

APPLIED RESEARCH

Fault Diagnosis Method for Bearing Based on Attention Mechanism and Multi-Scale Convolutional Neural Network

QIMIN SHEN¹ AND ZENGQIANG ZHANG²¹Jinzhong College of Information, Jinzhong 030800, China²Taiyuan Institute of Technology, Taiyuan 030008, China

Corresponding author: Qimin Shen (ty2023ty@163.com)

This work was supported by the Shanxi Province Higher Education Teaching Reform and Innovation Project J20221448.

ABSTRACT Convolutional neural networks (CNNs) serve as powerful feature extraction tools capable of effectively extracting information from complex environments, thus improving the accuracy of fault identification for bearing data. In this paper, we present a method for diagnosing bearing faults using an attention mechanism and a multi-scale convolutional neural network (MSCNN). Firstly, truncate and sample the rolling bearing data, and use continuous wavelet transform to generate corresponding time-frequency images, which will be used as inputs to the neural network. Next, the MSCNN, which includes efficient convolutional modules with residual structures, is utilized to extract features from the input data while maximizing the retention of valuable information. The extracted data then undergoes feature selection through the employment of an Efficient Convolutional Module (ECM) with channel attention. Finally, after being mapped through fully connected layers, the features are fed into a softmax layer for fault category prediction. In this study, the model results were tested and verified using the Case Western Reserve University (CWRU) dataset and the bearing dataset of Jiangnan University (JNU). A comparison was made with the LeNet model, ResNet model, LSTM model, and WDCNN model. The results showed that the classification accuracy of the ten types of bearing signals at the same speed can reach 100%, and the classification accuracy of the thirty types of bearing signals at different speeds can reach over 99.4%, significantly higher than the other models. The proposed method achieves the recognition of different fault states of rolling bearings under complex conditions, including multiple operating conditions and variable operating conditions. It is capable of extracting the global characteristic information of bearing faults, resulting in high diagnostic accuracy and good generalization ability. This method can provide reference for the diagnosis of rolling bearing faults under corresponding operating conditions.


INDEX TERMS CNN, bearing, attention mechanism, fault diagnosis.

I. INTRODUCTION

As one of the important components in mechanical equipment, bearings play a crucial role in the performance and reliability of the equipment. In reality, the failure rate of bearing components remains high due to the long continuous operation of mechanical equipment. Accurate and fast bearing fault diagnosis is one of crucial significance in preventing

equipment failures, reducing maintenance costs, and improving production efficiency.

Traditional methods for bearing fault diagnosis include analysis methods based on vibration signals, sound signals, and temperature signals, etc. These methods extract frequency domain features [1], time domain features [2], and time-frequency domain features [3] from the signals to determine the working condition and fault type of the bearing. However, traditional methods have certain limitations when applied in complex operating conditions and cannot meet the requirements of practical applications.

The associate editor coordinating the review of this manuscript and approving it for publication was Guillermo Valencia-Palomo .

With the systematization and complexity of data in the era of big data, data-driven intelligent diagnostic methods have been used for self-diagnosis and self-recognition of equipment faults. In terms of feature extraction, Zhang et al. [4], [5] proposed the use of wavelet transform for feature extraction of acquired signals and demonstrated the effectiveness of the method through experiments. He et al. [6] employed a feature extraction method based on cross-wavelet transform and variational Bayesian matrix decomposition, which can effectively localize different types of defects with high accuracy. In terms of diagnostic methods, researchers have proposed a fault diagnosis method based on Manhattan distance and voltage difference analysis [7], which can sensitively and reliably detect and isolate multiple faults. Additionally, machine learning and deep learning techniques have also emerged in this field.

In the application of machine learning, classic algorithms such as Artificial Neural Network(ANN) [8], K-Nearest Neighbors (KNN) [9], and Support Vector Machine (SVM) [10] have been cited and achieved certain effectiveness in the field of fault diagnosis. However, these algorithms inevitably encounter some issues when applied in practical fault diagnosis. First, there is the issue of feature input. These algorithms require a large amount of effective features as input to be effective, and extracting useful features from raw signals requires experienced professionals with rich knowledge, resulting in additional personnel and time costs. Second, there is the issue of training time. Traditional machine learning algorithms are usually based on serialized computing models, which can only process and infer one sample at a time. This means that training and predicting on large-scale datasets can be relatively slow. Lastly, the structure of machine learning algorithm models is relatively simple. While they may be effective for small datasets or low algorithm complexity, their classification performance is limited in complex environments and multiple operating conditions. Therefore, currently, deep learning is widely used for bearing fault diagnosis.

The deep learning techniques currently applied in bearing fault diagnosis include DNN [11], CNN [12], [13], and autoencoders [14], [15], among others. Compared to machine learning, deep learning itself has powerful and efficient feature extraction capabilities. Additionally, deep learning models have complex structures, which enable them to overcome various complex environments and operating conditions, resulting in accurate classification. Wang et al. [16] proposed converting one-dimensional vibration signals of bearings into two-dimensional grayscale images, which achieved good results after being processed by CNNs. Liang et al. [17] applied residual networks to the fault classification model, which deepened the convolutional network layers while avoiding the problem of vanishing gradients. Additionally, benefiting from the powerful time series information mining capability of LSTM networks [18], An et al. [19] proposed a rolling bearing fault diagnosis method based

on periodic sparse attention and LSTM. They compared this method with others and validated its effectiveness and superiority.

In the aforementioned studies, various deep learning network architectures were employed for fault diagnosis with the aim of attaining enhanced accuracy. However, for the actual extracted bearing fault signals, there is often a significant amount of noise mixed in, which can hinder the accuracy of the models' recognition. Moreover, as the number of layers and components of the CNNs increase, a large number of redundant parameters are introduced, leading to poor real-time performance due to lengthy training times.

To address these issues, Li et al. [20] introduced an attention mechanism into the diagnostic model, further enhancing fault-related features while reducing the weights of irrelevant parameters. Their proposed model further improved the accuracy of fault recognition. Zhang et al. [21] proposed a feature extractor based on a multi-scale attention mechanism, and the results showed that their strategy had good learning ability and diagnostic performance. Wu et al. [22] proposed a multi-source domain adaptation network with an attention mechanism, and the network model achieved outstanding fault diagnosis capability by leveraging its exceptional adaptability to samples. These research achievements demonstrate that the introduction of attention mechanisms can strengthen the adaptive capabilities of network models, enabling faster and more accurate identification of fault categories.

Inspired by the multi-scale attention mechanism [23], this article proposes a rolling bearing fault diagnosis method called MSCNN-ECM, which takes into account the interference of noise in actual samples. The method combines various models such as residual modules [17], [24], [25], multi-scale convolution [26], [27], [28], attention modules [29], [30], [31], and decomposed convolution modules to innovatively propose an efficient multi-scale convolution module for feature extraction of input signals, maximizing the retention of effective features from the original signal. Therefore, the main contributions and innovations of this paper are as follows:

a) In response to the issue of insufficient feature extraction, corresponding improvements are made to the model structure. In order to fully extract effective fault information, a three-branch structure is adopted to extract information at different scales, and large-scale convolutional kernels are used to enhance the effective receptive field of the network, thereby better capturing information between input data, while employing decomposed convolutions to reduce the computational load.

b) Improvements are made to address issues such as redundancy of multi-scale feature information and gradient vanishing. An attention mechanism is introduced after multi-scale convolutions to select feature information from different scales, and a residual structure is employed to ensure continuous gradient descent of the model.

II. RELATED WORK

This section will introduce the modules used in the fault diagnosis model, including CNN, ResNet, Decomposed Convolution, and Attention Mechanism.

A. CNN

CNNs are widely used deep learning models in computer vision and image processing tasks, typically for handling two-dimensional data. As we will preprocess one-dimensional time-varying bearing fault signals into two-dimensional representations using specific methods, CNNs can be migrated to the field of fault diagnosis. CNNs are named after the convolution operation, which is a special type of linear operation extensively applied in signal processing and image processing domains. In CNNs, the convolution operation is used to extract local features from input data, and by learning the parameters of the convolution kernels, specific patterns in the input signals or images can be captured. Convolution kernels of different sizes and shapes can be employed to extract features of various scales and types.

For two-dimensional images, the convolution operation can be understood as sliding a filter over the input image and performing element-wise multiplication, followed by summing all elements to obtain the value at the corresponding position in the output feature. The specific representation process is illustrated in Figure 1.

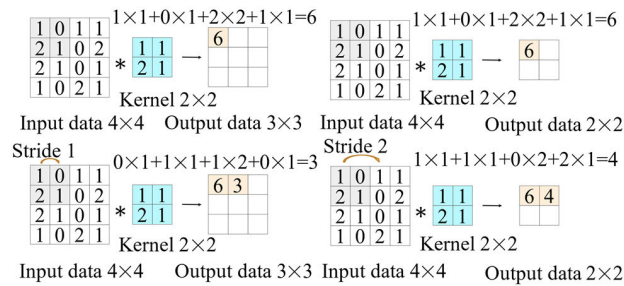


FIGURE 1. Convolution operation with different strides.

Stride is the step size at which the filter traverses. In each convolution operation, the filter slides horizontally to the right by a stride length. When the filter reaches the rightmost position, it starts sliding from the next row. Once the filter reaches the corresponding position in the bottom-right corner of the input image, the convolution operation is completed.

Figure 2 illustrates the padding scenario. By adding some extra pixels around the input image in the convolution operation, we can control the size of the output feature map to be the same as the input image, making network design and computation more convenient. Additionally, padding helps retain the information at the edges of the input signal. Without padding, the pixels at the edge of the signal would be involved in computation fewer times, leading to a loss of edge information. By applying padding, the participation of edge pixels

in the computation remains the same as that of central pixels, preserving edge information more effectively.

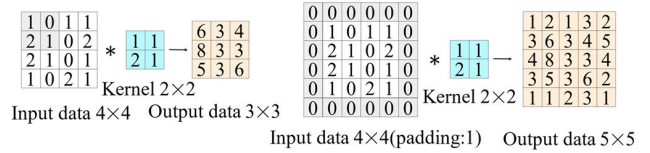


FIGURE 2. Padding operation in convolution.

Increasing the stride size reduces the output image dimensions, while increasing the padding layers can increase the output image dimensions. Suppose the input image has a height of H and width of W. After performing a convolution operation with a kernel size of k×k, the resulting output image has a height of h and width of w. This relationship can be expressed as follows:

$$h = \frac{H - k + 2p}{s} + 1 \tag{1}$$

$$w = \frac{W - k + 2p}{s} + 1 \tag{2}$$

In the equation, p represents the number of padding layers in the convolution process, and s represents the stride size of the filter. By using this formula, you can modify the relevant parameters to control the size of the output image.

B. RESIDUAL NETWORK STRUCTURE

In the early stages of researching neural network structures, it was widely believed that the depth of the network determined the accuracy and effectiveness of detection. Therefore, many researchers focused on increasing the network depth, and many classic models achieved good results by continuously deepening the network. However, it was later discovered that when the network becomes too deep, the performance actually starts to degrade, with issues such as vanishing or exploding gradients. To address this problem, the residual network structure was introduced.

The residual structure was initially proposed by He et al [32]. It allows the network to propagate larger gradient values back to the earlier layers during the backpropagation process. Additionally, ResNet has a flexible network structure, allowing for adjustments in the number of stacked residual blocks and the number of channels within each residual block. A basic residual block is shown in Figure 3.

In the above residual structure, the function F(x) represents the residual mapping, and it is defined as F(x) = y - x. BN represents batch normalization operation, ReLU is the activation function. During the process of residual learning, F(x) can be directly used as the optimization objective. By minimizing F(x) to approach zero, the optimal solution is obtained. This allows the model to maintain its best state even when the network is deepened. When the residual block takes x_n as input, the output can be represented as:

$$x_{n+1} = f(x_n + F(x_n, w_n)) \tag{3}$$

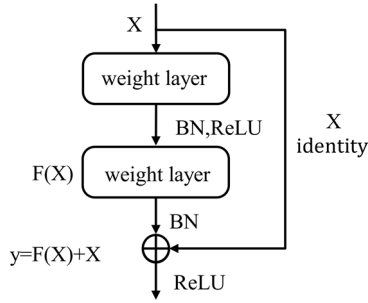


FIGURE 3. Basic residual block.

Here, $f(\cdot)$ is the activation function, w_n represents the corresponding weight parameters, and $F(\cdot)$ is the residual mapping function.

Compared to ordinary convolutional structures, the residual block adds an extra channel, allowing the input data to be directly passed to the output data. This enables the direct learning of residual values, simplifying the learning objective. This structure helps sustain gradient descent, which is beneficial for constructing deeper networks. The additional computational cost introduced by this structure can be ignored with GPU acceleration, but it greatly enhances the overall performance of the network.

C. DECOMPOSED CONVOLUTION

The principle of decomposed convolution [33] is to decompose a $k \times k$ standard 2D convolution kernel into a $k \times 1$ vertical convolution kernel and a $1 \times k$ horizontal convolution kernel. By reducing the dimensionality of the convolution kernel, it effectively reduces the floating point operations (FLOPs) required for convolution. Research has shown [34] that simply stacking small kernel convolutions does not significantly improve the effective receptive field [35] of the network. However, the introduction of decomposed convolution helps us mitigate the impact of large kernel convolutions [36] that lead to a large number of parameters. Decomposed convolution is implemented in two steps, as illustrated in Figure 4.

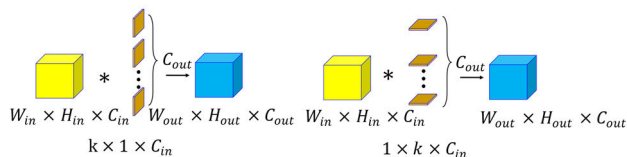


FIGURE 4. Principle of decomposed convolution.

For decomposed convolution, considering the bias, the corresponding FLOPs are as follows:

$$F_{FC} = 4kc_{in}c_{out}H_{out}W_{out} \quad (4)$$

where F_{FC} represents the FLOPs of decomposed convolution, k is the size of the convolution kernel, c_{in} is the number of input feature channels, c_{out} is the number of output feature

channels, H_{out} is the height of the convolution output feature, and W_{out} is the width of the convolution output feature.

For a regular convolution with a kernel size of $k \times k$, as shown in Figure 5, and considering the bias, the corresponding floating-point operations are:

$$F_{conv} = 2k^2c_{in}c_{out}H_{out}W_{out} \quad (5)$$

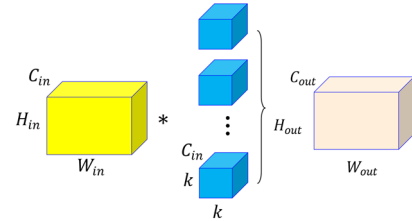


FIGURE 5. Standard convolution.

Decomposed convolution reduces the FLOPs of convolution by dimension reduction of the convolution kernel, changing the relationship between FLOPs and kernel size k from quadratic to linear. When the kernel size $k > 2$, the FLOPs of decomposed convolution will be smaller than the computational cost of standard convolution. Moreover, as the kernel size k and the number of output channels c_{out} increase, decomposed convolution saves more computational resources.

To combine residual modules and decomposed convolution, we propose a new efficient multi-scale convolution module called EMSCM (Efficient multi-scale convolution module). The structure of this new module is shown in Figure 6. This module ensures a large receptive field while reducing computational and parameter burdens, which meets the requirements of deep neural networks well.

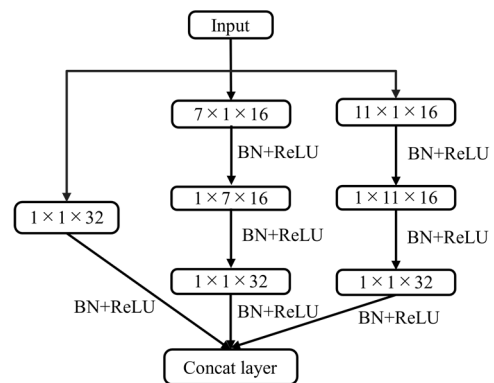


FIGURE 6. Efficient multi-scale convolution module.

D. ATTENTION MECHANISM

The introduction of attention mechanism [37] aims to enhance the model's focus and concentration on the input data, enabling the model to learn and process information that is more important for the current task. The reason for

incorporating attention modules in the fault diagnosis model is that there is a significant amount of noise in the original fault signals. By introducing attention, the model can pay more attention to the relevant fault signals in the original information and ignore most of the noise. Additionally, for relatively complex CNNs, the introduction of attention mechanism can simplify the model structure while maintaining performance, thereby reducing training parameters and time.

In this study, we employ the Efficient Channel Attention (ECA) module to enhance the weights of crucial features. Since our processed fault signals are two-dimensional, it is necessary to map the 2D signal into a three-dimensional matrix and then utilize the ECA attention mechanism module to obtain channel weight information. The ECA module utilizes a 1×1 convolutional layer after the global average pooling layer, eliminating the need for a fully connected layer while preserving the original feature map dimensions. This allows for effective utilization of inter-channel interaction information and captures local interaction relationships with surrounding channels. Moreover, the ECA module achieves excellent results with only a small number of parameters. The corresponding structure is illustrated in Figure 7.

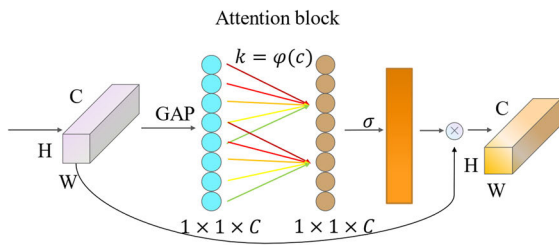


FIGURE 7. Efficient channel attention module.

ECA achieves cross-channel information interaction through one-dimensional convolution, which can reduce the model’s complexity while maintaining performance. The size of the convolutional kernel, denoted as k , is adapted using a function that allows layers with a larger number of channels to have more cross-channel interactions. The adaptive function is defined as follows:

$$k = \left\lfloor \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rfloor \quad (6)$$

where C represents the number of channels, γ and b are non-linear parameters, and $\gamma = 2, b=1$.

Additionally, the corresponding weight calculation formula is as follows:

$$\omega = \sigma(C_1 D_k(y)) \quad (7)$$

In this case, σ represents the sigmoid function. The output feature map is obtained by multiplying the weights with the corresponding elements of the original input feature map.

III. BEARING FAULT DIAGNOSIS MODEL

A. DATA PREPROCESSING METHODS

The vibration signals in the datasets from JNU and CWRU used in this study are in the form of one-dimensional time-varying non-stationary signals. In order to efficiently and quickly extract features using convolutional neural networks, sampling truncation and two-dimensional processing are required. It is worth noting that during the truncation process, it is best to ensure that each sample data contains the complete set of sampling points for one rotation of the bearing. The number of sample points required to cover a complete rotation can be obtained according to the following formula:

$$N = \frac{60}{n} f \quad (8)$$

where n is the rotation speed (r/min) and f is the sampling frequency (Hz).

Taking an electric motor speed of 1730 as an example, the number of sampling points contained in one rotation of the bearing is approximately 416, calculated as $60 \div 1730 \times 12000$. This study sets the number of sampling points in each sample data to 1024. In order to increase the number of sample data, overlapping sampling is adopted here: the first set of data $x_1(t)$ is sampled from 1 to 512, the second set of data $x_2(t)$ is sampled from 513 to 1024, and so on, as shown in Figure 8. After truncation, continuous wavelet transform is used to generate image data with time-frequency characteristics.

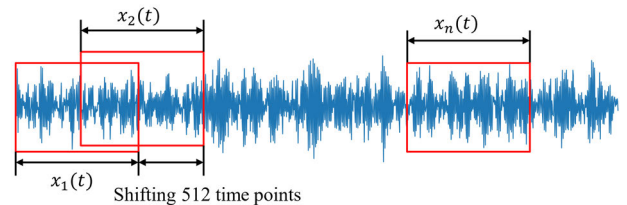


FIGURE 8. Signal expansion method used in this study.

Continuous wavelet transform is a signal processing method based on wavelet analysis, which can be used to analyze the time-frequency characteristics of non-stationary signals. In this study, the complex Gaussian function is chosen as the wavelet basis function. This function has the advantages of Gaussian sub-wavelets and exhibits superior resolution and concentration in the time-frequency representation.

B. MODEL DESIGN

In this paper, the proposed model is based on lightweight considerations and simplifies non-essential components to improve diagnostic efficiency while enhancing the feature extraction and resolution capabilities of the model. The entire model consists of efficient multi-scale convolution modules, pooling layers, normalization and activation layers, attention modules, fully connected layers, and a softmax layer. The specific model structure is shown in Figure 9.

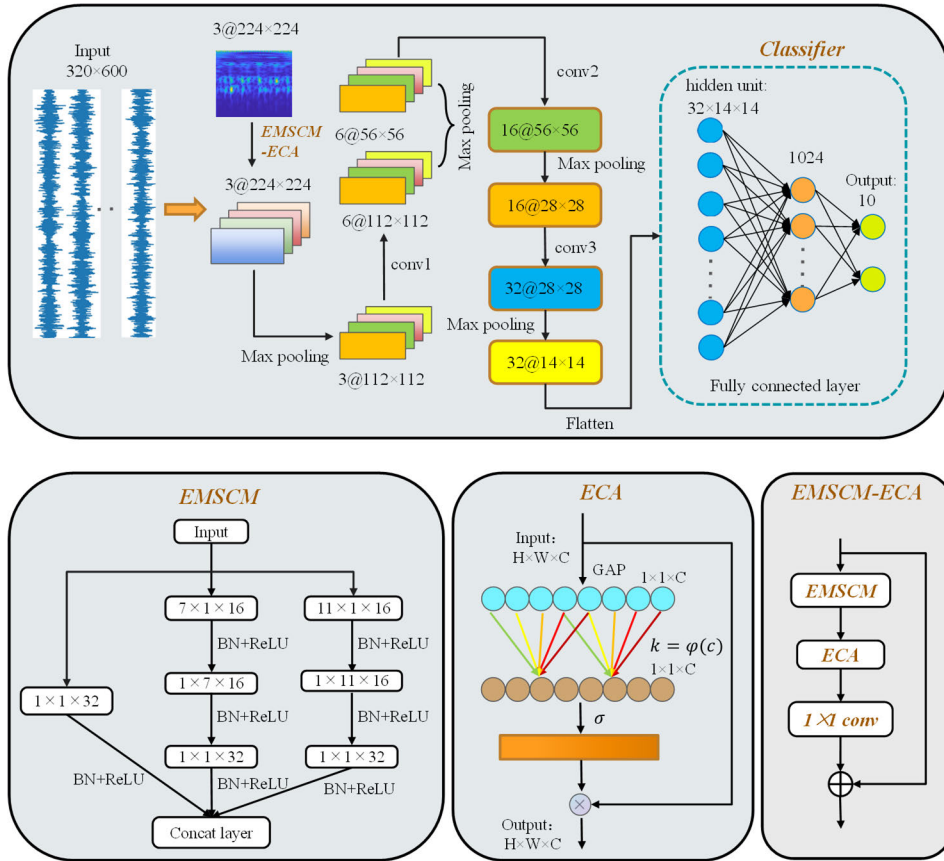


FIGURE 9. Adaptive CNN model based on attention mechanism.

The specific implementation process can be roughly divided into three steps: signal input, feature extraction and processing, classification output. Firstly, one-dimensional vibration data is transformed into a two-dimensional image with time-frequency features using continuous wavelet transform. The transformed two-dimensional image is then input into the model, and the time-frequency image features are extracted efficiently using a multi-scale convolution module, with attention mechanism used to filter the feature information. Secondly, the feature map of the feature map is further extracted using a regular convolution module, and normalization and activation function processing are performed between these two steps, followed by pooling operation to retain the main feature information. This process is repeated three times to obtain two-dimensional feature signals for classification. Finally, the two-dimensional signal is flattened and input into the fully connected layer for feature classification. The model parameter settings are shown in Table 1.

C. MODEL PARAMETERS AND LOSS FUNCTION

We set the sizes of two large-scale convolution kernels in the efficient multi-scale convolution module to 7 and 11, respectively, and reduce the parameter volume by decomposing the convolution. In the EMSCM module, 0/3/5 represents the padding of the three branches, and 1 × 1.7 × 7/11 × 11 represents the convolution kernel size of the corresponding branch

in the three-branch structure. After normalization and activation layers, the attention module is used to obtain the key channel information. This ensures that the width and height of the processed feature map remain consistent with the original input, while the number of channels remains consistent with the number of filters. Then, a max pooling layer is applied with a pooling window size of 3 × 3 and a stride of 2, reducing the size of the feature map by half in the width and height dimensions. The number of fully connected layers is set to 2, with 1024 hidden neurons and 10/4 output categories. “10” represents the number of classifications on the CWRU dataset, and “4” represents the number of classifications on the JNU bearing dataset. Specific parameter settings and module selections can be found in Table 1.

The chosen deep learning framework for this experiment is PyTorch. The computer setup includes a Linux operating system, an Intel Core i9-12900K CPU processor, and an NVIDIA GeForce RTX 3090 graphics card. The selected loss function is the cross-entropy loss function, expressed as follows:

$$L = - \sum_{i=1}^K y_i \lg \bar{y}_i \tag{9}$$

where L represents the error loss value, K represents the number of fault categories, y_i represents the true label for the i-th category, and \bar{y}_i represents the predicted probability value

TABLE 1. Specific model parameters.

| Number | Layer type | Kernel size | Stride | Padding | Channel | Output |
|--------|------------------|--------------------------------------|--------|---------|---------|----------------------------|
| 1 | EMSCM | $1 \times 1/7 \times 7/11 \times 11$ | 1 | 0/3/5 | 32 | $224 \times 224 \times 32$ |
| 2 | ECA | - | - | - | - | $224 \times 224 \times 32$ |
| 3 | Conv | 1×1 | 1 | 0 | 3 | $224 \times 224 \times 3$ |
| 4 | Max pooling1 | 3×3 | 2 | 0 | 3 | $112 \times 112 \times 3$ |
| 5 | Conv1 | 5×5 | 1 | 2 | 6 | $112 \times 112 \times 6$ |
| 6 | Max pooling2 | 3×3 | 2 | 0 | 6 | $56 \times 56 \times 6$ |
| 7 | Conv2 | 5×5 | 1 | 2 | 16 | $56 \times 56 \times 16$ |
| 8 | Max pooling3 | 3×3 | 2 | 0 | 16 | $28 \times 28 \times 16$ |
| 9 | Conv3 | 3×3 | 1 | 1 | 32 | $28 \times 28 \times 32$ |
| 10 | Max pooling4 | 3×3 | 2 | 0 | 32 | $14 \times 14 \times 32$ |
| 11 | Fully-connected1 | 1024 | - | - | 1 | 1024 |
| 12 | Fully-connected2 | 10/4 | - | - | 1 | 10/4 |

of the model indicating the likelihood of being the i -th fault category.

D. FAULT DIAGNOSIS PROCESS

The fault diagnosis process of the proposed model in this paper is shown in Figure 11, and the specific steps are as follows:

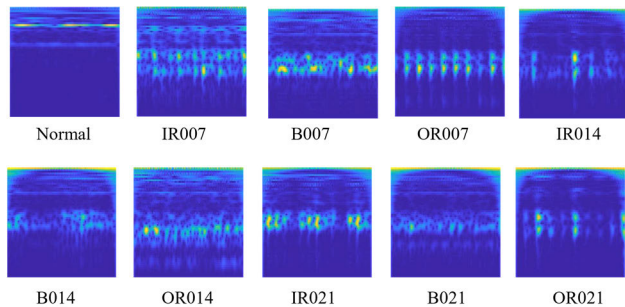


FIGURE 10. Time frequency images of each state under 1Hp load.

(1) Data information is extracted by truncating the data, and continuous wavelet transform is used to convert the one-dimensional data into two-dimensional form. Figure 10 shows the time-frequency diagrams for 10 states under a 1Hp load in the CWRU dataset.

(2) The dataset is divided according to a certain proportion.

(3) The network model is constructed and the model parameters are initialized before training. The batch size is set to 32 and the number of iterations is set to 50. The cross-entropy loss function and Adam optimizer are used. The initial learning rate is set to 0.001, and the learning rate is adjusted using the cosine annealing algorithm during actual training.

(4) The training set is input into the model to start training. The features are extracted through the convolutional module, and the extracted features are classified using the classifier.

(5) The model performance is validated using the validation set, and the model parameters are adjusted to achieve

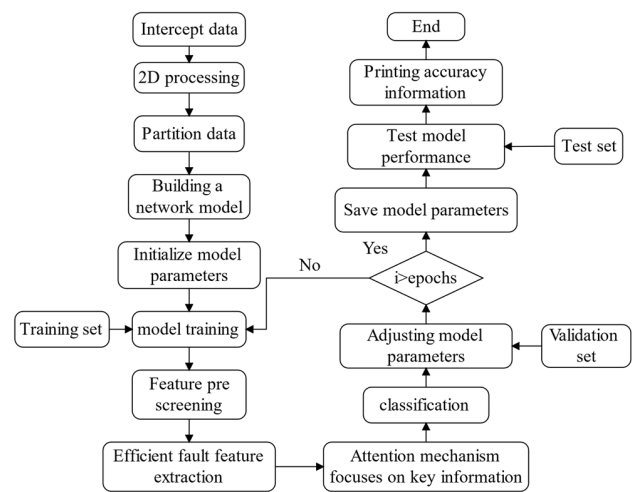


FIGURE 11. Fault diagnosis flowchart.

optimal performance. The best parameters obtained during the training process are saved after reaching the specified number of training iterations.

(6) The best parameters are imported, and the test set is used to obtain accuracy information corresponding to the model.

IV. EXPERIMENTAL ANALYSIS

This article uses the bearing vibration dataset from CWRU and the dataset from JNU for testing to validate the model's generalization and robustness.

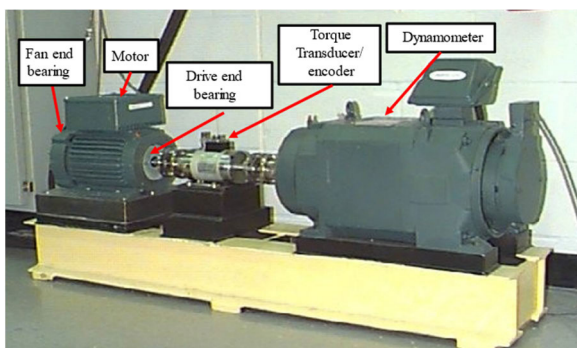
A. CWRU DATASET

The experimental setup of the CWRU dataset is shown in Figure 12, which includes a 1.5 kW motor, a torque sensor, a power analyzer, and an electronic controller. The experiment selected the data from the drive-end bearing, with the SKF6205 bearing model being tested. The bearing loads are

TABLE 2. CWRU dataset partitioning.

| Dataset | Number | Type | Training/Validation/Test Set |
|---------|--------|--------|------------------------------|
| A/B/C | 0 | IR007 | 240/30/30 |
| | 1 | IR014 | 240/30/30 |
| | 2 | IR021 | 240/30/30 |
| | 3 | OR007 | 240/30/30 |
| | 4 | OR014 | 240/30/30 |
| | 5 | OR021 | 240/30/30 |
| | 6 | B007 | 240/30/30 |
| | 7 | B014 | 240/30/30 |
| | 8 | B021 | 240/30/30 |
| | 9 | normal | 240/30/30 |

divided into 1HP, 2HP, and 3HP, corresponding to a sampling frequency of 12 kHz.

**FIGURE 12.** CWRU bearing test rig.

The fault in the experimental bearings was created using the electrical discharge machining technique to generate a single-point damage. The collected bearing data includes fault data for the inner race, outer race, and rolling elements positions. Each position includes damages with diameters of 0.18 mm, 0.36 mm, and 0.53 mm. Based on the different fault positions and diameters, the bearing data can be divided into ten categories: Normal, inner race faults (IR007, IR014, IR021), outer race faults (OR007, OR014, OR021), and rolling element faults (B007, B014, B021).

According to the data sampling plan, 300 samples were collected for each state under a single load condition. To conveniently represent the corresponding 10 states, they are replaced with the numbers 0-9. The three load conditions are denoted as A, B, and C. The training set, validation set and testing set are divided in an 8:1:1 ratio. The specific partitioning is shown in Table 2.

B. EVALUATION METRICS

In order to evaluate the performance of fault classification results, four metrics are used: accuracy (Acc), precision (Pre), recall (Rec), and F-score. Accuracy is the most direct evaluation metric, which represents the proportion of correctly predicted samples to the total number of samples. It is simple and easy to understand and reflects the overall accuracy of the classification results. However, when the sample distri-

bution is imbalanced, the evaluation of accuracy may be biased. Therefore, other evaluation metrics need to be used in conjunction with accuracy to assess the classification performance of the model. The formula for calculating accuracy is as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

where TP (True Positive) represents correctly classified positive samples, FP (False Positive) represents negative samples incorrectly classified as positive, TN (True Negative) represents correctly classified negative samples, and FN (False Negative) represents positive samples incorrectly classified as negative.

Precision represents the proportion of correctly predicted positive samples to all samples predicted as positive. It evaluates the accuracy of positive class predictions. The formula for calculating precision is as follows:

$$Pre = \frac{TP}{TP + FP} \quad (11)$$

Recall represents the proportion of correctly predicted positive samples to all actual positive samples. It is advantageous for evaluating the model's ability to capture positive class samples. However, it ignores false negatives and needs to be used in conjunction with precision to evaluate model performance. The formula for calculating recall is as follows:

$$Rec = \frac{TP}{TP + FN} \quad (12)$$

F1 score is the weighted harmonic mean of precision and recall, providing a balanced measure between accuracy and capture capability. Depending on the specific scenario, the use of F1 score as an evaluation metric needs to be determined. The formula for calculating F1 score is as follows:

$$F - Score = 2 \frac{Pre \cdot Rec}{Pre + Rec} \quad (13)$$

C. MODEL VISUALIZATION AND PERFORMANCE ANALYSIS

We used the t-SNE algorithm [38] for visualization, as it is a nonlinear dimensionality reduction algorithm that allows for intuitive analysis of the model's effectiveness. The t-SNE algorithm was applied to visualize a randomly sampled subset of 2000 examples from the dataset, with 200 samples for each class. In order to evaluate the performance of the proposed model in this paper for visualization-based classification, a comparison was made between the visualization results obtained in this paper and those of LeNet, LSTM, ResNet and WDCNN [39].

LeNet network consists of two convolutional layers and three fully connected layers. The size of the convolutional kernel is 5×5 , followed by a corresponding pooling layer after each convolution. LSTM network includes key components such as input gate, forget gate, and output gate. It effectively processes information in sequences by adaptively updating and managing the state of memory cells.

TABLE 3. Comparison of model diagnosis results.

| Method | Dataset A | Dateset B | Dataset C | Average |
|-----------------|-----------|-----------|-----------|---------|
| LeNet | 96.67% | 94.45% | 97.35% | 96.15% |
| LSTM | 91.53% | 89.85% | 95.64% | 92.34% |
| ResNet | 98.59% | 97.23% | 99.83% | 98.55% |
| WDCNN | 96.54% | 95.46% | 97.43% | 96.47% |
| Proposed method | 100% | 100% | 100% | 100% |

ResNet is a very deep convolutional neural network composed of basic convolutional layers, pooling layers, and batch normalization layers. It introduces residual connections to address the degradation problem in deep networks. Each residual block consists of two or three convolutional layers and a skip connection. The size of the convolutional layers is typically 3×3 . WDCNN consists of five convolutional layers and five maximum pooling layers. The first convolutional layer is a wide convolutional layer with a kernel size of 64×1 , while the subsequent layers have a kernel size of 3×1 .

The comparative graph is shown in Figure 13. From the graph, it can be observed that the proposed model in this paper accurately and correctly partitions the target dataset, while the other models still exhibit significant misclassification in the final classification.

Through training, the accuracy of all 10 classes in Table 3 can reach 100%, which is higher than other classification methods including LeNet, LSTM, ResNet, and WDCNN. The LSTM method and the WDCNN method both belong to one-dimensional convolution methods. In comparison, two-dimensional convolution can extract features in both row and column directions simultaneously, thereby obtaining more comprehensive feature representations. From the diagnostic results, it can be seen that using two-dimensional convolution for fault diagnosis has advantages. Additionally, the WDCNN model adopts a multi-scale convolution approach, which allows capturing fault features at different scales and obtaining better global structural information. Therefore, despite being in the form of one-dimensional convolution, the WDCNN achieves higher accuracy than the LSTM model. The proposed model in this paper takes advantage of the multi-scale structure and combines it with large-scale convolution kernels and attention mechanisms to more efficiently process information at different scales, thereby improving the diagnostic performance of the model. Furthermore, Figure 14 displays the confusion matrix for load condition A, clearly demonstrating that each label has been well classified.

In addition, we also compared the training and testing time of different models on dataset A, and the results are shown in Table 4.

According to the data in the table, it can be seen that the model proposed in this article consumed relatively less time. Referring to the classification accuracy in Table 2, the diagnostic accuracy of the proposed model in this article increased by 1.45% compared to ResNet. Compared with the

TABLE 4. Comparison of time consumption between different models.

| Model | Training time | Test time |
|--------|---------------|-----------|
| LSTM | 49s | 2.2s |
| LeNet | 45s | 2.1s |
| ResNet | 60s | 2.9s |
| WDCNN | 72s | 3.2s |
| Ours | 50s | 2.1s |

longest training time model, WDCNN, the training time was only 38s. Compared with the shortest training time model, LeNet, the training time was only 5s longer, but the training accuracy improved by 3.85%. The above data indicates that the proposed model in this article has low complexity, high computational efficiency, and excellent diagnostic performance.

D. EXPERIMENTAL VERIFICATION UNDER MULTIPLE OPERATING CONDITIONS

In practical engineering applications, data acquisition is influenced by surrounding environmental conditions. Additionally, the number of fault categories we need to classify is usually much greater than 10. Here, we obtain more categories and quantities of fault types by combining data. Specifically, we combine different bearing data at three original speeds. The basis for the combination comes from the differences in sample distribution under different load conditions, as shown in Figure 16. This results in a total of 30 fault categories and three times the data volume compared to a single operating condition.

To better evaluate the effectiveness and computational scale of this model, we compare its experimental results with those of LSTM, ResNet, LeNet and WDCNN models. The comparison results are shown in Figure 15, where it can be observed that our model outperforms other popular models, achieving good detection results in a short time.

The confusion matrix in Figure 17 reflects the specific label recognition results. It can be seen that misidentified label data mostly come from normal samples under different load conditions. This is because normal sample data have similar signal distributions and are less affected by speed differences, making them difficult to distinguish.

However, our model still achieves an accuracy of 99.46% in recognition. This indicates that our proposed method can maintain high fault recognition capability in multiple operating conditions, meeting the requirements of real-world engineering applications.

It should be noted that the processor requirements and computation time for specific industrial applications may differ from the experimental setup described in this paper. This is because factors such as the complex operating environment, the scale of data extraction, and the precision of sensors in industrial scenarios can all affect training time. The algorithm proposed in this paper involves attention mechanisms and large kernel convolution modules, which require

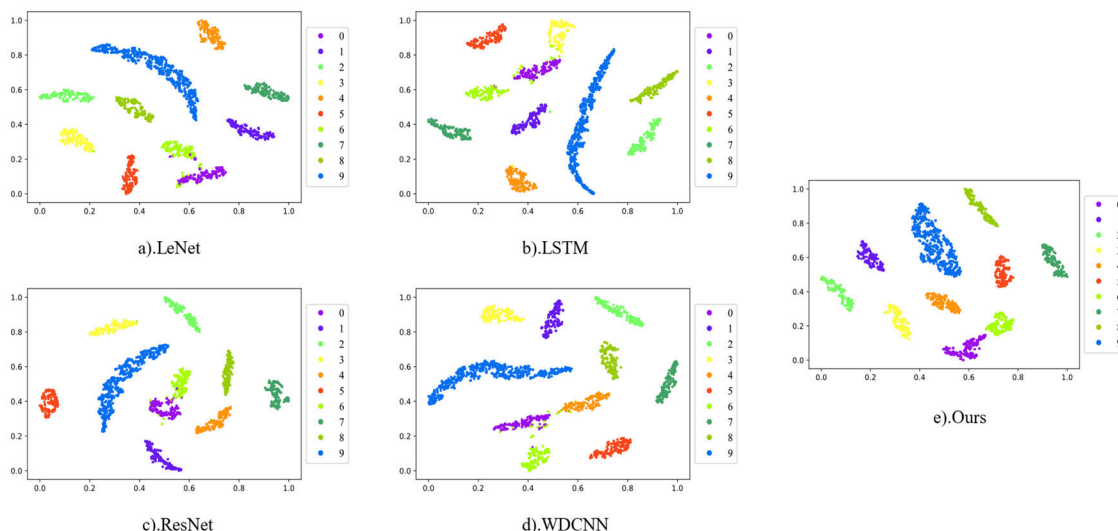


FIGURE 13. Comparison of visualization-based classification results for different models.

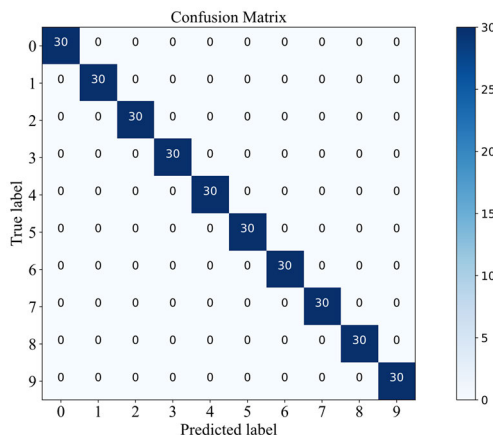


FIGURE 14. Confusion matrix for load condition A.

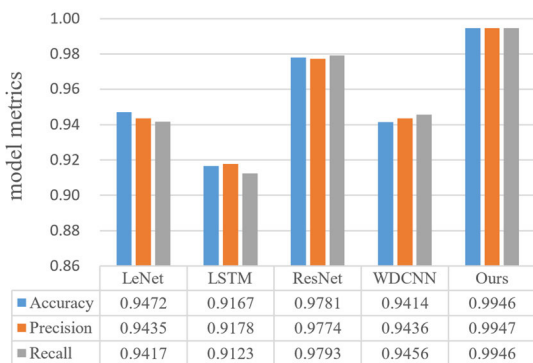


FIGURE 15. Comparison of different model parameters.

CUDA implementation for GPU acceleration. There are no specific requirements for CPUs. Considering the cost budget and computational power of GPUs in industrial scenarios, graphics card models such as GeForce GTX 650 or higher can be used.

The graphics card model used in this experiment, GeForce RTX 3090, has a computational capability that is 2.85 times higher than the previous one. Specifically, in terms of training time, using GeForce GTX 650 takes approximately 2.85 times longer than the time mentioned in this article, which is around 142 seconds. However, industrial data sets are usually larger in scale compared to the data set used in this article, which consists of 3000 samples. If we calculate based on an industrial scale of 10,000 samples, it would take approximately 473 seconds.

Therefore, when implementing this method on a large scale in the industrial field, it is necessary to use at least a GeForce GTX 650 or higher graphics card, with a corresponding computation time of approximately 473 seconds.

E. EXPERIMENTAL STUDY ON TRANSFER LEARNING UNDER VARIABLE LOAD

To verify the generalization ability of the proposed model in this paper, a fault diagnosis experiment on bearings under variable load was conducted in this section. The proposed method was compared with LeNet, ResNet, LeNet, and WDCNN, and the average values of five repeated experiments were taken as the experimental results. In the specific experiment, six transfer tasks can be established based on the data under three different loads. For convenience, “A → B” is used to represent the transfer diagnostic task with A as the source domain data and B as the target domain data. In the target domain, the training set and test set were divided in a 2:8 ratio. The specific transfer tasks and data division are shown in Table 5. The diagnostic results of each method under variable load conditions are shown in Figure 18.

From the figure, it can be seen that the average accuracy of the proposed method reached 98.35%, which is much higher than other models. Compared with the LSTM and WDCNN models adopted by the comparative methods, which use one-dimensional convolution with raw vibra-

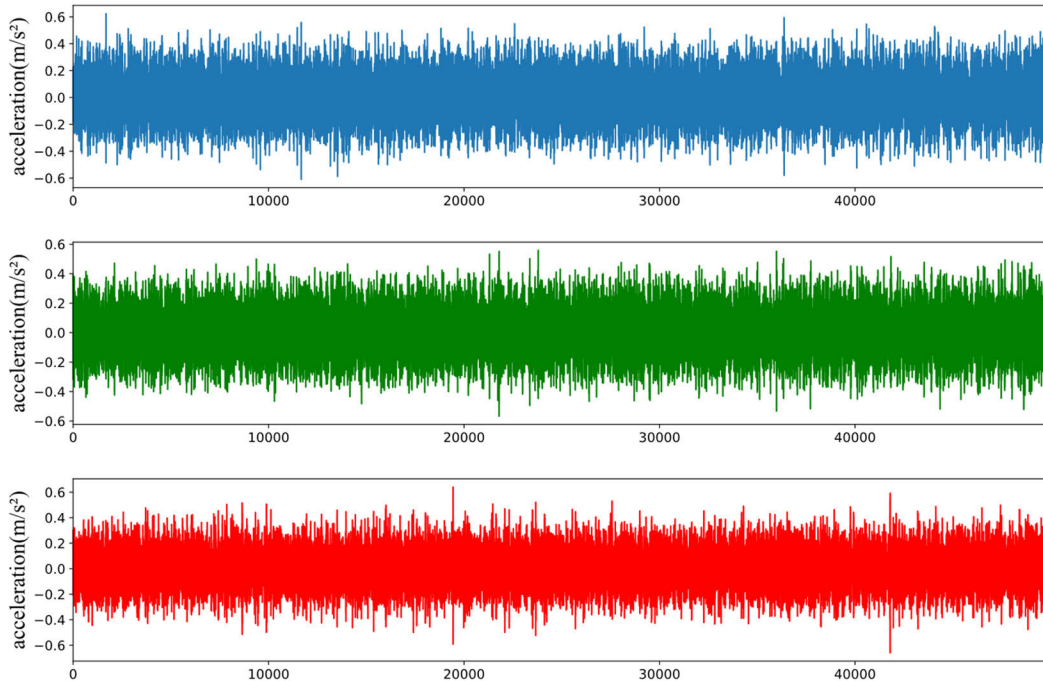


FIGURE 16. Fault signals under different operating conditions.

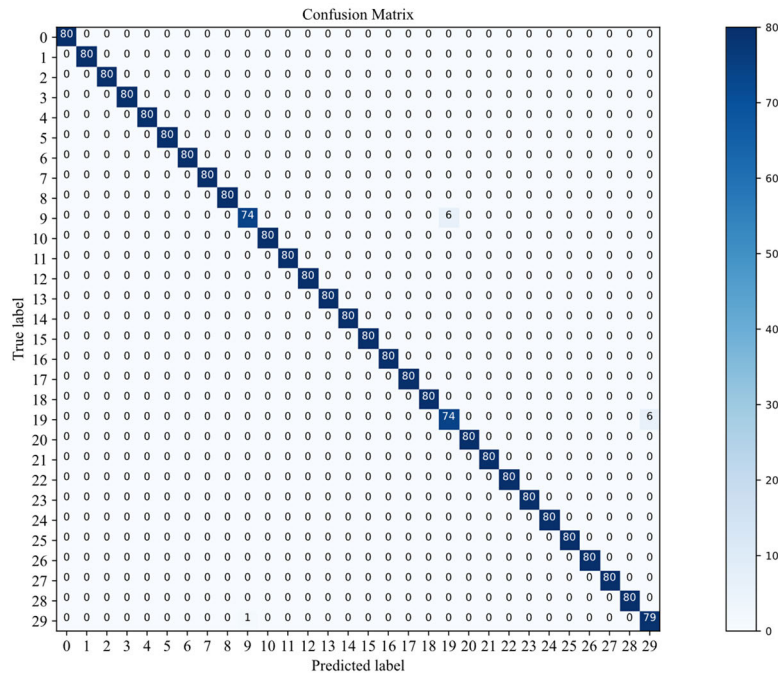


FIGURE 17. Confusion matrices for three load conditions.

tion signals as input, the diagnostic effect is limited. The LeNet and ResNet methods use continuous wavelet transform to extract the time-frequency features of the original signals and use two-dimensional convolution to efficiently process more abundant fault feature information, thereby

improving the diagnostic accuracy. The proposed method, by introducing multi-scale convolution and attention mechanism, can more efficiently extract effective fault information and still achieve good results even under variable load conditions.

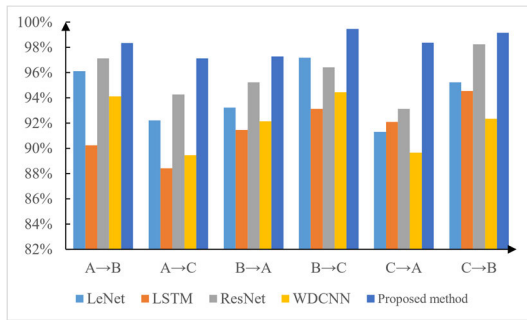


FIGURE 18. Diagnostic results for different migration tasks.

TABLE 5. Migration data partitioning.

| Migration task | Source domain | Target Domain | Train | | Test |
|----------------|---------------|---------------|---------------|---------------|------|
| | | | Source domain | Target Domain | |
| A→B | A | B | 3000 | 600 | 2400 |
| A→C | A | C | 3000 | 600 | 2400 |
| B→A | B | A | 3000 | 600 | 2400 |
| B→C | B | C | 3000 | 600 | 2400 |
| C→A | C | A | 3000 | 600 | 2400 |
| C→B | C | B | 3000 | 600 | 2400 |

TABLE 6. JNU dataset.

| Dataset | Location | Number of samples | Label |
|---------|-----------------|-------------------|-------|
| D/E/F | Inner ring | 400 | 0 |
| | Outer ring | 400 | 1 |
| | Rolling element | 400 | 2 |
| | Health | 400 | 3 |

F. GENERALIZATION EXPERIMENT ANALYSIS

To further verify the performance of the proposed method on other datasets, experiments were conducted using the bearing dataset from JNU. This dataset includes fault states such as normal, inner ring fault, outer ring fault, and rolling element fault, with a sampling frequency of 50kHz for vibration signals. Data collection was performed for each fault type at three different speeds: 600 r/min, 800 r/min, and 1000 r/min. The same overlapping sampling method was used, collecting 400 data samples for each state, and generating corresponding two-dimensional time-frequency images through continuous wavelet transform. The data collected at the three speeds are denoted as D, E, and F, respectively, with each speed containing 1600 data samples. The data partitioning of the dataset is shown in Table 6.

During the actual training, the number of categories was changed to 4 while keeping other parameters unchanged. In the experimental phase, five methods (LSTM, ResNet, LeNet, WDCNN, and the method proposed in this paper) were compared and tested, and the experimental results under different models are shown in Table 7.

By analyzing the data in the table, it can be concluded that the proposed method in this paper still demonstrates excellent

TABLE 7. Comparison of model diagnosis results.

| Method | Dataset D | Dateset E | Dataset F | Average |
|-----------------|-----------|-----------|-----------|---------|
| LeNet | 96.67% | 97.32% | 97.57% | 97.18% |
| LSTM | 92.46% | 93.56% | 93.64% | 93.22% |
| ResNet | 98.62% | 98.23% | 99.23% | 98.69% |
| WDCNN | 97.34% | 96.56% | 97.27% | 97.05% |
| Proposed method | 100% | 100% | 100% | 100% |

diagnostic performance when applied to different datasets. Compared to the ResNet network, the diagnostic accuracy has improved by 1.31%. These findings suggest that the algorithm proposed in this paper has good generalizability.

V. CONCLUSION

This paper proposes a multi-scale convolutional neural network model based on attention mechanism for bearing fault diagnosis. A reasonable truncation sampling method is proposed to address the form of fault signals, and the signals are transformed into two-dimensional matrices for processing through continuous wavelet transform. To improve the effective receptive field and reduce the number of model parameters, an efficient multi-scale convolution module is designed. At the same time, the design of residual module can effectively avoid the problem of gradient explosion and gradient disappearance. In order to address the interference problem of noise information in the data itself, an efficient channel attention module is proposed to focus on effective information. The algorithm was validated on the CWRU bearing vibration dataset and the JNU bearing dataset, and compared with other mainstream algorithm models. The results show that the proposed method performs well in both single and multiple working conditions, and outperforms other algorithms. In addition, to verify the generalization ability of the model, different transfer tasks were established using the CWRU dataset, and the experiment shows that the proposed model has good transfer performance.

REFERENCES

- [1] X. Zhang, S. Wan, Y. He, X. Wang, and L. Dou, "Bearing fault diagnosis based on iterative 1.5-dimensional spectral kurtosis," *IEEE Access*, vol. 8, pp. 174233–174243, 2020.
- [2] B. R. Nayana and P. Geethanjali, "Analysis of statistical time-domain features effectiveness in identification of bearing faults from vibration signal," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5618–5625, Sep. 2017.
- [3] D. Li, "Research of fault diagnosis of mine rolling bearing based on time-frequency analysis," in *Proc. J. Phys., Conf.*, 2023, vol. 2459, no. 1, Art. no. 012081.
- [4] C. Zhang, Y. He, L. Zuo, J. Wang, and W. He, "A novel approach to diagnosis of analog circuit incipient faults based on KECA and OAO LSSVM," *Metrology Meas. Syst.*, vol. 22, no. 2, pp. 251–262, Jun. 2015.
- [5] C. Zhang, Y. He, T. Yang, B. Zhang, and J. Wu, "An analog circuit fault diagnosis approach based on improved wavelet transform and MKELM," *Circuits, Syst., Signal Process.*, vol. 41, no. 3, pp. 1255–1286, Mar. 2022.
- [6] W. He, Y. He, B. Li, and C. Zhang, "Feature extraction of analogue circuit fault signals via cross-wavelet transform and variational Bayesian matrix factorisation," *IET Sci., Meas. Technol.*, vol. 13, no. 2, pp. 318–327, Mar. 2019.

- [7] C. Zhang, S. Zhao, Z. Yang, and Y. He, "A multi-fault diagnosis method for lithium-ion battery pack using curvilinear Manhattan distance evaluation and voltage difference analysis," *J. Energy Storage*, vol. 67, Sep. 2023, Art. no. 107575.
- [8] M. Dashtdar, R. Dashti, and H. R. Shaker, "Distribution network fault section identification and fault location using artificial neural network," in *Proc. 5th Int. Conf. Electr. Electron. Eng. (ICEEE)*, May 2018, pp. 273–278.
- [9] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-nearest neighbors and grid search CV based real time fault monitoring system for industries," in *Proc. IEEE 5th Int. Conf. Conver. Technol. (I2CT)*, Mar. 2019, pp. 1–5.
- [10] Q. Shi and H. Zhang, "Fault diagnosis of an autonomous vehicle with an improved SVM algorithm subject to unbalanced datasets," *IEEE Trans. Ind. Electron.*, vol. 68, no. 7, pp. 6248–6256, Jul. 2021.
- [11] F. Zhou, T. Sun, X. Hu, T. Wang, and C. Wen, "A sparse denoising deep neural network for improving fault diagnosis performance," *Signal, Image Video Process.*, vol. 15, no. 8, pp. 1889–1898, Jun. 2021.
- [12] S. Han, S. Oh, and J. Jeong, "Bearing fault diagnosis based on multi-scale convolutional neural network using data augmentation," *J. Sensors*, vol. 2021, pp. 1–14, Feb. 2021.
- [13] Y. Hao, H. Wang, Z. Liu, and H. Han, "Multi-scale CNN based on attention mechanism for rolling bearing fault diagnosis," in *Proc. Asia-Pacific Int. Symp. Adv. Rel. Maintenance Model. (APARM)*, Aug. 2020, pp. 1–5.
- [14] X. Lin, B. Li, X. Yang, and J. Wang, "Fault diagnosis of aero-engine bearing using a stacked auto-encoder network," in *Proc. IEEE 4th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Dec. 2018, pp. 545–548.
- [15] C. Liu, B. Chen, H. Zhang, and X. Wang, "Fault diagnosis application of short wave transmitter based on stacked auto-encoder," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 119–123.
- [16] X. Wang, D. Mao, and X. Li, "Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network," *Measurement*, vol. 173, Mar. 2021, Art. no. 108518.
- [17] H. Liang, J. Cao, and X. Zhao, "Multi-scale dynamic adaptive residual network for fault diagnosis," *Measurement*, vol. 188, Jan. 2022, Art. no. 110397.
- [18] W. Song, H. Liu, and E. Zio, "Long-range dependence and heavy tail characteristics for remaining useful life prediction in rolling bearing degradation," *Appl. Math. Model.*, vol. 102, pp. 268–284, Feb. 2022.
- [19] Y. An, K. Zhang, Q. Liu, Y. Chai, and X. Huang, "Rolling bearing fault diagnosis method base on periodic sparse attention and LSTM," *IEEE Sensors J.*, vol. 22, no. 12, pp. 12044–12053, Jun. 2022.
- [20] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal Process.*, vol. 161, pp. 136–154, Aug. 2019.
- [21] Q. Zhang, N. Tang, X. Fu, H. Peng, C. Bo, and C. Wang, "A multi-scale attention mechanism based domain adversarial neural network strategy for bearing fault diagnosis," *Actuators*, vol. 12, no. 5, p. 188, Apr. 2023.
- [22] Z. Wu, H. Jiang, H. Zhu, and X. Wang, "A knowledge dynamic matching unit-guided multi-source domain adaptation network with attention mechanism for rolling bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 189, Apr. 2023, Art. no. 110098.
- [23] Y. Wang, J. Liang, and X. Gu, "Multi-scale attention mechanism residual neural network for fault diagnosis of rolling bearings," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 236, no. 20, pp. 10615–10629, 2022.
- [24] K. X. Sang, J. Shang, and T. R. Lin, "Synchroextracting transform and deep residual network for varying speed bearing fault diagnostic," *J. Vib. Eng. Technol.*, vol. 11, no. 1, pp. 343–353, Jun. 2022.
- [25] J. Yan, J. Kan, and H. Luo, "Rolling bearing fault diagnosis based on Markov transition field and residual network," *Sensors*, vol. 22, no. 10, p. 3936, May 2022.
- [26] Y. Wang and G. Cao, "A multiscale convolution neural network for bearing fault diagnosis based on frequency division denoising under complex noise conditions," *Complex Intell. Syst.*, vol. 9, no. 4, pp. 4263–4285, Aug. 2023.
- [27] Y. Jin, C. Qin, Z. Zhang, J. Tao, and C. Liu, "A multi-scale convolutional neural network for bearing compound fault diagnosis under various noise conditions," *Sci. China Technol. Sci.*, vol. 65, no. 11, pp. 2551–2563, Nov. 2022.
- [28] X. Li, Z. Lei, and G. Wen, "Intelligent fault diagnosis with multi-scale convolutional dense network," in *Proc. J. Phys., Conf.*, vol. 2184, no. 1, May 2022, Art. no. 012009.
- [29] Y. Hou, J. Wang, Z. Chen, J. Ma, and T. Li, "Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106507.
- [30] J. Lu, K. Wang, C. Chen, and W. Ji, "A deep learning method for rolling bearing fault diagnosis based on attention mechanism and Graham angle field," *Sensors*, vol. 23, no. 12, p. 5487, Jun. 2023.
- [31] J. Wang, J. Guo, L. Wang, Y. Yang, Z. Wang, and R. Wang, "A hybrid intelligent rolling bearing fault diagnosis method combining WKN-BiLSTM and attention mechanism," *Meas. Sci. Technol.*, vol. 34, no. 8, Aug. 2023, Art. no. 085106.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [33] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.
- [34] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31 : Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11953–11965.
- [35] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. NIPS*, 2016, pp. 4905–4913.
- [36] S. Li, X. Wu, and Z. Wu, "Efficient multi-lane detection based on large-kernel convolution and location," *IEEE Access*, vol. 11, pp. 58125–58135, 2023.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [38] A. Chatzimparmpas, R. M. Martins, and A. Kerren, "T-viSNE: Interactive assessment and interpretation of t-SNE projections," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 8, pp. 2696–2714, Aug. 2020.
- [39] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.



QIMIN SHEN was born in Yicheng, Hubei, in 1983. She received the master's degree from Harbin Engineering University, in 2010.

From 2010 to 2021, she was a Designer with the Taiyuan Heavy Machinery Group. During her tenure, she undertook multiple research projects and obtained five patents. In 2019, she was awarded the title of a Senior Engineer. Since 2021, she has been with the Jinzhong College of Information. Her research interests include mechanical design and manufacturing and intelligent operation and maintenance technology.



ZENGQIANG ZHANG was born in Luoyang, Henan, China, in 1982. He received the master's degree from the Taiyuan University of Technology, in 2008. He is currently a Senior Engineer with the Taiyuan Institute of Technology. His research interests include mechanical design, manufacturing, and theoretical analysis.

...