Hindawi
Shock and Vibration
Volume 2021, Article ID 6616592, 16 pages
https://doi.org/10.1155/2021/6616592

Research Article

Vibration Images-Driven Fault Diagnosis Based on CNN and Transfer Learning of Rolling Bearing under Strong Noise

Hongwei Fan,[1,2] Ceyi Xue,[1] Xuhui Zhang,[1,2] Xiangang Cao,[1,2] Shuoqi Gao,[1] and Sijie Shao[1]

[1]*School of Mechanical Engineering, Xi'an University of Science and Technology, Xi'an 710054, China*
[2]*Shaanxi Key Laboratory of Mine Electromechanical Equipment Intelligent Monitoring, Xi'an 710054, China*

Correspondence should be addressed to Hongwei Fan; hw_fan@xust.edu.cn

Received 5 January 2021; Revised 12 March 2021; Accepted 19 March 2021; Published 12 April 2021

Academic Editor: Yi Qin

Deep learning-based fault diagnosis of rolling bearings is a hot research topic, and a rapid and accurate diagnosis is important. In this paper, aiming at the vibration image samples of rolling bearing affected by strong noise, the convolutional neural network- (CNN-) and transfer learning- (TL-) based fault diagnosis method is proposed. Firstly, four kinds of vibration image generation method with different characteristics are put forward, and the corresponding pure vibration image samples are obtained according to the original data. Secondly, using CNN as the adaptive feature extraction and recognition model, the influences of main sensitive parameters of CNN on the network recognition effect are studied, such as learning rate, optimizer, and L1 regularization, and the best model is determined. In order to obtain the pretraining parameters, the training and fault classification test for different image samples are carried out, respectively. Thirdly, the Gaussian white noise with different levels is added to the original signals, and four kinds of noised vibration image samples are obtained. The previous pretrained model parameters are shared for the TL. Each kind of sample research compares the impact of thirteen data sharing schemes on the TL accuracy and efficiency, and finally, the test accuracy and time index are introduced to evaluate the model. The results show that, among the four kinds of image generation method, the classification performance of data obtained by empirical mode decomposition-pseudo-Wigner–Ville distribution (EP) is the best; when the signal to noise ratio (SNR) is 10 dB, the model test accuracy obtained by TL is 96.67% and the training time is 170.46 s.

1. Introduction

The safe operation of mechanical equipment is an important guarantee for the modern industrial production. As an indispensable part of intelligent equipment, the fault diagnosis technology of rolling bearings has attracted great attention. Rolling bearing is generally composed of inner ring, outer ring, rolling body, and cage. After a long time of operation, various faults are easy to occur. Therefore, the rapid and accurate fault identification of rolling bearing is a great challenge and of great significance [1, 2].

The traditional fault diagnosis method of rolling bearings is mainly to obtain the characteristic information through the processing of vibration signals. Different analysis algorithms are selected for different types of vibration signal. Common analysis algorithms, such as wavelet transform and empirical mode decomposition (EMD), still play important roles in the fault diagnosis field and are constantly being improved [3, 4]. The multialgorithm fusion analysis can synthesize the integrated advantages and improve the analysis effect. Jiang et al. [5] used multiwavelet packet as the prefilter to refine the vibration signal combining with ensemble EMD (EEMD), which can implement the effective extraction of multifault features. Guo and Deng [6] used particle swarm optimization (PSO) algorithm to screen the optimal intrinsic modal function (IMF), and a multiobjective optimized EMD method was proposed to overcome the modal aliasing of EMD and EEMD. The authors in [7–10] adopted the comprehensive analysis method of EMD and pseudo-Wigner–Ville distribution (PWVD), which not only has the time-frequency focusing but also avoids the problem of cross-interference terms in multicomponent signal processing. The frequency band energy was divided according to time-frequency image, which was used as the characteristic index for the unsupervised clustering calculation and achieved a good classification effect [11].

With the rapid development of intelligent algorithms such as deep learning and cross-integration with various disciplines, the intelligent diagnostic methods are extensively used in mechanical fault diagnosis [2, 12]. Compared with the traditional fault diagnosis method, the intelligent diagnostic method has no strict prior knowledge requirements and can avoid the dependence on the artificial feature extraction [13, 14]. Shao et al. [15] used the dual-tree complex wavelet packet (DTCWPT) as the signal preprocessing, and a new adaptive deep belief network (DBN) with DTCWPT was proposed, which can eliminate the necessity of artificial feature selection and reliably identify different bearing faults. As a common type of network, the convolutional neural network (CNN) generally uses images as input samples. When facing the fault diagnosis problem of rotating machinery, choosing an appropriate image generation method is significant. Wang et al. [16] took the time-frequency image gained by wavelet analysis for the CNN training, which can achieve a good effect of fault identification. Xiao et al. [17] compared the effects of grayscale map samples with different sizes and different optimizers on the accuracy and robustness of training model. The performance of CNN model is often related to its own parameter setting, and Chen et al. [18] changed the extreme learning machine (ELM) as a strong classifier to improve the classification performance. In general, obtaining a group of highly qualified samples in the actual industrial environment is affected by various factors, and too few samples and the imbalance among different data will affect the training effect of the model. A framework based on the auxiliary classifier generative adversarial network (ACGAN) is proposed [19], which has the ability to generate the artificial raw data of mechanical signals. However, in the face of data affected by the strong background noise under the complex working conditions, the process of using generative adversarial network (GAN) to generate new samples is complicated. Transfer learning (TL) provides another solution for small sample cases.

The data corresponding to different working conditions are of different distributions. It is found that the classification accuracy of CNN is low when the distribution of training data set and test data set is different [20, 21]. Qiu et al. [22] performed an unsupervised and semisupervised dimensionality reduction, respectively, on the source and target domain data so that two domain data had a relatively similar distribution state, and then, TL was performed. Wang et al. [23] shortened the conditional distribution distance of vibration data acquired under different working loads for the intraclass adaptation. Fan et al. [24] used the designed image samples for pretraining and transferred the CNN model parameters to the tested samples, and after fine-tuning, the parameters a good classification effect was achieved.

In order to solve the problems of the sample effectiveness and the performance of TL under the strong noise, a method of vibration image-driven CNN- and TL-based model is proposed in this paper. In Section 2, four different types of image generation method are proposed and optimized. In Section 3, CNN is used for the adaptive feature extraction, three sensitive parameters in the model are studied, and the best model is determined. In Section 4, thirteen different TL schemes are designed for the noised samples, and finally, the recognition accuracy and efficiency for four types of image samples are evaluated.

2. Research Framework

2.1. Overall Scheme. The research process of this work is shown in Figure 1, and the main process is divided into three parts. Part 1 is the preprocessing of samples. Four kinds of image processing method are proposed, which are intrinsic mode functions arrangement (IMFA), empirical mode decomposition-pseudo-Wigner–Ville distribution (EP), symmetrical polar coordinates image (SPCI), and grayscale texture map (GTM) to convert the original vibration signal into images, which are used as the training samples of CNN model. Part 2 is the model training process, in which the CNN model is built and then the model parameters are shared for the noised vibration image samples and TL. Part 3 is the test and evaluation of the recognition model.

2.2. Convolution Neural Network. As a feedforward neural network, the CNN model is composed of convolution layer, pooling layer, and fully connected layer, in which the convolution layer is used to convolute the image to obtain the featured map. The convolution formula is shown as follows [24]:

$$O^l = \sum_{i=0}^{k}\sum_{j=0}^{k} W_{i,j}^{(l)} I_{n+i,m+j}^{(l-1)} + b^{(l)}, \tag{1}$$

where k is the size of the convolution kernel; i and j are the i-th row and j-th column of the convolution kernel, respectively; $W_{i,j}^{(l)}$ is the weight matrix of convolution kernel in layer l; $b(l)$ is the offset of layer l; $I_{n+i,m+j}^{(l-1)}$ is the featured map matrix of layer l-1; and its two dimensional are n, m; and O^l is the output featured map matrix of layer l.

After the convolution operation, the parameters are input into the activation function, which increases the nonlinearity of the model, so that the model can be applied to the complex classification problems. The activation function formula is shown as follows:

$$y^l = f_{\text{active}}\left(O^{l(i,j)}\right), \tag{2}$$

where f_{active} is the activation function and y^l is the output of the l-layer featured map matrix. The common activation functions are Sigmoid, Tanh, and Relu.

The pooling layer is to reduce the dimension of the input parameters, and the commonly used pooling operations are maximum pooling and average pooling. The fully connected

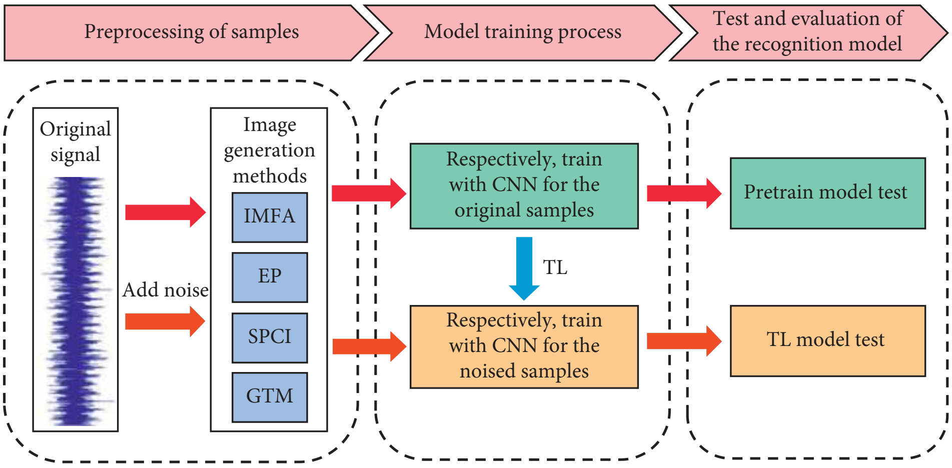

Preprocessing of samples
Model training process
Test and evaluation of the recognition model
Original signal
Add noise
Image generation methods
IMFA
EP
SPCI
GTM
Respectively, train with CNN for the original samples
TL
Respectively, train with CNN for the noised samples
Pretrain model test
TL model test

Figure 1: Overall research flowchart.

layer in the convolution network maps the distributed features to the sample space. In the whole network structure, the fully connected layer generally acts as a classifier, for example, the Softmax function, and the calculated result is the output of the whole CNN model, as shown in the following equations:

$$z_i^l = y_i^{(l-1)} W^{(l)} + b^{(l)}, \tag{3}$$

$$P(Y_i) = \frac{e^{z_i^{(l)}}}{\sum_{u=1}^{L} e^{z_u^{(l)}}}, \tag{4}$$

where $z_i^{(l)}$ is the eigenvalue of the fully connected layer, $y_i^{(l-1)}$ is the featured matrix before the fully connected layer, $P(Y_i)$ is the probability prediction of theY_i – th sample, L is the total number of sample categories, and u is the u-th sample category in order to distinguish the i in the numerator.

In the process of model training, the error between the predicted model output and the actual target is calculated by the loss function, and the commonly used loss function is the cross-entropy function, as follows [24]:

$$C = -\frac{1}{M} \sum_x p(x)\log(q(x)) + (1 - p(x))\log(1 - q(x)), \tag{5}$$

where assuming that $p(x), q(x)$ are the true target value and the Softmax output value, respectively, and M is the total number of samples.

The C value in equation (5) indicates how close the predicted value is to the target value. The smaller the value is, the closer the prediction is to the target, and on the contrary, the prediction deviates from the target. The C value obtained by each forward propagation is input into the optimizer, i.e., by reducing the C value to achieve the training effect, the commonly used optimizer has the gradient descent, momentum optimization, etc.

2.3. Transfer Learning Method. The TL is an effective method to solve small sample problem, which takes the model developed for task A as the starting point and reuses it in the process of developing the model for task B. In a practical research, TL is mainly divided into three kinds: the case-based, feature-based, and shared parameter-based transfer [25, 26]. With the development of deep learning, the combination of feature learning and TL is more and more applied. In order to solve the problem of low training accuracy and poor classification effect for low quality data affected by background noise, in this paper, a method of TL combined with CNN is used to classify rolling bearing faults. The original data collected are used as the pure training samples for pretraining, and then, the model parameters will be restored and shared to train the model after pretraining. A part of the convolution layer is frozen so that the parameters cannot be updated in the process of backpropagation. Only the designated training layer is retained to participate in the training process of noised samples to complete the parameter updating.

3. Research Basis

3.1. CNN Model Construction. This paper takes the LeNet-5 network as the basic model. The model has a shallow network layer and requires few training parameters, so it has fast training speed and good classification effect. For the low complexity of vibration image samples in this paper, it is appropriate to choose this network. The network structure is shown in Figure 2. The total number of layers is 7, including 3 convolution layers, 2 pooling layers, and 2 fully connected layers. Under the condition that the image sample is not affected by its color, the use of single-channel image can not only reduce the input parameters and the memory usage but also improve the operation efficiency. Therefore, the sample used in this paper is a single-channel image with a size of 256 ∗ 256, and the output is 4 nodes, which represent 4 bearing state categories. In the training process, the hyperparameters in the model are fine-tuned.

The initial CNN model parameters are shown in Table 1.

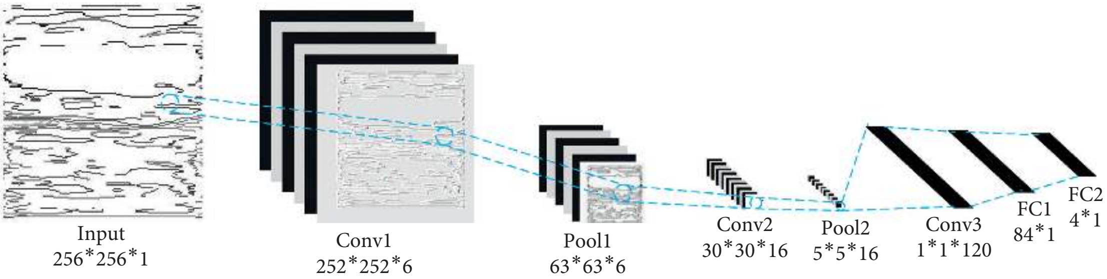


Figure 2: CNN model structure diagram.

Table 1: CNN model parameters preset.

Items	Parameters
Batch size	24
Convolution1 kernel	Length ∗ width ∗ number = 5 ∗ 5 ∗ 6, stride = 1
Convolution2 kernel	Length ∗ width ∗ number = 5 ∗ 5 ∗ 16, stride = 2
Convolution3 kernel	Length ∗ width ∗ number = 5 ∗ 5 ∗ 120
Pool1 set	Type = max pooling, length ∗ width = 4 ∗ 4, stride = 4
Pool2 set	Type = max pooling, length ∗ width = 6 ∗ 6, stride = 6
Activate function	Relu
Loss function	Cross entropy
Optimizer	Stochastic gradient descent
Initial learning rate	0.001
Regularization	None

3.2. Vibration Image Sample Preparation. The original vibration signals do not have the advantages of images, such as great differences between samples and more intuitive observation. Therefore, this paper carries out four image generation methods for vibration signals and finds which of 4 types of image has better classification performance. The samples used in this paper are the vibration data of rolling bearing from the Case Western Reserve University [27], and the test platform is shown in Figure 3.

The selected data are those with 1797 r/min, fault degree of 0.007 inch, sampling frequency of 12 kHz, and 4 kinds of working state including normal state, inner ring fault, outer ring fault, and rolling body fault. According to the rotational speed and sampling frequency, it is calculated that 400 sampling points are obtained for each rotation, so this paper converts 400 sampling points as a group of data for image conversion.

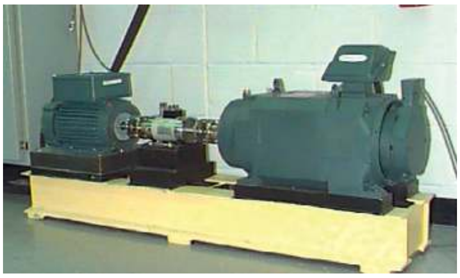


Figure 3: Bearing test platform of Case Western Reserve University.

3.2.1. IMFA Image Generation Method

(1). EMD. EMD [28] can decompose the signal into the sum of a series of IMFs with different time scales, and each IMF is a single component signal.

For any signal$x(t)$, all the extreme points are firstly determined, and then, the upper envelope $x_{\max}(t)$ and the lower envelope $x_{\min}(t)$ are obtained by third-order spline interpolation. If $m(t)$ is the average value of the upper and lower envelope, $h(t)$ is the difference between $x(t)$ and $m(t)$, then

$$m(t) = \frac{[x_{\max}(t) + x_{\min}(t)]}{2},$$
$$h(t) = x(t) - m(t). \tag{6}$$

Treat $h(t)$ as a new $x(t)$, and repeat the above operations until the $h(t)$ meets certain criteria; let $\mathrm{imf}_1 = h(t)$ and imf_1 is an IMF, then

$$r_1(t) = x(t) - \mathrm{imf}_1. \tag{7}$$

Treat $r_1(t)$ as a new $x(t)$, repeat the above process, and get the second IMF, and the third IMF, ..., n-th IMF. When imf_n or $r_n(t)$ satisfies the given stopping criteria, the screening process is terminated and the solution is obtained as follows:

$$x(t) = \sum_{i=1}^{n} \mathrm{imf}_i + r_n(t), \tag{8}$$

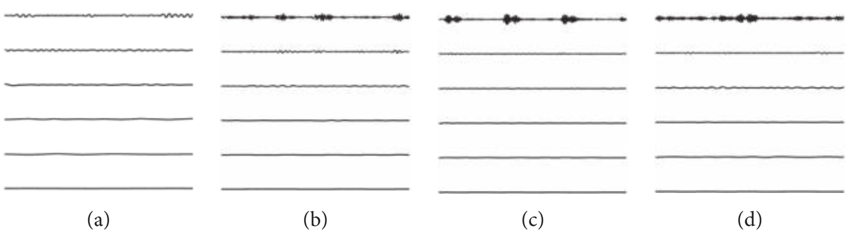


FIGURE 4: IMFA sample: (a) normal state; (b) inner ring fault; (c) outer ring fault; (d) rolling body fault.

where $r_n(t)$ is a residual function, which represents the average trend of the signal.

(2). Image generation. The vibration signal is intercepted according to 400 sampling points in each segment, and the signal is decomposed by EMD in turn, in which the first six IMFs contain the main components of the signal, so the first six groups of IMF are extracted and arranged in the form of six curves in order, as shown in Figure 4.

In the process of image generation, in order to show the fluctuation state of each IMF, the method of adaptively adjusting the vertical coordinate is adopted; i.e., the amplitude A_i with the largest absolute value in the i-th signal segment is obtained, and then, the vertical coordinate range is set as $-A_i \sim A_i$. The advantage of this process is that it can show the complete shape of the IMF component and make the visual contrast between the $\mathrm{IMF}_1 \sim \mathrm{IMF}_6$ and strengthen the sample feature difference.

3.2.2. EP Time-Frequency Analysis Method

(1). EP principle. EP uses the EMD to decompose the multicomponent signal into finite IMFs. For each IMF, the pseudo-Wigner–Ville distribution (PWVD) is carried out, in which the PWVD is shown as follows:

$$\mathrm{PW}_x(t,f) = \int_{-\infty}^{\infty} h(\tau)x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}\mathrm{d}\tau, \quad (9)$$

where t is the time-domain variable, f is the frequency-domain variable, and $h(\tau)$ is the window function.

Then, the Wigner time-frequency distribution of the original signal is obtained by superposing the PWVD results of different IMFs. The combination of EMD and PWVD not only effectively eliminates the cross interference but also retains the excellent time-frequency focusing [10].

(2). Image generation. The steps for generating the time-frequency distribution map are as follows:

(a) Vibration signal segmentation: processing 400 sampling points at a time.

(b) EMD decomposition: the above signals are decomposed by EMD, and the IMF components and residual components from high frequency to low frequency are obtained in turn. Only the first six IMFs are processed in the next step.

(c) EP time-frequency analysis: the first six IMFs are analyzed by PWVD, and then, the EP time-frequency distribution is obtained.

(d) Grayscale processing: in order to reduce the input parameters and improve the training efficiency, the grayscale processing of the generated samples is done.

According to the above steps and Figure 5, we can clearly see that, for the bearing data with different fault states, the energy distribution is different.

3.2.3. SPCI Method. The above two methods obtain the images from vibration signal processing, but the SPCI does not belong to this scope, and it represents the form of a mirror symmetrical image in polar coordinates and directly converts the sampled signal into an image. The graphic display is intuitive and has a strong ability to express features.

(1) SPCI principle. The schematic diagram of SPCI is shown in Figure 6; $\gamma(i)$ is the polar radius, and $\alpha(i)$ and $\beta(i)$ are the rotation angles along the initial line in the counterclockwise direction and clockwise direction, respectively. The principle is that, in the discrete sampling data series, the vibration parameters at i time are x_i, and the vibration parameters at $i+a$ time are x_{i+a}; substituting x_i and x_{i+a} into equations (10)–(12), it can transform vibration data into a point in the $P(\gamma(i), \alpha(i), \beta(i))$ polar coordinate space. By changing the rotation angle of the initial line, a mirror symmetrical image can be formed [29]:

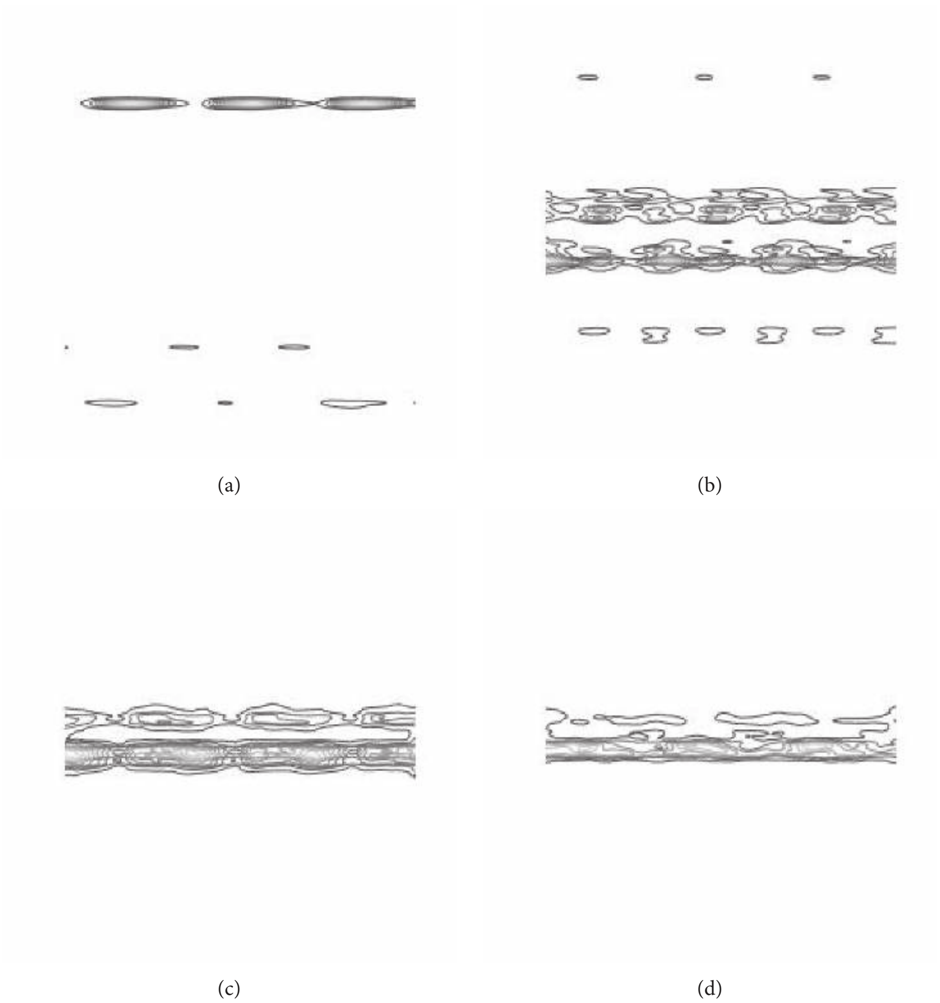
$$\gamma(i) = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad (10)$$

$$\alpha(i) = \varphi + \frac{x_{i+a} - x_{\min}}{x_{\max} - x_{\min}} b, \quad (11)$$

$$\beta(i) = \varphi - \frac{x_{i+a} - x_{\min}}{x_{\max} - x_{\min}} b. \quad (12)$$

In the above formula, $x_{\max}$ is the maximum value of vibration data, $x_{\min}$ is the minimum value, a is the time interval, φ is the initial line angle, and b is the angle magnification factor. It is often taken as $\varphi = 60°$, $a = 3 \sim 10$, and $b = 20° \sim 60°$.

(2) Image generation. In this paper, the image with high resolution is obtained by adjusting the parameters a and b.


(a)

(b)

(c)

(d)

Figure 5: EP sample: (a) normal state; (b) inner ring fault; (c) outer ring fault; (d) rolling body fault.

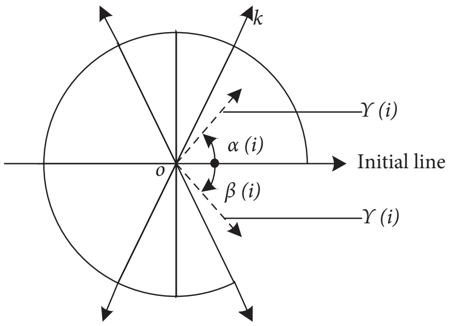


Figure 6: Schematic diagram of SPCI.

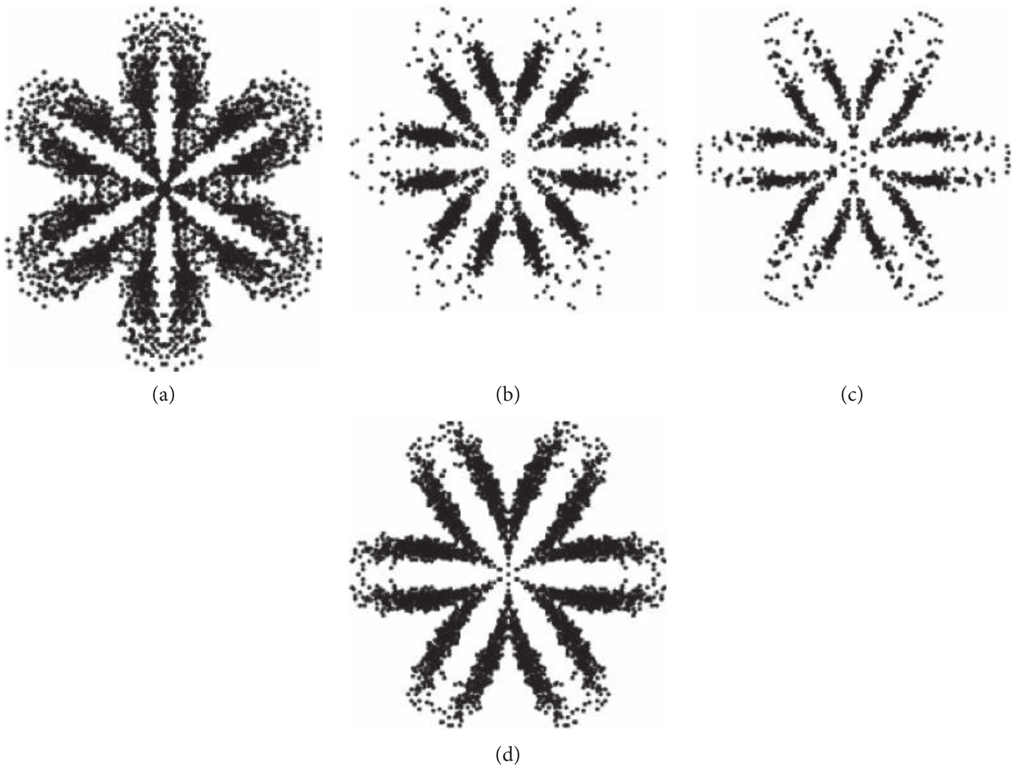
After trying, we take $a = 2$ and $b = 30^{\circ}$, and the SPCI samples generated are shown in Figure 7. It can be found that the SPCI samples of different data are different.

3.2.4. Improved GTM Method. In addition to the above three vibration images, the GTM is introduced in this paper. Before obtaining a GTM, it is necessary to convert the vibration signal into a grayscale image. The grayscale image is a data matrix, in which the subscript of each element corresponds to its position in the image, i.e., row and column coordinates, and the element value represents the luminance value of the corresponding position. The generation process of grayscale image is actually a process of data mapping. The maximum value "max" in the feature matrix is mapped to the gray level 255. The minimum value "min" is mapped to the gray level 0, as shown in Figure 8 [30]. The relationship between the feature matrix and grayscale value in the image is shown as follows:

$$G_{i,j} = \frac{d_{i,j} - \min}{\max - \min} * 255. \tag{13}$$

In the above equation, $d_{i,j}$ is the vibration signal data, in which i, j is the size of feature matrix, and $G_{i,j}$ is the gray value corresponding to $d_{i,j}$.

The data arrangement in traditional grayscale image is "sequential arrangement of vibration data," but this arrangement uses a large number of data points every time. If the image sample resolution is sacrificed and a small number of sample points are used to convert the grayscale image each time, it cannot highlight the characteristics of the original signal. Therefore, this paper uses vibration data to convert



(a) (b) (c)

(d)

Figure 7: SPCI sample: (a) normal state; (b) inner ring fault; (c) outer ring fault; (d) rolling body fault.

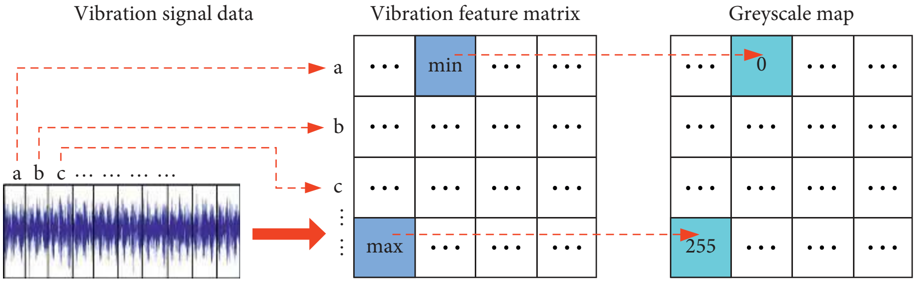


Figure 8: Mapping relationship of grayscale image.

the grayscale image according to the horizontal-vertical cross arrangement; i.e., the i-th vibration signal segment with 400 points is longitudinally copied 400 rows, and the matrix A with 400 ∗ 400 is obtained. The matrix B is obtained by transposing the matrix A into another matrix. The matrix C is obtained by averaging the A and B, and the new grayscale image is obtained by inputting the elements of C into equation (13). The grayscale image processed in this way has two advantages:

(1) Each grayscale image with 400 ∗ 400 needs only 400 signal points, and the data of samples are updated frequently. However, if the traditional grayscale image is used, 160000 signal points are needed to form a grayscale image with the same size.

(2) The horizontal-vertical cross arrangement can form a grid structure in the grayscale image, which can enhance the image texture feature.

After the above conversion, the grayscale image is extracted by the local binary pattern (LBP) [31], and the algorithm is shown as

$$\mathrm{LBP}\left(x_c, y_c\right) = \sum_{p=0}^{p-1} 2^p s\left(i_p - i_c\right), \tag{14}$$

where (x_c, y_c) represents a 3 ∗ 3 neighborhood center element, its pixel value is i_c, i_p represents other pixel values in the neighborhood, p represents the number of pixels in the neighborhood center, and $S(x) = \left\{\begin{matrix} 1, & x \geq 0, \\ 0, & x < 0 \end{matrix}\right\}$.

LBP is used to show the relationship between the pixel value of a certain point in the grayscale image and its surrounding pixel value. The image processed by LBP shows the texture information. Figure 9 is an example from grayscale image to GTM.

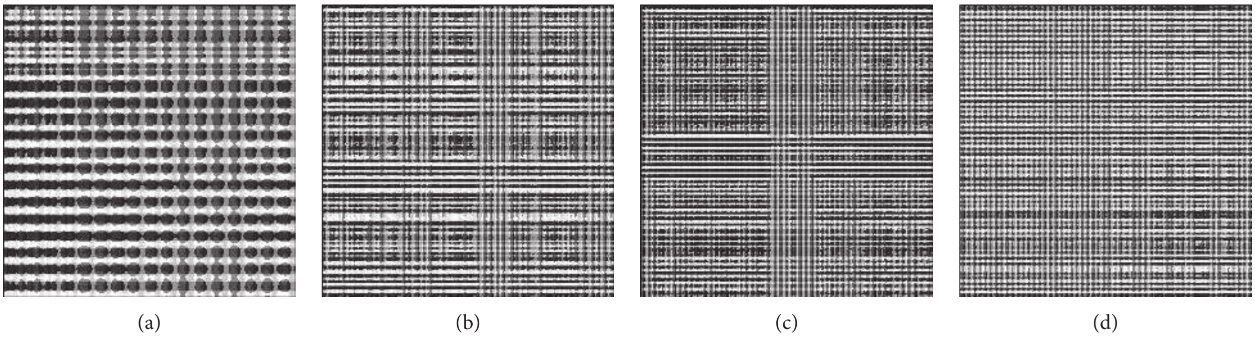


Figure 9: GTM sample: (a) normal state; (b) inner ring fault; (c) outer ring fault; (d) rolling body fault.

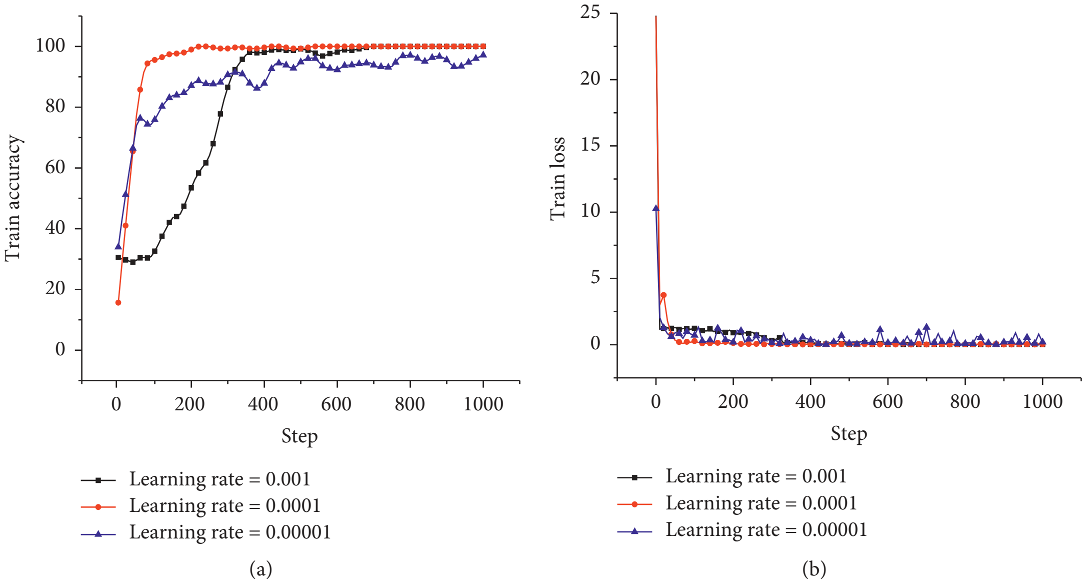


Figure 10: Comparison of training processes with different initial learning rates: (a) training accuracy; (b) training loss.

Table 2: Comparison of test results with different initial learning rates.

Learning rate	0.001	0.0001	0.00001
Test accuracy	98.75%	98.33%	95.42%

The four types of image generation method used in this paper focus on showing the different features between vibration data. In this paper, four kinds of image sample will be obtained, in which each kind of sample contains four kinds of bearing state, and each kind has 1200 samples. The samples is divided into the training set, validation set, and test set according to 6 : 2 : 2, in which the test set does not participate in the training process and only evaluates the final model.

4. Research on the Optimal CNN Model

The task of deep learning is divided into two stages: training and test, in which the training is the process of optimizing the model, such as improving the classification accuracy and strengthening the generalization ability. After many experiments, it is found that the model performance is highly sensitive to the initial learning rate, training optimizer, and regularization parameters, so this paper mainly focuses on these items. As the research methods for four types of image sample are similar, the following only shows the research process of EP, and the other gives their results in the end.

4.1. Determination of Pretraining Model Parameters

4.1.1. Selection of Initial Learning Rate. Learning rate is an important parameter of CNN training. For the training with fixed learning rate, if the learning rate value is small, the high training accuracy can be obtained, but the convergence speed will be affected. If the learning rate value is big, there are the opposite results. To avoid a fixed learning rate, this paper proposes a degenerative learning rate, i.e., to find an equilibrium between training speed and accuracy, and the learning rate decreases with the increase in the number of training steps. The formula is shown as follows:

$$\mathrm{DLR} = \mathrm{lr} * \mathrm{dr}^{(\mathrm{gs}/\mathrm{ds})}, \tag{15}$$

Table 3: Confusion matrix of test results with learning rate = 0.0001.

Bearing states	Normal	Inner ring fault	Outer ring fault	Rolling body fault
Normal	60	0	0	0
Inner ring fault	2	58	0	0
Outer ring fault	0	0	59	1
Rolling body fault	0	0	1	59

where dr is the decay index, lr is the initial learning rate, gs is the current number of iterative rounds, ds is the measure of decay in each iteration, which can be called decay speed, and DLR is the learning rate after decay.

In order to obtain the best training effect, this paper takes the learning rate as 0.001, 0.0001, and 0.00001, respectively. The accuracy and loss curves in the training process are shown in Figure 10. The models trained under the above three parameter settings are tested, and the test set includes 60 bearing samples in each of the 4 states. The test results are shown in Table 2.

In Figure 10(a), when the initial learning rate is 0.0001, the model can quickly converge and maintain a stable state, while the initial learning rate is 0.00001, it cannot smoothly converge and the training accuracy is low, and the early convergence rate at 0.001 is slower than that at 0.0001. In Figure 10(b), the loss value change at 0.00001 is unstable, while the losses at the other two cases remain stable and low, and 0.0001 is the best choice. In Table 2, when the learning rate is 0.001, 0.0001, and 0.00001, the test results are 98.75%, 98.33%, and 95.42%, respectively. After comparison, 0.0001 is selected as the initial learning rate. When the learning rate is 0.0001, the confusion matrix of the test results is shown in Table 3.

4.1.2. Optimizer Determination. In deep learning, there are many optimization methods to find the optimal solution of the model. In this paper, three kinds of optimization algorithms are used to select the most suitable optimizer.

(1). Gradient descent algorithm. It is often used to approximate the minimum deviation model, where the gradient descent direction is to use the negative gradient direction as the search direction, and the minimum value is solved along the gradient descent direction. The most frequently used gradient descent algorithm is stochastic gradient descent (SGD) [32], and SGD formula is shown as follows:

$$g_t = \Delta J_{i_s}\left(W_t, X^{(i_s)}, Y^{(i_s)}\right), \tag{16}$$

$$W_{t+1} = W_t - \eta_t g_t, \tag{17}$$

where $\Delta J_{i_s}(W_t, X^{(i_s)}, Y^{(i_s)})$ is the cost function; W_t is the model parameter at t time; $X^{(i_s)}, Y^{(i_s)}$ are the samples of each input and output, respectively; g_t is the correlation gradient of the cost function at t time; i_s represents a randomly selected gradient direction; and η_t is the learning rate at t time.

(2). Momentum optimization algorithm. Because the SGD optimizer frequently updates the variables, it will cause serious shock to the loss function, easy to fall into the local extremum, and easy to be trapped in the saddle point. Therefore, based on the gradient descent method, the momentum optimization is proposed. The commonly used momentum optimizer is momentum [33], and its formulas are as follows:

$$v_t = \alpha v_{t-1} + \eta_t \Delta J\left(W_t, X^{(i_s)}, Y^{(i_s)}\right), \tag{18}$$

$$W_{t+1} = W_t - v_t, \tag{19}$$

where v_t represents the acceleration accumulated at t time and α indicates the magnitude of the power; generally, $\alpha = 0.9$, which means the maximum speed is 10 times that of SGD.

Momentum adds the inertia in the gradient descent process, which makes the speed with the same gradient direction faster and the renewal speed of the dimension with the change in gradient direction slower so that it can speed up the convergence and reduce the oscillation.

(3). Adaptive learning rate optimization algorithm. The loss in deep learning is usually highly sensitive to some directions of the parameter space. The momentum algorithm can alleviate these problems to some extent, but at the cost of introducing another hyperparameter. And the traditional optimization algorithm needs to set the learning rate to a constant or adjust the learning rate according to the number of training, which greatly ignores the possibility of other changes in the learning rate. Adaptive moment (Adam) estimation is an adaptive optimization algorithm of learning rate [34], and its formulas are as follows:

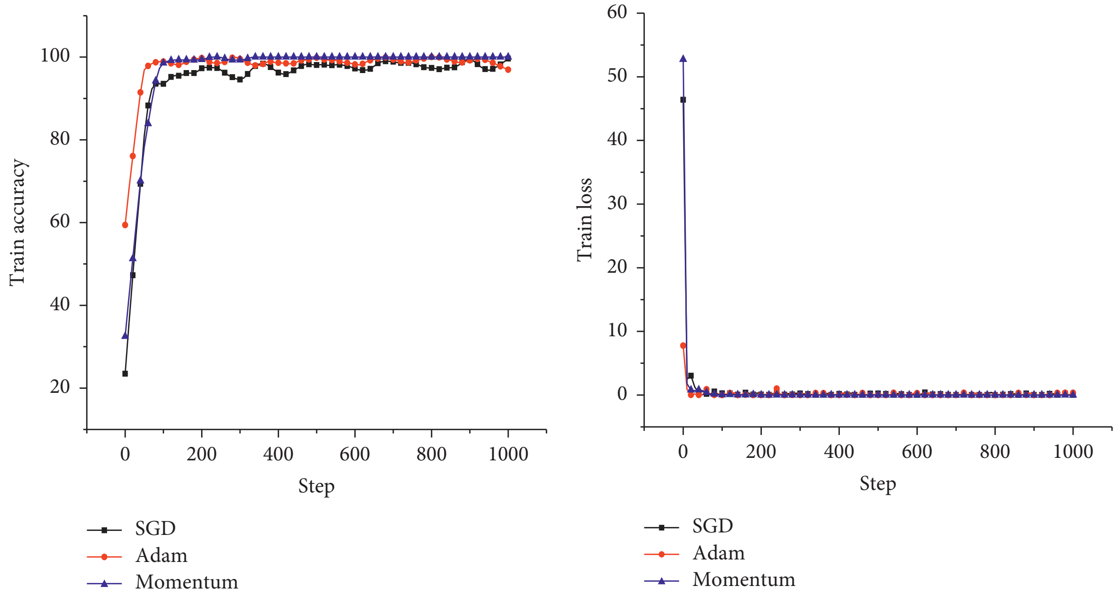
$$\begin{cases} m_t = \mu * m_{t-1} + (1-\mu) * g_t, \\ n_t = \nu * n_{t-1} + (1-\nu) * g_t^2, \end{cases} \tag{20}$$

$$\begin{cases} \widehat{m}_t = \dfrac{m_t}{1-\mu^t}, \\ \widehat{n}_t = \dfrac{n_t}{1-\nu^t}, \end{cases} \tag{21}$$

$$\Delta\theta_t = -\frac{\widehat{m}_t}{\sqrt{\widehat{n}_t} + \varepsilon} * \eta, \tag{22}$$

where m_t, n_t are the first-order and second-order moment estimation of the gradient, respectively; $\widehat{m}_t$, $\widehat{n}_t$ are the corrections to m_t, n_t; and $\Delta\theta_t$ is the learning rate subject to dynamic constraints.

In this paper, under the condition that the initial learning rate is 0.0001 and the other parameters are



(a)

(b)

Figure 11: Comparison of different optimizer training effects: (a) training accuracy; (b) training loss.

Table 4: Comparison of different optimizer test results.

Optimizer	SGD (%)	Adam (%)	Momentum (%)
Test accuracy	98.33	98.33	99.17

unchanged, the model optimizer is set as SGD, momentum, and Adam, respectively. The accuracy and loss comparison curves in the training process are shown in Figure 11. The model test results after training are shown in Table 4.

Compared with the three curves in Figure 11(a), when the Adam is used, the training accuracy slightly fluctuates below 100% with the increase in the number of iterative steps. The curve of training accuracy is stable after 100 steps with momentum. However, SGD curve constantly fluctuates within 1000 steps. There is no significant difference in three loss curves in Figure 11(b). In Table 4, when the optimizer is SGD, Adam, and momentum, the test results of the model are 98.33%, 98.33%, and 99.17%, respectively. After comparison, the momentum optimizer should be selected for this kind of samples. The confusion matrix of the test results with momentum is shown in Table 5.

4.1.3. Regularization Parameter Determination. There are two kinds of abnormal fitting in the training process: overfitting and underfitting. The overfitting means that the model established is too superior in the training samples, resulting in poor performance in the validation and test data sets, while underfitting generally means that the features extracted from the training samples are relatively few, resulting in the training model cannot match well, and the performance is very poor.

In order to solve the overfitting, the called regularization is introduced into the training process. The main purpose of regularization is to control the complexity of the model and reduce overfitting. The basic regularization method is to add a penalty term to the original loss function to "punish" the model with high complexity. In this paper, several commonly used regularization methods, such as $L1$, $L2$, regularization, and dropout, are studied in the training process, and it is found that $L1$ regularization performs the best, so $L1$ regularization is adopted. The sum of squares of weight parameters is added on the training loss function, as follows:

$$C_1 = C_0 + \lambda \sum_{W}^{n} |W|, \tag{23}$$

where C_1 is the final loss value, C_0 is the real loss value, W is the network learning parameter, and λ is the adjustable regularization parameter. In this paper, the effects are compared when λ is 0.1, 0.01, 0.001, and 0.0001 and has no regularization, and the contrast curves of accuracy and the final loss C_1 in the validation process are given, as shown in Figure 12. The model test results after training are shown in Table 6.

In Figure 12(a), all the validation curves under five parameters slightly fluctuate around 98% after 50 to 200 steps, in which $\lambda = 0.1$, 0.01, and 0.0001 are more efficient. Figure 12(b) shows the loss value in the validation process, in which the curve of $\lambda = 0.1$ has a downward trend, but its validation loss values keep higher within 1000 steps. The value of $\lambda = 0.1$ curve stabilizes at about 9 after 20 iterations. In Table 6, the test results of the models under the five regularization settings are 98.75%, 98.75%, 98.75%, 99.17%, and 99.17%, respectively. After comparison, $\lambda = 0.0001$ is selected. The confusion matrix of the test results with $\lambda = 0.0001$ is shown in Table 7. The training, validation accuracy, and loss curves are given, as shown in Figure 13.

Table 5: Confusion matrix of test results with momentum.

Bearing states	Normal	Inner ring fault	Outer ring fault	Rolling body fault
Normal	60	0	0	0
Inner ring fault	2	58	0	0
Outer ring fault	0	0	60	0
Rolling body fault	0	0	0	60

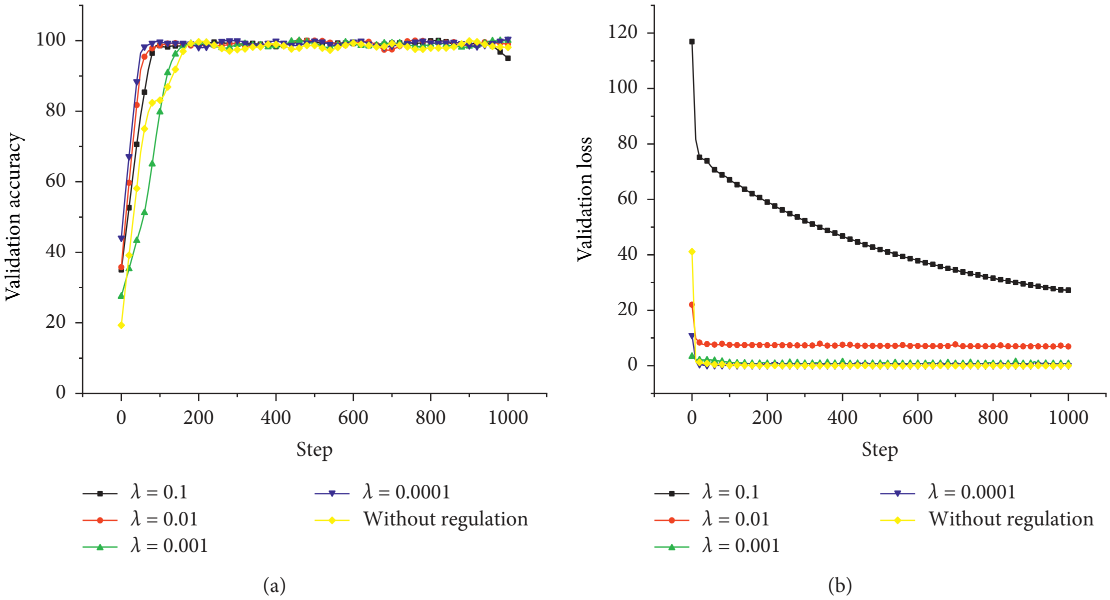


Figure 12: Comparison of validation results under different regularization parameters: (a) validation accuracy; (b) validation loss.

Table 6: Comparison of test results under different regularization parameters.

Regularization	$\lambda = 0.1$ (%)	$\lambda = 0.01$ (%)	$\lambda = 0.001$ (%)	$\lambda = 0.0001$ (%)	Without regulation (%)
Test accuracy	98.75	98.75	98.75	99.17	99.17

Table 7: Confusion matrix of test results with $\lambda = 0.0001$.

Bearing states	Normal	Inner ring fault	Outer ring fault	Rolling body fault
Normal	60	0	0	0
Inner ring fault	2	58	0	0
Outer ring fault	0	0	60	0
Rolling body fault	0	0	0	60

4.2. Optimal CNN Model Parameters for Four Image Generation Methods. Under the condition that other parameters remain unchanged, the other three types of image samples were trained and validated for the above three sensitive parameters, and the final parameters are shown in Table 8.

Under the best model parameters obtained above, four types of sample were tested, respectively, and the number of tested samples in each type is 240. This paper uses the accuracy, precision, recall, and $F1$-score of the test results as the evaluation indexes. The average values of the corresponding indexes for 4 types of samples are given in Table 9. Among them, the accuracy is the proportion of all predictions that are correct, and the precision is positive in all predictions; the recall rate is the proportion that the correct prediction is positive, while the F1-score considers both the accuracy rate and the recall rate to achieve the highest and balance at the same time.

From the test results, it can be found that the model trained by SPCI samples has a good classification effect, and the accuracy reaches 99.19%, which is the best among four kinds of image samples. The classification effect by EP samples is also good, with the test accuracy of 98.75%. GTM and IMFA samples have the test results of 96.67% and 93.75%, respectively.

5. Parameter-Based TL

In the actual industrial field, the collected vibration signals will be disturbed by background noise, and the signals polluted by noise will cover up the effective information in

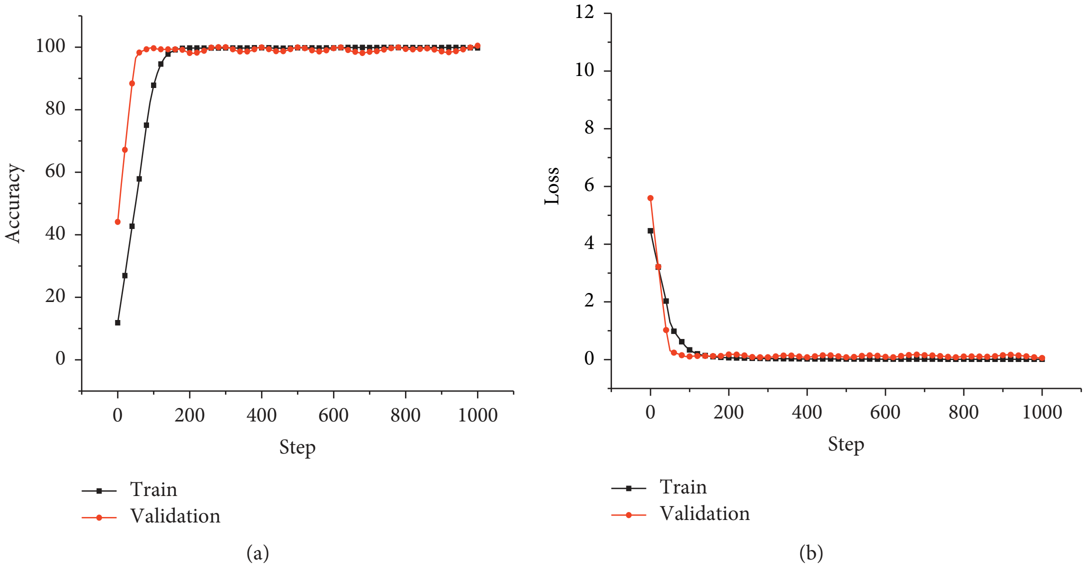


Figure 13: Performance curves of the model with $\lambda = 0.0001$: (a) accuracy; (b) loss.

Table 8: Best parameter determination of four models.

Sample categories	Initial learning rate	Optimizer	$L1$ regularization parameters
EP	0.0001	Momentum	0.0001
IMFA	0.0001	SGD	0.01
SPCI	0.0001	Momentum	0.0001
GTM	0.0001	Momentum	0.1

Table 9: Evaluation data of test results.

Sample categories	Accuracy (%)	Average precision (%)	Average recall rate (%)	Average $F1$-score (%)
EP	98.75	98.75	98.75	98.50
IMFA	93.75	94.25	93.75	94.00
SPCI	99.19	99.25	99.25	99.00
GTM	96.67	96.75	96.75	96.50

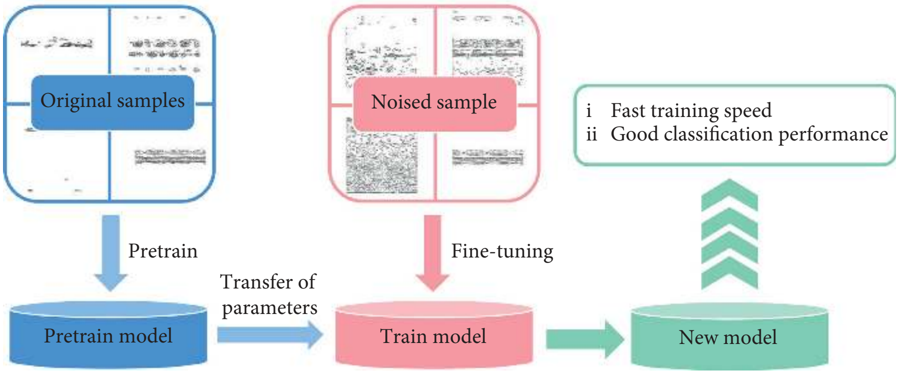
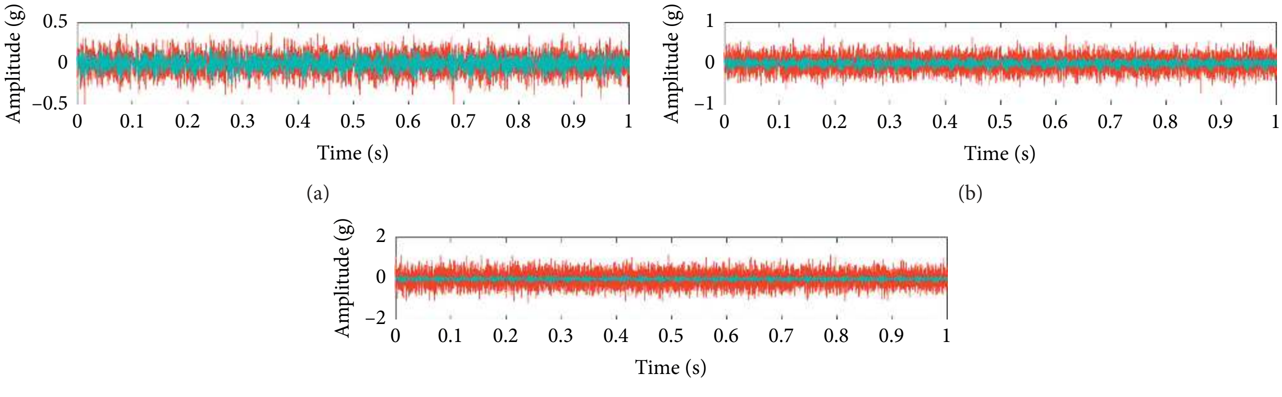


Figure 14: Parameter transfer process.

the original signals, so it is a significant work to identify the fault categories quickly and accurately for the noised signals. The samples used in the pretraining have a certain similarity with the noised samples, which increases the probability of successful parameter transfer. By transferring the model parameters obtained from the pretraining, the slow training process can be avoided and the model efficiency can be improved. In this paper, by adding Gaussian white noise

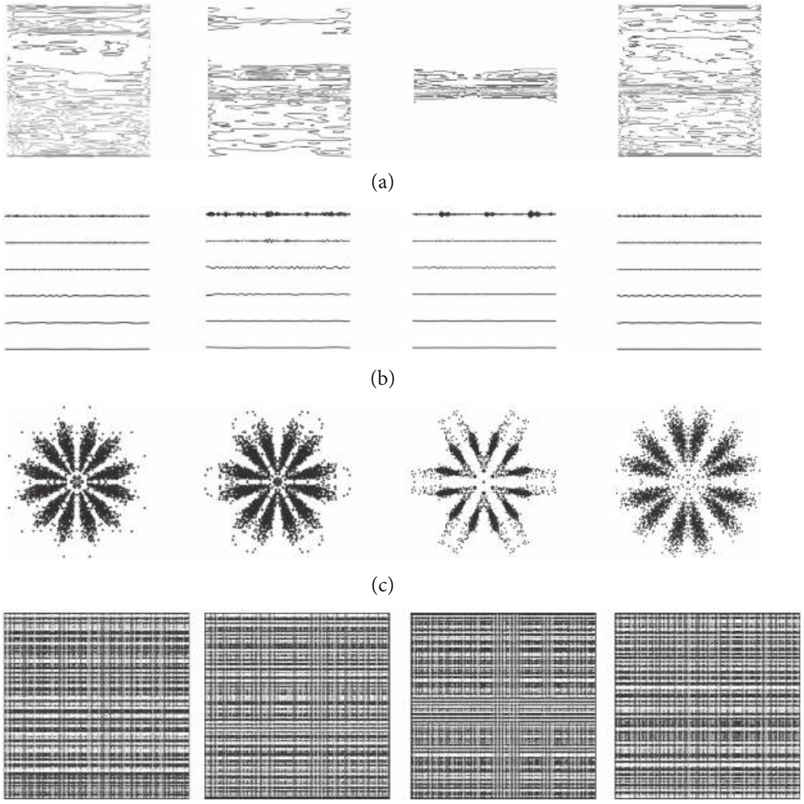


(a)

(b)

(c)

FIGURE 15: Time-domain comparison of noised signals with different SNRs: (a) 22 dB; (b) 16 dB; (c) 10 dB.



(a)

(b)

(c)

(d)

FIGURE 16: Noised image samples with a SNR of 10 dB: (a) EP; (b) IMFA; (c) SPCI; (d) GTM.

TABLE 10: Types of frozen layer combination.

Type	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Frozen layer	None	*C*1	*C*2	*C*3	*F*1	*C*1–*C*2	*C*1–*C*3
Type	*H*	*I*	*G*	*K*	*L*	*M*	*N*
Frozen layer	*C*1-*C*2-*C*3	*C*1-*F*1	*C*1-*C*2-*C*3-*F*1	*C*2-*C*3	*C*2*F*1	*C*2-*C*3-*F*1	*C*3-*F*1

(GWN) to the bearing data of Case Western Reserve University to simulate the actual field signal, the noised signal is converted into image. The designated layers of the training model using noised samples are frozen, and the pretraining parameters by the pure samples are shared. The TL flowchart is shown in Figure 14.

In order to simulate the influence of different degrees of noise on the signal, the GWN with slight, moderate, and severe SNR is added, respectively. Through a test, it is found

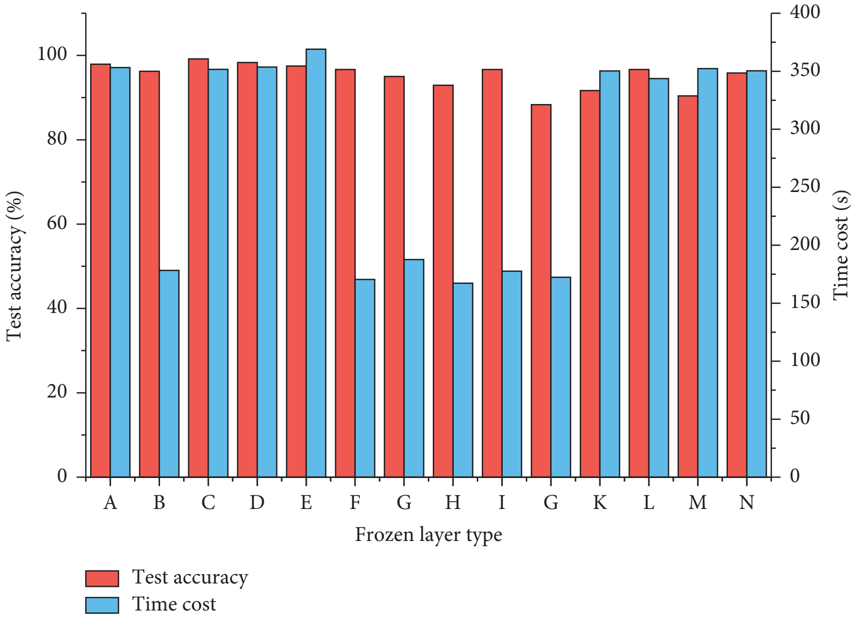


Figure 17: Test accuracy and training time.

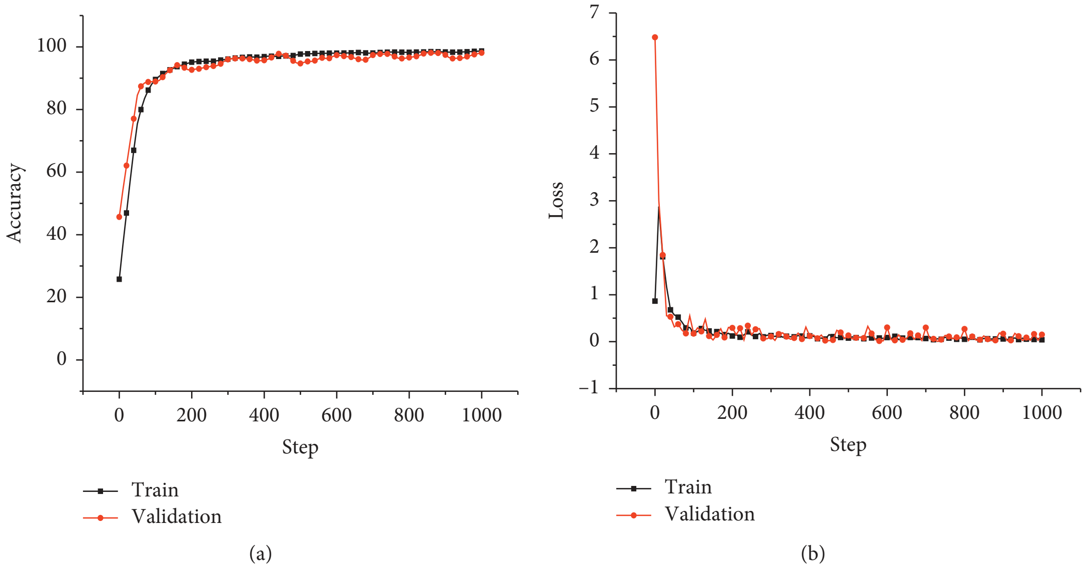


Figure 18: Performance curves of the model with type *F*: (a) accuracy; (b) loss.

Table 11: Summary of test results.

SNR (db)	Parameters	EP	IMFA	SPCI	GTM	Average
22	Freezing type	*H*	*I*	*I*	*H*	—
	Test accuracy	99.58%	97.75%	96.67%	91.67	96.42%
	Time cost (s)	161.52	179.44	194.86	165.77	175.40
16	Freezing type	*B*	*B*	*G*	*F*	—
	Test accuracy	97.92%	90.42%	83.33%	76.67%	87.09%
	Time cost (s)	170.68	171.54	194.14	181.55	179.48
	Relative reduction rate	1.67%	7.50%	13.80%	16.36%	9.68%
10	Freezing type	*F*	*F*	*B*	*G*	—
	Test accuracy	96.67%	78.33%	86.67%	59.58%	80.31 %
	Time cost (s)	170.46	165.36	194.19	175.27	176.32
	Relative reduction rate	1.28%	13.37%	-4.01%	22.29%	7.78%

that when the SNR is 22 dB, the noised signal can drown the pure signal slightly, so the GWN of 22 dB, 16 dB, and 10 dB is added. The time-domain comparison between the pure and noised signals is shown in Figure 15, in which the blue waveform is the pure signal and the red waveform is the noised signal. The conversions of four types of image were performed for the noised signals of 10 dB, as shown in Figure 16.

Firstly, the experiment for 10 db was carried out on the noised samples obtained by the EP method. The layers capable of being frozen are the convolution layer 1 (*C*1), convolution layer 2 (*C*2), convolution layer 2 (*C*3), and fully connected layer 1 (*F*1). The trainable parameters in these four layers are 156, 1516, 48120, and 10164, respectively. In order to study which layer parameters play an important role in the model training process, 13 model freezing schemes are carried out in this paper, and the nonfrozen layer (type A) is used as the contrast group, as shown in Table 10. Figure 17 summarizes the final test accuracy and the time required for each training process.

Among the above $A \sim N$ test results shown in Figure 17, the type *C* (freeze *C*2) has the highest test accuracy, reaching 99.17%, and the time-consuming is 351.61 s. The model with the least time consumption is type *H* (freeze *C*1-*C*2-*C*3), whose test accuracy is 92.92% and takes 167.25 s. In the industrial field, the ideal TL scheme should have a certain accuracy and be able to complete the training process in a short time, so this paper takes the top 7 with high test accuracy among the above 14 types and then selects the type that takes the shortest time, i.e., the ideal transfer type of this paper. Type *F* (*C*1-*C*2) can achieve the desired results, its test accuracy is 96.67%, and it takes 170.46 s; its training, validation, accuracy, and loss curves are shown in Figure 18.

The noised samples with a SNR of 16 dB and 22 db were tested in the same way, and the training and test results were sorted out for the best scheme. The results for the selected freezing type in each case are shown in Table 11. Relative reduction rate in Table 11 represents the reduction rate of the current test value relative to the last test value.

With the noise intensity increase in the original signal, more noise information is introduced into the converted image samples, and the difference between the noised samples and the original pure samples is greater. Because of the influence of noise on the characteristics of the original signal, it weakens the feature differences between different bearing samples, which makes it difficult for CNN to obtain the distinguishable features between different samples.

In the experiment, under the slight noise of 22 db, the four types of image samples have accurate classification results, and their average accuracy is 96.42%. When the noise intensity is increased to 16 dB, the test effect of EP samples is not significantly reduced, and the accuracy is still maintained at 97.92%, while IMFA, SPCI, and GTM are significantly reduced, whose accuracy is reduced by 7.5%, 13.8%, and 16.36%, respectively. Under the strong noise of 10 dB, the relative increase in SPCI samples is 4.01% compared to 16 dB case, and the decrease rates of EP, IMFA, and GTM are 1.28%, 13.37%, and 22.29%, respectively, compared to 16 dB case. It can be seen that IMFA and GTM are more sensitive to noise intensity. However, EP and SPCI can maintain the test accuracy of 96.67% and 86.67% under the strong noise, indicating that they have the good tolerance ability to noise.

It is found that, under the strong noise of 10 db, the training time of four kinds of sample is 170.46 s, 165.36 s, 194.19 s, and 175.27 s, respectively, and EP has the highest test accuracy of 96.67%. Others are 78.33%, 86.67%, and 59.58%. From the results, it can be concluded that the long feature extraction time can be avoided and the model training speed can be accelerated by pretraining the model parameters, but if only the parameter updating of the last fully connected layer is retained, the model learning ability will be weakened. Therefore, different schemes should be tried in the process of parameter transfer to obtain the best transfer learning efficiency, i.e., to improve the speed of network training on the premise of guaranteeing accuracy.

6. Conclusions

(1) In this paper, four vibration image generation methods are discussed, and in order to distinguish the features among different image samples and optimize the resolution, the adaptive IMFA and gridding GTM are proposed, which provide new approaches for vibration image sample preparation.

(2) In order to give full play to the learning efficiency of CNN model, the best model parameters are obtained by adjusting sensitive parameters including learning rate, optimizer, and regularization, and the trained model has accurate classification result when the samples obtained by EP and SPCI are used.

(3) Aiming at the samples with different GWN, the effect of 13 model freezing schemes on TL is studied. Under the strong noise, the model still has good classification effect. Through the specific TL schemes, the training time-consuming of the model is reduced; meanwhile, the test accuracy can be kept at a high level.

Data Availability

The data used to support the results of this study are available from the corresponding author upon request or can be downloaded at Case Western Reserve University website "http://csgroups.Case.edu/bearing/data/center/home."

Conflicts of Interest

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

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant nos. 51605380, 51875451, and 51834006) and Natural Science Basic Research Program of Shaanxi (Program no. 2021JM-391).

References

[1] G. B. Wang, Z. J. He, X. F. Chen, and Y. N. Lai, "Basic research on machinery fault diagnosis-what is the prescription," *Journal of Mechanical Engineering*, vol. 49, no. 1, pp. 64–72, 2013.

[2] Y. G. Lei, F. Jia, D. T. Kong, J. Lin, and S. B. Xing, "Opportunities and challenges of machinery intelligent fault diagnosis in big data era," *Journal of Mechanical Engineering*, vol. 54, no. 5, pp. 94–104, 2018.

[3] R. Q. Yan, R. X. Gao, and X. F. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing*, vol. 96, pp. 1–15, 2014.

[4] Y. H. Jiang, R. Q. Li, W. D. Jiao et al., "Feature extraction method based on empirical mode decomposition and bispectrum analysis," *Journal of Vibration, Measurement & Diagnosis*, vol. 27, no. 2, pp. 338–407, 2017.

[5] H. K. Jiang, C. L. Li, and H. X. Li, "An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 36, pp. 225–239, 2013.

[6] T. Guo and Z. M. Deng, "An improved EMD method based on the multi-objective optimization and its application to fault feature extraction of rolling bearing," *Applied Acoustics*, vol. 127, pp. 46–62, 2017.

[7] Y. P. Cai, A. H. Li, T. Wang, L. Yao, and P. Xu, "Time-frequency analysis of internal combustion engine vibration based on EMD-wigner-ville," *Journal of Vibration Engineering*, vol. 23, no. 4, pp. 430–437, 2010.

[8] J. Ma, "Wigner-ville distribution and research in machine fault diagnosing," *Instrumentation Technology*, vol. 1, pp. 54–56, 2011.

[9] Y. P. Cai, A. H. Li, L. S. Shi, P. Xu, and W. Zhang, "IC engine fault diagnosis method based on EMD-WVD vibration spectrum time-frequency image recognition by SVM," *Chinese Internet Combustion Engine Engineering*, vol. 33, no. 2, pp. 72–78, 2012.

[10] Y. J. Yue, G. Sun, Y. P. Cai, and X. Wang, "Time-frequency analysis of ICE vibration based on VMD-PWVD," *Journal of Wuhan University of Science and Technology*, vol. 39, no. 5, pp. 365–369, 2016.

[11] H. W. Fan, S. J. Shao, X. H. Zhang, X. Wan, X. G. Cao, and H. W. Ma, "Intelligent fault diagnosis of rolling bearing using FCM clustering of EMD-PWVD vibration images," *IEEE Access*, vol. 8, pp. 145194–145206, 2020.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] H. D. Shao, H. K. Jiang, Y. Lin, and X. Q. Li, "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders," *Mechanical Systems and Signal Processing*, vol. 102, pp. 278–297, 2018.

[14] H. D. Shao, H. k. Jiang, X. Q. Liu, and S. P. Wu, "Intelligent fault diagnosis of rolling bearing using deep wavelet autoencoder with extreme learning machine," *Knowledge-Based Systems*, vol. 140, pp. 1–14, 2017.

[15] H. D. Shao, H. K. Jiang, F. Wang, and Y. N. Wang, "Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet," *ISA Transactions*, vol. 69, pp. 187–201, 2017.

[16] P. Wang, Ananya, R. Q. Yan, and R. X. Gao, "Virtualization and deep recognition for system fault classification," *Journal of Manufacturing Systems*, vol. 44, pp. 310–316, 2017.

[17] X. Xiao, J. X. Wang, Y. J. Zhang, Q. Guo, and S. Y. Suo, "A two-dimensional convolutional neural network optimization method for bearing fault diagnosis," *Proceedings of the CSEE*, vol. 39, no. 15, pp. 4558–4567, 2019.

[18] Z. Y. Chen, K. Gryllias, and W. H. Li, "mechanical fault diagnosis using convolutional neural networks and extreme learning machine," *Mechanical Systems and Signal Processing*, vol. 133, pp. 1–20, 2019.

[19] S. Y. Shao, P. Wang, and R. Q. Yan, "Generative adversarial networks for data augmentation in machine fault diagnosis," *Computers in Industry*, vol. 106, pp. 85–93, 2019.

[20] F. Z. Zhuang, P. Luo, Q. He, and Z. Z. Shi, "Survey on transfer learning research," *Journal of Software*, vol. 26, no. 1, pp. 26–39, 2015.

[21] Z. H. Wu, H. K. Jiang, K. Zhao, and X. Q. Li, "An adaptive deep transfer learning method for bearing fault diagnosis," *Measurement*, vol. 151, pp. 107227–107240, 2019.

[22] N. J. Qiu, X. X. Wang, P. Wang, S. C. Zhou, and Y. C. Wang, "Research on convolutional neural network algorithm combined with transfer learning model," *Computer Engineering and Applications*, vol. 56, no. 5, pp. 43–48, 2020.

[23] X. Wang, C. Q. Shen, M. Xia et al., "Multi-scale deep intraclass transfer learning for bearing fault diagnosis," *Reliability Engineering and System Safety*, vol. 202, pp. 1–15, 2020.

[24] H. W. Fan, S. Q. Gao, X. H. Zhang et al., "Intelligent recognition of ferrographic images combining optimal CNN with transfer learning introducing virtual images," *IEEE Access*, vol. 99, pp. pp1-1, 2020.

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724, Columbus, OH, USA, June 2014.

[26] J. Yosinski, J. Jason, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proceedings of the International Conference on Neural Information Processing Systems*, MIT Press, Cambridge; MA, USA, December 2014.

[27] The Case Western Reserve University Bearing Data Center, "Bearings vibration data set," 2015, http://csgroups.case.edu/bearing/data/center/home.

[28] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings Mathematical Physical & Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[29] L. L. Zhang, J. C. Ren, H. J. Feng, Z. K. Zhu, and Y. K. Xiao, "Diagnosis of diesel engine crankshaft bearing fault based on symmetric polar coordinates and image recognition," *Chinese Internal Combustion Engine Engineering*, vol. 36, no. 4, pp. 144–149, 2015.

[30] L. L. Zhang and J. Xiao, *Case Study of Mechanical Fault Diagnosis Technology Based on Matlab*, Higher Education Press, Beijing, China, 2016.

[31] L. Liu, P. Fieguth, M. Pietikäinen, and S. Lao, "Median robust extended local binary pattern for texture classification," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1368–1381, 2016.

[32] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, https://arxiv.org/abs/1609.04747.

[33] Y. Heaton, I. G. Jeff, Y. Bengio, and A. Courville, "Deep learning, genetic programming and evolvable machines," 2017.

[34] D. P. Kingma and J. L. Ba, "Adam:a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, pp. 1–15, Vancouver, BC, Canada, May 2016.