Hindawi
Complexity
Volume 2019, Article ID 6943234, 17 pages
https://doi.org/10.1155/2019/6943234

Research Article

A Novel Method for Intelligent Fault Diagnosis of Bearing Based on Capsule Neural Network

Zhijian Wang[1,2], **Likang Zheng**[3], **Wenhua Du**[1], **Wenan Cai**[4], **Jie Zhou**[1], **Jingtai Wang**[1], **Xiaofeng Han**[1] and **Gaofeng He**[1]

[1] *School of Mechanical Engineering, North University of China, Taiyuan, Shanxi, 030051, China*
[2] *School of Mechanical Engineering, Xi'an Jiaotong University, Shanxi, 030619, China*
[3] *School of Energy and Power Engineering, North University of China, Taiyuan, Shanxi, 030051, China*
[4] *School of Mechanical Engineering, Jinzhong University, Jinzhong, Shanxi, 030600, China*

Correspondence should be addressed to Wenhua Du; dwh@nuc.edu.cn and Wenan Cai; caiwenan0008@link.tyut.edu.cn

Received 8 April 2019; Revised 17 May 2019; Accepted 3 June 2019; Published 20 June 2019

Academic Editor: Diego R. Amancio

In the era of big data, data-driven methods mainly based on deep learning have been widely used in the field of intelligent fault diagnosis. Traditional neural networks tend to be more subjective when classifying fault time-frequency graphs, such as pooling layer, and ignore the location relationship of features. The newly proposed neural network named capsules network takes into account the size and location of the image. Inspired by this, capsules network combined with the Xception module (XCN) is applied in intelligent fault diagnosis, so as to improve the classification accuracy of intelligent fault diagnosis. Firstly, the fault time-frequency graphs are obtained by wavelet time-frequency analysis. Then the time-frequency graphs data which are adjusted the pixel size are input into XCN for training. In order to accelerate the learning rate, the parameters which have bigger change are punished by cost function in the process of training. After the operation of dynamic routing, the length of the capsule is used to classify the types of faults and get the classification of loss. Then the longest capsule is used to reconstruct fault time-frequency graphs which are used to measure the reconstruction of loss. In order to determine the convergence condition, the three losses are combined through the weight coefficient. Finally, the proposed model and the traditional methods are, respectively, trained and tested under laboratory conditions and actual wind turbine gearbox conditions to verify the classification ability and reliable ability.

1. Introduction

The rolling bearing is the most commonly used part in mechanical equipment. In the working process, the bearing may be damaged due to improper assembly, poor lubrication, water, and foreign body invasion, corrosion, overload, etc. [1]. Due to the processing technology, working environment, and other reasons, the fault signal is nonlinear and nonstationary, which makes the dynamic mutation of the fault signal unable to be detected effectively. So, it is difficult to identify the fault type of bearing accurately and stably. Compared with other machine parts, the rolling bearing works badly, which causes the probability of failure to be high and the unpredictability strong. Therefore, the fault diagnosis of rolling bearings is of great significance to ensure the safety of equipment, personal property, and maintenance cost [2].

In recent years, methods based on signal processing [3] or deep learning [4, 5] are widely used to solve practical engineering problems [6–9]. Moreover, in the field of compound fault diagnosis of rotating machinery, with the continuous exploration of many researchers, novel intelligent fault diagnosis methods emerge in an endless flow. For example, the methods based on the entropy [10, 11] of signal processing include maximum kurtosis spectral entropy deconvolution (MKSED) [12], multipoint optimal minimum entropy deconvolution adjusted (MOMEDA) [13], modified multiscale symbolic dynamic entropy (MMSDE) [14], and minimum entropy deconvolution [15]. In addition, there are other ways, such as improved ensemble local mean decomposition (IELMD) [16, 17], kernel regression residual [18], and modified variable modal decomposition (MVMD) [19, 20]. Methods based on big data and machine learning or

deep learning include support vector machine (SVM) [21, 22], extreme learning machine (ELM) [23], kernel extreme learning machine (KELM) [24], deep belief network (DBN) [25], and convolutional neural network (CNN) [26, 27]. In general, these methods can solve most classification problems well. But for composite fault diagnosis, their fault test accuracy rate is not too high. Moreover, this kind of algorithm always fails when there is not enough data to meet the convergence condition or causes overfitting phenomenon, which will lead to low test accuracy [28].

For example, the traditional convolution neural network requires a lot of training samples to meet the convergence condition. Moreover, people may subjectively reduce the dimension of filter [29] on the pooling of convolution neural network layer, which can result in a substantial loss on the pooling layer information and even causes a phenomenon that the input has a small change but the output is hardly changed. However, as for time-frequency graphs, a very small change may be the different type of bearing fault type or the large change of fault size. To summarize, the traditional convolutional neural network is difficult to achieve a high fault test accuracy.

Based on these disadvantages of traditional convolutional neural network, the capsule neural network (CapsNet) architecture was proposed by Hinton and his assistants in November 2017 [30], which can retain the exact position, inclination, size, and other parameters of the feature in the time-frequency graphs when training the deep learning method, so as to make the slight changes in the input also bring about slight changes in the output. In the famous handwritten digital image data set (Minst), CapsNet has reached the most advanced performance of the current deep learning algorithms. CapsNeth architecture is made up of capsules rather than neurons. A capsule is a small group of neurons that can learn to examine a particular object in an area of an image. Its output is a vector, the length of each vector represents the estimated probability of the existence of the feature, and its direction records the object of attitude parameters, such as accurate position, inclination, and size. If the feature changes slightly, the capsule will output a vector with the same length but slightly different direction, which is helpful to improve the test accuracy of bearing fault diagnosis.

The input of the deep learning algorithm in the fault diagnosis is fault time-frequency graphs and the two common time-frequency analysis methods are short-time Fourier time-frequency analysis and wavelet time-frequency analysis. Short-time Fourier transform (STFT) used to play a dominant role in the field of signals and is an indispensable analysis method [31]. However, due to its own limitations, it is unable to deal with the nonstationary signals in real life, and there is a contradiction between noise suppression and signal protection in the process of signal denoising. After the idea of wavelet transform, the wavelet transform replaces the position of Fourier transform in signal processing. Firstly, wavelet has very good time-frequency characteristics and can decompose many different frequency signals in nonstationary signals into nonoverlapping frequency bands, which can solve the problems encountered in signal filtering, signal-noise separation, and feature extraction well [32]. Secondly, due to the time-frequency characteristic of localization, the choice of wavelet basis is flexible and the calculation speed is very fast, which makes wavelet transform a powerful tool for signal denoising. Wavelet denoising can effectively remove noise and retain the original signal, thus improving the signal-to-noise ratio of the signal. Therefore, the continuous wavelet transform can effectively separate the effective part of the signal from the noise, greatly improve the feature extraction performance of fault diagnosis [33], and finally improve the fault recognition rate.

In addition, improving the network structure can improve the learning ability and reliable ability of the neural network. For example, using the Inception of modules and convolution in GoogLeNet can help neural network in different areas to capture more target-oriented characteristic, accelerate the calculation speed, and increase the depth of the neural network [34, 35]. Besides, Xception module is the extreme version of Inception [36]; Xception module completely decoupled across the channel correlation and spatial correlation and has achieved the classification accuracy of 94.5% in the classification of ImageNet database [37].

In terms of the convergence condition of the deep learning algorithm, most of them only consider the classification loss as the only index of convergence and do not consider the influence on the model when the parameters change a lot or the reconstruction loss, which may make the model difficult to converge or require a lot of time to converge [38]. However, most samples are collected under ideal working conditions in the laboratory, which may lead to the contingency when verifying the feasibility of the deep learning method. In other words, this deep learning method can only diagnose the gearbox under the specific working conditions [39, 40], which means that the reliably is very poor.

Based on the above, a novel intelligent fault diagnosis method of capsules network combined with the Xception module was proposed and the weight coefficient of loss was taken into account, in order to improve the convergence speed of the neural network classification, robustness, and learning ability. In order to verify XCN model of classification ability and reliable ability, the ideal laboratory condition of samples and the actual work condition of samples were chosen to train and test, respectively, and compare with other deep learning methods.

2. The Basic Theory of the Model

2.1. Wavelet Transform and Time-Frequency Transform. Compared with the short-time Fourier transform (STFT) method, the wavelet basis of continuous wavelet transform is no longer a trig function of infinite length, but a wavelet basis function of finite length that will decay. The wavelet basis can be stretched, which solves the problem that the time resolution and the frequency resolution cannot be both [41], so it can better and effectively extract the effective information in the bearing fault time domain signal.

Assuming the function $\varphi(t) \in L^2(R)$, its Fourier transform $\varphi(\omega)$ satisfies the condition

$$C_\varphi = \int_0^\infty \frac{|\varphi(\omega)|^2}{\omega} d\omega < \infty \quad (1)$$

where the function $\varphi(t)$ is called the parent wavelet or wavelet basis. After scaling and shifting the parent wavelet, a wavelet function cluster can be generated, whose expression is as follows:

$$\varphi_{a,b}(t) = \frac{1}{\sqrt{a}} \varphi\left(\frac{t-b}{a}\right) \quad (2)$$

where a is the scale factor and b is the translation factor. The scale factor a is used to scale the wavelet basis, while the scale factor b is used to change the position of the window on the time axis.

The continuous wavelet transform of the signal $x(t)$ is defined as

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \overline{\varphi\left(\frac{t-b}{a}\right)} dt \quad (3)$$

However, most of the fault signals of the rolling bearings are impulse fault signals, whose time-domain waveform is damped and freely attenuated vibration, while the time-domain waveform feature of Morlet wavelet is vibration attenuation from the central position to both sides. And their signals are similar [42, 43]. Therefore, the Morlet of continuous wavelet transform is selected as a wavelet basis, and its function is as follows:

$$\varphi(t) = e^{-t^2/2} e^{i\omega_0 t} \quad (4)$$

After determining the wavelet base and scale, the actual frequency sequence f is combined with the time series to draw the wavelet time-frequency graphs.

2.2. The Principle of Capsule Network. Capsule network is a novel type of neural network proposed by Hinton and his assistant in October 2017, which uses the module length of the activation vector of the capsule to describe the probability of the existence of the feature, and uses the direction of the activation vector of the capsule to represent the parameters of the corresponding instance [30].

Unlike previous neural networks which are composed of nerve neurons, capsule neural networks are composed of many capsules with specific meanings and directions. Activation of neuronal activity within the capsule represents various properties of a specific feature presented in the image. These properties can include many different types of instantiation parameters, such as posture (position, size, and direction), deformation, velocity, reflectivity, tone, and texture.

At the network level, the capsule neural network is composed of many layers. The lowest level capsules are called vector capsules, and they use only a small portion of the image as input, respectively. The small area is called the perceptual domain and it attempted to detect whether a particular pattern exists and how it is posed. At higher levels, capsules called routing capsules are used to detect larger and more complex objects.

The output of the capsule is a vector, the length of each vector represents the estimated probability of the existence of the object, and its direction records the object of attitude parameters. If the object changes slightly, the capsule will also output a vector with the same length but slightly different direction. So, the capsules are isotropic. For example, if the capsule neural network outputs an eight-dimensional capsule, its vector length represents the estimated probability of the existence of the object, and its direction in the eight-dimensional space represents its various parameters, such as the exact position of the object or the number of rotation angles. Then, when the object rotates by a certain angle or moves by a certain distance, it only changes the direction of the output vector, not its length, so it has little effect on the recognition rate of capsule neural network. Moreover, this phenomenon is not found in traditional neural networks, such as convolutional neural networks.

Convolution neural network, such as the traditional neural network, mostly through pooling mechanisms which choose the maximum value of the region or in a fixed area average, extracts main features to next layers, which makes the neural network of subjectivity much bigger. Therefore, the pooling operation may reduce the recognition rate of the neural network greatly. The capsule neural network proposed a very significant mechanism called the dynamic routing mechanism. In the capsule neural network, the output of the capsule is set as a vector, which makes it possible to use a powerful dynamic routing mechanism to ensure that the output of the capsule is sent to the appropriate parent node in the above layer. Initially, the output is routed to all possible parent nodes after the coupling sum is reduced by a factor of 1. For each possible parent node, the capsule calculates the prediction vector by multiplying its own output by a weight matrix. If the scalar product of this prediction vector and the output of a possible parent node are larger than others, there is top-down feedback, which has the effect of increasing the coupling coefficient of this parent node and reducing the coupling coefficient of other parent nodes. This increases the contribution of the capsule to that parent node and further increases the scalar product of the capsule prediction vector and the output of that parent node. So, this operation is more efficient than the primitive form of routing implemented through pooling, where all feature detectors in the next layer are ignored, except for the most active feature detectors in the local layer. In Hinton's paper, he demonstrated that [30]. Then, he imported the images in Minst into the capsule neural network with dynamic routing and the convolution neural network with pooling, respectively, finally finding the capsule neural network in the digital recognition accuracy compared to convolution neural network, and capsule neural network is significantly higher than the convolutional neural network on the highly overlapping digital image recognition. So dynamic routing mechanism is a very effective way.

Take a three-layer capsule neural network architecture for identifying digital images in Minst, which is shown in Figure 1. The architecture can be simply represented as consisting of only two convolutional layers and one fully connected layer. Conv1 is activated by 256, 9×9 convolution kernels, stride 1, and ReLU function. The layer converted the

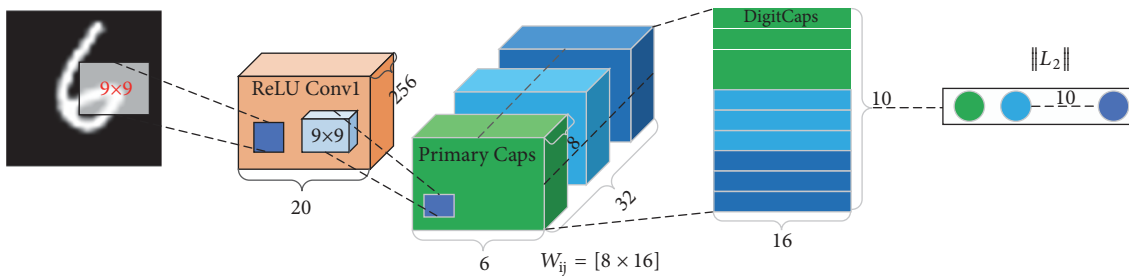

Figure 1: Three-layer capsule neural network.

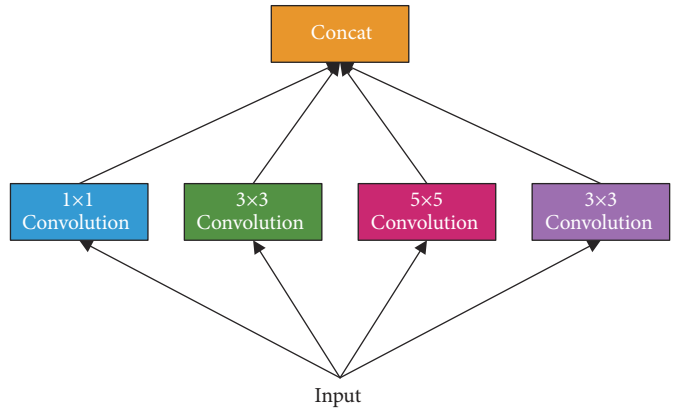

Figure 2: Inception module.

pixel intensity into a matrix. Then, the matrix is used as the activated part of the local feature detector input by the basic capsule. The length of the activation vector for each capsule in the DigitCaps layer represents the probability of the specific object and is used to calculate the classification loss.

W_{ij} is a weight matrix for each input in the Primary Capsule, and $\|L_2\|$ is the length of the longest capsule.

2.3. Xception Module. In order to obtain a deeper feature graph, the traditional neural network tends to increase the number of layers of convolution, which will bring too many parameters and make it difficult to train. More importantly, it increases the time and space consumption. Moreover, the convolution kernel of the same layer in the traditional convolutional neural network is single, but the correlation between different channels in the time-frequency graphs is not very much, which requires the use of different convolution kernel for different channels.

To solve these problems, Inception provides a different way to extract deeper feature maps. By using convolution kernels of different sizes in the same layer, different sizes of sensing fields can be obtained to improve the classification effect. Its architecture is shown in Figure 2.

An input feature graph is processed by the convolution kernel of 1×1, 3×3, and 5×5 at the same time, and the obtained features are combined, which can extract many kinds of features and obtain better results. However, there is a serious problem with this structure that the number of parameters is much larger than that of using a single convolution kernel, and such a huge amount of calculation will make the model inefficient. This was inspired by the 1×1 convolution kernel in the Network. In Network, 1×1 convolution kernel is added to Inception. Its architecture is shown in Figure 3.

To take an example which is shown in Figure 4, assuming that the dimension of the input feature graph is 256, the output dimension is also 256. There are two operations.

The first is that the 256-dimensional input goes through a convolution layer of 3×3×256 directly. And the output is a 256-dimensional feature graph, so the number of parameters is 256×3×3×256=589824.

The second is that firstly the 256-dimensional input passes through a convolution layer of 1×1×64, then through a convolution layer of 3×3×64, and finally through a convolution layer of 1×1×256. The output dimension is 256, but the number of parameters is 256×1×1×64+64×3×3×64+64×1×1×256=69632. That reduces the number of parameters for the first operation to one in nine.

For convolution, the convolution kernel can be viewed as a three-dimensional filter: channel dimension and spatial dimension (width and height of the feature graph). The conventional convolution operation is actually the joint mapping of channel correlation and spatial correlation. Inception module is based on the assumption that the channel correlation and spatial correlation between the channels in the convolutional layer can be decoupled. So, mapping the channel correlation and spatial correlation can get better results separately.

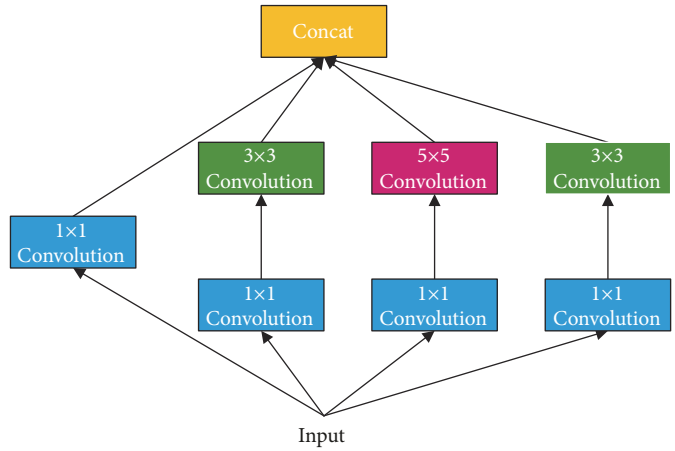

Figure 3: Inception module with 1×1 convolution.

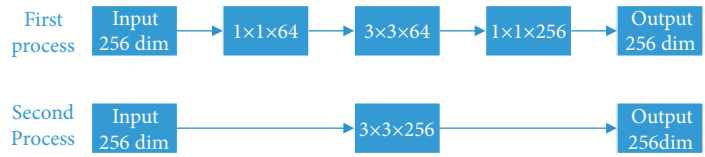

Figure 4: Comparison of the convolution process with and without 1×1 convolution.

Based on the Inception hypothesis, it can be found that when the number of channels increases, it is more reasonable to assume that cross-channel correlation and spatial correlation mapping are completely separated.

Based on the above findings and assumptions, the Xception module was proposed. Firstly, it uses 1x1 convolution map to map cross-channel correlation, and then there is an independent spatial convolution on each output channel of 1×1 convolution to map spatial correlation. The channel correlation convolution of 1×1 is carried out, and the number of convolution kernels with different types of subsequent convolving is the same as the number of output channels of 1×1 convolution. Various feature maps are output during the whole Xcepiton module process, and each output can be represented by the following function:

$$x_j^l = f\left(u_j^l\right) \tag{5}$$

$$u_j^l = \sum_i x_i^{l-1} \times w_{ij}^l + b_j^l \tag{6}$$

where x_j^l is the output of the convolution channel l of the convolution layer j, u_j^l is the net activation of the convolution channel l of the convolution layer j, w_{ij}^l is the weight coefficients between the layer i and the layer j, and b_j^l is the threshold offset term of the channel l of the convolution layer j.

Then, all outputs are connected together through a full connection operation; its function is as follows:

$$x^l = f\left(u^l\right) \tag{7}$$

$$u^l = w^l x^{l-1} + b^l \tag{8}$$

And the gradient descent method is used to reduce the error of the whole process; its function is as follows:

$$\nabla\varphi\left(u^l\right) = \frac{\partial\varphi}{\partial u^l} \tag{9}$$

where $\nabla\varphi(u^l)$ is the error $\varphi(u^l)$ with the changing of u^l.

Although the architecture of the Xception module is basically the same as Inception, the improvement of Xception module performance is not due to the increase of model capacity but due to the more efficient use of model parameters.

Because the parts of the image with a long distance do not matter much, different convolution kernels are adopted in the Xception module. Inspired by convolution decomposition, it is used for different convolution kernels. For example, the 7×7 convolution kernels are decomposed into two one-dimensional convolution kernels 1×7 and 7×1. And the 3×3 convolution kernel is same for two one-dimensional convolution kernels 1×3 and 3×1. Such operation can not only accelerate the calculation but also further increase the network depth, which is helpful to improve the learning ability of the neural network. This is because what is useful is the homology of the information, not the invariance, and the sorting of the information rather than the discarding of the information. In addition, it also routes every part of the input into neurons that know how to process it, which means finding the best path. However, pooling operation determined by human factors performs poorly in handling dynamic routing, so it is not good for the establishment of neural network.

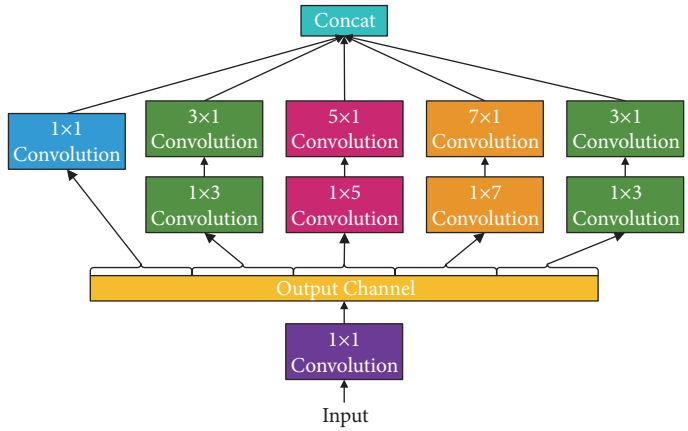

Figure 5: Xception module.

Building an improved Xcepiton module without pooling layer is shown in Figure 5.

3. Establishment of a Novel Capsule Neural Network with Xception Module (XCN)

3.1. Input and Dynamic Routing of Capsule Neural Network. When the time-frequency graphs of bearing fault are recognized by the capsule neural network, the selection of the structure parameters of the capsule neural network has a significant influence on the recognition results. Therefore, only when the appropriate parameters are selected can the classification and recognition performance of the capsule neural network for bearing faults be truly reflected. The input of the capsule neural network is time-frequency graphs whose size can be chosen as a variety of sizes. In this paper, for the sake of simplicity, the pixel size of all time-frequency graphs was chosen as 256×256. Then, multiple normal time-frequency graphs and failure time-frequency graphs were imported into the XCN model.

The length of the output vector of the capsule was used to represent the probability that the entity represented by the capsule exists in the current input. In addition, a nonlinear squashing function was used to ensure that the short vector is compressed to nearly 0 and the long vector is compressed to slightly less than 1:

$$v_j = \frac{\left\|s_j\right\|^2}{1+\left\|s_j\right\|^2}\frac{s_j}{\left\|s_j\right\|} \tag{10}$$

where v_j is the vector output of the capsule j and s_j is its total input.

In addition to the first layer of the capsule body, the total input s_j of the capsule is a weighted sum of the prediction vector $u_{j|i}$ of all the capsules from the next layer, which is generated by multiplying the output u_i of the following layer of capsules by the weight matrix W_{ij}.

$$s_j = \sum_i c_{ij}\hat{u}_{j|i} \tag{11}$$

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ik}\right)} \tag{12}$$

$$\hat{u}_{j|i} = W_{ij}u_i \tag{13}$$

where c_{ij} is the coupling coefficient determined by the iterative dynamic routing process. The sum of coupling coefficients between the capsule i and all capsules in the higher layer is 1, which is determined by routing softmax. The initial logic b_ij of routing softmax is the logarithmic prior probability; that is, the capsule i should be coupled with the capsule j. The logarithmic prior probability of the same time can be used as the discriminant learning of all other weights. They depend on the location and type of the two capsules, not on the current input image. Then, the initial coupling coefficient achieves iterative refinement by measuring the consistency between the current output v_j of each capsule j in the higher layer and the predicted $u_{j|i}$ of the capsule i. Consistency is the index dot product $a\_ij = v\_j \times u_{j|i}$. This consistency is considered to be a logarithmic likelihood ratio and is added to the initial logic b_ij before new values are calculated for all coupling coefficients between capsule i and higher level capsules.

In the convolutional capsule layers, each capsule outputs a vector local network to each type of capsule in the higher layer and uses a different transformation matrix for each part of the network and each type of capsule.

This operation, which Hinton calls dynamic routing between capsules, is used in the propagation of Primary caps layer to Digital caps layer. Figure 6 shows the dynamic routing mechanism.

Details of dynamic routing algorithm are shown in Procedure 1.

3.2. Output Vector Processing Method. The time-frequency graphs are imported into the XCN model. Finally, capsule vectors with many different meanings are obtained. The modules of all capsule vectors are calculated and the corresponding fault type of the capsule vector with the maximum module value is obtained.

(1) procedure Routing ($\hat{u}_{j|i}$,r,l)
(2) for all capsule i in layer l and capsule j in layer(l+1): $b_{ij} \longleftarrow 0$
(3) for r iterations do
(4) for all capsule i in layer l: $c_i \longleftarrow softmax(b_i)$
(5) for all capsule j in layer(l+1): $s_j \longleftarrow \sum_i c_{ij}\, \hat{u}_{j|i}$
(6) for all capsule j in layer(l+1): $v_j \longleftarrow squash(s_j)$
(7) for all capsule i in layer l and capsule j in layer(l+1): $b_{ij} \longleftarrow b_{ij} + \hat{u}_{j|i}.v_j$
return v_j

Procedure 1: Routing algorithm.

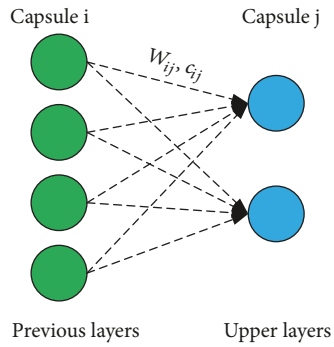

Figure 6: Dynamic routing mechanism.

In order to know the fault identification rate of XCN model intuitively, the specific capsules are generated fault time-frequency graphs through reconstruction. In the process of training, all the vectors in the Digital Caps layer except the correct capsule vector are shielded; that is, only the correct capsule is kept and the other capsules are set to zero. Then, the Digital Caps layer is used to reconstruct the input graph. The output of the digital capsule is fed into a decoder composed of three inverse routing iterations which are formed of the full connection layer, and the linear unit activation function is added in each layer of the full connection layer to ensure that the feature graph of the output of each layer is positive. Then, the original fault time-frequency graph and the reconstructed fault time-frequency graph are fused into a 256×256 target graph. The reconstructed flow chart is shown in Figure 7.

In order to measure the difference between reconstructed graph and truth graph in texture details and structure and minimize the loss of texture details and structure information of reconstructed graph, in this paper, the reconstruction loss function of XCN model uses L_2 norm to measure the distance between reconstructed graph and truth graph. The formula is

$$l_\gamma = \frac{d\left(G\left(Z\right) - Z\right)}{Z} = \frac{\left\|G\left(Z\right) - Z\right\|^2}{Z} \tag{14}$$

where Z represents the input original graph, $G(Z)$ represents the reconstructed graph, and l_γ represents the reconstruction loss.

3.3. Adjustment of Capsule Neural Network. Cross entropy measures the difference information between two probability distributions by describing the distance between them. The method of capsule neural network to solve the multiclassification problem is to set output capsules n, where n is the number of categories, and each capsule represents a different category. For each sample, an n-dimensional array formed of the length of each output capsule can be obtained by the capsule neural network. Each dimension in the array corresponds to a category. Ideally, if a sample belongs to the category k, the output value of the output node corresponding to that category should be 1, while the output value of all other nodes should be 0.

Take an example of identifying handwritten Numbers in Minst, which fall into 10 categories, from 0 to 9. When the number 1 is recognized, the closer the output 10-dimensional array of the capsule neural network [0, 1, 0, 0, 0, 0, 0, 0, 0, 0], the better. The ideal probability distribution on handwritten number classification is defined as $p(x)$ and the probability distribution of the output of the capsule neural network as $q(x)$. Then the cross entropy of this sample classification is

$$H\left(p, q\right) = -\sum p\left(x\right) \times \log_{10} q\left(x\right) \tag{15}$$

The cross entropy represents the uncertainty of the random variable or the whole system. The higher the entropy is, the greater the uncertainty of the random variable or the whole system will be. However, cross entropy describes the distance between two probability distributions. While the output length of the capsule neural network is not necessarily a probability distribution nor is necessarily between 0 and 1, softmax regression is a very useful way to transform the results of the forward propagation of the capsule neural network into a probability distribution.

Assuming the output of the neural network is $y_1, y_2, y_3, \cdots, y_n$, the output after softmax regression processing is

$$softmax\left(y\right)_i = y_i' = \frac{e^{yi}}{\sum_{j=1}^{n} e^{yj}} \tag{16}$$

Softmax transformed the output of the capsule neural network into a probability distribution. In this way, the output of the capsule neural network is also turned into a probability distribution, and the distance between the predicted probability distribution and the actual answer probability distribution can be calculated by cross entropy.

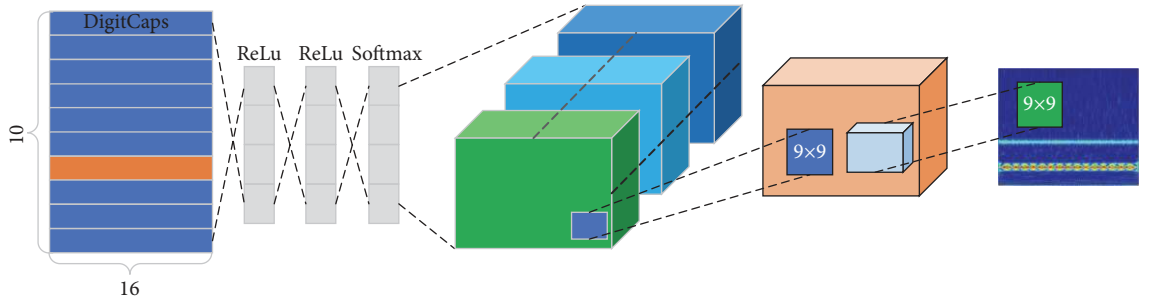

Figure 7: Reconstruction mechanism.

Then the cross entropy loss function of the capsule neural network is

$$H(p, y) = -\sum_{i}^{n} p(x_i) \times \log_{10} y_i \tag{17}$$

where x_i is its correct output, $p(x_i)$ is its probability, and y_i is the actual probability.

In the neural network, the minimal fluctuation of some parameters of weights will often lead to the change of the value of the loss function, which will lead to the overfitting of the model or the long convergence time, which will affect the prediction performance. In order to improve this phenomenon, the L_2 norm is introduced as the punishment for the parameters whose weight changes greatly each time the structure of the capsule neural network is adjusted. L_2norm refers to the sum of squares of the weight of ownership divided by the number of samples. The expression is as follows:

$$\|\theta\|_2 = \sqrt{\sum_{k=1}^{k} \sum_{l=1}^{l} \theta_{kl}^2} \tag{18}$$

where θ_{kl} is the weight of the parameter l in the layer k.

Then, combine the loss function, penalty function of the XCN model, and the reconstruction loss in the reconstruction through the regularization term coefficient α and β, so that when the gradient of the adjustment function L reaches the minimum value, it can be considered that the capsule neural network has been convergent. The expression of the adjustment function L of the capsule neural network is as follows:

$$L = H(p, y) + \alpha \|\theta\|_2^2 + \beta l_\gamma \tag{19}$$

where L is the adjustment function of the capsule neural network, $H(p, y)$ is the loss function, α and β are the weight coefficients, $\|\theta\|_2^2$ is the square of the penalty function, and l_γ is the reconstruction loss.

In order to ensure the accuracy of the classification and avoid over-fitting during training, α and β should be appropriately selected. For example, the size of $H(p, y)$ and $\|\theta\|_2^2$ is between 0 and 1. If the loss function is set in a large proportion, it will lead to the nonconvergence of the training process of the capsule neural network. If the penalty function is set in a large proportion, it will lead to the low accuracy of classification. Therefore, the weight coefficient α must be set a series of values, and then test its performance and adjust through experiments. Similarly, the selection of the weight coefficient β also needs to go through the appropriate selection.

The flowchart of XCN model is shown in Figure 8.

3.4. Feasibility Discussion of XCN Model. When the XCN model is used to identify bearing fault signals, the time-domain signals collected by the sensor need to be converted into time-frequency graphs. In the process of fault time-frequency graphs processing, the signal noise reduction and feature extraction are carried out firstly. And then the time-frequency graphs with pixel adjustment are taken as the input of the capsule neural network. In this paper, a capsule neural network was established with a convolutional layer, a dynamic routing layer, a full connection layer, and a reconstructed decoder. In order to improve the depth and dimension of the neural network, the Xception module was added to the neural network to improve the fault classification accuracy. When using XCN model to identify time-frequency graphs, different parameter settings have a great impact on the identification results of the capsule neural network, including the number of iterations, batch size, size and number of convolution kernels, the number of layers of Xception module, the selection of convolution parameters, and the weight coefficient. In order to discuss the feasibility of the XCN model, the bearing failure data set provided by Case Western Reserve University was used to train and test the XCN model. Figure 9 shows the experimental station of Case Western Reserve University. All the algorithm program code was written in and run on a computer with CPU i7-4790K@4.00GHz, RAM 16.00G, GPU Nvidia Geforce GTX960, and operating system Win7.

The experiment was carried out with a 2 HP motor and fan end bearing data was all collected at 12,000 samples per second, and the acceleration data were measured near and away from the motor bearing. In this data set, all the data are from the bearings damaged by the test bench. In other words, the load of each bearing is the same (in this paper, test bearing data under no-load condition was selected), but the health condition of the bearing is different. For the acquisition of faulty bearings, three different kinds of damage were caused to the outer ring, inner ring, or balls of the bearing by electrical discharge machining and their failure diameter was selected for a class of test bearing data sets of

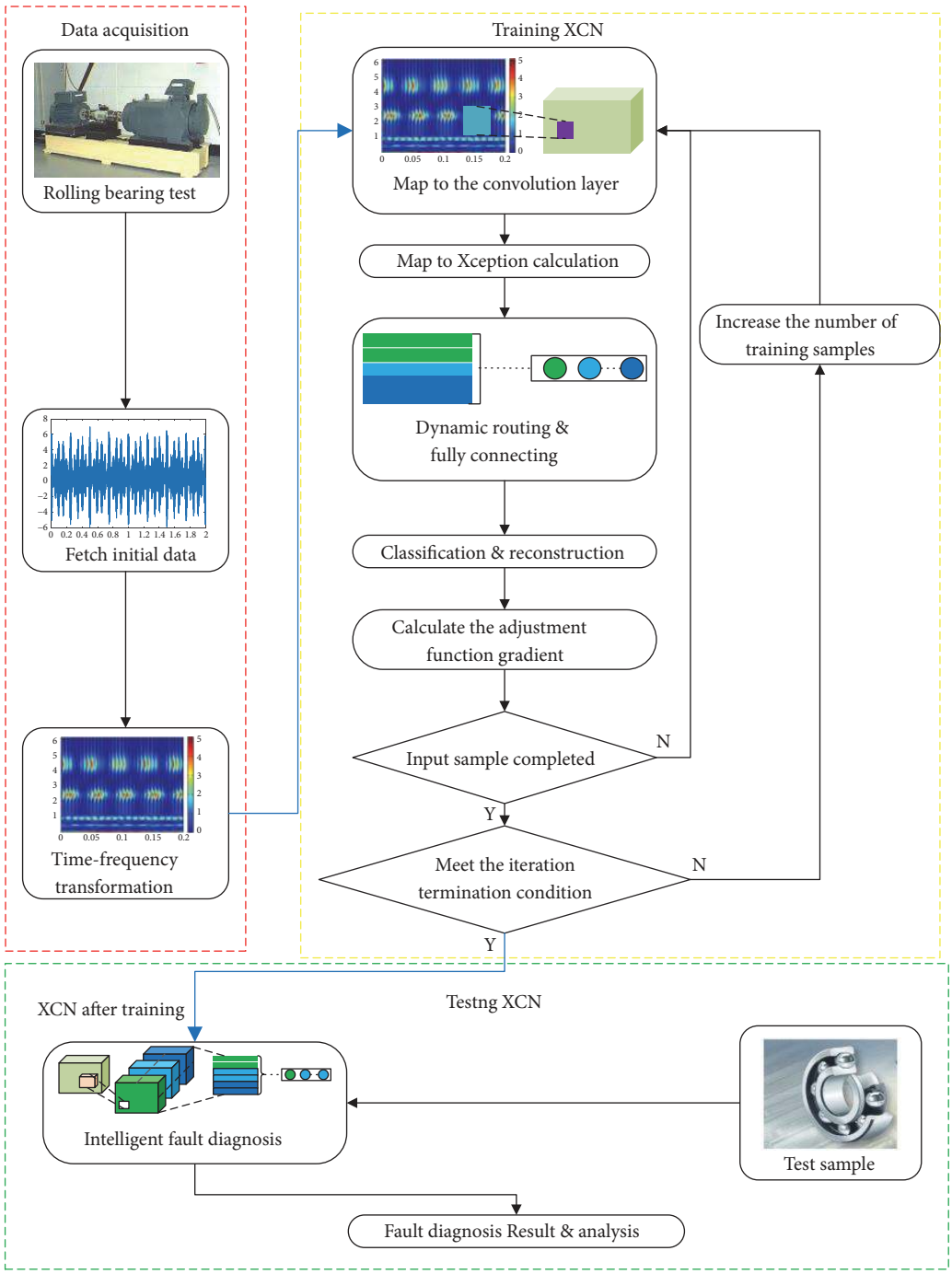

Figure 8: Intelligent fault diagnosis flow chart based on XCN.

0.007 inches. Therefore, there are 4 types of time-frequency graphs of bearings, including the healthy bearings.

In view of the rich information contained in the time-frequency graphs and the fact that the input of the capsule neural network must be a two-dimensional matrix, when studying the fault identification performance of the XCN model, time-frequency conversion processing of time-domain signals is required firstly. In this paper, the time-frequency conversion method is chosen as the continuous wavelet transform, and the wavelet basis is the Morlet wavelet.

One sample of four different types in the data set is taken out for continuous wavelet transform, respectively, and the results are shown in Figure 10.

In order to reduce the influence of noise, the zero mean normalization method was adopted, which not only ensures the distribution of the original signal but also eliminates the influence of dimension and converts the data into the same distribution. The formula is as follows:

$$x^* = \frac{x - \mu}{\sigma} \tag{20}$$

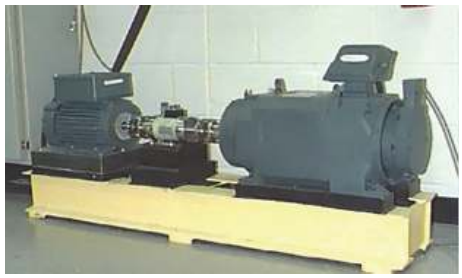

Figure 9: Experimental station of Case Western Reserve University.

where x is the input, x^* is the output of the time-frequency graph, and μ and σ are the mean and variance of x, respectively.

The input time-frequency graphs pixel size of bearing failure was selected as 256×256. The selection of capsule vector should be carried out first when the time-frequency graphs with appropriate pixel adjustment are imported into the capsule neural network for training. In this paper, the tonal distribution of time-frequency graphs and its change speed and direction are selected to transform the characteristics. Eighty percent of the samples from bearing data set provided by Case Western Reserve University were randomly selected for training to select parameters, and twenty percent of the samples were tested to verify the reliability of the selected parameters.

(1) Selection of Batch Size. Batch size refers to the process of training the neural network, in which a certain number of samples are randomly selected for batch training each time. Then the parameters of weight are adjusted once until all training samples are input. This process is called the completion of an iteration. If a larger batch size value is selected, the memory utilization can be improved through parallelization, so as to improve the running speed. In other words, the convergence speed will be faster, but the times of adjusting the weight will be less, thus reducing the classification accuracy. While choosing a smaller batch size can improve classification accuracy, it will lead to longer computing time. Therefore, in the case of limited memory capacity, when selecting the batch size, the requirements of classification accuracy and computing time should be weighed to ensure that the time cost can be reduced as much as possible under the premise of sufficient classification accuracy. The selection principle of batch size number must first meet the requirement that the number of training samples can be divided. Therefore, the selected batch sizes are 2, 4, 5, 8, 10, 20, 25, 30, and 50, respectively. When discussing the influence of batch size on classification results, in order to simplify the complexity of the discussion, the preliminary assumption of the number of iterations is 10, the preliminary assumption of the convolution kernel is 3×3, 5×5, and 7×7 whose number is 1, respectively, and the weight coefficient is 0.4 and 0.6, respectively. The experiment was repeated for ten times, and the average value of the classification results of ten times was taken as the final classification accuracy. Then the relationship between the classification accuracy and batch size was shown in Figure 11. As can be seen from Figure 11, when the batch size is within 10, the change of the batch size has little influence on the classification accuracy, while when the batch size is larger than 10, the classification accuracy obviously decreases with the increase of the batch size. As a result, the total sample batch size must meet the first condition that it can be divided exactly by the training sample in certain cases. Secondly, choosing the smaller batch size can increase the recognition rate of fault, which contributes to the failure of judgment. Such as the trial, when the batch size is selected as 5, 8, 10, or 20, the classification accuracy is the highest. Combined with the time cost factor, the batch size was selected as 20.

(2) Selection of Iteration Times. In essence, iteration is a process of continuous approximation and fitting. If the number of iterations is too small, the fitting effect will be not ideal. When the number of iterations increases to a certain degree, the fitting error will no longer decrease, but the time cost will increase with the increase of the number of iterations. Therefore, it is necessary to select an appropriate number of iterations, which can achieve a relatively low time cost under the condition of satisfying a certain recognition rate. When discussing the influence of iteration times on classification results, in order to simplify the complexity of discussion, the batch size is selected as 20 according to the previous section, the convolution kernel is preliminarily assumed to be 3×3, 5×5, and 7×7 whose number is 1, respectively, and the weight coefficients are 0.4 and 0.6, respectively. The experiment was repeated for five times, and the average of the five classification results was taken as the final classification accuracy. And the relationship between the classification accuracy and the number of iterations is shown in Figure 12. As can be seen from Figure 12, with the increase of iteration times, the classification accuracy increases gradually. When the number of iterations reached 6, the classification accuracy was more than 96%; when the number of iterations was more than 10, the classification accuracy was more than 98.5%; and with the increase of the number of iterations, the classification accuracy tended to be stable. Therefore, under the conditions of the number of samples and the size of graphs in this paper, the selection of 10 iterations can not only meet the requirements of high classification accuracy, but also reduce the time cost.

(3) Selection of the Size and Quantity of Convolution Kernel. The larger the size of the convolution kernel is, the larger the representable feature space of the network will be and the stronger the learning ability will be. However, there will be more parameters to be trained and the calculation will be more complex, which will lead to the phenomenon of overfitting easily. Meanwhile, the training time will be greatly increased. 1×1 convolution can increase across the channel correlation to improve the utilization rate of XCN model parameters, which can be used in the Xception module but does not make sense in the feature extraction. As the size of an even number of convolution kernels even symmetrically for zero padding operations, there is no guarantee that the input graph size and output characteristics graph size remain

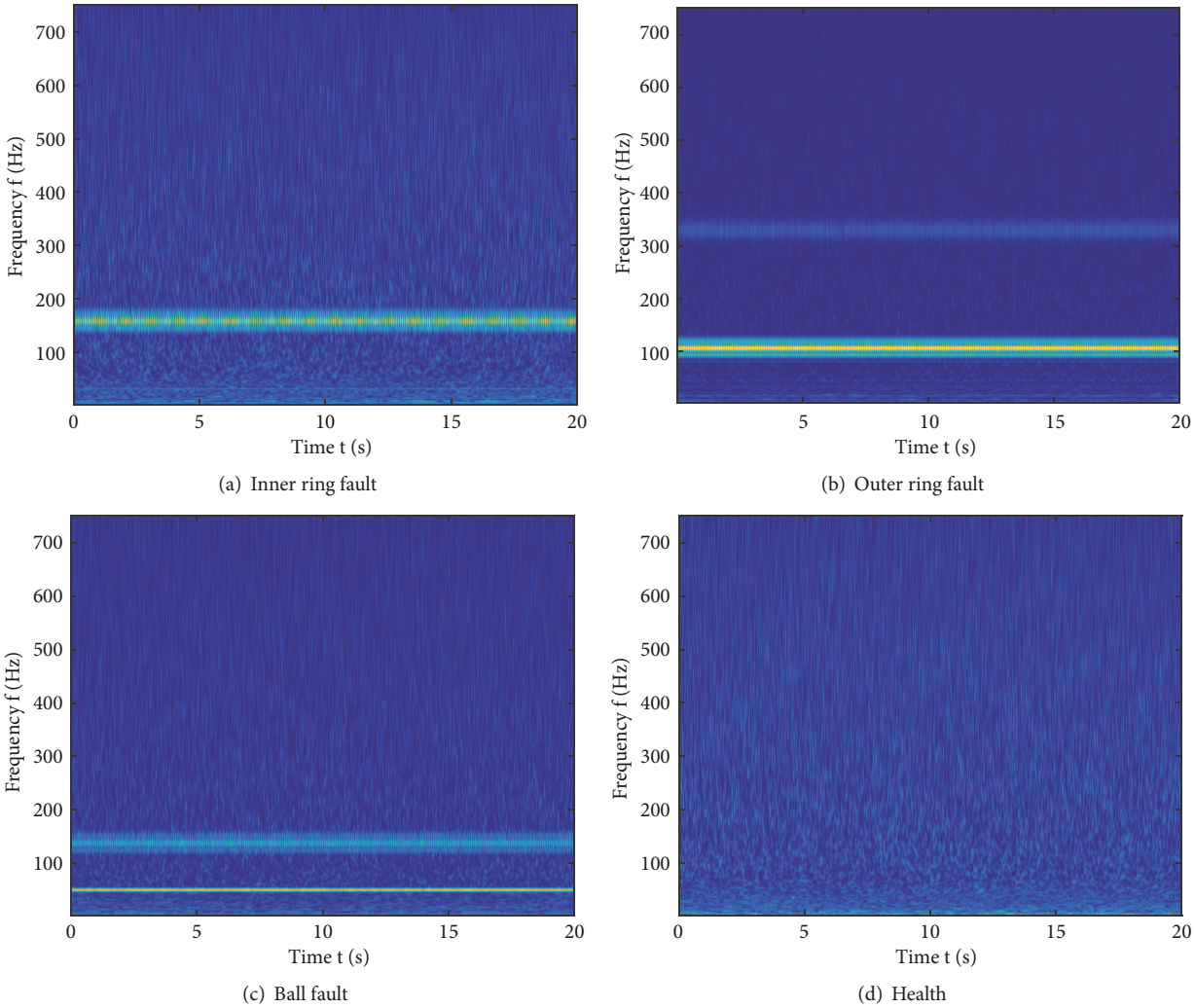

Figure 10: Signal time-frequency graphs of the experimental station of Case Western Reserve University.

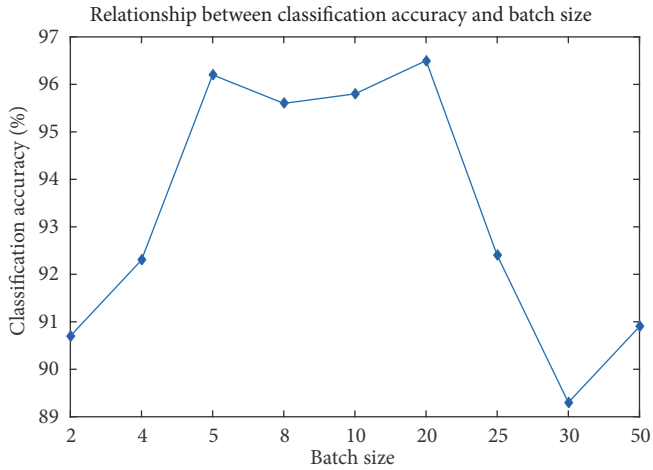

Figure 11: Relationship between classification accuracy and batch size.

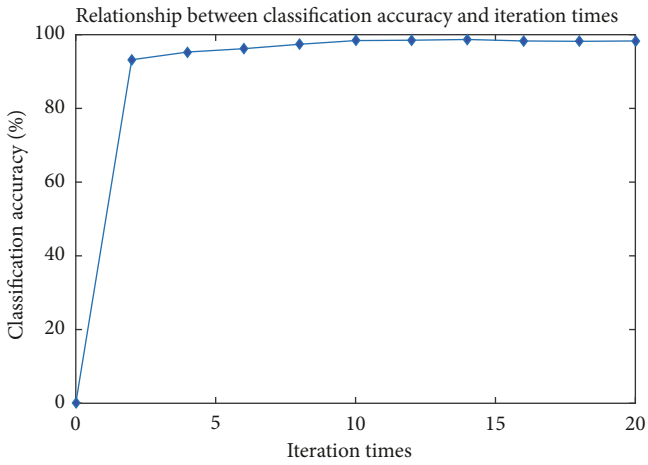

Figure 12: Relationship between classification accuracy and iteration times.

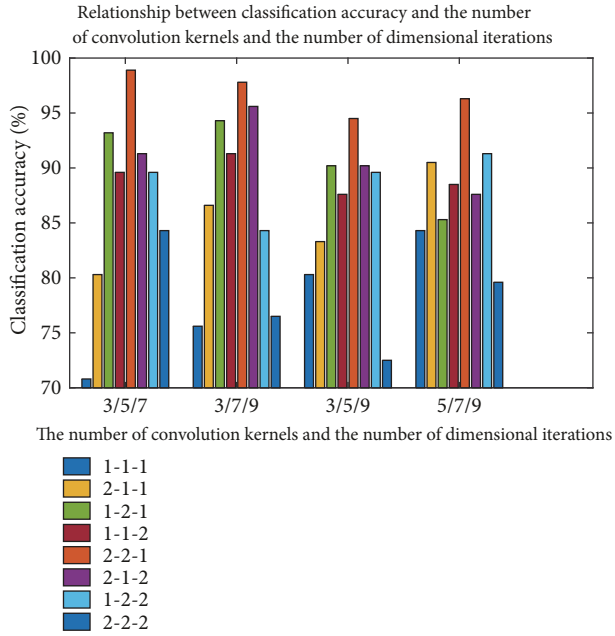

Figure 13: Relationship between classification accuracy and the number of convolution kernels and the number of dimensional iterations.

the same. In order to ensure that the size of the input and output characteristics graph is changeless, the convolution kernel size is chosen as the positive odd except 1. In the case of reaching the same receptive field, the smaller the size of the convolution kernel, the faster the operation. To simplify the complexity of the discussion, the maximum convolution kernel is selected as 9×9, the number of different convolution kernels is selected as 3, and the number of a single convolution kernels is no more than 2. The size and number of convolution kernels of each convolution layer constitute a set of parameters, in which the optimal combination will largely determine the performance of the neural network. The batch size is selected as 20 according to the previous section, the number of iterations is 10, and the weight coefficient is 0.4 and 0.6, respectively. The experiment was repeated for five times, and the average value of the five classification results was taken as the final classification accuracy. The relationship between the classification accuracy and the size and quantity of the convolution kernel is shown in Figure 13. In Figure 13, 3/5/7 represents the convolution kernel combination of 3×3, 5×5, and 7×7. And label 1-2-1 indicates that, in the convolution kernel combination, the number of the first convolution kernel is 1, the number of the second convolution kernel is 2, and the number of the third convolution kernel is 1. And the same is true for other labels. According to Figure 13, when the combination of convolution kernels is 3/5/7, overall classification accuracy is higher than other combination. In addition, in the combination of 3/5/7 and label is 2-2-1, the classification accuracy reaches 98.9%, which shows that the 3×3, 5×5, and 7×7 convolution kernels are better than other combinations to extract features. Furthermore, in the convolution kernels, combination of 3×3 and 5×5 is able to extract features more than 7 x 7 convolution kernels.

(4) Selection of Weight Coefficients α and β. The weight value of the weight coefficient α represents the size of the degree of punishment on the XCN model when the parameters are adjusted every time. And choosing the appropriate weight coefficient α can enhance the robustness of the XCN model. The weight of the weight coefficient β indicates the proportion of reconstruction loss. And choosing the appropriate weight coefficient β can enhance the classification accuracy of XCN model. The loss function $H(p, y)$, the penalty function $\|\theta\|_2^2$, and the reconstruction loss l_γ are all from 0 to 1, because it reduces the impact of dimensionality, so the range of the weight coefficients α and β is selected to be from 0 to 1. In order to simplify the complexity of the discussion, the weight coefficients α and β are reserved as one decimal place. According to the above section, the batch size is selected as 20, the number of iterations is selected as 10, the convolution kernel combination is selected as 3/5/7, and the quantity allocation is selected as 2-2-1. The experiment was repeated for five times, and the average value of the classification results of five times was taken as the final classification accuracy. Then the relationship between the classification accuracy and the weight coefficient is shown in Figure 14. It can be seen from Figure 14 that, with the increase of α, the classification accuracy tends to increase. With the increase of β, the classification accuracy first increases and then decreases. When the selection α is 0.7 and β is 0.6, the classification accuracy reaches 98.5%, while the classification accuracy of XCN is the highest.

In order to verify the reliability of the selected parameters, the parameters of the XCN model are selected according to the above conclusions as follows: batch size 20, iteration number 10, convolution kernel combination 3/5/7, quantity allocation 2-2-1, and weight coefficients 0.7 and 0.6, respectively. The test was repeated for ten times, and the test results are shown in Table 1.

As can be seen from Table 1, under the selected XCN model parameters, when the time-frequency graphs of three

Table 1: Training accuracy and test accuracy of each type of fault.

The fault types	Average accuracy of training sample set (%)	Average accuracy of test sample set (%)
Inner ring fault	99.8	99.2
Outer ring fault	100	99.7
Ball fault	98.5	96.3

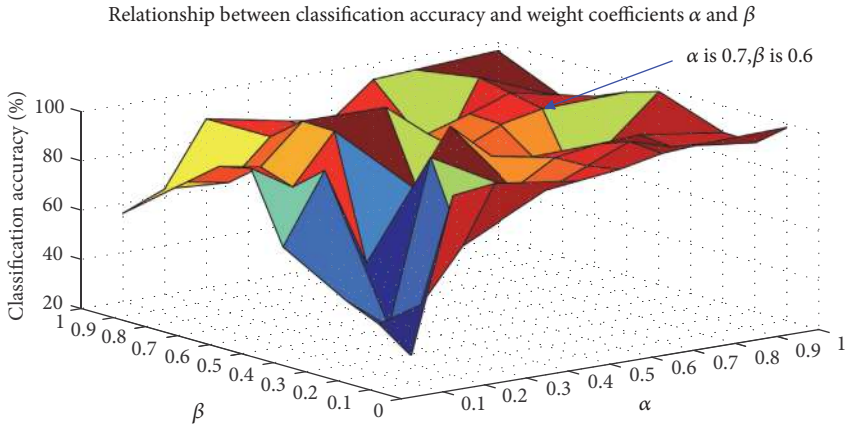

Figure 14: Relationship between classification accuracy and weight coefficients α and β.

kinds of fault signal were identified, the test accuracy and the training accuracy are similar. And the diagnosis accuracy is 96.3% in the diagnosis of outer ring fault, which may be related to time-frequency methods, samples, or other issues, and the rest were over 99%, which shows the reliability of parameter selection and XCN feasibility of the model.

3.5. Reliably of the XCN Model

Experiment 1 (the test bearing data set of Western Reserve University). In order to verify the superiority of the proposed method, the XCN model proposed in the previous section is compared with other deep learning algorithms for nearly three years: DWAE+ELM [23] which is based on deep wavelet autoencoder with extreme learning machine, CapsNet which is based on standard capsule neural network, MPE+ISVM+BT [22] which is based on multiscale permutation entropy and improved support vector machine based binary tree, AE+ES+CNN [26] which is based on an acoustic emission analysis-based bearing fault diagnosis invariant under fluctuations of rotational speeds using envelope spectrums and a convolutional neural network, and DBN [25] which is based on the standard deep belief network. Finally, the test accuracy of each algorithm is shown in Table 2 and Figure 15.

As can be seen from Table 2 and Figure 15, the fault diagnosis accuracy of these six models in the bearing data set tested by the Western Reserve University is generally higher than 90%, and the diagnostic accuracy of XCN model is significantly higher than the other models. The recognition rates of XCN model in outer ring fault, inner ring fault, and ball fault are 99.2%, 99.7%, and 96.3%, respectively. By comparing XCN model with CapsNet, it can be seen that the former has higher diagnostic accuracy than the latter in the three fault types, which proves that Xception can help improve the classification accuracy of CapsNet. By comparing XCN model with the other four models, except in the outer ring fault diagnosis, the classification accuracy of SVM is 100% higher than XCN, and the classification accuracy of XCN is all higher than other methods. Therefore, it can be preliminarily concluded that XCN has certain advantages over other methods in bearing fault diagnosis.

Experiment 2 (the test bearing data set of wind turbine gearbox in actual working conditions). For gearbox under the actual working conditions, there are always some mechanical working changes, such as speed and load. Due to the lack of reliable ability, these deep learning methods can only be effective under conditions similar to training data. In other words, these methods may fail if the working condition of the gearbox changes.

In order to discuss the reliability of the XCN model and compare it with other deep learning methods, they were imported to the samples obtained from the wind turbine gearbox under actual working conditions to train and test.Table 3 is the parameters of a wind turbine gearbox under actual working conditions.

Figure 16 is a real picture of three faults of a wind power gearbox. The sampling frequency is 5333Hz, and the gearbox transmission ratio is 113.4. The three fault types of wind turbine gearbox are shown in Figure 16.

Firstly, the collected time-domain signals were transformed into time-frequency graphs through continuous wavelet time-frequency transformation. Secondly, the time-frequency graphs were imported into the XCN model after adjusting the pixel size. Finally, the same input was imported

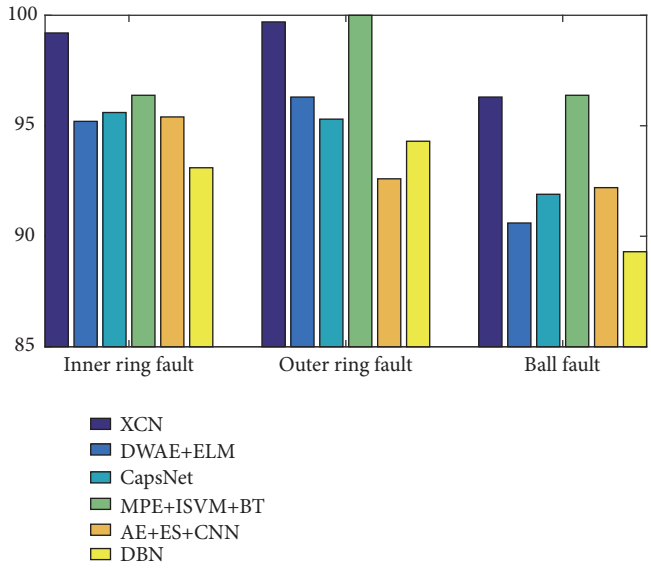

Figure 15: Average classification accuracy of different algorithms based on the test bearing data set of Western Reserve University.

Table 2: Average classification accuracy of different algorithms.

Average accuracy of test sample set (%)	XCN	DWAE+ELM	CapsNet	MPE+ISVM+BT	AE+ES+CNN	DBN
Inner ring fault	99.2	95.2	95.6	96.4	95.4	93.1
Outer ring fault	99.7	95.2	95.3	100	92.6	94.3
Ball fault	96.3	95.2	91.9	96.4	92.2	89.3

Table 3: Parameters of wind power gearbox under a certain actual working condition.

Parameters of a wind power gearbox (including three planetary wheels)		
Part name	Number of teeth	Rotate speed (r/min)
Wind turbines	/	92
The ring gear	92	0
The planets gear	37	-23.78
The sun gear	16	108
Intermediate gear	91	108
Intermediate gear shaft	20	491.4
Final big gear	96	491.4
Output gear shaft	26	1814.4

into other algorithms such as DWAE+ELM [23], CapsNet, MPE+ISVM+BT [22], AE+ES+CNN [26], and DBN [25] and tested the failure diagnosis performance of these methods.

In this experiment, sixty percent of the adjusted time-frequency graphs were used for algorithm training and forty percent were used for testing. The experiment was repeated five times.Table 4 and Figure 17 show the fault identification rates of the three algorithms.

It can be seen from Table 4 and Figure 17 that XCN model in these kinds of fault identification is obviously better than the other methods. In addition, XCN model in the recognition rate of outer ring fault and inner ring fault reached 98.7% and 97.2%, respectively, but the recognition rate of tapered roller fault is only 94.5%, which may be because the noise is too large and the time-frequency method is poor. By comparing the fault identification rates of these algorithms, it can be seen that the XCN model has great advantages in the fault identification of the wind turbine gearbox. Moreover, combined with the experiment in the previous section, it can be shown that the XCN model has a certain reliable ability rather than being limited to the working environment of the laboratory.

4. Conclusion

This paper takes the intelligent diagnosis of bearing faults as the research object, combines the capsule neural network which belongs to the category of deep learning with the Xception module, and then applies it to the fault identification of the wind turbine gearbox. Firstly, the time-domain

Table 4: Average classification accuracy of different algorithms.

Average accuracy of test sample set (%)	XCN	DWAE+ELM	CapsNet	MPE+ISVM+BT	AE+ES+CNN	DBN
Inner ring fault	97.2	88.6	95.6	87.6	90.4	89.1
Outer ring fault	98.7	91.3	94.3	90.3	92.6	91.3
Tapered roller fault	94.5	88.6	91.9	92.3	87.2	84.3

(a) Inner ring fault (b) Outer ring fault (c) Tapered roller fault

Figure 16: Fault types of wind turbine gearbox under actual working condition.

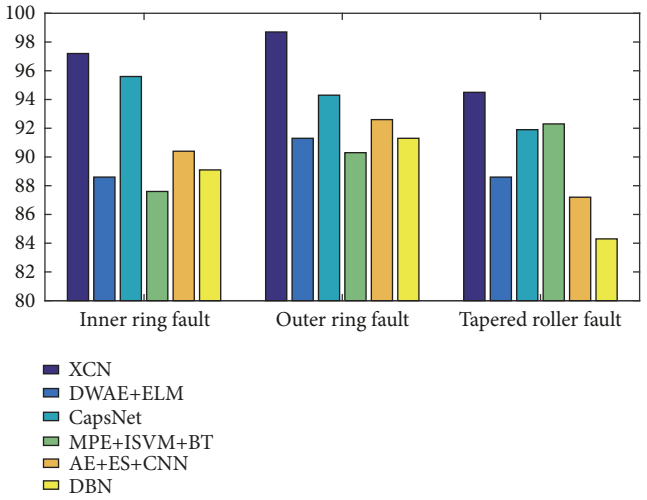

Figure 17: Average classification accuracy of different algorithms based on the test bearing data set of wind turbine gearbox in actual working conditions.

signals of faulty bearings provided by Case Western Reserve University were transformed into time-frequency graphs through a continuous wavelet transform of Morlet wavelet basis. Then, the pixel size of the time-frequency graphs was normalized and adjusted. Secondly, XCN model was trained to select better model parameters. Then, the trained XCN model was tested to study its classification effect. As can be seen from the above, the average classification accuracy is over 96%. Thirdly, the classification effect of XCN model on the fault of the wind turbine gearbox under actual working conditions was studied. And the average classification accuracy is over 94%, which shows the reliability of XCN model and provides a novel method for the fault diagnosis of the gearbox. Finally, other fault diagnosis algorithms based on big data were applied to the fault diagnosis of wind turbine gearbox. It can be seen from the experimental results that the classification accuracy of XCN model is significantly higher than other algorithms on the three fault types. By comparing CapsNet and CNN, it can be seen that, except for rolling fault diagnosis, CapsNet is obviously better than CNN in the other two types of fault diagnosis, which also indicates the inadequacy of such subjective operation as pooling. By comparing XCN and CapsNet, it can be seen that, among the three fault diagnoses, XCN is obviously better than CapsNet, which indicates that the Xception module is of great help to improve the classification accuracy. However, it can also be seen that, in terms of ball fault diagnosis and tapered roller fault, the classification accuracy of all algorithms is low, which may be related to the high noise caused by the harsh working environment of wind turbine gearbox or the weak ability of time-frequency conversion method to extract effective features. And it also points out the direction of future research. In the era of big data, there will be a great demand for intelligent fault diagnosis methods based on deep learning. With the development of deep learning, it will become the main method to solve the gearbox fault diagnosis. However, there are still many problems to be solved to improve the recognition rate and reliable ability of these algorithms. In future work, this topic is still a research focus.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Shanxi Provincial Natural Science Foundation of China under Grants 201801D221237, 201801D121186, and 201801D121187 and in part by the Science Foundation of the North University of China under Grant XJJ201802. At the same time, we would like to thank Professor Wang Junyuan from North University of China for his guidance in the experimental section.

References

[1] B. R. Randall and A. Jérôme, "Ball bearing diagnostics - a tutorial," *Mechanical Systems & Signal Processing*, vol. 25, no. 2, pp. 485–520, 2011.

[2] Y. G. Lei, J. Lin, M. J. Zuo, and Z. J. He, "Condition monitoring and fault diagnosis of planetary gearboxes," *Measurement*, vol. 48, no. 2, pp. 292–305, 2014.

[3] C. Shen, J. Yang, J. Tang, J. Liu, and H. Cao, "Parallel processing algorithm of temperature and noise error for micro-electro-mechanical system gyroscope based on variational mode decomposition and augmented nonlinear differentiator," *Review of Scientific Instruments*, vol. 89, no. 7, Article ID 076107, 2018.

[4] S. Chong, S. Rui, L. Jie et al., "Temperature drift modeling of MEMS gyroscope based on genetic-Elman neural network," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 897–905, 2016.

[5] H. Cao, Y. Zhang, C. Shen, Y. Liu, and X. Wang, "Temperature energy influence compensation for MEMS vibration gyroscope based on RBF NN-GA-KF method," *Shock and Vibration*, vol. 2018, Article ID 2830686, 10 pages, 2018.

[6] X. Guo, J. Tang, J. Li, C. Shen, and J. Liu, "Attitude measurement based on imaging ray tracking model and orthographic projection with iteration algorithm," *ISA Transactions*, 2019.

[7] X. Guo, J. Tang, J. Li, C. Wang, C. Shen, and J. Liu, "Determine turntable coordinate system considering its non-orthogonality," *Review of Scientific Instruments*, vol. 90, no. 3, Article ID 033704, 2019.

[8] C. Shen, X. Liu, H. Cao et al., "Brain-like navigation scheme based on MEMS-INS and place recognition," *Applied Sciences*, vol. 9, no. 8, p. 1708, 2019.

[9] H. Cao, Y. Zhang, Z. Han et al., "Pole-zero-temperature compensation circuit design and experiment for dual-mass mems gyroscope bandwidth expansion," *IEEE/ASME Transactions on Mechatronics*, vol. 24, 2019.

[10] Y. Li, X. Wang, Z. Liu, X. Liang, and S. Si, "The entropy algorithm and its variants in the fault diagnosis of rotating machinery: a review," *IEEE Access*, vol. 6, pp. 66723–66741, 2018.

[11] Y. Li, X. Wang, S. Si, and S. Huang, "Entropy based fault classification using the case western reserve university data: a benchmark study," *IEEE Transactions on Reliability*, pp. 1–14, 2019.

[12] Z. Wang, J. Zhou, J. Wang et al., "A novel fault diagnosis method of gearbox based on maximum kurtosis spectral entropy deconvolution," *IEEE Access*, vol. 7, pp. 29520–29532, 2019.

[13] Z. Wang, W. Du, J. Wang et al., "Research and application of improved adaptive MOMEDA fault diagnosis method," *Measurement*, vol. 140, pp. 63–75, 2019.

[14] Y. Li, Y. Yang, G. Li, M. Xu, and W. Huang, "A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection," *Mechanical Systems and Signal Processing*, vol. 91, pp. 295–312, 2017.

[15] S. Wang, J. Xiang, H. Tang, X. Liu, and Y. Zhong, "Minimum entropy deconvolution based on simulation-determined band pass filter to detect faults in axial piston pump bearings," *ISA Transactions*, vol. 88, pp. 186–198, 2019.

[16] Z. Wang, J. Wang, W. Cai et al., "Application of an improved ensemble local mean decomposition method for gearbox composite fault diagnosis," *Complexity*, vol. 2019, Article ID 1564243, 17 pages, 2019.

[17] Y. Gao, F. Villecco, M. Li, and W. Song, "Multi-scale Permutation entropy based on improved LMD and HMM for rolling bearing diagnosis," *Entropy*, vol. 19, no. 4, article 176, 2017.

[18] H. Liu and J. Xiang, "Kernel regression residual decomposition-based synchroextracting transform to detect faults in mechanical systems," *ISA Transactions*, vol. 87, pp. 251–263, 2019.

[19] Z. Wang, G. He, W. Du et al., "Application of parameter optimized variational mode decomposition method in fault diagnosis of gearbox," *IEEE Access*, vol. 7, pp. 44871–44882, 2019.

[20] Z. Wang, J. Wang, and W. Du, "Research on fault diagnosis of gearbox with improved variational mode decomposition," *Sensors*, vol. 10, p. 3510, 2018.

[21] Y. Li, Y. Yang, X. Wang, B. Liu, and X. Liang, "Early fault diagnosis of rolling bearings based on hierarchical symbol dynamic entropy and binary tree support vector machine," *Journal of Sound and Vibration*, vol. 428, pp. 72–86, 2018.

[22] Y. B. Li, M. Q. Xu, Y. Wei, and W. H. Huang, "A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree," *Measurement*, vol. 77, pp. 80–94, 2016.

[23] S. Haidong, J. Hongkai, L. Xingqiu, and W. Shuaipeng, "Intelligent fault diagnosis of rolling bearing using deep wavelet auto-encoder with extreme learning machine," *Knowledge-Based Systems*, vol. 140, pp. 1–14, 2018.

[24] K. Li, L. Su, J. Wu, H. Wang, and P. Chen, "A rolling bearing fault diagnosis method based on variational mode decomposition and an improved kernel extreme learning machine," *Applied Sciences (Switzerland)*, vol. 7, no. 10, Article ID 1004, 2017.

[25] Z. Shang, X. Liao, R. Geng, M. Gao, and X. Liu, "Fault diagnosis method of rolling bearing based on deep belief network," *Journal of Mechanical Science and Technology*, vol. 32, no. 11, pp. 5139–5145, 2018.

[26] D. K. Appana, A. Prosvirin, and J.-M. Kim, "Reliable fault diagnosis of bearings with varying rotational speeds using envelope spectrum and convolution neural networks," *Soft Computing*, vol. 22, no. 20, pp. 6719–6729, 2018.

[27] L. Jing, M. Zhao, P. Li, and X. Xu, "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox," *Measurement*, vol. 111, pp. 1–10, 2017.

[28] Z. Tao, H. Muzhou, and L. Chunhui, "Forecasting stock index with multi-objective optimization model based on optimized neural network architecture avoiding overfitting," *Computer Science and Information Systems*, vol. 15, no. 1, pp. 211–236, 2018.

[29] X. Liu, Y. Yang, and J. Zhang, "Resultant vibration signal model based fault diagnosis of a single stage planetary gear train with an incipient tooth crack on the sun gear," *Journal of Renewable Energy*, vol. 122, pp. 65–79, 2018.

[30] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems, NIPS 2017*, pp. 3857–3867, USA, December 2017.

[31] S. Meignen and D.-H. Pham, "Retrieval of the modes of multicomponent signals from downsampled short-time Fourier transform," *IEEE Transactions on Signal Processing*, vol. 66, no. 23, pp. 6204–6215, 2018.

[32] T. Abuhamdia, S. Taheri, and J. Burns, "Laplace wavelet transform theory and applications," *Journal of Vibration and Control*, vol. 24, no. 9, pp. 1600–1620, 2018.

[33] A. Cardinali and G. P. Nason, "Locally stationary wavelet packet processes: basis selection and model fitting," *Journal of Time Series Analysis*, vol. 38, no. 2, pp. 151–174, 2017.

[34] M. Haris, M. R. Widyanto, and H. Nobuhara, "Inception learning super-resolution," *Applied Optics*, vol. 56, no. 22, pp. 6043–6048, 2017.

[35] M. F. Haque and D. Kang, "Multi scale object detection based on single shot multibox detector with feature fusion and inception network," *The Journal of Korean Institute of Information Technology*, vol. 16, no. 10, pp. 93–100, 2018.

[36] M. Mahdianpari, B. Salehi, M. Rezaee, F. Mohammadimanesh, and Y. Zhang, "Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery," *Remote Sensing*, vol. 10, no. 7, p. 1119, 2018.

[37] L. Bai, Y. Zhao, and X. Huang, "A CNN accelerator on FPGA using depthwise separable convolution," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 10, pp. 1415–1419, 2018.

[38] R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro, "Tradeoffs between convergence speed and reconstruction accuracy in inverse problems," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1676–1690, 2018.

[39] Z. Chen, F. Han, L. Wu et al., "Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents," *Energy Conversion and Management*, vol. 178, pp. 250–264, 2018.

[40] R. Paul, R. Shandilya, and R. K. Sharma, "Comparative study and analysis of pulse rate measurement by vowel speech and EVM," in *Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering 2018 (ISMAC-CVB)*, pp. 137–146, 2019.

[41] H. Yang and H. Pan, "The adaptive analysis of shock signals on the basis of improved morlet wavelet clusters," *Shock and Vibration*, vol. 2018, Article ID 9892713, 13 pages, 2018.

[42] K. Deák, T. Mankovits, and I. Kocsis, "Optimal wavelet selection for the size estimation of manufacturing defects of tapered roller bearings with vibration measurement using Shannon Entropy Criteria," *Strojniski Vestnik: Journal of Mechanical Engineering*, vol. 63, no. 1, pp. 3–14, 2017.

[43] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573–3587, 2018.