IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 22, 2022, accepted March 19, 2022, date of publication March 23, 2022, date of current version April 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161835

Automated Traffic Incident Detection: Coping With Imbalanced and Small Datasets

TIAN XIE, QIANG SHANG, AND YANG YU
School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China

Corresponding author: Qiang Shang (shangqiang@sdut.edu.cn)

This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2021MF109, and in part by the Ministry of Education in China (MOE) Project of Humanities and Social Sciences under Grant 21YJC630110.

ABSTRACT Automatic incident detection (AID) has always been one of the focus issues in the field of transportation. However, due to the contingency and randomness of traffic incident, traffic incident samples are scarce and far less than non-incident samples. Therefore, unlike other scenarios using large-scale deep networks, traffic incident detection tackle at small and imbalanced sample size. Imbalanced, small sample data sets, inappropriate and incomplete initial variable sets make the AID model insensitive to incident samples, resulting in unsatisfactory model performance (low detection rate or high false alarm rate). Therefore, a hybrid AID method (SASYNO-RF-RSKNN) is proposed using self-adaptive synthetic over-sampling, random forest and random subspace k nearest neighbor. First, the spatial-temporal and real-time characteristics of traffic stream are used for the selection of appropriate initial variables to construct a relatively complete set of initial variables. Second, the SASYNO oversampling method is used to expand the original imbalanced sample database, so that the number of minority class samples is consistent with the number of most class samples. Then, feature variables are selected from the initial variables using the RF algorithm. Finally, the RSKNN ensemble algorithm with feature variables as input is employed to detect traffic incident. In addition, six indexes are used to evaluate model performance, including accuracy (ACC), false alarm rate (FAR), detection rate (DR), precision, Matthews correlation coefficient (MCC) and F1-score. Simultaneously, we also designed horizontal and vertical contrast experiments, and the experimental results show that SASYNO-RF-RSKNN model has superior performance. It is worth mentioning that experiments are implemented on two real-world datasets. Most indexes of the proposed model are the best compared with other five excellent machine learning algorithms. On the whole, the proposed model has a dependable and high-performance for traffic incident detection.

INDEX TERMS Traffic incident detection, ensemble-learning, spatial and temporal characteristic, SASYNO, random subspace k nearest neighbor.

I. INTRODUCTION

With the rapid increase of urban population and other urbanization activities, traffic conditions have deteriorated and the frequency of traffic accidents has increased significantly. Traffic incidents refer to non-repetitive incidents, such as traffic accidents, vehicle stalls, overflow loads, temporary construction and maintenance activities. The commonness of these incidents is the fact that they disturb the normal flow of traffic. Timely detection of incidents and take appropriate measures can reduce the risk of secondary incidents, improve the response and intervention of government and traffic managers, and minimize the loss of money and time in the incident.

Many advanced traffic management and information systems (ATMIS) has automatic detection algorithms to assist operators in detecting accidents. Automatic Incident Detection (AID) can be divided into direct detection method and indirect detection method according to data sources. The direct detection method refers to the use of algorithms to complete target recognition and tracking through the input of videos and images, and then to determine whether the incident occurs. Indirect detection method refers to the analysis of the change of traffic stream to deduce whether there is a traffic

The associate editor coordinating the review of this manuscript and approving it for publication was Razi Iqbal.

incident. Its data comes from numerical traffic information collected by various types of sensors, such as volume, speed and occupancy.

Nowadays, in the context of traffic big data, massive traffic information data will be generated at all times. Therefore, compared with the statistical method based on strict mathematical assumptions and function structure, machine learning method is more flexible in structure, and can deal with the complex and highly nonlinear relationship between dependent variables and independent variables, which is brilliant in traffic automatic detection. In this way, it seems that using the deep learning algorithm of large deep network in the machine learning algorithm can achieve better results. However, the accidental and random occurrence of traffic incidents makes the number of incident samples far less than that of non-incident samples, and the incident samples may also be scarce. Therefore, when faced with imbalanced and small sample scenarios, the deep learning network is not unique, and the selection of AID algorithm that can properly handle this scenario is the key.

AID currently has a wealth of research results, according to different types of algorithms to distinguish roughly the following:

A. TRADITIONAL ALGORITHM

California algorithms is one of the earliest algorithms used to detect sudden traffic incidents, which were developed from 1965 to 1970.It sets the threshold by comparing the change of occupancy between adjacent detectors. When the mutation exceeds the threshold, it is considered that there may be an incident. Subsequently, someone improved it and released 10 improved algorithms. The results showed that algorithms No. 7 and No. 8 performed best and further reduced the false alarm rate [1], [2]. Persaud *et al.* established the McMaster algorithm based on mutation theory. It uses the flow and occupancy data to construct a standardized model of the 'flow–occupancy' distribution relationship. By comparing the difference between the observed data and the template, it determines whether there is a sudden traffic incident [3]. In 2015, Cheng *et al.* considered detecting incidents near ramps with frequent interlaced flows on urban expressways. Based on the geometric conditions and detector locations, the expressway is divided into short segments. The equivalent upstream and downstream traffic flow density difference is defined and calculated using the loop detector data. A detection logic based on the pattern of the density difference fluctuation is then proposed. Compared with the aforementioned California 8 algorithm, the average detection time is shortened, but the detection rate on the ramp and the weaving area is significantly reduced [4].

In general, the traditional classical algorithm uses the mutation phenomenon of one or more variables between adjacent detectors to identify the occurrence of emergencies by comparing the normal template.

B. STATISTICAL ALGORITHM

The standard deviation method (SND) was proposed by the Texas Transportation Institution (TTI) in 1970-1975. Firstly, the standard deviation of traffic parameters within 3 min or 5 min is calculated based on statistical analysis theory. Then a traffic incident is determined if it is greater than a predetermined threshold in one cycle or two consecutive sampling cycles [5]. The double exponential smoothing algorithm similar to the standard deviation method was first proposed by Cook in 1974, which uses more complex prediction methods to predict traffic parameters. The algorithm gives the nearest traffic parameters greater weight coefficient and the double exponential smoothing value as the predicted value and then compares with the real value to define the tracking signal value. If it exceeds the predetermined threshold, the alarm triggers. In this way, changes in weather or flow will not easily send false alarms. 13 types of parameter tests show that the detection effect of flow and occupancy is better [6]. Chassiakos and Stephanedes proposed the low-pass filtering algorithm in 1993. The moving average method is used to remove the noise and high frequency components in the measured data of traffic parameters, and only the low frequency data is retained. By comparing the spatial occupancy difference of adjacent detectors in 3 minutes to determine whether traffic incidents occur. It has low false alarm rate and high detection rate, but the average detection time is longer [7]. Wang *et al.* designed a highway incident detection model based on partial least squares regression (PLSR) in 2007. The PLSR models are built with the components extracted from the training dataset, and it distinguish incidents state from normal traffic state according to the output whether exceeding the threshold predefined. The performance is better than the neural network and support vector machine. The model is sensitive to rare samples, and it is most important to select typical examples in practical use to construct the model [8]. Kinoshita *et al.* proposed an automatic detection algorithm based on a probability model in 2016. The probability model is introduced to describe the traffic state of various roads, and the expectation-maximization algorithm is used to learn the normal flow model. Several divergence discriminant templates are used to evaluate the difference between normal flow state and real-time state [9]. Such methods use the theory of statistical analysis to predict or combine traffic parameters into new statistics, and set the prediction range of new data according to the trend of real-time data. Comparing the real-time data with the new data, if it exceeds the scope, it is considered to be a traffic incident.

C. MACHINE LEARNING

Machine learning is dedicated to studying how to improve the performance of the system itself by means of calculation and experience. Since the 1990s, a number of AID algorithms based on machine learning have emerged. Its principle is to regard traffic incident detection as a binary classification problem, normal operation or emergency. Using the previous

historical traffic data as input, the computer system outputs 0 (normal) or 1 (incident) for real-time traffic state discrimination according to the learning algorithm. Artificial neural network (ANN) was first explored by Chew *et al.* in 1991 to apply to highway congestion detection, imitating human neuronal excitation transfer mode. Later, in 1993, Chew successfully applied ANN to the automatic incident detection of a mile section of urban expressway using the simulation model data [10]. In 1995 Chew and Ritchie proposed three new ANN-based neural network models, multi-layer feedforward (MLF), the self-organizing feature map (SOFM) and adaptive resonance theory 2 (ART2). The multi-layer feedforward neural network uses upstream and downstream traffic flow, speed and occupancy input, and the false alarm rate is much lower than that of California, McMaster and Minnesota algorithms [11]. Later, from 1997 to 1999, many scholars have developed and verified this neural network model in various data sets. The performance evaluation results clearly show that the neural network model has substantially improved the performance of incident detection, which can provide rapid and reliable accident detection for highways [12], [13]. In 2004, Srinivasan evaluated multi-layer feed-forward neural network (MLF), basic probabilistic neural network (BPNN) and constructive probabilistic neural network (CPNN) from the aspects of classification accuracy, adaptability and network size. The results show that the MLF model has the highest classification accuracy, and the CPNN model is superior to the other two models in adaptability and flexible structure [14]. A hybrid model combining partial least squares method and neural network was proposed in 2011 [15].

Support Vector Machine (SVM) [16], Bayesian network classifier [17], [18] and Decision tree [19] are excellent machine learning algorithms that are also introduced into traffic incident detection. Compared with traditional classical algorithms and statistical analysis theories, machine learning algorithms have more prospects and vitality. Wang *et al.* combined time series analysis (TSA) with SVM in 2013. The time series component predicts traffic, and the SVM component detects the incident according to real-time traffic, predicted normal traffic and the difference between them. Compared with the past, the average detection time is further shortened, and the false alarm rate (FAR) is similar [20]. In 2019, Li *et al.* proposed an incident detection model based on GAN-RF-SVM under small sample conditions. The generative adversarial network (GAN) was used to generate new incident samples, and the random forest (RF) algorithm was used to select variables. Finally, SVM was used as the incident detection model. Solving the problem of small sample size, unbalanced sample size and timeliness in the incident detection system, and reducing the false alarm rate of traffic incident detection [21]. In 2020, Jiang *et al.* also used factor analysis (FA) method to reduce the dimension of the initial correlation variables for the imbalance of incident data. Random forest (RF) was used to train the data set, and Matthews correlation coefficient (MCC) is calculated for the classification results as a new weight value to test data, so as to improve the overall classification performance of random forest algorithm for unbalanced data. The experimental results demonstrate that the model based on FA-WRF has better classification effect and is more competitive in dealing with imbalanced data classification [22]. In 2020, a hybrid AID method using Random Forest-Recursive Feature Elimination (RF-RFE) algorithm and Long-Short Term Memory (LSTM) network optimized by Bayesian Optimization Algorithm (BOA) was proposed. Experiments are conducted using real data. Compared with several advanced AID methods, this method has achieved good performance in almost all evaluation indexes [23].

In recent years, deep learning has been a rapidly developing field in the research of artificial intelligence, optimization and pattern recognition. At present, it has been widely used in image recognition technology. At the same time, many scholars have studied how to apply it on AID. For large-scale data sets, it is a more efficient framework and a hot research method. In 2020, Li used the generative adversarial network (GAN) to expand the sample size and balance the data set, and a temporal and spatially stacked autoencoder (TSSAE) is used to extract temporal and spatial correlations of traffic flow and detect incidents. This model can not only increase the amount of incident samples, but also balance the data set, and improve the real-time performance of detection [24]. In 2020, Jiang proposed a Long short-term memory (LSTM) based framework considering traffic data of different temporal resolutions (LSTMDTR) for crash detection. LSTM is an effective deep learning method for capturing the long-term dependence and dynamic transition of pre-collision conditions. Compared with machine learning methods and LSTM models with one or two temporal resolutions, the LSTMDTR model has been validated to perform better on crash detection and transferability [25].

D. ENSEMBLE LEARNING

Ensemble learning is a branch of machine learning algorithm. It completes the learning task by building and combining multiple learners. Ensemble learning algorithm performs well in classification, regression, outlier detection and other issues. It often combines some weak learning machines to obtain better learning patterns than single learning machine. In 2008, Cai *et al.* designed an automatic incident detection and alarm system based on the concept of multi-core SVM. According to seven different types of SVM learners, seven different input variables were processed respectively. Finally, alarm information was output through a combination layer. Its effectiveness and portability are analyzed by simulation data, and good results are achieved on multiple data sets [26]. In 2014, Liu *et al.* took into account the excellent performance of standard naive Bayes in incident detection. In order to improve the detection efficiency, an ensemble Bayesian classifier was constructed. Compared with standard naive Bayes and Bayesian decision tree algorithm, the ensemble Bayesian classifier has good robustness [18]. In 2019, Xiao designed a SVM and KNN ensemble classifier to solve the

problem that the current incident detection algorithm has obvious differences in detection results for different data sets. Firstly, a single SVM and KNN model are trained, and then SVM is used as the principal model. The strategy of KNN supplemented combines the two. It is shown that its robustness is better than other algorithms on both I-880 and PeMs datasets [27].

In summary, there have been a lot of traffic incident detection model based on the above several different areas of research theory. However, as previously mentioned problem traffic incident detection often faces imbalanced and small sample size scenarios. Reviewing the above several types of methods, it is found that this problem is rarely considered or solved specifically. From this perspective, automatic traffic incident detection has two problems unresolved.

Firstly, this paper is an article on machine learning algorithms, which is one of the four types mentioned above. In many machine learning applications, there is a significant difference between the prior probabilities of different classes, i.e., between the probabilities with which an example belongs to the different classes of the classification problem. This situation is known as the class imbalance problem and it is common in many real problems from telecommunications, web, finance-world, transportation, biology, medicine not only, and which can be considered one of the top problems in data mining today [28].

In traffic incident detection, the number of incident samples is far less than that of non-incident samples, and the number of incident samples is very small. Moreover, the two types of samples are seriously unbalanced, which often leads to poor incident detection results. However, the above machine learning algorithms rarely consider the scarcity and imbalance of incident samples.

Secondly, the previous algorithm for the establishment of initial variables is not really attention, a good initial variable set will greatly affect the final detection results. The basic parameters of traffic stream (flow, speed and occupancy) are often directly used as the initial variables of traffic incident detection, and they cannot fully characterize the traffic stream changes caused by traffic incidents. It is the basis of traffic incident detection that traffic incidents cause dramatic changes in traffic stream.

Based on these two problems, this paper proposes a solution:

For Problem 1: for the scarcity of samples, there are three common strategies: 1) improve the algorithm itself to apply to the dataset [22], 2) cost-sensitive learning approaches, 3) change the size of the sample. In contrast, because the third strategy is more simple to use, the application scope is more widely and popular. There are two main methods about third strategy: 1) minority class samples oversampling, 2) majority class samples undersampling.

Synthetic minority oversampling technique (SMOTE) is a common oversampling technique, which uses k nearest neighbor samples of incident samples to generate interpolation and then obtains new k samples [29]. But the new sample only uses the original incident sample, without considering the difference between the original incident sample and the normal incident sample. Self-adaptive synthetic oversampling technique (SASYNO) considers both the incident sample and the normal sample, and new samples generated by SASYNO are more reliable and comprehensive [30]. Therefore, SASYNO algorithm is used to generate traffic incident samples, expand the amount of incident samples, and balance the amount of incident samples and non-incident samples to obtain better detection results.

For Problem 2: Previous algorithms only rely on three traffic parameters of traffic volume, speed and occupancy to study, and then there are algorithms to combine the three traffic parameters to get new variables, such as California algorithm. Shang [23] and Jiang [22] added the prediction variables to the initial variable set. Li fully considered the real-time nature of traffic incident, and established a relatively comprehensive initial variable set with the traffic parameters five minutes before the incident as new variables [21]. On the basis of previous studies, a more comprehensive initial variable set with 57 variables is established based on the spatial-temporal characteristics and real-time nature of traffic incidents. Compared with previous studies, the constructed new variable set is more reasonable and reliable. The high-dimensional data structure makes the algorithm have to adapt to and learn this data structure to play its advantages. Random forest (RF) and random subspace k-nearest neighbor (RSKNN) are mature machine learning algorithm, but also ensemble learning algorithm. The high stability of ensemble learning algorithm is one of the important advantages of RF and RSKNN algorithm. The idea of them is to randomly select samples, randomly select sample space and construct multiple learners, so it can make good use of high-dimensional data. Therefore, RF is used for feature selection to 57 initial variables, and RSKNN is used to identify traffic incident.

In this paper, based on loop data of PORTAL highway trunk, a new traffic incident detection framework using SASYNO, random forest (RF) and random subspace k-nearest neighbor (RSKNN) ensemble learning method is proposed. SASYNO to obtain new samples, random forest feature selection and RSKNN classification is expected to obtain better incident detection performance. The three main contributions of this paper are as follows:

1. On the basis of previous studies, a relatively complete set of initial variables is established according to the spatial-temporal correlation and real-time characteristics of traffic flow during the impact period of the traffic incident.

2. A new automatic incident detection framework is established. Firstly, SASYNO algorithm is used to balance the huge gap between the numbers of two types of samples in the sample database. Then, the importance of the newly established complete initial variable set is extracted by RF.

Finally, the ensemble learning algorithm RSKNN is used for incident detection. It is worth noting that SASYNO and RSKNN are both first used in traffic incident detection.

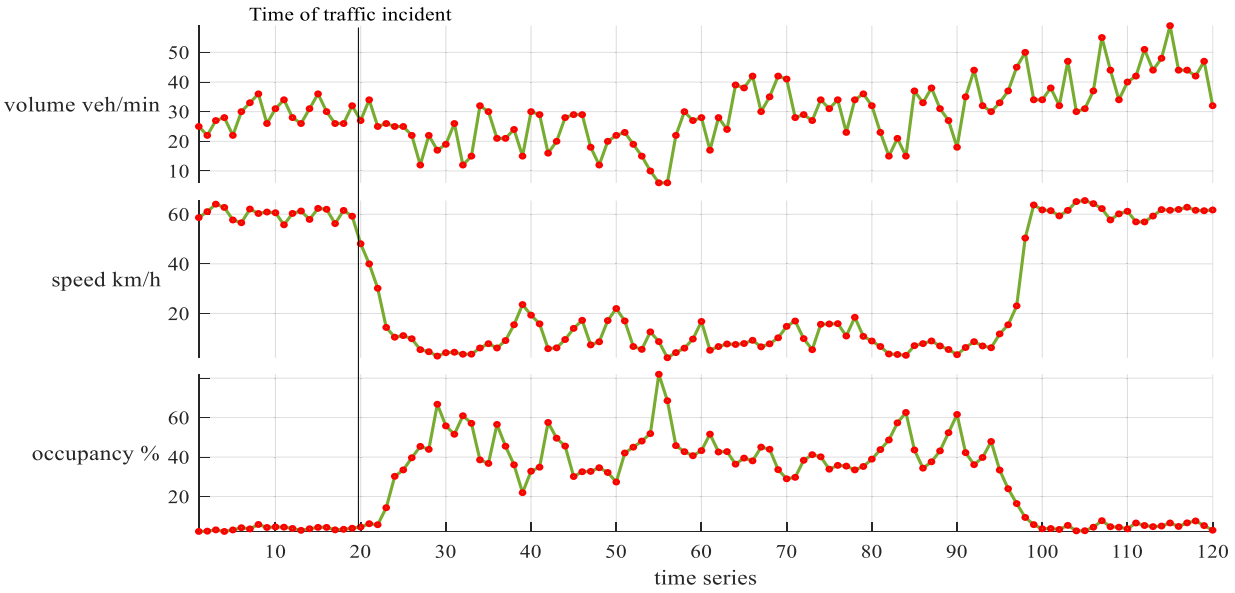

FIGURE 1. Time variation diagram of upstream detector volume, speed and occupancy.

3. Experiments are carried out on two real-world datasets to verify the practicability and robustness of the proposed method. At the same time, a horizontal comparison experiment and a vertical comparison experiment are implemented, which compare many state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 illustrates the methodology, in which the construction rules of the initial variable set are given. Then the principle of SASYNO method balancing the sample database, the selection of feature variables based on RF algorithm and the classification steps of RSKNN ensemble algorithm are illustrated. Finally, the whole process of the proposed method is presented. Section 3 is devoted to experiments, including data description and preprocessing, evaluation indexes, experimental design, and experimental results and analysis. In the end, Section 4 gives the conclusions and future work.

II. METHODOLOGY

A. CONSTRUCTION OF THE INITIAL VARIABLES SET

In previous algorithms, traffic parameters (traffic volume, speed and occupancy) are the main data for AID. When a major impact of traffic incident occurs, the traffic parameters will change dramatically. For example, due to congestion at the accident point, traffic volume and speed at upstream locations will decrease and occupancy increases. On the contrary, the downstream traffic volume and occupancy will decrease, and the speed will be improved, as shown in FIGURE 1 and FIGURE 2. Therefore, these parameters are often combined by some methods to get more sensitive traffic variables [1], [2]. In order to obtain a more complete initial variables set, 57 variables are selected from the following aspects in this study.

1) EARLY WARNING BEFORE THE INCIDENT

Before an incident, the traffic parameters are same as that of normal conditions. However, once an incident occurs, traffic parameters change drastically in a short period of time. Therefore, the traffic parameters within a short period of time before and after the incident start time are used as initial variables. Specifically, the traffic parameters collected by the upstream and downstream detectors at 1 min, 2 min, and 3 min before and after the incident start time are used as initial variables respectively. In summary, 36 variables are selected in this part.

2) MEASURED AND PREDICTED VALUES OF TRAFFIC PARAMETERS AT THE TIME OF THE INCIDENT

The predicted values of traffic parameters when a traffic incident occurs are obtained based on the actual measured data a few minutes before the incident. The predicted value can reflect the normal trend of the measured data a few minutes before the incident. However, the traffic incident will cause this abnormal trend, resulting in a significant difference between the measured values and the predicted values of traffic parameters, and this difference can be used for traffic incident detection. In this study, the moving average method is used to predict the traffic parameters based on the measured data of the upstream and downstream detectors in the first 3 minutes. In summary, 12 variables are selected in this part.

3) THE DIFFERENCE BETWEEN THE TRAFFIC PARAMETERS COLLECTED BY THE UPSTREAM AND DOWNSTREAM DETECTORS

The upstream and downstream traffic parameters of the incident site show obvious different trends, and the difference between the upstream and downstream traffic parameters can

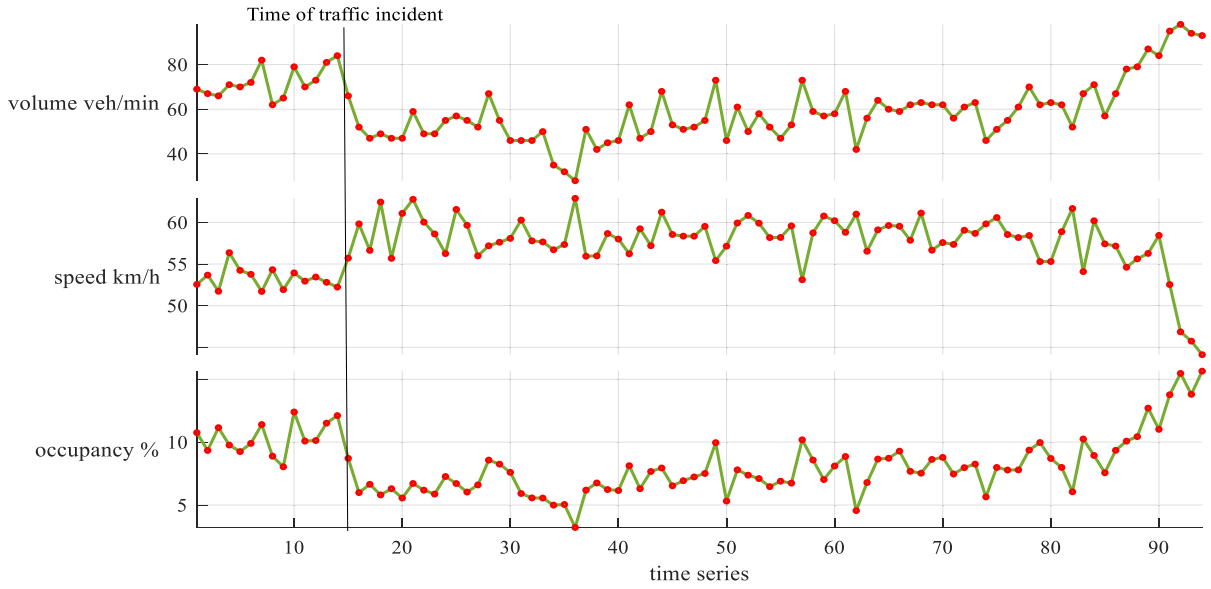

FIGURE 2. Time variation diagram of downstream detector volume, speed and occupancy.

reflect the abnormal condition caused by the incident to a certain extent, which is helpful for traffic incidents. In this part, 3 variables are constructed as initial variables.

4) THE DIFFERENCE BETWEEN THE MEASURED AND PREDICTED VALUES OF TRAFFIC PARAMETERS COLLECTED BY THE SAME DETECTOR

The difference between the measured values and the predicted values of traffic parameters can reflect the impact of the traffic incident. If there is no impact of the traffic incident, the predicted values will usually not deviate significantly from the measured values. In this part, 6 initial variables are constructed as initial variables.

Finally, as shown in TABLE 1, a relatively complete initial variables set with 57 variables is constructed. Subsequently, feature selection of initial variables using RF will be introduced in subsection C.

B. PROCESSING IMBALANCED DATA USING SASYNO

The scarcity of traffic incident samples has been mentioned earlier. Compared with non-incident samples, it is more difficult to obtain incident samples. In order to deal with the problem of imbalanced samples, a novel oversampling method called SASYNO is used in this study, which was first proposed in 2020 [30].

Popular approaches for imbalance learning generally can be categorized into three major types: 1) data sampling, 2) cost-sensitive learning and 3) algorithmic modification. Data sampling approaches rebalance the data sets by sampling, which is achieved by over-sampling the minority class, under-sampling the majority class or a hybrid of both. Cost-sensitive learning approaches incorporate the costs of misclassifying minority class samples into function minimization. Algorithmic modification approaches are the modifications of commonly-used machine learning algorithms to achieve better performance with imbalanced data set.

Currently, data sampling approaches are the dominant solutions to address the class imbalance problem because they are more generic and can be employed by standard classification methods.

The oversampling method like SMOTE that it randomly selects a small number of samples and creates linear interpolation between them and their neighbors. However, this strategy does not necessarily expand the database, and is more likely to overlap between the extended minority classes and the original majority classes, especially in the case of complex data structure [30]. On the contrary, the key idea of SASYNO is to select them according to the distance between adjacent minority samples, and create interpolation and extrapolation methods around the adjacent samples to synthesize data. The main steps of the method are as follows.

Step 1 (Identifying Pairwise Neighbouring Samples): Firstly, the average distance of minority class samples is calculated according to Equation (1), and then based on the average distance γ, the decision condition (2) is set up to identify the adjacent sets in pairs.

$$\gamma = \frac{1}{n}\sum_{i\neq j}\|x_i - x_j\| \tag{1}$$

where $\|x_i - x_j\| = \sqrt{(x_i - x_j)^T (x_i - x_j)}$ represents the Euclidean distance between two minority samples, and n is the number of such paired samples.

$$if \ \left(\|x_i - x_j\| < \gamma\right) \quad then \ \left(x_i, x_j\right) = (p, q) \subseteq P \tag{2}$$

where x_i, x_j are minority class samples, P is a set of pairs of adjacent samples, p, q is a set of x_i, x_j.

TABLE 1. Initial variables set for traffic incident detection.

Initializing variables	Descriptions	No.		
Upstream forward 3 minute volume/speed/occupancy	$up.f.3\ v/s/o$	1	17	33
Upstream forward 2 minute volume/speed/occupancy	$up.f.2\ v/s/o$	2	18	34
Upstream forward 1 minute volume/speed/occupancy	$up.f.1\ v/s/o$	3	19	35
Upstream measured volume/speed/occupancy	$up.m\ v/s/o$	4	20	36
Upstream back 1 minute volume/speed/occupancy	$up.b.1\ v/s/o$	5	21	37
Upstream back 2 minute volume/speed/occupancy	$up.b.2\ v/s/o$	6	22	38
Upstream back 3 minute volume/speed/occupancy	$up.b.3\ v/s/o$	7	23	39
Upstream predicted volume/speed/occupancy	$up.p\ v/s/o$	8	24	40
Downstream forward 3 minute volume/speed/occupancy	$down.f.3\ v/s/o$	9	25	41
Downstream forward 2 minute volume/speed/occupancy	$down.f.2\ v/s/o$	10	26	42
Downstream forward 1 minute volume/speed/occupancy	$down.f.1\ v/s/o$	11	27	43
Downstream measured volume/speed/occupancy	$down.m\ v/s/o$	12	28	44
Downstream back 1 minute volume/speed/occupancy	$down.b.1\ v/s/o$	13	29	45
Downstream back 2 minute volume/speed/occupancy	$down.b.2\ v/s/o$	14	30	46
Downstream back 3 minute volume/speed/occupancy	$down.b.3\ v/s/o$	15	31	47
Downstream predicted volume/speed/occupancy	$down.p\ v/s/o$	16	32	48
the difference between the measured values of upstream and downstream volume	$up-downm.m\ v$		49	
the difference between the measured values of upstream and downstream speed	$up-downm.m\ s$		50	
the difference between the measured values of upstream and downstream occupancy	$up-downm.m\ o$		51	
the difference between the measured value and the predicted value of the upstream volume	$up.m-p\ v$		52	
the difference between the measured value and the predicted value of the upstream speed	$up.m-p\ s$		53	
the difference between the measured value and the predicted value of the upstream occupancy	$up.m-p\ o$		54	
the difference between the measured value and the predicted value of the downstream volume	$down.m-p\ v$		55	
the difference between the measured value and the predicted value of the downstream speed	$down.m-p\ s$		56	
the difference between the measured value and the predicted value of the downstream occupancy	$down.m-p\ o$		57	

Step 2 (Creating Explorations by Gaussian Disturbance): In this stage, the algorithm randomly selects a pair of neighbouring samples (p_k, q_k) from the collection P and apply Gaussian disturbance to create extrapolations in the data space. As shown in equation (3).

$$(p_k, q_k) = \left(p_k + G_i, q_k + G_j\right) = \left(\hat{p}_k, \hat{q}_k\right) \tag{3}$$

where $G_i = [g_1, g_2, \cdots, g_m]$ are m dimensional randomly generated vectors following the Gaussian distributions. $g_l \sim N(0, \sigma_l^2), l = 1, 2, \cdots, m$.

$$\sigma_l = \frac{1}{u} \sum_{i \neq j} \left| x_{i,l} - x_{j,l} \right| \tag{4}$$

where $x_{i,l}$ represents the lth variable of sample i, and $|\cdot|$ represents the absolute value.

Step 3 (Creating Interpolations for Synthetic Data Generation): Finally, according to the extrapolation value $(\hat{p}_k, \hat{q}_k)$ obtained in the previous step, random interpolation is created between the two samples to generate new samples. As shown in equation (5).

$$x_{new} = \hat{p}_k + rand\ (0, 1) \times \left(\hat{q}_k - \hat{p}_k\right) \tag{5}$$

It follows that the uniqueness of SASYNO comes from the following two aspects:

1) The approach selects out the most proper candidates from minority class samples and uses them for data synthesis only. This allows SASYNO to precisely expand the minority class avoiding possible overlaps with the majority class.

2) The approach employs Gaussian disturbance to create extrapolations from existing data samples for synthetic data generation, which gives SASYNO an extra degree of freedom for expanding the knowledge base.

C. VARIABLE IMPORTANCE RANKING BASED ON RF

RF algorithm is an ensemble learning algorithm based on decision trees. The main idea is that it constructs the random sample subspace and the random feature variable subspace.

Thus, each decision tree classifier generated by these two strategies is relatively independent and the final classification

result is built on the maximum of all decision tree results. The main steps of importance ranking based on RF algorithm are as follows.

Step 1 (Random Sample Subspace): The Bootstrap method was used to randomly select N data samples from the original data set M and repeat k times to generate k training sets. Data that is not selected in each extraction process constitutes out-of-bag (OOB) data and forms a test set

Step 2 (Random Feature Subspace): Select l features as candidate features randomly from L feature dimensions ($l < \mathrm{L}$), maximize the growth of each decision tree according to CART algorithm, repeat the above operation for k training sets respectively, and finally obtain k decision trees and constitute RF.

Step 3 (Classification Using Out-of-Bag Data): Use k decision trees in RF to make decisions on test set (OOB) and obtain the classification accuracy T_k.

Step 4 (Adding Noise to Out-of-Bag Data and Classification): Each initial variable in the training set is recorded as $\lambda_s (s = 1, 2, \cdots, 57)$, and random noise is added to λ_s of the OOB data L_k^{OOB} to obtain a new OOB data $\hat{L}_k^{OOB}$, and the classification accuracy $\hat{T}_k$ of each decision tree using the new OOB data $\hat{L}_k^{OOB}$ is calculated.

Step 5: Calculate the importance of each variable according to equation (6)

$$VI = \frac{1}{k}\left(T_k - \hat{T}_k\right) \tag{6}$$

D. RANDOM SUBSPACE K NEAREST NEIGHBOR CLASSIFIER (RSKNN)

Random subspace method (RSM), also called feature bagging, is one of ensemble learning. Random subspace trains each classifier by using randomly selected partial features rather than all features to reduce the correlation between each classifier. In fact, the random forest algorithm is a decision tree algorithm using RSM and bagging. Similarly, RSM can be supported in other classifiers such as KNN. Then, KNN is used to predict each subspace, and the corresponding results of each classifier are obtained. Final results obtained by majority voting.

KNN algorithm is a relatively mature machine learning algorithm. The basic idea of KNN classification algorithm is that it finds k labeled samples with the nearest neighbors of the samples to be classified in the feature space, and the class of the most diverse samples is the class of the samples to be classified. In classification decision, the algorithm mainly depends on the nearest neighbor samples to determine the class, rather than relying on the classification hyperplane. Therefore, KNN classification algorithm is more suitable for samples with certain cross or overlapping feature space of different classes. FIGURE 3 shows the classification process of RSKNN.

First, the sample database after feature selection is split into training samples and test samples. Second, 15 feature variables are randomly selected for all training samples, and 30 different random subspaces are produced by repeating this operation for 30 times. Third, KNN model is used to train these 30 subspaces. Finally, ensemble KNN model is tested using the test data, and most of the classes are voted for the final results.

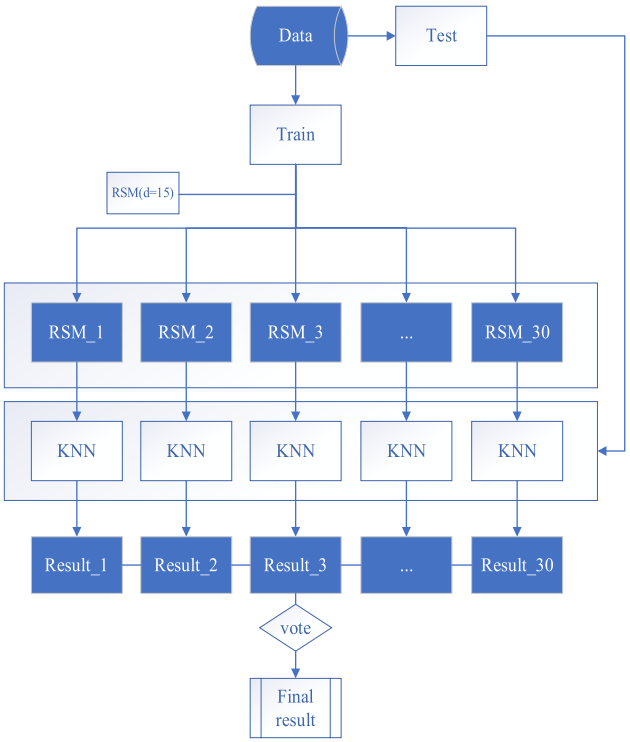

FIGURE 3. Flowchart of RSKNN algorithm.

The RSKNN algorithm has two important parameters. One is the number of subspaces, namely the number of learners, and the other is the dimension of each subspace, namely the number of feature variables. For each data set, this pair of parameters is diverse. The 15 feature variables and 30 learners selected are also for I-205 data sets, which is not the case for another I-880 data set in this paper. In order to determine this pair of parameters, we also conducted experiments to find the optimal classification accuracy under different dimensions and different learners, as shown in TABLE 2. As can be seen from TABLE 2, the classification accuracy of 15 features variables and 30 learners is the highest. At the same time, the lower the dimension, the higher the classification accuracy will need more learners. However, when the learner reaches a certain number, it cannot further improve the accuracy but will increase the running time.

E. THE PROPOSED METHOD (SASYNO-RF-RSKNN)

The flowchart of the proposed method is illustrated in FIGURE 4. As can be seen from FIGURE 4, the application of the proposed AID method includes six steps. The specific steps are as follows.

Step 1: Collection of highway loop detectors data including traffic flow, speed and occupancy.

TABLE 2. Classification accuracy corresponding to different subspace dimensions and different subspace numbers on I-205 highway.

Subspace dimension	Number of subspaces									
	5	10	15	20	25	**30**	35	40	45	50
25	92.9	92.9	91.4	92.9	92.9	92.9	94.3	92.9	94.3	92.9
25	91.4	92.9	94.3	94.3	94.3	94.3	94.3	94.3	94.3	92.9
15	91.4	91.4	91.4	94.3	94.3	**95.7**	94.3	94.3	94.3	92.9
10	88.6	90	91.4	91.4	94.3	95	94.3	95.7	94.3	94.3
5	90	90	90	91.4	91.4	90	88.6	94.3	91.4	91.4

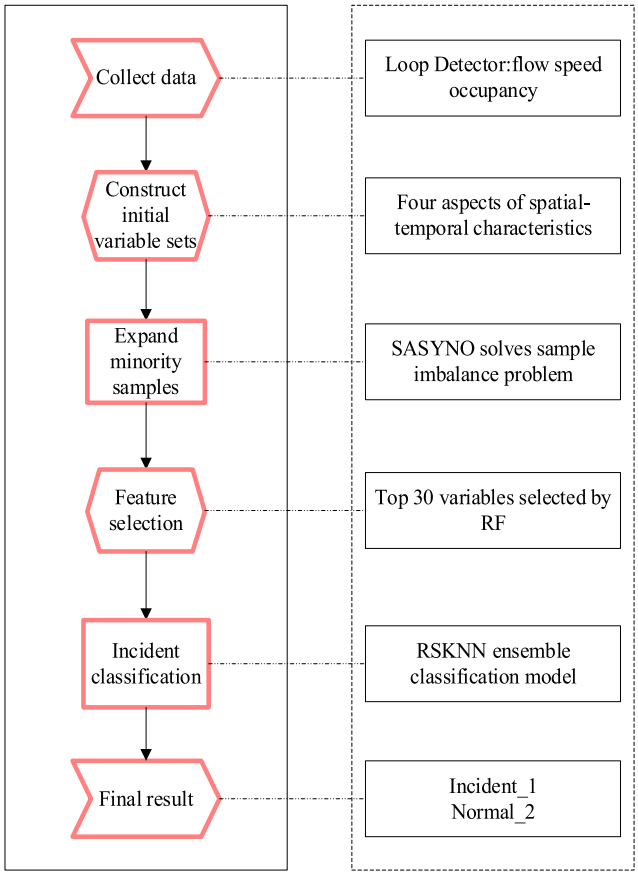

FIGURE 4. Flowchart of SASYNO-RF-RSKNN based AID method.

Step 2: The initial sample set is constructed according to the above four variable extraction rules.

Step 3: SASYNO method is used to balance incident samples and non-incident samples in the database.

Step 4: RF algorithm is used to select feature top 30 variables for AID.

Step 5: RSKNN ensemble learning algorithm is used to classify incident database.

Step 6: SASYNO-RF-RSKNN incident detection model output classification results 1 or 2. Among them, 1 indicates an incident, and 2 indicates normal (non-incident).

III. EXPERIMENTS

This section is the experimental part. Firstly, the data source of this experiment is introduced, and the over-sampling technology and data standardization preprocessing method are used to split up the test training set. Secondly, performance evaluation indexes of this study are introduced, including accuracy, false alarm rate, detection rate, precision, MCC and F1-score. Then we will introduce the design idea of this experiment, according to the principle of single variable set the horizontal contrast experiment and the vertical contrast experiment. Finally, the test results were compared and analyzed to bring to the conclusion.

A. DATA DESCRIPTION AND PREPROCESSING

The experimental data are from the test data set project of the Federal Highway Administration (FHWA) of Portland State University (PSU). PORTAL is an official transport data archive for the Portland-Vancouver metropolitan area.

In this study, I-205 highway loop data of PORTAL from 15 September 2011 to 15 November 2011 are used. The section of I-205 NB covered by this test data set is 10.09 miles long and ranges from the Sunnyside Road ramp at milepost 14.32 to milepost 24.41, approximately one mile past the end of the detection. The section of I-205 SB covered by this test data set is 12.01 miles long and runs from Sunnyside Road ramp at milepost 14.58 to milepost 26.59, approximately one mile past the end of the detection.

There is a total of 18 loop detection stations in the north-south direction. Each detector records traffic parameters such as traffic volume, speed and occupancy. The sampling frequency is 20s, and the detector is about 1.5 miles apart. FIGURE 5 shows the position of the detection station on I-205 highway. Accident database records in detail the location and start-stop time of the incident, as well as the causes and impact of the incident. According to the time, location information and influence degree of the incident sample, the typical incident with significant influence on I-250 highway is selected. Because the number of normal samples is large, in addition to randomly selecting normal samples, the data of the previous and following days at the same time and place is also used as normal samples. Being dependent on

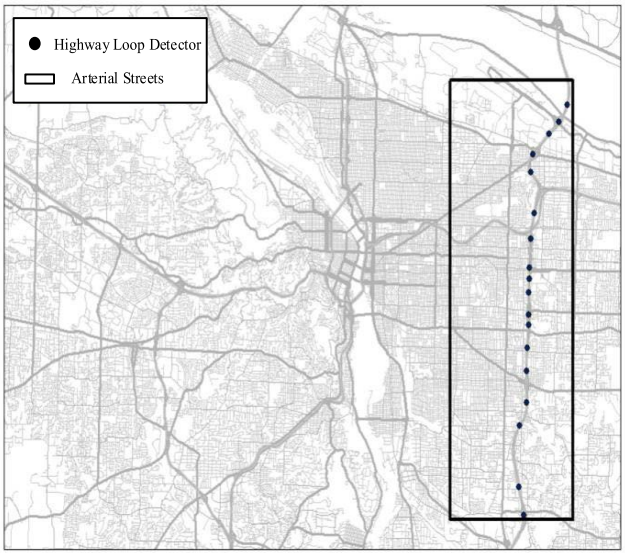

FIGURE 5. I-205 highway detector layout.

the construction method of the initial variable set, 18 incident samples and 118 normal incident samples are finally obtained, and the incident samples account for 13% of the total samples. The data imbalance in the original sample will seriously affect the classification results, so SASYNO is used to oversample and constructing the following input matrix.

$$Input = \left[\lambda_{i,j}\ Y_i\right] = \begin{bmatrix} l\lambda_{1,1} & \lambda_{1,2} & \cdots & \lambda_{1,57} & Y_1 \\ \lambda_{2,1} & \lambda_{2,2} & \cdots & \lambda_{2,57} & Y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{m,1} & \lambda_{m,2} & \cdots & \lambda_{m,57} & Y_m \end{bmatrix} \tag{7}$$

where m represents the total sample size, $\lambda_{i,j}$ represents the jth variable of the i sample, Y_i represents the class attribute of the i sample, 1 represents the incident sample, 2 represents the normal sample, where $m = 136$. Finally, the output matrix similar to the input matrix is obtained, and the total sample $m = 236$.

$$Output = \begin{bmatrix} Input \\ Newsample \end{bmatrix} \tag{8}$$

In order to eliminate the influence of different dimensions, improve the training speed and classification accuracy, the data is normalized to the interval [0, 1]. The normalization formula is as follows.

$$\lambda_{new} = \frac{\lambda_{old} - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \tag{9}$$

where λ_{new} is the standardized variable, λ_{old} is the original variable, $\lambda_{\max}, \lambda_{\min}$ are the maximum and minimum values of the original variables respectively.

In order to reflect the practicability and robustness of our method, we also prepared a new data set I-880 highway data. I-880 is a specific highway in San Francisco Bay Area, which can download data from the website [27]. The data recorded the flow, speed and occupancy of a 49,700-foot road from Marina to Whipple. There are 20 stations, each about 0.5 miles apart. More detailed information is available in the Reference [27], [31].

In order to obtain more reasonable experimental results, the total sample library is randomly divided into a test set and training set at 3:7. At the same time, the training set is trained by a 5-fold cross-validation method. The training set is randomly divided into five samples of the same size. Four of them are selected as the training samples during each training, and the remaining one is used as the verification sample, which is repeated five times. So far, all the samples are used as the verification samples. The two partitioned sample databases are shown in TABLE 3 and TABLE 4.

TABLE 3. Overview of I-205 sample database.

	Number of incident samples	Number of normal samples	Total samples
Original samples	18	118	136
SASYNO samples	118	118	236
Training set	81	85	166
Testing set	37	33	70

TABLE 4. Overview of I-880 sample database.

	Number of incident samples	Number of normal samples	Total samples
Original samples	24	120	144
SASYNO samples	120	120	240
Training set	82	86	168
Testing set	38	34	72

B. EVALUATION INDEXES

In this study, incident detection can be regarded as a binary classification problem, and the confusion matrix for the binary classification problem is one of the most frequently used results, which can visually reflect the performance of the model. It is shown in TABLE 5.

True Positive (TP) represents the number of actual incident samples predicted as incident samples. False Negative (FN) represents the number of actual incident samples predicted as normal samples. Similarly, False Positive (FP) and True Negative (TN) denote the number of normal samples in actual

TABLE 5. Confusion matrix.

True	Prediction	
	incident	normal
incident	TP	FN
normal	FP	TN

classes predicted as incident samples and normal samples. Through the confusion matrix can be very clear to calculate the six performance indicators needed in this study, accuracy (ACC), false alarm rate (FAR), detection rate (DR), precision rate, Matthews correlation coefficient (MCC) and F1-score.

Accuracy is a very common evaluation index in classification, which reflects the overall detection accuracy of the model, that is, all the predicted samples, how much the correct probability is predicted, whether it is the incident sample or the normal sample. The calculation formula is as follows.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

The false alarm rate is relative to the incident sample, which indicates the proportion of false prediction in the actual normal sample. The calculation formula is as follows.

$$FAR = \frac{FP}{FP + TN} \tag{11}$$

Detection rate is also called recall rate, which means that the proportion of incident samples is correctly predicted in the real class incident samples. The calculation formula is as follows.

$$DR = \frac{TP}{TP + FN} \tag{12}$$

Precision rate is also called accurate-checking rate, which measures the ability of classifier to correctly identify incident samples. It means that the proportion of correct prediction in the samples predicted as incident samples. The calculation formula is as follows.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

Matthews correlation coefficient (MCC) is a comprehensive index used in machine learning classification, which considers true positive, true negative, false positive and false negative. Even when the sample content of the two classes is very different, it can also be applied. It describes the correlation coefficient between the actual classification and prediction classification. Its value range is $[-1, 1]$. When the value is closer to 1, the prediction is more perfect and 0 means random prediction. When the value is -1, it means completely irrelevant. The calculation formula is as follows.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

F1-score is the harmonic average of precision and recall. Obviously, this is also a comprehensive evaluation index, so it is also applicable to the situation of unbalanced data distribution. The value range is located between [0, 1], and 1 represents the best classification. The calculation formula is as follows.

$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{15}$$

C. EXPERIMENTAL DESIGN

In this experiment, two control groups were set up according to the principle of single variable to test the performance of SASYNO-RF-RSKNN incident detection model. They are horizontal contrast and vertical contrast.

In horizontal contrast, the same preprocessing procedure for data oversampling, standardization and RF feature selection techniques are performed firstly. Then, in order to reflect the advantages of the ensemble learning algorithm RSKNN, we compare a classical classification algorithm SVM [16] and a newly proposed eigenvalue classification (EigenClass) model [32]. EigenClass model is a supervised machine learning algorithm based on the input matrix eigenvalue calculation and threshold setting proposed by Erkan in 2020. It runs and tests 30 time in 20 different datasets. The results demonstrate that EigenClass has the best classification performance for 15 datasets in each index, reflecting its superior accuracy and strong stability. Similarly, the classical oversampling method SMOTE is also compared with SASYNO. The SMOTE method has been introduced in the previous section, so it is not elaborated here. The rest of the two methods are consistent. Without special emphasis, the above methods are implemented under the same conditions, when they are used to complete the same process. The hyperparameters settings of each method are shown in TABLE 6 below.

TABLE 6. The hyperparameters of horizontal comparison methods.

Methods	Hyperparameters
SMOTE-RF-RSKNN	The nearest neighbor k of SMOTE is set to 5. The number of incident samples is consistent with the number of normal samples.
SASYNO-RF-RSKNN	For RF, the number of random feature variables is the square root of the total number of feature variables, and the number of decision trees is set to 1000.
SASYNO-RF-SVM	SVM kernel function is a linear kernel function.
SASYNO-RF-EigenClass	EigenClass calculation method is determined by the size of input matrix.

Many excellent machine learning algorithms are mentioned in the previous research review. In the vertical contrast part, we also selected five machine learning algorithms including three deep learning algorithms that are often cited in recent years. They are ensemble SVM and KNN algorithm (E-SVM-KNN) in 2019 [27], GAN-RF-SVM deep learning

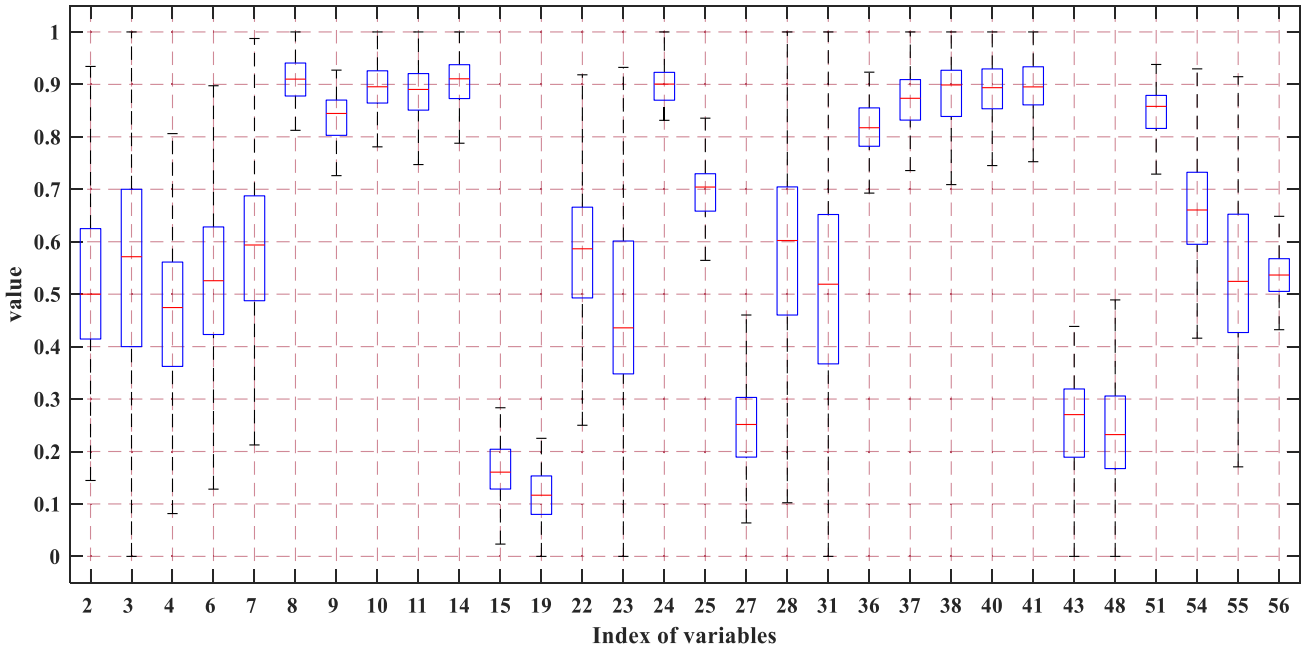

FIGURE 6. Distribution of original variables.

algorithm in 2019 [21], FA-WRF ensemble learning algorithm in 2020 [22], GAN-TSSAE deep learning algorithm in 2020 [24], LSTMDTR deep learning algorithm in 2020 [25]. The introduction of these methods has been involved in the previous method summary. By comparing these excellent algorithms in recent years to verify whether the SASYNO-RF-RSKNN incident detection model proposed in this paper has excellent performance. In order to ensure the performance of the comparison method, the parameters of these methods are set and optimized according to the corresponding literature. Similarly, in order to assure the fairness of comparison, all methods are set under the same sample database. The specific sample input form is also set according to the form mentioned in the literature. The hyperparameters settings of each method are shown in TABLE 7 below.

TABLE 7. The hyperparameters of vertical comparison methods.

Methods	Hyperparameters
SASYNO-RF-RSKNN	For RSKNN, the subspace dimension is 15 and the classifier is set to 30.
E-SVM-KNN	The k values of KNN and SVM kernel function need to be manually selected according to the sample library. Here we choose Polynomial kernel function of degree 2, k = 10.
GAN-RF-SVM	The generator and the discriminator are three-layer fully connected layer network structures. RF selects 28 variables, and SVM kernel function is a linear kernel function.
FA-WRF	FA selected seven effective factors, WRF subspace dimension is 3, the number of decision trees is 600.
GAN-TSSAE	The generator and the discriminator are three-layer fully connected layer network structures. For TSSAE, the number of hidden layers was set to 3 with 39, 20, and 10 hidden nodes.
LSTMDTR	The number of neurons in LSTM is 64; the dropout rate is 0.4; the number of neurons in the second fully-connected layer is 64; the epoch is 100; the batch size is 50; the optimization algorithm is Adam.

D. EXPERIMENTAL RESULTS AND ANALYSIS

1) EXPERIMENTAL RESULTS OF SASYNO ALGORITHM

In the data description and preprocessing section, we describe how to construct an input matrix using the SASYNO method, as detailed in formula (7). To evaluate the generated dataset, box plots are used to describe the distribution of a given variable, including minimum, lower quartile, median, upper quartile, and maximum. It is noteworthy that in the pre-processing phase we have used standardized techniques to eliminate the impact of different dimensions. Similarly, in order to eliminate the randomness of the over-sampling algorithm, we carried out many experiments, and took the average as the final result. In order to avoid duplication, we only show the data results of I-205 highway. FIGURE 6 and FIGURE 7 show the distribution of the original data and the sample distribution after oversampling. The variable set presented here is 30 variables after feature selection. The number of the X axis corresponds to the number of the initial variable set in the second section, and the Y axis is the numerical size. Because standardization to the [0, 1], the maximum value of the Y axis is 1.05. It can be observed in these two figures that the distribution of original data and generated data is very similar. The median of the generated data is almost the same

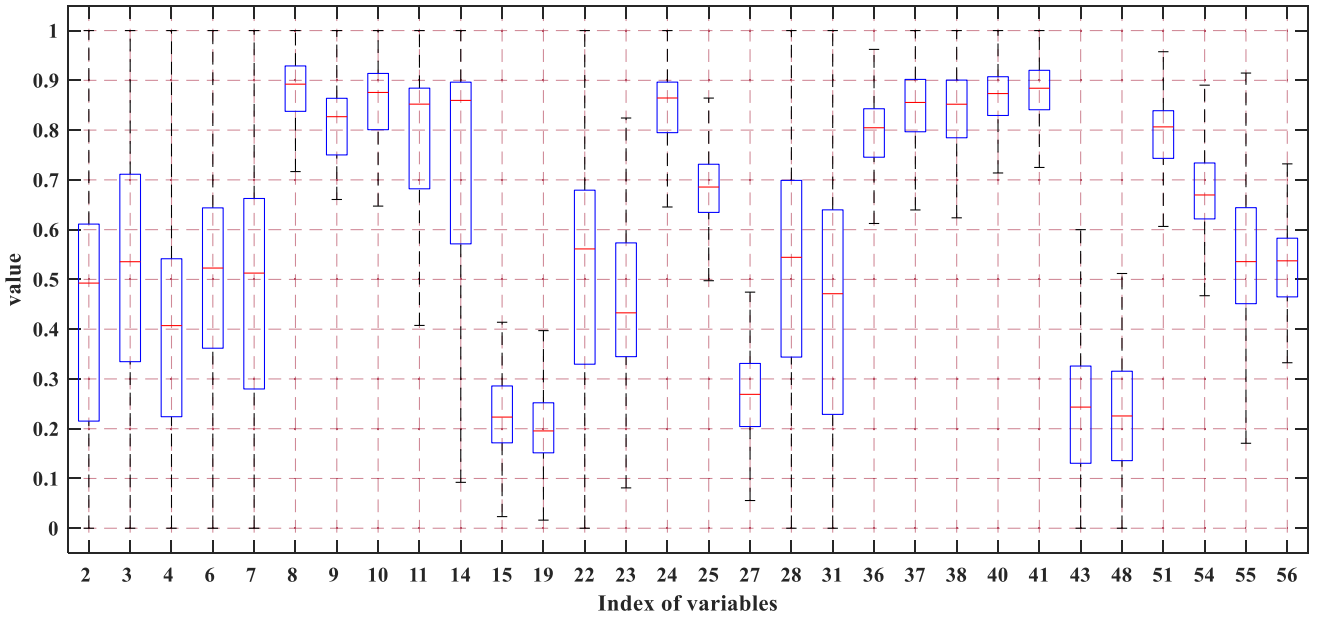

FIGURE 7. Distribution of generate sample variables.

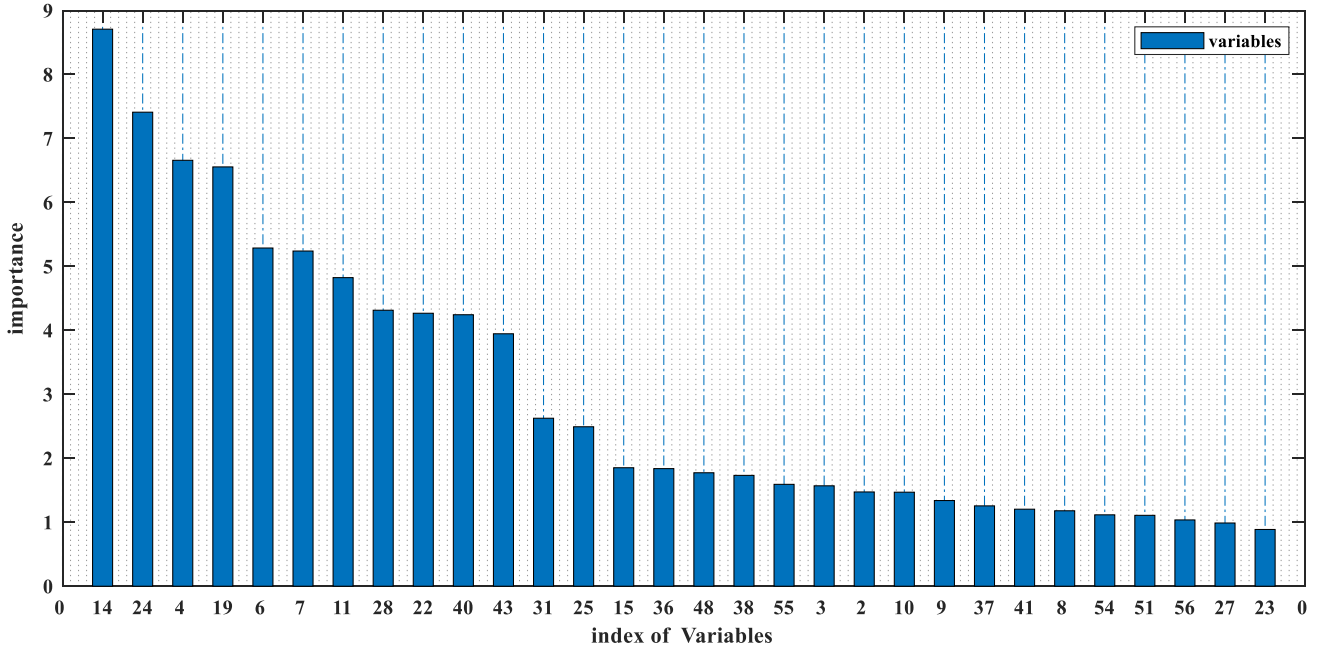

FIGURE 8. Top 30 variables selected by random forest of I-205 highway.

as the original data. The lower quartiles of the generated data are similar except for 7, 14 and 22 variables, and the upper quartiles of all variables have not changed much. Compared with the original data, the maximum and minimum values of No.4, No.11 and No.14 variables are expanded, which should be that the variable values are relatively dense and large, so the data sensitivity is easy to make the oversampling method appear abnormal values. Based on the above analysis, it can be said that SASYNO algorithm has excellent stability and can be used as an effective means to deal with unbalanced samples in traffic incident detection.

2) FEATURE SELECTION RESULTS OF RF ALGORITHM

RF related theories have been introduced above. The parameters to be determined in this RF feature selection are the dimension of subspace and the number of decision trees. According to the suggestion [33], in this study, the subspace dimension is 8, and the number of decision trees is set to

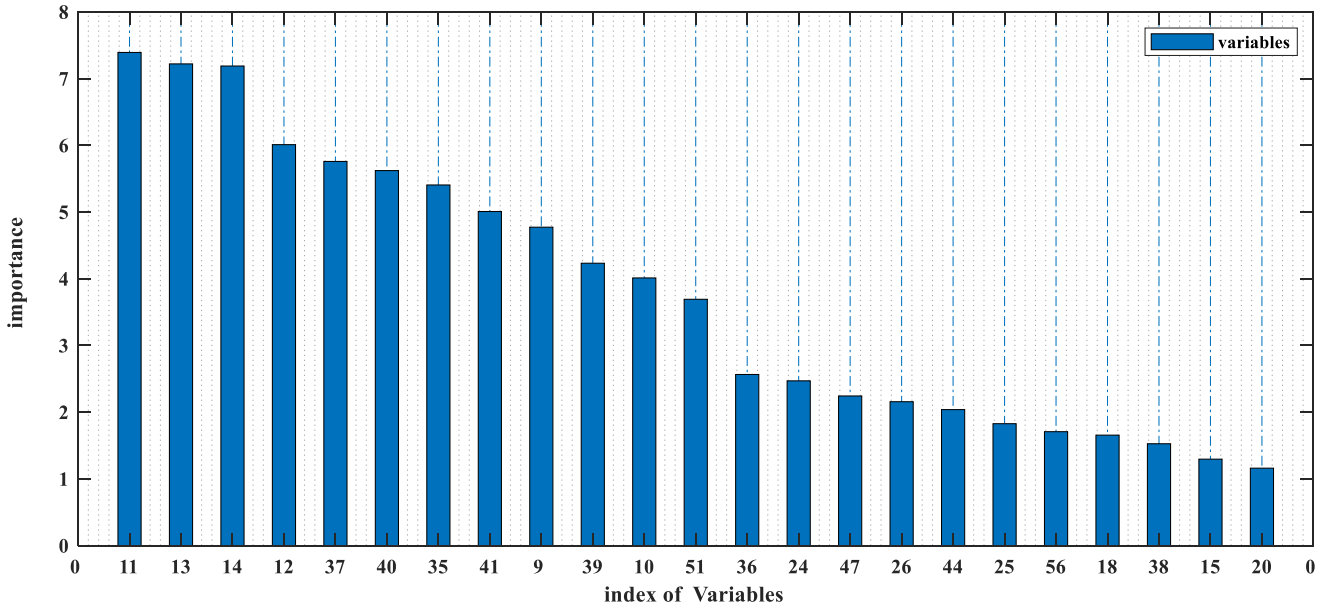

FIGURE 9. Top 23 variables selected by random forest of I-880 highway.

30. The ranking of the top 30 importance variables on I-205 highway is [14, 24, 4, 19, 6, 7, 11, 28, 22, 40, 43, 31, 25, 15, 36, 48, 38, 55, 3, 2, 10, 9, 37, 41, 8, 54, 51, 56, 27, 23] and their importance percentages are [8.7, 7.4, 6.6, 6.5, 5.2, 4.8, 4.3, 4.2, 4.2, 3.9, 2.6, 2.4, 1.8, 1.8, 1.7, 1.7, 1.5, 1.5, 1.4, 1.4, 1.3, 1.2, 1.2, 1.2, 1.1, 1.1, 1.0, 1.0, 0.9]. Factor analysis and principal component analysis are also commonly used feature dimension reduction methods. They choose factors with cumulative variance greater than 85% as reasonable factors to represent all variables. Here, the cumulative importance of our top 30 variables is 89.98%, which meets the conditions of more than 85%. The total number of initial variables is 57, and the variables are reduced by nearly half after feature selection, which obviously contributes to improving the efficiency of model detection and does not lose data characteristics. Similarly, according to the above variable selection rules, we selected 23 variables for I-880 highway, and the cumulative importance is 86 %. FIGURE 8 and FIGURE 9 intuitively show the importance of variables. The abscissa is the index of variables, and the ordinate is the importance in percentage form.

3) HORIZONTAL CONTRAST RESULTS

FIGURE 10 shows the confusion matrix of the horizontal comparison algorithm of I-205 highway. The TP, TN, FP, FN values are clearly displayed, and the error prediction results are distinguished by different colors, and the larger the value, the deeper the color. 1 represents the incident sample and 2 represents the normal sample. For incident detection, we always want to get the highest DR, ACC and the lowest FAR. TP and TN are as large as possible, while FP and FN are best 0. It can be seen from the FIGURE 10 that the proposed SASYNO-RF-RSKNN model has the highest TP and TN values compared with other methods, and FN and FP are also the lowest in the method of using SASYNO technology. For SMOTE-RF-RSKNN, although its FN is 0, its FP value is the highest in all algorithms, which means that it will produce more error warning information. And it does not contribute to traffic managers to monitor traffic conditions. This also shows that the classical SMOTE over-sampling technology only considers the incident sample data to generate interpolation as a new sample, so that the sample database does not have complete information characteristics, which leads to limited learning ability of the classifier. SASYNO improves this problem by combining interpolation and extrapolation. With the same SASYNO technology, the FP values of SVM, EigenClass and RSKNN are the same, but the FN of SVM comes to 11, which indicates that SVM is not sensitive to incident samples and cannot well distinguish incident samples from normal samples. The DR is low, which should be caused by the inadaptability of SVM to high-dimensional variables. EigenClass and RSKNN are sensitive to incident samples and can well learn the information characteristics of high-dimensional variables, so as to distinguish incident samples from normal samples.

FIGURE 11 shows the confusion matrix of the horizontal comparison algorithm of I-880 highway. Similar to I-205 highway, it can be seen that the proposed SASYNO-RF-RSKNN model has the highest TP and TN values compared with other methods, and FN and FP are also the lowest. Since the detector spacing of I-880 dataset is more intensive, it can better reflect the traffic flow changes caused by incidents. Therefore, on the whole, the confusion matrix of the four methods of I-880 is significantly better than that of I-205. Nevertheless, it can be seen from FIGURE 11 (a) and (b) that the FP values of SMOTE and SASYNO are 0 on I-880,

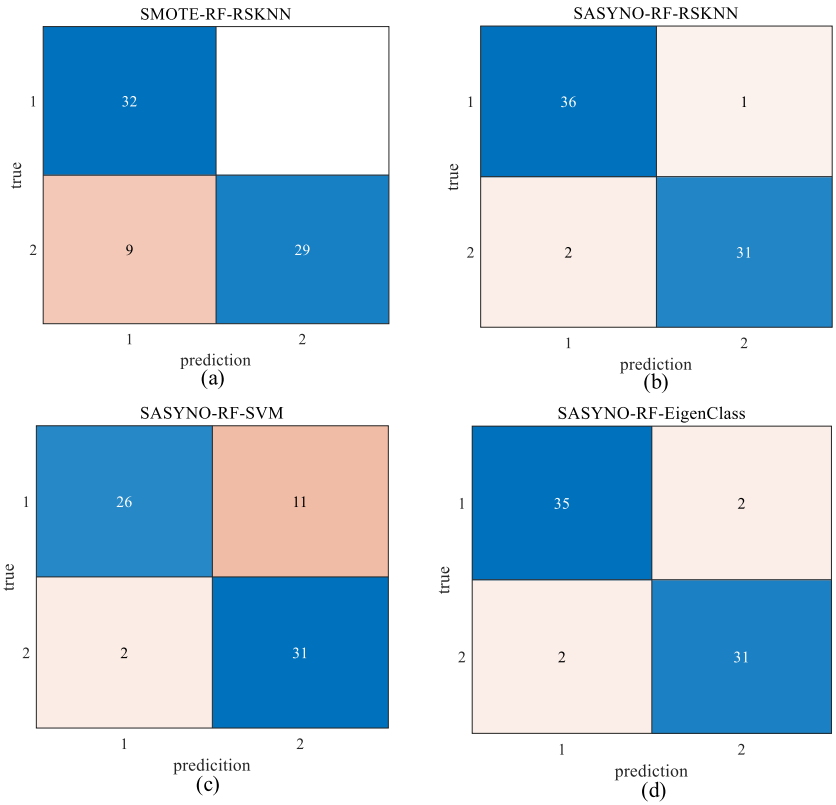

FIGURE 10. The confusion matrix of horizontal contrast method of I-205 highway, (a) SMOTE-RF-RSKNN, (b) SASYNO-RF-RSKNN, (c) SASYNO-RF-SVM, (d) SASYNO-RF-EigenClass.

indicating that the generated new samples have good effect and no error information. But the FN value of SASYNO is lower, so it has better detection effect. It can be seen from FIGURE 11 (b), (c) and (d) that with the same SASYNO and RF algorithms, RSKNN is superior to SVM in all aspects, and slightly inferior to EigenClass in detection rate. The confusion matrix shows that our incident detection model has the best performance, followed by EigenClass method.

TABLE 8 and TABLE 9 show the six performance index (ACC, FAR, DR, Precision, MCC and F1-score) of the horizontal comparison algorithm of I-205 highway and I-880 highway. It can be seen that the proposed method is the best in the five indicators except DR, and DR is also in the second place. On I-205 highway, Similar to the results expressed by the confusion matrix, SMOTE algorithm obtains 100% DR and its FAR reaches the maximum value in all algorithms, which is much higher than that of other algorithms. SASYON reduced the FAR by 17% by sacrificing the DR of 3%, and its performance was better. SVM, EigenClass and RSKNN have little difference in Precision and FAR.

However, there is a 15%-27% gap between the DR and ACC of SVM and the other two methods, indicating that the proposed method is a comprehensive method. On the basis of ensuring accuracy and precision, there is a lower FAR and a higher DR. MCC and F1-score are comprehensive indicators that balance DR and Precision. It can be seen from forward four indicators that SASYNO-RF-RSKNN and SASYNO-RF-EigenClass are the best algorithms with little difference. The MCC and F1-score values of RSKNN algorithm are close to 1, which indicates that it has a performance close to the best classification. At the same time, MCC of RSKNN is about 3% ahead of EigenClass, which indicates that the performance of the two is similar, and RSKNN is slightly better.

On I-880 highway, the performances of SMOTE and SASYNO are different from those of I-205 highway. The precision and FAR of SMOTE algorithm are consistent with those of SASYNO algorithm, but the comprehensive performances of MCC and F1-score are inferior to those of SASYNO algorithm. Similarly, the performance of EigenClass algorithm and SVM algorithm on I-880 highway has changed, but ACC, MCC and F1-score of RSKNN algorithm are optimal on both data sets. It follows that the robustness of SASYNO-RF-RSKNN algorithm for data sets is verified, which is an efficient and stable machine learning algorithm.

4) VERTICAL CONTRAST RESULTS

Similar to the above part, we also draw the confusion matrix in the vertical comparison part, as shown in FIGURE 12 and 13. It is worth noting that different methods set up the sample database according to the requirements of the literature, resulting in inconsistent confusion matrix size. Absolute values are not comparable.

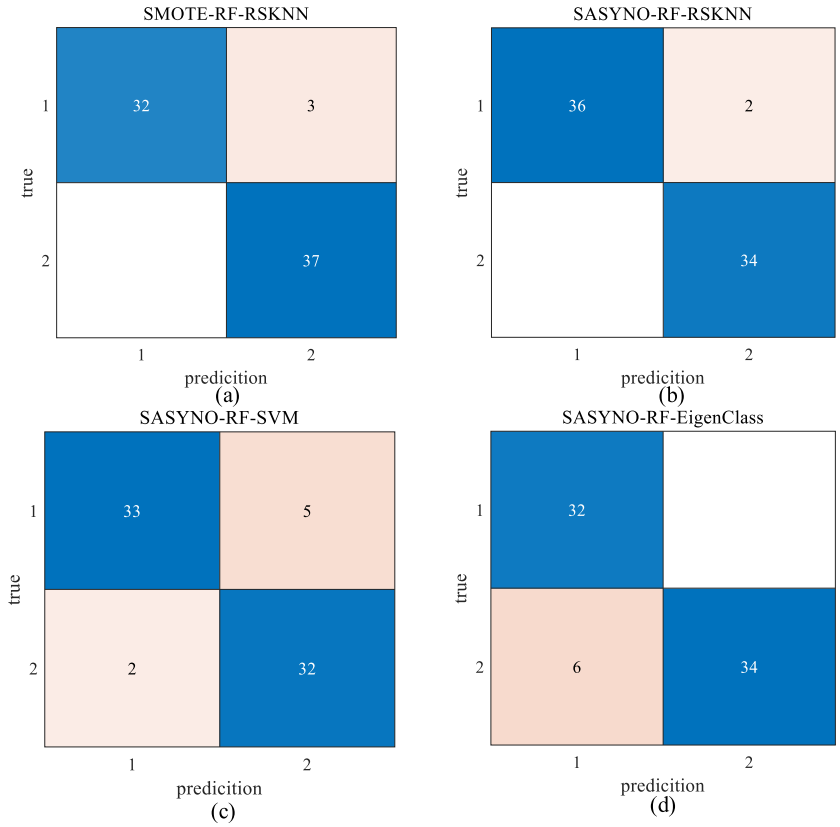

FIGURE 11. The confusion matrix of horizontal contrast method of I-880 highway, (a) SMOTE-RF-RSKNN, (b) SASYNO-RF-RSKNN, (c) SASYNO-RF-SVM, (d) SASYNO-RF-EigenClass.

TABLE 8. Indexes of Horizontal contrast of I-205 highway.

Methods	Precision	FAR	DR	ACC	MCC	F1-score
SASYNO-RF-RSKNN	**0.947**	**0.061**	0.973	**0.957**	**0.914**	**0.960**
SMOTE-RF-RSKNN	0.780	0.237	**1.000**	0.870	0.772	0.877
SASYNO-RF-SVM	0.928	**0.061**	0.703	0.814	0.654	0.800
SASYNO-RF-EigenClass	0.946	**0.061**	0.946	0.943	0.885	0.946

TABLE 9. Indexes of Horizontal contrast of I-880 highway.

Methods	Precision	FAR	DR	ACC	MCC	F1-score
SASYNO-RF-RSKNN	**1**	**0**	0.947	**0.972**	**0.945**	**0.9723**
SMOTE-RF-RSKNN	**1**	**0**	0.914	0.958	0.919	0.955
SASYNO-RF-SVM	0.942	0.058	0.868	0.902	0.808	0.904
SASYNO-RF-EigenClass	0.842	0.15	**1**	0.916	0.846	0.914

Among all the algorithms excluding FA-WFR, the sum of TP and TN values of the SASYNO-RF-RSKNN model is the largest, and the same sum of FP and FN is the smallest. E-SVM-KNN and SASYNO-RF-RSKNN all are ensemble algorithms. The latter selects 15 random variables from 30 variables to construct 30 different learners, which

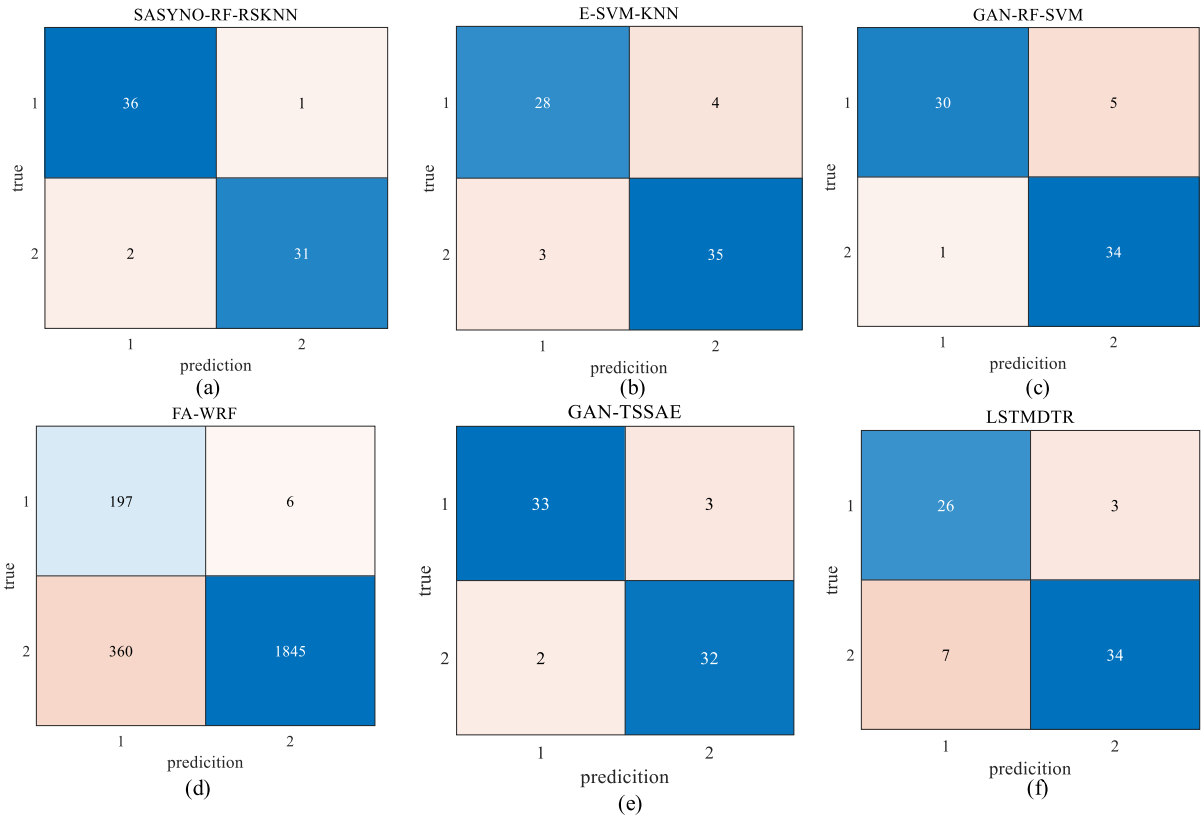

FIGURE 12. The confusion matrix of Vertical contrast method of I-205 highway, (a) SASYNO-RF-RSKNN, (b) E-SVM-KNN, (c) GAN-RF-SVM, (d) FA-WRF, (e) GAN-TSSAE, (f) LSTMDTR.

fully excavates the information characteristics between multivariate data and is superior to E-SVN-KNN method in terms of confusion matrix results. GAN-RF-SVM is a kind of deep learning method, which uses GAN to generate samples and SVM to classify. Due to the characteristics of GAN, it can produce new samples that cannot be imitated according to random noise. It can be said that its samples contain more information characteristics in it, but it will also produce more interference information, which has a high requirement for the construction of deep learning network. From the results of the confusion matrix, it can be seen that compared with the other methods, its FN value is the highest, which also confirms the analysis just now. It will lead to the insensitivity of the model to the incident samples and lower detection rate. SASYNO algorithm has obvious advantages in this respect.

It can be seen from FIGURE 12 and FIGURE 13 that the detection rate and accuracy of the two deep learning algorithms GAN-TSSAE and LSTMDTR on the I-205 highway are lower than those of RSKNN. At the same time, on the I-880 highway, because the overall effect of the data set is good, the overall performance of all algorithms is significantly improved, and the improvement of RSKNN with significant effect on I-205 highway is relatively small. So, on I-880 highway, FP of RSKNN is lower than GAN-TSSAE and LSTMDTR, FN is higher than both. This also shows that in most cases the false alarm rate and detection rate are a pair of contradictory values.

TABLE 10 and TABLE 11 also show the six performance indexes (ACC, FAR, DR, Precision, MCC and F1-score) of the vertical comparison algorithm. On I-205 highway, except for Precision and FAR, the other indexes of our algorithm are also the best. GAN-RF-SVM has good performance on Precision and FAR, which is better than SASYNO-RF-RSKNN 2% and 3.2% respectively, but its DR and ACC are lower than that of SASYNO-RF-RSKNN 12% and 4%.

In the aspect of incident detection, FAR and DR are a pair of contradictory values. The high DR often means that the FAR is also very high. SASYNO-RF-RSKNN has achieved a 12% increase in DR by increasing FAR by 3.2%. This conversion efficiency is very wonderful, and the FAR is also controlled at about 5%. The remaining four methods are ensemble learning algorithms and deep learning algorithms, in which E-SVM-KNN has an average performance and FA-WRF algorithm does not use over-sampling and under-sampling techniques for imbalanced data sets. Instead, it improves the algorithm and proposes a weighting strategy to ensure the algorithm performance. Its DR is 97%, but other Precision, FAR and ACC are the lowest. However, its ACC and MCC are the highest on the I-880 highway, and the classification effect can also be obtained due to the more obvious characteristics of the I-880 highway. It can

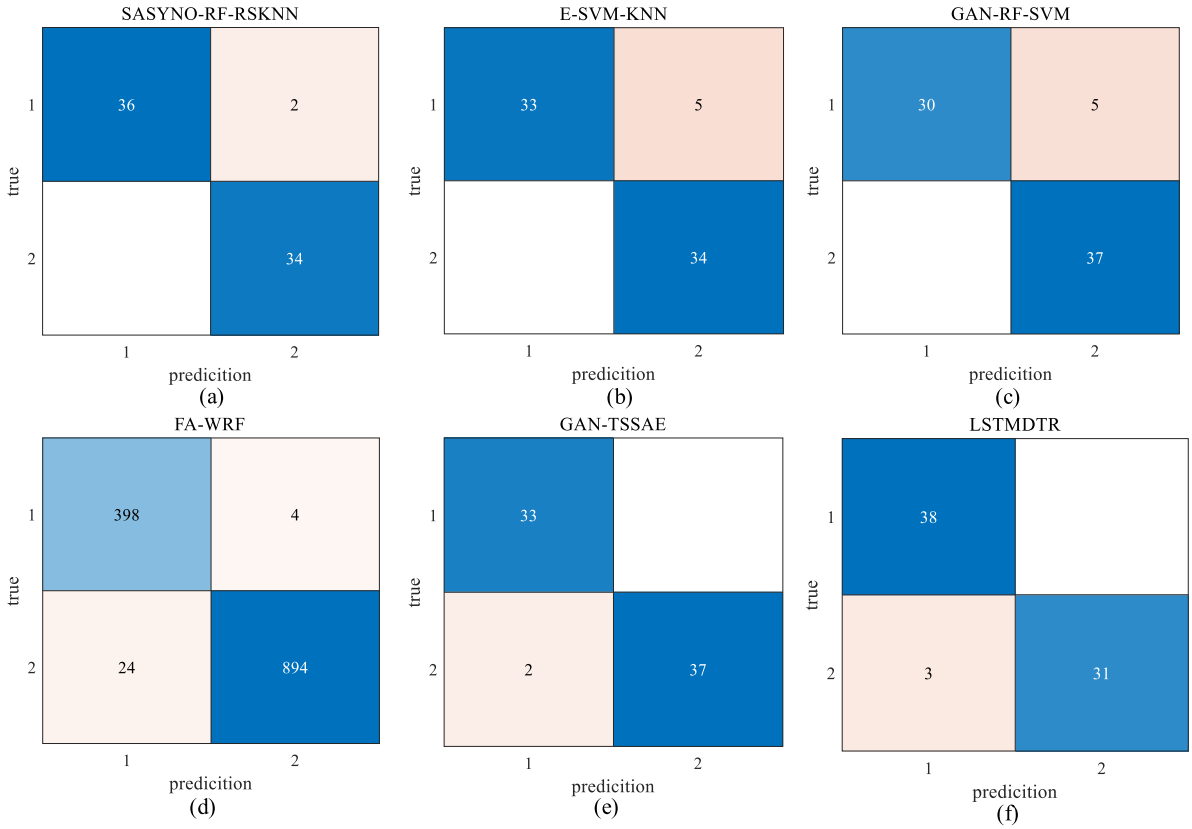

FIGURE 13. The confusion matrix of Vertical contrast method of I-880 highway, (a) SASYNO-RF-RSKNN, (b) E-SVM-KNN, (c) GAN-RF-SVM, (d) FA-WRF, (e) GAN-TSSAE, (f) LSTMDTR.

TABLE 10. Indexes of Vertical contrast of I-205 highway.

Methods	Precision	FAR	DR	ACC	MCC	F1-score
SASYNO-RF-RSKNN	0.947	0.061	**0.973**	**0.957**	**0.914**	**0.960**
E-SVM-KNN	0.903	0.079	0.875	0.900	0.798	0.889
GAN-RF-SVM	**0.967**	**0.029**	0.857	0.914	0.834	0.910
FA-WRF	0.350	0.163	0.970	0.848	0.531	0.518
GAN-TSSAE	0.942	0.058	0.916	0.928	0.857	0.929
LSTMDTR	0.787	0.170	0.896	0.857	0.716	0.838

be said that compared with the over-sampling calculation such as SASYNO, the strategy robustness of the improved algorithm for unbalanced data is poor, and it cannot be well applied to other data sets. Similarly, the proposed algorithm also has a great lead in comprehensive indicators such as MCC and F1-score, which are 6% and 3% higher than the second.

In addition to the FAR of GAN-TSSAE is better than SASYNO-RF-RSKNN on I-205 highway, the other indicators of GAN-TSSAE and LSTMDTR are not as good as our proposed algorithms. LSTMDTR and GAN-TSSAE achieved 100% detection rate on I-880 highway. ACC, MCC and F1-score of GAN-TSSAE are basically the same as SASYNO-RF-RSKNN, and FAR and Precision are poor.

Therefore, SASYNO-RF-RSKNN model has the best classification performance under the condition of small imbalance sample data. It is reasonable to use the model framework of SASYNO oversampling, RF based feature selection and RSKNN based classification. Combined with the horizontal comparison results in the previous section, it can be said that the practicability and robustness of the proposed method are well verified.

TABLE 11. Indexes of Vertical contrast of I-880 highway.

Methods	Precision	FAR	DR	ACC	MCC	F1-score
SASYNO-RF-RSKNN	**1**	**0**	0.947	0.972	0.945	**0.972**
E-SVM-KNN	**1**	**0**	0.868	0.930	0.870	0.929
GAN-RF-SVM	**1**	**0**	0.857	0.930	0.868	0.923
FA-WRF	0.943	0.026	0.990	**0.978**	**0.951**	0.966
GAN-TSSAE	0.942	0.051	**1**	0.972	0.945	0.970
LSTMDTR	0.926	0.088	**1**	0.958	0.919	0.962

IV. CONCLUSION

AID has attracted much attention in the field of transportation in recent decades. However, the number of traffic incident samples is small and the number of incident samples and non-incident samples is extremely imbalanced, which often has a negative impact on the accuracy of the incident detection model. From this perspective, a method to address data imbalance under small sample conditions and incomplete initial variables problem was proposed in this study. It is a novel incident detection framework named SASYNO-RF-RSKNN. Firstly, a relatively complete set of initial variables is constructed by using the spatial-temporal and real-time characteristics of traffic stream. Second, the sample database is balanced to solve small and imbalanced sample size by SASYNO oversampling technology, and then RF is used for feature selection. Finally, the ensemble learning algorithm RSKNN is used to identify traffic incident. To validate the proposed approach, two real-world traffic data on the I-205 highway and I-880 highway are used for empirical analysis. In addition, six excellent machine learning algorithms are evaluated by six indexes: ACC, FAR, DR, Precision, MCC and F1-score. In the horizontal comparison part, the proposed algorithm is the best except DR. In the vertical comparison, precision, FAR and F1-score are also the best. It shows that SASYNO-RF-RSKNN algorithm is an excellent AID framework.

Notably, this study is an ensemble learning algorithm based on small-scale datasets. Deep learning is more effective framework for prediction and classification related tasks, especially for large-scale datasets. Therefore, it is also expected that deep learning algorithm does not achieve the best effect in performance evaluation. In the future, it can be considered to use deep learning algorithms such as LSTM, CNN and other deep learning networks for traffic automatic incident detection on large-scale data sets.

REFERENCES

[1] H. Payne and S. Tignor, "Freeway incident-detection algorithms based on decision trees with states," *Transp. Res. Rec.*, vol. 682, pp. 30–37, Jan. 1978.

[2] A. Karim and H. Adeli, "Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm," *J. Transp. Eng.*, vol. 128, no. 1, pp. 21–30, Jan. 2002.

[3] B. N. Persaud, F. L. Hall, and L. M. Hall, "Congestion identification aspects of the McMaster incident detection algorithm," *Transp. Res. Rec.*, no. 1287, pp. 167–175, 1990.

[4] Y. Cheng, M. Zhang, and D. Yang, "Automatic incident detection for urban expressways based on segment traffic flow density," *J. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 205–213, Jan. 2015.

[5] C. L. Dudek, C. J. Messer, and N. B. Nuckles, "Incident detection on urban freeways," *Transp. Res. Rec.*, vol. 495, no. 495, pp. 12–24, 1974.

[6] A. R. Cook and D. E. Cleveland, "Detection of freeway capacity reducing incidents by traffic-stream measurements," *Transp. Res. Rec.*, vol. 495, pp. 1–11, Jan. 1974.

[7] Y. J. Stephanedes and A. P. Chassiakos, "Application of filtering techniques for incident detection," *J. Transp. Eng.*, vol. 119, no. 1, pp. 13–26, Jan. 1993.

[8] W. Wang, S. Chen, and G. Qu, "Incident detection algorithm based on partial least squares regression," *Transp. Res. C, Emerg. Technol.*, vol. 16, no. 1, pp. 54–70, Jun. 2008.

[9] A. Kinoshita, A. Takasu, and J. Adachi, "Real-time traffic incident detection using a probabilistic topic model," *Inf. Syst.*, vol. 54, pp. 169–188, Dec. 2015.

[10] S. G. Ritchie and R. L. Cheu, "Simulation of freeway incident detection using artificial neural networks," *Transp. Res. C, Emerg. Technol.*, vol. 1, no. 3, pp. 203–217, Sep. 1993.

[11] R. L. Cheu and S. G. Ritchie, "Automated detection of lane-blocking freeway incidents using artificial neural networks," *Transp. Res. C, Emerg. Technol.*, vol. 3, no. 6, pp. 371–388, Dec. 1995.

[12] H. Dia and G. Rose, "Development and evaluation of neural network freeway incident detection models using field data," *Transp. Res. C, Emerg. Technol.*, vol. 5, no. 5, pp. 313–331, Oct. 1997.

[13] S. Ishak and H. Al-Deek, "Performance of automatic ANN-based incident detection on freeways," *J. Transp. Eng.*, vol. 125, no. 4, pp. 281–290, Aug. 1999.

[14] D. Srinivasan, X. Jin, and R. L. Cheu, "Adaptive neural network models for automatic incident detection on freeways," *Neurocomputing*, vol. 64, pp. 473–496, Mar. 2005.

[15] J. Lu, S. Chen, W. Wang, and H. van Zuylen, "A hybrid model of partial least squares and neural network for traffic incident detection," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 4775–4784, Apr. 2012.

[16] F. Yuan and R. L. Cheu, "Incident detection using support vector machines," *Transp. Res. C, Emerg. Technol.*, vol. 11, nos. 3–4, pp. 309–328, Jun. 2003.

[17] K. Zhang and M. A. P. Taylor, "Effective arterial road incident detection: A Bayesian network based algorithm," *Transp. Res. C, Emerg. Technol.*, vol. 14, no. 6, pp. 403–417, Dec. 2006.

[18] Q. Liu, J. Lu, S. Chen, and K. Zhao, "Multiple Naïve Bayes classifiers ensemble for traffic incident detection," *Math. Problems Eng.*, vol. 2014, pp. 1–16, Apr. 2014.

[19] S. Chen and W. Wang, "Decision tree learning for freeway automatic incident detection," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 4101–4105, Mar. 2009.

[20] J. Wang, X. Li, S. S. Liao, and Z. Hua, "A hybrid approach for automatic incident detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1176–1185, Sep. 2013.
[21] Y. Lin, L. Li, H. Jing, B. Ran, and D. Sun, "Automated traffic incident detection with a smaller dataset based on generative adversarial networks," *Accident Anal. Prevention*, vol. 144, Sep. 2020, Art. no. 105628.
[22] H. Jiang and H. Deng, "Traffic incident detection method based on factor analysis and weighted random forest," *IEEE Access*, vol. 8, pp. 168394–168404, 2020.
[23] Q. Shang, L. Feng, and S. Gao, "A hybrid method for traffic incident detection using random forest-recursive feature elimination and long short-term memory network with Bayesian optimization algorithm," *IEEE Access*, vol. 9, pp. 1219–1232, 2021.
[24] L. Li, Y. Lin, B. Du, F. Yang, and B. Ran, "Real-time traffic incident detection based on a hybrid deep learning model," *Transportmetrica A*, 2020, doi: 10.1080/23249935.2020.1813214.
[25] F. Jiang, K. K. R. Yuen, and E. W. M. Lee, "A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions," *Accident Anal. Prevention*, vol. 141, Jun. 2020, Art. no. 105520.
[26] C. Zhili, J. Guiyan, and D. Qiushi, "Study on automated incident detection algorithms based on multi-SVM classifier," in *Proc. Chin. Control Decis. Conf.*, Yantai, China, Jul. 2008, pp. 1358–1362.
[27] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Phys. A, Stat. Mech. Appl.*, vol. 517, pp. 29–35, Mar. 2019.
[28] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[30] X. Gu, P. P. Angelov, and E. A. Soares, "A self-adaptive synthetic over-sampling technique for imbalanced classification," *Int. J. Intell. Syst.*, vol. 35, no. 6, pp. 923–943, Jun. 2020.
[31] J. Xiao, X. Gao, Q.-J. Kong, and Y. Liu, "More robust and better: A multiple kernel support vector machine ensemble approach for traffic incident detection," *J. Adv. Transp.*, vol. 48, no. 7, pp. 858–875, May 2013.
[32] U. Erkan, "A precise and stable machine learning algorithm: Eigenvalue classification (EigenClass)," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 5381–5392, Sep. 2020.
[33] Y. Liu, J.-W. Bi, and Z.-P. Fan, "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms," *Expert Syst. Appl.*, vol. 80, pp. 323–339, Sep. 2017.

TIAN XIE received the bachelor's degree from the Transportation College, Changsha University of Science and Technology, Changsha, China, in 2018. He is currently pursuing the Ph.D. degree with the Shandong University of Technology, Zibo, China. His research interests include traffic incident detection and intelligent transportation systems.

QIANG SHANG received the Ph.D. degree from the Transportation College, Jilin University, Changchun, China, in 2017. He is currently a Lecturer with the Shandong University of Technology, Zibo, China. He has authored over 20 academic articles in journals. His research interests include traffic data analysis, traffic model, and intelligent transportation systems.

YANG YU received the bachelor's degree from the Transportation College, Shandong University of Technology, Zibo, China, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include short-term traffic flow prediction and traffic data analysis.