vehicles were slightly affected by the external vehicles. Collision risk primarily arises from other vehicles in the same group, as shown in Figure 3.

Close car-following implies that the following distance in vehicle groups should be less than a critical value. However, the car-following distance cannot completely characterize the mutual influence between vehicles. Short spatial and temporal distances between vehicles indicate that the vehicles demonstrate a significant level of mutual influence. Therefore, time and distance parameters must be considered to effectively classify the vehicles into different categories. In this study, the time headway and distance along the lane line (vehicle position) were considered to identify the time and space characteristics. The vehicle positions reflect the distance between all vehicles on the road. The hierarchical clustering method was employed to classify the vehicle groups, as shown in Figure 4. The classification was based on the specified threshold and the distance among clusters. The number of clusters need not be determined in advance. The smaller the distance, the more likely the vehicle groups will be classified into the same category. The distance and time headway between adjacent vehicles are smaller than those between nonadjacent vehicles. Therefore, vehicles in the same group are guaranteed to be adjacent to each other, as shown in Figure 3. The vehicle group composition should be dynamic because the driving state of a vehicle changes dynamically. Therefore, the vehicle groups were categorized in real time in this study, which may cause the vehicles to be classified into different groups during different time slices. The frame rate of the data set was 24 frames per second, which enables the real-time categorization of the vehicle group.

The temporal and spatial attributes of a vehicle were defined as X_i (x_i, y_i) in this study, where x_i and y_i are the time headway and the distance along the lane line, respectively. In the selected data set, the headway and distance distributions were relatively concentrated. The maximum-minimum method is a classic normalization method that is widely used in traffic safety research [44] and data processing prior to clustering [45]. Compared with other normalization methods, this method can distribute the data set selected in this study more evenly in the interval [0, 1] and maintain the relative linear relationships of their values [46]. Therefore, to eliminate the effects of dimensional differences and consider the data set structure, the maximum-minimum normalization method was used to normalize the data to the interval [0, 1]. Subsequently, the similarity between vehicles was measured using the Euclidean distance d (X_1, X_2). The proximity of *clusters a* and *b* can be measured by the average distance d (a, b) between them.

The similarity between vehicles can be measured as follows:

$$d(X_1, X_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{1}$$

The similarity between clusters can be obtained as follows:

$$d(a, b) = \frac{1}{n_a n_b} \sum_{p \in a} \sum_{p' \in b} |p - p'|, \tag{2}$$

where n_a and n_b are the numbers of samples in *clusters a* and *b*, respectively, and p and p' are the data in a and b, respectively.

Statistical methods typically used for determining the threshold include the 85% quantile and interquartile range methods [47]. In this study, the 25% quantile ($D_{25\%}$), median

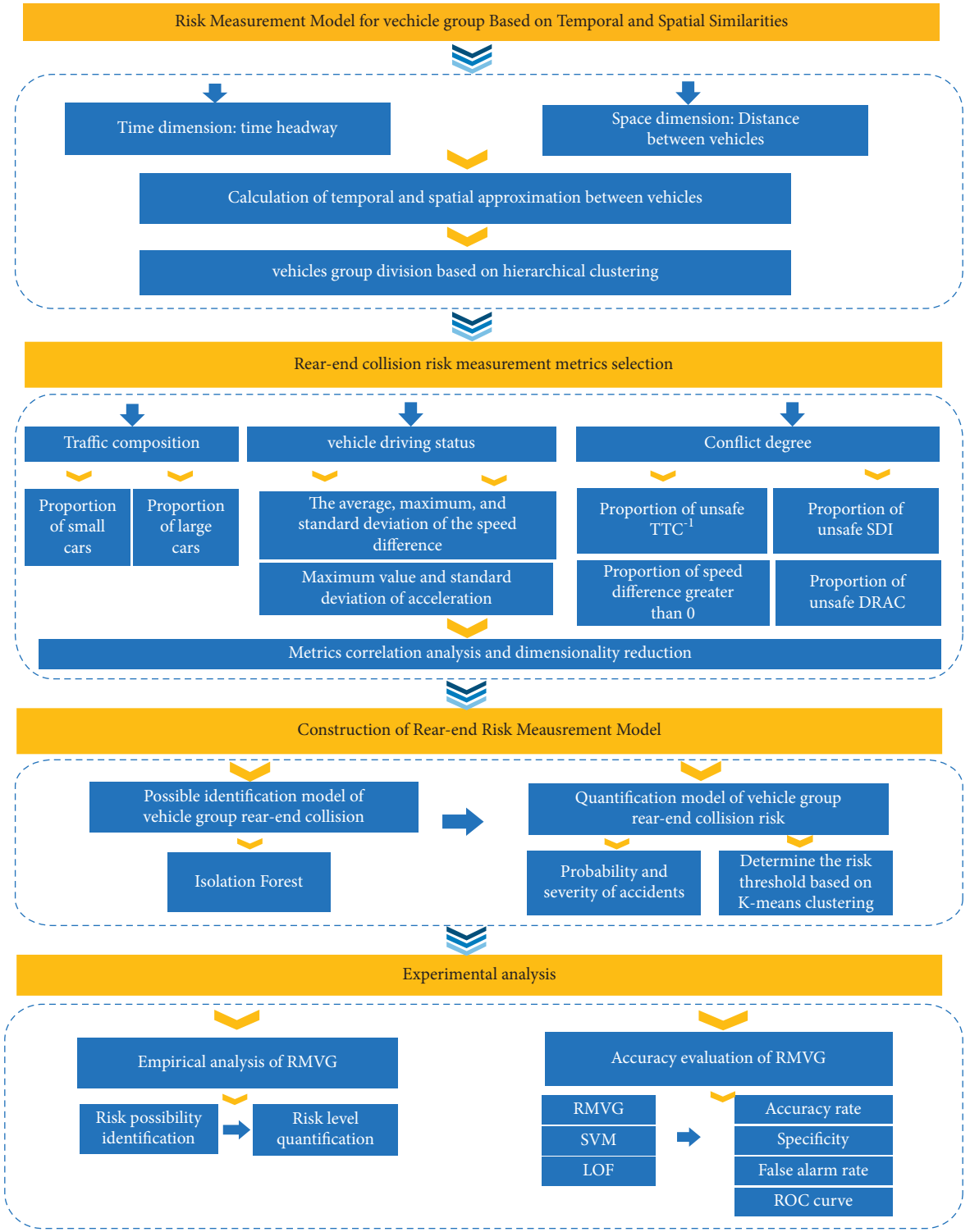

Figure 1: Framework of vehicle group risk measure model based on temporal and spatial similarities.

($D_{50\%}$), 75% quantile ($D_{75\%}$), and 85% quantile ($D_{85\%}$) of the temporal and spatial similarities were used as thresholds to categorize the vehicle group. The composition of the vehicle group changed dynamically over time. The results at a certain time are shown in Figure 5, where numbers 1–6 in the legend represent the different vehicle groups. All vehicles in the data set were categorized into groups. The results show that $D_{50\%}$ is the preferable threshold. When $D_{25\%}$ was used as the threshold, the categorization conditions were extremely strict, which rendered it difficult to categorize vehicles with high proximity into different groups. It is a challenge to limit the driving risk within a group. When $D_{75\%}$ or $D_{85\%}$ was used as the threshold, the classification conditions were extremely lenient; all vehicles can be easily categorized into the same

Figure 2: Top view of the study section captured by the drone.

Table 1: Example of trajectory data set.

Vehicle number	Lane number	Time (s)	Distance perpendicular to the lane (m)	Distance along the lane (m)	Speed (km/h)	Acceleration (m/s^2)	Vehicle length (m)	Vehicle width (m)
1	9	7.55	−1.30	108.94	54.04	0.44	12.05	4.05
110	6	41.45	−0.23	194.97	54.88	0.00	5.93	2.82
337	8	89.87	−0.59	2.58	51.14	−1.04	4.90	2.17
576	9	158.38	−0.97	1.28	42.67	0.31	4.80	2.26
708	6	233.19	0.53	203.29	21.21	0.52	4.80	2.17
879	7	248.79	1.41	29.84	25.11	0.27	5.56	2.82

vehicle group, which resulted in a significantly different degree of interaction between the internal vehicles. Hence, the appropriate categorization could not be achieved. However, vehicles with proximate time headways and positions are classified into the same cluster if $D_{50\%}$ is used as the threshold. The time headway and position between the clusters differed significantly. In other words, vehicles with high interactions can be classified into the same vehicle group more easily. The effect between the vehicle groups was insignificant, and the risk was associated with the vehicles within the group. Therefore, the median $D_{50\%}$ was used as the threshold in this study.

5. Establishment of Vehicle Group Rear-End Collision Risk Measurement Model

5.1. Selection of Rear-End Collision Risk Measurement Metrics. Rear-end collision risks are associated closely with various factors, such as traffic composition, driving status, drivers' risk perception, traffic conflict, and road conditions. Therefore, risk metrics $Q_{ij}(t)[q_{i1}(t), q_{i2}(t)...q_{i11}(t)]$ were extracted in this study while considering three aspects: traffic composition, vehicle driving status, and conflict degree, as listed in Table 2.

$$SDI(t) = SSD_{i-1}(t) - SSD_i(t) + d(t) - l_{i-1}, \tag{3}$$

where $SSD_{i-1}(t)$ and $SSD_i(t)$ denote the stopping sight distance of PV and FV at time t, respectively. $d(t)$ represents the distance between these two vehicles at time t. l_{i-1} is the length of the PV.

$$SSD(t) = \frac{v(t)^2}{254 \times (f \pm g)} + t_r \times v(t) \times 0.278, \tag{4}$$

where $v(t)$ is the vehicle speed at time t. f is the road friction coefficient. According to the friction coefficient standard of dry pavement, f is valued as 0.6. g is the road gradient, which is temporarily valued as 0. t_r is the driver's perception reaction time.

$$\mathrm{DRAC}(t) = \frac{\left[v_i(t) - v_{i-1}(t)\right]^2}{x_{i-1}(t) - x_i(t) - l_{i-1}}, \tag{5}$$

where $v_{i-1}(t)$ and $v_i(t)$ denote the speed of PV and FV at time t, respectively; $x_{i-1}(t)$ and $x_i(t)$ denote the positions of the PV and FV at time t, respectively; and l_{i-1} is the length of the PV.

5.2. Correlation Analysis of Metrics. Multiple variables were selected for the RMVG. However, the redundant features of the variables affected the accuracy of the results without providing any new information to the model. Therefore, the

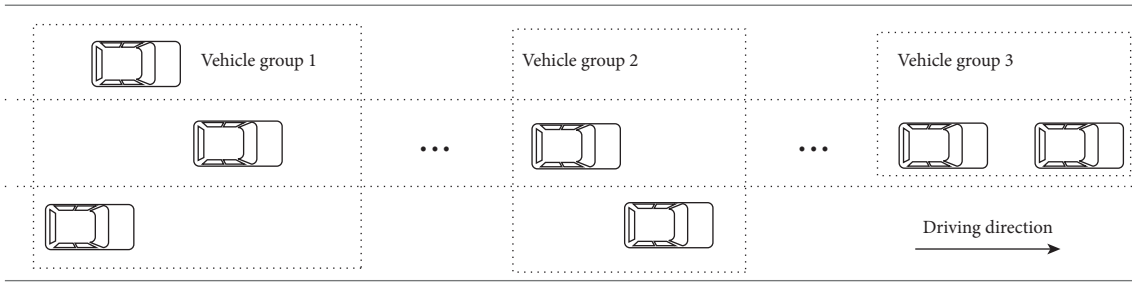

Figure 3: Schematic diagram of vehicle group division.

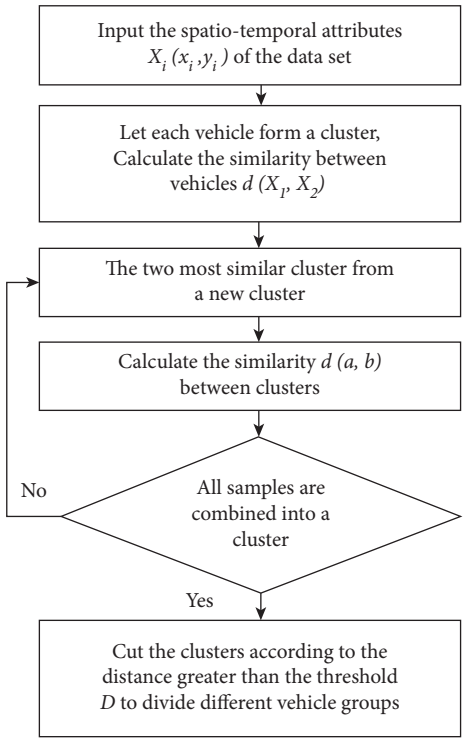

Figure 4: Vehicle group division steps.

correlation between $Qi(t)$ must be analyzed, and overlapping information must be removed. Reshef proposed the maximum information coefficient (MIC), which can be used to measure the linear and nonlinear relationships between variables in big data, as well as to determine their nonfunctional dependencies [50]. In this study, an analysis was performed to determine whether a correlation exists and the strength degree among variables. Although the correlation between the variables is dynamic, the average correlation between them can be determined by calculating the MIC when the sample size is sufficiently large.

The mutual information for random variables X and Y can be calculated as follows:

$$I(X,Y) = \int\int p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}dx\,dy, \tag{6}$$

where $p(x, y)$ is the joint probability distribution of X and Y and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

In calculating the MIC, the sample data are first placed in a two-dimensional space. Next, meshing is performed. Subsequently, random variables X and Y are selected from data set Q to form set D. Random variables X and Y are equally classified into x and y, respectively. The probability of each grid (x_i, y_j) is calculated as follows:

$$p(x_i, y_j) = \frac{n(x_i, y_j)}{n}, \tag{7}$$

where $n(x_i, y_i)$ is the number of data points in the (x_i, y_j) grid. n is the total number of data points. Similarly, $p(x_i)$ and $p(y_j)$ can be calculated.

The probability distribution under the current categorization method is denoted as $D|_{x*y}$. Mutual information $I(D|_{x*y})$ can be calculated using equation (6). First, the maximum mutual information value is $\max I(D|_{x*y})$ for all categorizations under the same segmentation scale. Next, let $I'[D(x, y)] = \max I(D|_{x*y})$, and standardize it.

$$M(D)_{x,y} = \frac{I'[D(x,y)]}{\lg(\min\{x,y\})}. \tag{8}$$

Subsequently, the MIC of random variables X and Y at different segmentation scales can be calculated as follows:

$$\mathrm{MIC}(X,Y) = \max\left\{M(D)_{x,y}\right\}. \tag{9}$$

The MIC between variables in $Q_i(t)$ is calculated as shown in Figure 6.

The calculated MIC indicates the existence of correlations between the variables. Principal component analysis was performed to extract effective information and simplify the calculation of the model. The variance of each component in $Q_i(t)$ is presented in Table 3. Currently, a popular method for determining the number of principal components is based on eigenvalues. Components with eigenvalues greater than 1 are identified as principal components. Another widely used approach is to select principal components based on the cumulative percent variance according to the amount of information to be retained [51]. In this study, a slight difference in the variables affected the effectiveness of risk identification. Therefore, the premise of selecting principal components should be to preserve effective information to the greatest extent. To retain most of the information while reducing the dimensions, the cumulative percentage variance of the principal components must be greater than 90% [52]. The variance of each component in $Q_i(t)$ is presented in Table 3.

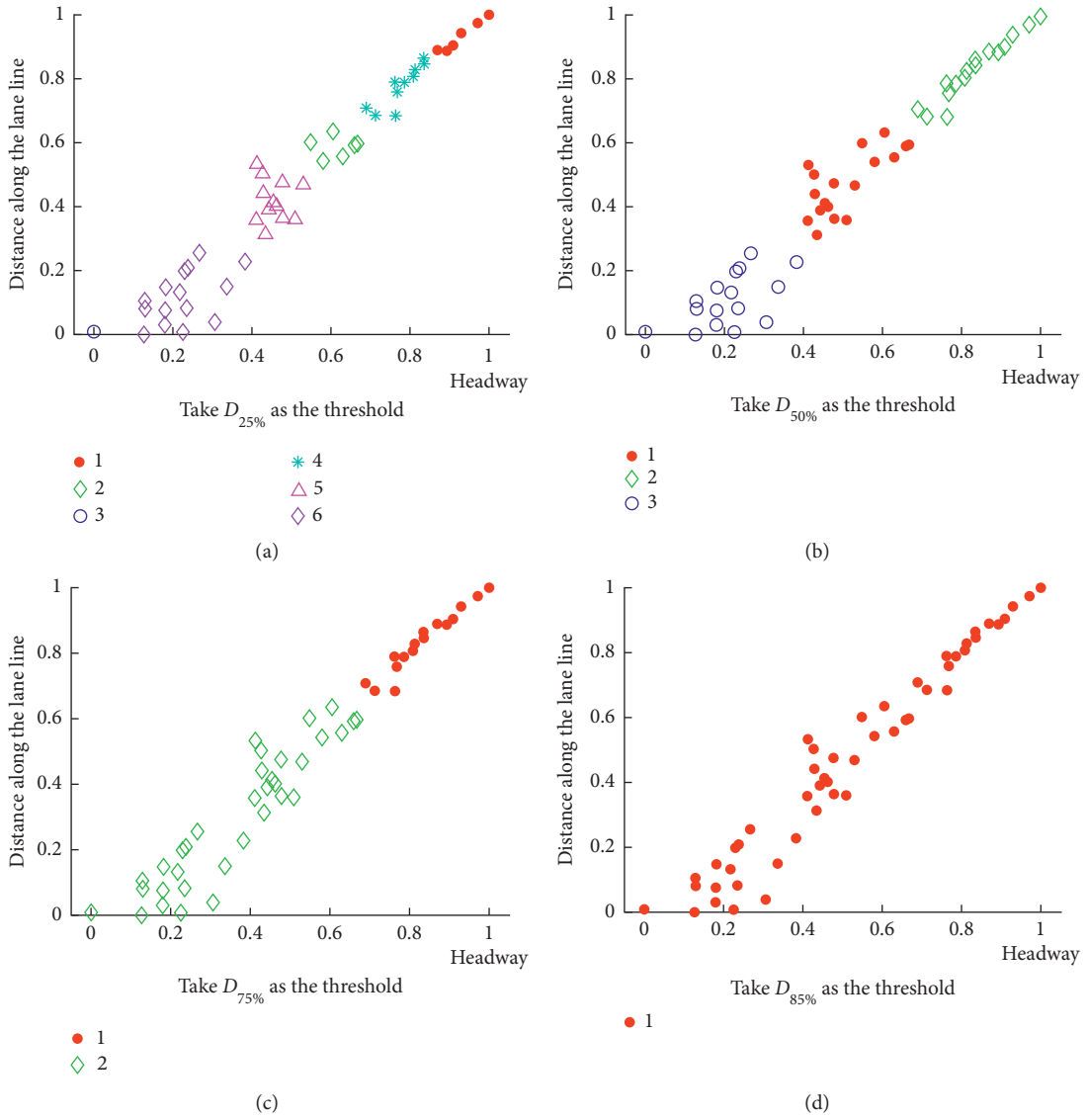

Figure 5: Division results for different thresholds.

Considering that the cumulative percent variance exceeded 90%, the first six components $U_{ik}(t)$ $[u_{i1}(t), u_{i2}(t),\ldots u_{i6}(t)]$ were selected in this study. The eigenvectors E_k $[e_1, e_2,\ldots e_6]$ of $U_{ik}(t)$ are listed in Table 4, and the calculation formulas are presented in equation (10).

$$\begin{aligned}
u_{i1}(t) &= 0.075 \times q_{i1}'(t) - 0.085 \times q_{i2}'(t) \ldots\ldots + 0.414 \times q_{i11}'(t),\\
u_{i2}(t) &= -0.218 \times q_{i1}'(t) + 0.221 \times q_{i2}'(t) \ldots\ldots - 0.023 \times q_{i11}'(t),\\
u_{i3}(t) &= 0.647 \times q_{i1}'(t) - 0.643 \times q_{i2}'(t) \ldots\ldots + 0.073 \times q_{i11}'(t),\\
u_{i4}(t) &= 0.148 \times q_{i1}'(t) - 0.157 \times q_{i2}'(t) \ldots\ldots - 0.269 \times q_{i11}'(t),\\
u_{i5}(t) &= 0.039 \times q_{i1}'(t) - 0.027 \times q_{i2}'(t) \ldots\ldots + 0.177 \times q_{i11}'(t),\\
u_{i6}(t) &= 0.061 \times q_{i1}'(t) - 0.038 \times q_{i2}'(t) \ldots\ldots - 0.378 \times q_{i11}'(t).
\end{aligned} \tag{10}$$

5.3. Development of Rear-End Risk Measurement Model. During the driving process, affected by several factors including road and traffic conditions, vehicles may exhibit abnormal driving behaviors, such as trailing extremely closely or decelerating rapidly. Generally, abnormal driving behavior causes a single-vehicle collision; however, this abnormal behavior may result in a pile-up if it significantly affects the surrounding vehicles. Collisions are more likely to occur when the driving state is abnormal. The rear-end collision risk measurement for a vehicle group is used to determine the possibility of a collision and to quantify its risk level. The RMVG model is realized via two procedures: first, the safe state of the vehicle group is determined based on trajectory data. Second, the risk level is quantified based on the possibility and severity of the collision.

5.3.1. Possible Identification Model for Vehicle Group Rear-End Collision Based on Isolation Forest (IF). IF is an unsupervised machine learning method that isolates outliers by continuously segmenting the data set [53]. This algorithm uses the isolated structure of a binary tree (*i*Tree). By randomly selecting sample features without replacement, the

Table 2: Vehicle group rear-end collision risk measurement metrics.

Variable	Category	Name	Description	Units
q_{i1} (t)	Traffic composition	The proportion of small cars	The proportion of small cars in the vehicle group.	—
q_{i2} (t)		The proportion of large cars	The proportion of large vehicles in the vehicle group.	—
q_{i3} (t)	Vehicle driving status	Average speed difference	The difference between speeds of the following vehicle and preceding vehicle.	km/h
q_{i4} (t)		The standard deviation of speed difference		—
q_{i5} (t)		Maximum speed difference		km/h
q_{i6} (t)		The standard deviation of acceleration	Acceleration is the absolute value of each car's acceleration in the vehicle group.	—
q_{i7} (t)		Maximum acceleration		m/s^2
q_{i8} (t)	Conflict degree	The proportion of unsafe TTC^{-1}	The proportion of vehicles in the vehicle group whose TTC^{-1} is greater than the safety threshold (0.25/s) [48]. The calculation method of TTC is shown in equation (15).	—
q_{i9} (t)		The proportion of unsafe stopping distance index (SDI)	The proportion of vehicles in the vehicle group whose SDI is less than the safety threshold (0) [49]. The calculation method of SDI is shown in equation (3).	—
q_{i10} (t)		The proportion of unsafe deceleration rate to avoid a crash (DRAC)	The proportion of vehicles in the vehicle group whose DRAC is greater than the maximum available deceleration rate [26]. The calculation method of DRAC is shown in equation (5).	—
q_{i11} (t)		The proportion of the following vehicle (FV) speed greater than the preceding vehicle (PV) speed	The proportion of vehicles in the vehicle group whose speed is greater than that of the preceding vehicle.	—

data set is segmented continuously until each sample is isolated. Because the outliers present are few, distinct, and sparsely distributed, the path is extremely short during isolation. Therefore, abnormal points are isolated closer to the root of the tree, whereas normal points are isolated from deeper regions of the tree. Compared with other methods, the IF can provide an abnormal probability to each sample, thus reflecting the possibility of a collision [54].

The dimensionality reduction metrics $U_i(t)$ in Section 5.2 are the input variables of identification models based on the IF. The output results indicate the possibility and assessment of collisions within the vehicle group. The realization process is shown in Figure 7.

(1) *Training Phase*. The model was trained to build isolated trees (*iTrees*) and an isolated forest (*i*Forest).

Step 1: Randomly select φ subsamples from the data set without replacement.

Step 2: Randomly select the characteristic attribute q as the starting node. Subsequently, select a split value p between the maximum and minimum values of q.

Step 3: Assign subsamples with attribute values less than p to the left branch of the binary tree; otherwise, assign them to the right branch.

Step 4: Repeat steps 2 and 3 until the segmentation is completed or the desired tree depth is reached. The depth limit is calculated using

$$l = \text{ceiling}\left(\log_2^{\varphi}\right). \tag{11}$$

Step 5: Repeat steps 1–4 until the number of *i*Trees reaches the limit. These *i*Trees are joined to form an *i*Forest = [$i\text{Tree}_1$, $i\text{Tree}_2$. . .$i\text{Tree}_n$].

(2) Testing Phase. After the construction is completed, the iForest can be used to identify data abnormalities based on the abnormal scores.

Step 6: Allow the test sample to traverse *i*Trees in *i*Forest and compute the average path length when the traversal halts.

Step 7: Calculate the abnormal probability of the sample and determine whether it is abnormal.

The formula to calculate the abnormal score is

$$S(x,n) = 2^{-(E(h(x))/c(n))}, \tag{12}$$

where $E\ (h\ (x))$ is the average path length required to separate sample x in *i*Forest and $c\ (n)$ is the average tree length, which is calculated as follows:

$$C(n) = 2H(n-1) - \frac{2(n-1)}{n},$$
$$H(i) = \ln(i) + 0.5772156649. \tag{13}$$

When $E\ (h(x))$ approaches (n), S tends to 0.5, and the sample is regarded as normal; in this case, the model outputs a judgment result of "1." When $E(h\ (x))$ approaches 0, S tends to 1, and the sample is regarded as abnormal; in this case, the model outputs a judgment result of "−1."

5.3.2. Quantification Model for Vehicle Group Rear-End Collision Risk. In the typically used quantification method, the risk value is calculated based on the probability and severity of a collision. The abnormality degree reflects the possibility of collision in a vehicle group. Meanwhile, the risk severity depends on the coupling between various influencing factors. Therefore, the abnormal score, as calculated in Section 5.3.1, was adopted in this study to reflect the possibility of a collision, and risk metrics $Q_i\ (t)$ were used to reflect the severity.

For vehicle group i at time t, the risk can be quantified using

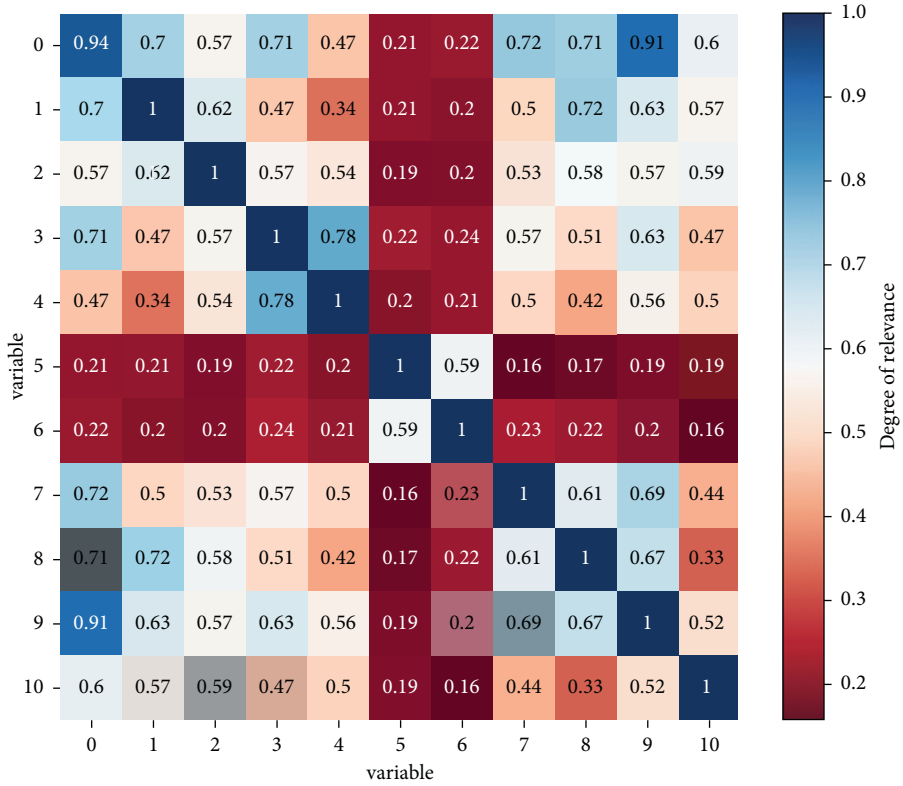

Figure 6: MIC of the variables.

Table 3: Variance of principal components.

Component	Percent of variance (%)	Cumulative percent variance (%)
1	29.00	29.00
2	21.21	50.21
3	17.94	68.15
4	10.38	78.53
5	8.59	87.11
6	5.58	92.69
7	4.03	96.72
8	1.51	98.23
9	0.86	99.09
10	0.62	99.70
11	0.30	100.00

$$H(t)_i = \sum_{j=2}^{11} \left(s(t)_i \times q(t)_{ij}' \right), \tag{14}$$

where $H(t)_i$ is the risk value, $s(t)_i$ is the abnormal score, and $q(t)_{ij}'$ is the driving risk metric after normalization.

After calculating the risk values, K-means clustering was adopted to separate the risk levels and thresholds [55].

(1) Calculate the silhouette coefficient for different cluster numbers to select the most appropriate risk level n.

(2) Determine the cluster centers $[c_1, c_2, \ldots, c_n]$. Consider e between (c_i, c_{i+1}) $(i \in n)$ with a step size of 0.01 to classify the sample and calculate the accuracy rates. Select e with the highest accuracy rate as the categorization threshold.

6. Experimental Analysis

6.1. Data Processing. First, a Kalman filter was adopted to denoise the original data set. Subsequently, based on the categorization rules in Section 4, all vehicles on the road segment were categorized into different vehicle groups. Finally, the rear-end collision risk measurement metrics were calculated based on the data set, and the results are listed in Table 5.

6.2. Empirical Analysis of RMVG. In this study, 537 vehicle groups were selected from the data set, including 12,075 trajectory data points. The empirical analysis comprised three stages: (1) identifying the possibility of rear-end collisions in vehicle groups based on the RMVG, (2) quantifying the degree of rear-end collision risk based on the RMVG and classifying the risk level, (3) analyzing the feasibility of the model via accuracy evaluation.

6.2.1. Vehicle Group Rear-End Collision Possibility Identification. A vehicle group rear-end collision identification model was established based on the IF algorithm through Python. To construct *i*Forest, 70% of the data were randomly selected as the training set, and the remaining 30% were used as the test set. This model can automatically identify the collision probability and output either "−1" or

Table 4: The eigenvectors of U_i (t).

Variable	e_1	e_2	e_3	e_4	e_5	e_6
q_{i1} (t)	0.075	−0.218	0.647	0.148	0.039	0.061
q_{i2} (t)	−0.085	0.221	−0.643	−0.157	−0.027	−0.038
q_{i3} (t)	0.227	−0.029	−0.038	−0.332	0.819	0.103
q_{i4} (t)	0.465	0.066	0.010	−0.267	−0.380	0.180
q_{i5} (t)	0.461	0.023	0.018	−0.301	−0.316	0.281
q_{i6} (t)	0.402	0.188	−0.080	0.507	0.134	−0.008
q_{i7} (t)	0.360	0.183	−0.159	0.574	0.085	0.077
q_{i8} (t)	−0.005	0.581	0.241	−0.126	−0.016	−0.232
q_{i9} (t)	−0.191	0.412	0.139	−0.050	0.154	0.748
q_{i10} (t)	−0.089	0.567	0.228	−0.055	−0.037	−0.332
q_{i11} (t)	0.414	−0.023	0.073	−0.269	0.177	−0.378

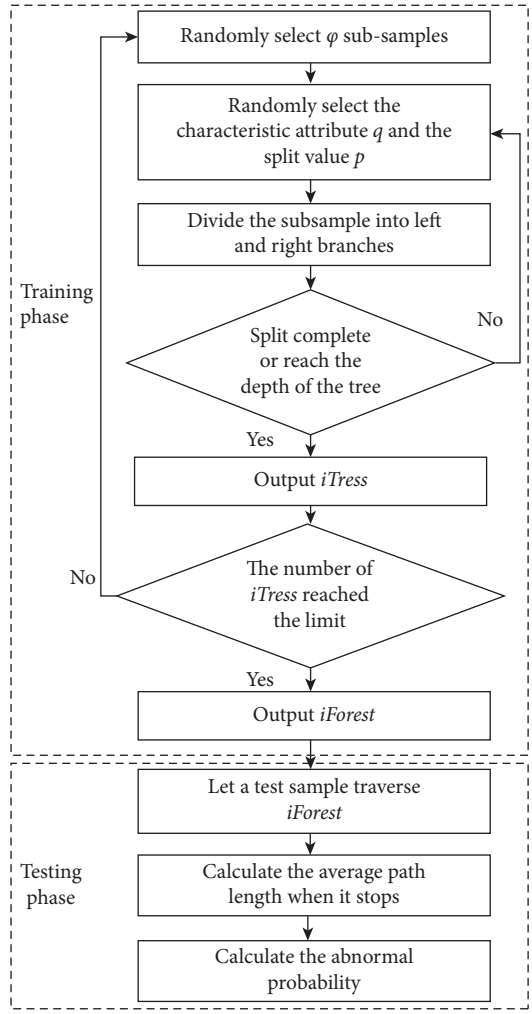

Figure 7: Algorithm flowchart.

"1." The results are presented in Table 6, where "1" and "−1" represent vehicle groups with lower and higher probabilities of collision, respectively. The higher the abnormal score, the greater the possibility of an anomaly.

In the aforementioned analysis, time and space risks were indicated during vehicle group driving; a small distance between cars and a short TTC may cause collisions. Therefore, the TTC and margin to collision (MTC) can be used as time and space risk evaluation indicators, respectively, to identify whether the vehicle is susceptible to a rear-end collision [56].

The TTC is the predicted time of a collision between PV and FV when the two vehicles maintain the current relative velocity. Collisions are more likely to occur when the TTC is

Table 5: Data processing results.

Vehicle group ID	Number of vehicles	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}
1703	19	0.95	0.05	0.10	5.87	9.54	0.63	2.28	0.32	0.42	0.21	0.53
830	7	0.86	0.14	0.49	7.71	9.81	0.71	1.94	0.43	0.43	0.43	0.57
3500	22	0.95	0.05	−0.15	4.51	6.86	0.52	1.77	0.05	0.14	0.05	0.55
3808	21	0.95	0.05	0.28	5.14	8.63	0.80	2.62	0.19	0.33	0.14	0.48
4545	15	0.93	0.07	0.22	3.97	6.02	0.63	2.26	0.00	0.07	0.00	0.40
8956	24	1.00	0.00	−0.51	4.34	6.07	0.44	1.54	0.21	0.29	0.13	0.46
9379	19	0.95	0.05	−1.01	2.67	3.24	0.51	1.70	0.16	0.16	0.16	0.26
9712	22	0.91	0.09	−0.77	3.43	6.87	0.58	1.88	0.14	0.18	0.14	0.27
10344	18	0.94	0.06	−0.25	2.99	5.91	0.55	1.95	0.17	0.17	0.17	0.44
11109	26	1.00	0.00	−0.18	4.18	7.52	1.01	4.82	0.19	0.23	0.15	0.54

between 0 and 5 s [48]. Furthermore, researchers have shown that when the TTC is less than 5 s, drivers tend to feel nervous and perform more incorrect actions [57].

$$TTC(t) = \frac{x_{i-1}(t) - x_i(t) - l_{i-1}}{v_i(t) - v_{i-1}(t)}, \quad (15)$$

where $x_{i-1}(t)$ and $x_i(t)$ denote the positions of the PV and FV at time t, respectively; $v_{i-1}(t)$ and $v_i(t)$ denote the speed of the PV and FV at time t, respectively; and l_{i-1} is the length of the PV.

MTC indicates the final relative position of the PV and FV when the two vehicles decelerate abruptly. An MTC of less than 1 indicates that the stopping distance of the FV is greater than the summation of the intervehicular distance and stopping distance of the PV. In this case, a collision may occur between the vehicles. The lower the MTC, the higher the probability of collision.

$$\mathrm{MTC}(t) = \left(D(t) + \frac{v(t)_{i-1}^2}{2 \times a}\right) \times \left(v(t)_i \times t_0 + \frac{v(t)_i^2}{2 \times a}\right)^{-1}, \quad (16)$$

where $D(t)$ is the distance between vehicles at time t; a is the braking deceleration, which was set as 6.86 m/s^2 in this study [58]; $v_{i-1}(t)$ and $v_i(t)$ denote the speed of the PV and FV at time t, respectively; and t_0 is the driver's reaction time, which was set as 1.5 s in this study [58].

The vehicle group represented a whole. When the proportion of serious conflicts is high, the internal driving situation is chaotic. In this situation, conflicts significantly affect the internal vehicles and are more likely to cause crashes. Therefore, in this study, the collision probability of the vehicle group was measured based on the proportion of severe conflicts. The proportions of TTC less than 5 s and MTC less than 1 in the vehicle group were calculated, and the results are listed in Table 6. Vehicle groups with a high proportion of abnormal TTC or MTC were identified as anomalous and indicated high abnormal scores. Additionally, it can be seen that the RMVG model has effective risk identification capabilities.

6.2.2. Accuracy of Rear-End Collision Possibility Identification Method. It is considered that the vehicle group has a comparatively higher possibility of collision when its TTC proportion for less than 5 s or its MTC proportion for less than 1 exceeds 25%. Among the 537 selected vehicle groups, 51 indicated a high possibility of collision, including 183 trajectory data points. A total of 486 vehicle groups exhibited a low collision probability, including 11,892 trajectory data points. Vehicle group rear-end collision identification models were established based on the RMVG, SVM, and LOF by selecting 70% of the data as the training set and 30% as the test set. The model evaluation indicators were calculated as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \\ \text{Specificity} &= \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}, \\ \text{False Alarm Rate} &= \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}, \end{aligned} \quad (17)$$

where TN refers to observations correctly identified as unsafe, TP is the correct prediction of safe conditions, FN is the incorrect labeling of safe samples as unsafe, and FP is the incorrect prediction of unsafe samples as safe.

The accuracy levels determined by calculating the confusion matrix of the prediction results are presented in Table 7. The RMVG exhibited the highest accuracy rate, specificity, and the lowest false alarm rate. The receiver operating characteristic (ROC) curves of the three algorithms based on their sensitivity and specificity are shown in Figure 8. Comparative analysis shows that the area under the curve (AUC) of the RMVG was 0.93, which was higher than that of the SVM (AUC = 0.74) and LOF (AUC = 0.90). Additionally, the RMVG demonstrated better recognition ability than the other models under the same data conditions.

6.2.3. Vehicle Group Rear-End Collision Risk Quantification. The method discussed in Section 6.2.1 can only be used to identify the possibility of rear-end collisions in the vehicle group; however, the risk degree remains ambiguous. Therefore, the rear-end collision risk was quantified based on the quantitative model proposed in Section 5.3.2, and the results are presented in Table 6.

The silhouette coefficient results obtained using the risk classification method presented in Section 5.3.2 are shown in Figure 9. When the cluster numbers were 2, 3, 4, and 5, the contour coefficients were 0.597, 0.558, 0.576, and 0.581,

Table 6: RMVG model results.

Vehicle group ID	Number of vehicles	Proportions of TTC < 5 s	Proportions of MTC < 1	RMVG results	Abnormal score	Risk quantification value
8907	3	0.33	0	−1	0.651	2.772
8888	3	0.33	0	−1	0.598	2.182
8860	4	0.25	0	−1	0.613	2.205
22505	4	0.25	0	−1	0.582	1.798
16250	3	0	0.33	−1	0.524	1.213
8939	3	0.33	0	−1	0.621	2.284
22543	4	0.25	0	−1	0.582	1.742
22492	4	0.25	0	−1	0.598	1.778
8844	4	0.25	0	−1	0.634	2.704
8864	4	0.25	0	−1	0.616	2.105
16728	5	0	0	1	0.517	0.822
9706	8	0	0	1	0.451	1.308
16746	5	0	0	1	0.495	0.856
16725	5	0	0	1	0.485	0.824
6895	15	0	0	1	0.434	1.292
9700	22	0	0.05	1	0.497	1.46
6856	31	0	0	1	0.419	1.437
16770	5	0	0	1	0.492	1.004
16737	5	0	0	1	0.500	1.018
16781	45	0	0	1	0.434	1.556

Table 7: Performance comparison of three machine learning algorithms.

Algorithm	AUC	Accuracy rate (%)	Specificity (%)	False alarm rate (%)
RMVG	0.93	95.68	88.89	3.47
SVM	0.74	89	55.56	6.94
LOF	0.90	95.06	83.33	3.47

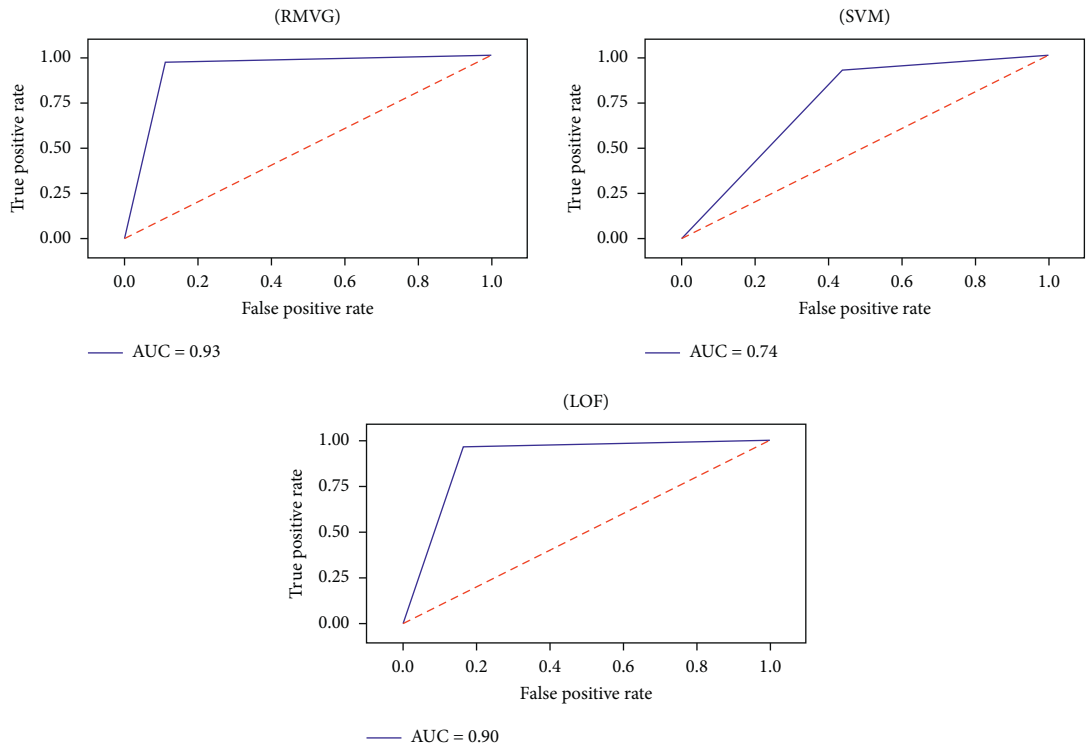

Figure 8: ROC and AUC curves.

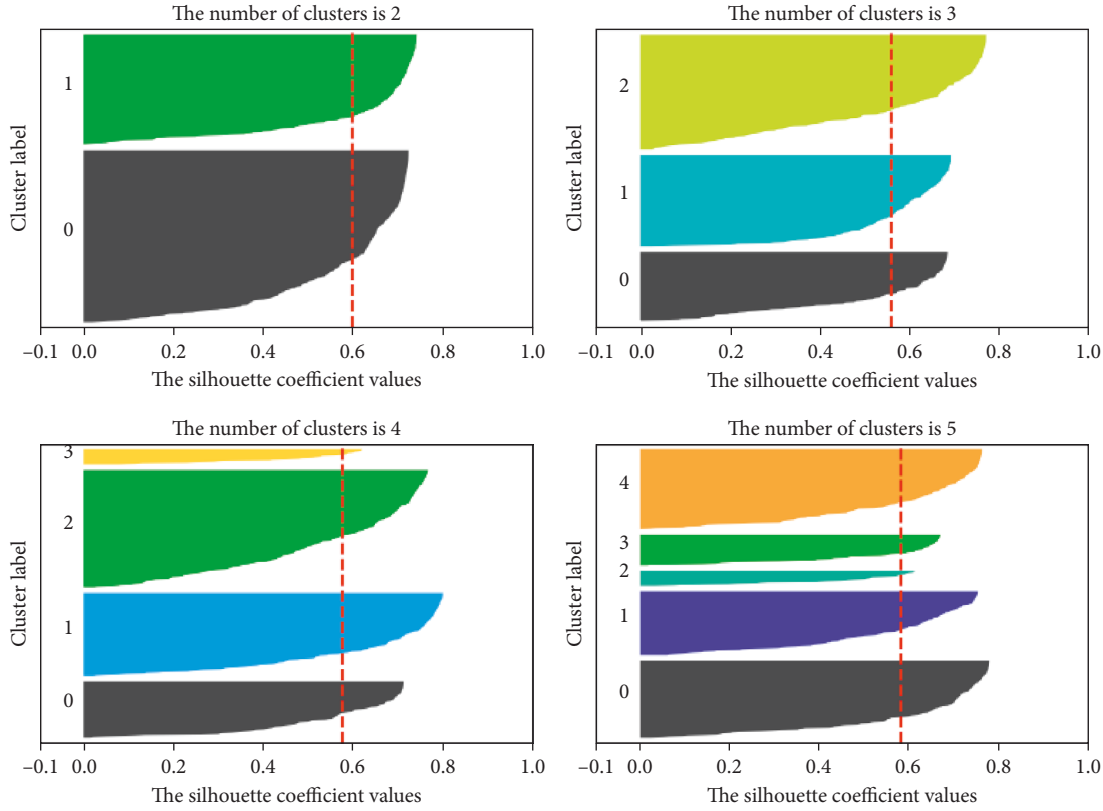

Figure 9: Silhouette coefficient under different cluster numbers.

Table 8: Risk levels.

Risk level	Risk value interval
Security	[0, 1.12)
Low risk	[1.12, 1.51)
Moderate risk	[1.51, 1.87)
Higher risk	[1.87, 2.32)
Danger	[2.32, ∞]

respectively. However, if the number of risk levels is low, then the difference between risk values is difficult to define. The rear-end collision risk of the vehicle group was categorized into five levels; the higher the risk quantification value, the greater the risk. The thresholds were $e_1 = 1.12$, $e_2 = 1.51$, $e_3 = 1.87$, and $e_4 = 2.32$, and the accuracy rates were 99.70%, 99.72%, 100%, and 100%, respectively. The classification results are listed in Table 8.

As shown in Tables 6 and 8, the risk level is directly proportional to the risk value. However, the risk level is a combination of the collision probability and severity. Therefore, when the RMVG recognizes that the collision probability of a vehicle group is high, it may indicate different risk levels.

For example, the RMVG recognizes that vehicle group 16250 has a high collision probability; however, its risk value is 1.213. Its internal characteristics are as follows: the number of internal vehicles, 3; maximum speed difference, 1.61 km/h; maximum acceleration, 0.72 m/s^2; and the proportion of unsafe SDI, 0.67. This shows that high traffic conflicts occurred within the vehicle group, that is, the probability of collision is high. However, because the dispersion of speed and acceleration is low and the number of internal vehicles is small, the collision severity is low. By contrast, the RMVG recognizes that the collision probability of vehicle group 16781 is low; however, its risk level is 1.556. The characteristics of this group are as follows: the number of vehicles, 45; maximum speed difference, 24.289 km/h; and maximum acceleration, 1.4972 m/s^2. The proportions of unsafe TTC^{-1}, SDI, and DRAC are 0.09, 0.11, and 0.02, respectively. This indicates that the degree of traffic conflict is relatively low. However, owing to the large dispersion of acceleration and speed, the driving stability of the vehicle group can degrade easily. Once a collision occurs, the severity is comparatively great, and it may appear as a multivehicle rear-end collision.

For the road section with a length of 427 m in this study, all the vehicles were categorized into multiple groups. Owing to the different driving characteristics within each group, the risks at various spatial locations along the road segment were different. Moreover, the composition of a vehicle group changes dynamically with the vehicle position and driving status. Therefore, the risk is dynamic and varies over time and space, as shown in Figure 10. Based on the real-time categorization of vehicle groups, dynamic risk detection was realized in this study for different spatial positions in long road sections.

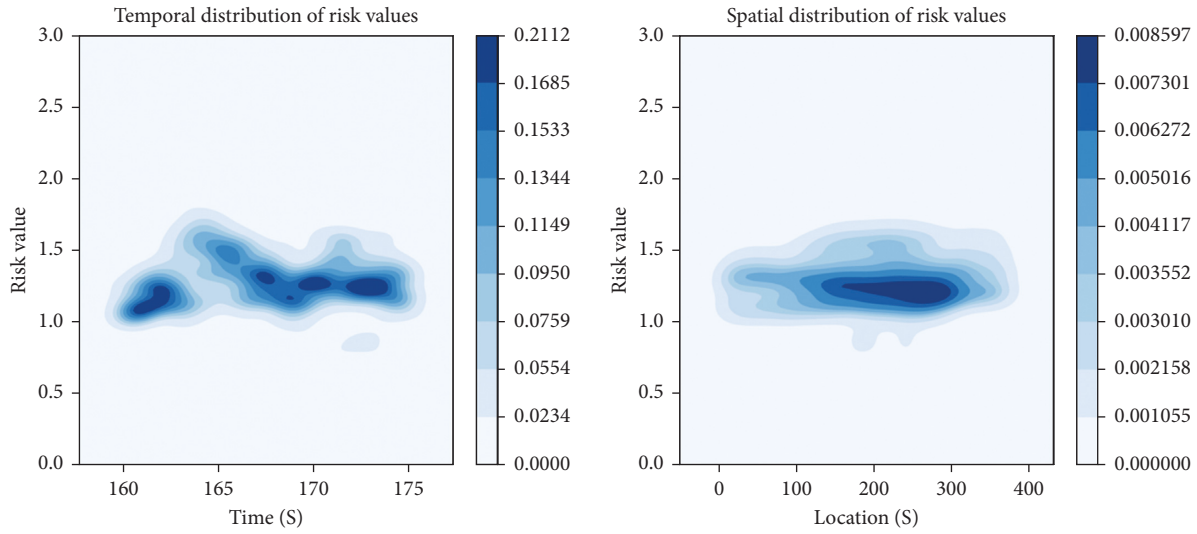

Figure 10: Temporal and spatial variation of risk values based on kernel density estimation.

7. Conclusions

A risk measurement model for vehicle groups was proposed herein based on temporal and spatial similarities. In contrast to conventional macrorisk identification, which focuses on traffic flow, or microrisk identification, which focuses on individual vehicles, the research object of this study was a vehicle group. It can narrow the recognition range as well as comprehensively consider the effects of individual behaviors and the interaction among surrounding vehicles on rear-end collisions.

First, vehicles that trailed closely and showed group responses to random interference factors were categorized into the same vehicle group. Rear-end collisions can be expressed as a process in which the stable and close car-following state in the vehicle group is discontinued. After the categorization, the effect of external vehicles on the vehicles inside a group becomes less significant. The risk primarily arises from internal vehicles within groups. Considering vehicle groups as research objects can provide a new perspective for investigating traffic safety problems. Subsequently, based on the IF, an RMVG was established, which considers the probability and severity of a collision to identify and quantify the risk of rear-end collisions. Additionally, the k-means clustering algorithm was used to separate the risk level and threshold. Finally, the RMVG was tested using a vehicle trajectory data set published by the ITS Laboratory at Southeast University. The results showed that the AUC, accuracy, specificity, and false alarm rate of the RMVG were 0.93, 95.68%, 88.89%, and 3.47%, respectively. This study provides a theoretical basis and technical support for the effective prevention of rear-end collisions, thereby reducing traffic crashes and economic losses. As connected vehicles and holographic road technologies are further developed, the results of this study can provide useful suggestions for drivers and road traffic management authorities. Vehicle trajectory data can be obtained from holographic roads, whereas vehicle communication can be accomplished via a connected vehicle environment. Combining these technologies with the RMVG allows rear-end conflicts to be monitored in real time. For road traffic management, the RMVG considers both the individual behavior and group responses of vehicles. It is more effective than a single strategy for identifying driving risks. For drivers, receiving traffic management information in real time allows them actively avoid risks, thereby improving driving safety.

The limitations of the proposed method are as follows: (1) it is temporarily impossible to validate the risk clusters with the ground truth owing to insufficient data. (2) It may not be suitable for a traffic congestion state because the categorization method will be invalid. When the traffic flow is smooth, this method can flexibly identify vehicles with a high degree of mutual influence, thereby effectively categorizing them into groups. However, when the traffic is congested, all vehicles are categorized into the same group. In this case, it will be meaningless to categorize the vehicle groups. Consequently, a possible research direction would be to improve the methods based on these scenarios. (3) The penetration rate of large vehicles, vehicle driving status, and conflict degree were used in this study to quantify the risk of rear-end collisions; however, certain limitations were indicated. Therefore, future research can also attempt to quantify the severity of crashes based on kinetic energy loss. The change in kinetic energy can be calculated from the velocities and collision angles based on more comprehensive data. (4) Driving safety is affected by weather and road conditions, in addition to driving conditions and surrounding vehicles. In future studies, weather conditions, road linearity, and other factors should be integrated to form a more comprehensive risk measurement metric system to improve accuracy. Moreover, the extreme value theory (EVT) framework is widely used for crash prediction and has been demonstrated to be effective. Hence, EVT could be combined with machine learning in the future to improve the accuracy of collision prediction.

Data Availability

The data set used to support the findings of this study was obtained from the ITS Laboratory of Southeast University. The data set is publicly available and can be downloaded and accessed from http://seutraffic.com/#/download.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, grant no. 52172345.

References

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

[2] National Highway Traffic Safety Administration, *Traffic Safety Facts A Compilation of Motor Vehicle Crash Data*, National Highway Traffic Safety Administration, Washington DC, 2019.

[3] X. Wang, M. Zhu, M. Chen, and P. Tremont, "Drivers' rear end collision avoidance behaviors under different levels of situational urgency," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 419–433, 2016.

[4] H. Zhang, S. Li, C. Wu, Q. Zhang, and Y. Wang, "Predicting crash frequency for urban expressway considering collision types using real-time traffic data," *Journal of Advanced Transportation*, vol. 2020, pp. 1–8, 2020.

[5] L. Yu, B. Du, X. Hu, L. Sun, L. Han, and W. Lv, "Deep spatio-temporal graph convolutional network for traffic accident prediction," *Neurocomputing*, vol. 423, pp. 135–147, Jan. 2021.

[6] Y. Wu, M. Abdel-Aty, Q. Cai, J. Lee, and J. Park, "Developing an algorithm to assess the rear-end collision risk under fog conditions using real-time data," *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 11–25, 2018.

[7] W. Qi, Z. Wang, R. Tang, and L. Wang, "Driving risk detection model of deceleration zone in expressway based on generalized regression neural network," *Journal of Advanced Transportation*, vol. 2018, pp. 1–8, 2018.

[8] L. Zheng, T. Sayed, and F. Mannering, "Modeling traffic conflicts for use in road safety analysis: a review of analytic methods and future directions," *Analytic methods in accident research*, vol. 29, p. 100142, 2021.

[9] N. Mascret, M. Nicolleau, and C. Martha, "The predictive role of achievement goals adoption on sensation-seeking and risk taking in driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 79, pp. 1–10, 2021.

[10] J. Sun and J. Sun, "A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 176–186, May 2015.

[11] Q. Ali, M. R. Yaseen, and M. T. I. Khan, "The causality of road traffic fatalities with its determinants in upper middle income countries: a continent-wide comparison," *Transportation Research Part A: Policy and Practice*, vol. 119, pp. 301–312, 2019.

[12] C. Dong, C. Shao, J. Li, and Z. Xiong, "An improved deep learning model for traffic crash prediction," *Journal of Advanced Transportation*, vol. 2018, pp. 1–13, 2018.

[13] Y. Yang, H. Chung, and J. S. Kim, "Local or neighborhood? Examining the relationship between traffic accidents and land use using a gradient boosting machine learning method: the case of suzhou industrial park, China," *Journal of Advanced Transportation*, vol. 2021, pp. 1–30, Jan. 2021.

[14] R. J. Smeed, "Variations in the patterns of accident rates in different countries and their causes," *Traffic Engineering and Control*, vol. 10, pp. 364–371, 1968.

[15] K. s. Ng, W. t. Hung, and W. g. Wong, "An algorithm for assessing the risk of traffic accident," *Journal of Safety Research*, vol. 33, no. 3, pp. 387–410, Oct. 2002.

[16] M. B. A. Rabbani, M. A. Musarat, W. S. Alaloul et al., "A comparison between seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ES) based on time series model for forecasting road accidents," *Arabian Journal for Science and Engineering*, vol. 46, no. 11, pp. 11113–11138, 2021.

[17] B. Dong, X. Ma, F. Chen, and S. Chen, "Investigating the differences of single-vehicle and multivehicle accident probability using mixed logit model," *Journal of Advanced Transportation*, vol. 2018, pp. 1–9, 2018.

[18] J. C. Hayward, "Near miss determination through use of a scale of danger," *Highway Research Record*, vol. 384, pp. 24–34, 1972.

[19] C. Hydén, *The Development of a Method for Traffic Safety Evaluation: The Swedish Traffic Conflict Technique [doctoral Thesis]*, Lund University, Department of Traffic Planning and Engineering, 1987.

[20] V. E. Balas and M. M. Balas, "Driver assisting by inverse time to collision," in *Proceedings of the 2006 World Automation Congress*pp. 1–6, Budapest, Hungary, July 2006.

[21] M. M. Balas, V. E. Balas, and J. Duplaix, "Optimizing the distance-gap between cars by constant time to collision planning," in *Proceedings of the IEEE International Symposium on Industrial Electronics*, pp. 304–309, ISIE '07, June 2007.

[22] A. Tarko, G. A. Davis, N. Saunier, and T. Sayed, *Surrogate Measures of Safety*, Emerald, Bingley, U.K, 2018.

[23] A. Tarko, *Measuring Road Safety with Surrogate Events*, Elsevier, Amsterdam, The Netherlands, 2019.

[24] M. Hu, Y. Liao, W. Wang, G. Li, B. Cheng, and F. Chen, "Decision tree-based maneuver prediction for driver rear-end risk-avoidance behaviors in cut-in scenarios," *Journal of Advanced Transportation*, vol. 2017, pp. 1–12, 2017.

[25] L. Zheng and T. Sayed, "A bivariate Bayesian hierarchical extreme value model for traffic conflict-based crash estimation," *Analytic methods in accident research*, vol. 25, no. 4, p. 100111, 2020.

[26] T. Chen, X. Shi, and Y. D. Wong, "Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data," *Accident Analysis & Prevention*, vol. 129, pp. 156–169, 2019.

[27] M. Yang, X. Wang, and M. Quddus, "Examining lane change gap acceptance, duration and impact using naturalistic driving data," *Transportation Research Part C: Emerging Technologies*, vol. 104, pp. 317–331, 2019.

[28] D. D. Hema and K. A. Kumar, "Levenberg–marquardt –lstm based efficient rear-end crash risk prediction system optimization," *International Journal of Intelligent Transportation Systems Research*, vol. 20, no. 1, pp. 132–141, Apr. 2022.

[29] Z. Fan, C. Liu, D. Cai, and S. Yue, "Research on black spot identification of safety in urban traffic accidents based on machine learning method," *Safety Science*, vol. 118, pp. 607–616, 2019.

[30] Á. Török, K. Varga, and J.-M. Pergandi, "Towards a cognitive warning system for safer hybrid traffic," *Intelligent Decision Technologies*, vol. 11, no. 4, pp. 431–439, 2017.

[31] Y. Djenouri, A. Zimek, and M. Chiarandini, "Outlier detection in urban traffic flow distributions," in *Proceedings of the 2018 IEEE International Conference on Data Mining*, pp. 935–940, ICDM, Singapore, November 2018.

[32] Z. Elamrani Abou Elassad, H. Mousannif, and H. Al Moatassime, "A real-time crash prediction fusion framework: an imbalance-aware strategy for collision avoidance systems," *Transportation Research Part C: Emerging Technologies*, vol. 118, p. 102708, 2020.

[33] G. Shen, L. Guan, J. Tan, and X. Kong, "DeepTSW: an urban traffic safety warning framework based on bayesian deep learning," in *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*, H. Ning and F. Shi, Eds., vol. 1329, pp. 50–63, Springer Singapore, Singapore, 2020.

[34] S. Amin, "Backpropagation–artificial neural network (BP-ann): understanding gender characteristics of older driver accidents in west midlands of United Kingdom," *Safety Science*, vol. 122, p. 104539, 2013.

[35] Y. Tian, D. Xiao, L. Wang, and H. Chen, "Expressway traffic safety early warning system based on cloud architecture," *Computer Communications*, vol. 171, pp. 140–147, 2021.

[36] Q. Chen, W. Wang, K. Huang, S. De, and F. Coenen, "Multimodal generative adversarial networks for traffic event detection in smart cities," *Expert Systems with Applications*, vol. 177, p. 114939, 2021.

[37] D. Xu, C. Wei, P. Peng, Q. Xuan, and H. Guo, "GE-GAN: a novel deep learning framework for road traffic state estimation," *Transportation Research Part C: Emerging Technologies*, vol. 117, p. 102635, 2020.

[38] X. Wang, J. Liu, T. Qiu, C. Mu, C. Chen, and P. Zhou, "A real-time collision prediction mechanism with deep learning for intelligent transportation system," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9497–9508, 2020.

[39] X. Shi, Y. D. Wong, C. Chai, and M. Z.-F. Li, "An automated machine learning (AutoML) method of risk prediction for decision-making of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 7145–7154, 2021.

[40] J. Yuan, M. Abdel-Aty, Y. Gong, and Q. Cai, "Real-time crash risk prediction using long short-term memory recurrent neural network," *Transportation Research Record*, vol. 2673, no. 4, pp. 314–326, 2019.

[41] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," in *Proceedings of the Learning and Technology Conference (L&T)*, pp. 40–45, New York, NY, USA, 2018.

[42] W. Ma, Z. He, L. Wang, M. Abdel-Aty, and C. Yu, "Active traffic management strategies for expressways based on crash risk prediction of moving vehicle groups," *Accident Analysis & Prevention*, vol. 163, p. 106421, 2021.

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

[44] J. Yao and Y. Ye, "The effect of image recognition traffic prediction method under deep learning and naive Bayes algorithm on freeway traffic safety," *Image and Vision Computing*, vol. 103, p. 103971, 2020.

[45] R. Saha, M. T. Tariq, M. Hadi, and Y. Xiao, "Pattern recognition using clustering analysis to support transportation system management, operations, and modeling," *Journal of Advanced Transportation*, vol. 2019, pp. 1–12, 2019.

[46] J. H. Chen, H. H. Wei, C. L. Chen, H. Y. Wei, Y. P. Chen, and Z. Ye, "A practical approach to determining critical macroeconomic factors in air-traffic volume based on K-means clustering and decision-tree classification," *Journal of Air Transport Management*, vol. 82, p. 101743, 2020.

[47] J. Lu, K. Wang, and Y. M. Jiang, "Real time identification method of abnormal road driving behavior based on vehicle driving trajectory," *Journal of Traffic and Transportation Engineering*, vol. 20, no. 6, pp. 227–235, 2020.

[48] X. Wang, X. Zhang, F. Guo, Y. Gu, and X. Zhu, *Effect of daily car-following behaviors on urban roadway rear-end crashes and near-crashes: a naturalistic driving study*, vol. 164, p. 106502, 2022.

[49] J. Wu, H. Wen, and W. Qi, "A new method of temporal and spatial risk estimation for lane change considering conventional recognition defects," *Accident Analysis & Prevention*, vol. 148, p. 105796, 2020.

[50] D. N. Reshef, Y. A. Reshef, H. K. Finucane et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[51] S. Valle, W. Li, and S. J. Qin, "Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods," *Industrial & Engineering Chemistry Research*, vol. 38, no. 11, pp. 4389–4401, 1999.

[52] Z. Wang, M. Liang, and D. Delahaye, "A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 280–294, 2018.

[53] F. Liu, K. Ting, and Z. Zhou, "Isolation forest," in *Proceedings of the 8th International Conference on Data Mining*pp. 413–422, Pisa, Italy, 2008.

[54] Z. Lin, X. Liu, and M. Collu, "Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105835, 2020.

[55] X. Ji, S. Xie, and W. Qin, "Dynamic prediction of traffic accident risk in risky curve sections based on vehicle trajectory data," *China Journal of Highway and Transport*, pp. 1–15, 2020, http://kns.cnki.net/kcms/detail/61.1313.U.20201017.1414.002.html.

[56] Q. Xue, K. Wang, J. J. Lu, and Y. Liu, "Rapid driving style recognition in car-following using machine learning and vehicle trajectory data," *Journal of Advanced Transportation*, vol. 2019, p. 11, 2019.

[57] J. Peng and Y. Shao, "Intelligent method for identifying driving risk based on V2V multisource big data," *Complexity*, vol. 2018, p. 1, 2018.

[58] L. Ferreira, M. S. Hoque, and A. Tavassoli, "Application of proximal surrogate indicators for safety evaluation: a review of recent developments and research needs," *IATSS Research*, vol. 41, no. 4, pp. 153–163, 2017.