Mathematical Biosciences and Engineering

MBE, 19(8): 8057–8080.
DOI: 10.3934/mbe.2022377
Received: 13 April 2022
Revised: 18 May 2022
Accepted: 23 May 2022
Published: 02 June 2022

http://www.aimspress.com/journal/MBE

Research article

Bearing fault diagnosis based on wavelet sparse convolutional network and acoustic emission compression signals

Jinyi Tai[1,2], Chang Liu[1,2,*], Xing Wu[1,3] and Jianwei Yang[1,2]

[1] Key Laboratory of Advanced Equipment Intelligent Manufacturing Technology of Yunnan Province, Kunming University of Science & Technology, Kunming 650500, China
[2] Faculty of Mechanical & Electrical Engineering, Kunming University of Science & Technology, Kunming 650500, China
[3] Yunnan Vocational College of Mechanical and Flectrical Technology, Kunming 650203, China

* **Correspondence:** Email: liuchang3385@gmail.com.

Abstract: A bearing is an important and easily damaged component of mechanical equipment. For early fault diagnosis of ball bearings, acoustic emission signals are more sensitive and less affected by mechanical background noise. To cope with the large amount of data brought by the high sampling frequency and high sampling points of acoustic emission signals, a compressed sensing processing framework is introduced to research data compression and feature extraction, and a wavelet sparse convolutional network is proposed for resolved diagnosis and evaluation. The main research objective of this paper is to maximize the compression rate of the signal under the constraint of ensuring the reconstruction error of the acoustic emission signal, which can reduce the data volume of the acoustic emission signal and reduce the pressure of data analysis for subsequent fault diagnosis. At the same time, a wide convolution kernel based on a continuous wavelet is introduced when designing the neural network, and the energy information of different frequency bands of the signal is extracted by the wavelet convolution kernel to characterize the fault characteristics of the equipment. The energy pooling layer is designed to enhance the deep mining ability of compressed features, and the regularized loss function is introduced to improve the diagnostic accuracy and robustness through feature sparseness. The experimental results show that the method can effectively extract the fault characteristics of the bearing acoustic emission signal, improve the analysis efficiency and accurately classify the bearing faults.

Keywords: acoustic emission signal; compressed sensing; wavelet transform; convolution kernel; energy pooling layer; sparse feature

1. Introduction

Mechanical equipment often fails, and bearing failure is the main factor causing mechanical equipment failure [1]. In rotating machinery, 70% of failures are caused by rolling bearings. Among all kinds of failures in the gearbox, bearing failures are second only to gears and account for 19%, and 80% of motor failures are motor bearing failures. Therefore, if the bearing fails, it will cause downtime, cause serious economic losses and even lead to major safety accidents in addition to the failure of machine. Studying the common fault types and fault mechanisms of bearings and carrying out fault monitoring, diagnosis and evaluation of bearings can effectively prolong the service life of bearings and mechanical equipment and ensure the normal operation of mechanical equipment [2]. One of the difficulties in the field of condition monitoring is accurately detecting the early faults of bearings and carrying out timely maintenance and repair [3]. Therefore, it is of great practical significance to study the causes of bearing faults and fault diagnosis techniques. In recent years, many new bearing fault diagnosis methods have appeared and have achieved good results. Ni et al. proposed a fault information-based VMD (FIVMD) method to extract weak bearing repetitive transients, which can successfully fault early fault signals, impulse noise signals and planetary bearing signals under complex operating conditions of rolling bearing diagnosis [4]. Ji et al. proposed a new robust IFB selection method based on correlation entropy fault energy (FECgram for short), which can capture fault symptoms for fault diagnosis without being affected by fault-independent impulses and cyclostationary disturbances [5].

When the bearing fails in the early stage, its fault signal characteristics are extremely weak, and the surrounding noise signals can easily cover them. It is difficult to effectively diagnose the bearing using the vibration signal [6]. Compared with the vibration signal, the acoustic emission signal has weak fault features that are more sensitive, so acoustic emission signals are also widely used in fault diagnosis [7]. However, the high sampling frequency of acoustic emission signals requires high sampling equipment which increases the amount of data. A large amount of data needs to occupy a large data storage space, which affects the efficiency of fault diagnosis. In recent years, the theory of compressed sensing has begun to rise, which provides a new approach for massive data processing. Donoh et al. proposed a new signal acquisition and processing method based on sparse representation and signal approximation theory, namely, compressed sensing [8]. The principle of compressed sensing is to sample the signal much lower than the Nyquist sampling theorem suggests and to reconstruct the original signal with high probability. In other words, the original signal is recovered from the compressed data, so the compressed sensing method can greatly reduce the redundancy of the data while retaining the information in the original signal, improving the data transmission efficiency during fault diagnosis, and saving storage space. An algorithm for bearing fault diagnosis through compressed sensing and match pursuit (MP) rebuilding was proposed by Zhang et al. [9]. Gang et al. proposed detecting characteristic harmonics from sparse measurement data using a compressed MP strategy in the case of incomplete signal reconstruction [10]. Wang et al. proposed a fault diagnosis means for compressing sparse time-frequency feature indication through compressed sensing, which can refactor the time-frequency features of fault signals from a small quantity of compressed sampling data including noise [11]. Tang et al. proposed a method combining sparse indication and random dimension reduction to acquire and classify rotating machinery fault features [12]. Liu et al. proposed a means to immediately acquire compression characteristics of acoustic emission signals from compressed sensing data for the evaluation of rotating machinery operating conditions [13].

Traditional shallow learning-based fault diagnosis methods rely too much on the experience of diagnostic experts in signal processing and feature extraction. A better solution to this problem has

been proposed using deep learning. Deep learning has a strong automatic feature extraction ability and can automatically learn representative features from data. Therefore, by combining compressed sensing with deep learning, compressed sensing greatly reduces data redundancy while retaining most of the information. Margin, with the help of deep learning to strengthen feature mining and extraction, improves the accuracy of fault diagnosis and is used in the current research in the field of fault diagnosis. Ahmed et al. proposed a method to classify bearing faults in compression measurements using sparse overcomplete features and deep neural networks (DNNs) combined with a sparse-autoencoder-(SAE)-based unsupervised feature learning algorithm [14]. Shao et al. proposed a novel method based on compressed sensing (CS) to improve the convolutional deep belief network (CDBN), which enhanced the feature learning ability of mechanical equipment vibration compressed data [15]. Yuan et al. proposed a bearing fault diagnosis means combining compressed sensing and a heuristic neural network [16]. Aiming at the nonstationarity and nonlinearity of the vibration signal of a planetary gearbox, Chen et al. proposed a fault state identification method for planetary gearboxes based on convolutional neural networks and discrete wavelet transforms [17].

Based on the above analysis, this paper performs random dimension reduction projection on the acoustic emission signal to obtain the compressed signal as the input of the neural network, designs a continuous wavelet wide convolution kernel (CWConv) to improve the first layer structure of the CNN and designs an energy pool according to the characteristics of the energy enrichment of the compressed signal frequency band. The regularization penalty term is introduced into the network loss function to improve the sparsity of features, improve the network learning ability and improve the diagnosis accuracy. The effectiveness of the method is proven by a thrust ball bearing experiment. The experimental results show that the method proposed in this paper has great advantages in accuracy and robustness. The main contributions of this paper include the following:

1) Combining compressed sensing theory with acoustic emission signal analysis while retaining effective information for diagnosis reduces the demand for acoustic emission data for model analysis and improves model training and diagnosis efficiency.
2) The traditional random convolution layer is replaced with a continuous wavelet wide convolution layer, a wavelet is used to extract the frequency band information of the compressed signal, and the compressed domain information of different frequency bands is extracted.
3) The energy pooling layer is designed to mine the band information energy features of the compressed signal, enhance the ability of the 1D convolutional network to mine signal energy features and fully mine the hidden features of the data.
4) The regularization penalty term is added to improve the model loss function so that the network can learn sparsity features during the learning process.

2. Theoretical background

2.1. Compressed sensing

Compressed sensing theory comes from norm function analysis and approximation theory. It is a theory based on functional analysis, matrix analysis, sparse representation, optimal reconstruction and other theories. The sparseness of signals is an important prerequisite for compression. However, most signals, in reality, do not have sparsity, so it is necessary to find a specific orthonormal basis $\psi \in R^{N\times N}$ so that the signal $x \in R^{N\times 1}$ can be represented on the basis ψ :

$$x = \psi\theta . \tag{1}$$

In the formula, the transformation coefficient θ is sparse, containing only a small number of nonzero items, and most of the other items have zero values. The observation base $\Phi \in R^{M \times N}$ is selected, and the signal is projected to the observation base Φ to obtain the dimensionality-reduced projection data to realize data compression:

$$y = \Phi x . \tag{2}$$

$y \in R^{M \times 1}$ is the compressed signal obtained after compression. According to the sparsity property, when y, ψ and Φ are known, the reconstruction of the signal is achieved by optimizing the norm l_1 and a greedy algorithm [18,19].

2.2. Random energy preservation properties

According to compressed sensing theory, for the signal $x \in R^{N \times 1}$, the $M \times N$ dimensional Gaussian random matrix Φ is used as the projection matrix to compress the signal. The compressed signals $y = \Phi x$ and $y = R^{M \times 1}$, where $M << N$, are obtained after compression. According to the projection matrix isometric (distance preserving property, DPP) condition [20], the random projection distance maintenance of the time domain signal x can be described as

$$(1-\varepsilon)\|x\|_2^2 \le \|y\| \le (1+\varepsilon)\|x\|_2^2 , \tag{3}$$

where $\varepsilon \in (0,1)$ in the formula. From the perspective of energy, the l_2-norm can be expressed as the energy of a vector, so the energy of the signal obtained from the above formula is approximately unchanged before and after projection. Accordingly, the energy characteristic parameter E that defines the time domain signal and the energy characteristic parameter $\hat{E}$ of the compressed domain signal are described as

$$E = \|x\|_2^2 = \sum_{i=1}^{N} x_i^{\,2}, \hat{E} = \|y\|_2^2 = \sum_{i=1}^{N} y_i^{\,2} . \tag{4}$$

The random projection energy retention can be described as

$$\left|\hat{E} - E\right| \le C , \tag{5}$$

where C is a constant with a small value.

According to Eq (5), the signal energy remains approximately constant during the random projection process, that is, the energy characteristic E of the signal in the compression domain and the energy characteristic E of the time domain signal are approximately the same. Based on this property, it can be seen that in the process of signal analysis of the bearing, the results of analysis and diagnosis using the energy characteristics of the compressed data are approximately consistent with the diagnosis results obtained from the energy characteristic parameters of the original signal in the time domain.

2.3. Continuous wavelet transform

The principle of the continuous wavelet transform is to use the scale parameter $s > 0$ and translation parameter $r \in R$ of the mother wavelet $\psi(\cdot)$ to represent a general wavelet dictionary $\{\psi_{r,s}(t)\}$ in the time domain [21–23]:

$$\psi_{r,s}(t)=\frac{1}{\sqrt{s}}\psi(\frac{t-r}{s}). \tag{6}$$

In the formula, s is the scale parameter, which is inversely proportional to the frequency, r is the translation parameter, and t is the time. The scale parameter s represents that the signal can be scaled and transformed. The scale parameter determines the amount of information we obtain. The larger the scale parameter is, the less detailed information we obtain. In contrast, the smaller the scale parameter is, the more detailed information we obtain [24]. The inner product of the signal $x(t)$ and the wavelet dictionary $\psi_{r,s}(t)$ is the process of the continuous wavelet transform of the signal, which can be expressed as follows:

$$CWT_f(r,s)=\left\langle x(t),\psi_{r,s}(t)\right\rangle=\frac{1}{\sqrt{s}}\int x(t)\psi^*(\frac{t-r}{s})dt, \tag{7}$$

where $\psi^*(\cdot)$ is the complex conjugate of the mother wavelet $\psi(\cdot)$. Equation (7) shows that the CWT (continuous wavelet transform) and the Fourier transform process are roughly the same, and the Fourier transform is the inner product of the signal and the trigonometric function. The Fourier transform is usually carried out over the entire frequency band, so it does not have multiscale properties, whereas the wavelet transform can achieve the multiscale analysis of the nonstationary signal by scaling parameters and shifting parameters [24]. The CWT (continuous wavelet transform) mainly includes three steps, as shown in Figure 1:

1) A wavelet basis function is chosen, and the $CWT_f(r,s)$ formula is used to calculate the similarity factors with some signals.

2) The wavelet is continuously shifted to the right by r, and the similarity factors are calculated until all parts of the signal are calculated.

3) The scale parameter s of the wavelet is altered, and the above procedures are duplicated to complete the analysis of all scales.

Through the above steps, the signal $x(t)$ is decomposed into a series of similarity coefficients and then combined with the wavelet dictionary, and the signal is projected onto the two-dimensional (2-D) time dimension and scale dimension.

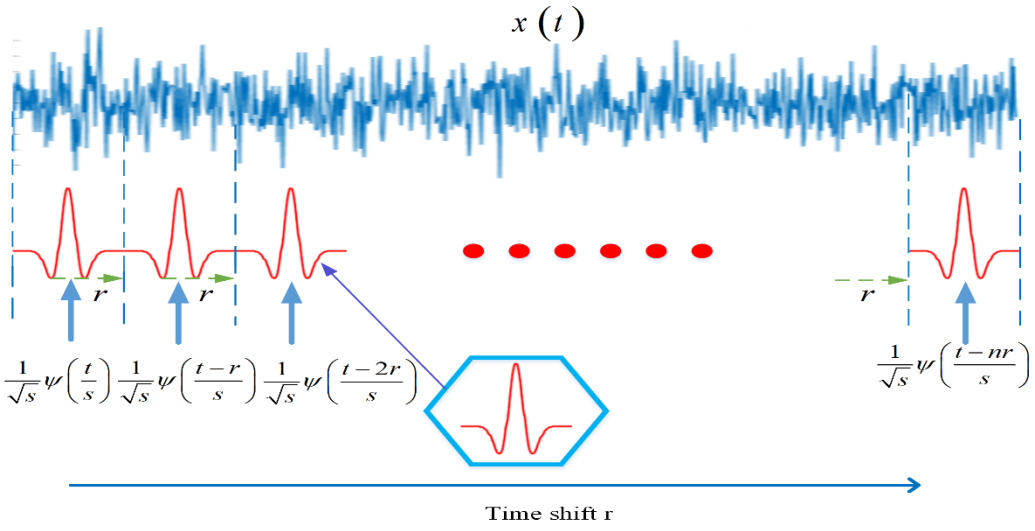


Figure 1. Process of continuous wavelet transform.

3. One-dimensional convolutional neural network improved wavelet sparse convolutional network

One-dimensional CNNs have features such as local connections, weight sharing and downsampling and have been widely used in the field of fault diagnosis [25,26]. An input layer, multiple hidden layers and an output layer together form a one-dimensional CNN, where the input layer is used to input data, the hidden layer is mainly used to extract features, and the output layer is used to output results. Typically, hidden layers include one-dimensional convolutional layers, one-dimensional pooling layers, activation layers and fully connected layers.

3.1. Continuous wavelet convolutional layer

In a standard one-dimensional convolution layer, each randomly initialized convolution kernel is convolved with the input signal, and the process of the convolution operation can be expressed as

$$h_k^l = w_k^l * x(t) + b_k^l, \tag{8}$$

where w_k^l represents the weight of the k-th convolution kernel of the l-th layer, $x(t)$ represents the input signal, b_k^l represents the deviation, and $*$ is the convolution operator. An activation function is required after the convolution operation, and the activation process can be expressed as follows:

$$y_k^l = f(h_k^l), \tag{9}$$

where $f(x)$ is the activation function of the l-th convolutional layer, and y_k^l is the output of the feature map after activation.

The first convolutional layer is displaced by a continuous wavelet convolutional layer (CWConv). By choosing basis functions $\psi_{r,s}(t)$ of different scales, the first convolutional layer of the convolutional neural network is replaced by a continuous wavelet convolutional layer, where the wavelet convolutional layer is composed of wavelet kernels with different scale parameters and translation parameters [24,29], as shown in Figure 2.

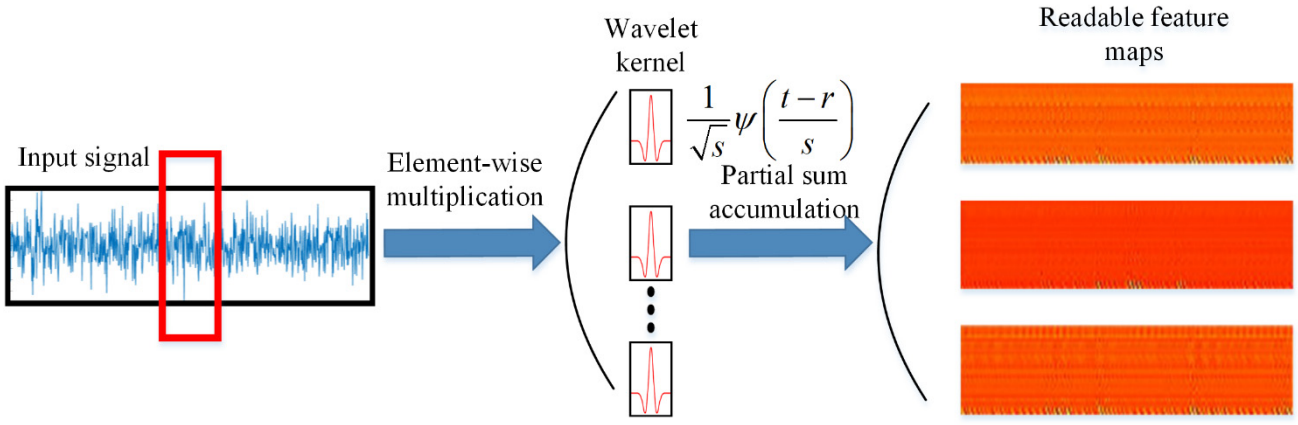


Figure 2. Schematic diagram of wavelet convolution layer operation.

The time domain is converted to the frequency domain by convolving the signal with the defined $\psi_{r,s}(t)$ and using convolution kernels of different scales to obtain different frequency band information of the signal. This function only depends on two learnable parameters, r and s, as follows:

$$h = \psi_{r,s}(t) * x(t) \tag{10}$$

where $x(t)$ represents the input signal, and the basic functions $\psi_{r,s}(t)$ of different scales can be wavelet functions with time-domain expressions, representing the result of continuous wavelet convolution operations. The signal will distribute some local feature information in different frequency bands. Combining the feature information in all frequency bands can effectively improve the fault classification accuracy.

In the process of backpropagation [27], the CWConv layer, as the first layer, only needs to update the parameters r and s, and we have

$$\begin{cases} \delta_{r_k} = \dfrac{\partial G}{\partial r_k} = \dfrac{\partial G}{\partial z_k}\dfrac{\partial z_k}{\partial g_k}\dfrac{\partial g_k}{\partial \psi_{r,s}^k}\dfrac{\partial \psi_{r,s}^k}{\partial r_k} \\ \delta_{s_k} = \dfrac{\partial G}{\partial s_k} = \dfrac{\partial G}{\partial z_k}\dfrac{\partial z_k}{\partial g_k}\dfrac{\partial g_k}{\partial \psi_{r,s}^k}\dfrac{\partial \psi_{r,s}^k}{\partial s_k} \end{cases}, \tag{11}$$

$$r_k = r_k - \lambda\delta_{r_k}, s_k = s_k - \lambda\delta_{s_k}, \tag{12}$$

where λ is the derivative operator, $\psi_{r,s}^k$ is the k-th wavelet filter of the first layer of length L, and r_k and s_k are the translation parameters and scale parameters, respectively. Parameters r and s will be updated by subtracting the product of learning rate λ and gradient δ. For example, the partial derivatives $\partial\psi_{r,s}/\partial r$ and $\partial\psi_{r,s}/\partial s$ of the Laplace wavelet dictionary (Laplace) [28] can be derived by the following formulas. First, the Laplace wavelet is a complex exponential wavelet number with unilateral decay, expressed as

$$\psi(t) = Ae^{-((\beta/\sqrt{1-\beta^2})+u)w_c t}, \quad t > 0, \tag{13}$$

$$\psi(t) = 0, \quad t > 0, \tag{14}$$

where β corresponds to the frequency bandwidth of the wavelet in the frequency domain, and frequency w_c determines the number of significant oscillations of the wavelet in the time domain and corresponds to the center frequency of the wavelet in the frequency domain. A is an arbitrary scale factor used to normalize the wavelet function. Second, the resulting wavelet dictionary is

$$\psi_{r,s}(t) = \frac{1}{\sqrt{s}} A e^{\frac{-((\beta/\sqrt{1-\beta^2})+u)w_c(t-r)}{s}}. \tag{15}$$

We introduce $\partial\psi_{r,s}/\partial r$ and $\partial\psi_{r,s}/\partial s$ into Eqs (11) and (12) to update the parameters s and r of the Laplace wavelet convolutional layer. The continuous wavelet convolution layer can automatically mine the fault feature information of signals in different frequency bands [29] and accomplish adaptive feature extraction for local feature information of different frequency bands.

3.2. Energy pooling layer

For the standard one-dimensional pooling layer, the maximum pooling method is generally used, and the calculation method of convolution is as follows:

$$p_i^l(u) = \max_{(u-1)w+1 \le t \le uw} \left\{ q_i^{l-1}(t) \right\}. \tag{16}$$

In the formula, $p_i^l(u)$ is the value corresponding to the neuron in the l layer, w is the width of the pooling area, and $q_i^{l-1}(t)$ is the output value of the i-th neuron in the $l-1$ layer, where $t \in \left[(u-1)w+1, uw\right]$. Based on the principle of compressed sensing, the continuous wavelet convolution is mined to the energy information of different frequency bands of the signal to represent the equipment fault characteristics, and the energy pooling layer is designed to mine the energy characteristics of the frequency bands by using the energy information of different frequency bands [27]. Root mean square (RMS) is a time-domain statistical feature used to describe signal energy. It has the characteristics of stability and good repeatability in diagnostic indicators. It is an important indicator for judging equipment operating status and diagnosing component failures. For signal X, the energy calculation formula for the RMS value is

$$X_{rms} = \sqrt{\frac{\sum_{i=1}^{M} x_i^2}{M}}. \tag{17}$$

The continuous wavelet convolution layer mines the fault feature information of signals in different frequency bands and uses the energy pooling method to mine the energy characteristics of different frequency bands in the fault characteristic information. For signal X, the calculation formula of the signal channel energy pooling method is

$$f(X) = \sqrt[p]{\frac{\sum_{x \in X} x^p}{M}}. \tag{18}$$

The main function of the pooling layer is to reduce the dimension, compress the features and reduce the amount of calculation. Combining Eqs (17) and (18), when p is infinity, it is equivalent to a max pooling operation; when $p=1$, it is equivalent to an average pooling operation. In this paper, $p=2$ is taken. When $p=2$, the energy pooling calculation method is similar to the root mean square value of the input channel. It can be used to describe the band energy index and reflect the abnormal situation in the nonstationary signal. Maximum pooling mainly reduces the dimension by taking the maximum feature points in the neighborhood, average pooling mainly reduces the dimension by averaging the feature points in the neighborhood, and the energy pooling calculation method extracts the average value of the input channel. The square root energy statistical value is used to reduce the feature dimension, which is used to describe the signal and the root mean square value of the energy statistical characteristic index. It has the characteristics of stability and good repeatability in the diagnosis index and is an important indicator of failures such as component wear.

3.3. Fully connected layer

The convolutional layer extracts local features, and the fully connected layer reassembles the previous local features extracted by the convolutional layer and the pooling layer into a complete column through the weight matrix, which is expressed as

$$\delta_k^l = f(w_k^l p_k^l + b_k^l). \tag{19}$$

In the formula, $k \in (1,n)$ with a total of n neurons, l represents the number of layers, w_k^l represents the weight of the k-th neuron in the l-th layer, b_k^l represents the threshold of the k-th neuron in the l-th layer, δ_k^l represents output of the k-th neuron in the l layer, and $f(x)$ is the activation function.

3.4. Softmax classification layer

The model of the convolutional network usually has a softmax layer at the end. The main function of this layer is to output the probability of each class. The calculation formula is

$$s(i) = \frac{e^{\delta_i}}{\sum_j^J e^{\delta_j}}. \tag{20}$$

In the formula, $s(i)$ is the probability of each output, and the sum of all $s(i)$ is 1; J is the number of categories of multiclassification problems.

3.5. Regular penalty term for the loss function

In the process of model training, the convolutional network generally selects cross entropy as the loss function, which can be expressed as

$$L(\hat{y}, y) = -[y \cdot \log \hat{y} + (1-y) \cdot \log(1-\hat{y})]. \tag{21}$$

In deep learning, it is often seen that an extra item is added after the loss function. There are generally two common extra items. One is l_2 regularization, and the other is l_1 regularization. l_1 regularization and l_2 regularization can be regarded as penalty terms of the loss function, and the penalty terms impose some restrictions on some parameters in the loss function. l_1 regularization can produce a sparse weight matrix, that is, a sparse model that can be used for feature selection. A sparse matrix refers to a matrix in which many elements are 0 and only a few elements are nonzero values, that is, most of the coefficients of the obtained linear regression model are 0. Usually, there are many features in deep learning, and it is difficult to select too many features during classification. However, if the model obtained by substituting these features is a sparse model, it means that only a few features contribute to the model, and most features do not contribute or contribute very little (because the coefficients in front of them are 0 or very small values; even if they are removed, it does not affect the model). At this time, we can only focus on the features whose coefficients are nonzero values. l_2 regularization can prevent model overfitting (overfitting); to a certain extent, l_1 can also prevent overfitting.

Therefore, to extract the sparsity feature from the compressed data, the l_1 penalty term is added to the loss function. $\Omega(h) = \lambda \sum_i |h_i|$ at this time, and the loss function becomes

$$L(\hat{y}, y) + \Omega(h) = -[y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y})] + \lambda \sum_i |h_i|, \tag{22}$$

where h_i represents the activation value of the output of each neuron in layer n-1. All the neurons are also located in network, and λ is the regularization coefficient. y is the true label, and $\hat{y}$ is the model prediction. When 1 represents that the true label is true, the closer the predicted output is to the true label 1, the smaller the loss function L; the closer the predicted output is to 0, the larger the loss function L. When 0 represents that the true label is true, it also conforms to the true situation.

4. Fault diagnosis of thrust ball bearing acoustic emission signals based on a wavelet sparse convolutional network

In this paper, the fault diagnosis of thrust ball bearings is carried out based on the acoustic emission signal of compressed sensing. Combining compressed sensing theory with a wavelet sparse convolutional network, compressed sensing effectively solves the problem of data redundancy in acoustic emission signals, but some time-domain features will be lost in the process of compression, and similar random features will be generated. At this time, the conventional acoustic emission signal analysis method can no longer meet the requirements. Because the compression of the signal is based on the energy preservation property of random projection, the compressed data maintain the same signal structure and energy as the original signal, and the wavelet sparse convolutional network can realize the independent selection of the frequency band energy features, including the primary feature information, avoiding subjective parameter settings and uncertain effects on the experimental results. After compressing and sampling the original signal, the amount of data is greatly decreased. The compressed data of the signal are used as input, and the feature information of different frequency bands is extracted from the input signal through wavelet convolution. Finally, the energy pooling layer is used for mining the energy characteristics of each frequency band. To extract sparsity features from compressed data, an $l1$ penalty term is added to the loss function $\Omega(h) = \lambda \sum_i |h_i|$ at this time. This solves the problem of a large amount of acoustic emission signal data in fault diagnosis, resolves the issues of faint fault features in compressed data and extracts compressed domain features. Finally, the softmax classifier is used to classify and predict thrust ball bearings with different fault types. The basic flow of the method proposed in this paper is shown in Figure 3.

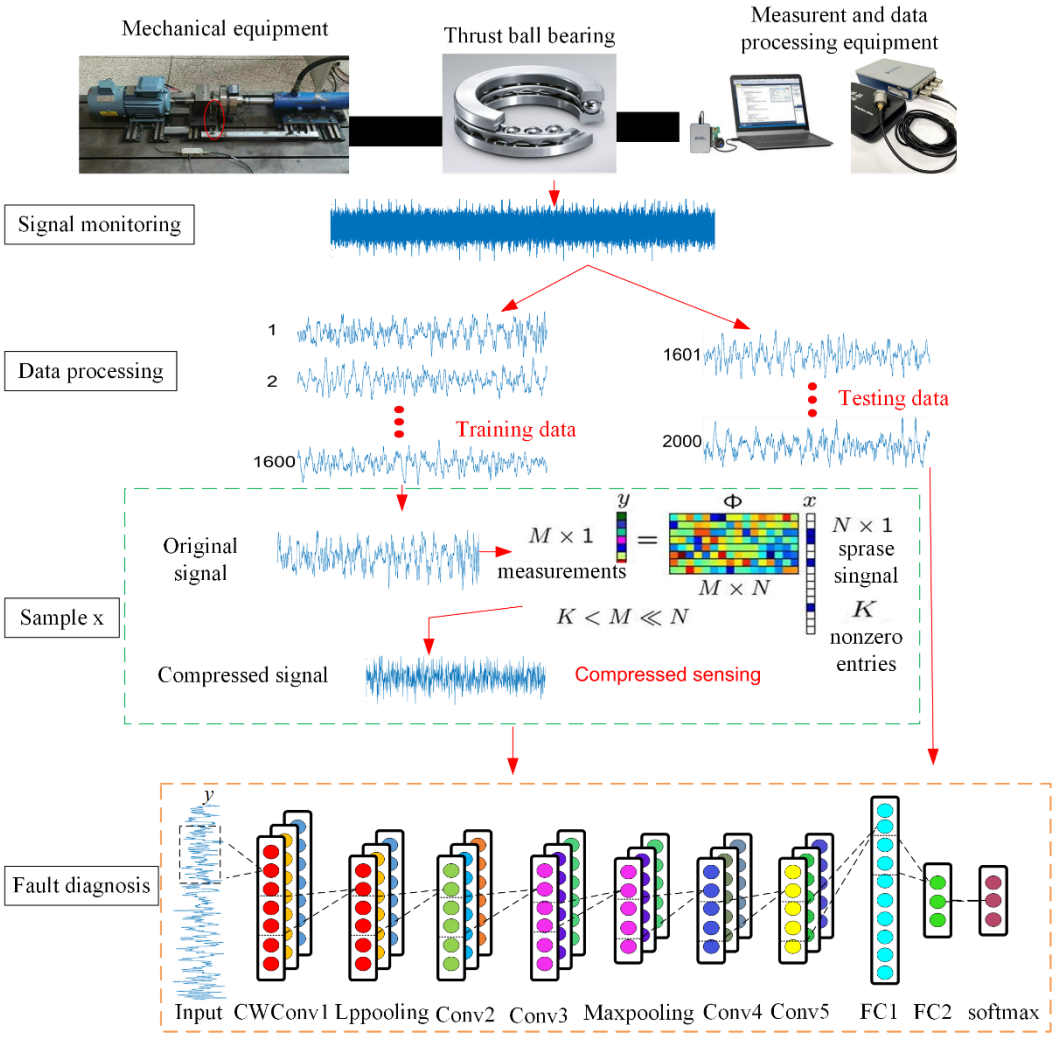


Figure 3. Wavelet sparse convolutional network flow chart.

4.1. Wavelet sparse convolution model structure

In this section, the structural parameters of the wavelet sparse convolutional network are set as shown in Table 1. The first convolutional layer is used as a continuous wavelet convolutional layer, and the remaining multiple convolutional layers are standard one-dimensional convolutional layers. The continuous wavelet convolutional layer can obtain the feature information of different frequency bands of the signal in the first layer, while the remaining convolutional layers and deeper network structures can fully exploit the signal's hidden fault features.

To suppress overfitting and improve the generalization ability of the model, the dropout function is added; to solve the problem of gradient disappearance or explosion, batch normalization (BN) is added after each convolutional layer. The training process stipulates that the initial learning rate is 0.001, the decay rate is 0.9, the batch learning size is 100, and all parameters are updated using backpropagation and Adam optimization algorithms.

When training the model in this paper, the cross-entropy loss function combined with the l_1 regular penalty term is used for measuring the difference distribution between the predicted value and the actual value. The loss function is defined as follows:

The last layer of the network uses the softmax function to output $\hat{y}$, where m represents the total number of samples, y is the fault class label corresponding to the sample, and h_i is the excitation value of the i-th neuron.

$$\begin{aligned} Loss &= \frac{1}{m}\sum_{x}(L(\hat{y},y)+\Omega(h)) \\ &= -\frac{1}{m}\sum_{x}([y\cdot\log\hat{y}+(1-y)\cdot\log(1-\hat{y})]+\lambda\sum_{i}|h_i|) \end{aligned} \tag{23}$$

Table 1. Network structure parameters.

No.	Layer	Kernel Size	Activation	Output Size
1	CWConv 1	1×81×54	ReLU	768×81
2	LPPooling	4×1		
3	Conv 2	81×81×55	ReLU	192×81
4	Maxpool 1	4×1		
5	Conv 3	81×81×55	ReLU	48×81
6	Maxpool 2	4×1		
7	Dropout			
8	Conv 4	81×64×55	ReLU	12×64
9	Maxpool 3	4×1		
10	Conv 5	64×64×55	ReLU	3×64
11	Dropout			
12	Flatten			
13	FC 1	192×64	ReLU	64×1
14	Dropout			
15	FC 2	64×3	Softmax	3×1

5. Experimental verification and analysis

To examine the effectiveness of the proposed method in this thesis, the thrust ball bearing is taken as the object of the simulated experiment, and three states of the bearing are simulated: namely, normal bearing, rolling element failure, and race failure. The thrust ball bearing is installed on a bearing failure simulation test bench, which is mainly composed of a driving device, gearing, an experimental device, a loading device, a guide rail device and an anti-overload preservation device. The acoustic emission sensor is used to collect the acoustic emission signal of the bearing, and the acoustic emission system consists of five parts: sensor, preamplifier, data acquisition card, host and display. The installation position of the test facilities and sensor is exhibited in Figure 4.

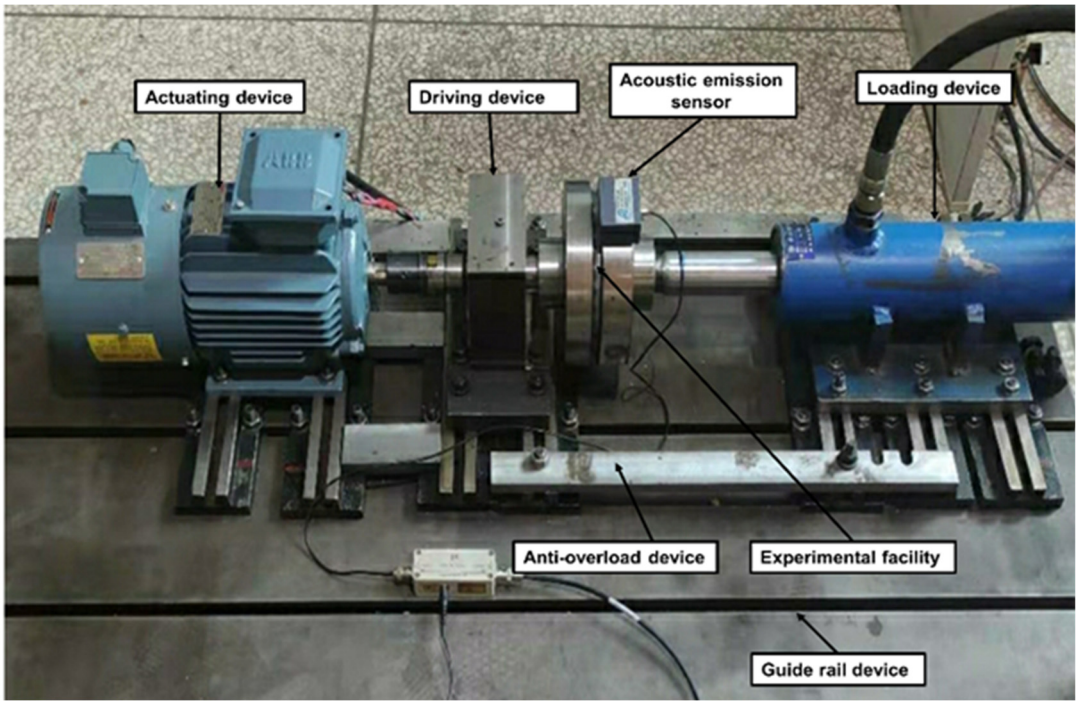


Figure 4. Bearing failure test bench.

When simulating the weak fault of the bearing, EDM (electric discharge machining) technology is used to process the single-point dimple at the center position of the bearing race and rolling element, as shown in Figure 5. Part of the fault location is enlarged, as exhibited in Figure 6. The spindle speed of the experimental equipment is 400 revolutions per minute, the sampling frequency is 1 MHz, and the number of data points is 20,000,000.

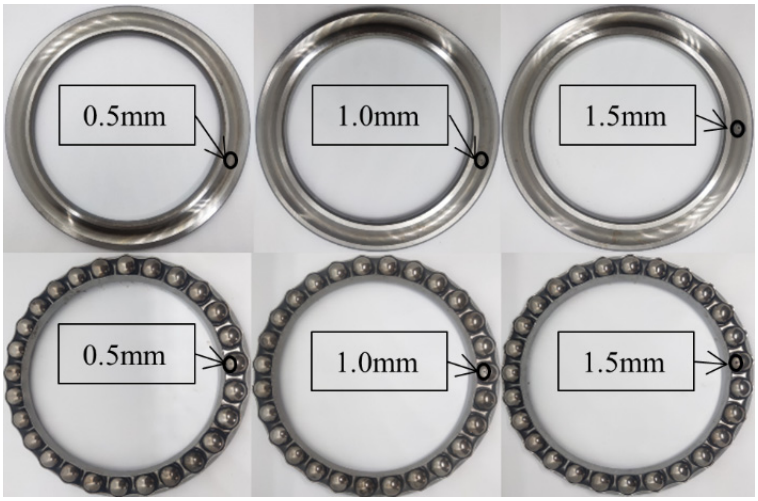


Figure 5. Schematic diagram of bearing component failure.

Figure 6. Partially enlarged view of bearing component failure.

5.1. Acoustic emission signal compression

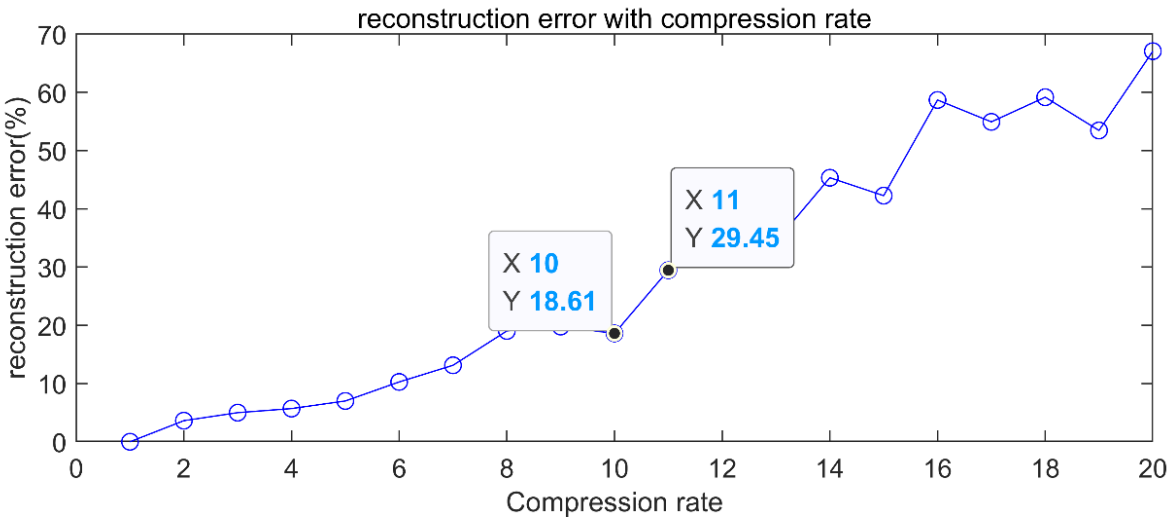


Figure 7. Reconstruction error curves under different compression ratios.

For compressed signals, the compression ratio is an indicator reflecting the degree of signal compression and an important parameter for data compression. The larger the compression ratio R (original signal data length / compressed data length), the more obvious the degree of data compression, the greater the data volume is reduced, and the greater the original signal features are lost in the compressed domain signal. Therefore, it is necessary to choose an appropriate compression ratio that can significantly reduce data while retaining key information and significantly reducing computational costs. In this paper, the discrete cosine transform (DCT) is used to sparsely represent the original acoustic emission signal, and the orthogonal matching pursuit algorithm (OMP) is used to reconstruct the signal. The influence of different compression ratios on the reconstruction error is studied. The reconstruction error is defined as $Error = \|x - \hat{x}\| / \|x\|$. The changing trend of the reconstruction error at the lower rate is used to select an appropriate compression rate. The compression ratio range is selected as

$R \in (1,20)$, and the reconstruction error curves under different compression ratios are obtained, as shown in Figure 7.

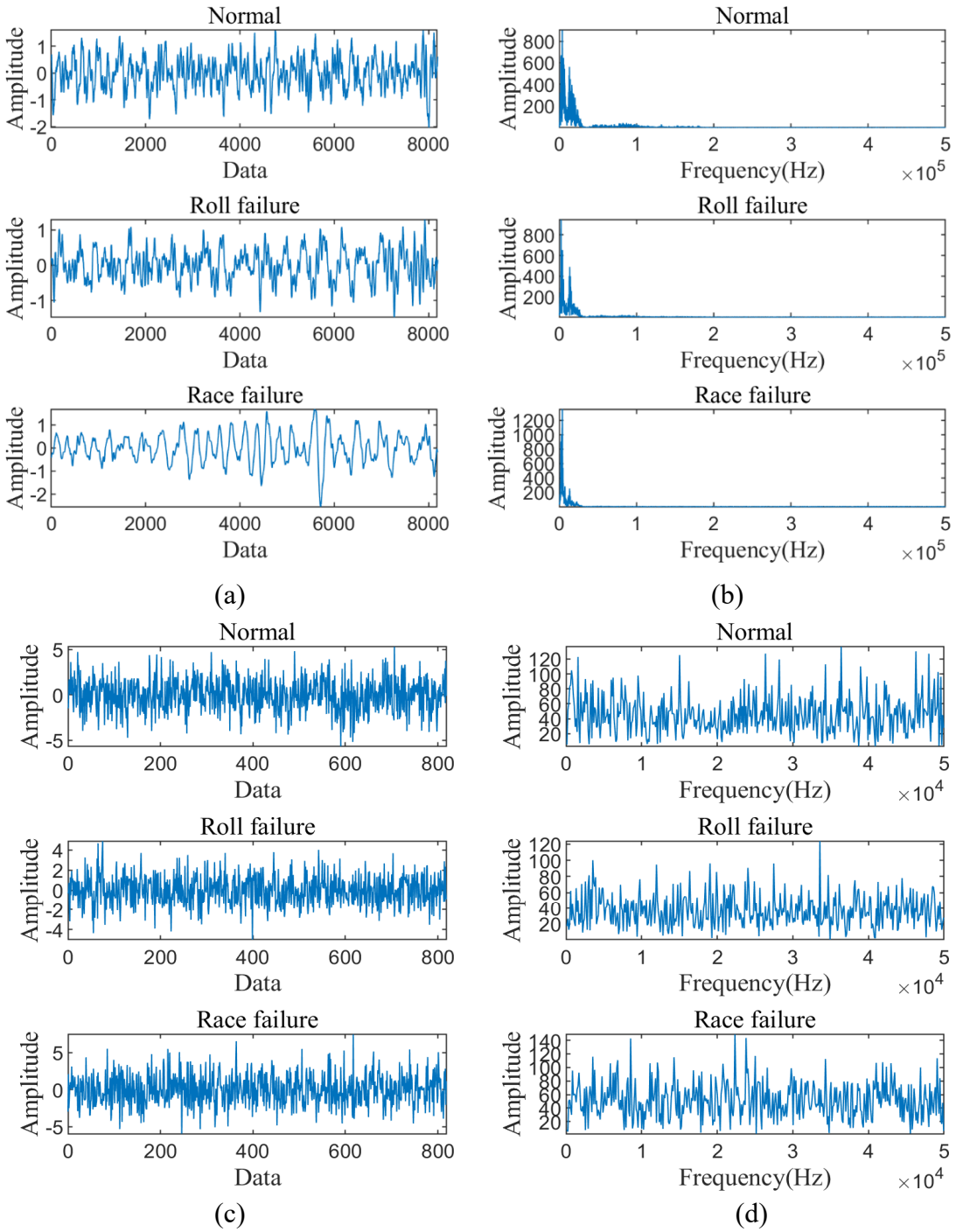


Figure 8. The time-domain waveform diagrams and spectrograms of the acoustic emission original signal and compressed signal in different states. (a) Time-domain waveform of the original acoustic emission signal; (b) spectrogram of the original acoustic emission signal; (c) time-domain waveform of compressed signal; (d) spectrogram of a compressed signal.

It is evident from the figure that the greater the compression ratio is, the greater the reconstruction error. According to the analysis of the reconstruction error curve shown in Figure 7, the relationship between the compression ratio and the reconstruction error is similar to a linear relationship. According to actual experience, a reconstruction error $E\text{rror} \le 25\%$ can meet the actual needs. As shown in Figure 7, when the compression rate $R \le 10$, the reconstruction error $Error \le 21.73\%$, which is acceptable in practical applications. Therefore, a compression ratio of 10 times is selected to compress the original signal. In this way, the compressed data can not only retain the important information in the original signal but also reconstruct the signal with low reconstruction error. It is also ready for the network to extract enough feature information in the compressed domain in the future.

In the process of generating compressed data, the property of approximately equidistant projection can ensure that the structure of the signal is approximately unchanged before and after compressed projection, while random projection, as a commonly used linear projection method, has been studied in many fields, especially in compressed sensing. In research, since the Gaussian random matrix is irrelevant to the vast majority of sparse matrices, the Gaussian random matrix is often used as the observation matrix for signal compression sampling in compressed sensing. Since the Gaussian random matrix satisfies the property conditions of the approximate equidistant projection, the Gaussian random matrix is selected. The random matrix is used as the projection matrix [30]. In this paper, the original signal is randomly dimensionally reduced by constructing a Gaussian random matrix to obtain compressed data. According to the error of the reconstructed error curve under different compression ratios in Figure 7, a compression ratio of 10 times is selected to compress the original signal. Figure 8 shows the time domain waveform and frequency spectrum of the acoustic emission signal of the thrust ball bearing before and after compression under three states: normal, rolling element fault and race fault. Figure 8 shows that the strength of the shock signal carried by the original signal is different. Comparing the compressed acoustic emission signal and the original acoustic emission signal shown in Figure 8, the time-domain waveforms and spectrograms in the three different states obviously show similar randomness characteristics. This is because some time-domain features of the signal are lost in the process of compression measurement, so the conventional acoustic emission signal analysis methods are unable to extract and accurately analyze the signal features.

The energy of the signal remains approximately unchanged in the random projection process, that is, the energy characteristic $\hat{E}$ of the compressed signal is approximately the same as the energy characteristic E of the original signal. Based on this property, the following conclusions are drawn. In the process of analyzing the acoustic emission signal of the thrust ball bearing, using the energy characteristic parameters of the compressed data for analysis and diagnosis, the results obtained are approximately consistent with the results obtained using the energy characteristic parameters of the original data.

Typically, ordinary convolutional networks perform a set of temporal convolutions between the input signal and some finite impulse response filters in the first layer. However, the shock components in the signal can not be effectively extracted from the compressed data through ordinary convolution operations, and the extracted features have no effective information. To extract effective energy features from compressed data, a deep model combining continuous wavelet convolution and energy features is proposed in the compressed domain, and a continuous wavelet convolutional layer is constructed to replace the first layer of ordinary convolutional networks. The wavelet convolution layer is convolved with the original signal to obtain signals in different frequency bands. Only the scale parameters and transformation parameters in the CWConv layer can be learned directly from the raw data, which is a very efficient way to obtain a filter bank suitable for the input data. The wavelet convolution structure is used to mine the frequency domain information of the signal, and after wavelet

convolution, energy pooling is used to extract energy features from multiband signals. The compressed data have sparse characteristics. Referring to the structural design of the sparse feature extraction in the sparse autoencoder, the l_1 regularization penalty is added to the design of the loss function in this method so that the network can learn the sparse features during the learning process.

To optimize the operation speed of the model and make the data samples contain more periodic signals, three types of data are used: normal, roller fault and race fault. Each type of raw data selects 16,384,000 data points for compression, each type of data point is divided into 2000 samples, and the size of the dataset becomes 2000 × 8192. The compressed data length becomes 1,638,000 data points, which are also divided into 2000 samples, and the data size becomes 2000 × 819. The size of the compressed data becomes $1/10$ that of the original data. According to the ratio settings of 80 and 20%, the dataset is divided into a training dataset and a test dataset. The training dataset in the original data contains 1600 × 3 samples, and the test dataset contains 400 × 3 samples. The sample segmentation information is shown in Tables 2 and 3.

Table 2. Original dataset description.

Sample type	Number of sample points	The number of training set samples	The number of samples in the test set	Category tag
Normal	2000 × 8192	1600	400	0
0.5 Roller	2000 × 8192	1600	400	1
0.5 Seat ring	2000 × 8192	1600	400	2

Table 3. Compressed dataset description.

Sample type	Number of sample points	The number of training set samples	The number of samples in the test set	Category tag
Normal	2000 × 819	1600	400	0
0.5 Roller	2000 × 819	1600	400	1
0.5 Seat ring	2000 × 819	1600	400	2

5.2. Experimental results and discussion

5.2.1. Different types of fault diagnosis

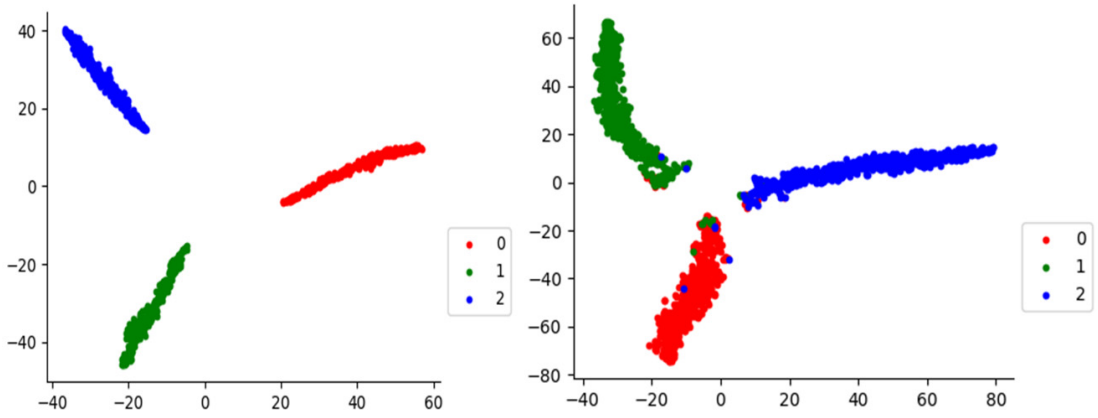
Since the compressed data show randomness and sparsity of the amplitude, to obtain more feature information from the signal, this paper uses wavelet convolution and energy pooling combined with $l1$ regularization to improve the model structure and then uses the compressed data as the model input. To prove the effectiveness of this method in this thesis, different methods are used to diagnose the faults of the thrust ball bearings. In method 1, the original signal is directly input into the convolutional neural network for fault diagnosis. The compressed signal is directly fed into the model proposed in this thesis for learning. Under the same conditions of each parameter, the diagnosis results are shown in Table 4. It can be seen from the table that the method proposed in this thesis can exhibit 98.50% accuracy in diagnosis. The effect of the proposed method is almost the same as that using the original

signal for direct fault diagnosis, but it only takes 1/13 of the time of method 1, which obviously improves the efficiency, proving that the method proposed in this paper can effectively extract compressed signal features and perform fault diagnosis.

Table 4. Diagnostic results of different methods.

Method	Accuracy (%)	Training time (s)
Method 1	100.00	2752.65
The method of this paper	98.75	194.71

The t-SNE method proposed by Laurens [28] is used to visualize the features learned by two different methods. Figure 9(a) is the mapping result output by method 1 in the fully connected layer, showing obvious classification results; Figure 9(b) is the mapping result output by the method in this paper in the fully connected layer. It can also be seen that the basic classification has been completed, and the classification effect has been achieved.



(a)Method1 (b)The method of paper

Figure 9. t-SNE dimensionality reduction feature visualization.

5.2.2. Fault diagnosis of different sizes

Table 5. Parameters of each component.

Fault state	Rotating speed (rpm)	Fault size (mm)
Roller	400	0 mm
	400	0.5 mm
	400	1 mm
Seat ring	400	1.5 mm
	400	0 mm
	400	0.5 mm
	400	1 mm
	400	1.5 mm

To further illustrate the advantages of the proposed method in extracting compressed data features and saving time, the diagnostic effect is verified on roller faults and race faults of different sizes. The parameters of each component are shown in Table 5. Method 1 is used to directly input the original signal into the convolutional neural network for fault diagnosis, and the method in this paper directly sends the compressed signal to the rewritten neural network model structure for fault diagnosis. Under the same conditions of each parameter, the diagnosis results are shown in Table 5. It can be seen from the table that the diagnostic accuracy of the method proposed in this paper is almost the same as that of method 1 using the original signal for direct fault diagnosis, the time consumption is significantly reduced, and the efficiency is significantly improved.

Table 6. Diagnosis effects for different sizes.

Fault state	Method	Accuracy (%)	Training time (s)
Roller (0, 0.5, 1, 1.5 mm)	Method 1	97.38%	3327.95
	The method of this paper	97.88%	255.17
Seat ring (0, 0.5, 1, 1.5 mm)	Method 1	99.94%	3093.05
	The method of this paper	98.81%	261.11

The t-SNE method is also used for visualizing the features learned by the two different methods, as shown in Figure 10. It can be seen from the figure that both methods show obvious classification effects.

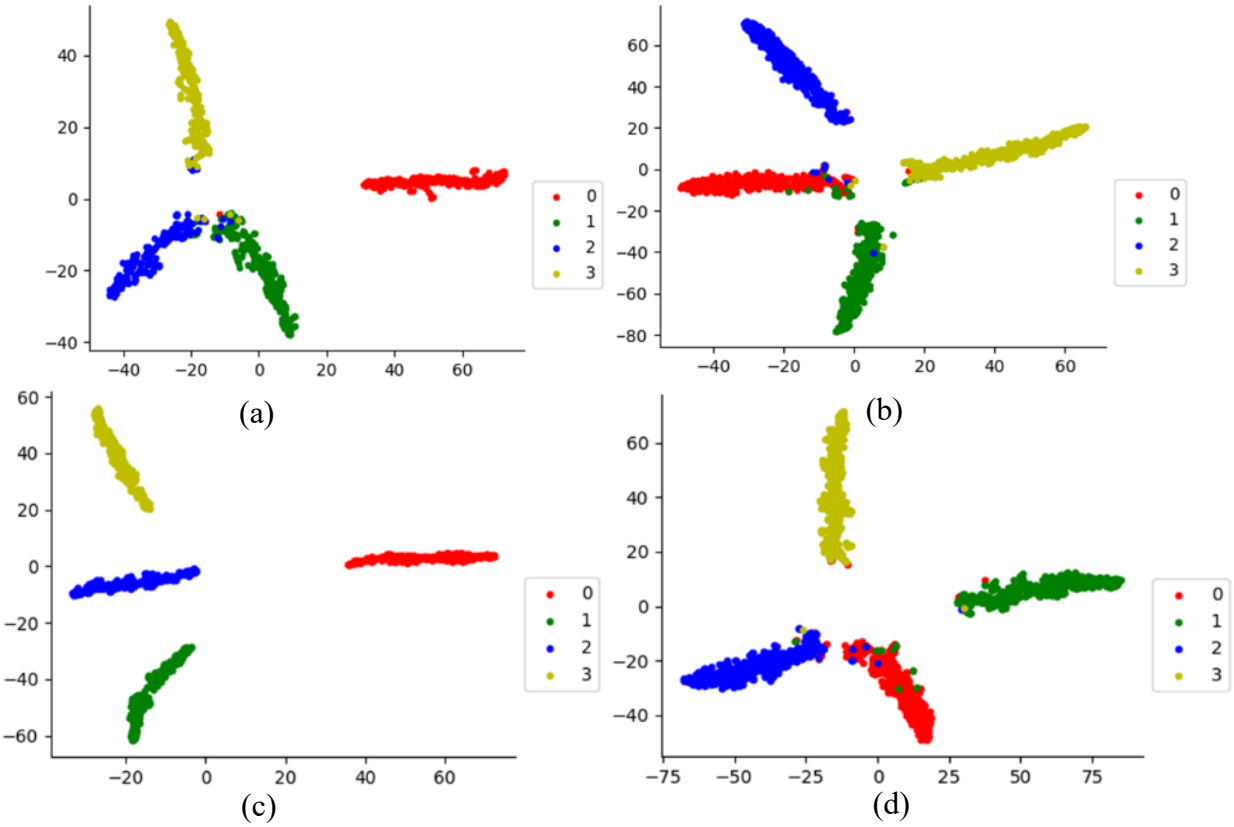


Figure 10. (a) Roller failure–Method 1; (b) Roller failure–Method proposed here; (c) Race failure–Method 1; (d) Race failure–Method proposed here.

To more clearly show the recognition results of the method in this paper on various types of signals in the test set, a confusion matrix is introduced for three experiments (three different fault types,

four different sizes of roller faults and four different sizes of race faults). It can be seen from the confusion matrix that the model can achieve a recognition accuracy of more than 95% for each state, showing excellent performance, as shown in Figure 11.

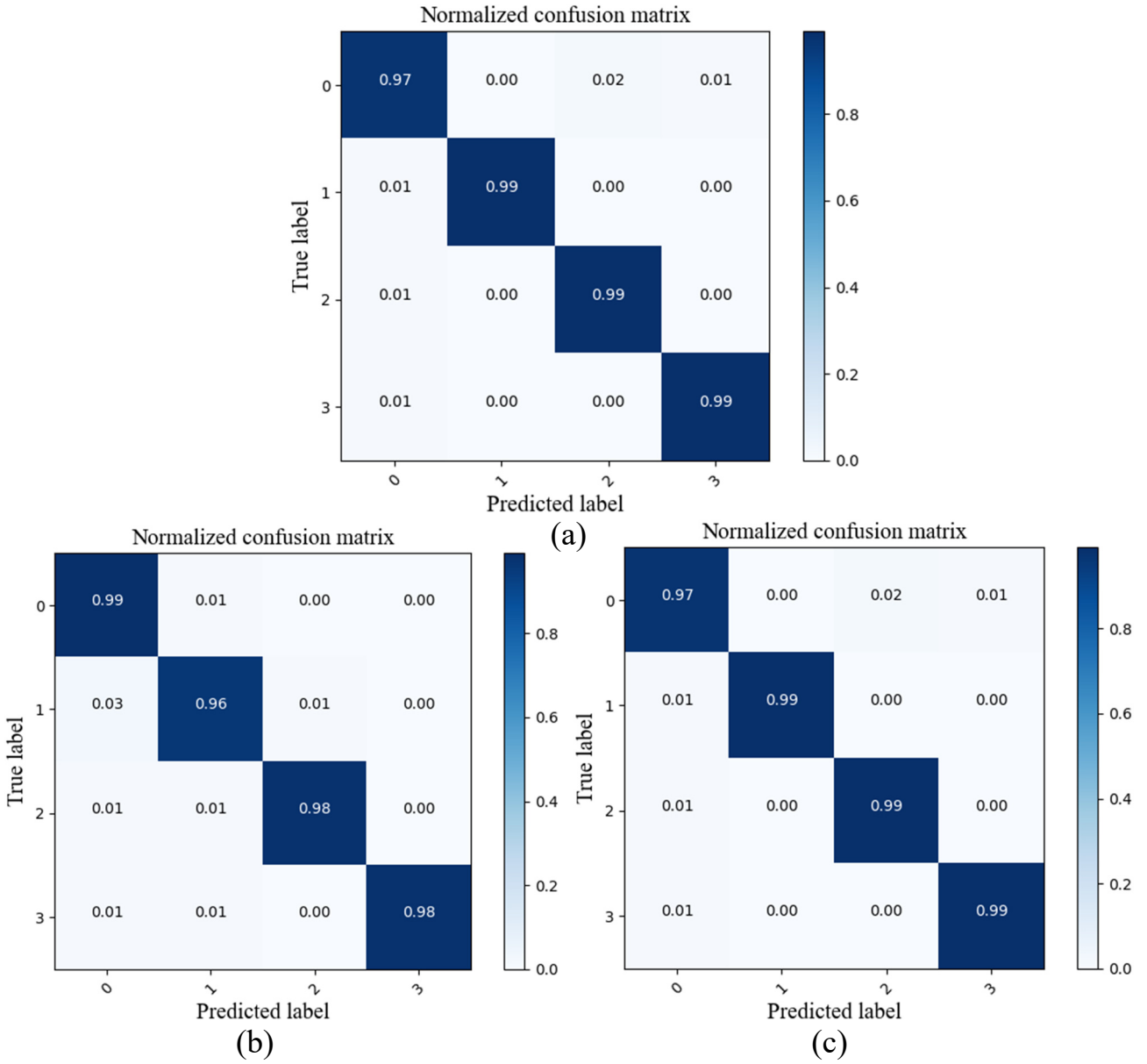


Figure 11. (a) Three types of faults; (b) four roller faults; (c) four race faults.

To further illustrate the advantages of the proposed method in extracting features from compressed data, three fault types are utilized, as shown in Table 7. The method in this paper is compared with the sparse autoencoder (SAE) and support vector machine (SVM) fault diagnosis methods. The radial basis function is chosen as the kernel function of the SVM, and the classification features used by the SVM algorithm are 9 eigenvalues of the mean square value, effective value, skewness, kurtosis, peak value, peak-to-peak value, form factor, impulse factor and crest factor. The diagnosis results are shown in Figure 12. As seen from the figure, the diagnostic results of our method are higher than those of SAE and SVM. This paper combines compressed sensing theory with wavelet convolution, energy

pooling and sparsity to generate new network structures. While retaining most of the feature information of the signal, the method in this paper adaptively selects the frequency band information to enhance the faint fault features of the compressed signal and can use fewer data to achieve higher classification accuracy. Meanwhile, SAE mainly extracts features from the input data, which are passed to the decoder to reconstruct the original data. For the compressed data of the processing object in this paper, it is impossible to obtain all the information of the signal accurately, resulting in a low fault diagnosis and recognition rate of the network. In addition, the performance of the traditional shallow model of SVM relies heavily on subjective artificial feature extraction, so the fault diagnosis accuracy and generalization ability when combined with the SVM classifier are low. The diagnosis results of the three methods are shown in Figure 12.

Table 7. Three fault types.

Type-1	Type-2	Type-3
Three different types of failures	Four different sizes of roller failure	Four different sizes of seat ring failure

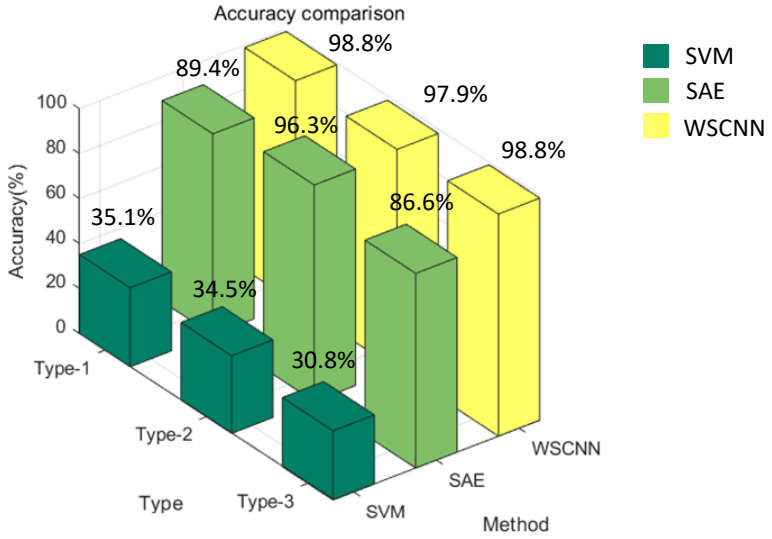


Figure 12. Accuracy comparison.

6. Conclusions

In this paper, a method for diagnosing acoustic emission signals of thrust ball bearings based on compressed sensing combined with deep learning is proposed. This method projects the acoustic emission time domain signal into the compressed domain to obtain the compressed signal, which greatly reduces the data redundancy and improves the analysis efficiency while retaining most of the information. The compressed signal is taken as the network input, and the wavelet sparse convolution network is used to carry out wavelet changes on the signals in the compressed domain to obtain the information of each frequency band. Then, the energy pooling layer is used for mining the energy features of each frequency band, and the l_1 regularization penalty term is added in the design of the loss function to make the network learn the sparse features in the learning process. This solves the problem of the large amount of acoustic emission signal data in fault diagnosis and solves the problem of difficulty

in extracting fault features in the compressed domain due to the randomness of the signal due to compression. Finally, the experimental results show that the proposed method gives better diagnostic results than other traditional methods. In the meantime, it offers a new diagnosis means for equipment fault diagnosis under modern big data and can carry out fault diagnosis more efficiently under the condition of ensuring high diagnosis accuracy.

Acknowledgments

This work was supported in part by the Key Scientific Research Projects of Yunnan Province under Grant 202102AC080002.

Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this paper.

References

1. Y. Lv, J. Luo, C. Yi, Enhanced orthogonal matching pursuit algorithm and its application in mechanical equipment fault diagnosis, *Shock Vib.*, **2017** (2017), 1–13. https://doi.org/10.1155/2017/4896056
2. T. R. Kurfess, S. Billington, S. Y. Liang, Advanced diagnostic and prognostic techniques for rolling element bearings, *Springer London*, (2006), 137–165. https://doi.org/10.1007/ 1-84628-269-1_6
3. C. Liu, K. Gryllias, A Semi-supervised support vector data description-based fault detection method for rolling element bearings based on cyclic spectral analysis, *Mech. Syst. Sig. Process.*, **140** (2020), 106682. https://doi.org/10.1016/j.ymssp.2020.106682
4. Q. Ni, J. C. Ji, K. Feng, B. Halkon, A fault information-guided variational mode decomposition (FIVMD) method for rolling element bearings diagnosis, *Mech. Syst. Sig. Process.*, **164** (2022), 108216. https://doi.org/10.1016/j.ymssp.2021.108216
5. Q. Ni, J. C. Ji, K. Feng, B. J. Halkon, A novel correntropy-based band selection method for the fault diagnosis of bearings under fault-irrelevant impulsive and cyclostationary interferences, *Mech. Syst. Sig. Process.*, **153** (2020), 107498. https://doi.org/10.1016/ j.ymssp.2020. 107498
6. H. An, W. Liang, Y. Zhang, Y. Li, Y. Liang, J. Tan, Rotate vector reducer crankshaft fault diagnosis using acoustic emission techniques, in *2017 5th International Conference on Enterprise Systems (ES)*, (2017), 294–298. https://doi.org/10.1109/ES.2017.55
7. Y. Zhang, W. Lu, F. Chu, Planet gear fault localization for wind turbine gearbox using acoustic emission signals, *Renewable Energy*, **109** (2017), 449–460. https://doi.org/10.1016/ j.renene.2017.03.035
8. D. L. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory*, **52** (2006), 1289–1306. https://doi.org/10.1109/TIT.2006.871582
9. X. Zhang, N. Hu, C. Zhe, A bearing fault detection method base on compressed sensing, engineering asset management-systems, *Eng. Asset Manage.-Syst.*, (2015), 789–798. https://doi.org/10.1007/978-3-319-09507-3_69

10. G. Tang, W. Hou, H. Wang, G. Luo, J. Ma, Compressive sensing of roller bearing faults via harmonic detection from under-sampled vibration signals, *Sensors (Basel, Switzerland)*, **15** (2015), 25648–25662. https://doi.org/10.3390/s151025648
11. Y. Wang, J. Xiang, Q. Mo, S. He, Compressed sparse time–frequency feature representation via compressive sensing and its applications in fault diagnosis, *Measurement*, **68** (2015), 70–81. https://doi.org/10.1016/j.measurement.2015.02.046
12. Q. Y. G. Tang, H. Q. Wang, G. Luo, J. Ma, Sparse classification of rotating machinery faults based on compressive sensing strategy, *Mechatronics*, **31** (2015), 60–67. https://doi.org/10.1016/j.mechatronics.2015.04.006
13. L. Chang, W. Xing, J. Mao, X. Liu, Acoustic emission signal processing for rolling bearing running state assessment using compressive sensing, *Mech. Syst. Sig. Process.*, **91** (2017), 395–406. https://doi.org/10.1016/j.ymssp.2016.12.010
14. H. Ahmed, M. Wong, A. K. Nandi, Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete features, *Mech. Syst. Sig. Process.*, **99** (2018), 459–477. https://doi.org/10.1016/j.ymssp.2017.06.027
15. H. Shao, H. Jiang, H. Zhang, W. Duan, T. Liang, S. Wu, Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing, *Mech. Syst. Sig. Process.*, **100** (2018), 743–765. https://doi.org/10.1016/j.ymssp. 2017.08.002
16. H. Yuan, X. Wang, X. Sun, Z. Ju, Compressive sensing-based feature extraction for bearing fault diagnosis using a heuristic neural network, *Meas. Sci. Technol.*, **28** (2017), 065018. https://doi.org/10.1088/1361-6501/aa6a07
17. R. Chen, X. Huang, L. Yang, X. Xu, X. Zhang, Y. Zhang, Intelligent fault diagnosis method of planetary gearboxes based on convolution neural network and discrete wavelet transform, *Comput. Ind.*, **106** (2019), 48–59. https://doi.org/10.1016/j.compind.2018.11.003
18. E. Candes, T. Tao, Near optimal signal recovery from random projections: Universal encoding strategies?, *IEEE Trans. Inf. Theory*, **52** (2004), 5406–5425. https://doi.org/10.1109/TIT.2006.885507
19. J. I. Xiu-Xia, X. X. Bian, Study on performance of greedy algorithms for signal reconstruction, *Comput. Simul.*, 2013.
20. M. A. Davenport, M. F. Duarte, Y. C. Eldar, G. Kutyniok, Introduction to compressed sensing, *Citeseer*, (2012), 1–64. https://doi.org/10.1017/CBO9780511794308.002
21. J. Schmidhuber, Deep learning in neural networks: An overview, *Neural Networks*, **61** (2015), 85–117. https://doi.org/10.1016/j.neunet.2014.09.003
22. R. Socher, Y. Bengio, C. D. Manning, Deep learning for NLP (without magic), *Acl. Tuto.*, **2012** (2013), 5–5. https://dl.acm.org/doi/pdf/10.5555/2390500.2390505
23. G. Hinton, L. Deng, D. Yu, G. E. Dahl, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Sig. Process.Mag.*, **29** (2012), 82–97. https://doi.org/10.1109/MSP.2012.2205597
24. T. Li, Z. Zhao, C. Sun, L. Cheng, R. X. Gao, WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis, *IEEE Trans. Syst.*, **52** (2019), 2302–2312. https://doi.org/10.1109/TSMC.2020.3048950
25. F. Y. Guo, Y. C. Zhang, Y. Wang, P. Wang, P. J. Ren, R. Guo, X. Y. Wang, Fault detection of reciprocating compressor valve based on one-dimensional convolutional neural network, *Math. Probl. Eng.*, **2020** (2020). https://doi.org/10.1155/2020/8058723

26. D. T. Hoang, T. T. Xuan, M. Van, H. J. Kang, A deep neural network-based feature fusion for bearing fault diagnosis, *Sensors*, **21** (2021), 244. https://doi.org/10.3390/s21010244
27. D. E. Rumelhart, Learning internal representations by error propagation, In D. E. Rumelhart, J. L. McCelland and PDP Research Group, *Parallel Distrib. Process.*, **1** (1986). https://dl.acm.org/doi/10.5555/104279.104293
28. K. F. Al-Raheem, A. Roy, K. P. Ramachandran, D. K. Harrison, S. Grainger, Application of the laplace-wavelet combined with ANN for rolling bearing fault diagnosis, *J. Vib. Acoust.*, **130** (2008), 3077–3100. https://doi.org/10.1115/1.2948399
29. M. Liao, C. Liu, C. Wang, J. Yang, Research on a rolling bearing fault detection method with wavelet convolution deep transfer learning, *IEEE Access*, **9** (2021), 45175–45188. https://doi.org/10.1109/ACCESS.2021.3067152
30. C. Wang, C. Liu, M. Liao, Q. Yang, An enhanced diagnosis method for weak fault features of bearing acoustic emission signal based on compressed sensing, *Math. Biosci. Eng.*, **18** (2021), 1670–1688. https://doi.org/10.3934/mbe.2021086

AIMS Press