Risk Level Identification of Train Dispatch Safety Based on CGAN-KPCA-BO-LSSVM

Hao Peng, Xuelei Meng, Siyu Cao, Doudou Wang , and Tianshu Qi

Abstract—**In train dispatching, risk identification plays a crucial role in ensuring both system safety and operational efficiency. This study addresses two key issues: the imbalance in risk level data and the excessive number of risk factors influencing train dispatching, both of which lead to suboptimal risk prediction performance. To tackle these challenges, a novel risk identification model for train dispatching is proposed. The identified risk factors are classified into four main categories: human factors, equipment factors, train factors, and environmental factors. A Conditional Generative Adversarial Network (CGAN) is employed to generate samples for underrepresented risk levels, mitigating data imbalance. Subsequently, Kernel Principal Component Analysis (KPCA) is employed for feature dimensionality reduction. Finally, this study employs the Least Squares Support Vector Machine (LSSVM) algorithm combined with Bayesian optimization (BO-LSSVM) for risk level identification. The case study shows that the proposed CGAN-KPCA-BO-LSSVM model improves identification accuracy by 8.54% compared to the standard LSSVM algorithm.**

Index Terms—**train dispatching; risk identification; CGAN; kernel principal component analysis; BO-LSSVM algorithm**

I. Introduction

IN modern railway transportation systems, the safety and efficiency of the dispatching process are paramount. Railway transportation dispatching is a complex process that involves the coordinated interaction of multiple factors, including human operations, equipment conditions, and environmental factors. Any failure or malfunction in any of these aspects may lead to serious consequences, such as train delays, traffic accidents, or even casualties. Among the various components of the railway dispatching system, train dispatching is particularly critical, as it directly influences the safe and efficient operation of trains. However, the dispatching process is susceptible to risks arising from equipment failures, human errors, and dynamic environmental changes[1]. These risks not only threaten the safety of train operations but also have the potential to trigger significant social and economic repercussions. Consequently, the systematic identification and assessment of risks in the train dispatching process are essential to ensure the reliability and sustainability of railway transportation systems.

As the importance of risk management in railway systems continues to grow, numerous methods have been applied to risk identification research in railway transportation[2]. Traditional approaches, such as Fault Tree Analysis (FTA) and Event Tree Analysis (ETA), largely rely on statistical analysis and expert judgment, focusing on the classification and assessment of risk factors. Bohus Leitner developed a risk assessment model for railway systems, evaluating the frequency of hazardous events based on historical accident data and employing safety techniques such as fault tree analysis and event tree analysis for structured expert judgment[3]. Jun Lai et al. proposed a fault probability assessment method that integrates Integrated Fault Tree and Fault Event Tree Analysis (IFFTA) with Networked Bayesian Networks (NGBN) to quantify derailment risk in rail transport (RT). This approach provides a comprehensive evaluation of derailment risk and effectively identifies key contributing factors[4]. However, as the complexity of railway systems increases, these traditional methods face limitations in addressing nonlinear relationships and multi-factor risks. In particular, when dealing with large-scale data and complex environments, traditional models often struggle to accurately capture the intricate relationships between risk factors. To overcome these challenges, some researchers have turned to artificial intelligence (AI) technologies for risk identification[5]. Hui Peng Liu et al. proposed a novel aviation safety risk level identification model, employing Generative Adversarial Networks (GANs) to generate underrepresented class samples from ASRS (Aviation Safety Reporting System) data. The integration of Bayesian optimization algorithms for hyperparameter tuning further improved the model's performance in risk identification[6]. Jicheng Liu et al. proposed a risk assessment model based on the Kernel Principal Component Analysis-Tunicate Swarm

Manuscript received November 28, 2024; revised May 6, 2025.

This work was supported by the Science and Technology Program of Gansu Province (Project No. 24JRRA865) and the Gansu Province Central Government-Guided Local Science and Technology Development Fund Project (Project No. 25ZYJA015).

Hao Peng is a postgraduate student at School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China. (e-mail: 2356838062@qq.com).

Xuelei Meng is a Professor at School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China. (corresponding author , e-mail: mxl@mail.lzjtu.cn).

Siyu Cao is a postgraduate student at School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China. (email: 1509382720@qq.com).

Doudou Wang is a postgraduate student at School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China. (email: 3215600544@qq.com).

Tianshu Qi is a postgraduate student at School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China. (email: 1224511046@qq.com).

Optimization-Least Squares Support Vector Machine (KPCA-TSO-LSSVM) algorithm. This model uses KPCA for dimensionality reduction of the original data, thereby improving the classification accuracy of the model[7]. Keyang Liu et al. developed an automated risk identification method combining machine learning, deep learning, and natural language processing technologies, utilizing identified risk factors for data-driven risk assessment[8]. In railway transportation, Wencheng Huang et al. employed enhanced Support Vector Machine (SVM) methods, including Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Gradient Search (GS), to identify risks in the hazardous goods transportation system, with the assessment scores of each sub-risk factor represented in interval numbers[9]. Sadiq Khan et al. used a semi-quantitative risk matrix method, combining historical accident data and expert experience, to perform a comprehensive risk assessment of Pakistan Railways[10]. Yujie Huang developed a new causal modeling framework for studying railway intrusion risks, based on text mining technology and fuzzy rule modeling[11]. Chang Liu et al. constructed a hazard and accident knowledge graph using text mining technology, applying it to railway hazard identification and risk assessment[12]. Keping Li et al. proposed a complex network-based risk monitoring model for identifying accident causal factors and analyzing their interactions[13]. Jun Liu et al. proposed an AI-driven railway control and scheduling system, utilizing large model technologies with advanced learning capabilities, efficient associative abilities, and linkage analysis features[14]. Although these studies have made significant strides in railway transportation and risk identification, research specifically focused on risk identification in railway dispatching, particularly concerning train dispatch risks, remains relatively limited.

Based on this foundation, this paper proposes a systematic method for identifying risk levels in train dispatching. The method categorizes risk factors into four groups: personnel, equipment, train, and environmental factors. To address data-related challenges, the proposed approach employs CGANs for data augmentation and KPCA for feature selection and dimensionality reduction. These preprocessing steps enhance the quality and reliability of the dataset, thereby improving the accuracy of risk assessment. Furthermore, the paper introduces an optimized Least Squares Support Vector Machine (LSSVM) algorithm for risk level identification. To optimize model performance, Bayesian optimization is employed to fine-tune the hyperparameters of the LSSVM model. This integrated approach not only enhances the accuracy of risk identification but also strengthens the robustness and generalizability of the proposed method.

II. Analysis of Factors for Identifying Safety Risks in Train dispatching

A. Composition of risk factors

The train dispatching process is a complex and crucial component influenced by numerous factors. Several potential risk factors can affect the normal operation of train dispatching, thereby impacting the safety, efficiency, and reliability of the railway system [15][16]. To accurately identify the influence of various risk factors on accidents during railway traffic scheduling, this paper establishes a set of potential risk factors based on historical data analysis. In general, the primary risk factors encountered during train dispatching can be classified into three broad categories: human, equipment, and environmental factors. However, due to the significant impact of train-related factors on train dispatching, this paper further divides the risk factors into four categories: personnel factors, equipment factors, train factors, and environmental factors. Each category is then subdivided into specific risk indicators, as shown in Table I.

B. Data preprocessing

This study analyzes historical risk events and daily train dispatch records, identifying key hazard factors that influence system safety based on statistical data. The data is then subjected to specific processing and preprocessing steps to ensure its compatibility with machine learning algorithms for effective risk identification.

The data is categorized into discrete and continuous types. To reduce the uncertainty and ambiguity of risk factors that may influence risk prediction, continuous data is represented as intervals. For integer-valued data, it is indicated that $x_{ijk}^{a_0} = x_{ijk}^{b_0}$. $x_{ijk}^0 = [x_{ijk}^{a_0}, x_{ijk}^{b_0}]$ represents the data value corresponding to the j-th risk in the i-th risk element of the k-th incident, where $x_{ijk}^{a_0}$ and $x_{ijk}^{b_0}$ denote the lower and upper bounds of the interval data, respectively, satisfying $x_{ijk}^{a_0} \le x_{ijk}^0 \le x_{ijk}^{b_0}$. First, the obtained data is normalized. Let x_{ijk}^0 represent the original data, and let x_{ijk}^a , x_{ijk}^b denote the lower and upper bounds of the data after normalization. Here, $x_{ijk}^{\min}$ represents the minimum data value corresponding to the j-th risk within the i-th risk element of the k-th incident, while $x_{ijk}^{\max}$ represents the maximum data value for the same risk element. The normalization method for continuous data is as follows:

$$x_{ijk}^a = \frac{x_{ijk}^{a_0} - x_{ijk}^{a_{\min}}}{x_{ijk}^{a_{\max}} - x_{ijk}^{a_{\min}}} \tag{1}$$

$$x_{ijk}^b = \frac{x_{ijk}^{b_0} - x_{ijk}^{b_{\min}}}{x_{ijk}^{b_{\max}} - x_{ijk}^{b_{\min}}} \tag{2}$$

Finally, the normalized risk data is obtained as $x_{ijk}^c = [x_{ijk}^a, x_{ijk}^b]$, where $0 \le x_{ijk}^a \le 1$, $0 \le x_{ijk}^b \le 1$.

Discrete data refers to values that can only take specific, fixed categories. For instance, in this study, train type is classified into three distinct categories: passenger train, freight train, and mixed passenger-freight train. The normalization method for discrete data is defined as follows:

$$x_{ijk}^d = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

Where x represents the original data; $x_{\max}$ and $x_{\min}$ denote the corresponding maximum and minimum data values, respectively; x_{ijk}^d represents the normalized result of discrete data.

$x_{ijk}^c = [x_{ijk}^a, x_{ijk}^b]$ represents the normalized continuous risk assessment data. By solving the following equation, the actual assessment data for each risk factor can be derived:

$$\overline{x}_{ijk}^{c} = \frac{x_{ijk}^{a} + x_{ijk}^{b}}{2} - \varepsilon \frac{x_{ijk}^{b} - x_{ijk}^{a}}{x_{ijk}^{a} + x_{ijk}^{b}} \quad (4)$$

Where $0 < \varepsilon \le \left[\frac{\left(x_{ijk}^{a} + x_{ijk}^{b} \right)^{2}}{2\left(x_{ijk}^{a} - x_{ijk}^{b} \right)} \right]_{\min}$, $\forall i, j, k$. ε is the penalty factor for decision-makers, $\frac{x_{ijk}^{a} + x_{ijk}^{b}}{2}$ represents the midpoint value of x_{ijk}^{c} [9].

After normalization, discrete data does not require further processing, and the actual evaluation data is obtained, where $\overline{x}_{ijk}^{d} = x_{ijk}^{d}$.

In summary, the normalized actual evaluation data is $\overline{x}_{ijk} = \left\{ \overline{x}_{ijk}^{c}, \overline{x}_{ijk}^{d} \right\}$.

TABLE I
STATISTICS OF RISK FACTOR INDICATORS

Risk factors	Sub-indicators	Specific risk factor details	Explanation	
H	n_{11}	Average years of experience of the dispatcher	The average years of experience of the dispatcher are retained to one decimal place.	Personnel factors
	n_{12}	Dispatcher's experience and skills	The dispatcher's experience and skills can be represented by work performance records, categorized into 5 levels, ranging from 1 to 5.	
	n_{13}	Statistics of train driver operational errors	The statistics of train driver operational errors can be obtained through driving record devices and records from the dispatch center.	
	n_{14}	Train driver emergency response ability	The data on train driver emergency response ability is expressed in interval form and can be assessed based on driving data, represented by score results.	
	n_{15}	Train driver years of experience	The train driver's years of experience are retained to one decimal place.	
M_1	n_{21}	Track condition	Track condition is assessed through track inspection data and sensor-recorded data, classified into five levels from 1 to 5.	Equipment factors
	n_{22}	Signal equipment condition	The status of signaling equipment is classified into two categories: normal and abnormal.	
	n_{23}	Mobile Signaling Equipment Failure Frequency	Mobile signaling equipment failure frequency represents the number of failures occurring in all mobile signaling equipment within a year.	
	n_{24}	Switch operating condition	Switch condition is assessed through switch inspection data and sensor-recorded data, classified into five levels from 1 to 5.	
M_2	n_{31}	Train type	This paper categorizes trains based on their transport types into passenger trains, freight trains, and mixed passenger-freight trains.	Train factors
	n_{32}	Train running speed	Running speed is expressed in interval form.	
	n_{33}	Train total weight	Different types of trains have different total weights, with the train total weight expressed in interval form.	
	n_{34}	Statistics of train failure occurrences	The number of train failures can directly reflect the presence of safety hazards.	
E	n_{41}	Weather condition	Including clear, cloudy, heavy rain, fog, snow, and other weather conditions.	natural environmental factors
	n_{42}	Temperature condition	Both excessively low and high temperatures can affect train operation.	
	n_{43}	Humidity condition	Excessively high humidity may affect the normal operation of railway equipment, potentially leading to safety hazards.	
	n_{44}	Visibility condition	Visibility data is expressed in interval form.	

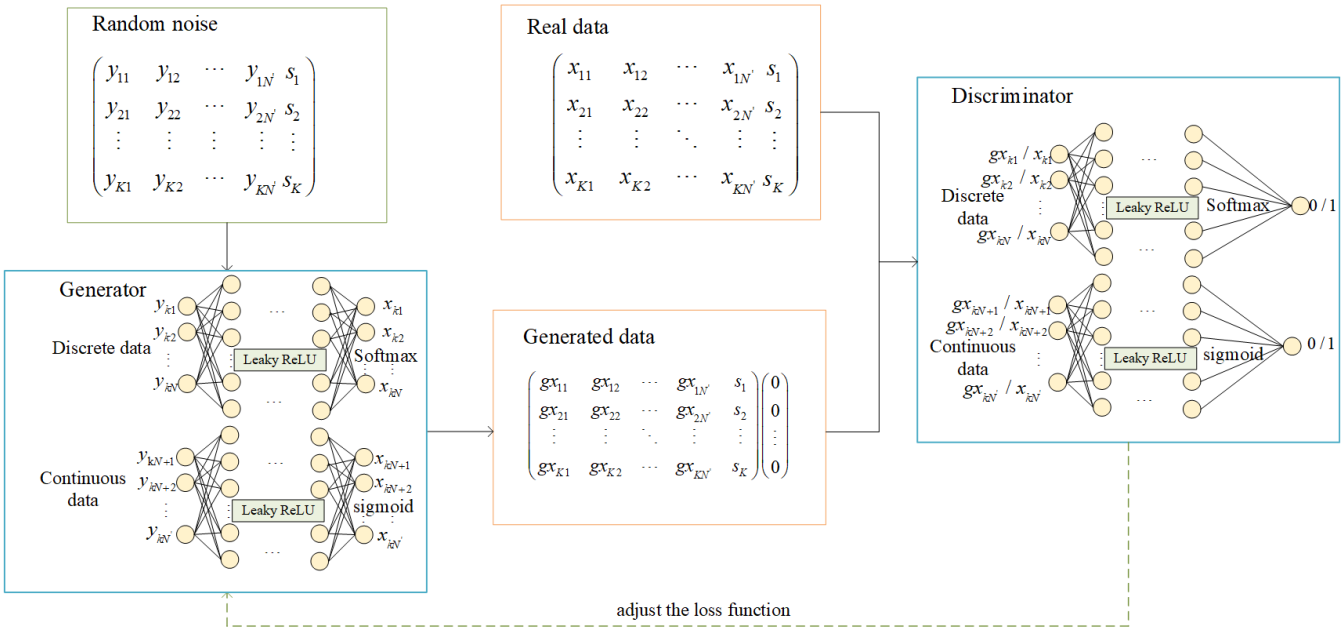

Fig.1. Implementation process of Conditional Generative Adversarial Networks (CGAN)

1) Handling imbalanced data

Through the analysis of historical accident data in railway transportation dispatching, it was found that high-risk data samples constitute a very small proportion of the dataset, leading to an imbalance. This imbalance may undermine the model's ability to accurately identify these high-risk samples during the risk identification process. To address this issue, the study employs CGANs to generate additional samples for the underrepresented classes, thereby improving data diversity and mitigating the risk of overfitting[17][18][19].

The CGAN is a deep learning model rooted in game theory. It extends the GAN by incorporating additional conditional information to guide the data generation process[20][21]. The model employs adversarial training, where a generator and a discriminator compete against each other to produce realistic data samples. The generator takes random noise as input and creates synthetic samples that mimic the distribution of real data. In contrast, the discriminator receives both real data and the synthetic samples produced by the generator, outputting a probability that indicates the likelihood that the input data is real.

In this study, the conditional variable is set as the risk level, and the objective function constructed is as follows:

$$\min_G \max_D V(D,G) = E_{\bar{x}_{ijk}\sim pdata(\bar{x}_{ijk})}\left[\log D\left(\bar{x}_{ijk},c\right)\right] + E_{b\sim p_b(\bar{x}_{ijk})}\left[\log\left(1-D\left(G\left(b,c\right),c\right)\right)\right] \tag{5}$$

Where $E_{\bar{x}_{ijk}\sim pdata(\bar{x}_{ijk})}\left[\log D\left(\bar{x}_{ijk},c\right)\right]$ is the prediction result of the discriminator for real data, $E_{b\sim p_b(\bar{x}_{ijk})}\left[\log\left(1-D\left(G\left(b,c\right),c\right)\right)\right]$ is the prediction result of the discriminator for generated data, $G\left(b,c\right)$ represents the generator, and $D\left(\bar{x}_{ijk},c\right)$ represents the discriminator, $\bar{x}_{ijk}$ represents preprocessed data, b represents the noise vector input to the generator, c is the conditional variable, representing the risk level s_k.

The primary role of the discriminator is to distinguish between real samples and generated samples. Guided by its loss function, the discriminator aims to maximize the accurate classification of real samples while minimizing the misclassification of generated ones. The corresponding loss function is defined as follows:

$$\Gamma_D = E_{\bar{x}_{ijk}\sim pdata(\bar{x}_{ijk})}\left[\log D\left(\bar{x}_{ijk}\mid c\right)\right] + E_{b\sim p_b(\bar{x}_{ijk})}\left[\log\left(1-D\left(G\left(b\mid c\right)\mid c\right)\right)\right] \tag{6}$$

Where Γ_D is the loss function of the discriminator, $\bar{x}_{ijk}$ represents the preprocessed data, and $pdata(\bar{x}_{ijk})$ is the distribution of $\bar{x}_{ijk}$ real data.

The generator's role is to generate realistic sample data with the goal of convincing the discriminator that these samples are authentic. Its loss function directs the generator to produce more realistic samples, and is expressed as follows:

$$\Gamma_G = E_{b\sim p_b(\bar{x}_{ijk})}\left[\log\left(1-D\left(G\left(b\mid c\right)\mid c\right)\right)\right] \tag{7}$$

Where Γ_G represents the loss function of the generator, $b\sim p_b(\bar{x}_{ijk})$ is a noise vector randomly sampled from the $p_b(\bar{x}_{ijk})$ noise distribution.

Through alternating training of the generator and discriminator, both models enhance their capabilities via continuous adversarial learning, ultimately leading to the generation of highly realistic sample data. The steps of the algorithm implementation are illustrated in Fig. 1, where N' represents the sum of the four types of risk factors H, M_1, M_2, E, with 1 to N representing discrete risk factors and $N+1$ to N' representing continuous risk factors.

2) Feature selection

Due to the numerous risk factors affecting the safety of train scheduling and the high dimensionality of feature data, there is a risk that excessively high dimensions could lead to increased model complexity and potential overfitting. To mitigate this risk, this paper employs Kernel Principal Component Analysis (KPCA) for dimensionality reduction,

thereby improving the model's generalization and predictive capabilities[22][23].

KPCA is a nonlinear extension of traditional PCA, which uses the kernel trick to map data into a higher-dimensional feature space before performing linear dimensionality reduction[24][25][26]. This allows it to effectively capture the nonlinear relationships within the data. The implementation of KPCA is as follows:

$$K(x_{k_1}, x_{k_2}) = \exp(-\frac{\left\| x_{k_1} - x_{k_2} \right\|^2}{2\sigma^2}) \tag{8}$$

Where $\left\| x_{k_1} - x_{k_2} \right\|$ represents the Euclidean distance between samples x_{k_1} and x_{k_2}, and σ denotes the parameter of the kernel function.

Next, a kernel matrix K is constructed between the samples, where each element $k(x_{k_1}, x_{k_2})$ represents the similarity between data points x_{k_1} and x_{k_2}.

To eliminate any data bias, the kernel matrix K needs to be centered. The centralized kernel matrix can be computed using the following formula:

$$K_{centered} = K - 1_n K - K 1_n + 1_n K 1_n \tag{9}$$

Where 1_n is an $n \times n$ matrix of ones, and this step ensures that the mean of each data point is zero, thereby eliminating any global bias.

The eigenvalue decomposition is performed on the centered kernel matrix $K_{centered}$, yielding eigenvalues λ_1, λ_2, $\cdots$, λ_n, and their corresponding eigenvectors $v_1, v_2, \cdots, v_n$, which satisfy the following eigen equation:

$$K_{centered} v_i = \lambda_i v_i (i = 1, 2, \cdots, n) \tag{10}$$

Where the eigenvalues λ_i represent the importance of the data along each direction, while the eigenvectors v_i describe the distribution of the data along these directions.

Finally, the risk factors that contribute significantly to the data variation are selected based on the magnitude of the eigenvalues, resulting in the number of risk feature factors used for risk identification as U.

The number of risk feature factors after dimensionality reduction is determined by the cumulative contribution rate:

$$C(u) = \frac{\sum_{i=1}^{u} \lambda_i}{\sum_{j=1}^{N} \lambda_j} \geq T \tag{11}$$

Where T is a constant, and $C(u)$ represents the cumulative contribution rate.

3) Constructing a risk matrix

First, by processing the feature data of various risk factors, the initial risk matrix can be obtained.

$$R_i = \begin{bmatrix} \overline{x}_{i11} & \overline{x}_{i21} & \cdots & \overline{x}_{im1} \\ \overline{x}_{i12} & \overline{x}_{i22} & \cdots & \overline{x}_{im2} \\ \vdots & \vdots & \vdots & \vdots \\ \overline{x}_{i1K} & \overline{x}_{i2K} & \cdots & \overline{x}_{imK} \end{bmatrix}, \forall i \in \{H, M_1, M_2, E\} \tag{12}$$

Next, each risk event is assigned a comprehensive risk score based on its severity and expert recommendations. The obtained scores are then classified into different risk levels according to established risk classification standards. Let $s_k = \{1,2,3,4,5\}$ represent the risk levels of the various risk events, with each s_k corresponding to a v_k.

In summary, the overall risk matrix can be represented as follows:

$$R^* = \begin{bmatrix} R_H & R_{M_1} & R_{M_2} & R_E & S \end{bmatrix} \tag{13}$$

Where $S = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{bmatrix}$.

After data augmentation using the CGAN and feature selection through the KPCA method[27][28], the final risk matrix is derived as follows:

$$R = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1U} & s_1 \\ z_{21} & z_{22} & \cdots & z_{2U} & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{K'1} & z_{K'2} & \cdots & z_{K'U} & s_{K'} \end{bmatrix} \tag{14}$$

Where U represents the number of new features after feature selection, and $z_{k'u}$ denotes the risk assessment data for the u-th new feature in the k'-th train dispatching risk event, K' represent the number of samples after data augmentation.

III. BO-LSSVM Algorithm for Risk Identification Model

A. Model selection and construction

The risk level identification problem in train dispatching involves systematically identifying, documenting, and analyzing potential risks that could arise during the railway scheduling process. The goal is to prevent and mitigate accidents, delays, and other adverse impacts.

In this study, the BO-LSSVM algorithm is employed for risk identification, utilizing a multi-classification approach to assess the risk levels of various factors. Based on the identification results, the severity of each risk event can be evaluated, allowing for the implementation of targeted railway accident prevention measures[29][30]. This method adopts an RBF kernel function, with Bayesian optimization applied to determine the optimal parameters c and γ for the RBF kernel.

The Least Squares Support Vector Machine (LSSVM) is a variant of the Support Vector Machine (SVM) that leverages the least squares method to optimize computation. Its core principle is to determine the decision hyperplane by minimizing the squared error, thereby effectively distinguishing different classes of samples. Compared to traditional SVMs, LSSVM reformulates the original quadratic programming problem into a linear system of equations, significantly improving computational efficiency[31]. The implementation steps of this algorithm are outlined as follows:

Step 1: For a given training dataset($z_{k'u}, s_{k'}$), where k' represents the risk events, $u \in \{1,2,\cdots U\}$, $k' \in \{1,2,\cdots K'\}$, $z_{k'u}$ denotes the input feature vectors, $s_{k'} \in \{1,2,3,4,5\}$

denotes the output labels, LSSVM seeks the decision hyperplane by solving the following optimization problem:

$$\min z = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_{k'u}^2 \tag{15}$$

$$s_{ik'}(w^T \phi(z_{k'u}) + b) = 1 - e_{k'u}, k' = 1, 2 \cdots K' \tag{16}$$

Where w is the weight vector, b is the bias, $e_{k'u}$ is the error variable, γ is the regularization parameter, and $\phi(z_{k'u})$ is the nonlinear mapping of the input features.

Step 2: By applying the Lagrange multiplier method, the above problem can be transformed into a dual problem:

$$L(w,b,e,a) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_{k'u}^2 - \sum_{i=1}^{n} \alpha_{k'u} \left[s_{k'u}(w^T \phi(z_{k'u}) + b) - 1 + e_{k'u} \right] \tag{17}$$

Step 3: By taking the partial derivatives of w, b, e and setting them to zero, we obtain the following system of linear equations:

$$\begin{bmatrix} 0 & 1^T \\ 1 & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ s \end{bmatrix} \tag{18}$$

Where Ω represents the kernel matrix, which denotes the similarity matrix between the training set sample, I represents the identity matrix, α represents the Lagrange multiplier vector.

By solving these equations, the values of the model parameters α and b can be obtained. With these parameter values, the weight vector w and the bias b can then be calculated.

Step 4: To address nonlinear problems, a kernel function is introduced to map the data into a high-dimensional feature space. Intelligent algorithms are then utilized to optimize the parameters c and γ within the kernel function. Here, c represents the regularization parameter, and γ denotes the kernel parameter, which controls the width of the kernel function. The most common kernel functions include those listed in Table II.

This paper selects the Gaussian Radial Basis Function kernel for computation and utilizes Bayesian optimization to optimize the parameters c and γ [32]. The structural framework of the LSSVM is shown in Fig. 2.

Step 5: For a new sample z', compute the predicted value:

$$f(z') = \sum_{i=1}^{K'} \alpha_i K(z_i, z') + b \tag{19}$$

The classification result can be determined based on the sign of the predicted value $f(z')$. This paper addresses a multi-class classification problem and employs a one-vs-one strategy for classification[33][34].

The flowchart of the BO-LSSVM algorithm implementation is shown in Fig. 3.

Table II
THE DISTRIBUTION OF COMMON KERNEL FUNCTIONS INCLUDES

Name	Expression	Applicable Scope
Linear Kernel Function	$K(z_i, z_j) = z_i \cdot z_j$	Suitable for linearly separable data
Polynomial Kernel Function	$K(z_i, z_j) = (\gamma(z_i \cdot z_j) + r)^d$	Suitable for data where there are polynomial relationships between features
Gaussian Radial Basis Function (RBF) Kernel	$K(z_i, z_j) = \exp(-\gamma \lVert z_i - z_j \rVert^2)$	Suitable for nonlinear data and data with high feature dimensions
Tanh Kernel Function	$K(z_i, z_j) = \tanh(\gamma(z_i \cdot z_j) + r)$	Performs well when the data exhibits a nonlinear relationship similar to neural network activation functions

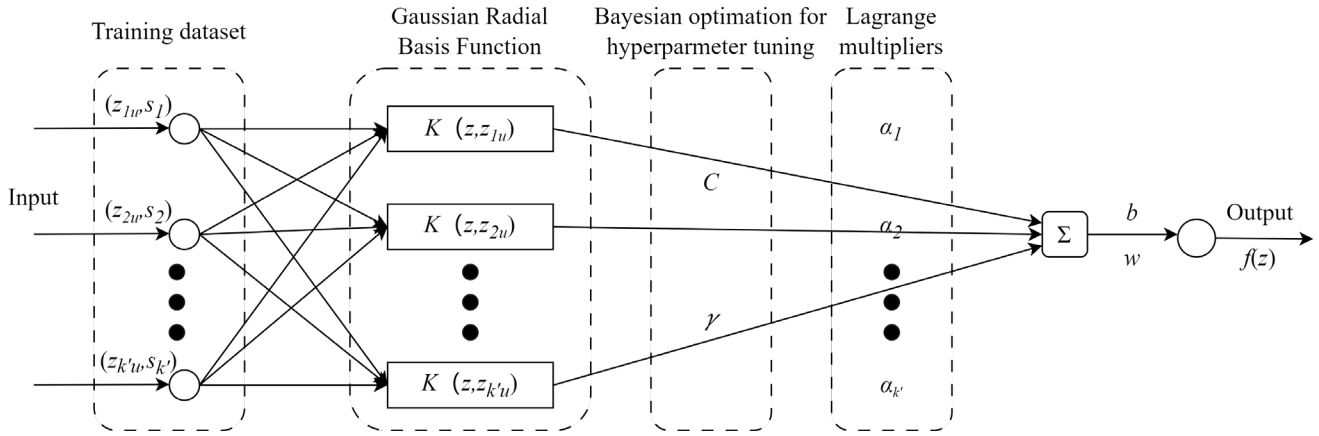

Fig. 2. LSSVM structural framework

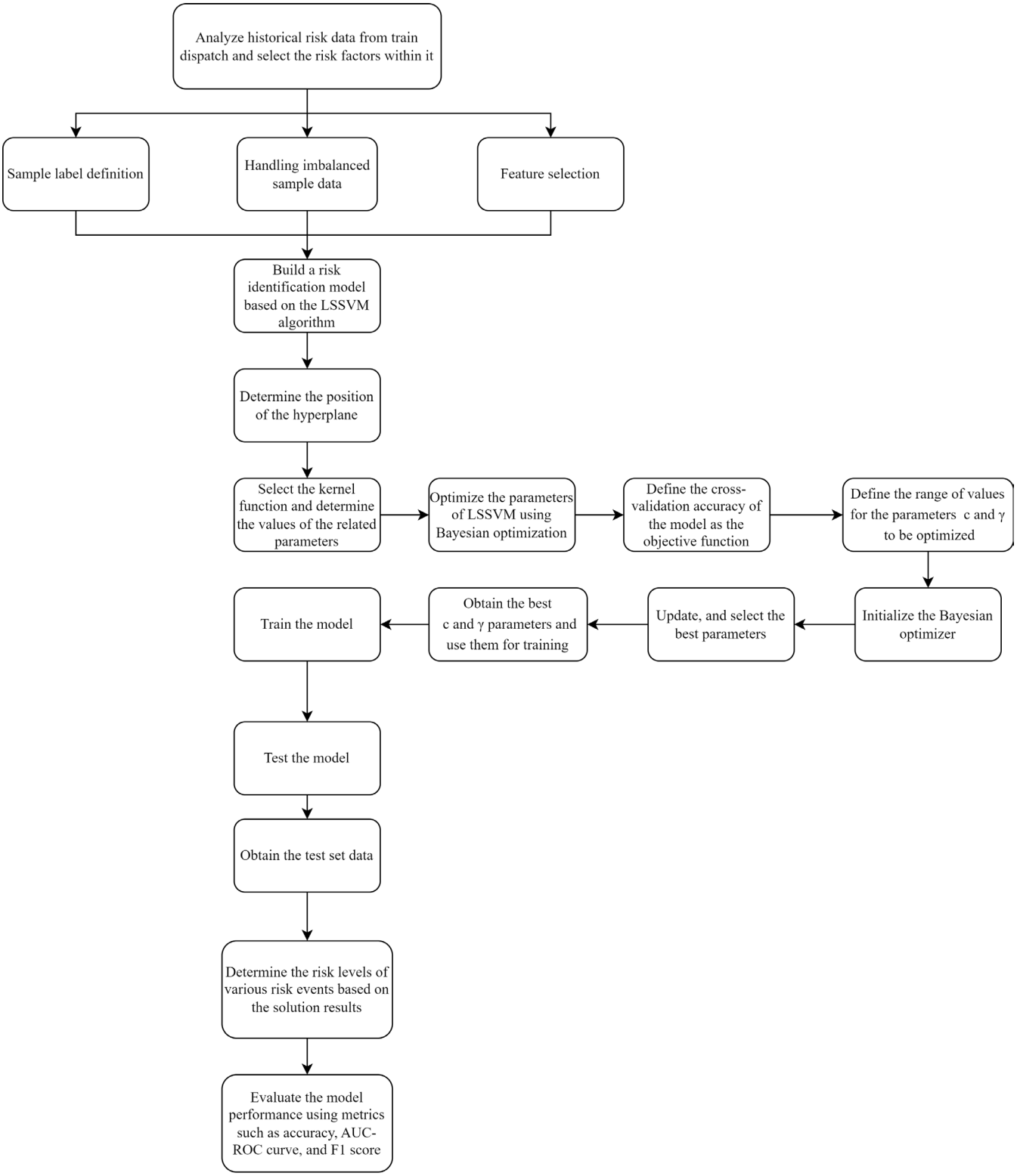

Fig. 3. Steps for risk identification based on the BO-LSSVM algorithm

B. Model evaluation and validation

After solving the algorithm, it is crucial to evaluate and validate the model's performance to ensure its effectiveness and generalization in practical applications. In this study, evaluation metrics include accuracy, precision, recall, F1 score, and the AUC-ROC curve. Accuracy, precision, recall, and F1 score are computed based on four key indicators: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Additionally, macro-averaging is employed to calculate precision, recall, and F1 score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{precision + Recall} \tag{23}$$

$$MacroAvg_j = \frac{\sum_{i=1}^{N} Metric_{ij}}{N} \quad (24)$$

Where TN represents the number of correctly identified negative labels, TP represents the number of correctly identified positive labels, FP represents the number of incorrectly identified negative labels, and FN represents the number of incorrectly identified positive labels. N represents the number of risk levels, and $Metric_{ij}$ represents the j-th evaluation metric (e.g. Precision, Recall, F1-score) for the i-th risk category.

The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is a widely used metric for evaluating the performance of binary classification models. The ROC curve illustrates the model's performance across different threshold settings, while the AUC value quantifies the area under the curve, ranging from 0 to 1. A higher AUC value indicates better classification performance[35].

To evaluate the model's generalization ability, this study employs k-fold cross-validation to assess its stability and robustness. The dataset is partitioned into k subsets, where each subset serves as the validation set while the remaining subsets are used for training. This process is repeated k times, and the final evaluation metrics are obtained by averaging the results from all iterations.

IV. Case Analysis and Results Discussion

A. Sample data processing

This study simulated potential risk events in train dispatching based on recent years' data, developed case studies, and assessed the risk identification performance of the CGAN-KPCA-BO-LSSVM algorithm. Initially, 1,005 train dispatching risk events were selected as data samples. These included 247 samples classified as risk level 1, 400 samples as risk level 2, 253 samples as risk level 3, 84 samples as risk level 4, and 21 samples as risk level 5. To mitigate overfitting and enhance sample diversity, the CGAN was employed to generate 150 additional samples for risk level 4 and 200 for risk level 5. Fig. 4 illustrates the performance changes of various models after data augmentation using CGANs. After data augmentation, the identification performance of each model significantly improved for risk levels 4 and 5. However, some models exhibited a decline in performance for risk levels 1, 2, and 3. One possible explanation is that, during data augmentation with CGANs, some noise or bias may have been introduced into the feature space, affecting the model's ability to generalize effectively to the non-augmented risk levels. Despite the decrease in identification accuracy for risk levels 1, 2, and 3, the decline was minimal and remained within an acceptable range.

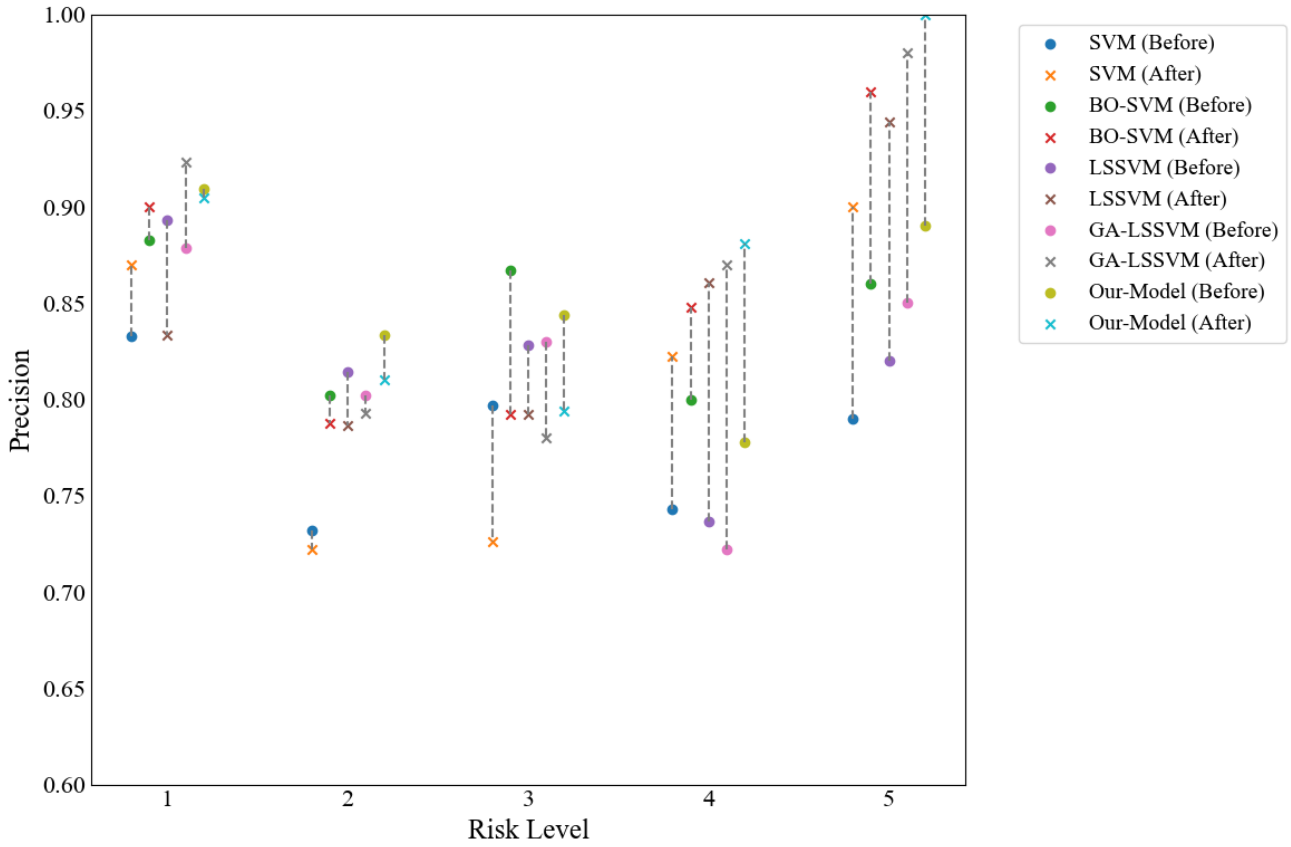

Fig. 4. Comparison of model identification performance before and after data augmentation

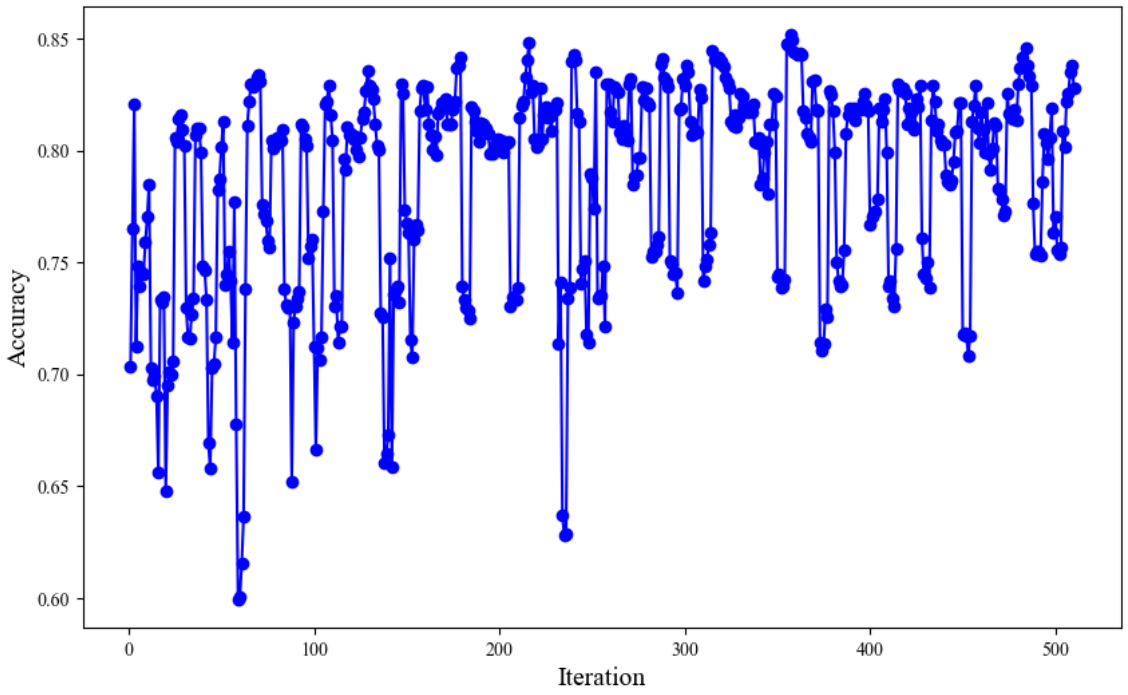

Fig. 5. Accuracy curve of the BO-LSSVM algorithm

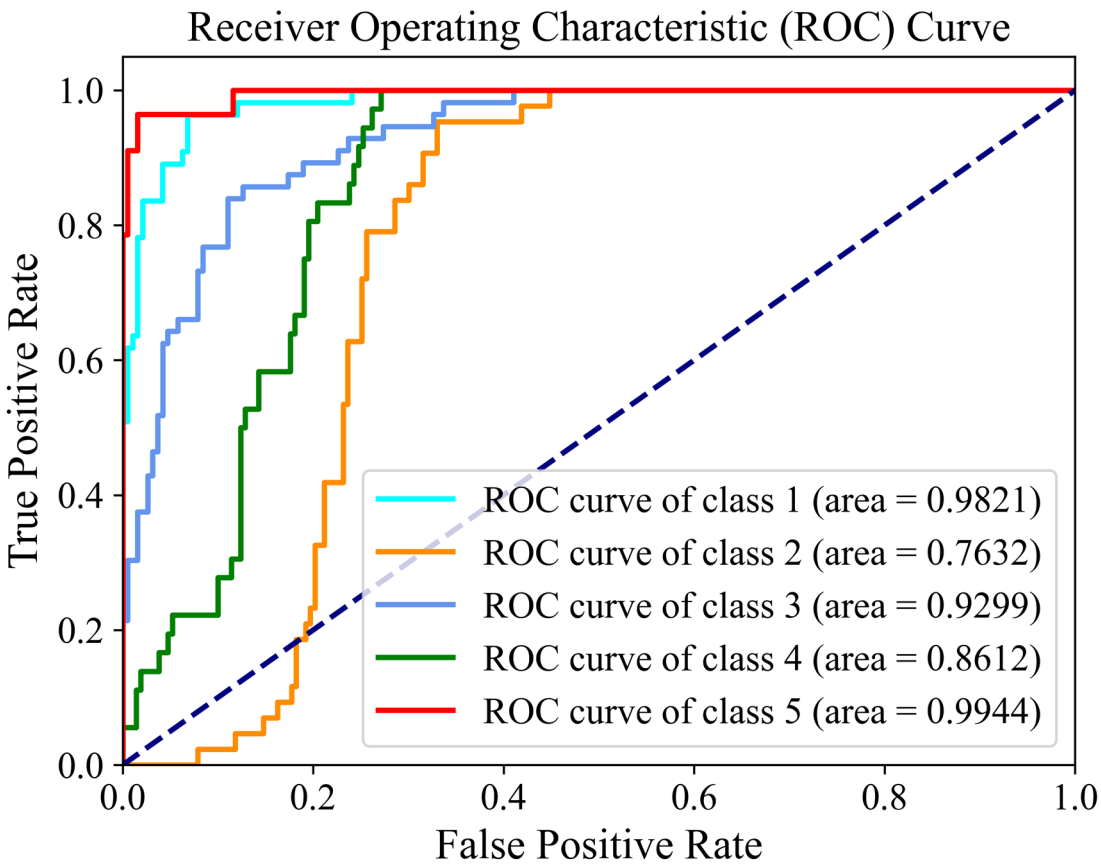

Fig. 6. AUC-ROC curves of the BO-LSSVM algorithm

B. Feature selection processing

The original feature data in this study consists of 17 different types of features. Some of these features exhibit complex nonlinear relationships, such as weather conditions, temperature, humidity, and visibility. The Kernel Principal Component Analysis (KPCA) method effectively captures these nonlinear relationships while simultaneously reducing

the dimensionality of the data. Based on the analysis of these factors, this study employs KPCA to fuse the 17 feature types. Subsequently, the LSSVM algorithm is used to evaluate the effectiveness of the feature fusion process. The fusion results for the different feature types are presented in Table III.

TABLE III
FEATURE SELECTION RESULTS

Number	Main risk factors	Contribution rates %	Cumulative contribution rates %
1	n_{24}	13.12	13.12
2	n_{23}	11.87	24.99
3	n_{15}	9.75	34.74
4	n_{21}	6.98	41.72
5	n_{13}	6.43	48.15
6	n_{22}	6.21	54.36
7	n_{14}	6.13	60.49
8	n_{34}	5.98	66.47
9	n_{12}	5.66	72.13
10	n_{11}	5.54	77.67
11	n_{33}	5.40	83.07
12	n_{32}	5.27	88.34
13	n_{44}	4.98	93.32
14	n_{41}	3.71	97.03

The collected data was subjected to KPCA to calculate the contribution rate and cumulative contribution rate of each risk feature factor. As shown in Table 3, the cumulative contribution rate of the first 14 principal components reached 97.03%, surpassing the 95% threshold. This suggests that these principal components can effectively capture the characteristics of the original data. Therefore, the top 14 primary risk feature factors were selected for subsequent risk level identification.

C. Case study analysis

The selected Bayesian optimization algorithm parameters in this study are as follows: the maximum number of iterations is set to 500, the range for parameter c is (0.01, 100), and the range for parameter γ is (0.0001, 10). 20% of the dataset is used as the test set, and 5-fold cross-validation is employed to evaluate the stability and generalization ability of the model. As shown in Fig. 5, with the increase in the number of iterations, the optimization process using the Bayesian algorithm gradually converges, reaching the highest accuracy at the 355th iteration.

Fig. 6 presents the AUC-ROC curves of the proposed CGAN-KPCA-BO-LSSVM based risk identification model, illustrating the classifier's performance in distinguishing five different risk levels. The True Positive Rate (TPR) represents the proportion of correctly identified positive samples among all positive cases, while the False Positive Rate (FPR) denotes the proportion of negative samples incorrectly classified as positive among all negative cases. The five distinct colored curves in the figure correspond to the ROC curves for each category. Curves positioned closer to the top-left corner indicate superior classifier performance, demonstrating a stronger ability to differentiate between positive and negative cases. Additionally, the Area Under the Curve (AUC) serves as a key metric for evaluating classifier performance, with values approaching 1 indicating better predictive accuracy. As shown in Fig. 6, it can be observed that the model exhibits strong discrimination capability for risk levels 1 and 5, with AUC values approaching 1.

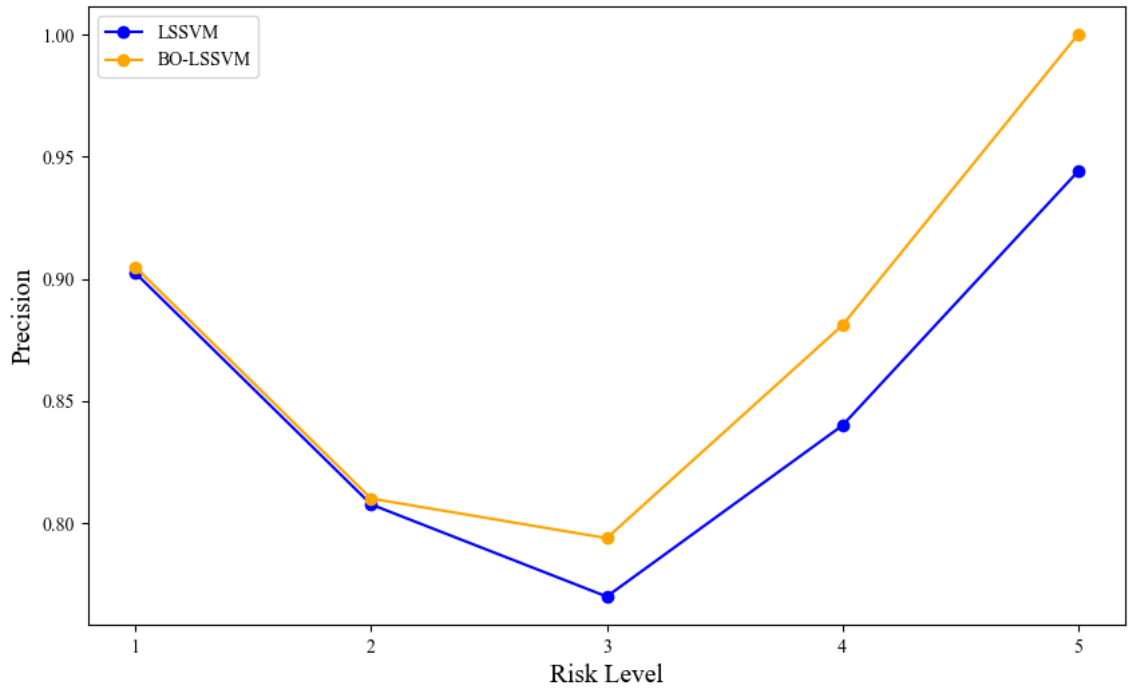

Fig. 7. Performance evaluation before and after algorithm optimization

To evaluate the effect of hyperparameter optimization using the Bayesian optimization algorithm on the performance of the proposed risk identification model, we examined the precision metrics for each risk level in the validation set. As shown in Fig. 7, after hyperparameter optimization, the updated risk identification model exhibits improved performance in identifying various risk levels, especially for risk levels 4 and 5.

D. Hyperparameter performance evaluation

Fig. 8 shows the curves of the accuracy as hyperparameters c and γ vary after Bayesian optimization. The selected parameter ranges are $c \in (0.01,100)$ and $\gamma \in (0.0001,10)$. The points marked in the figure indicate the points where the accuracy is highest. As illustrated, the highest accuracy is achieved when c is 25.0166 and γ is 0.4454.

V. Comparative Study

To further validate the risk identification performance of the BO-LSSVM algorithm, as well as the effects of data augmentation using CGAN and feature selection via KPCA, this paper compares the performance of the proposed model with that of seven other optimized algorithm models. Specifically, the Accuracy, Precision, Recall, and F1-score of each model are compared. The results are shown in Table IV.

From the prediction results, it is evident that the BO-LSSVM algorithm outperforms others across various evaluation metrics. Notably, the RBF kernel achieves higher scores on all metrics when compared to the Linear LSSVM and Polynomial LSSVM algorithms. Furthermore, by addressing the data imbalance using CGAN and applying feature selection through the KPCA, the accuracy of all algorithms showed noticeable improvement.

TABLE IV
Comparison of Algorithm Prediction Results

Model	Before Data Processing			
	Accuracy	$Precision_{macro}$	$Recall_{macro}$	$F1-Score_{macro}$
SVM	76.00%	80.10%	57.67%	58.59%
GA-SVM	82.00%	87.02%	66.47%	66.66%
Linear-LSSVM	80.50%	83.99%	65.40%	64.20%
Polynomial -LSSVM	82.00%	82.84%	79.75%	81.51%
LSSVM	79.00%	85.12%	75.52%	78.15%
GA-LSSVM	82.50%	86.35%	68.31%	67.04%
Our Model	82.51%	84.61%	82.83%	83.52%
Model	After Data Processing			
	Accuracy	$Precision_{macro}$	$Recall_{macro}$	$F1-Score_{macro}$
SVM	79.85%	83.45%	78.89%	80.02%
GA-SVM	85.21%	86.02%	84.68%	84.82%
Linear-LSSVM	84.46%	84.32%	83.36%	83.88%
Polynomial -LSSVM	83.71%	84.63%	83.45%	83.89%
LSSVM	84.64%	84.32%	83.76%	84.03%
GA-LSSVM	85.78%	84.52%	83.69%	84.01%
Our Model	87.54%	87.01%	86.52%	86.67%

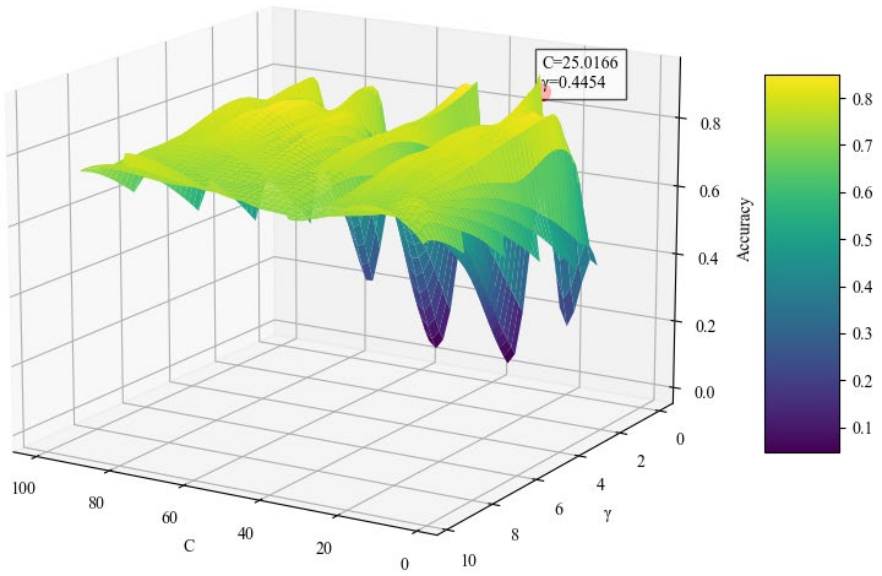

Fig. 8. Hyperparameter optimization curves

VI. Conclusion

This paper proposes a risk identification model based on the CGAN-KPCA-BO-LSSVM algorithm to address safety issues in railway transportation dispatching. First, by analyzing historical risk event data from railway dispatching, key risk factors affecting the system were identified, and these factors were primarily categorized into personnel, equipment, train, and natural environment factors. To address data imbalance and high-dimensional feature data issues, CGAN and KPCA were employed. The LSSVM algorithm was used for classification, with Bayesian optimization applied to fine-tune the RBF kernel parameters.

Through case analysis, the proposed CGAN-KPCA-BO-LSSVM algorithm demonstrated an 11.54% improvement in accuracy compared to the traditional SVM algorithm, validating the effectiveness of this method in solving the risk level identification problem in train dispatching.

In conclusion, the proposed risk level identification model enhances the efficiency of risk identification in railway transportation scheduling to some extent, providing a novel solution for managing high-dimensional risk feature data in this field. It also offers valuable insights for risk assessment in other sectors.

However, despite the positive results in identifying risks within railway transportation scheduling, certain aspects require further exploration. Due to the limited sample data used in this study, the robustness of the model could not be fully validated. Future research could introduce more diverse sample data and risk factors to further assess the model's applicability. Additionally, this paper focuses on 17 identified risk factors, but future studies may explore additional factors that impact railway transportation safety.

References

[1] Yumou Ren, Yanhao Sun, Shuxin Ding, Zhi Li and Xiaozhao Zhou, "Research on risk identification of human errors of train dispatchers in high-speed railway," In Seventh International Conference on Traffic Engineering and Transportation System (ICTETS 2023), vol. 13064, pp6-13, 2024.

[2] Nafiseh Esmaeeli, Fereshteh Sattari, Lianne Lefsrud, and Renato Macciotta, "Assessing the Risks Associated with the Canadian Railway System Using a Safety Risk Model Approach," Transportation Research Record, vol.2678, no.2, pp795-808, 2024.

[3] Leitner, and Bohus, "A general model for railway systems risk assessment with the use of railway accident scenarios analysis," Procedia engineering, vol. 187, no.187, pp150-159, 2017.

[4] Jun Lai, Kai Wang, Jingmang Xu, Ping Wang, Rong Chen, Shuguo Wang, and Michael Beer, "A failure probability assessment method for train derailments in railway yards based on IFFTA and NGBN," Engineering Failure Analysis, vol. 154, suppl C, pp10765, 2023.

[5] Ruifan Tang, Lorenzo De Donato, Nikola Bešinović, Francesco Flammini, Rob M.P. Goverde, Zhiyuan Lin, Ronghui Liu, Tianli Tang, Valeria Vittorini, and Ziyulong Wang, "A literature review of Artificial Intelligence applications in railway systems," Transportation Research Part C: Emerging Technologies, vol. 140, suppl C, pp103679, 2022.

[6] Huipeng Liu, Minghua Hu, and Lei Yang, "A new risk level identification model for aviation safety," Engineering Applications of Artificial Intelligence, vol. 136, part A, pp108901, 2024.

[7] Jicheng Liu, Yanan Song, and Xue Yu, "Risk assessment study of hydrogen energy storage system based on KPCA-TSO-LSSVM," International Journal of Hydrogen Energy, vol. 79, no.0, pp931-942, 2024.

[8] Keyang Liu, Baoping Cai, Qibing Wu, Mingxin Chen, Chao Yang, Javed Akbar Khan, Chenyushu Wang, Hasini Vidumini Weerawarna Pattiyakumbura, Weifeng Ge, and Yonghong Liu, "Risk identification and assessment methods of offshore platform equipment and operations," Process Safety and Environmental Protection, vol. 177, pp1415-1430, 2023.

[9] Wencheng Huang, Hongyi Liu, Yue Zhang, Rongwei Mi, Chuangui Tong, Wei Xiao, and Bin Shuai, "Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM," Applied Soft Computing, vol. 109, pp107541, 2021.

[10] Sadiq Khan, Amjad Pervez, Yinggui Zhang, Suleman Ahmad, Hijratullah Sharifzada, Emad A.A. Ismail, and Fuad A. Awwad, "Comprehensive Risk Assessment of Pakistan Railway Network: A Semi-Quantitative Risk Matrix Approach," Heliyon, vol. 10, no.12, ppe32682, 2024.

[11] Yujie Huang, Zhipeng Zhang, Yu Tao, and Hao Hu, " Quantitative risk assessment of railway intrusions with text mining and fuzzy Rule-Based Bow-Tie model," Advanced Engineering Informatics, vol. 54, suppl C, pp101726, 2022.

[12] Jintao Liu, Keyi Chen, Huayu Duan, and Chenling Li, "A knowledge graph-based hazard prediction approach for preventing railway operational accidents," Reliability Engineering & System Safety, vol. 247, pp110126, 2024.

[13] Keping Li, and Shanshan Wang, "A network accident causation model for monitoring railway safety," Safety science, vol. 109, no.0, pp398-402, 2018.

[14] Jun Liu, Gehui Liu, Yu Wang, and Wanqiu Zhang, " Artificial-intelligent-powered safety and efficiency improvement for controlling and scheduling in integrated railway systems," High-speed Railway, vol. 2, no.3, pp172-179, 2024.

[15] Johanna T\"ornquist, "Computer-based decision support for railway traffic scheduling and dispatching: A review of models and algorithms," OASIcs: OpenAccess Series in Informatics, 2006, Doi: 10.4230/OASIcs.ATMOS.2005.659.

[16] Changfeng Zhu, Yu Wang, Qingrong Wang, Jinhao Fang, Jie Wang, and Linna Cheng, "Research on Traffic Accident Prediction Based on KG-CWT-RGCNN-BiLSTM," Engineering Letters, vol. 31, no.4, pp1402-1414, 2023.

[17] Guangyu Zhao, Peng Liu, Ke Sun, Yang Yang, Tianyu Lan, and Han Yang, "Research on data imbalance in intrusion detection using CGAN," Plos one, vol. 18, no.10, 2023.

[18] Hongtao Guan, Yijie Wang, Xingkong Ma, Yongmou Li, "DCIGAN: A Distributed Class-Incremental Learning Method Based on Generative Adversarial Networks," In 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, pp768-775, 2019.

[19] Xinyi Li, Chunhao Wang, Yang Sheng, Jiahan Zhang, Wentao Wang, Fang-Fang Yin, Qiuwen Wu, Q Jackie Wu, and Yaorong Ge, "An artificial intelligence- driven agent for real- time head- and- neck IMRT plan generation using conditional generative adversarial network (CGAN)," Medical physics, vol. 48, no.6, pp2714-2723, 2021.

[20] Dongping Li, Yingchun Yang, Shikai Shen, Jun He, Haoru Shen, Qiang Yue, Sunyan Hong, and Fei Deng, "Research on Fault Diagnosis based on Improved Generative Adversarial Network under Small Samples," IAENG International Journal of Computer Science, vol. 50, no.1, pp7-13, 2023.

[21] Xiaoyuan Dang, Guorui Liu, Xianlun Tang, Shifei Wang, Tianzhu Wang, and Mi Zou, "Motor Imagery EEG Recognition Based on Generative and Discriminative Adversarial Learning Framework and Hybrid Scale Convolutional Neural Network," IAENG International Journal of Applied Mathematics, vol. 52, no.4, pp946-954, 2022.

[22] Yiqian Sun, Meiqi Song, Chunjing Song, Meng Zhao and Ramzan Talib, "KPCA-based fault detection and diagnosis model for the chemical and volume control system in nuclear power plants," Annals of Nuclear Energy, vol. 211, no.o, pp110973, 2025.

[23] Mohsena Ashraf, Ruichao Jia, Renren Zhang, Shangyu Yang, and Guoming Chen, "A KPCA-BRANN based data-driven approach to model corrosion degradation of subsea oil pipelines", Reliability Engineering & System Safety, vol. 219, Suppl C, 2022.

[24] Juanxia He, Liwen Huang, Yao Xiao, Wen Li, Jiamei Yin, Qingshan Duan and Linna Wei, "Prediction model of continuous discharge coefficient from tank based on KPCA-DE-SVR," Journal of Loss Prevention in the Process Industries, vol. 89, pp105316, 2024.

[25] Andrej Gisbrecht, and Barbara Hammer, "Data visualization by nonlinear dimensionality reduction," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 2, pp51-73, 2015.

[26] Huizhi Gou, and Yuncai Ning, "Forecasting Model of Photovoltaic Power Based on KPCA-MCS-DCNN," CMES-Computer Modeling in Engineering & Sciences, vol. 128, no.2, pp803-822, 2021.
[27] Yiqian Sun, Meiqi Song, Chunjing Song, Meng Zhao and Yanhua Yang, "KPCA-based fault detection and diagnosis model for the chemical and volume control system in nuclear power plants," Annals of Nuclear Energy, vol. 211, no.0, pp110973, 2025.
[28] Mohammad Zarei, Bruce Hellinga, and Pedram Izadpanah, "Application of Conditional Deep Generative Networks (CGAN) in empirical bayes estimation of road crash risk and identifying crash hotspots," International Journal of Transportation Science and Technology, vol. 13, pp258-269, 2024.
[29] Said Elbostani, and Rachid El Jid, "A Meshless Method Based on the Moving Least Squares Approach for Approximate Solution of the Generalized 2-D Nonlinear Benjamin-Bona-Mahony-Burgers Equation," IAENG International Journal of Applied Mathematics, vol. 54, no. 9, pp1734-1746, 2024.
[30] Haifeng Wang, and Dejin Hu, "Comparison of SVM and LS-SVM for regression," In 2005 International conference on neural networks and brain, vol. 1, pp279-283, 2005.
[31] Yangyu Deng, and Yakun Liu, "Prediction of Depth-Averaged Velocity for Flow Though Submerged Vegetation Using Least Squares Support Vector Machine with Bayesian Optimization," Water Resources Management, vol. 38, no.5, pp1675-1692, 2024.
[32] Lijia Chen, Peiyi Yang, Shengwei Li, Yanfei Tian, Guangquang Liu, and Guozhu Hao, "Grey-box identification modeling of ship maneuvering motion based on LS-SVM," Ocean Engineering, vol. 266, Part 3, pp0, 2022.
[33] Xishan Dong, Meili Sun, Ting Zhang, Qiaolian Liu, and Weikuan Jia, "A Reliable Ensemble Classification Algorithm by Genetic Neural Network based on Multiple Regression," IAENG International Journal of Computer Science, vol. 50, no.4, pp1269-1278, 2023.
[34] Xinying Chen, Xupeng Liang, Weiguo Yi, Xudong Song, Di Wang, and Yina Zhang, "A Multi-label Classification Algorithm Combining Feature Screening and Label Correlation," IAENG International Journal of Computer Science, vol. 50, no.4, pp1578-1585, 2023.
[35] Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," Patterns, vol. 5, no.6, pp100994, 2024.