IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 1, 2019, accepted July 23, 2019, date of publication July 29, 2019, date of current version August 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931616

Rolling Bearing Fault Diagnosis Algorithm Based on FMCNN-Sparse Representation

FENG-PING AN
School of Physics and Electronic Electrical Engineering, Huaiyin Normal University, Huai'an 223300, China
School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

e-mail: anfengping@163.com

This work was supported in part by the National Natural Science Foundation of China under Grant 61701188, in part by the China Postdoctoral Science Foundation under Grant 2019M650512, and in part by the Natural Science Foundation of Shanxi under Grant 201801D221171.

ABSTRACT The time-frequency analysis of vibration signals is an effective means to analyze the fault characteristics of rolling bearings. The traditional pattern recognition method is difficult to adapt to the complex mapping relationship between the high-dimensional feature space and the state space. The deep learning method has high-dimensional feature adaptive analysis ability, which is suitable for the intelligent analysis of the high-dimensional feature space in fault states. The feedforward deep convolutional neural network (CNN) has achieved some success in mechanical fault diagnosis. However, the rolling bearing fault signal is complex, and there are many interference factors. The CNN relying on the simple feedforward method cannot effectively meet the actual needs in the field of fault diagnosis. Although there are some CNNs with feedback methods, the CNNs of these feedback methods cannot systematically obtain the characteristic information of rolling bearing faults. Therefore, they do not solve the feature extraction problem of rolling bearing faults well. In view of this, this paper provides a specific mathematical definition of the feedback mechanism for constructing the feedback mechanism in the deep CNN, models the feedback mechanism into an optimization problem, determines the basic framework of the feedback mechanism, and an effective feedback mechanism calculation model is proposed. Based on this, a solution algorithm based on the gradient descent method is proposed. Then, an effective supervised feature extraction method based on sparse expression is proposed. It maps the sample features to the feature domain through the effective transform method. In the process, the wavelet packet transform (WPT) transform is used as the basis function to construct a dictionary with structural effects, and mixed penalty terms are introduced to further optimize the performance of structural sparse expression. Finally, the sparse expression is combined with the feedback mechanism CNN (FCNN) to establish a sub-module fault diagnosis network so that a diagnosis can determine the fault severity while assessing the bearing fault location. The example shows that the method proposed in this paper has high accuracy in determining the state of rolling bearings and has great application potential in engineering.

INDEX TERMS Fault diagnosis, rolling bearing, sparse representation, convolutional neural networks, feedback mechanism, gradient descent.

I. INTRODUCTION

Once the mechanical equipment or key components therein fail, the operation becomes abnormal, which leads to the collapse of the entire mechanical system, which in turn leads to serious economic losses or major disasters. Relevant information indicates that the application of fault diagnosis technology to the diagnosis of mechanical equipment can effectively reduce the cost of the enterprise [1]–[5]. In rotating machinery, rolling bearings are one of the key components. According to statistics, in common rotating machinery failures, rolling bearing failures account for 30%-40% [6], [7]. Therefore, it is of great theoretical significance and engineering value to study the intelligent fault diagnosis technology and rolling bearings method to ensure the safe and efficient operation of equipment [7], [8].

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

Rolling bearing fault diagnosis methods are generally divided into fault diagnosis methods based on time-frequency analysis, fault diagnosis methods based on feature dimension reduction, and fault diagnosis methods based on artificial intelligence. For the fault diagnosis method based on time-frequency analysis, for example, Gao *et al.* [9] proposed fault feature extraction and identification of rolling bearings based on short-time Fourier transform and nonnegative matrix factorization. Unlike the methods mentioned above, empirical mode decomposition (EMD) can automatically divide no stationary, nonlinear signals into several different intrinsic mode functions (IMFs). The EMD method has received extensive attention and application in the field of fault diagnosis [10], [11]. Although EMD is widely used in the analysis of no-stationary signals, it has the problems of extreme point extraction, envelope extraction, interpolation technique, endpoint effect and modal aliasing [12], [13]. Aiming at this problem, choosing fault-sensitive statistical features as the basis of subsequent fault analysis has been the focus of in-depth research by many researchers. The second type of fault diagnosis method based on feature dimension reduction has been formed. Among them, principal component analysis and linear discriminant are two classic linear feature dimension reduction methods. Harmouche *et al.* [14] proposed principal component analysis (PCA) for early fault diagnosis and used information gain to measure the dissimilarity between different modal principal components. It has the following problems [15], [16]: over-fitting problems; high computational complexity.

To solve the above problems, relevant scholars introduced this kind of method into fault diagnosis and then formed a fault diagnosis method based on artificial intelligence for rolling bearings. One is a fault diagnosis method based on traditional artificial neural networks, wavelets and other artificial intelligence algorithms. Rai and Upadhyay [17] combined a nonlinear autoregressive neural network (NARX-NN) with wavelet filtering technology for state estimation of rolling bearings. Oliveira *et al.* [18] proposed a fault diagnosis model based on weightless neural networks (WNN). However, the shallow machine learning method does not extract all the fault information of the rolling bearing. It also does not allow for adaptive features based on the fault characteristics of different rolling bearings. In view of the high learning ability of deep learning, another type of fault diagnosis method based on deep learning was introduced. Shao *et al.* [19] proposed a deep belief network (DBN) for intelligent state monitoring of induction motors. The DBN is used to automatically extract the relevant features of the vibration signal for state recognition. AlThobiani and Ball [20] proposed the combination of energy operator demodulation feature extraction and DBN to realize fault diagnosis of reciprocating compressor valves. Janssens *et al.* [21] proposed a fault diagnosis method based on convolutional neural network (CNN), which extracts local feature information directly from the original vibration signal through a multi-layer convolution-pooling structure. It enables fault diagnosis of rotating machinery. Zhang *et al.* [22] proposed a deep CNN model for the diagnosis of rolling bearings, which uses wide kernel extraction features and suppresses high-frequency noise interference in the first layer of the model network. Other network layers use small convolution kernels to build nonlinear mappings. It can improve the bearing fault diagnosis rate as a whole. Jiang *et al.* [23] proposed a new multilayer deep learning CNN for fault diagnosis of rolling bearings. The normalized preprocessing method was used to preprocess the vibration signals. The processed signal is sent to the multilayer CNN for pattern recognition. He *et al.* [24] proposed the envelope spectrum information of bearing vibration signal as the feature vector, and constructed a Gaussian-limited Boltzmann machine model to classify high-dimensional feature vectors. It improves the recognition accuracy of faults. Although deep learning has achieved satisfactory results in the diagnosis of rolling bearing faults, it cannot be further applied and popularized due to the problems inherent in deep learning. The main reasons are as follows: First, the learning of deep neural networks belongs to supervised learning. To accomplish a certain task, it is necessary to rely on a large number of manually labeled data. Second, the interpretability is poor. There is no corresponding theory to support network structure initialization and parameter initialization. Third, the task is singular, and different tasks are usually completed by different network frameworks and learning methods. These defects limit the deep application and promotion of deep learning in the diagnosis of rolling bearing faults.

However, almost all deep CNN models are feedforward neural networks; the information transfer between neurons is unidirectional. In the human visual neural network, in addition to the feedforward connection, there is a large number of feedback connections and lateral connections [25]–[27]. Studies have shown that the number of feedback connections is several times the number of feedforward connections [28]–[30]. The feedforward neural network can perform certain sensing tasks, but there are still the problems of sensitivity to noise, relying on a large number of samples, poor interpretability, lack of adaptability and robustness, etc. Some recent efforts have also made some attempts to introduce feedback mechanisms in neural networks [31]–[33]. Both the depth Boltzmann machine (DBM) [31], [32] and the deconvolution network [32]–[34] attempt to define the feedback mechanism as a reconstruction process during the training phase. At the same time, there are some work that uses the recurrent neural network and Long Short-Term Memory (LSTM) to capture attention-shifting signals in a dynamic environment and implement feedback mechanisms through reinforcement learning [35]–[37]. Some scholars have proposed a robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning. However, the actual application effect is not satisfactory. It can be seen from the above literature that a convolutional neural network with a feedback mechanism can better identify and process fault signals. This is because the feedback mechanism can better optimize and adjust the convolutional neural network itself so that it can be more adaptive to the subsequent analysis and processing of the fault signal. Therefore, the feedback mechanism convolutional neural network is better than traditional convolutional neural networks in the diagnosis of rolling bearing faults. However, these architectures do not effectively capture the high-level semantic concepts of fault diagnostic signals and obtain all the important feature information associated with them. So, in-depth study of the mechanism of the feedback connection will give the deep CNN more potential and flexibility and identify a solution to overcome the many defects encountered by the current deep neural network, which can greatly accelerate the development of deep learning theory in the field of fault diagnosis. Because the activation function of the convolutional neural network has an important influence on the actual effect of the model, the convolutional neural network constructing the feedback mechanism will weaken the problem to a certain extent. If the convolutional neural network with good feedback mechanism is constructed, the feature information will be more accurate and rich. At the same time, the rolling bearing failure mode introduces various components, such as transient periodic shock, nonstationary features and noise in the vibration signal. Traditional one-dimensional time-domain or frequency-domain analysis cannot capture the intrinsic structure in the bearing signal, which bases the signal analysis for rolling bearings on two-dimensional time-frequency analysis technology. The advantage of data representation of sparse expression is due to the structural dictionary used. The clearer the basic elements of the dictionary grasp the structure of the signal, the more the expression results obtained will highlight the characteristic patterns in the signal [27], [29], [30].

Therefore, this paper proposes a fault diagnosis algorithm for rolling bearings based on feedback mechanism convolutional neural network-sparse representation. The basic idea is as follows: first, construct a feedback connection around the deep convolutional neural network feedback target in the convolutional neural network, give the mathematical definition of the feedback adjustment mechanism problem based on the feedback connection and the feedback target, and finally, abstract the feedback optimization problem. And then, a new framework of a feedback-convolution neural network based on gradient descent is proposed. The sparse coding technique is used to sparsely express the deep learning model, thus effectively removing redundant information. It seeks a simple representation of the essence of the data, reduces the corresponding calculation process, and reduces the difficulty of data analysis. Finally, the algorithm is used to analyze and summarize the rolling bearing fault signal.

Section II of this paper describes the convolutional neural network with the feedback mechanism proposed in this paper. Section III illustrates the sparse expression technique proposed in this paper to sparsely express the deep learning model established in Section II. Section IV introduces a fault state recognition framework and method based on feedback mechanism convolutional neural network-sparse representation. Section V analyzes the fault diagnosis algorithm proposed in this paper and compares it with the mainstream fault diagnosis algorithm. Finally, the full text is summarized and discussed.

II. FEEDBACK CONVOLUTION NEURAL NETWORK BASED ON GRADIENT DESCENT

A. MATHEMATICAL MODELING OF THE FEEDBACK ADJUSTMENT MECHANISM

In this section, we will implement a decision rule R by constructing feedback (the decision rule: the decision rule can be used to eliminate all connection paths that are not associated with the target signal, and it can locate or even segment the target signal of interest in a top-down manner). In addition, the problem of cutting the connection path is transformed into a neuron screening problem. This section first describes the convolutional neural network from the perspective of information selection.

1) NEW EXPLANATION OF DEEP NEURAL NETWORKS

Deep convolutional network models are constructed by stacking simple operational layers, including the convolutional layer, the ReLU layer, and the max layer. For each layer, assuming the input is X, it is known that it is not the signal itself or the output of the previous layer. It is assumed that x is composed of C channels, the length and width are represented by S and T, that is $x \in R^{S\times T\times C}$, the output thereof is assumed to be y and is composed of C' channels, and the length and width are S' and T', respectively, that is $y \in R^{S\times T\times C}$. On this basis, this paper can formulate the convolutional layer, ReLU layer and max layer separately.

Convolutional layers are used to extract different features of the input. The convolutional layer consists of C' convolution kernels, each convolution kernel $k \in R^{K\times K\times C}$, and then the operation of the convolutional layer is described by the following formula:

$$y_{C'} = \sum_{c=1}^{C} k_{c'c} * x_c, \quad \forall c' \tag{1}$$

The ReLU layer is mainly used to increase the nonlinearity of the network without affecting the receptive field of the convolutional neurons. Its corresponding input and output function relationship is as follows:

$$y = \max(0, x) \tag{2}$$

The max layer is mainly used to reduce the dimensions of the output vector and to obtain a degree of invariance to ensure that similar structures can achieve the same output. Max acts on the neighborhood N of each signal (i, j), specifically:

$$y_{i,j,c} = \max_{u,v\in N} x_i + u, j + v, c, \quad \forall i, j, c \tag{3}$$

Selectivity in the feedforward process. To better understand how selectivity works in neural networks and the model feedback mechanisms, we need to reinterpret the role of the ReLU and max layers. In this paper, the max( ) operation in

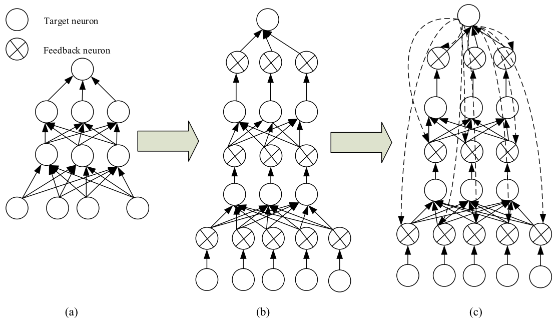


FIGURE 1. Schematic diagram of the CNN feedback connection constructed in this paper.

formulas (2) and (3) can be replaced by a series of binary switches $z \in \{0, 1\}$. Therefore, the ReLU and max layers can be represented in the form of $y = zox$. More specifically, the ReLU layer is represented as $y = zox$, O represents the multiplication of the signal class level, the max layer is represented as $y = z * x$, $*$ represents the convolution operation, and z represents the convolution kernel with a value of 0, 1.

By reinterpreting the ReLU and max layers as gate operations controlled by input x, convolutional neural networks can be understood as a bottom-up approach for selecting the useful information for decision making in the feedforward process by these gate operations and discarding those that contribute little to the decision. Then it makes the final decision. To ensure versatility and generalization, a large amount of information can be filtered through the ReLU and max layers, and thus, a large number of neurons are activated. However, activated neurons may be useful or detrimental to the final decision and often introduce a large amount of noise, such as input signals collected from complex scenes, and then the signal background will be very complex. This background information also reaches the classification decision layer through the network, which can cause serious interference to the characteristic information in the identification signal.

2) BASIC IDEAS AND MATHEMATICAL MODELING OF THE FEEDBACK ADJUSTMENT MECHANISM

To ensure the versatility and generalization of the model, the deep convolutional neural network opens up almost all gate operations for the input signal and allows as much information as possible to pass through the entire network. Then, when targeting signals or targets for feature semantics, to improve discriminability at the feature level, an easy method is to turn off those neurons that are not associated with a given feature signal. This strategy is the key to biasing competition theory to explain neuron selectivity.

In this paper, a binary switch type hidden variable, which is called a feedback neuron, is introduced for each hidden layer of neurons. The control of these new neurons is realized by constructing the feedback connection between the target neurons and all the feedback neurons. Fig. 1 shows this process briefly. Fig. 1(a) shows the original CNN, and Fig. 1(b) shows the addition of switch-type feedback neurons to each hidden layer neuron. Fig. 1(c) shows the construction of a feedback connection between the target neuron and all feedback neurons. Fig. 1(a) has no feedback neurons. Fig. 1(b) does not establish feedback neurons, and Fig. 1(c) is a true feedback CNN. This configuration will improve the feature extraction capabilities of CNN. That is, a feedback layer is introduced in the traditional convolutional neural network to implement the selection function.

Bottom-up: It inherits the feature selectivity of the ReLU and max layers and passes the image information to the next layer.

Top-down: It is implemented by the feedback layer, which passes high-level semantic information to the data layer through gate operations. These gate operations only allow neurons associated with the target to be activated.

Thus far, this paper has introduced the basic idea of the feedback adjustment mechanisms in deep convolutional neural networks. Given an input signal, all neurons associated with a given target signal can be successfully screened out from the activated neurons. Then, the connecting pathway formed by these neurons becomes the connecting pathway that we need to screen out. Therefore, this paper needs to further clarify the feedback neuron state control problem.

As mentioned earlier, this paper introduces a large number of switch-type feedback neurons in deep convolutional neural networks. Simultaneously, a simple feedback connection is constructed between the target neuron and these feedback neurons to indicate that the state of the feedback neuron is controlled by the target neuron. By introducing the binary switch a, this paper further transforms the feedback mechanism into a numerical optimization problem. Given a signal I and a well-learned neural network, the parameter is w and a set $Z \in \{0, 1\}$ of binary switches in the network. This paper assumes that the target neuron output is S, and the mapping function of signal I to the target neuron S is $f(I, Z)$. This article attempts to maximize the target output by adjusting the switch state of all the feedback layers. The specific description is as follows:

$$\begin{aligned} &\max_{Z} S = f(I, Z) - \lambda \|Z\| \\ &s.t.z^{l}_{ijc} \in \{0, 1\}, \quad \forall i, j, c, l \\ &\qquad type(l) = ReLU \quad or \quad Max \end{aligned} \tag{4}$$

where z^{l}_{ijc} represents the binary switch of the position coordinate of the cth channel of the lth feedback layer being (i, j). Because our goal is to maximize the target output by activating the fewest neurons, this paper uses the L_1 norm to constrain the number of activations of z. Thus far, this paper applies the mathematical model of the feedback mechanism to the deep convolutional neural network. The problem described in formula (4) is the feedback optimization problem. This objective function is an integer-programming problem, and this is an NP-hard problem; in addition, in the model, the mapping function d is determined by the structure and parameters of the network, and it is impossible to write an explicit mathematical expression. The solution to the feedback optimization problem is not easy, and it is difficult to obtain the globally optimal solution. For the construction of the feedback connection mentioned above, the purpose of the feedback connection path is to transmit a feedback control signal to the feedback neuron; then the feedback neuron works in a predetermined manner. Therefore, this paper can construct the feedback connection in the process of solving the feedback optimization problem. However, from the feedback problem, it is difficult to obtain the globally optimal solution; different solution methods mean that the calculation method of the feedback control signal is different. So, it requires different feedback adjustment mechanisms. And it can be expected that different solutions to the feedback optimization problem will have different implementation effects. This problem will give the solution ideas and specific processes in the following content.

B. FEEDBACK OPTIMIZATION PROBLEM GRADIENT DESCENT METHOD

From formula (4), the optimization problem of formula (4) is an integer programming problem and is an NP-hard problem in the case of a given neural network structure. To facilitate the solution, this paper makes a linear approximation of the optimization problem and reduces the difficulty of the solution. First, from the analysis in Section II-A, it can be seen that both the ReLU and max layers can be regarded as gate control operations, while the max layer is usually connected after the ReLU layer. Essentially, it is a target-driven screening behavior that regulates the ReLU and max layers, and simultaneous optimization can complicate the problem. In fact, optimizing one of them can also achieve top-down neuron screening. Therefore, to simplify the problem, we retain the selection behavior of the max layer, add only a feedback layer behind the ReLU layer, and adjust the neuron screening behavior of the ReLU layer through the feedback layer to achieve target-driven information extraction. Second, $z \in \{0, 1\}$ makes the solution space of the feedback optimization problem discrete and the conventional optimization method is difficult to perform. Therefore, we relax the value constraint of Z and transform the optimization space into $0 \leq z^l_{ijc} \leq 1$; that is, the optimization space of the feedback optimization problem becomes continuous and steerable, so the above problem is transformed into the optimization problem of formula (5).

$$\begin{aligned} \max_{Z} S &= f(I, Z) - \lambda\|z\| \\ s.t. 0 &\leq z^l_{ijc} \leq 1, \quad \forall i, j, c, l \\ type(l) &= ReLU \end{aligned} \tag{5}$$

Therefore, the solution of the feedback optimization problem is no longer an NP-hard problem. It can solve the feedback optimization problem shown in formula (5) by the gradient descent method. The gradient of the objective function to each feedback neuron is calculated to simultaneously update all feedback neurons. Specifically, the entire network structure is continuously optimized in an iterative manner by formula (6).

$$z_{t+1} = z_t + \alpha \cdot \left(\left.\frac{\partial S}{\partial z}\right|_{z_t} - \lambda \right) \tag{6}$$

Among them, $\partial\lambda \, \|z\|_1 / \partial z_i = \lambda$ because this article limits $0 \leq z^l_{ijc} \leq 1$. The initialization state of the feedback layer neurons is consistent with the activation state of the ReLU layer corresponding to the first forward propagation. If the switch z exceeds the range of [0, 1] during the optimization process, it will be cut off. In formula (6), although the mathematical form of the function f cannot be directly written, the gradient solution in formula (6) can be derived by the chain rule and calculated by the error back propagation of the neural network. This process shows that the construction of the feedback connection in this paper does not need to display the construction but can share the information channel with the original neural network.

1) IMPLEMENTATION DETAILS

After each ReLU layer in a traditional neural network, we add a feedback layer. The switching amount of each feedback layer is initialized to z=1 to ensure that the first forward propagation can proceed normally. This paper updates the switch state based on the gradient of each feedback neuron backpropagation. The learning rate of the feedback layer is set to 0.1, and the switching states of all the feedback layers are updated at the same time. Each iteration requires a forward propagation and a backpropagation. Each signal needs to be iteratively updated 15-120 times, and the number of iteration updates is related to the characteristics of the processed signal. Because the number of iterations is related to the signal characteristics, after the multiple signals are tested, the optimal number of iterations is 15-120. In the range of this iteration update, the late feedback convolutional neural network will obtain a satisfactory effect.

The entire working process of the feedback convolutional neural network (FCNN) is shown in Fig. 2. For the first iteration, the network structure proposed in this paper is the same as that of the convolutional neural network, in which some neurons are not activated, that is, they are in the closed state. The black nodes in the figure represent the neurons that are turned off. The network then identifies a target neuron, lter2, which causes the feedback layer neurons to update their state through a feedback adjustment mechanism. Thereby to maximize the target neuron output. Then, in lter2, it maintains the state of the neuron that was turned off during the last feedback adjustment, repeating the feedforward and feedback adjustment process. This continues to iterate until convergence. Only one feedback layer update is shown in the figure for illustration. The working mechanism of the feedback convolutional neural network in this paper is compatible with this. The feedback convolutional neural network obtains a certain degree of screening. In the feedback iteration, it can retain the neurons related to the target and suppress the

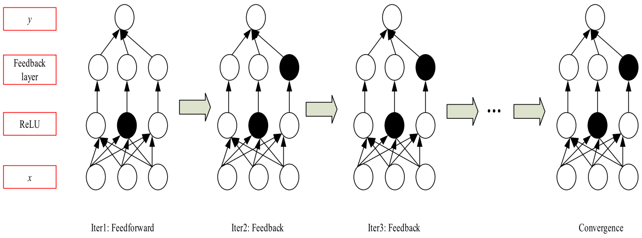


FIGURE 2. The FCNN workflow constructed in this paper.

neurons unrelated to the target signal, and can initially simulate the visual attention mechanism.

III. EFFECTIVE SUPERVISED FEATURE EXTRACTION METHOD BASED ON SPARSE EXPRESSION

A. STRUCTURE SPARSE REPRESENTATION

Sparse expression has great potential application in the field of fault diagnosis as an emerging signal processing technology. However, the general sparse expression method with a norm measure of sparsity only considers the feature selection at the atomic level, ignoring the structural effects inside the signal features (i.e., the strong correlation or irrelevance of one type of feature with other class features). This effect is that the key technology to simplifies pattern recognition. Inspired by this, it extends the idea of structural sparse expression, which assumes that the group structure (a group of strongly correlated samples) in the training sample is known. The coding process performs feature selection at the group level in groups instead of feature selection at a single atomic level in general sparse expression. The purpose of this paper is to study the fault feature extraction method of rolling bearings based on the structure sparse representation algorithm. In the process, the WPT transform is used as the basis function to construct a dictionary with structural effects, and the mixed penalty term is introduced to further optimize the performance of structural sparse expression.

1) OVERVIEW OF STRUCTURAL SPARSENESS

Sparse structure is a natural extension of the standard sparse theory. This paper understands the atoms in the dictionary as well-designed inputs, which are known information. On this basis, there is a theory that some of these inputs have related features, and the atoms belong to a specific group. Assume that there is an input $X = \cup_{j=1}^{C} G^j$, where $G^i \mathrm{I} G^j = \theta$, and $i \neq j$, the structure sparseness reflects the feature group structure by introducing a penalty term that makes the group sparse, specifically:

$$\beta = \arg\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \gamma \sum_{g=1}^{C} \left\|\beta_g\right\|_2 \tag{7}$$

where, β_g is the coefficient vector of the gth feature group Gg and γ is the weight parameter. Comparing the norm penalty terms in the standard sparse objective function and the structure sparse optimization objective function formula (7), the l_1 norm is the sparsity of the expression of the promotion coefficient, the l_2 norm is the feature group selection at the group level, and the l_2 norm does not make the coefficients in the feature group sparse. However, to accurately characterize the intrinsic data, the algorithm in this paper not only selects an important feature set but also makes it possible to further filter out the important features in the group, namely, the sparse group feature. A straightforward way to achieve this sparse group feature selection is to introduce a linear combination of two norm penalty terms (l_1 and l_2 norm) into the objective function, which is also called elastic net regularization. Therefore, it can obtain a new objective function, specifically:

$$\beta = \arg\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + v \|\beta\|_1 + \gamma \sum_{g=1}^{C} \left\|\beta_g\right\|_2 \tag{8}$$

B. SPARSE STRUCTURE SOLUTION

Set $\Omega(\beta) = \gamma \sum_{g=1}^{C} \left\|\beta_g\right\|_2$. The difficulty in the optimization of formula (7) is its inseparability in nonsmooth terms. In response to this problem, this paper uses a smooth near-end gradient iterative algorithm to solve this problem. The reason for choosing the algorithm is (1) the efficiency brought by the convergence speed $O\left(\frac{1}{\mu}\right)$, (2) the scalability of the step method, and (3) the ease of use. The main purpose of the algorithm is to find a smooth approximation so that the gradient can be calculated.

Thus, equivalent to the following formula:

$$\begin{aligned}\Omega(\beta) &= \gamma \sum_{g=1}^{C} \left\|\beta_g\right\|_2 = \gamma \sum_{g=1}^{C} \max_{\|\alpha_g\|_2 \le 1} \alpha_g^T \beta_g \\ &= \max_{\|\alpha_g\|_2 \le 1} \sum_{g=1}^{C} \gamma \alpha_g^T \beta_g = \max_{\|\alpha_g\|_\infty \le 1} \gamma \alpha^T \beta \end{aligned} \tag{9}$$

where α_g is the β_g auxiliary variable, $\alpha = \left[\alpha_1^T, \alpha_2^T, \mathrm{L}, \alpha_g^T\right]^T$. Using Nesterov smoothing, $\Omega(\beta)$ can be approximated as a smoothing function as follows:

$$\Omega(\beta) \approx f_\mu(\beta) = \max_{\|\alpha\|_\infty \le 1} (\gamma \alpha^T \beta - \mu d(\alpha)) \tag{10}$$

In the formula, $\mu \geq 0$, it represents the smoothing parameter and defines $d(\alpha)$ as $\frac{1}{2} \|\alpha\|_2^2$. According to the Nesterov theory, $f_\mu(\beta)$ is a continuous and differentiable convex function. Therefore, there is a gradient of $f_\mu(\beta)$ with respect to β, as follows:

$$\nabla f_\mu(\beta) = \gamma \alpha^* \tag{11}$$

where α^* is composed of α_g^*, $g = \{1, 2, \mathrm{L}, C\}$, which is the optimal solution of $f_\mu(\beta)$. The calculation result is as follows:

$$\alpha_g^* = S_2\left(\frac{\gamma \beta_g}{\mu}\right) \tag{12}$$

S_2 represents the projection operator, its role is to project $\frac{\gamma \beta_g}{\mu}$ onto the l_2 ball, and the specific formula is:

$$S_2(u) = \begin{cases} \dfrac{u}{\|u\|_2} & , \|u\|_2 > 1 \\ u, \|u\|_2 < 1 \end{cases} \tag{13}$$

Given the required precision ε, calculate $\mu = \varepsilon / C$. Therefore, formula (8) can be converted into:

$$\beta = \arg\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + v \|\beta\|_1 + f_\mu(\beta) \tag{14}$$

Thus, the original complex formula becomes formula (14), which is the l1 optimization problem. Extract the smoothing term in formula (14), specifically:

$$h(\beta) = \frac{1}{2} \|y - X\beta\|_2^2 + f_\mu(\beta) \tag{15}$$

Therefore, it can calculate the corresponding gradient, specifically:

$$\nabla h(\beta) = X^T (X^T \beta - y) + \gamma \alpha^* \tag{16}$$

According to the above formula, formula (14) can be solved by the FISTA iterative algorithm to obtain specific results. L in formula (17) is a step size parameter that can be obtained by a linear search method. For the problem presented in this paper, there is $L = \lambda_{\max}(X^T X) + \frac{\gamma}{\mu}$, where $\lambda_{\max}(X^T X)$ represents the maximum eigenvalue of matrix $X^T X$.

$$\beta^{t+1} = \arg\min h(z^t) + (\beta - z^t, \nabla h(z^t)) + v \|\beta\|_1 + \frac{L}{2} \|\beta - z^t\|_2^2 \tag{17}$$

The basic steps for solving it are as follows:

(1) Input and output settings, input: y, X, precision ε; output: $\beta = \beta^{t+1}$.

(2) Set the parameters, $\mu = \varepsilon / C$, $\theta_0 = 1$, $z^0 = \beta^0$.

(3) Iterative process: $t = 0, 1, \ldots,$. Until β^t converges, $\nabla h(z^t)$ is calculated according to formula (16), and β, $\beta^{t+1} = \arg\min_{\beta} h(z^t) + \langle \beta - z^t, \nabla h(z^t) \rangle + v \|\beta\|_1 + (L/2) \|\beta - z^t\|_2^2$ is updated by the following formula. Then set $\theta_{t+1} = (1 + \sqrt{1 + 4t^2})/2$ and update $z^{t+1} = \beta^{t+1} + (\theta_{t-1}/\theta_{t+1})(\beta^{t+1} - \beta^t)$.

C. FEATURE EXTRACTION METHOD BASED ON STRUCTURE SPARSE EXPRESSION

As seen from the foregoing, the solution to formula (8) addresses the limitation of standard sparse expression and general group sparse expression. It can achieve sparse coding at the feature group level, and the obtained expression highly expresses the essential information of the signal. Specifically, only the group coefficients that are highly correlated with the signal are characteristically reflected in the solution, and most other group coefficient values that are not related to the signal are zero. This expression, therefore, provides a comprehensive description of the signal with minimal space. In addition, this coefficient distribution is directly related to the signal type, and the intrinsic structure can reveal the class features and the structural features of the signal. In summary, the sparse structure can be further used for the fault feature extraction studied in this paper. The entire feature extraction process is WPT transformed into preprocessing. To meet the application requirements of the algorithm, the WPT transform result needs to be rearranged into a vector x according to the sub band frequency from low to high, and the subsequent operations are performed based on the sample WPT transformed and rearranged data. If there is a vibration signal dataset containing the C health status types, according to the above theory, it is first necessary to prepare a sample set S containing all types of data and to arrange the samples into a dictionary according to the category label as follows:

$$X_{m\times n} = \begin{bmatrix} x_{1,1}, x_{1,2}, \mathrm{L}, x_{1,n_1}(X_1), x_{2,1}, x_{2,2}, \\ \mathrm{L}, x_{2,n_2}(X_2), \mathrm{L}, x_{C,1}, x_{C,2}, \mathrm{L}, x_{C,nC}(X_C) \end{bmatrix} \tag{18}$$

where $n = n_1 + n_2 + \ldots + n_C$ and $X_i (i = 1, 2, 3, \ldots, C)$ represents a sub dictionary corresponding to the i-th type of data. Similarly, the corresponding coefficients can be expressed as:

$$\beta = \begin{bmatrix} \beta_{1,1}, \beta_{1,2}, \mathrm{L}, \beta_{1,n_1}(\beta_1), \beta_{2,1}, \beta_{2,2}, \\ \mathrm{L}, \beta_{2,n_2}(\beta_2), \mathrm{L}, \beta_{C,1}, \beta_{C,2}, \mathrm{L}, \beta_{C,nC}(\beta_C) \end{bmatrix}^T \tag{19}$$

When a new sample y belonging to category τ is given, it can use the algorithm to solve the coefficient vector β, as shown in formula (20).

$$\beta = \Big[0, \cdots, 0, \underbrace{\beta_{\tau,1}, 0, \cdots, 0, 0, \beta_{\tau,k}, \cdots, \beta_{\tau_i, n_i},}_{\beta_i} \cdots, 0, \cdots, 0 \Big] \tag{20}$$

The above description shows that the coefficient vector β is a representation of the input signal y mapped to the sparse domain, and its nonzero region can clearly indicate the category information of the signal y. To further enhance the representativeness of the features and improve the efficiency of subsequent diagnosis, this paper calculates the l_1 norm and rearranges the sparse expression coefficients of the structure and obtains the structure-based sparse wavelet feature (SSW) proposed in this paper. The formula is as follows.

$$SSW = \left[\|\beta_1\|_1, \|\beta_2\|_1, \mathrm{L}, \|\beta_C\|_1 \right] \tag{21}$$

It can be known from formula (21) that the dimension of the SSW feature is equal to the number of feature groups, that is, the number of data types contained in the entire data. This means that the SSW has a lower feature size. It ensures the efficiency of subsequent classification work.

In addition, the high recognition and strong adaptability of the SSW features are also obvious application advantages.

This feature can make different categories of data correctly recognized and have lower classification errors, so the degree of recognition determines whether the features are high or low for the classifier and the feature classification effect. There is a one-to-one mapping between the distribution of sparsely expressed structures and data structure types, and the SSW features calculated from this distribution are the most direct representation of data structure information. Therefore, even if the number of classifications is large, the SSW feature is still very strong, and the requirements for subsequent classifiers are also reduced.

Thus far, this paper has theoretically explained the advantages of SSW features for mechanical fault diagnosis. The next section will introduce a method based on FCNN and sparse expression to establish a rolling bearing fault diagnosis method.

IV. ROLLING BEARING FAULT DIAGNOSIS METHOD BASED ON FCNN-SPARSE EXPRESSION

A. ROLLING BEARING FAULT DIAGNOSIS METHOD BASED ON FCNN-SPARSE EXPRESSION

Section II elaborates on the FCNN model proposed in this paper. When solving the complex signal feature extraction problem through the FCNN model, to improve the learning ability of the FCNN network, the corresponding number of hidden layer units must be increased. This will complicate the deep learning model calculation process, which will increase the difficulty of signal feature analysis and overwhelm some signals with typical characteristics. The sparse representation technique introduced in Section III can effectively eliminate redundant information and retain valid feature information, which provides a more concise expression for the data. Thus, this paper proposes a fault diagnosis method for rolling bearings based on FCNN-sparse expression.

Here, it is necessary to add a sparse penalty term to the log-likelihood objective function in the FCNN so that the deep learning model hidden layer unit maintains a low degree of activation. Assuming a given sample set $v = \left(v_1, v_2, \mathrm{L}, v_{n_v}\right)$, the optimization problem expression is as follows:

$$\min imize_{\{w_{i,j}, b_i, a_j\}} - \sum_{l=1}^{n_v} \ln P(v;\theta) + \lambda \sum_{j=1}^{n_h} \left| p - \frac{1}{n_v} \sum_{l=1}^{n_v} E\left[h_j | v\right] \right|^2 \qquad (22)$$

where $E[\cdot]$ is the conditional probability given a training sample, λ is a regularization parameter, and p is a parameter that controls the average activation degree of the hidden layer unit.

The parameter calculation of the sparse FCNN can be calculated by the gradient method. The CD algorithm can update the parameter calculation of formula (22). And then the parameter is updated again by the gradient of the added penalty item until convergence (It can be known from the proof that it is only necessary to update the offset parameter a_j which controls the degree of activation of the hidden layer).

TABLE 1. Rolling bearing fault diagnosis process based on sparse FCNN.

Step1	The rolling bearing data collected by various sensors is preprocessed and divided into training sets and test sets.
Step 2	Feature extraction is performed on the training set and test set data obtained in step 1.
Step 3	A classifier model for rolling bearing signals is constructed based on the sparse FCNN theory.
Step 4	The sparse FCNN classifier obtained in step 2 is trained using the feature information obtained from the training set.
Step 5	The classification error of step 4 is analyzed, and the evaluation of the sparse FCNN classification effect is given.
Step 6	The FCNN classifier model trained in step 4 is used for subsequent rolling bearing fault diagnosis.

B. APPLICATION OF SPARSE FCNN IN FAULT DIAGNOSIS

Data preprocessing is a necessary step for feature extraction. The sparse FCNN used in this section optimizes the network parameters in a global parameter fine-tuning manner, the learning results are sparse, and the essential representation of the data can be abstracted from a large amount of data. Therefore, simple data preprocessing can achieve the application requirements. Sparse FCNN-based fault diagnosis is divided into training and testing steps, both of which require a large number of samples to be prepared in advance. To improve training efficiency, the sample will be further divided into "mini-batch" processing. The general steps for the health status diagnosis of rolling bearings based on sparse FCNN are shown in Table 1, in which steps 3 and 4 are cycled until the number of trainings is reached.

C. BLOCK FAULT DIAGNOSIS SYSTEM

This section will use the sparse FCNN model to build a block fault diagnosis network. Different from the traditional fault diagnosis system, the task of the block fault diagnosis network proposed in this section is not limited to accurately identifying the fault categories of rolling bearing parts, but it innovatively builds a hierarchical diagnostic system so that the network can further diagnose the severity of the fault. This is the key to discovering the weak links in mechanical systems. If the fault diagnosis process cannot detect the weak points of the system or find the serious faults in time and the system performance will drop rapidly. Therefore, the diagnosis of the severity of mechanical components is particularly important for the reliability design of the mechanical equipment.

1) STEPS FOR BUILDING A BLOCK FAULT DIAGNOSIS NETWORK

The data used in the validity verification experiment of the network are the rolling bearing data. To clarify the whole network construction step, the bearing classification task is taken as an example to illustrate the complete network

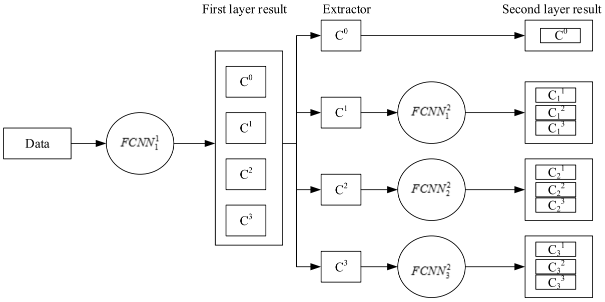


FIGURE 3. Schematic diagram of the sparse FCNN block fault diagnosis network.

construction steps. The rolling bearing dataset contains a set of normal signals, three sets of fault signals, and all types of faults contain three fault sizes. The submodule fault diagnosis network structure includes two independent functional layers ($L1$ and $L2$) to determine the fault form and the fault degree of the data. A schematic diagram of the structure of the block fault diagnosis structure for the current data is shown in Fig. 3. The network consists of two sparse FCNN modules connected successively. The operation of sparse $FCNN_1^1$ is the fault type identification, and sparse $FCNN_t^2, t = 1, 2, 3$ further identifies the fault degree.

Of course, the first prerequisite for completing the submodule fault diagnosis network is to learn the modules of the sparse *FCNNs* with the subfunctions shown in Fig. 3 according to the sparse *FCNN* learning method in Section IV-B and the sparse *FCNN* training steps described in Table 1. Specifically, the sample used for training $FCNN_1^1$ needs to include all fault types and fault levels, and the samples for training $FCNN_t^2$ need to include multiple fault degree samples under the corresponding fault type. According to Fig. 3, the first module and the second module work continuously without interruption. The result of the first level diagnosis is input into the corresponding subclassifier of the second level for further identification. Therefore, to achieve this function, it is necessary to add a data extractor before each subclassifier of the second-level diagnosis and assign the sample to the corresponding $FCNN_t^2$ according to the sample class label obtained by the first-level identification for further fault degree judgment.

2) CLASSIFICATION ACCURACY OF THE SEGMENTATION FAULT DIAGNOSIS NETWORK

The definition formula of the classification accuracy rate is the number of correctly determined samples compared to the total number of discriminated samples. Considering the two-level diagnostic structure of the submodule fault diagnosis network of Fig. 3, the network contains four sparse *FCNNs*, corresponding to four classification accuracy rates. By definition, the identification accuracy of $L1$ can be easily obtained based on the classification result of sparse $FCNN_1^1$. Taking into account the particularity of the network structure, the data extractor transfers the sample to the $FCNN_t^2$ corresponding to the class label according to the identification result of the $L1$ module. That is, the data received by $FCNN_t^2$ contains the error samples caused by the $L1$ module misjudgment, and the classification of these misjudgment samples in the $L2$ module is invalid. Therefore, the classification result of $L1$ directly affects the classification accuracy of the $L2$ module. This article uses the simplest two-category data identification submodule diagnostic example to illustrate. Assuming that there is no classification error in the $L1$ layer, $FCNN_2^1$ and $FCNN_2^2$ of the $L2$ layer will each obtain m samples. If the $L1$ module is classified, each r and s sample is misclassified into the relative categories. The next $FCNN_2^1$ and $FCNN_2^2$ are extracted to $(m + s\text{-}r)$ and $(m + r\text{-}s)$ samples, as shown in Fig. 4. In particular, the newly added r and s samples are invalid samples generated in the $L1$ module, which will lead to additional diagnostic errors to $L2$. Therefore, the classification error of the $L2$ module consists of two parts: the $L1$ fault type is judged correctly, the $L2$ fault degree misjudgment error and the $L1$ fault type misjudgment error. Assume that the number of misjudgment samples of the first part of the error generated by $FCNN_2^1$ and $FCNN_2^2$ of L2 is α and β, respectively. Based on the above theory, the accuracy of $FCNN_2^1$ is calculated as 1-$((s + \alpha)/(m + s\text{-}r))$, the accuracy of $FCNN_2^2$ is 1-$((r + \beta)/(m + r\text{-}s))$ and the classification accuracy of the network given in Fig. 3. It is 1-$((r + s + \alpha + \beta)/2m)$.

TABLE 2. Rolling bearing structural parameters.

Outer diameter /mm	Inner diameter. /mm	Number of balls	Contact angle (^{0})	Ball diameter /mm
51.99	25	9	0	7.94

V. EXPERIMENT ANALYSIS

A. EXPERIMENT 1

To verify the effectiveness of the proposed solution, the rolling bearing data of the Case Western Reserve University bearing data center website was used for analysis. These data have been widely studied and applied by scholars in the field of fault diagnosis at home and abroad, and they have become a standard dataset for verifying new methods of state recognition of mechanical equipment [38]–[43]. The experimental device includes an electrical control device, a motor, and a torque sensor/angular speed encoder. The rolling bearing model is 6205-SKF, and the bearing structural parameters are shown in Table 2.

Under the condition that the driving motor speed is 1797 r/min, the piezoelectric acceleration sensor and its supporting magnetic seat are used to rigidly connect the sensor with the supporting seat where the fault bearing is located. The original vibration signals of the rolling bearing under normal conditions, rolling element failure, inner ring failure and outer ring failure are collected. The bearing outer ring

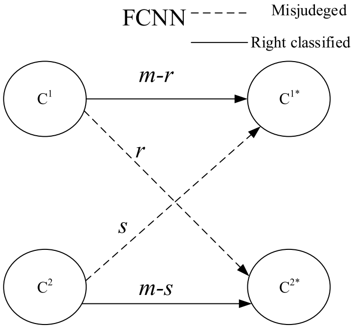
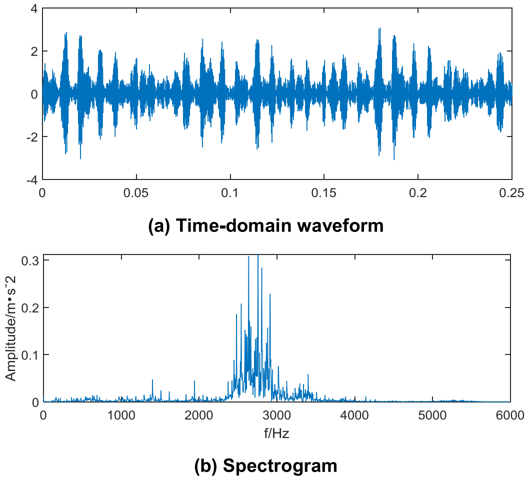


FIGURE 4. FCNN error diagram.



(a) Time-domain waveform

(b) Spectrogram

FIGURE 5. Time-domain waveform and spectrum of the outer ring fault signal.

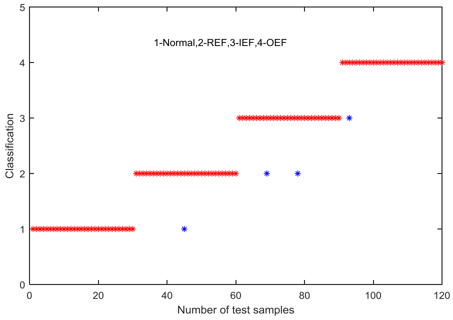


FIGURE 6. This paper's method classification result graph.

fault is now studied. The sampling points are 3000 points, the sampling frequency is 12 kHz, and the sampling time is 0.25 seconds (144.dat). The time-domain waveform and spectrum of the measured signal are shown in Fig. 5.

In order to analyze the fault signal, the theory proposed in Section III of this paper is used to sparsely express the above fault signal; next, the key signal information is identified, and then, the redundant information is eliminated. Finally, under the four working conditions (normal state, rolling element fault, inner ring fault and outer ring fault, expressed by Normal, REF, IRF and ORF, respectively), 100 sets of data are randomly selected for sparse expression preprocessing and used as training samples. The remaining 60 data points are used as test samples. Then, the fault diagnosis method proposed in Section IV of this paper is used for fault diagnosis. To verify the advantages of this method, the training and test data are diagnosed by the CNN, References [33], [44], SVM and artificial neural network. The specific test results are shown in Fig. 6-Fig. 10. Normal, REF, IRF, and ORF are replaced with the numbers 1, 2, 3, and 4, respectively.

It can be seen from Fig. 6-Fig. 10 that the fault diagnosis method proposed in this paper has the highest average correctness rate, the correct rate of pattern recognition under various working conditions reaches over 93%, and the recognition rate of bearings in the normal state is as high as 100%. The average recognition accuracy of the CNN method is only 1% lower than the method in this paper. This is because the CNN method can adaptively learn most of the information of the fault signal. However, the CNN method does not perform sparse expression processing on the fault signal, which increases the interference signal. Therefore, the recognition rate of the CNN method is lower than the method proposed in this paper. The average recognition accuracy of the SVM method for the four operating conditions is only 90%. This is because the SVM does not have a good anti-interference ability, and it cannot be adaptively learned based on the fault signal, thus performing classification and recognition.

At the same time, in order to better demonstrate the advantages of the feedback convolutional neural network proposed in this paper compared with other feedback convolutional

TABLE 3. Statistics of three algorithms for fault diagnosis results.

Method type	Number of train	Numb er of test	Classification accuracy of different working conditions (%)				Average accuracy (%)
			Normal	REF	IRF	ORF	
Ours	100	30	100	96.7	93.3	96.7	96.7
[33]	100	30	100	96.7	90.0	96.7	95.9
[44]	100	30	100	96.7	93.3	93.3	95.8
CNN	100	30	100	96.7	90.0	93.3	95.0
SVM	100	30	96.7	93.3	86.7	86.7	90.9

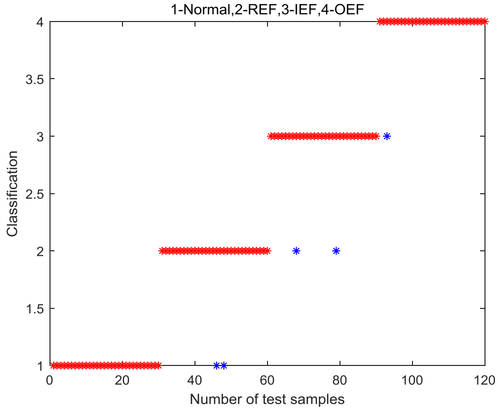


FIGURE 7. [44] method classification result graph.

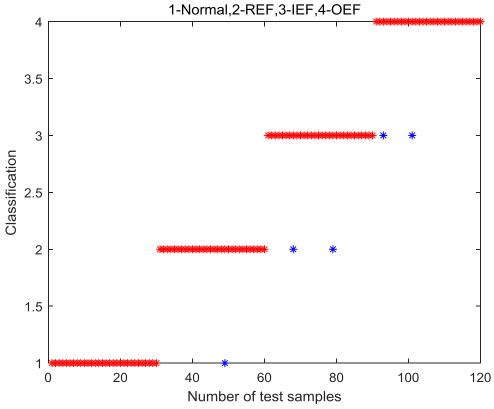


FIGURE 8. [45] method classification result graph.

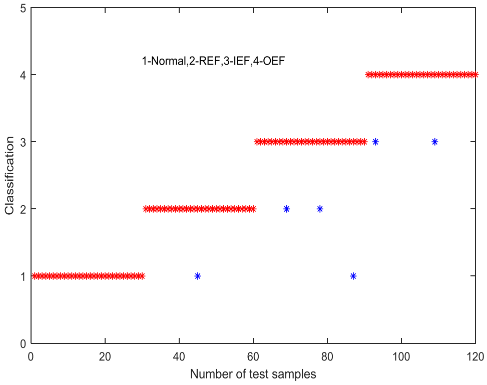


FIGURE 9. CNN method classification result graph.

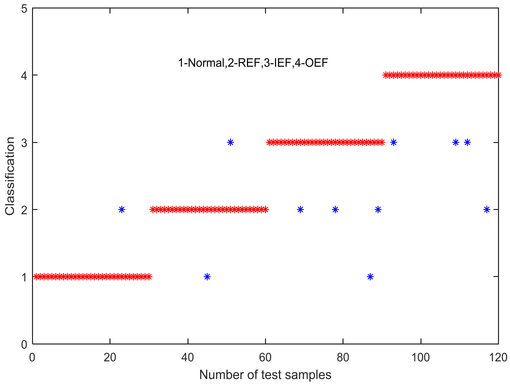


FIGURE 10. SVM method classification result graph.

neural networks, the method proposed in [33], [44] is added for comparison. The comparison results are shown in Table 3. It can be seen that the method proposed in this paper is more accurate than the reference [33], [44]. This is because the proposed method not only introduces a feedback adjustment mechanism, but also performs corresponding optimization operations. And it use sparse expression to find more effective feature information.

The results also show that in the case of equal training samples, the proposed method has higher accuracy and better diagnostic efficiency than the CNN and SVM methods. The method not only improves the means of fault diagnosis of rolling bearings but can also be applied to the field of fault diagnosis of other rotating machinery (rotor, gear, etc.), which has great engineering application value.

TABLE 4. Three methods calculation time comparison table.

Method type	Training time (ms)	testing time (ms)
Ours	38487.15	3.41
CNN	37962.36	3.45
SVM	7.98	13.47

In order to further compare the calculation time of the three methods. Table 4 shows the calculation time required for these three methods. The experimental software and hardware environment processes are the same. The training time is the processing time of 100 sets of training samples in the experimental data set. The test time is the recognition time of 60 sets of test samples in the experimental data set.

In this experiment, the data is transmitted to the fault diagnosis analysis module for analysis through the network method. As can be seen from Table 4, the SVM method has the least training time, but its testing time is too long. Compared with the CNN method, although the method in this paper is longer than the CNN method, the training time gap is small. However, this method is shorter than the CNN method. This shows that the proposed method can adaptively identify the fault signal more quickly. At the same time, it also shows that the method of this paper is of great value to practical engineering applications. The feedback convolutional neural network mentioned in this paper takes more time to construct the model and sparse process. But the test process is directly using the trained model for testing. At this point, it takes less time, and it also reflects the need for feedback convolutional neural network modeling to take more time.

B. EXPERIMENT 2

The data used in the experiment are the Case Western Reserve University (CWRU) [41]. The experimental data of this group include four types of bearing status: normal, outer ring fault, inner ring fault and roller fault. The faulty bearing is a single damage condition, which is artificially introduced by EDM. The diameter of the damage is of the following four types: 0.007, 0.014, 0.021 and 0.028 inches. The fault depth is 0.011 inches. The data were collected at four motor load speeds of 0, 1, 2, and 3 hp (corresponding speeds of 1720-1797 rpm). The specific experimental conditions were as follows: the drive end bearing adopts 6205-2RS JEM SKF deep groove ball bearing, the EDM bearing has single point damage, the damage diameter is 0.1778 mm, and the sampling frequency is 12 KHz.

For any sample in the above dataset, using the block fault diagnosis network system proposed in this paper to analyze these datasets can not only diagnose the fault but can also determine the degree of the fault. To facilitate experimental analysis, the vibration signal for each state is segmented into

TABLE 5. Experimental rolling bearing data information table (the listed samples are added to the load condition of 0-3 hp).

Sample type	Fault size	Sample collection	Number of samples	Sample length
Normal	0	Normal	200	5012
Inner ring failure	0.007	Inner ring07	200	5012
	0.014	Inner ring14	200	5012
	0.021	Inner ring21	200	5012
Outer ring fault	0.007	Outer ring7	200	5012
	0.014	Outer ring14	200	5012
	0.021	Outer ring21	200	5012
Roller failure	0.007	Roller7	200	5012
	0.014	Roller14	200	5012
	0.021	Roller21	200	5012

TABLE 6. Block fault diagnosis network experimental data statistics table.

Parameter name	C^0 (Normal)	C^1(Outer ring fault)			C^2(Inner ring failure)			C^3(Roller failure)		
		C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	C_3^2	C_1^3	C_2^3	C_3^3
Fault size	0	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021
Number of train	100	100	100	100	100	100	100	100	100	100
Number of test	100	100	100	100	100	100	100	100	100	100

a time series of 5012 no-sampling points. At the same time, to ensure the rationality and generality of the dataset, the same type of normal data collected under different loads is merged into the dataset to form an experimental sample, and the number of samples is 200, as shown in Table 5.

As a result, 50% (100) of the samples are randomly extracted from the sample data of each state as training data, and the rest are used as test data. Therefore, the training sample and the test sample are each 100. The specific experimental data are shown in Table 6. In this experiment, the data is transmitted to the fault diagnosis analysis module for analysis through the network method.

According to the above, the sparse $FCNN_1^1$ of the first module is trained with all the training samples, and the sparse $FCNN_k^2$ of the corresponding class is trained by the samples (200) of different damages under the same fault. The parameters of all the sparse *FCNNs* in the network pretraining are configured according to the applicable degree of the signals, as shown in Table 7. The mini-batch size is usually the same as the number of categories. The fine-tuning process of the network weight parameter is followed by the pretraining step of the network to further improve the classification performance of the network.

To further demonstrate the application advantages of the submodule fault diagnosis network based on sparse FCNN, we compare it with the BP neural network, support vector machine and CNN method. To ensure the generality of the experimental process, the same structure of the submodule fault diagnosis network proposed in this paper is used to replace the sparse *FCNN* in the diagnostic network with the BP neural network, SVM or CNN to complete a similar block fault diagnosis. In the experiment, the BP neural network structure consists of 6 hidden layers, 12 neurons per layer, and the weight and bias of the network are updated by the Levenberg-Marquard optimization algorithm. The multiclass SVM classifier uses the radial basis function as the kernel function and the training method to select the one-against-all method. According to the data extraction mode of this experiment, half of the data listed in Table 2 were randomly selected as a training sample, and the rest were used as a test sample. To obtain more accurate and general results, we repeat the data extraction 100 times and diagnose the 100 networks of the diagnostic networks constructed by the above three classification algorithms and obtain the average fault diagnosis accuracy of the three algorithms. The details are shown in Table 8.

TABLE 7. Statistical data of the block fault diagnosis network training parameters proposed in this paper.

Parameter name	SparseFCNN_1^1	SparseFCNN_1^2	SparseFCNN_2^2	SparseFCNN$_3^2$
Number of hidden layers	4	4	4	4
Number of neurons per layer	70	60	50	50
Learning rate	0.00075	0.00075	0.0007	0.013
Number of training	300	200	200	200
Mini-batch size	4	3	3	3

TABLE 8. Statistics of the four algorithms fault diagnosis results.

Network Type	L1 Accuracy (%)	L2 Accuracy (%)			Average accuracy (%)
		C^1	C^2	C^3	
Method of this paper	99.69	99.72	99.58	96.21	98.8
CNN	98.75	97.57	96.54	95.29	97.04
BP neural network	94.36	93.68	92.37	96.31	94.18
SVM	93.87	95.9	96.17	95.46	95.35

It can be seen from Table 8 that for the diagnosis result of the L1 module, the sparse FCNN obtains the highest classification accuracy of 98.8% among the four algorithms, and the identification performance for the bearing fault location is significantly better than the BP neural network (BPNN, 94.18%) and SVM (95.35%). The results show that through effective training, sparse FCNN can improve the confusion of various feature patterns due to different degrees of failure. This method has certain advantages over CNN (97.04%). This is because sparse FCNN has eliminated redundant signal information and avoids interference of invalid signals, thus eliminating the influence of training accuracy in subsequent deep learning. For the maintenance and condition monitoring of mechanical equipment, an accurate understanding of the type of fault is the primary task, namely, the first module of the fault diagnosis network proposed in this paper. Therefore, in the task completion of the L1 module, the classification

TABLE 9. Four methods calculation time comparison table.

Method type	Training time (ms)	testing time (ms)
Ours	45981.39	3.98
CNN	45102.93	4.03
BPNN	15.36	14.92
SVM	9.32	15.95

performance of the sparse FCNN is much better than the BP neural network and the SVM intelligent algorithm, which is also better than the traditional CNN deep learning method.

As explained above, due to the successive connections of the front and rear modules, the misjudgment of the first module will have a direct negative impact on the second module. It causes the second module classification result to include the error of the previous module due to the fault type in addition to the classification error of the module. For the diagnosis results of the L2 module, the sparse FCNN obtains the highest classification accuracy among the four algorithms. In the three categories of L2 fault diagnosis, the method has the highest classification accuracy, and there is no considerable fluctuation. This shows that the method has good adaptive characteristics and self-learning characteristics. The other three methods have certain fluctuation, which indicates that these three methods do not have the ability of adaptive learning to some extent.

To further analyze the calculation time of these four methods, Table 9 lists the calculation times for these four methods. The experimental software and hardware environment processes are the same. The training time is the processing time of the training samples in the experimental data set, and the test time is the recognition time of the test samples in the experimental data set.

It can be seen from Table 9 that the test time of this method is the least, and the training time of the SVM method is the least, but its test time is longer. Compared with the CNN method, although the method has a small difference between the training time and the CNN method, the method is shorter than the CNN method. This also shows that the model trained in this method can adaptively identify the test set. The feedback convolutional neural network mentioned in this paper takes more time to construct the model and sparse process. But the test process is directly using the trained model for testing. At this point, it takes less time, and it also reflects the need for feedback convolutional neural network modeling to take more time.

In summary, the submodule fault diagnosis network proposed by sparse FCNN can realize the comprehensive fault diagnosis of rolling bearings that locates both the fault location and the bearing fault degree. Compared with CNN, BP neural network and the SVM intelligent classifier, it is found that the sparse FCNN-based submodule fault diagnosis network is better for the function of each module. Therefore, the deep learning algorithm represented by FCNN has great mining value in the field of mechanical fault diagnosis.

VI. CONCLUSION

In this paper, fault feature extraction and intelligent diagnosis of rolling bearings in core parts of rotating machinery are presented. Based on the sparse expression and convolutional neural network theory of current research hotspots, a fault diagnosis method based on feedback convolutional neural network-sparse expression is proposed. First, the feedback connection is constructed in the convolutional neural network around the deep convolutional neural network feedback target, and the mathematical definition of the feedback adjustment mechanism problem is given according to the feedback connection and the feedback target. Finally, the feedback optimization problem is abstracted. A new framework of feedback convolutional neural networks based on gradient descent is proposed. Then, sparse coding technology is used to sparsely express the deep learning model, eliminate redundant information, and seek a simple representation of the outstanding essence of the data to reduce the difficulty of data analysis. The main results are as follows:

(1) In this paper, the vibration signal of a rolling bearing is affected by noise interference and other factors. The sparse expression of various state data will produce intraclass mode differences. Therefore, a new structural sparse representation algorithm is introduced to preprocess the signal. It addresses the application limitation of standard sparse expression only selecting features at the atomic level. It takes into account the structural properties inside the signal. It can get a more meaningful coefficient distribution. It can then be used to identify the signal. The experiment proves that the feature has strong stability and recognition and has great application potential in the fault diagnosis of rolling bearings.

(2) In the deep convolutional neural network, the calculation model of the feedback adjustment mechanism is proposed first. This paper reinterprets the composition of convolutional neural networks from the perspective of feedback, highlights the existence of a stimulus-driven neuron screening mechanism and its existing problems in the feedforward process of convolutional neural networks, and proposes the basic framework for the operation of feedback adjustment mechanisms. The mathematical model of constructing the whole feedback mechanism is completed, and the feedback optimization problem is clarified. So, a gradient descent method is proposed to solve the feedback optimization problem we constructed, and the corresponding feedback convolutional neural network architecture is given.

(3) In this paper, the sparse representation of the structure and the feedback convolutional neural network method are proposed. The fault diagnosis method of rolling bearing based on feedback convolutional neural network-sparse representation is proposed. The diagnostic method was applied to the analysis of rolling bearing data of the Case Western Reserve University Bearing Data Center website and compared with other existing mainstream rolling bearing fault diagnosis methods. The experimental results show that the proposed

method can not only identify the rolling bearing fault type but also identify the rolling bearing fault degree. The recognition effect is better than other mainstream rolling bearing fault diagnosis algorithms. The method proposed in this paper has the shortest test time for the rolling bearing signal. In other words, it can accurately identify whether the rolling bearing signal is faulty in the shortest time.

COMPLIANCE WITH ETHICAL STANDARDS

A. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

B. DATA AND CODE AVAILABILITY STATEMENT

The data and code used to support the findings of this study are included within the paper.

REFERENCES

[1] I. Attoui, N. Fergani, N. Boutasseta, B. Oudjani, and A. Deliou, "A new time–frequency method for identification and classification of ball bearing faults," *J. Sound Vib.*, no. 397, pp. 241–265, Jun. 2017.

[2] T. Han and D. Jiang, "Rolling bearing fault diagnostic method based on VMD-AR model and random forest classifier," *Shock Vib.*, vol. 2016, Jun. 2016, Art. no. 5132046.

[3] J. B. Ali, N. Fnaiech, L. Saidi, B. Chebel-Morello, and F. Fnaiech, "Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals," *Appl. Acoust.*, vol. 89, pp. 16–27, Mar. 2015.

[4] H. Shao, H. Jiang, X. Zhang, and M. Niu, "Rolling bearing fault diagnosis using an optimization deep belief network," *Meas. Sci. Technol.*, vol. 26, no. 11, Sep. 2015, Art. no. 115002.

[5] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mech. Syst. Signal Process.*, vol. 85, pp. 746–759, Feb. 2017.

[6] T. A. Harris, *Rolling Bearing Analysis*. Hoboken, NJ, USA: Wiley, 2001.

[7] T. A. Harris and M. N. Kotzalas, *Advanced Concepts of Bearing Technology: Rolling Bearing Analysis*. Boca Raton, FL, USA: CRC Press, 2006.

[8] H. Cao, L. Niu, S. Xi, and X. Chen, "Mechanical model development of rolling bearing-rotor systems: A review," *Mech. Syst. Signal Process.*, vol. 102, pp. 37–58, Mar. 2018.

[9] H. Gao, L. Liang, X. Chen, and G. Xu, "Feature extraction and recognition for rolling element bearing fault utilizing short-time Fourier transform and non-negative matrix factorization," *Chin. J. Mech. Eng.*, vol. 28, no. 1, pp. 96–105, 2014.

[10] L. Saidi, J. B. Ali, and F. Fnaiech, "Bi-spectrum based-EMD applied to the non-stationary vibration signals for bearing faults diagnosis," *ISA Trans.*, vol. 53, no. 5, pp. 1650–1660, 2014.

[11] L. Lu, J. Yan, and C. W. de Silva, "Dominant feature selection for the fault diagnosis of rotary machines using modified genetic algorithm and empirical mode decomposition," *J. Sound Vib.*, vol. 344, pp. 464–483, May 2015.

[12] Z. Feng, M. Liang, and F. Chu, "Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 165–205, Jul. 2013.

[13] H. Sun, Z. He, and Y. Zi, "Multiwavelet transform and its applications in mechanical fault diagnosis—A review," *Mech. Syst. Signal Process.*, vol. 43, no. 2, pp. 1–24, Feb. 2014.

[14] J. Harmouche, C. Delpha, and D. Diallo, "Incipient fault detection and diagnosis based on Kullback–Leibler divergence using principal component analysis: Part II," *Signal Process.*, vol. 109, pp. 334–344, Apr. 2015.

[15] S. Dong, J. Sheng, B. Tang, L. Zhong, X. Xu, L. Chen, Q. Hu, J. Luo, L. Zhao, and R. Chen,"Bearings in simulated space conditions running state detecton based on Tsallis entropy-KPCA and optimized fuzzy c-means model," *Noise Control Eng. J.*, vol. 65, no. 2, pp. 62–70, Apr. 2017.

[16] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, nos. 1–2, pp. 108–126, Feb. 2013.

[17] A. Rai and S. H. Upadhyay, "The use of MD-CUMSUM and NARX neural network for anticipating the remaining useful life of bearings," *Measurement*, vol. 111, pp. 397–410, Dec. 2017.

[18] J. C. M. Oliveira, K. V. Pontes, I. Sartori, and M. Embiruçu, "Fault detection and diagnosis in dynamic systems using weightless neural networks," *Expert Syst. Appl.*, vol. 84, pp. 200–219, Oct. 2017.

[19] S. Shao, W. Sun, P. Wang, R. X. Gao, and R. Yan, "Learning features from vibration signals for induction motor fault diagnosis," in *Proc. Int. Symp. Flexible Automat. (ISFA)*, Aug. 2016, pp. 71–76.

[20] V. T. Tran, F. Al Thobiani, and A. Ball, "An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4113–4122, Jul. 2014.

[21] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016.

[22] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, pp. 425–437, Feb. 2017.

[23] H. Jiang, F. Wang, H. Shao, and H. Zhang, "Rolling bearing fault identification using multilayer deep learning convolutional neural network," *J. Vibroeng.*, vol. 19, no. 1, pp. 138–149, Feb. 2017.

[24] X.-H. He, D. Wang, Y.-F. Li, and C.-H. Zhou, "A novel bearing fault diagnosis method based on Gaussian restricted Boltzmann machine," *Math. Problems Eng.*, vol. 2016, Dec. 2016, Art. no. 2957083.

[25] V. A. F. Lamme, H. Supèr, and H. Spekreijse, "Feedforward, horizontal, and feedback processing in the visual cortex," *Current Opinion Neurobiol.*, vol. 8, no. 4, pp. 529–535, 1998.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[27] S. Loffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017, *arXiv:1610.02357*. [Online]. Available: https://arxiv.org/abs/1610.02357

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[31] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[32] A. Shrivastava and A. Gupta, "Contextual priming and feedback for faster R-CNN," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 330–348.

[33] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2956–2964.

[34] U. R. Acharya, S. L. Oh, Y. Hagiwara, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Comput. Biol. Med.*, vol. 100, pp. 270–278, Sep. 2018.

[35] Y. Pan and H. Yu, "Biomimetic hybrid feedback feedforward neural-network learning control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1481–1487, Jun. 2017.

[36] C. Wu, J. Wang, J. Liu, and W. Liu, "Recurrent neural network based recommendation for time heterogeneous feedback," *Knowl.-Based Syst.*, vol. 109, pp. 90–103, Oct. 2016.

[37] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, ‘‘Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval,’’ *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 694–707, Apr. 2016.

[38] T. W. Rauber, F. de A. Boldt, and F. M. Varejão, ‘‘Heterogeneous feature models and feature selection applied to bearing fault diagnosis,’’ *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 637–646, Jan. 2015.

[39] X. Ding, Q. He, and N. Luo, ‘‘A fusion feature and its improvement based on locality preserving projections for rolling element bearing fault classification,’’ *J. Sound Vib.*, vol. 335, pp. 367–383, Jan. 2015.

[40] W. Mao, L. He, Y. Yan, and J. Wang, ‘‘Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine,’’ *Mech. Syst. Signal Process.*, vol. 83, pp. 450–473, Jan. 2017.

[41] K. A. Loparo. (Oct. 16, 2011). *Case Western Reserve University Bearing Data Center Seeded Fault Test Data [EB/OL][2015-06-09]*. [Online]. Available: http;cse-groups.case.edu/bearing data center pages /12K -drive-end- bearing-fault-data

[42] Y.-J. Cha, W. Choi, and O. Büyüköztürk, ‘‘Deep learning-based crack damage detection using convolutional neural networks,’’ *Comput. Aided Civil Infrastruct. Eng.*, vol. 32, no. 5, pp. 361–378, May 2017.

[43] X. Ding and Q. He, ‘‘Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis,’’ *IEEE Trans. Instrum. Meas.*, vol. 66, no. 8, pp. 1926–1935, Aug. 2017.

[44] S. Li, G. Liu, X. Tang, J. Lu, and J. Hu, ‘‘An ensemble deep convolutional neural network model with improved d-s evidence fusion for bearing fault diagnosis,’’ *Sensors*, vol. 17, no. 8, pp. 1729–1738, Aug. 2017.

FENG-PING AN received the B.S. degree from the School of Economic and Management, Hefei University, Hefei, China, in 2008, the M.S. degree from the School of Economics and Management, Hebei University of Engineering, Handan, China, in 2011, and the Ph.D. degree from the School of Computer and Communication Engineering, Beijing University of Science and Technology. He was with the Huaiyin Normal of University. At the same time, he was a Postdoctoral Researcher with the Beijing Institute of Technology. His research interests include image processing, deep learning, artificial intelligence, and pattern recognition.