Received: 18 September 2022 | Revised: 12 January 2023 | Accepted: 13 March 2023

IET Intelligent Transport Systems

DOI: 10.1049/itr2.12357

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Fault diagnosis of ZDJ7 railway point machine based on improved DCNN and SVDD classification

Zengshu Shi[1] | **Yiman Du**[2] | **Xinwen Yao**[1]

[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China

[2]Beijing SWJTU RichSun Tech Co. Ltd., Beijing, China

Correspondence
Yiman Du, Beijing SWJTU RichSun Tech Co. Ltd., Beijing, China.
Email: duyiman@swjtu-richsun.com

Funding information
China State Railway Group Co., Ltd., Grant/Award Number: N2018G062

Abstract

Problems such as poor noise immunity and overfitting are prone to occur when convolutional neural network (CNN) is exploited in the fault diagnosis of ZDJ7 railway point machine. In addition, some fault features are unbalanced and have the features of multiple tags, which lead to low diagnosis accuracy. Therefore, an improved deep convolutional neural network (DCNN) and support vector data description (SVDD) classification is proposed. First, the depthwise separable convolution in the Xception structure is used to optimize the extraction of fault features. Second, the adaptive batch normalization processing (AdaBN) is performed to improve the noise immunity. Meanwhile, the global average pooling layer (GAP) is used instead of the fully connected layer to improve the generalization ability of the network. Aiming at the unbalanced features of the railway point machine sample, an improved quantity learning algorithm for hypersphere coordinate mapping based on SVDD is proposed. The classification is realized under unbalanced samples. The experiment shows that the accuracy based on the improved DCNN and SVDD is 96.59%. It has a good anti-noise performance under different convolution kernels and SNRs. When the sample distribution is unbalanced, the performance indexes obtained by the proposed model are the best.

1 | INTRODUCTION

With the rapid development of high-speed railways worldwide [1], the railway point machine plays an increasingly important role in maintaining the safety of railway transportation [2]. The working state of its equipment directly determines the safety of train operation. However, the poor working environment of the railway point machine and the most frequent actions make the highest fault rate in the railway signal infrastructure. At present, the main artificial experience diagnosis method has the problems of a low efficiency, high cost, and high misdiagnosis rate. In turn, they cannot meet the field requirements. Therefore, the diagnosis of this equipment needs to be as efficient and accurate as possible.

At present, the automatic diagnosis of railway point machine faults has been studied by many scholars at home and abroad [3, 4]. It can be divided into diagnosis models based on machine learning [5–7] and deep learning algorithms. Machine learning models typically use simple formulas to define the association model. Therefore, the features need to be extracted manually. The modelling process is too idealistic, and the robustness is not good. Moreover, the fault features extracted manually are not comprehensive. On one hand, the features that reflect minor faults are easily deleted by mistake or concealed by noise. On the other hand, the extracted features are primarily used to solve specific fault problems [8, 9]. They have poor universality and are difficult to complete in the context of big data. The main reason for the above defects is that most of the models used in the existing fault diagnosis algorithms are shallow structures. The feature extraction ability is weak. However, diagnosis models based on deep learning have been widely used because they can be trained to extract features automatically [10]. Furthermore, the CNN is representative of deep learning methods. The original signal is directly used as input to simulate the complex relationship between input and output. It reduces the dependence on expert experience and does not need to extract features manually. It is increasingly used in mechanical fault diagnosis. Moreover, the deep learning-based model for

extracting fault features can extract the sensitive fault feature. Although the CNN has good performance for feature extraction, it is sensitive to the problem of the unbalanced sample [11]. In addition, valuable fault information is scarce in fault diagnosis. Also, the essence of fault diagnosis is a pattern recognition problem. Correspondingly, accurate feature extraction is critical for fault diagnosis [12, 13]. However, a critical classifier or cluster is also required. Also, this process is quantified as the final diagnosis process. Besides, insufficient training samples and an unbalanced sample distribution will result in overfitting issues. Therefore, we will propose an improved DCNN structure to solve the above mentioned fault feature extraction problem in this study.

Consequently, effective fault feature extraction is accomplished using improved DCNN structure. However, fault and fault samples are scarce because of the lack of on-site fault data. In turn, it easily causes overfitting and poor noise immunity. Consequently, the normalization ability of the training model is poor. However, the condition monitoring and diagnosis of mechanical equipment is a very typical unbalanced classification problem. We urgently need to solve the problem of classifying the unbalanced samples. Fortunately, a lot of researches on unbalanced fault diagnosis has already appeared [41–43]. Although the traditional methods of unbalanced sample classification can achieve certain results. However, they still have the following shortcomings [44–47]. Specifically, the following: Limited by the poor collection and production conditions, the number of training set has difficulty meeting the above assumptions. In turn, the accuracy and reliability of the trained classification model are poor. Therefore, the support vector data description (SVDD) is used to construct a single class of hyperspheres for fault features in this paper. An improved quantity learning algorithm for hypersphere coordinate mapping is proposed.

After the above analysis and research, our improvement and the innovations of this study are as follows. The improvement of the CNN network is mainly divided into the following two steps. First, the depthwise separable convolution in the Xception structure is used to optimize the extraction of fault features. Then, the adaptive batch normalization processing (AdaBN) is performed to improve the noise immunity. Meanwhile, the global average pooling layer (GAP) is used instead of the fully connected layer to improve the generalization ability of the network. Finally, an improved quantity learning algorithm for hypersphere coordinate mapping based on SVDD is proposed. The classification of unbalanced samples is realized. Compared to other contractual models, the following points present the contributions of our method:

- We designed an improved DCNN model with better performance. Specifically, the depthwise separable convolution in the Xception structure is used to optimize the extraction of fault features. Further, the AdaBN strategy is proposed to optimize the performance of DCNN. In addition, we use GAP instead of the fully connected layer to improve the generalization ability of the model.

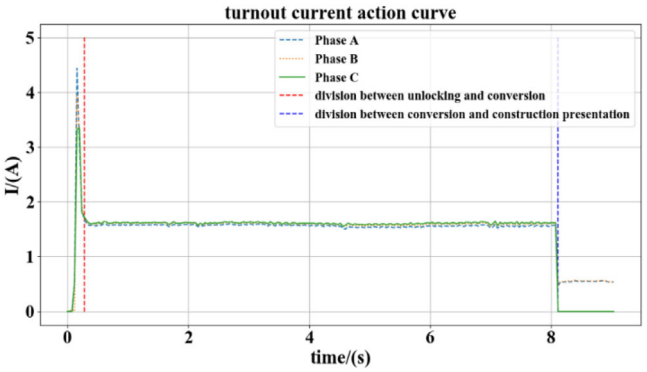

FIGURE 1 Curve of ZYJ7 railway point machine.

- Aiming at the unbalanced features of the railway point machine sample, an improved quantity learning algorithm for hypersphere coordinate mapping based on SVDD is proposed. In turn, the classification of unbalanced samples is realized.
- Experimental results on fault datasets show that our model has excellent diagnosis performance in unbalanced sample separation.

The remainder of this paper is organized as follows. The backgrounds of fault diagnosis of railway point machine and the idea of the proposed method is introduced in Section 1. In Section 2, we describe the related works of railway point machine diagnosis. In Section 3, the realization process of fault diagnosis method based on improved DCNN and SVDD classification is introduced. In Section 4, we set up a series of experiments to verify the performance of the proposed method. In Section 5, we summarize the work and propose the next step of the research plan.

2 | RELATED WORKS

The following content will be introduced from the ZDJ7 railway point machine curve analysis, machine learning diagnosis method, deep learning diagnosis method and unbalanced sample classification.

2.1 | ZDJ7 railway point machine manoeuvre curve and fault type analysis

According to the entire process of railway point machine conversion, the action current and total power curve collected are divided into four phases. The normal curve is shown in Figure 1.

The four phases are the T1 unlock zone, T2 action zone, T3 lock zone and T4 buffer zone. T1 (unlocking zone): The motor in the railway point machine barely starts to move. The current changes radian greatly, which can complete the function of unlocking the railway point machine. T2 (action zone): The power supply drives the motor to rotate normally and

pull the railway point machine to the specified position. In the absence of other external forces, the railway point machine rotates smoothly, and the current is stable. T3 (locked area): The railway point machine enters the locked state. T4 (buffer zone): After the railway point machine is converted to the specified position, 1DQJ remains in the buffer phase. In a 380 V three-phase power supply, the two phases formed a loop through the representation circuit. The two phases are dependent on the existing current. At the end of the 1DQJ cache, the three-phase current is 0. The three-phase current of the alternating current is expressed in terms of A, B, and C. The power is expressed in terms of P. When the fault occurs, its action current curve would also be fault. Several common railway point machine faults and their causes are given in Table 1.

2.2 | Fault diagnosis based on traditional machine learning

With the continuous rise of machine learning research, data-driven intelligent fault diagnosis had gradually become the mainstream method in the fault diagnosis [6, 7, 14]. The traditional intelligent fault diagnosis method usually includes three steps: Feature extraction, feature selection, and fault classification. Feature extraction was the key link of fault diagnosis. However, machine learning requires manual feature extraction, and the key to the success was the selection of a feature extractor and classifier. Therefore, feature extraction requires much work and high requirements for designers. The shallow network learning ability was insufficient. Therefore, the diagnosis accuracy was not high. Other commonly used methods include wavelet transform (WT) [15], spectral analysis (SA) [16] and empirical modal decomposition (EMD) [17].

Although these methods can achieve certain diagnosis results, their accuracy was relatively low. Feature selection removed the features with low sensitivity and utility from the extracted features. Thus, the aim of reducing the number of features and redundancy of fault features was achieved. Common methods primarily include principal component analysis (PCA) [18] and independent component analysis (ICA) [19]. Fault classification inputs selected features into fault classifiers for recognition and trains the fault classification results by repeated iterations. The BP neural network (BPNN) [20], support vector machine (SVM), and K-nearest neighbour method (KNN) have been widely used as back-end fault classifiers [21]. In 2016, an intelligent diagnosis method for high-speed railway turnouts based on SVM was proposed. Also, it used expert experience and knowledge to compute feature vectors [22]. In 2017, a hybrid feature extraction method was proposed. It extracted information in the time and frequency domain from the turnout signal, and the SVM was used for signal recognition. The introduction of low frequency features made fault diagnosis more sensitive to more types of faults [23]. As shown by ref. [24], an SVM was proposed as the core. Moreover, different algorithms were adopted to achieve fault diagnosis. As shown by ref. [25], EMD and WT signal analysis methods were combined to extract signal envelope features. The most effective features were used for classification to realize fault diagnosis.

In 2012, a method combining Bayesian network and fault diagnosis was proposed for fault diagnosis of railway point machine. However, it required high prior knowledge [26]. In 2014, a diagnosis method based on BPNN was proposed. It was based on the change rule of the curve to be detected for the turnout. The mapping dataset was established, and the classification was processed through the neural network [27]. Ref. [28] proposed an efficient Bayesian network based on rough set reduction. The complex relation between fault mode and fault cause was modelled to realize fault diagnosis. Ref. [29] combined neural network technology with grey correlation analysis to establish the corresponding fault feature for the power curve in common fault modes. First, the grey relational degree between fault curves and power curves was computed by a grey neural network. Second, the fault mode of the equipment was judged according to the correlation degree. Finally, the intelligent fault diagnosis of the railway point machine was realized. Ref. [30] proposed a fault diagnosis method with strong computing power to address uncertain knowledge reasoning problems based on a standard Bayesian network. Although this model can obtain some effect, it easily falls into the locally optimal and reduced search space. Moreover, the model only analyses the circuit fault of the railway point machine. In the actual fault record of a railway point machine, most of the faults were caused by mechanical reasons. Therefore, circuit faults had no practical diagnosis significance. As shown by ref. [31], four diagnosis models were compared. The optimal conclusion of the multilayer bidirectional model was obtained according to the experiment. Finally, the integrated attention mechanism increased the accuracy of fault classification to 97.6%.

To conclude, the above algorithms have poor feature extraction capability due to their shallow network structure, which makes it difficult to mine. The tiny features in the deeper level of fault data is difficult to extract. Moreover, scholars usually adopted the method of combining manual feature extraction with a shallow machine learning model for fault diagnosis. Although they has some effect, they has many shortcomings. On one hand, feature selection must depend on the experience and professional knowledge of engineers. In turn, it is subjective and blind. It can be seen although traditional machine learning methods can achieve certain results in fault diagnosis, the accuracy is low.

2.3 | Fault diagnosis based on deep learning

Deep learning (DL) methods have a strong feature extraction ability and can automatically extract features from massive data. The mapping relationship between input data and fault categories was well represented by establishing a deep learning model. It was very suitable for monitoring data diagnosis with a large data volume, feature non-linearity, and high dimensionality. In 2006, Hinton et al. [32] proposed deep learning theory. Specifically, it used a deep neural network structure to extract

TABLE 1 Common fault types of ZYJ7 railway point machine

Fault case	Fault phenomenon	Curve type	Fault analysis
Gap stuck	Railway point machine no-effect	The operation time of the railway point machine is the same as the normal conversion time, but there is no "small tail" line in the easing zone.	There is no "small tail" in the cache phase, indicating that 1DQJ has not entered the cache state. When the railway point machine to the specified position, there is no railway point machine position information. The automatic railway point machine is in the four-open state.
There are obstacles in the conversion.	Railway point machine no-effect	The railway point machine is not locked after being pulled to the position. The transition process is idling curve, and the action time is very short.	The railway point machine is not converted to the specified position.
Not locking	Railway point machine no-effect	The idling curve appears during the railway point machine conversion. The turning time is slightly shorter than the normal switching time of the railway point machine.	The gauge changes and the pressure is high.
Poor start cylinder plunger	The conversion time is long	When the railway point machine switches, the operation time is longer than normal.	The turnout action is slow, and the tension is reduced. There is the problem of oil transmission system.
Disconnect protector plug, two-phase or three-phase power fault	Railway point machine no-effect	The collected operating current of the railway point machine is always 0 A.	The three-phase power supply is cut off. The power supply has no output and the railway point machine cannot operate normally.
Poor insulation between wires	Fault operating current of the railway point machine	The action current of the turnout appears burr phenomenon. After the railway point machine is in place, the turnout position is indicated.	The poor insulation of the cable results in the phenomenon of semi-short circuit during the conversion, which leads to the phenomenon of instantaneous high burr in the action current curve.
Continuous circuit connection Poor point contact	Fault operating current of the railway point machine	In the buffer zone, the two-phase current suddenly rises to 5 A and then quickly falls back to 0 A, indicating the position of the railway point machine is fault when it is converted into position.	The two traction points are out of sync and the contacts in the continuation circuit are in bad contact.
Continuation circuit interrupt fault	Railway point machine no-effect	The switching time of the railway point machine is short, and there is no "small tail" in the buffer. The railway point machine position is not represented. When the railway point machine is pulled back, the conversion time is short. The conversion is in place with position information.	The motor continues to operate after the first traction point is converted into position. The poor cable communication between the two traction points makes the railway point machine short of two phases in the continuation circuit. So the continuation circuit cannot be closed. Therefore, the time is short, there is no 1DQJ easing curve.
Disconnect wiring between connector and motor	Railway point machine no-effect	The position information of the railway point machine is not shown, and the action current curve is fault.	When dealing with the open circuit fault of railway point machine circuit, first judge whether there is voltage in the circuit of indoor and outdoor connection. Indoor fault is determined, so the time of fault judgment is shortened.

the features of the input sample layer by layer. The layer-by-layer greedy algorithm was used to learn the internal relationship between data and tags. The dependence on manual feature extraction and expert experience was removed. The inherent defects of traditional intelligent diagnosis methods have been overcome. At present, it had been successfully applied in speech recognition, image recognition, and natural language processing. In the last 5 years, it had attracted the attention of scholars in the fault diagnosis. Jia et al. [33] proposed a fault diagnosis method based on DAE, which completed the fault diagnosis of rolling bearings and planetary gearboxes. However, the original data needed to be converted into spectral signals. Tamilselvan et al. [34] proposed a fault diagnosis method based on a deep belief network (DBN). The fault diagnosis of railway point machine was completed, and good diagnosis results were obtained. Li et al. [35] applied a three-layer DBN network to fault diagnosis experiments. However, bearing faults were simulated by groove cutting, and the fault features belonged to the category of significant faults. Therefore, it was difficult to reflect the effect of intelligent diagnosis.

The CNN is one of the important branches of deep learning. It has powerful feature extraction ability and the potential

to extract intelligent features. In recent years, scholars have applied CNN in the fault diagnosis. A CNN model was used in the fault diagnosis of asynchronous motors. Although both of fault diagnosis used the CNN, the traditional feature extraction method was still needed to extract the original vibration data [36]. Therefore, it failed to make full use of CNN's powerful feature extraction capability, which limited the further improvement of the fault diagnosis effect. Qu et al. [37] proposed an algorithm based on a one-dimensional CNN for fault diagnosis. The training parameters of the fully connected layer of the CNN model were too large, which leaded to a long training and testing time. As shown by ref. [38], a feature learning model based on a CNN was proposed. Fault diagnosis was performed through self-learning of useful features. The better performance was obtained than those of manual feature extraction and random forest combination. Wang et al. [39] proposed an adaptive CNN and applied it to the fault diagnosis of railway point machines. The feature was extracted automatically by a CNN to avoid artificial feature extraction. Particle swarm optimization was used to determine the main parameters. Ki Bum Lee et al. [40] proposed a CNN model. The fault features of multi-sensor signals were extracted along the time axis to correlate the output. Thus, the variable and time information representing the fault was found. Lu et al. [41] used a CNN to directly learn features from frequency data of vibration signals. The performance differences caused by extracting features from three types of data (raw data, spectrum, and time-frequency combination data) were compared. In recent years, the Inception network structure and attention mechanism were major breakthroughs in the development process of deep learning. It had been applied in many fields, showing great superiority. The possibility of simultaneous application in the fault diagnosis was indicated. Szegedy et al. [42] achieved the best result by stacking the Inception network structure to form GoogLeNet for optimization analysis. Chen et al. [43] applied the attention mechanism to CNN and proposed a spectral analysis CNN (SA-CNN). Spatial attention and channel wise attention have been added to CNN to complete the task of fault diagnosis.

As mentioned in the above analysis, the methods of CNN were increasingly used in the fault diagnosis. Although these methods were more accurate than traditional machine learning methods, they also had their own limitations and disadvantages. The generalization ability of most of the previous CNN models was not strong. Therefore, they could only be applied to the parts under the current load. Once the load changed (such as from noise changes, work environment changes), the accuracy would be reduced. In addition, some network model designs had too many layers. On one hand, it was difficult to train or even suffer degradation when the network was deep. On the other hand, the deeper the network was, the more prone it was to overfitting. Therefore, the training set accuracy was high, while the test set accuracy was low. Compared with pictures or other high-dimensional data, one-dimensional fault data were relatively concise and clear. The relatively simple models could be used to fully express the information behind the data. Moreover, the output stability of some network models was not ideal. The accuracy was difficult to maintain at a high level. The improvement of neural networks had become the focus of current research and development directions. As shown by ref. [44], the autoencoder was used to first denoise the signal. Then, it was input into the one-dimensional CNN model to realize fault diagnosis. However, two models needed to be trained. Also, the signal-to-noise ratio of the test signal was only 1 dB. As shown by ref. [45], a deep learning model with certain anti-noise properties was proposed. However, it could not processes the data. Also, the signal-to-noise ratio of the test signal was only 0 dB at the lowest. The strong noise interference below 0 dB was not studied, or the diagnosis effect was poor.

It can be seen from the above analysis that there are still many problems with the improvement and application of CNN in the fault diagnosis.

2.4 | Unbalanced sample classification based on single classification

The so-called unbalanced sample was that the number of samples in each category varies greatly or even there are missing samples. Sufficient training samples with a uniform category distribution were an important prerequisite for ensuring that the classification. When unbalanced dataset was used to pattern recognition, there would be the following problems: First, the network could not better learn the features of minority samples. Therefore, the accuracy rate of the classifier decreased to a few classes, and the classification performance decreased. The probability of misclassifying a few classes into many classes would be greatly increased. Then, it would be meaningless to take the accuracy rate as the index. For example, there were 99 normal samples and one fault sample. When the classifier returns all samples to the normal, its accuracy rate was still 99%. Finally, noise also had a great effect on the features of a few classes in unbalanced classification [46, 47]. Therefore, how to improve the fault diagnosis performance in the unbalanced samples was a key problem. At present, the single classification was the most researched on the unbalanced sample. The many normal samples were used for training to achieve better fault diagnosis performance. Wang et al. [48] introduced a one-class SVM into the field of network security and obtained better diagnosis performance in the traditional kernel method. Ma et al. [49] proposed a one-class SVM based on divergence for network fault diagnosis. They converted the original data into a new space according to the divergence representation of the data. A one-class SVM model was established in the space, and the data features were reduced by the PCA method. Chen et al. [50] proposed a one-class SVM that combines statistical density information. In this method, an undirected graph was constructed using normal data points. The sample ordering was obtained by computing the average KNN distance of the sample. Moreover, the sorting comparison results of the paired samples were fed back to the one-class SVM. In the diagnosis stage, the sorting threshold of the data samples was used to determine whether there were any faults.

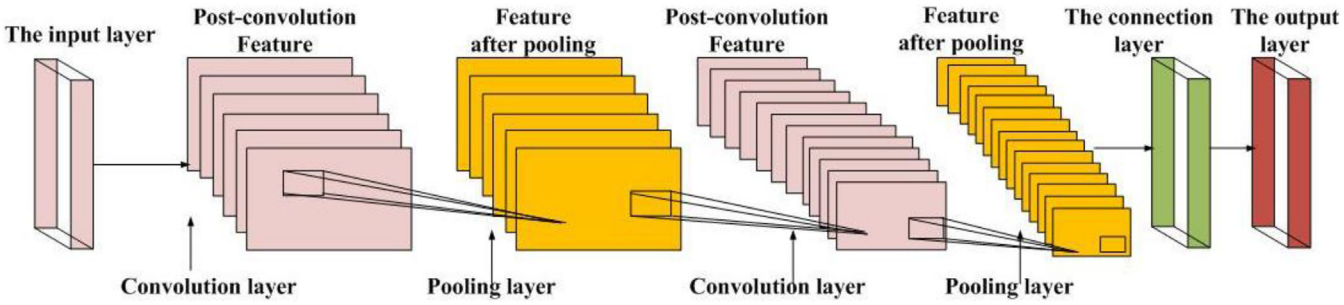

FIGURE 2 Structure of CNN.

3 | IMPROVEMENT METHODS

3.1 | Improved DCNN

The traditional CNN is generally composed of the input layer, conv layer, pooling layer, activation layer, density, and output layer. The essence of the CNN is to construct multiple filters that can extract features. Moreover, the input data are convolved layer by layer to extract the topological structure features hidden in the data. With the deepening of the network structure, the feature representation of input data is finally obtained. Meanwhile, sparse connections and parameter weight sharing mechanisms can be combined to reduce dimension sampling in time and space. Thus, the data dimension is reduced. Also, the number of training parameters are reduced. The algorithm overfitting is effectively avoided. It has invariance to scaling, tilting, translation, and other forms of transformation. Therefore, it is widely used.

Figure 2 shows a typical CNN structure. The CNN can be built by stacking multiple convolutions, activation, and pooling combinations. However, complex industrial environment diagnosis may lead to errors in training sets and weak generalization ability of new samples during the training process. Therefore, CNN has a poor ability to extract features. The training set deviation can be solved by expanding the training sample and selecting the appropriate pretreatment feature. It can also be solved by adding a neural network to the CNN to improve the training depth. However, the depth selection is not appropriate. And, the variance remains large. In addition, the structural design of the fully connected layer in CNN occupies too many computing resources. The test time of the CNN model in the fully connected layer is too long because of the excessive number of parameters. In turn, it is not conducive to rapid diagnosis.

To further illustrate the fault features in the railway point machine, the modal component diagram and frequency diagram of the sample during normal and fault are given in Figures 3, 4. Figure 3a shows that the sample signal is a normal state. The decomposed mode is relatively clear. Although there is noise interference, the noise amplitude is small. When in the fault state, the amplitude of signal noise interference is large (Figure 3b).

In turn, the obtained mode is basically submerged by high-frequency noise. The modes are prone to severe aliasing, and the

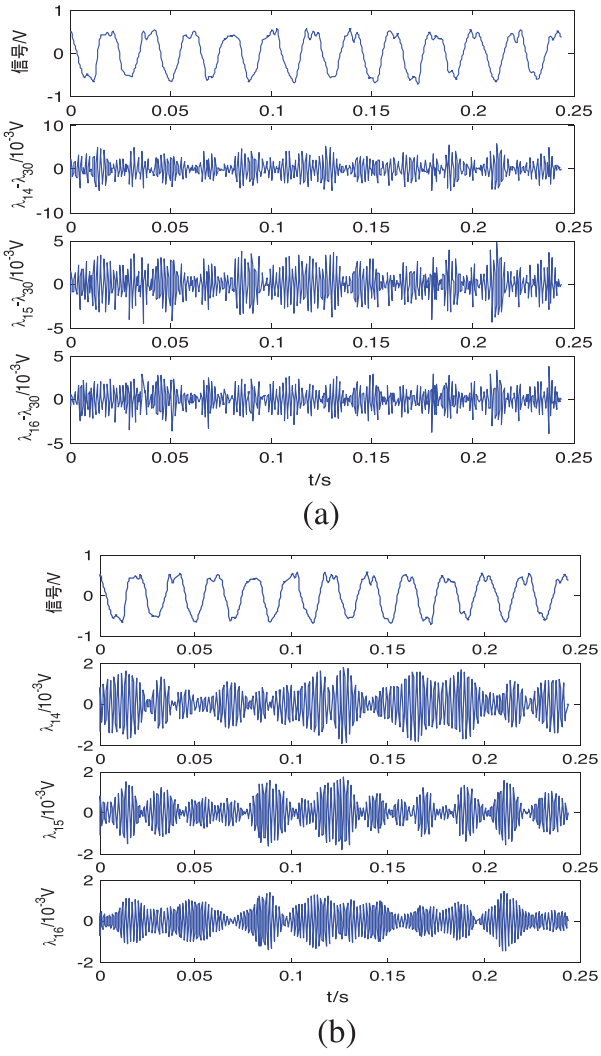

FIGURE 3 Comparison of sample modal components of the railway point machine. (a) The modal components of a normal sample (b) The modal component of the fault sample.

periodic pulse in the signal time domain is not obvious. Meanwhile, Figure 4a,b clearly shows that no interference component is in the small frequency band of the normal signal. The fault signal will be interfered with by the fundamental frequency.

Therefore, an improved DCNN structure (Figure 5) is proposed. We improve the effective performance of CNN in fault

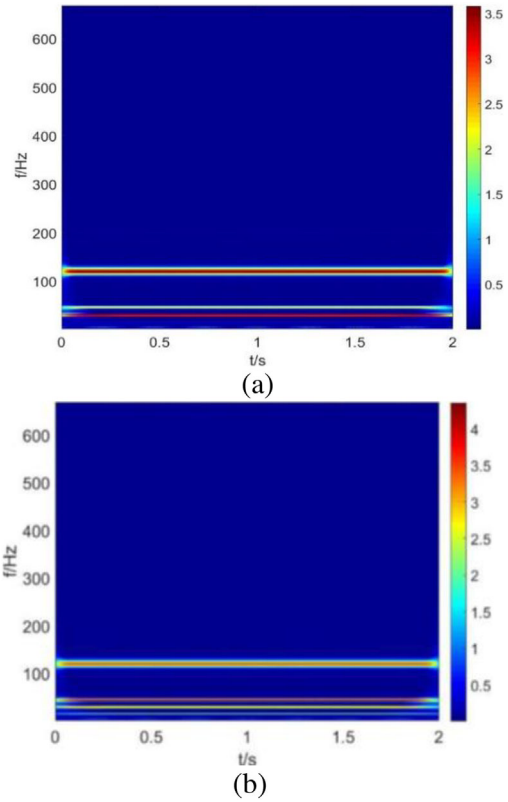

FIGURE 4 Time-frequency diagram. (a) Time-frequency diagram of normal signals (b) Time-frequency diagram of fault samples.

feature extraction by designing some optimization strategies in a single-layer CNN model. Specifically, it mainly includes the four steps. Step 1: After the input sample is normalized, the output is used as the input of the fault feature preprocessing to perform feature optimization processing. Step 2: After being processed, the output is used as the input of the Xception for feature extraction. Meanwhile, variable Dropout is used to further optimize the model and reduce deep network defects. Step 3: The AdaBN is performed after each layer of convolution operation. Meanwhile, the GAP is used instead of the fully connected layer to improve the generalization ability of the network. Step 4: After L2 regularization and Adam optimization, the features are output. Further, we design improved DCNN and SVDD models for complete fault diagnosis of ZDJ7 railway point machine under unbalanced samples. Therefore, we describe Xception structure optimization in Section 3.1.1, AdaBN strategy optimization in Section 3.1.2, GAP optimization in Section 3.1.3, and Adam optimization in Section 3.1.4.

3.1.1 | Optimization of Xception structure

After Lenet-5 is proposed, almost all model generalization capabilities of CNN is realized by deepening the depth of the network. Therefore, the possibility of overfitting and gradient disappearance will be greatly increased. The deeper the network layer is, the more difficult the optimization will be. Therefore, the Inception structure is introduced in terms of the composition of the convolution kernel. The features of various levels of the network are extracted using the convolution kernel of multiple scales on the same layer. In addition, the features between different convolutions are not cross-computed. In turn, the network computation can be reduced to a certain extent. However, the original Inception module has the shortcoming that the 5*5 convolution kernel in the structure requires too much computation.

Here, 1*1 convolution is used to reduce the dimension of the feature graph. In the end, the number of parameters decreases, and the expression ability of the network becomes superior. To further improve the overfitting situation, the Inception-V3 module is continued to divide the two-dimensional convolutional layer into two one-dimensional convolutional layers. After comprehensive analysis, the Inception-V3 model is set up as shown in Figure 6. The GAP is inserted behind the Concat layer in the model to improve the accuracy of feature extraction. GAP is used to reduce dimensionality by the pooling layer. The spatial features and contextual semantic information extracted from the previous convolutional and pooling layer are retained. However, the general convolution computation in the model is too large. In turn, it leads to a long training time.

To further improve the accuracy of the sample training, the Xception module is continued for optimization. The depthwise separable convolution (DSConv) [51] is adopted by the model to replace the original convolution operation in Inception-v3. Meanwhile, traditional convolution is decomposed into layer-by-layer and point-by-point convolution to reduce the amount of computation. In the layer-by-layer convolution, the input M channel feature maps are convolved with M filters. And, M feature maps are output. In the layer-by-layer convolution, the main focus is the extraction of spatial plane features. The convolution computation is only limited to the interior of each channel. Therefore, the information between channels is not fused. In point-by-point convolution, an eigen graph is output through the convolution layer by layer to output the M eigen graphs. In terms of computation amount, the amount of traditional convolution is as follows:

$$D_k \times D_k \times M \times N \times D_F \times D_F \tag{1}$$

D_k is the convolution size, and the D_F is the size of the input feature graph. M is the number of input feature graphs, and N is the number of output feature graphs. The traditional convolution is decomposed into a layer-by-layer and point-by-point. Moreover, the amount is the sum of the two, which is expressed as:

$$D_k \times D_k \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{2}$$

In (2), $M \times N \times D_F \times D_F$ is the amount of layer by layer. $D_k \times D_k \times M \times D_F \times D_F$ is the amount of point by point. Specifically, the computation amount difference before and

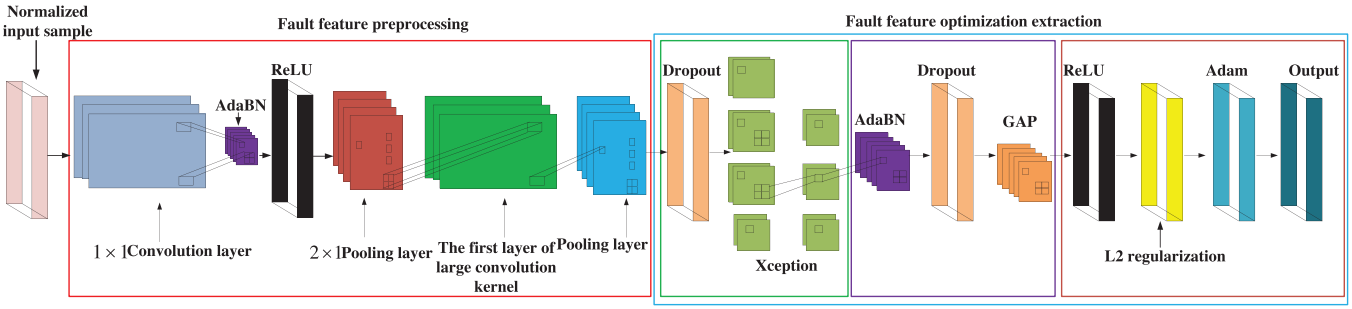

FIGURE 5 Improved DCNN.

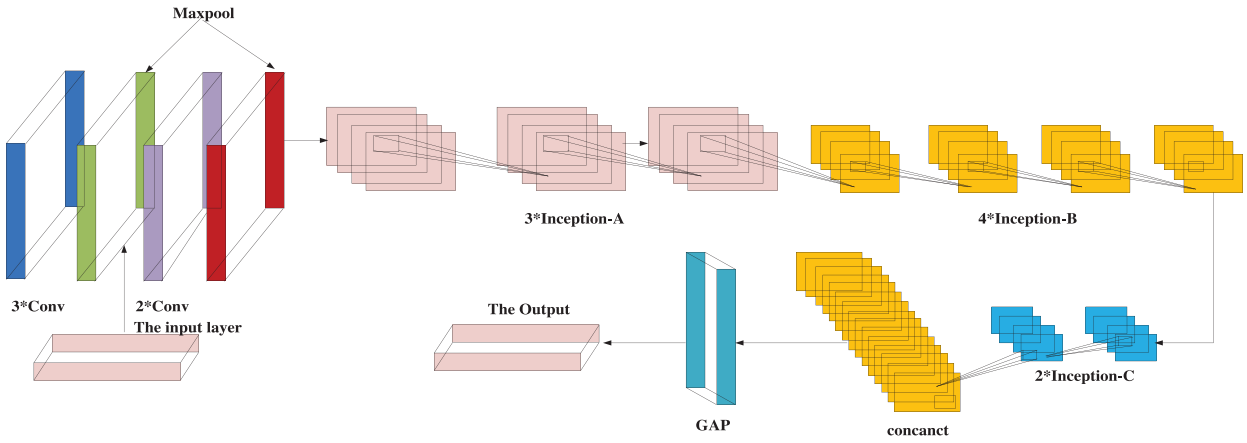

FIGURE 6 Inception-V3 model framework.

after optimization can be expressed as follows:

$$\frac{D_k \times D_k \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_k \times D_k \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_k^2} \tag{3}$$

It is seen that most of the computation is occupied by the optimized point-by-point convolution. There are N convolution kernels belonging to $1 \times 1 \times M$ in the point-by-point convolution. Also, the feature diagram of N channels is finally output. In this process, the feature fusion of the input feature graph at the channel level is realized. The internal implementation of traditional convolution is changed by DSConv. Therefore, the feature dimensions of the previous input and output can still be maintained externally. Specifically, the final design of the Xception structure is shown in Figure 7.

3.1.2 | Optimization algorithm based on AdaBN

When there is a difference between the distributions of training and test set, the generalization ability of the neural network will be greatly reduced. However, small changes accumulate continuously when passing through the multilayer network after normalization processing. Also, it leads to an excessive training loss rate. Therefore, a batch normalization (BN) layer [52] will be added to the convolutional layer and the activation layer. The same operation will be performed between the fully connected layer and the activation layer. Moreover, the main step is to subtract the mean value of the minibatch from the convolutional layer and the fully connected layer. In turn, the model can accelerate the training by the standard deviation. When the input value is limited to a narrow range, the expressive power of the network will be reduced. Therefore, BN processing can be added before the activation function to standardize the input of the activation function. The training efficiency of the network is improved. Although the generalization ability of the model can be effectively improved by BN, the fitting ability of the model is reduced to some extent. All data after operation are in the unsaturated region of the function, and the feature distribution will be destroyed. To recover the original data, the original feature distribution will be reconstructed. Also, the reconstruction process will introduce two learnable parameters γ and β, which can be expressed as:

$$\overset{*}{y}{}^{l(i)} = \frac{y^{l(i)} - \mu_B}{\sqrt{\left(\delta_B^2 + \varepsilon\right)}} \tag{4}$$

$$z^{l(i)} = \gamma^{l(i)} \overset{*}{y}{}^{l(i)} + \beta^{l(i)} \tag{5}$$

$$\mu_B = E_B\left[y^{l(i)}\right] = \frac{1}{m}\sum_{k=1}^{m} y_k^{l(i)} \tag{6}$$

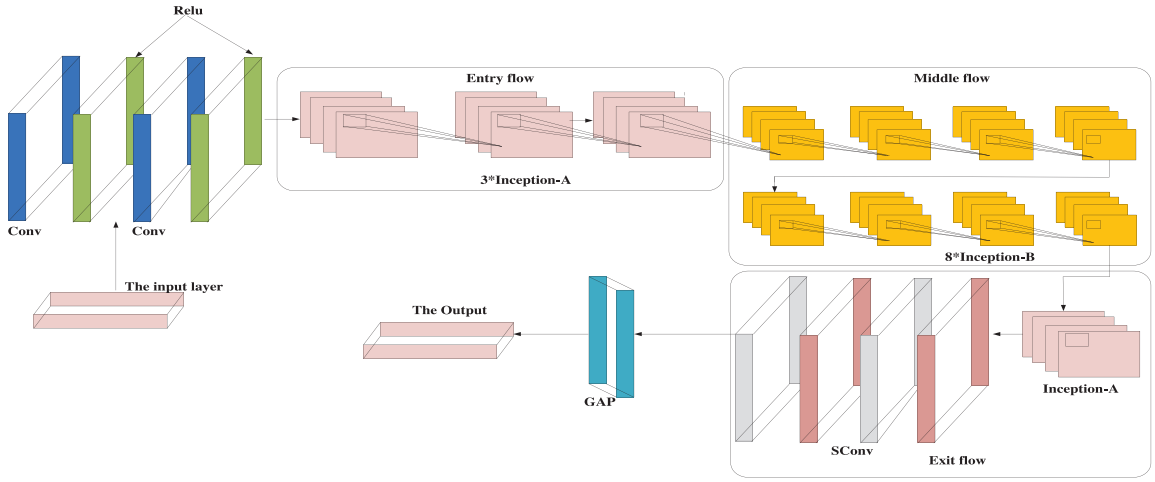

FIGURE 7 Structure of Xception model.

$$\delta_B^2 = \mathrm{Var}\left[y^{l(i)}\right] = \frac{1}{m}\sum_{k=1}^{m}\left(y_k^{l(i)} - \mu_B\right)^2 \tag{7}$$

In (4)–(7), $x^{l(i,j)}$ represents the output of the BN layer. $\gamma^{l(i)}$ represents the scaling of the BN layer. $\beta^{l(i)}$ represents the bias of the batch normalization layer, and ε is a constant.

When the distributions of the test and training samples substantially differ, the performance of the CNN model will decline. To suppress this decline, the method of AdaBN has been proposed to improve the field adaptability of improved DCNN. The AdaBN is based on the BN domain adaptive algorithm, which uses the μ_t and σ_t^2 of the target domain samples at each BN layer. The μ_s and σ_s^2 are computed from the source domain sample. Moreover, BN can reduce internal covariates. The source and target domain are adjusted to a new distribution space through AdaBN's effect on target domain samples. In this space, the approximate consistency enables the domain adaptive purpose to be achieved. Furthermore, the adaptive capacity of the improved DCNN is enhanced under the circumstance of noise and load changes. Specifically, the AdaBN algorithm is as follows:

Algorithm 1: AdaBN algorithm

Input: The signal ρ in the target field and the expression $x_t^{(i)}(\mathrm{p}) \in \mathrm{x}_\mathrm{t}^{(i)}$ of the i neuron in the BN; $\mathrm{x}_\mathrm{t}^{(i)} = \{\mathrm{x}_\mathrm{t}^{(i)}(1), \ldots \mathrm{x}_\mathrm{t}^{(i)}(\mathrm{n})\}$ the scaling and translation parameters $\gamma_\mathrm{S}^{(i)}$ and $\beta_\mathrm{S}^{(i)}$ of the training are given for the i neuron; Output: Adjusted improved DCNN network

For Each neuron *i* and each signal p in the target domain

Calculate the mean and variance of all samples in the target field: $\mu_\mathrm{t}^{(i)} \leftarrow \mathrm{E}[\mathrm{x}_\mathrm{t}^{(i)}]$, $\sigma_\mathrm{t}^{(i)} \leftarrow \mathrm{Var}[\mathrm{x}_\mathrm{t}^{(i)}]$

Calculate the output of the BN layer $x_t^{*(i)}(p) = \dfrac{\mathrm{x}_\mathrm{t}^{(i)}(\mathrm{p}) - \mu_\mathrm{t}^{(i)}}{\sigma_\mathrm{t}^{(i)}}$,

$\mathrm{y}_\mathrm{t}^{(i)}(\mathrm{p}) = \gamma_\mathrm{S}^{(i)} x_t^{*}(i)(\mathrm{p}) + \mathrm{fi}_\mathrm{S}^{(i)}$

End for

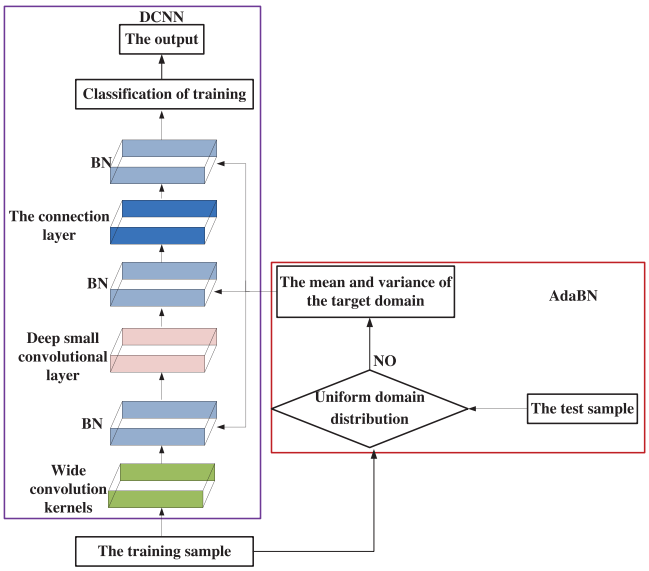

FIGURE 8 AdaBN optimized DCNN flowchart.

The flowchart of DCNN optimized by AdaBN is shown in Figure 8. First, training samples are used to completely train the model. If the domain distribution of the training set is inconsistent with that of the test set, the test set is input into the model. Only the data are propagated forward. Moreover, the mean variance of all BN layers is replaced by the mean variance of the test set. Moreover, other network parameters remain unchanged. The domain adaptive process is completed, and the improved DCNN can be used to diagnose each sample of the test set.

3.1.3 | GAP optimization

The GAP is a new technique for solving the problem of too many parameters. It is similar to the pooling layer operation in

taking the global average value of the feature output from the last convolution operation. The GAP is directly connected to the classifier in the improved DCNN. The preceding layer is the convolutional layer of the CNN. For the classification problem of n, the output dimension of the kernel in the last convolutional layer can be set as n. Thus, n output feature maps are obtained. And then, n global average pooling cores are adopted. The size and step size of each pooling core are identical to the size of the output feature map. Then, the average value of each pooling check should be computed to generate a value for each feature map. The n value equal to the output effect of the fully connected layer can be obtained. Finally, it is input to the Softmax classifier. Specifically, the expression for GAP is expressed as:

$$S^l_{\text{avg-pooling}} = \frac{1}{c}\sum_{i=1}^{c} X^l_{1:h,1:w,i} \tag{8}$$

In (8), $S^l_{\text{avg-pooling}}$ represents the result obtained by global average pooling in layer l. $X^l_{1:h,1:w,i}$ represents the range of the output feature map corresponding to the average pooling kernel from the pixels in the horizontal direction from row $1-h$. In the dropout algorithm application, a neuron is set to stop working. In other words, the activation value of neuron nodes can be "discarded" with fixed probability. In turn, it can effectively restrain the overfitting in the neural network. The random zero of the training sample can add noise to the input signal. During each minibatch training, the variation dropout rate is subject to a uniform distribution of 0.1–0.8. The dropout rate is set to 1 in every 5 training epochs. According to the back-propagation formula and chain derivation rule, the weight partial derivative formula of objective function L with respect to the convolution kernel of the first layer and other layers can be expressed as:

$$\begin{cases} \dfrac{\partial L}{\partial K_i^l(j^*)} = \displaystyle\sum_{k=1}^{m}\sum_{j}\left(\frac{x^{l(j)}}{p}\right) * \left(\frac{\partial L}{\partial y_k^{l(i,j)}}\right), & l = 1 \\ \dfrac{\partial H}{\partial K_i^l(j^*)} = \displaystyle\sum_{k=1}^{m}\sum_{j}\left(x^{l(j)}\right) * \left(\frac{\partial L}{\partial y_k^{l(i,j)}}\right), & l > 1 \end{cases} \tag{9}$$

In (9),$\tilde{K}_i^l$ is the weight of the convolution kernel after variation dropout rate.

3.1.4 | Adam optimizes the training network

The loss function and gradient computation of the traditional gradient descent method are performed on the entire data set scale. The problem that the computation is too large, and the memory is difficult to carry. Therefore, Adam [53] is used to optimize the network training. The first and second moment estimates of the gradient are computed to design independent adaptive learning rates for different parameters. Specifically, parameters are updated by computing the gradient exponential moving mean. The attenuation rate of these moving means is controlled using superparameters. Adam's updating criteria can be expressed as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{10}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{11}$$

$$\tilde{m}_t = \frac{m_t}{1-\beta_1^t} \tag{12}$$

$$\tilde{v}_t = \frac{v_t}{1-\beta_2^t} \tag{13}$$

$$w_{t+1} = w_t - \frac{\gamma}{\sqrt{\tilde{v}_t} + \varepsilon}\tilde{m}_t \tag{14}$$

In (10)–(14), m_t and v_t represent the first and second order moment estimates of the gradient. $\tilde{m}_t$ and $\tilde{v}_t$ are the correction terms of m_t and v_t. m_t and v_t are corrected by Adam. In turn, the iteration of the learning rate is controlled within a relatively clear range.

3.2 | Unbalanced classification based on SVDD parameter optimization

To ensure that the training model has high accuracy and reliability, it is generally assumed that the training and the test set is independent and identical distributions. Meanwhile, the training set are evenly distributed. Limited by the poor collection conditions, the number of training set has difficulty meeting the above assumptions. Therefore, the accuracy and reliability of the classification model are poor. In the unbalanced sample classification, the support vector data description (SVDD) is used to construct a single class of hyperspheres for fault features in this paper. An improved quantity learning algorithm for hypersphere coordinate mapping is proposed. The classification of unbalanced samples are realized by the increment of the hypersphere. Moreover, the superset of the support vector set as small as possible is constructed by extracting boundary samples based on SVDD. In turn, the information redundancy is reduced to improve the training accuracy. Therefore, we describe Penalty coefficient analysis in Section 3.2.1, SVDD sphere based on the original centre in Section 3.2.2.

3.2.1 | Penalty coefficient analysis

In the specific classification, the default linear kernel function is selected to take different values of the penalty coefficient C. The classification effect is compared. Moreover, the value range of the SVDD penalty coefficient is 0–1. Therefore, a classification result with a feature dimension of 200 is first presented.

Specifically, the accuracy is defined as follows:

$$\text{accuracy} = \begin{cases} \dfrac{\underset{Z_i \in C_n, n=1,\ldots k}{\text{num}} \left(|Z_i - \text{a}_j| \le R_j \right),}{k}, & i = j \\ \dfrac{\underset{Z_i \in C_n, n=1,\ldots k}{\text{num}} \left(|Z_i - \text{a}_j| > R_j \right)}{\text{k}}, & i \neq j \end{cases} \tag{15}$$

In (15), i is the class of the test set. j is the class of the training set. C_n is the test set. Z_i is the sample of the test set. k is the number of samples. $num()$ is the total number of samples. a_j is the centre of the hypersphere. R_j is the radius of the hypersphere.

3.2.2 | SVDD sphere based on the original centre

In the SVDD model, the support vector is abstract model knowledge and contains the complete information of the training model. When solving hyperspheres α, the Karush Kuhn Tucker conditions (KKT) condition is the necessary and sufficient condition for the optimal solution of non-linear programming problems.α is the optimal hypersphere if and only if all samples satisfy the KKT condition, which can be expressed as:

$$\begin{cases} a_{\text{i}} = 0 \rightarrow |f_{x(i)}| \ge 1 \\ 0 < a_i < \dfrac{1}{vl} \rightarrow |f_{x(i)}| = 1 \\ \text{a}_{\text{i}} = \dfrac{1}{vl} \rightarrow |f_{x(i)}| \le 1 \end{cases} \tag{16}$$

In (16), $a_i$is the Lagrangian vector corresponding to sample x_i.v $\in [0, 1]$ is the upper bound of the proportion of error samples in the total sample. l is the total number of samples. When $a_i = 0$, the corresponding sample is located outside the supersphere. When $0 < 1/vl$, the corresponding sample is located on the hypersphere. When $a_i = 1/vl$, the sample is in the hypersphere. After the preliminary screening of samples, the redundant information in the remaining samples should be judged. In SVDD model training, redundant information comes from a certain amount of sample aggregation. Therefore, the cluster non-boundary samples have no effect on the classification model training. The number of samples and the degree of sample aggregation can be used as indicators to evaluate the redundant information. Moreover, the degree of sample aggregation is indicated by the Hopkins statistical scale. Specifically, the computation steps are as follows:

1. First, $n$points $(p_1, p_2, \cdots p_n)$ are uniformly extracted from the sample set D. The nearest neighbour of each point p_i in D is found. Moreover, the nearest neighbour distance between p_i and D is x_i. Therefore, the computation can be expressed as:

$$x_i = \min_{v \in D} \left\{ \text{dist}(p_i, v) \right\} \tag{17}$$

2. In a similar way, n points $(q_1, q_2, \cdots q_n)$ are uniformly extracted from the sample set D. The nearest neighbour of each point q_i in $D - q_i$ is found. Moreover, the nearest neighbour distance between q_i and D is y_i. The computation can be expressed as follows. Specifically, Hopkins statistic y_i is computed:

$$\text{y}_i = \min_{v \in D - \{q_i\}} \{\text{dist}(q_i, v)\} \tag{18}$$

3. Hopkins statistic H is computed:

$$\text{H} = \frac{\sum_{\text{i}=1}^{\text{n}} \text{y}_{\text{i}}}{\sum_{\text{i}=1}^{\text{n}} \text{x}_{\text{i}} + \sum_{\text{i}=1}^{\text{n}} \text{y}_{\text{i}}} \tag{19}$$

If the sample D is uniformly distributed, then $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} y_i$ will be very close. H will be approximately 0.5. If D is highly clustered, then $\sum_{i=1}^{n} y_i$ will be substantially less than $\sum_{i=1}^{n} x_i$. Hwill be close to 0. Therefore, Hopkins statistics can be used to evaluate the clustering trend of samples in space. Moreover, the classification process based on the SVDD is shown in Figure 9. A boundary extraction is used to solve the information loss lower boundary extraction from a single centre. The observation centre is set in the entire observation area. Also, the sample information of the boundary is analysed by setting different radii of the hypersphere. Moreover, the adaptability of the algorithm to different samples is improved. The specific steps are as follows: First, the sample is observed according to the size of the sphere radius in the SVDD sphere based on the original centre. Then, the observation points of the samples are analysed to determine the radius of the hypersphere with the minimum loss of boundary information. Finally, test samples are selected for boundary extraction to verify the effectiveness of the multicentre method in actual diagnosis.

4 | EXPERIMENT AND RESULT ANALYSIS

4.1 | Datasets and evaluation indicators

These datasets are collected from the historical monitoring data of the current power of the ZDJ7 AC railway point machine in a subway depot in Chengdu. Moreover, the test rig is shown in Figure 10.

The sampling frequency of the monitoring equipment is 40 ms. Moreover, the point manoeuvre time under normal conditions is 5.8 s, which corresponds to approximately 145 points. In addition, a set of time series data is composed. To correspond to 11 types of faults in the railway point machine, 800 training sets are selected for each type of fault in the selection process of

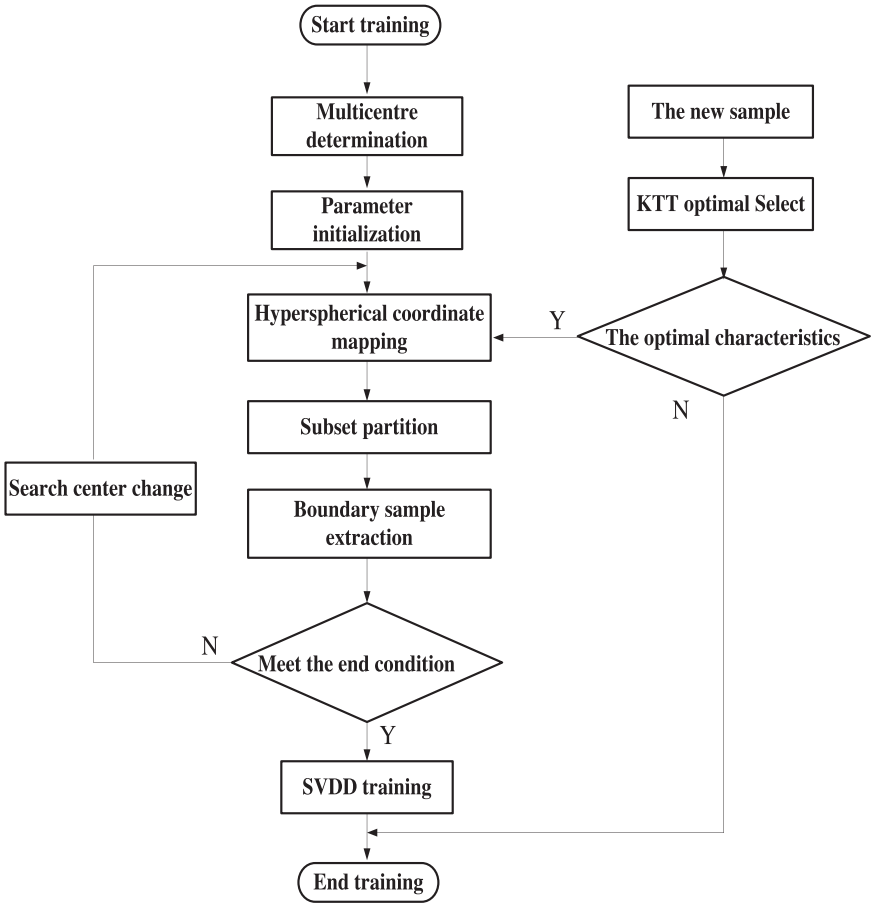

FIGURE 9 Classification process.

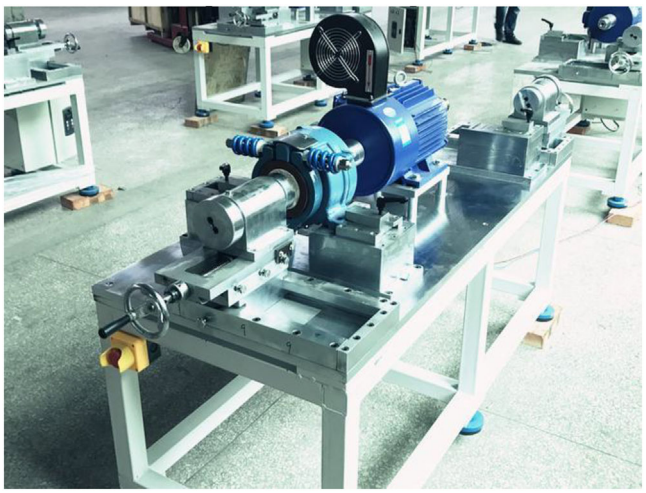

FIGURE 10 ZDJ7 AC railway point machine test rig.

TABLE 2 Setting of training dataset

Label	0	1	2	3	4	5
Size	800	800	800	800	800	800
Label	6	7	8	9	10	11
Size	800	800	800	800	800	800

TABLE 3 Test data set settings

Label	0	1	2	3	4	5
Size	200	105	52	46	116	92
Label	6	7	8	9	10	11
Size	51	47	53	86	48	44

training sets (the normal samples corresponding to each type of fault were consistent). Therefore, a total of 11 types of common high-frequency faults are selected. The labels 0–11 are added for these 11 types of faults and the normal state. Each dataset includes a time series of 4 dimensions of three-phase currents A, B, C, and the total power. The training set (Table 2) is enhanced by generating an antagonism network. Each class is set to contain 800 pieces of data. The test set is composed of the collected set, and composition is shown in Table 3.

In the experiment, accuracy (A), precision (P), recall (R), false positive rate (FPR), F-score, and AUC are used to test the method. The parameter computation and definitions are referred to in ref. [10]. A is expressed as Equation (20). The higher its value is, the better the overall performance of the classifier will be. P is defined in (21). The higher its value is, the lower the FPR of the classifier will be. The definition of R is shown in (22). The higher its value is, the lower the classification

TABLE 4 Specific parameters of Xception structure

Number	Layers	Model structure	Note
1	0	–	Training all layers
2	36	Entry flow	Freeze the entry flow module and the preceding layer
3	56	Middle flow(2)	Freeze the second middle flow module and preceding layers
4	76	Middle flow(4)	Freeze the second middle flow module and preceding layers
5	96	Middle flow(6)	Freeze the second middle flow module and preceding layers
6	116	Middle flow(8)	Freeze the second middle flow module and preceding layers
7	132	Exit flow	Freeze exit flow module and preceding layers

TABLE 5 Improved DCNN structure specific parameters

Number	Layers	Model structure	Step	Kernel size	Output size
1	Convolution layer 1	64 × 1/16 × 1	2	16	128 × 16
2	Pooling layer 1	2 × 1/2 × 1	1	16	64 × 16
3	Dropout layer1	Dropout(0.2)	–	–	–
3	Convolution layer 2	3 × 1/1 × 1	2	32	64 × 32
4	Pooling layer 2	1 × 1/2 × 1	1	32	32 × 32
5	Dropout layer2	Dropout(0.1)	–	–	–
6	Global average pooling layer	1 × 5/4 × 1	4	4	4 × 32

loss rate is. Equation (23) gives the FPR. The larger its value is, the worse the classification performance will be. Equation (24) gives the harmonic average value of P and R. The larger its value is, the better the classifier can detect the fault.

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

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

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

$$FPR = \frac{FP}{TN + FP} \tag{23}$$

$$F\text{-}score = \frac{2 * P * R}{P * R} \tag{24}$$

4.2 | Comparison of improved DCNN training epochs

To verify the influence of parameters on the diagnosis model, the training method of the CNN model is referred. Moreover, the fault set of the railway point machine are used for training comparison. The influence of training epochs is primarily compared. Considering the length of the paper and the importance of this section for the whole research work, we do not give experiments on the choice of parameters for the improved DCNN structure in this study. The parameters of the Xception module in the improved DCNN in the experiment are shown in Table 4. The overall parameter settings of the improved DCNN are shown in Table 5. The improved DCNN model contains convolution and pooling layers, and the convolution kernel of the first layer is 64 × 1. The number of neurons in the hidden layer is 100. The GAP is used to replace the fully connected layer in the experiment. The convolution kernel size is 4 × 1.

The comparison curve of the loss values of the training sets under different training epochs is shown in Figure 11a–f. Meanwhile, the loss values of the validation and the test set are also compared. When the overall number of epoch is set to 15, 20 and 25, the loss value of the training set decreases with the increasing number of epochs. However, the optimal number of overall epoch is 15 under the three training models. Also, the training loss is 0.024875. Moreover, the convergence of the curve is good. It is difficult for the model to classify the samples correctly because of the small number of epochs. With the increase in epochs, the loss value will increase because of overfitting. When the overall number of training epoch is set to 30, 35 and 40, the loss value of the training set also decreases with the increasing number of epochs.

Moreover, the optimal number of overall training epoch is 30 under the three training models. Also, the training loss is 0.014653. The convergence of the curve is good for the same reason as above. When the number of training epoch is set to 30, the training output is the best. The error rate is the lowest.

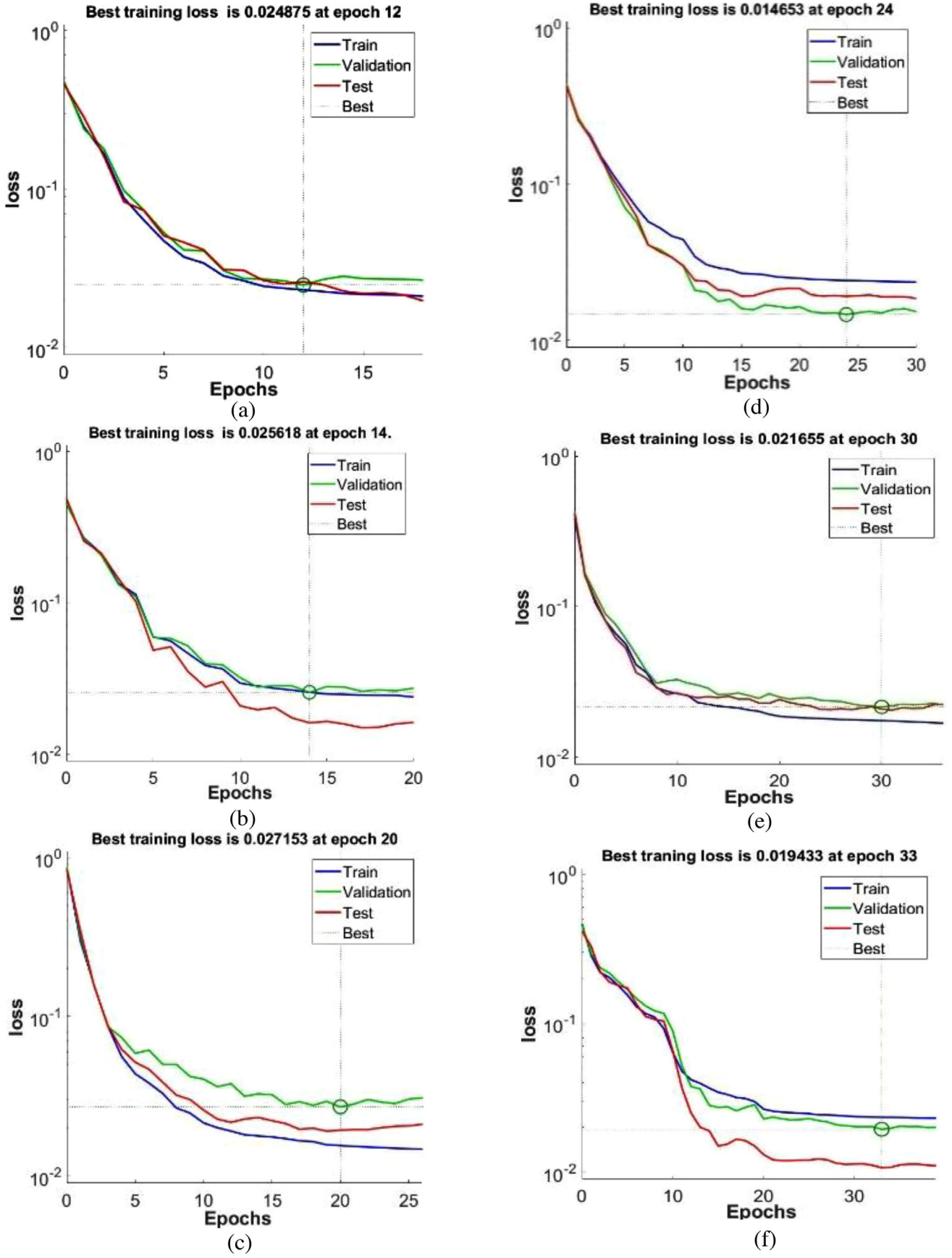

FIGURE 11 Comparative analysis of training epochs. (a) The epoch is 15. (b) The epoch is 20. (c) The epoch is 25. (d) The epoch is 30. (e) The epoch is 35. (f) The epoch is 40.

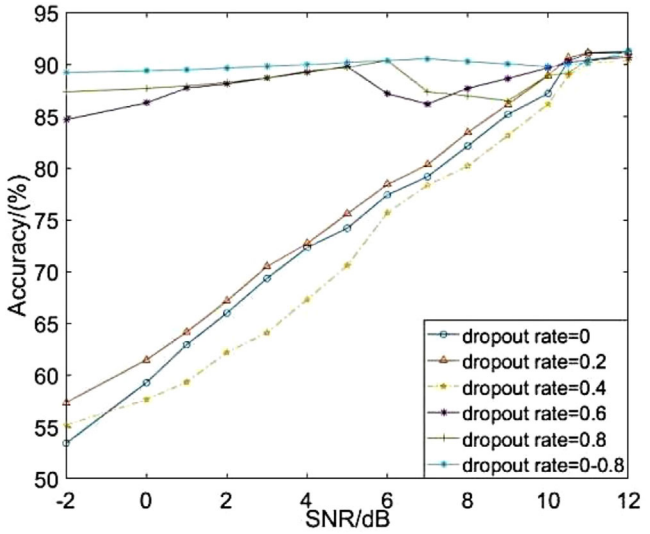

FIGURE 12 Accuracy comparison of variation dropout rate.

With this experiment, we first validate the excellent training performance of our method on the dataset.

4.3 | Parameters setting experimental verification

4.3.1 | Variation dropout rate setting verification

To verify the accuracy of the fault diagnosis is affected by variation dropout rate, a single dropout rate input between 0–0.8 and the variation dropout rate are selected to train the model. Moreover, the training accuracy is shown in Figure 12.

The conclusions obtained from Figure 12 are as follows: When the variation dropout rate is used for training, the model test results are good. Its accuracy is higher than the model with a single dropout rate. The results show that the variation dropout rate plays a certain role in improving the performance of the model.

4.3.2 | Training set setup validation

To verify the influence of the optimization process and dataset on training, the improved DCNN model is used under different proportions of the training and test set (70% training set + 30% test set, 75% training set + 25% test set). The comparison of training loss rate after regularization of L1 and L2 with or without Adam optimization is given in Figure 13a–d.

The following conclusions can be drawn from the Figure 13: First, the training loss after L2 regularization is small and within a relatively stable range. When L1 is used for regularization, the training loss is large. The floating range is wide. Therefore, the training stability is poor. When the model is optimized without the Adam method, the training loss of the L1 and L2 regularization methods is relatively large. After optimization with the Adam method, the training performance is greatly improved.

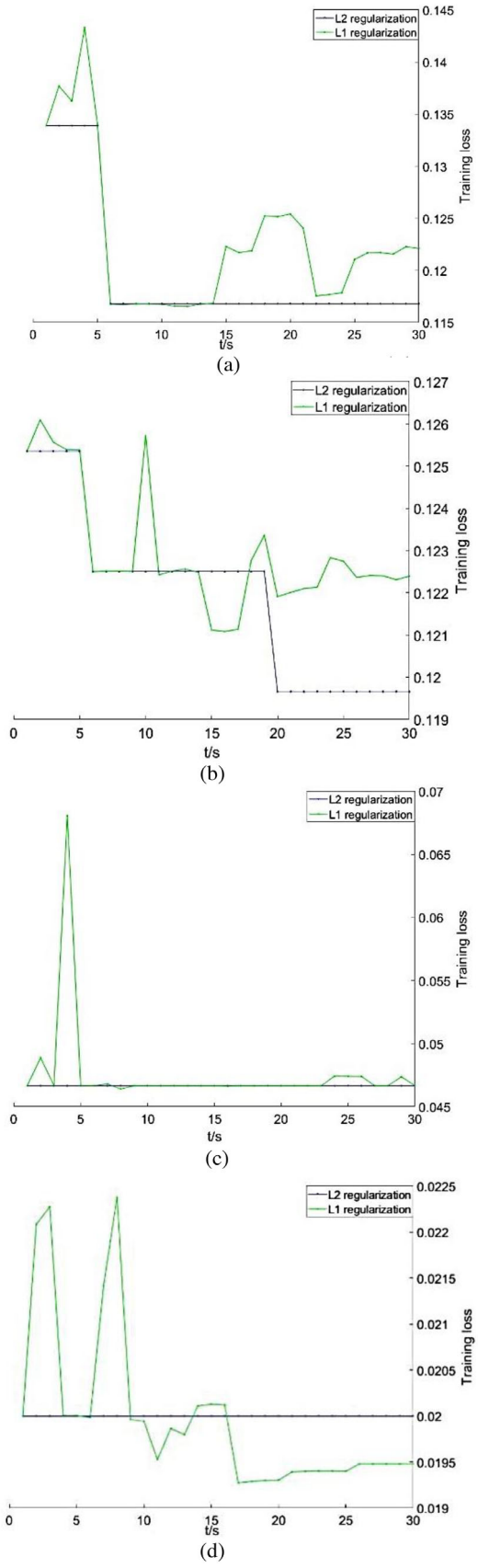

FIGURE 13 Optimization analysis. (a) Training set 70% + test set 30% + no Adam. (b) Training set 70% + test set 30% + Adam. (c) Training set 75% + test set 25% + no Adam. (d) Training set 75% + test set 25% + Adam.

TABLE 6 Accuracy under different penalty coefficients

	Penalty coefficient *C*				
Result	0.001	0.01	0.1	0.15	0.9
Normal	35.68%	80.09%	**96.18%**	93.10%	90.64%
Fault	94.23%	91.97%	98.33%	69.76%	59.26%

Therefore, the necessity of adding the Adam method is proved. When the training set is 75% and the test set is 25%, the training loss is the smallest. The loss of L2 regularization is approximately 0.2. Although the loss of L1 regularization decreases to 0.0195 with increasing time, the overall training process is unstable. Therefore, the end result is not good. Through the above experiments, we have verified the necessity of Adam's optimization strategy and the effect of dataset distribution on the training results.

4.3.3 | Penalty coefficients setup validation

To verify the effect of penalty coefficient C on SVDD performance, we set up the following experiments. The overall accuracy is compared under different penalty coefficients in Table 6. The conclusions obtained from Table 6 are as follows: When the C is too small, the classification results of normal samples are poor. When the penalty coefficient is too large, the classification results of fault samples are poor. When C is 0.1, the classification accuracy of normal and fault samples is the highest. Therefore, 0.1 is adopted in this paper for subsequent experiments.

4.3.4 | Verification of the noise resistance of the first convolution kernel

To verify the anti-noise performance of the improved DCNN model and the AdaBN method, the two comparative experiments are given in Figure 14a,b.

When the improved DCNN is not optimized by the AdaBN method, the highest variance of the diagnosis data is ≈0.35–0.45. When the time and learning rate are continuously increasing, the mean diagnosis variance is approximately 0.2. When the learning rate continues to increase, the diagnosis variance begins to increase again. When the improved DCNN is optimized by the AdaBN method, the highest variance of the diagnosis data is ≈0.30–0.35. When the time and learning rate are continuously increasing, the mean diagnosis variance is approximately 0.1. When the learning rate continues to increase, the diagnosis variance does not change considerably. The optimal performance of the AdaBN method for the improved DCNN is verified.

To further verify the stability of the improved DCNN, the accuracy of the SNR value changing from −4 to 10 dB is tested. To verify the necessity of the first-layer large convolution kernel in the improved DCNN model, the diagnosis rates of the improved DCNN and AdaBN under different intensity band noises. Specifically, the first-layer convolution kernel changes from 16 to 128 are also given. The results are shown in Tables 7, 8.

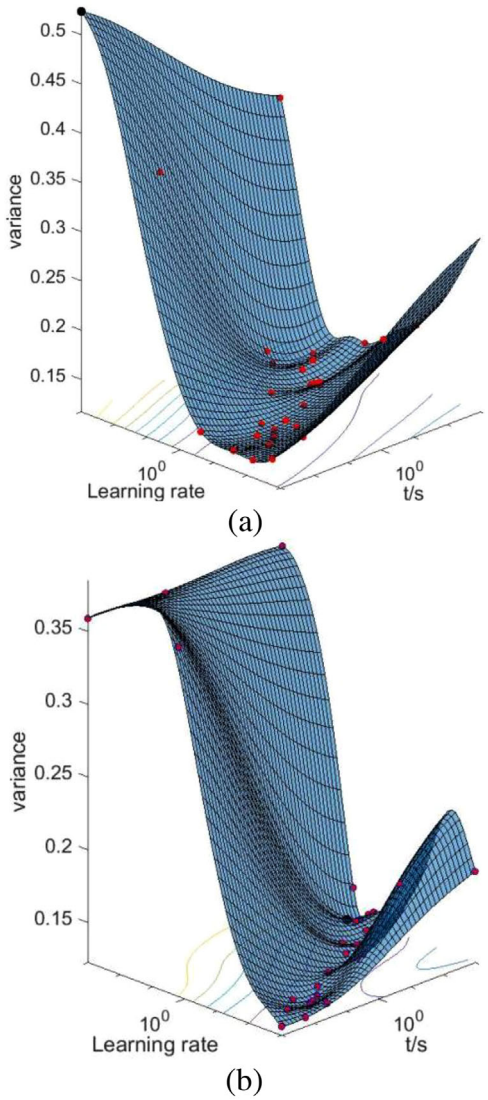

FIGURE 14 Comparative analysis. (a) Unused AdaBN, (b) AdaBN optimization.

The conclusions obtained from Table 7 are as follows: In the low SNR environment, the improved DCNN has a higher accuracy rate when the convolution kernel is large. When the first convolution kernel is 112×1, the accuracy of the improved DCNN model when SNR = 0 dB can reach more than 90%. When the size is 16 × 1, the diagnosis rate is only 57.49%. In addition, the accuracy rate is low in the case of SNR < 0 dB when the size is 128 × 1. When the convolution kernel is too small, the features of low and medium frequencies are difficult to capture. In turn, fault diagnosis is easily disturbed by high-frequency noise. When the convolution kernel is large, the missing detailed features are caused

TABLE 7 Accuracy of improved DCNN under different noise intensity

	SNR (dB)							
Kernel size	**−4**	**−2**	**0**	**2**	**4**	**6**	**8**	**10**
16	28.32%	40.78%	57.49%	75.81%	84.52%	92.37%	96.54%	98.17%
24	33.68%	47.66%	74.32%	82.64%	85.29%	92.59%	96.88%	98.44%
32	39.27%	49.23%	75.68%	83.36%	86.93%	92.97%	96.97%	98.68%
40	44.16%	53.55%	79.27%	83.95%	87.88%	94.12%	97.39%	98.81%
48	47.64%	58.64%	82.14%	84.64%	90.24%	96.09%	97.48%	99.25%
56	52.31%	64.28%	83.63%	85.92%	93.67%	97.18%	97.77%	**99.62%**
64	52.93%	65.37%	85.66%	88.67%	95.25%	97.41%	98.62%	99.24%
72	53.62%	66.92%	87.29%	90.28%	96.73%	97.75%	98.94%	99.14%
80	56.03%	69.49%	88.93%	92.57%	97.14%	98.47%	99.18%	99.17%
88	56.01%	71.75%	89.92%	95.18%	97.35%	98.22%	99.24%	99.32%
96	64.79%	78.41%	90.36%	96.51%	97.89%	98.81%	**99.49%**	**99.62%**
104	62.97%	79.74%	90.78%	**97.25%**	**98.40%**	**99.26%**	99.27%	99.20%
112	**67.82%**	**81.29%**	**91.60%**	96.37%	97.32%	99.11%	99.12%	99.09%
120	61.26%	76.33%	88.68%	96.75%	97.25%	98.72%	99.05%	99.13%
128	60.88%	77.85%	89.34%	96.99%	97.01%	99.04%	**99.49%**	99.08%

TABLE 8 Accuracy of improved DCNN (AdaBN) under different noise intensities

	SNR (dB)							
Kernel size	**−4**	**−2**	**0**	**2**	**4**	**6**	**8**	**10**
16	81.37%	90.18%	93.17%	96.33%	97.11%	98.29%	99.01%	99.28%
24	85.29%	90.27%	93.49%	96.57%	97.26%	98.34%	99.04%	99.32%
32	87.78%	90.41%	93.82%	96.91%	97.43%	98.47%	99.09%	99.36%
40	88.92%	90.55%	94.39%	97.28%	97.86%	98.65%	99.24%	99.42%
48	90.27%	90.62%	94.61%	97.84%	98.25%	98.79%	99.31%	99.45%
56	91.83%	92.66%	95.18%	98.15%	98.67%	98.91%	99.37%	99.53%
64	92.34%	93.74%	95.84%	98.43%	98.82%	99.18%	99.39%	**99.78%**
72	92.67%	94.80%	96.07%	98.52%	99.04%	99.24%	99.41%	99.49%
80	93.21%	95.12%	96.65%	98.79%	99.06%	**99.39%**	99.42%	99.52%
88	94.54%	95.64%	97.18%	99.04%	99.14%	99.35%	99.44%	99.54%
96	**95.39%**	**96.31%**	**97.94%**	**99.18%**	**99.21%**	99.26%	99.52%	99.56%
104	95.21%	96.10%	97.53%	99.05%	99.13%	99.28%	99.53%	99.58%
112	94.40%	96.01%	97.56%	99.06%	99.15%	99.31%	**99.55%**	99.61%
120	94.67%	95.75%	97.60%	99.14%	99.17%	99.26%	99.48%	99.64%
128	95.11%	95.96%	97.71%	99.08%	99.19%	99.24%	99.41%	99.69%

by the decrease in temporal resolution. Moreover, the accuracy decreases rather than increases. Therefore, a larger convolution kernel should be selected when the noise is larger. The width of the convolution kernel should be at least twice the step length.

Table 8 shows that AdaBN can greatly improve the accuracy of the improved DCNN model. The accuracy of the improved DCNN (AdaBN) is over 90% when SNR≥ -4 dB. The accuracy of the SNR $= -4$ dB sample is $\approx$30% higher than that of the improved DCNN model. When the convolution kernel is small, the accuracy is substantially increased by the improvement. The validity of AdaBN for the improved DCNN model optimization is verified. Better performance is achieved in noise fault samples when the AdaBN structure is added to the DCNN model. This also shows again the effectiveness of our proposed strategy for CNN improvement.

TABLE 9 Comparison of diagnosis indexes under unbalanced sample combinations

		Method					
Samples	**Indexes**	**DBPN+SVDD**	**CNN3+SVDD**	**RNN3+SVDD**	**DBN3+ SVDD**	**One-class SVM**	**Improved DCNN+SVDD**
normal(450) +fault(50)	G-mean	0.627	0.635	0.658	0.695	0.766	**0.784**
	MAUC	0.654	0.649	0.674	0.716	0.774	**0.789**
normal(450) +fault(100)	G-mean	0.635	0.652	0.676	0.738	0.781	**0.787**
	MAUC	0.661	0.675	0.681	0.759	0.795	0.893
normal(450) +fault(150)	G-mean	0.723	0.683	0.692	0.760	**0.796**	0.792
	MAUC	0.754	0.726	0.738	0.798	0.813	**0.819**
normal(450) +fault(200)	G-mean	0.796	0.784	0.714	0.786	0.819	**0.823**
	MAUC	0.807	0.795	0.762	0.827	**0.837**	0.834
normal(500) +fault(100)	G-mean	0.618	0.613	0.735	0.813	0.838	**0.846**
	MAUC	0.634	0.626	0.781	0.838	0.855	**0.869**
normal(500) +fault(150)	G-mean	0.642	0.637	0.789	0.826	0.850	**0.867**
	MAUC	0.656	0.645	0.814	0.842	**0.877**	0.875
normal(500) +fault(150)	G-mean	0.749	0.724	0.812	0.847	0.862	**0.879**
	MAUC	0.763	0.746	0.836	0.861	0.879	**0.886**
normal(500) +fault(250)	G-mean	0.775	0.767	0.847	0.874	0.879	**0.897**
	MAUC	0.798	0.780	0.865	0.893	0.906	**0.923**

4.4 | Unbalanced sample classification analysis

To evaluate the proposed model, the common algorithms in unbalanced sample are also used to perform diagnosis tests on the data under different sample combinations. The G-mean and MAUC indexes under different normal and fault samples are primarily compared. G-mean is an index to comprehensively test the accuracy of the two categories. Only when the accuracy of both categories are high can a higher G-mean be obtained. MAUC is the average AUC of classification system. The larger two values are, the better classification will be. The algorithms include the DBPN+SVDD, CNN3+SVDD, RNN3+SVDD, DBN3+SVDD, One-classSVM, improved DCNN+SVDD (proposed), and the specific results are shown in Table 9.

The conclusions obtained from Table 9 are as follows: In terms of the G-mean performance, the value obtained by the proposed method is lower than that of the One-class SVM method except in the normal (450) + fault (150) combination. In the other combinations, the proposed method is best. In terms of MAUC performance, MAUC values are lower than those of the One-class SVM method except for normal (450) + fault (200) and normal (500) + fault (150) combinations. In the other combinations, the proposed method is best. The two dimensions of the sample and feature are analysed by One-Class SVM by data distribution. The fault features are constructed by improved SVDD into a single class of hyperspheres. The classification of unbalanced sample is realized by the increment of the hypersphere. Moreover, the information redundancy is reduced by extracting boundary samples. Therefore, the classification effect is improved. In summary, the classification validity of the proposed method for unbalanced samples is verified by the G-mean and MAUC performance indexes. The improved DCNN model in our method can fully capture the fault features compared to other diagnosis methods combining shallow networks (CNN3, RNN3 and DBN3) with SVDD. Compared with One-class SVM, our designed SVDD model can achieve better performance in unbalanced sample fault diagnosis. When the improved DCNN and SVDD models are combined together, they can achieve excellent performance in fault diagnosis of unbalanced samples.

To compare the classification effect more intuitively, we set up a classification visualization experiment in this section for the above method. The specific classifications are shown in Figure 15a–f. The conclusions can be drawn from the figure: When DPBN+SVDD is added (Figure 15a), the normal and fault sample are disordered and discrete. The samples cannot be effectively classified.

When CNN3+SVDD (Figure 15b) is used for classification, the effect is improved to a certain extent compared with that mentioned above. However, normal and fault samples still fail to get effective separation. When the RNN3+SVDD (Figure 15c) is added, there is a certain degree of diffusion in the sample. Therefore, the effect is still poor. When the DBN3+SVDD (Figure 15d) is added, the effective classifications of normal and attack are still not completed. After one-class SVM (Figure 15e) is added, normal and fault features are effectively classified. However, the samples are mixed and the convergence effects of the sample are poor. When the proposed method (Figure 15f) is added, normal and fault sample have been fully classified. Also, the sample convergence effect is good. This again shows the validity of our improvements to the CNN. Specifically,

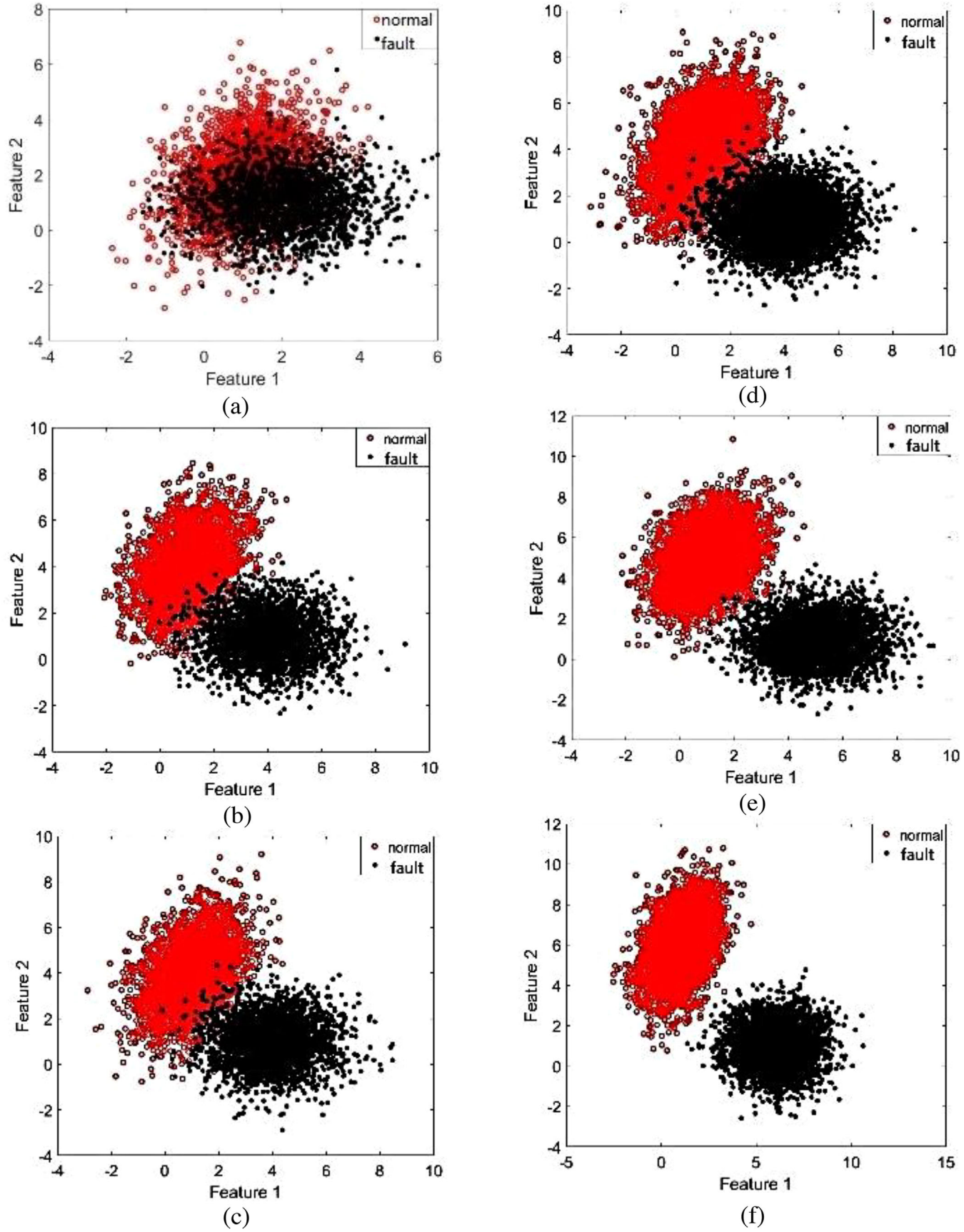

FIGURE 15 Comparative analysis. (a) DBPN+SVDD (b) CNN3+SVDD (c) RNN3+SVDD (d) DBN3+SVDD (e) one-class SVM (f) improved DCNN+SVDD.

TABLE 10 Comparison of accuracy of different methods under unbalanced samples

Model	KNN	DT	SVM	[10]	DBPN	[11]
Acc (%)	76.34	78.95	81.46	92.95	86.78	93.47
Model	CNN3	[38]	RNN3	[40]	DBN3	Proposed
Acc (%)	80.40	92.08	89.27	94.97	94.68	96.59

the depthwise separable convolution is used to optimize the extraction of fault features. Then, the adaptive batch normalization processing (AdaBN) is performed to improve the noise immunity. Meanwhile, the global average pooling layer (GAP) is used instead of the fully connected layer to improve the generalization ability of the network. Finally, an improved quantity learning algorithm for hypersphere coordinate mapping based on SVDD is proposed. The classification of unbalanced samples is realized.

4.5 | Fault diagnosis results

The diagnosis results of the different methods for normal and fault samples under two settings are given in Table 10. The three machine learning methods are divided into KNN (nearest neighbour method), DT (decision tree) and SVM (support vector machine). There are 8 deep learning methods: DBPN (deep artificial neural network), CNN3 (3-layer CNN), RNN3 (3-layer recurrent neural network), DBN3 (3-layer deep belief network). Moreover, our method is built on the basis of CNN. The advantages of our method can be demonstrated only by comparing with advanced CNN methods. Therefore, we choose the refs. [10, 11] and Refs. [38, 40] for our experiments.

Table 10 shows that the highest accuracy of the proposed method is 96.59%. However, the accuracy of the CNN3 network is only 80.40%. Also, it is lower than that of the SVM method. Therefore, the diagnosis limitation of the CNN network itself is explained. This suggests that traditional machine learning methods are generally effective in extracting deeper fault features. Thus leading to a lower final accuracy rate. Among the deep learning based diagnosis methods, CNN, RNN and DBN methods have average performance. The accuracy rates are both lower than those in [10, 11] and in [38, 40]. This is due to the large dataset and the overfitting problem of the algorithm itself. The proposed optimization strategy has the potential to overcome the inherent limitations of CNN. Finally, a more accurate diagnosis is achieved. Our method also achieves good diagnosis performance when faced with unbalanced fault datasets. Therefore, we first validate the diagnosis effectiveness of the proposed method with the accuracy rates obtained from the unbalanced datasets.

To further verify the diagnosis performance of the proposed method, fault samples are recorded as fault, and normal samples are recorded as normal. The ROC curve comparison experiment is conducted. The training and test ROC curves of the railway point machine dataset under the 6 methods is shown in Figures 16, 17. The six methods are the DBPN, CNN3, RNN3, DBN3, ref. [10], and the proposed method.

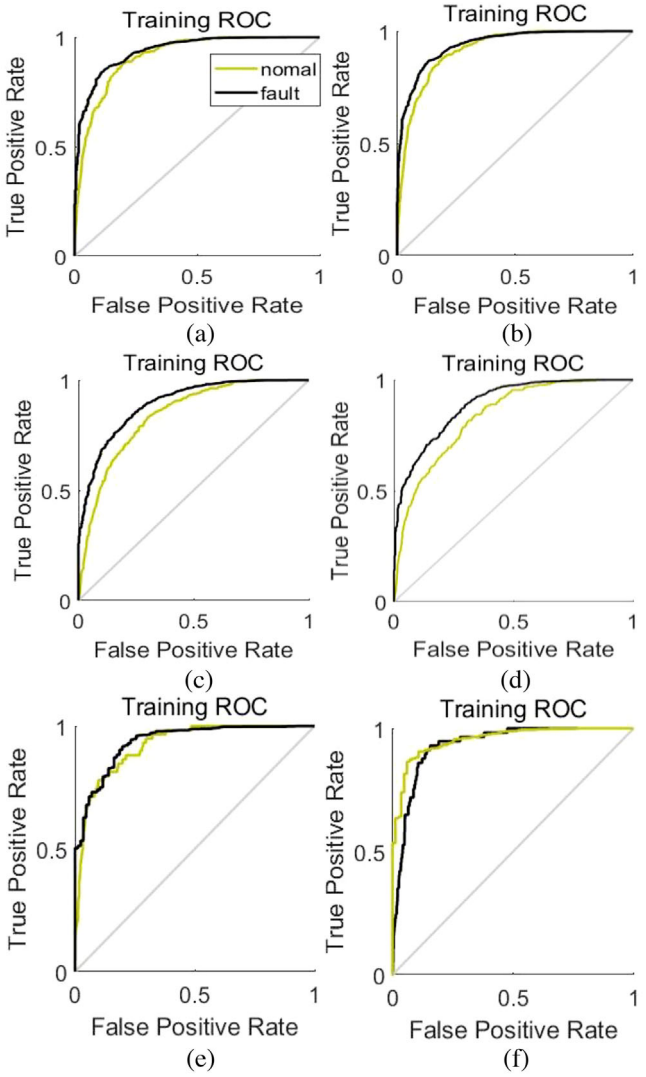

FIGURE 16 ROC curve comparison of training. (a) DBPN, (b) CNN3, (c) RNN3, (d) ref. [10], (e) DBN3, (f) proposed.

Figures 16 and 17 show that the AUC values obtained by the DBPN and CNN are relatively low in the ROC training and test curves. Moreover, the BPNN with four hidden layers has a low accuracy of direct training of the original data. The training accuracy of the BPNN with only one hidden layer is very low, and overfitting is severe. When the number of layers exceeds 5, the accuracy of the DBPN is difficult to improve under the influence of gradient disappearance. The overfitting is prominent, which is consistent with the conclusions drawn from Table 10. After the improvement of the CNN, the AUC value obtained by the proposed method is relatively high in the training and test ROC curves. After the improvement of the CNN, the AUC value is relatively high in the training and test ROC curves. Moreover, the AUC of the proposed method is superior to the normal and fault sample. In turn, it highlights the necessity of the optimization structure. However, this does not fully explain the effective performance of our method. There-

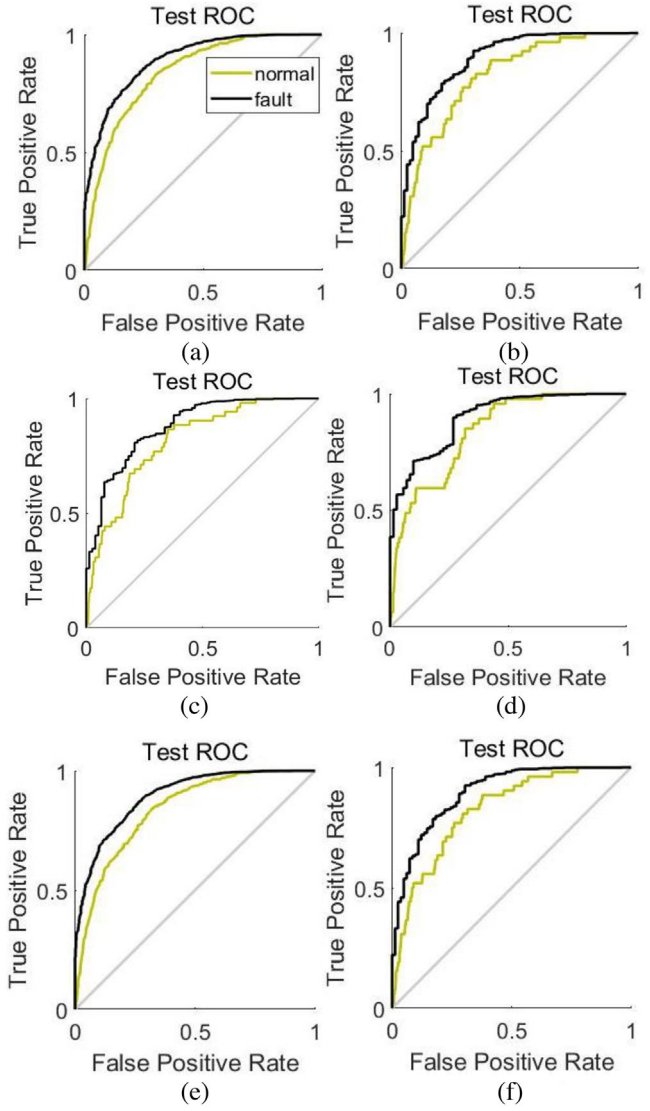

FIGURE 17 ROC curve comparison test. (a) DBPN, (b) CNN3, (c) RNN3, (d) ref. [10], (e) DBN3, (f) proposed.

TABLE 11 Indexes on normal sample

Model	Precision (%)	Recall (%)	FPR (%)	*F*-score (%)	AUC
KNN	76.454	74.335	19.436	75.380	0.757
DT	78.433	76.278	17.078	77.340	0.784
SVM	79.789	79.977	14.334	79.883	0.826
Ref. [10]	91.195	90.284	7.789	90.737	0.913
DBPN	86.224	85.427	11.120	85.824	0.865
Ref. [11]	93.457	92.365	5.118	92.908	0.934
CNN3	81.340	81.003	16.234	81.171	0.822
Ref. [38]	92.789	92.204	5.007	92.496	0.925
RNN3	89.711	89.121	9.545	89.415	0.897
Ref. [40]	94.102	93.698	3.434	93.900	0.941
DBN3	93.369	92.415	4.108	92.889	0.935
Proposed	**96.454**	**95.552**	**2.209**	**96.001**	**0.961**

TABLE 12 Indexes on fault sample

Model	Precision (%)	Recall (%)	FPR (%)	F-score (%)	AUC
KNN	77.525	77.921	18.220	77.723	0.771
DT	79.321	79.854	16.745	79.587	0.792
SVM	80.224	80.783	14.361	80.503	0.833
Ref. [10]	92.331	92.896	6.584	92.613	0.925
DBPN	87.106	87.630	10.169	87.367	0.871
Ref. [11]	93.962	94.451	4.337	94.206	0.938
CNN3	81.864	82.235	15.560	82.049	0.827
Ref. [38]	93.417	93.897	4.741	93.656	0.934
RNN3	89.928	89.993	8.238	89.960	0.902
Ref. [40]	94.635	94.892	3.006	94.763	0.946
DBN3	93.840	93.937	3.789	93.888	0.938
proposed	**96.993**	**97.321**	**2.001**	97.157	**0.964**

fore, we will again validate the performance of our method from the perspective of quantitative analysis.

To further verify the effectiveness of the proposed method, we select secondary evaluation indexes for a comparison of the methods that appear in Table 10. The precision rate, recall rate, false positive rate (FPR), F-score and AUC value are used for comparative analysis of under unbalanced distribution with a combination of normal and fault samples. For the calculation and definition of these indexes, please refer to ref. [10]. The specific results are shown in Tables 11 and 12.

According to Table 11, the following conclusions can be drawn: First, the precision rate of the three machine learning methods is not more than 80% for the diagnosis of normal sample. Moreover, it is lower than that of the nine deep learning methods. Therefore, the stability is general, and the classification effect is poor. The BPNN and CNN3 methods have poor precision rate in the nine deep learning methods. Also, their precision rates are not more than 90%. This result indicates that the training loss of the neural network will result when the CNN method is used solely for fault diagnosis. The necessity of improving the CNN method is highlighted. Moreover, the precision rate of the remaining methods is above 90%. However, the highest is that of the proposed method. In the recall rate, the average recall rate of the three machine learning methods do not exceed 80%. All recall rates are lower than those of the nine deep learning methods. The proposed method has the highest recall rate among the nine deep learning methods. Therefore, it has great advantages. Regarding the FPR, the FPR obtained by the proposed method is the lowest and far lower than those of the other methods. Also, the optimal number of training epochs is selected according to the training parameters of the experiment. The training loss is the lowest, and the overall FPR is the lowest. CNN3 is only higher than the KNN, DT and SVM in the comparison of the harmonic mean F-score. Specifically, the value does not exceed 85%. Also, the situation once again illustrates the limitation of the CNN. Because CNNs and are shallow neural networks, most of the deep fault features are

not captured. Also, the problems such as gradient disappearance and overfitting inherent to these two networks affect the training performance. In turn, this leads to poor final diagnosis results. The training overfitting is caused by increasing the number of layers of the CNN. It also more fully explains the necessity of the proposed method for improving the CNN. The largest average harmonic value of the F-score value is obtained by the proposed method. In the comparison of AUC values, all AUC values of the proposed method are better because the appropriate number of training epoch is selected, and the loss rate is low. This result is also consistent with the conclusion drawn from Figures 16 and 17. Similarly, the following conclusions can be drawn from Table 12. First, in the diagnosis of fault samples, the diagnosis performance is average regardless of whether three machine learning methods or nine deep learning methods are used. The average precision of the three machine learning methods is not more than 80%. Specifically, the average recall rate, average F-score value, and average AUC value are far lower than those of the proposed method. The accuracy rate of the nine deep learning methods is lower than 92%. However, the performance of the proposed method remains the best. The classification effect is relatively ideal compared with those of the other methods.

In summary, the effectiveness of the proposed method can be obtained by comparing several types of parameters under different methods. This also shows again that our proposed improved DCNN network is effective for CNNs. Our proposed improvement strategy in the improved DCNN network is able to address the inherent drawbacks of the CNN. Thereby, it makes the model perform better in capturing the fault features. Aiming at the unbalanced features of the railway point machine sample, an improved quantity learning algorithm for hypersphere coordinate mapping based on SVDD is proposed. In turn, the classification of unbalanced samples is realized.

4.6 | Ablation experiments

In the previous section, the diagnosis performance of the proposed method and the current methods is mainly compared. The proposed method contains several improvement strategies. Therefore, an ablation experiment is established to analyse the respective effects. In the experiment, the fault dataset of the railway point machine is used as the baseline dataset. The baseline network is CNN network. When the improvement strategy is successively added, the accuracy of fault and normal sample is given in Table 13. The baseline experiment is carried out, and the specific idea is: Xception, AdaBN, GAP, L2 regularization, Adam, and Dropout optimization methods are not added. In the fault and normal sample diagnosis tasks, the classification accuracy under this experiment is 65.96% and 70.35%, respectively. Compared with the optimal result of the dataset, there is a big gap. This indicates that the network generalization ability of pure CNN is poor. Therefore, the following contents verify the necessity of each step in the method in this paper, which are mainly as follows:

First, the Xception optimization structure is added in the experiment. In the fault and normal sample diagnosis tasks, the classification accuracy under this experiment is 69.12% and 72.63%, respectively. Compared with the baseline experiment, the improvement is 3.16% and 2.28%, respectively. Xception can solve the overfitting and gradient disappearance in the network proved by the result. Second, the AdaBN optimization strategy is added in the experiment. Compared with the baseline experiment, the improvement is 4.12% and 3.66%, respectively. Third, the fully connected layer is replaced by GAP layer in the experiment. Compared with the baseline experiment, the improvement is 7.51% and 5.99%, respectively. When the original fully connected layer is replaced by GAP layer, the results show that the network complexity can be effectively reduced. Fourth, the L2 regularization optimization strategy is used in the experiment. As mentioned above, the classification accuracy under the experimental setting is improved by 5.44% and 4.84%, respectively. This result indicates that unimportant feature variables are weakened by L2 regularization, while important features are extracted from many feature variables. Fifth, the Adam optimization strategy is added in the experiment. The classification accuracy under the experimental setting increases by 4.15% and 3.82%, respectively. When the variation Dropout optimization strategy is added, the classification accuracy under the experimental setting increased by 3.13% and 3.15%, respectively. The overfitting condition in the network can be well optimized by variation Dropout structure.

To further verify the necessity of each step of the proposed method, we remove each strategy and carry out the experiment. The specific results are shown in Table 14. It can be observed from Table 14 that each strategy of the proposed method will affect the accuracy of fault and normal sample diagnosis task. Taking Xception as an example, the accuracy of the two samples decrease by 2.21% and 1.41% when the Xception module is removed, respectively. There is also an impact on accuracy when removing other modules. This also means that none of our proposed strategies can be removed. Each proposed strategy can affect the diagnosis performance of our method. In turn, this again validates the effectiveness of the proposed strategies of our model.

4.7 | Robustness and model complexity verification

To demonstrate the advantages of the proposed method over others, the robustness and model complexity of these methods are compared in terms of the average accuracy, standard deviation of accuracy, training time, and test time when damaged samples are found in the training samples. The 600-point machine samples and 200 damaged samples are collected. The 70% samples are used for training, and the remainder are used for testing. To avoid the interference of random factors, the above experiments are conducted on different test sets 20 times. The specific parameter values are obtained. Specifically, the experimental results are shown in Table 15.

TABLE 13 Add each improvement strategy in turn

Xception	AdaBN	GAP	L2 regularization	Adam	Variation dropout	Fault	Normal
						65.96 ± 0.37	70.35 ± 0.48
√						69.12 ± 0.42	72.63 ± 0.70
√	√					73.24 ± 0.31	76.29 ± 0.65
√	√	√				80.75 ± 0.66	82.28 ± 0.31
√	√	√	√			86.19 ± 0.80	87.12 ± 0.38
√	√	√	√	√		90.34 ± 0.33	90.94 ± 0.42
√	√	√	√	√	√	**93.47 ± 0.54**	**94.09 ± 0.27**

TABLE 14 Remove each module

Xception	AdaBN	GAP	L2 regularization	Adam	Variation dropout	Fault	Normal
√	√	√	√	√	√	**93.47 ± 0.54**	**94.09 ± 0.27**
	√	√	√	√	√	91.26 ± 0.39	92.68 ± 0.81
√		√	√	√	√	92.04 ± 0.92	91.02 ± 0.66
√	√		√	√	√	90.43 ± 0.17	91.95 ± 0.43
√	√	√		√	√	91.86 ± 0.75	92.07 ± 0.62
√	√	√	√		√	90.89 ± 0.21	90.81 ± 0.29
√	√	√	√	√		90.34 ± 0.33	90.94 ± 0.42

TABLE 15 Comparison indexes with damage samples

Method	Average accuracy/%	Standard deviation of accuracy/%	Training time/s	Test time/s
KNN	**73.284**	3.018	96.32	1.258
DT	76.397	**4.254**	89.54	**2.367**
SVM	82.642	2.693	100.57	1.964
Ref. [10]	90.415	2.016	112.53	1.961
DBPN	81.639	2.968	152.78	2.078
Ref. [11]	91.424	1.821	141.79	1.908
CNN3	76.158	4.529	**154.63**	**1.924**
Ref. [38]	91.726	1.635	157.07	1.384
RNN3	87.654	1.857	**168.24**	0.846
Ref. [40]	92.029	1.023	176.44	0.953
DBN3	92.419	1.206	162.31	0.681
Proposed	**95.637**	**0.852**	**105.39**	**0.582**

Conclusions can be drawn by comparing the results in Table 15. When the damaged samples of other non-railway point machine samples are mixed, the performance of the improved DCNN is obviously better than those of the others. In terms of the comparison of average accuracy, the SVM method has the highest accuracy (82.642%) among the three machine learning methods. However, the average accuracy of the BPNN with one hidden layer for direct training of original data is only 80.371%. The training accuracy is very low, and overfitting is severe. When the number of DBPN layers exceeds 5, it is difficult to improve the accuracy under the influence of gradient disappearance and overfitting. The average accuracy is only 1.268% higher than that of the BPNN, which is a small increase. Increasing the number of network layers of the CNN, RNN, and DBN, it can improve the fault diagnosis accuracy. However, the range is small. It is apparent by comparing the CNN2 and CNN3 that simply increasing the number of network layers will also cause CNN overfitting. In turn, it leads to low diagnosis accuracy. Although the training time of the three machine learning methods is much less than that of the improved DCNN, the improved DCNN has more advantages for real-time online monitoring and diagnosis in terms of test time. In conclusion, the proposed method is obviously superior to other methods and has better stability. This also shows that our method can be used for real-time fault diagnosis of ZDJ7 railway point machine with better efficiency.

However, this does not fully validate the superiority of our method in terms of computational complexity. Therefore, we set up the following experiments: In addition to the training time and testing time given in Table 15, we also use the number of parameters, multiplication addition operations (FLOPs), and PPL values as indexes [54]. These three indexes are used to test the computational power, and the smaller their values represent the greater computational power. To verify the efficiency of our model in the presence of noise, we also give the indexes under two settings (with and without noise), as shown in Table 16. From the results, our model obtain better indexes by virtue of

TABLE 16 Computational complexity comparison

	Without noise			With noise		
Model	**Parameters/M**	**FLOPs/G**	**PPL value**	**Parameters/M**	**FLOPS/G**	**PPL value**
KNN	69.341	68.630	62.412	72.847	71.352	63.557
DT	73.519	99.528	65.378	75.313	102.113	66.769
SVM	85.127	78.341	72.503	87.009	81.717	73.258
Ref. [10]	74.134	112.207	84.604	76.751	115.003	85.569
DBPN	81.829	90.028	65.341	83.632	93.112	66.427
Ref. [11]	93.424	134.434	93.129	95.085	137.285	94.367
CNN3	50.113	88.112	57.738	52.667	90.138	58.414
Ref. [38]	88.756	137.525	106.057	90.438	140.419	107.112
RNN3	51.434	93.709	70.023	53.521	96.668	71.085
Ref. [40]	78.525	122.057	114.415	80.790	125.361	115.234
DBN3	50.792	86.713	76.224	52.417	89.852	77.298
Proposed	**54.438**	**81.715**	**72.037**	**55.126**	**82.347**	**72.861**

the number of parameters (54.438 M), FLOPs (81.715 G), training time (105.39 s), test time (0.582 s) and PPL value (72.037) compared to others in the absence of noise. Also, better indexes are obtained in the presence of noise. Although the number of parameters and FLOPs have increased, the increase is not significant. Although the indexes obtained by our model are higher than those of the three models (CNN3, RNN3 and DBN3), the difference is not significant. This difference is in an acceptable range for practical fault diagnosis. Therefore, this indicates that our model is able to perform fault diagnosis of ZDJ7 railway point machine with strong noise background under unbalanced samples distributions.

5 | CONCLUSIONS

An intelligent fault diagnosis method for railway point machines is proposed. The method is applied to the actual railway point machine fault dataset to verify the accuracy and superiority of the method, and the conclusions are as follows:

1. The proposed method has a higher fault diagnosis accuracy and anti-noise ability. Moreover, it is applied to intelligent diagnosis of the fault experimental data. It has a higher accuracy than methods based on manual feature extraction. Moreover, the proposed method has a shorter test time. Also, it is more conducive to the application of online real-time monitoring and rapid intelligent diagnosis.
2. The traditional CNN is improved by adopting the GAP, Xception structure, and AdaBN. The number of training parameters and computation time of the model are effectively reduced. Also, the overfitting of the model is prevented. Moreover, the proposed method does not need to perform any manual feature extraction on the original fault data. In turn, the results show that the "end-to-end" model structure has better operability and generality.
3. The designed optimized boundary extraction algorithm based on SVDD can classify unbalanced fault samples with high accuracy. The accuracy rate, recall rate, FPR, F-score and AUC value are better than other methods. Moreover, it has a strong robustness and superior diagnosis performance when samples are destroyed.
4. Although our model is able to achieve some performance in fault diagnosis with unbalanced samples, it is no longer applicable in the face of sample sparsity and few-shot sample distribution. Considering that domain adaption can solve the fault diagnosis problem with sparse samples, we will use domain adaption theory to solve the few-shot sample fault diagnosis problem in the next phase of our work.

AUTHOR CONTRIBUTIONS

Zengshu Shi: Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing – original draft; Yiman Du: Conceptualization, Data curation, Methodology, Supervision, Writing – original draft, Writing – review & editing; Xinwen Yao: Investigation, Software, Visualization.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

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

REFERENCES

1. Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., Zerhouni, N.: Prognostics and health management for maintenance practitioners-Review, implementation and tools evaluation. Int. J. Progn. Health Manage. 8(3), 1–31 (2017)
2. Eker, O.F., Camci, F., Guclu, A., Yilboga, H., Sevkli, M., Baskan, S.: A simple state-based prognostic model for railway turnout systems. IEEE Trans. Ind. Electron 58(5), 1718-1726 (2011)

3. Ciabattoni, L., Ferracuti, F., Freddi, A., Monteriu, A.: Statistical spectral analysis for fault diagnosis of rotating machines. IEEE Trans. Ind. Electron. 65(5), 4301–4310 (2017)
4. Javed, K., Gouriveau, R., Zerhouni, N., Nectoux, P.: Enabling health monitoring approach based on vibration data for accurate prognostics. IEEE Trans. Ind. Electron. 62(1), 647–656 (2015)
5. Chen, H., Miao, F., Chen, Y., Xiong, Y., Chen, T.: A hyperspectral image classification method using multifeature vectors and optimized KELM. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2781–2795 (2021). https://doi.org/10.1109/JSTARS.2021.3059451
6. Freeman, C., Kuli´c, D., Basir, O.: An evaluation of classifier-specific filter measure performance for feature selection. Pattern Recognit. 48(5), 1812–1826 (2015)
7. Coble, J., Hines, J.W.: Identifying optimal prognostic parameters from data: A genetic algorithms approach. In: Proceedings of Annual Conference of the Prognostics and Health Management Society. pp. 1–11. San Diego, CA, USA (2009)
8. Zhou, X., Hongjiang, M., Gu, J.: Parameter adaptation-based ant colony optimization with dynamic hybrid mechanism. Eng. Appl. Artif. Intell. 114, 105139 (2022). https://doi.org/10.1016/j.engappai.2022.105139
9. Yao, R., Guo, C., Deng, W.: A novel mathematical morphology spectrum entropy based on scale-adaptive techniques. ISA Trans. 126, 691–702 (2022). https://doi.org/10.1016/j.isatra.2021.07.017
10. Zhao, H., Liu, J., Chen, H.: Intelligent diagnosis using continuous wavelet transform and gauss convolutional deep belief network. IEEE Trans. Reliab. 10, 1-11 (2022). https://doi.org/10.1109/TR.2022.3180273
11. Jia, F., Lei, Y., Lu, N.: Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. Mech. Syst. Sig. Process. 110, 349–367 (2018). https://doi.org/10.1016/j.ymssp.2018.03.025
12. Han, T., Li, Y.-F.: Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles. Reliab. Eng. Syst. Saf. 226, 108648 (2022). https://doi.org/10.1016/j.ress.2022.108648
13. Chen, Y., Rao, M., Feng, K.: Physics-Informed LSTM hyperparameters selection for gearbox fault detection. Mech. Syst. Sig. Process. 171(15), 108907 (2022). https://doi.org/10.1016/j.ymssp.2022.108907
14. Lee, W., Park, C.G.: Double fault detection of cone-shaped redundant IMUs using wavelet transformation and EPSA. Sensors. 14(2), 3428–3444 (2014)
15. Wen, C.L., Lv, F.Y., Bao, Z.J.: A review of data driven-based incipient fault diagnosis. Acta Autom. Sin. 42(09), 1285–1299 (2016)
16. Zhao, J., Liao, Y.Y., Yang, S.P.: An extension of unscented Kalman filter to dynamic Bayesian wavelet transform in fault diagnosis of rolling element bearings. J. Vibr. Shock. 39(11), 53–62 (2020)
17. Yan, R.W., Lin, C., Gao, S.X.: Fault diagnosis and analysis of circuit breaker based on wavelet time-frequency representations and convolution neural network. J. Vibr. Shock. 39(10), 198–205 (2020)
18. Yu, L.A., Zhang, Y.D.: Weight-selected attribute bagging based on association rules for credit dataset classification. Syst. Eng.-Theory Pract. 40(2), 366–372 (2020)
19. Nai, Z., Wen, Q.: Application of hybrid operating rooms for clipping large or giant intracranial carotid-ophthalmic aneurysms. World J. Clin. Cases. 8(21), 5149–5158 (2020)
20. Ardakani, H.D., Lucas, C., Siegel, D., Chang, S., Dersin, P., Bonnet, B., Lee, J.: PHM for railway system-A case study on the health assessment of the point machines. In: *IEEE Conference on Prognostics and Health Management (PHM)*. pp. 1–5. Denver, CO, USA (2012)
21. Toklu, E., Altinisik, M., Elbay, A., Koytak, A.: Comparison of postoperative anterior segment changes associated with pars plana vitrectomy with and without vitreous base shaving. Int. J. Ophthalmol. 13(11), 1745–1752 (2020)
22. Gao, B., Zhao, P.F., Lu, Y.X.: Study on recognition and classification of blood fluorescence spectrum with BP neural network. Spectrosc. Spectral Anal. 38(10), 3136–3143 (2018)
23. Lee, J., Choi, H., Park, D., Chung, Y., Kim, H.Y., Yoon, S.: Fault detection and diagnosis of railway point machines by sound analysis. Sensors. 16(4), 114-118 (2016)
24. Soualhi, A., Medjaher, K., Zerhouni, N.: Bearing health monitoring based on Hilbert-Huang transform, support vector machine, and regression. IEEE Trans. Instrum. Meas. 64(1), 52–62 (2015)
25. Guo, Z., Ye, H., Dong, W.: A hybrid feature extraction method for fault detection of turnouts. In: 2017 Chinese Automation Congress (CAC). Jinan, pp. 540–545 (2017)
26. Guan, Q.: High-speed railway switch failure diagnosis based on FOA-LSSVM. J. Sci. Technol. 31(4), 230–232 (2015)
27. Lei, Y.T.: Railway point machine diagnostic system based on WPT And EMD incorporated feature extraction. Mach. Build. Autom. 46(2), 219–222 (2017)
28. Zhang, K.: The railway turnout fault diagnosis algorithm based on BP neural network. In: 2014 IEEE International Conference on Control Science and Systems Engineering. Yantai, pp. 135–138 (2014)
29. Liu, M., Yan, X., Sun, X.: Fault diagnosis method for railway turnout control circuit based on information fusion. In: *IEEE Information Technology Networking Electronic and Automation Control Conference*. pp. 315–320. Chongqing, China (2016)
30. Xiao, M., Zhai, C., Pan, C.L.: Research on S700K railway point machine fault diagnosis based on fast Bayesian network. J. Railway Sci. Eng. 000(002), 414–418 (2015)
31. Wang, R.F., Chen, W.B.: Research on fault diagnosis method for S700K railway point machine based on grey neural network. J. China Railway Soc. 38(6), 68–72 (2016)
32. Huang, Y.P., McMurran, R., Dhadyalla, G.: Probability based vehicle fault diagnosis: Bayesian network method. J. Intell. Manufact. 19(3), 301–311 (2008)
33. Dong, H.Y., Li, N.: Fault diagnosis method for railway point machine based on D-S evidence theory information fusion. J. Test Measur. Technol. 27(1), 1–7 (2013)
34. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science. 313(5786), 504–507 (2006)
35. Jia, F., Lei, Y., Lin, J.: Deep neural networks: A processing tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech. Syst. Signal Proces. 72-73, 303–315 (2016)
36. Tamilselvan, P., Wang, P.: Failure diagnosis using deep belief learning based health state classification. Reliab. Eng. Syst. Saf. 115(7), 124–135 (2013)
37. Li, W.H., Shan, W.P., Zeng, X.Q.: Bearning fault identification based on deep belief network. J. Vib. Eng. Technol. 29(02), 340–347 (2016)
38. Li, H., Zhang, Q., Qin, X.R.: Fault diagnosis method for rolling bearings based on short-time Fourier transform and convolution neural network. J. Vib. Shock. 37(19), 124–131 (2018)
39. Qu, J.L., Yu, L., Yuan, T.: Adaptive fault diagnosis algorithm for rolling bearings based on one-dimensional convolutional neural network. Chin. J. Sci. Instr. 39(07), 134–143 (2018)
40. Janssens, O., Slavkovikj, V., Vervisch, B.: Convolutional neural network based fault detection for rotating machinery. J. Sound Vib. 377, 331–345 (2016)
41. Wang, F., Jiang, H., Shao, H.: An adaptive deep convolutional neural network for rolling bearing fault diagnosis. Meas. Sci. Technol. 28(9), 005-016 (2017)
42. Lee, K.B., Cheon, S., Chang, O.K.: A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes. IEEE Trans. Semicond. Manuf. 30(2), 135–142 (2017)
43. Ji, L.Y., Zhao, M., Li, P., Xu, X.Q.: A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. Measurement. 111, 1–10 (2017). https://doi.org/10.1016/j.measurement.2017.07.017
44. Szegedy, C., Liu, W., Jia, Y.Q.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston (2015)
45. Chen, L., Zhang, H., Xiao, J.: SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu (2017)
46. HE, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21(9), 1263–1284 (2009)
47. Li, T., Duan, L.X., Zhang, D.N.: Application of adaptive convolutional neural network in rotating machinery fault diagnosis. J. Vib. Shock. 39(16), 275–282 (2020)

48. Wang, Y., Wong, J., Miner, A.: Anomaly intrusion detection using one class SVM. In: Proceedings of the Fifth Annual IEEE SMC on Information Assurance Workshop. pp. 358–364 (2004)
49. Ma, J., Dai, G.Z., Xu, Z.: Network anomaly detection using dissimilarity-based one-class SVM classifier. In: Proceedings of the International Conference on Parallel Processing. pp. 409–414. Vienna, Austria (2009)
50. Chen, Y.T., Qian, J., Saligrama, V.: A new one-class SVM for anomaly detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3567–3571. Vancouver, BC (2013)
51. Prasetyo, E., Purbaningtyas, R., Adityo, R.D.: Combining MobileNetV1 and depthwise separable convolution bottleneck with expansion for classifying the freshness of fish eyes. Inf. Process. Agric. 9(4), 485–496 (2022). https://doi.org/10.1016/j.inpa.2022.01.002
52. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. ar Xiv preprint ar Xiv:1602.07261 (2016)
53. Kingma, D., Adam, B.J.: A method for stochastic optimization. ar Xiv preprint ar Xiv:1412.6980 (2014)
54. Ansari, M.S., Cockburn, B.F., Han, J.: An improved logarithmic multiplier for energy-efficient neural computing. IEEE Trans. Comput. 4(70), 614–625 (2020). https://doi.org/10.1109/TC.2020.2992113

How to cite this article: Shi, Z., Du, Y., Yao, X.: Fault diagnosis of ZDJ7 railway point machine based on improved DCNN and SVDD classification. IET Intell. Transp. Syst. 17, 1649–1674 (2023). https://doi.org/10.1049/itr2.12357