Hindawi
Journal of Sensors
Volume 2022, Article ID 1809482, 20 pages
https://doi.org/10.1155/2022/1809482

Research Article

An Improved Metalearning Framework to Optimize Bearing Fault Diagnosis under Data Imbalance

Xinqian Hu,[1] **Junfeng Man**,[2] **Hengfu Yang**,[2] **Jiangmin Deng**,[3] **and Yi Liu**[3]

[1]*School of Computer Science, Hunan University of Technology, Zhuzhou 412000, China*
[2]*School of Computer Science, Hunan First Normal University, Changsha 410205, China*
[3]*CRRC Zhuzhou Electric Locomotive Co., Ltd., Zhuzhou 412000, China*

Correspondence should be addressed to Junfeng Man; mjfok@qq.com

Received 23 June 2022; Revised 5 September 2022; Accepted 27 September 2022; Published 18 October 2022

Academic Editor: Sangsoon Lim

The intelligent diagnosis of rotating machinery with big data has been widely studied. However, due to the variability of working conditions and difficulty in marking fault samples, it is difficult to obtain enough high-quality fault marking data for training bearing fault diagnosis models in practical industrial application scenarios. Aiming at the problem of training data imbalance caused by lack of fault samples, a novel metalearning fault diagnosis method (MOFD) is proposed to get the bearing fault diagnosis solution under data imbalance. Firstly, in order to enhance the variety of fault samples, a Feature Space Density Adaptive Synthetic Minority Oversampling Technique (FSDA-SMOTE) is proposed in this paper, which takes the density difference of minority samples in the spatial domain within the class as the constraint of local neighbor similarity to generate new fault samples for data augmentation. In addition, in order to strengthen the model's learning ability and diagnosis performance under limited fault samples, a residual-attention convolutional neural network (RA-CNN) was constructed to identify the deep features of fault signals, and a metalearning strategy based on parameter gradient optimization was applied to RA-CNN for refining the learning process of the diagnosis model. Finally, the reliability of the proposed method is verified through experimental analysis of public bearing dataset.

1. Introduction

In the process of modern industrial production and application, rolling bearing plays an essential role in the work efficiency of mechanical equipment as one of the core components of industrial rotating machinery equipment [1]. The components of rolling bearings have complex structure and operate in a poor environment. They are in a state of high load operation for a long time. Once the bearing operation failure occurs, it will increase huge operation and maintenance costs and unknown security risks [2, 3]. Therefore, the application of fault diagnosis based on rolling bearing can monitor and maintain the equipment in a more real-time and effective manner to avoid industrial production accidents. In recent years, the application of data-driven intelligence in the field of industrial machinery fault diagnosis has developed rapidly. Traditional fault diagnosis techniques rely on professional mechanism knowledge and expert experience, combined with signal processing methods [4–6] and pattern recognition techniques [7–9] to extract effective features from sensor data, so as to improve fault diagnosis accuracy. Ye et al. [10] proposed a signal processing technology based on variational mode decomposition (VMD) and a machine learning fault diagnosis model based on particle swarm optimization (PSO-SVM). The proposed method reconstructed vibration signals and calculated their multiscale displacement entropy (MPE) to construct multidimensional feature data and optimized penalty parameter C and kernel parameter g by PSO to improve the performance of SVM classifier. Lu et al. [11] used stack denoising autoencoder (SDAE) to extract fault features from vibration signals containing complex environmental noise and operation state fluctuations for fault identification and classification. The successful implementation of the above traditional fault diagnosis methods often depends on the feasibility of classifying the features extracted from the training data in low-dimensional space. However, for sensor data extracted under complex working conditions, it is difficult to obtain representative features through prior knowledge to distinguish different fault categories of data.

With the great success of deep learning model in the field of computer vision, the adaptive features learning method of CNN in high-dimensional space has attracted the attention of scholars [12–14], Wen et al. [15] proposed a data preprocessing method that converted the time series signal into gray images, and fault diagnosis was carried out by adaptive feature extraction based on Lenet-5 Convolutional neural network. Yao et al. [16] and Ravikumar et al. [17] introduced residual learning blocks to train the deep neural network to ensure that the model has sufficient depth and alleviate the problems of gradient disappearance and overfitting. However, the vibration signal sequence is too long and lacks correlation with local spatial features, which leads to the deviation of feature extraction in time and space dimension. To solve this problem, Wang et al. [18] proposed an attention mechanism (AM) for image classification. It improved the receptive field of the underlying features through multiple up-downsampling operations, so that the deep network could also capture various local dependencies and obtain rich context features, and the model could extract more representative deep features. Ye et al. [19] proposed a time convolutional network (TCN) based on attention mechanism, which extracted effective local features through causal convolution residual blocks and used attention mechanism to make the network tend to pay attention to fault features, so that the model could detect and diagnose fault types more quickly and improve the efficiency of fault diagnosis. Men et al. [20] proposed Res-CBAM model for hyperspectral information recognition by combining residual module and attention module. The classification performance of the model was improved by introducing CBAM to calculate channel and spatial model attention and reassigned weight parameters. However, the success of most of the above deep learning model applications relies on expensive computing resources and large amounts of balanced and annotated training data [21], but in practical industrial production applications, there are not enough fault data to support the training of the neural network model. The reasons are as follows: (1) the internal structure of industrial machinery and equipment is very complicated, and its failure will seriously hamper the progress of industrial production. (2) The degradation cycle time of rolling bearing is very long, which brings huge difficulties for collect sufficient fault data; (3) the nonstationary dynamic characteristics of bearings under variable working conditions greatly cause more trouble for fault data collection and correct identification; current machine learning classifier algorithms are inertially inclined to majority samples. However, in practical application, it is more valuable to correctly classify minority samples [22]. Therefore, the study on the imbalance of training data has become one of the focus in the current field of industrial fault diagnosis [23–25].

At present, some research results have been achieved in data imbalance. For example, methods such as signal translation, noise addition, time stretching, and resampling are proposed to increase the number and diversity of training data artificially [26–29]. Synthetic minority oversampling technique (SMOTE) [30] is one of the most widely used data augmentation methods. This method uses feature space composed of a few samples and their K-nearest neighbors to synthesize new samples, which can effectively reduce the overfitting phenomenon, but easily magnifies noise and leads to fuzzy boundary of adjacent category data. Wei et al. [31] proposed a cluster-based majority-weighted minority oversampling technique (cluster-MWMOTE), which further improved the model's adaptation to the internal imbalance of fault instances. Yi et al. [32] enhanced the data of minorities on the basis of clustering, in which the sample of minority group was aggregated into several clusters, and new samples were generated by linear interpolation between adjacent clusters. However, with the existence of boundary effect, it is easy to generate noise samples based on the nearest neighbor principle, which leads to sample overlap between classes, and the interference model learns the true spatial distribution of the original data samples. Therefore, in order to solve the problem that traditional data synthesis methods cannot mine deep features of data and generate the high-quality sample, generative adversarial network (GAN) was proposed by Goodfellow et al. [33]. As the most popular data enhancement method in recent years, GAN can generate pseudoimages, audio, or video based on real datasets to solve the problem of training data imbalance. Zhao and Yuan [34] proposed an improved GAN model. The improved GAN introduced an auxiliary classifier to facilitate the training process and an autoencoder-based method to estimate the similarity of the generated samples, thus improving the quality and diversity of the generated samples. Although the applications of fault diagnosis based on GAN are very wide, problems such as instability, modal collapse, and weak gradient in GAN training process [35–38] make it difficult to apply these models in practical engineering. Therefore, in order to overcome the challenge of faulty data imbalance, the research community needs to focus on developing an efficient computing model with faster learning ability that can be fit to task requirements even when datasets are unbalanced.

Recently, the superior performance of metalearning in solving the problem of few-shot fault diagnosis has gradually attracted the focus of scholars [39–42]. Model agnostic metalearning (MAML) is a metalearning method based on parameter optimization proposed by Finn et al. [43], which had demonstrated excellent generalization ability in image recognition for processing new tasks under a small number of training samples. Yu et al. [44] proposed a metalearning fault diagnosis model based on gradient optimization [45], which optimized the initial parameters of the model network through the scenario training mechanism, so that the model can also perform fault diagnosis efficiently and quickly under the condition of limited training data. Zhang et al. [21] applied Siamese network to fault diagnosis with data imbalanced. Siamese network is used to measure the distance between the same or different sample pairs to determine their similarity, so as to achieve high precision fault diagnosis with limited samples. Vinyals et al. [46] proposed

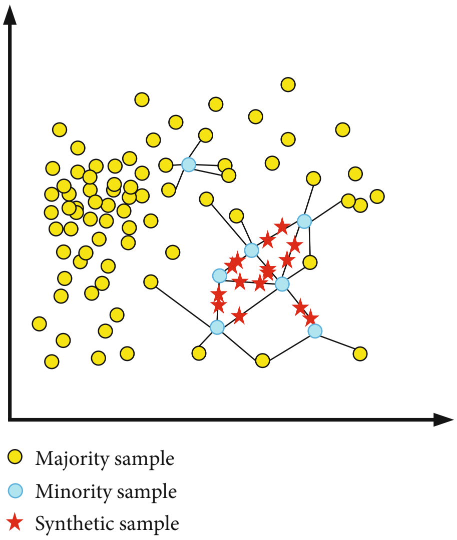


Figure 1: The process of SMOTE synthesizes minority samples.

a matching network for few-shot classification, which designed a short- and long-term memory network to realize metric-based metalearning architecture, so as to avoid fine-tuning requirements of new tasks. Chen et al. [47] proposed an improved prototype network to better improve the fault identification performance of the classification model in the case of data imbalance by setting a multiscale feature extractor and an appropriate distance measurement function. The above metalearning methods can solve the problem of data imbalance well in current industrial intelligent fault identification. The advantage of few-shot learning can further alleviate the problem of unbalanced training data and enhance the learning ability of the model for complex working conditions. We improve the framework of bearing fault diagnosis under the metalearning mechanism and summarize the following contributions:

(1) This paper proposes a metalearning network (RA-CNN) that integrates residual learning modules and attentional mechanisms. This method can effectively extract fault features with the limited fault data and avoid model overfitting

(2) This paper proposes FSDA-SMOTE for data augmentation. Compared with the traditional data augmentation method, this method can reduce the interference of data noise more effectively and avoid the fuzzy classification boundary

(3) This paper proposes a k-way N-shot episodic training mechanism to refine the feature learning process of network models. This method enables the model to learn more general fault information from multiple different metatasks

The rest of this paper are as follows: Section 2 outlines the background on relevant theories. Section 3 describes in detail the proposed methodology and framework. Section 4 and Section 5 are the experimental analysis and conclusions, respectively.

2. Related Research Work

2.1. Metalearning. Metalearning is referred to as “learning-to-learn,” which is generally utilized for tackling few-shot image classification. In recent years, metalearning research can be generally divided into three different categories, including optimizing the initialization parameters of the metalearner to quickly adapt to new tasks [43–45], generating the metric-learning network by judging the feature similarity between sample pairs [21, 46], and learning a recurrent neural network model with memory storage function [47]. Different from traditional deep learning methods, metalearning is a flexible framework that learns prior experience from multiple relevant tasks, which relies on the obtained experience to improve its performance on target tasks without training from scratch [42]. Metalearning is aimed at generating a general algorithm based on the ability to learn metaknowledge θ_i from different tasks T_i. For a given task $\{ T_1, T_2,\cdots,T_i\} \subseteq T$, this method makes the model have stronger generalization ability and faster adaptability for task sets $p(T)$ with different distributions. The description is shown in

$$\underset{\theta}{\operatorname{argmin}} \sum_{T\sim p(T)} \operatorname{Loss}(T\,; f(\theta)), \tag{1}$$

where $\operatorname{Loss}(T\,; f(\theta))$ represents the loss function obtained by training data in task T using the model $f(\cdot)$ of initializing network parameter θ. It is worth noting that θ in metalearning is constantly updated across multiple metatasks in the learning process, and the θ obtained after tasks learning can reduce the loss of new tasks as much as possible.

2.2. Synthetic Minority Oversampling Technique (SMOTE). SMOTE, a common oversampling method, was proposed by Chawla et al. in 2002 [30]. SMOTE not only solves the duplication problem caused by oversampling but also effectively reduces overfitting by creating a positive composite in the characteristic space of a few class instances and their K nearest neighbors (usually, K is 5 for SMOTE). Suppose a dataset contains m positive and n negative samples ($m > n$), in order to balance the positive and negative samples, m-n synthetic negative samples must be created. First, sample x_i is randomly selected from n negative samples. Second, choose 6 nearest neighbors of x_i in n instances which are identified through the Euclidean distance as shown in

$$d_k\left(x_i, x_j\right) = \sqrt{\sum_{l=1}^{n} \left(x_i^l - x_j^l\right)^2}, \tag{2}$$

$$x_{\text{new}} = x_i + \text{rand}\ (0, 1) \cdot \left(x_i - d_k\left(x_i, x_j\right)\right), \tag{3}$$

where l denoted the identity of the sample point and rand $(0, 1)$ is a random number from 0 to 1, which guarantees that a new minority sample x_{new} is generated by linear interpolation as shown in Figure 1. The above steps are repeated to obtain additional synthetic minority samples. Thus, SMOTE has been proven effective in dealing with imbalanced data.

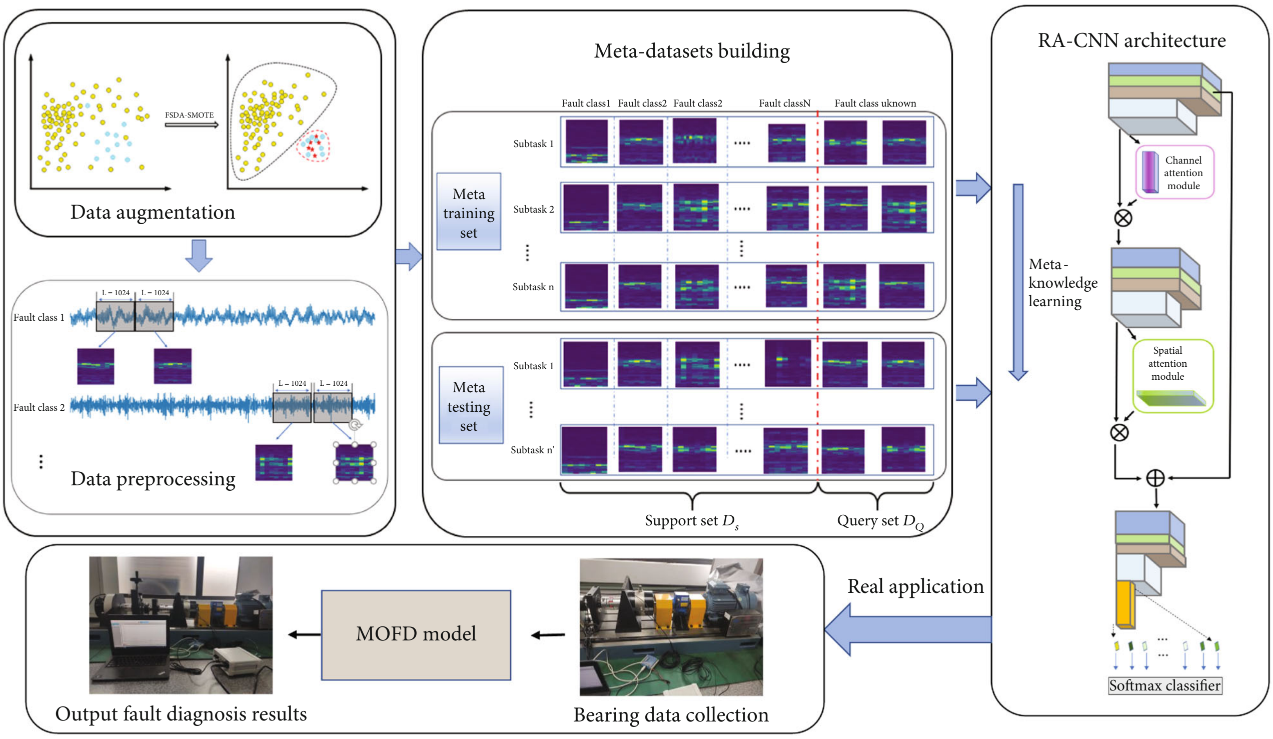


Figure 2: The flowchart of the metaoptimization fault diagnosis model.

3. Proposed Methodology

The bearing fault diagnosis under data imbalance can be treated as a typical few-shot classification problem, which is difficult for traditional deep-learning models to solve. Therefore, MOFD method based on advanced metalearning theory is proposed, and the flowchart is shown in Figure 2. The whole process is divided into four parts: (1) in data augmentation, bearing vibration data of different fault categories is collected, FSDA-SMOTE is used to generate auxiliary data, then the dataset is expanded and balanced; (2) in data preprocess, the original balanced vibration data is converted to time-frequency image (TFI) by STFT; (3) in metadataset building, the TFI dataset is used to organize metalearning tasks according to the episodic training mechanism; (4) finally, RA-CNN is used to fine-tuning its network parameters until the optimal parameters are achieved by extracting the deep fault features of TFIs from multiple metatraining tasks and metatest tasks, and the MOFD with the optimal network parameters can effectively realize fault classification under data imbalance.

3.1. Data Augmentation. A new oversampling technique called FSDA-SMOTE is proposed to improve the shortcomings of the traditional KNN-based oversampling method [32]. FSDA-SMOTE filters outlier samples out by averaging the intraclass distance [27]. In the meantime, in order to avoid the synthesized new samples from fuzzy boundary between classes, the distance weighting method is adopted to select the original samples for synthesis, and the minority samples are clustered. Finally, a new minority class sample is synthesized according to the spatial density difference of samples in the minority class cluster. The algorithm can not only enrich minority class samples in the inner space of a clusters but also avoid the confusion of the boundary between different classes.

The details of the FSDA-SMOTE algorithm for synthesizing new minority samples are represented in Algorithm 1. For convenience, these collections of composited minority patterns are represented by $S_{\text{min\_New}}$. The samples in $S_{\text{min\_New}}$, S_{minf}, and S_{majf} are combined to get a new set, which is represented by S_{new}, where S_{minf} and S_{majf} represent the minority class samples and the majority class sample after denoising. For Algorithm 1, lines 1-15 of the above pseudocodes are suited for denoising data samples. Lines 16-20 are clustering process for a minority samples. Lines 21-30 are density adaptive data generation based on sample density differences within the class.

3.2. Data Preprocessing. Using accelerometer sensor to collect bearing original vibration signal is the most common method in fault diagnosis field [41]. However, the original signal cannot directly show the relationship between the time information and the fault characteristics. Therefore, it is more beneficial for the model to learn the fault characteristics by converting it into a frequency domain signal with stable frequency domain characteristics through fast Fourier transform (FFT). However, only relying on one-dimensional analysis in time and frequency domain, the model cannot capture the internal information of bearings in various fault states to learn and transfer applications. Therefore, two-dimensional time-frequency analysis has become an important tool to analyze mechanical vibration signals for fault

```
Input: majority data $S_{maj}$; minority data $S_{min.}$
Initialize: nearest neighbors $K_1$; means clustering $K_2$
output: the set $S_{new}$ of new samples
obtain the $K_1$ nearest neighbors of $x_i$ and compose of the set $NN_{min}(x_i)$. $x_i \in S_{min}$
$S_{min\_noise} = \{x_i \mid x_i \notin NN_{min}(x_i)\}$
$S_{minf} = S_{min\_noise} \cap S_{min}$
obtain the $K_1$ nearest neighbors of $y_i$ and compose of the set $NN_{maj}(y_i)$. $y_i \in S_{maj}$
$S_{maj\_noise} = \{y_i \mid y_i \notin NN_{maj}(y_i)\}$
$S_{majf} = S_{maj\_noise} \cap S_{maj}$
for$x_i$ in $S_{minf}$do:
    for$y_i$ in $S_{majf}$do:
      $d_{min}(x_i)$ = the minimum Euclidean distance between $x_i$ and $y_i$
         dist($x_i$, $x_j$) = the Euclidean distance between $x_i$ and $x_j$, $i \neq j$
         ifdist($x_i$, $x_j$) < $d_{min}(x_i)$:
           Cluster $J_i = \{x_j \mid \mathrm{dist}(x_i, x_j) < d_{min}(x_i), i \neq j\}$
         end if
     end for
  end for
  for$i$ in the number of samples in $S_{minf}$ do:
    if$x_i \in J_i$:
      new cluster $Q_i = \{J_i \mid J_j \text{ contains } x_i, i \neq j\}$
    end if
  end for
  for$j$ in the number of clusters Q do:
    get the center $cy_i$ by the $K$-means cluster algorithm
    for each sample $x_i$ in $Q_j$do:
         radius ($x_i$) = the mean distance of $x_i$'s $k$-nearest neighbors
         spreadnumber ($x_i$) = the number of $x_i$'s contained in a circle of radius
         spreadability ($x_i$) = the sum of $x_i$'s spreadnumber in the current circle
         density $\rho(x_i)$ = spreadability($x_i$)/spreadability$_{max}(x)$
         synthesize new sample $x_{new} = cy_i + \rho(x_i) * (cy_i - x_i)$.
         $S_{new} \longleftarrow x_{new}$
      end for
  end for
```

Algorithm 1: The procedure of FSDA-SMOTE.

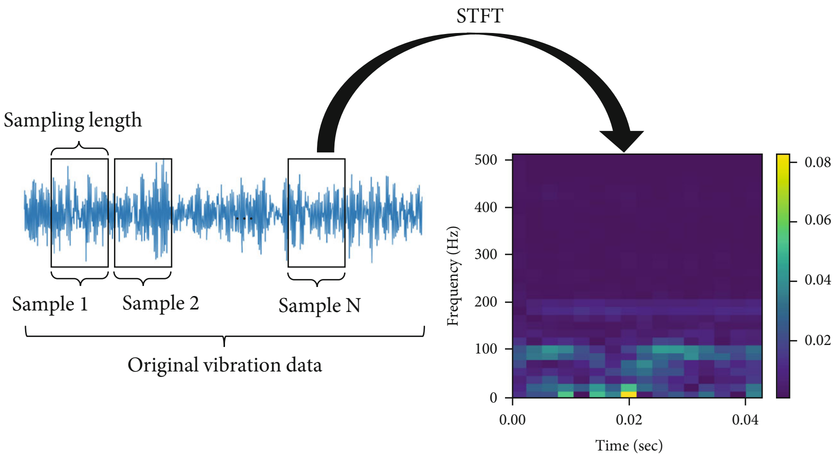


Figure 3: The process of STFT.

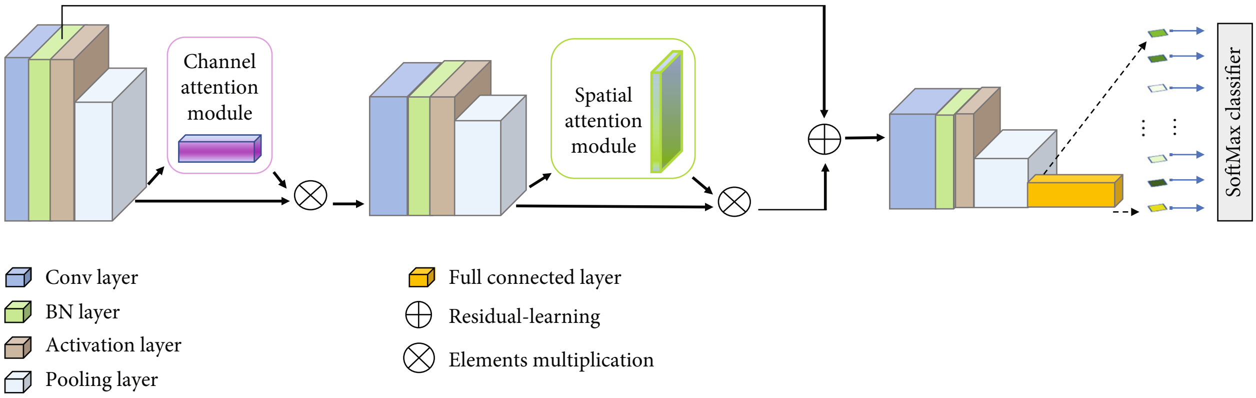


Figure 4: The network structure of RA-CNN.

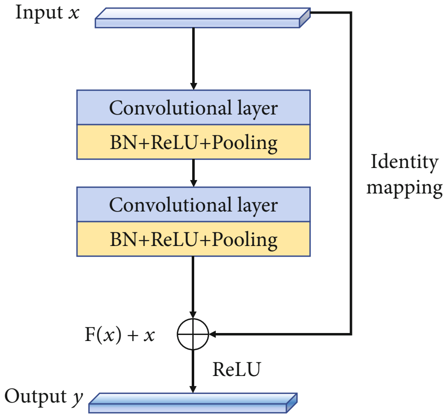


Figure 5: The process of residual learning.

diagnosis [40, 42]. The time-frequency analysis method cited in this paper is short-time Fourier transform (STFT), and its process is shown in Figure 3. In Figure 3, the x-axis, y-axis, and brightness of TFIs represent time domain, frequency domain, and amplitude, respectively.

STFT cuts the original signal $f(t)$ segment by segment through the finite scale window function $\varphi(t)$ and applies the FFT to each segment of the time axis at τ. Finally, TFI with time-frequency domain information is formed. The mathematical formula of STFT is shown in

$$\mathrm{STFT}(\omega, t) = \int_{-\infty}^{+\infty} f(t)\varphi(t-\tau)e^{-j\omega t}dt. \tag{4}$$

In STFT processing, the time-frequency resolution in the spectrogram is determined by the scale of the window function. Therefore, the best time-frequency resolution can be obtained by selecting appropriate window function widths. The time-frequency resolution P is calculated as:

$$P = \left\lfloor \frac{N_s - N_p}{N_w - N_p} \right\rfloor \times \left\lfloor \frac{N_f}{2} + 1 \right\rfloor, \tag{5}$$

where $\lfloor \cdot \rfloor$ means rounding down and N_s represents the length of segments of the signal. In order to reduce spectrum leakage, the scale of the window function $N_w = 64$ or $N = 128$ applied in this paper. N_p is the number of overlapping points, and N_f is the number of points participating in the Fourier transform.

3.3. Metadataset Building. In k-way, n-shot episodic training mechanism, "k" represents the number of categories that the task must classify, and "n" represents the number of images available for learning from each category. For example, the model has a total of 50 images (5 for each class) in 10-way, 5-shot learning problem, which are used to learn and classify the test images in the 10 classes. To ensure that RA-CNN learns more general fault information from multiple different tasks, this paper proposes to use TFI before and after data enhancement as $D_{\text{meta-test}}$ and $D_{\text{meta-train}}$, respectively. During the preparation of the metatraining set, M TFIs from each type of failure are randomly sampled, and K TFIs are used to form a 10-way K-shot support set. The support set of each task is used to train the fault classification model and provide corresponding loss feedback, and the data of the query set is used to verify the classification effect of the trained model. Repeating this way many times, the tasks of the entire metatraining phase is assigned, and the tasks of the metatest set are obtained in the same way, but it is worth noting that the $D_{\text{meta-train}}$ is obtained from the balanced dataset after data enhancement, while the $D_{\text{meta-test}}$ is obtained from the initial unbalanced dataset.

3.4. RA-CNN Architecture. The CNN has been proved to have a good performance when the original vibration data is processed by STFT for feature extraction before neural network training [44], while many kinds of CNNs cannot learn features well from the TFI of limited original rotating machinery fault data. Therefore, the proposed metalearning network model RA-CNN for high-precision intelligent fault diagnosis is combined with 2-dimensional convolutional neural network (2D-CNN), residual learning, and attention mechanism. In the case of limited training data, in order to facilitate the model to extract more representative image features, we combine the attention mechanism and 2D-CNN to make the model strengthen the learning weight of fault type features with minority samples and improve its learning ability for fault features. Then, in order to avoid model overfitting, we add the residual learning modules to

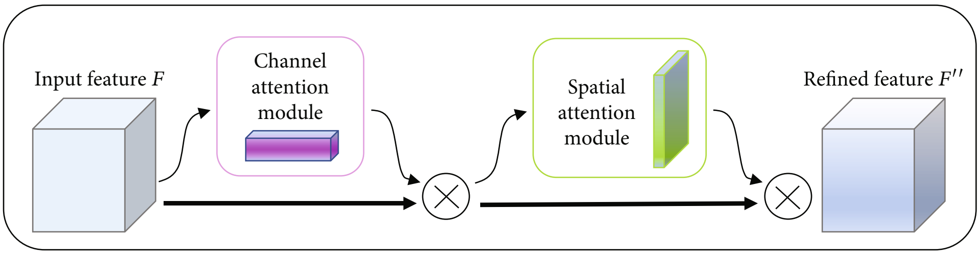


FIGURE 6: The flowchart of CBAM.

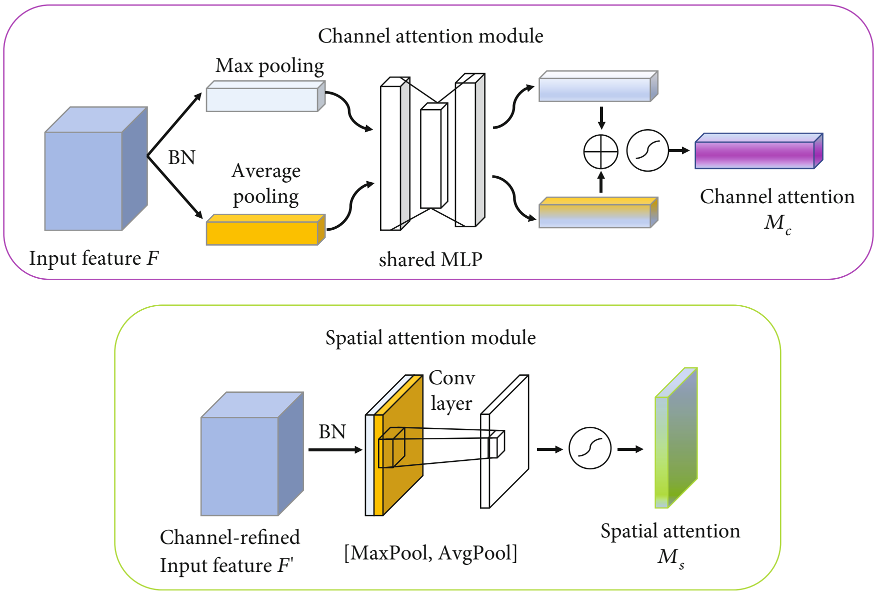


FIGURE 7: Diagram of channel and spatial attention.

the proposed model. Since metalearning focuses on the optimization of lightweight networks, the backbone of RA-CNN is a CNN with four convolution layers, as shown in Figure 4, which is decided by several simulation experiments and the special structure of the ResNet [20].

3.4.1. Residual Learning. The residual learning techniques provide an easy way to train neural networks with deep layers. However, with the expansion of layers, a neural network with deeper layers may not be able to learn some simple functions; at some epochs, the accuracy may begin to stagnate or even decrease, leading to model overfitting problems. In the case of limited sample training network model, the problem of overfitting becomes more serious. Therefore, residual learning block is introduced to train the network, and the reconstruction of neural network layer is simplified according to the residual function of input layer to avoid the problem of model overfitting. The residual learning process was shown in Figure 5, and the functional expression of residual learning is shown in Equation (6).

$$y = F(x, W) + x, \tag{6}$$

where x and y, respectively, represent the input and output vectors of the residual block, $F(\cdot)$ is the residual function, and W is the weight value inside the residual block.

In the forward propagation of the feature vector of CNN, in order to ensure the consistency of the dimensions of input and output vectors, if scale changes occur during the convolution operation inside the residual block, the transformation matrix W_s is used to adjust the dimension of the identity mapping part of input x. At the same time, considering that increasing the proportion of residual errors can reduce the difficulty of model training, coefficients are added on the basis of the original structure to adjust the value of the sum of output and residual errors, as shown in

$$y = \lambda F(x, W) + (1 - \lambda) W_s x. \tag{7}$$

3.4.2. Attention Mechanism. In this paper, attention mechanism CBAM is introduced into the network structure of metalearning to extract more expressive features, as shown in Figure 6. CBAM is composed of two different attention modules, and these two attention modules complement each other in defects and make the model pay more attention to feature representation and spatial definition. The two infer

```
Input:$D_{\text{origin}}, S_{\text{new}}, f_{\text{RA-CNN}}(\theta)$ inner and outer learning rate $\alpha$ and $\beta$, number of adaptation steps $m$, number of few-shot tasks $N$, metaepochs $M$
output: the classification accuracy of bearing fault diagnosis
1: Random initialize $\theta$
2: Split $D_{\text{origin}}$ into $D_{\text{meta\_train}}$ and $D_{\text{meta\_test}}$
    /*Metatraining stage                                        */
3:  forepoch $= 1, 2, \cdots M$do
4:      Randomly create n tasks from $\{D_{\text{meta\_train}}, S_{\text{new}}\}$
5:      for$i = 1, 2, \cdots n$do
6:          Create support set $D_s$ and query set $D_Q$ from task $i$
7:          Evaluate $L_{\text{tr}(\mathbf{T}_i)}(f_\theta)$ from Formula (11)
8:          Update $\theta$ to $\theta_i'$ using inner learning rate $\alpha$ and loss $L_{\text{tr}(\mathbf{T}_i)}(f_\theta)$
9:      end for
10:     Optimal parameter $\theta'$ update from Formula (13)
11: end for
    /*Metatesting stage                                         */
12: forepoch $= 1, 2, \cdots M$do
13:     Randomly create $n$-way, $k$-shot few-shot tasks from $D_{\text{meta\_test}}$
14:     repeat steps 5 to 8, to obtain the updated $\theta_i'$
15:     obtain optimal parameter $\theta'$
16:     predict fault class labels by $f(\theta')$ based on the $D_Q$ of $D_{\text{meta\_test}}$
17:     Compute the classification accuracy
18: end for
```

Algorithm 2:The implementation of the MOFD.

Table 1: RA-CNN parameters list.

No.	Layer name	Kernel number	Kernel size	Stride	Padding
1	Input	/	/	/	/
2	2D Conv1	32	3×3	1	Yes
3	ReLU	/	/	/	/
4	MaxPool	32	2×2	1	Yes
5	AvgPool_ channel	64	2×2	1	Yes
6	MaxPool_ channel	64	2×2	1	Yes
7	MLP	/	/	/	/
8	ADD	/	/	/	/
9	2D Conv2	64	3×3	1	Yes
10	ReLU	/	/	/	/
11	MaxPool	64	2×2	1	Yes
12	AvgPool_ spatial	64	2×2	1	Yes
13	MaxPool_ spatial	64	2×2	1	Yes
14	ADD	/	/	/	/
15	2D Conv3	128	7×7	1	Yes
16	2D Conv4	128	5×5	1	Yes
17	ReLU	/	/	/	/
18	MaxPool	128	2×2	1	Yes
19	FC		256, 96		
20	Softmax	/	96, 10	/	/

the attention weight from the channel dimension and the spatial dimension in series and then multiply it with the convolution result of the residual network to adjust the features, highlight the target features in the feature map, and enhance the recognition ability of the model for fault features.

Channel Attention Module (CAM) according to the dependence of different channels in the feature map on the response degree of the recognition target adjusts the feature map according to the different response degree and calculates the weight of each channel using multilayer perceptron. A channel with a high response degree indicates that it is similar to the recognition target and is assigned a higher weight; a channel with a low response degree indicates a large gap from the recognition target and is assigned a lower weight. The structure of CAM is shown in Figure 7, and the implementation steps are as follows: (1) the input feature map is subject to maximum pooling and average pooling, respectively. The average pooling realizes the compression of channel features, and the maximum pooling can collect the feature information of the target; (2) the pooled feature map is sent to the multilayer perceptron composed of the full connection layer, the average pooling layer, and the maximum pooling layer for parameter sharing; (3) the multilayer perceptron output results are multiplied and summed, and then, the channel attention feature map is output through sigmoid activation function. So, the calculation formula of channel attention feature map $\boldsymbol{M_c}(F)$ is shown in

$$\begin{aligned}\boldsymbol{M_c}(F) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{F}))) \\ &= \sigma\left(\boldsymbol{W_1}\left(\boldsymbol{W_0}\left(\boldsymbol{F}^{\boldsymbol{c}}_{\mathbf{avg}}\right)\right) + \boldsymbol{W_0}(\boldsymbol{W_1}(\boldsymbol{F}^{\boldsymbol{c}}_{\mathbf{max}}))\right),\end{aligned} \tag{8}$$

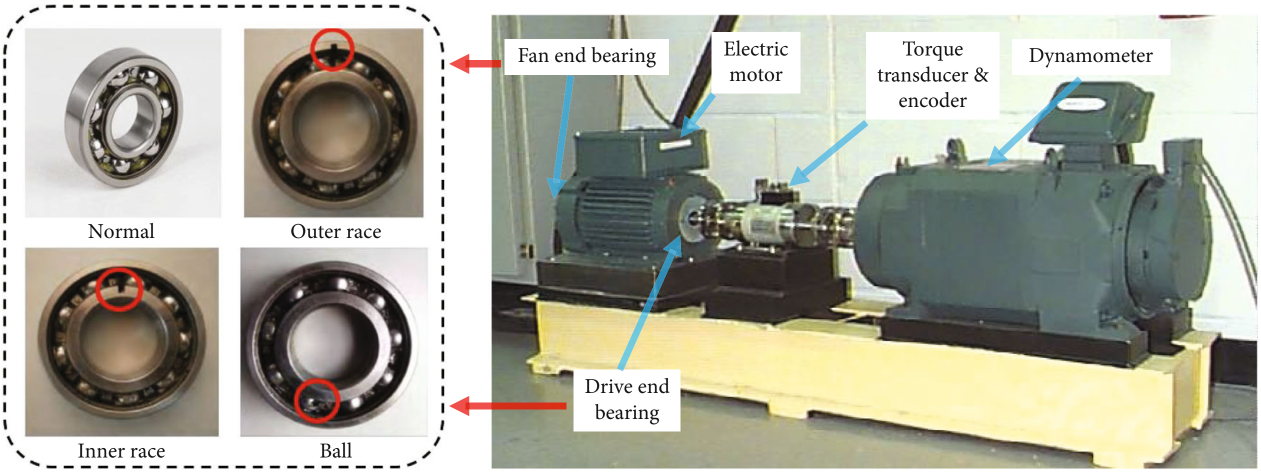


Figure 8: Bearing devices and failure types of CWRU experiments.

Table 2: The details of CWRU bearing datasets.

Fault state	Fault diameters	A	B	C	A′	B′	C′	D
Normal	0.000	720	720	720	720	720	720	100
Ball	0.007	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Inner race	0.007	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Outer race	0.007	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Ball	0.014	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Inner race	0.014	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Outer race	0.014	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Ball	0.021	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Inner race	0.021	36	72	144	36 (+684)	72 (+648)	144 (+576)	100
Outer race	0.021	36	72	144	36 (+684)	72 (+648)	144 (+576)	100

where σ denotes the sigmoid function, $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ represent the weight of shared network MLP. $\boldsymbol{F}$ represents an input feature map.

The neural network extracts the features for learning based on image details from the spatial dimension information of the feature map through the spatial attention module (SAM), to supplement CAM's output. Its structure is shown in Figure 7. The implementation steps are as follows: (1) the input feature maps are sequentially pooled and averaged and spliced the two feature maps with the channel; (2) the connected feature map is sent to the convolution layer for feature extraction, and then, the spatial attention feature map is finally output through the sigmoid activation function. In general, the calculation formula of channel attention feature map $\boldsymbol{M}_c(F)$ is shown in

$$\begin{aligned} \boldsymbol{M}_s(F) &= \sigma\left(f^{7\times7}([\mathrm{AvgPool}(\boldsymbol{F})\,;\mathrm{MaxPool}(\boldsymbol{F})])\right) \\ &= \sigma\left(f^{7\times7}\left(\boldsymbol{F}^S_{\mathrm{avg}}\,;\boldsymbol{F}^S_{max}\right)\right), \end{aligned} \tag{9}$$

where $f^{7\times7}(\cdot)$ indicates a convolution with the 7×7 filter size.

3.5. Metalearning Process. During the implementation of the proposed MOFD, the learning process is characterized by sequence of episodes, where each episode contains the metatraining and metatesting. RA-CNN as a basic learner is expressed as $f(\theta)$. Here, θ are weights of the neural network, and x is the feature vectors of the input. During metatraining, the metaparameter θ initializes the classification model $y = f(x\,;\theta_i)$, where $\theta_i = \theta$ and θ of each task T_i is different. T_i is divided into support sets D_S and query sets D_Q, and Adam is used to update parameters θ_i' based on D_S. Finally, the metaparameter θ_i is updated according to the learning of multiple metatasks in $\mathrm{T}_{\text{meta-train}}$ and obtains the optimal parameter θ' of metatraining as the shown in Equation (12).

$$\theta' \longleftarrow \theta - \alpha\nabla_\theta L_{\mathrm{tr(t)}}(f_\theta), \tag{10}$$

where m denotes the number of metatasks. α denotes the inner-loop learning rate.

Among them, the loss function $L_{\mathrm{tr}(\mathbf{T}_i)}(f_\theta)$ applies the loss function to measure the performance of the updated RA-CNN model in bearing fault classification with unbalanced

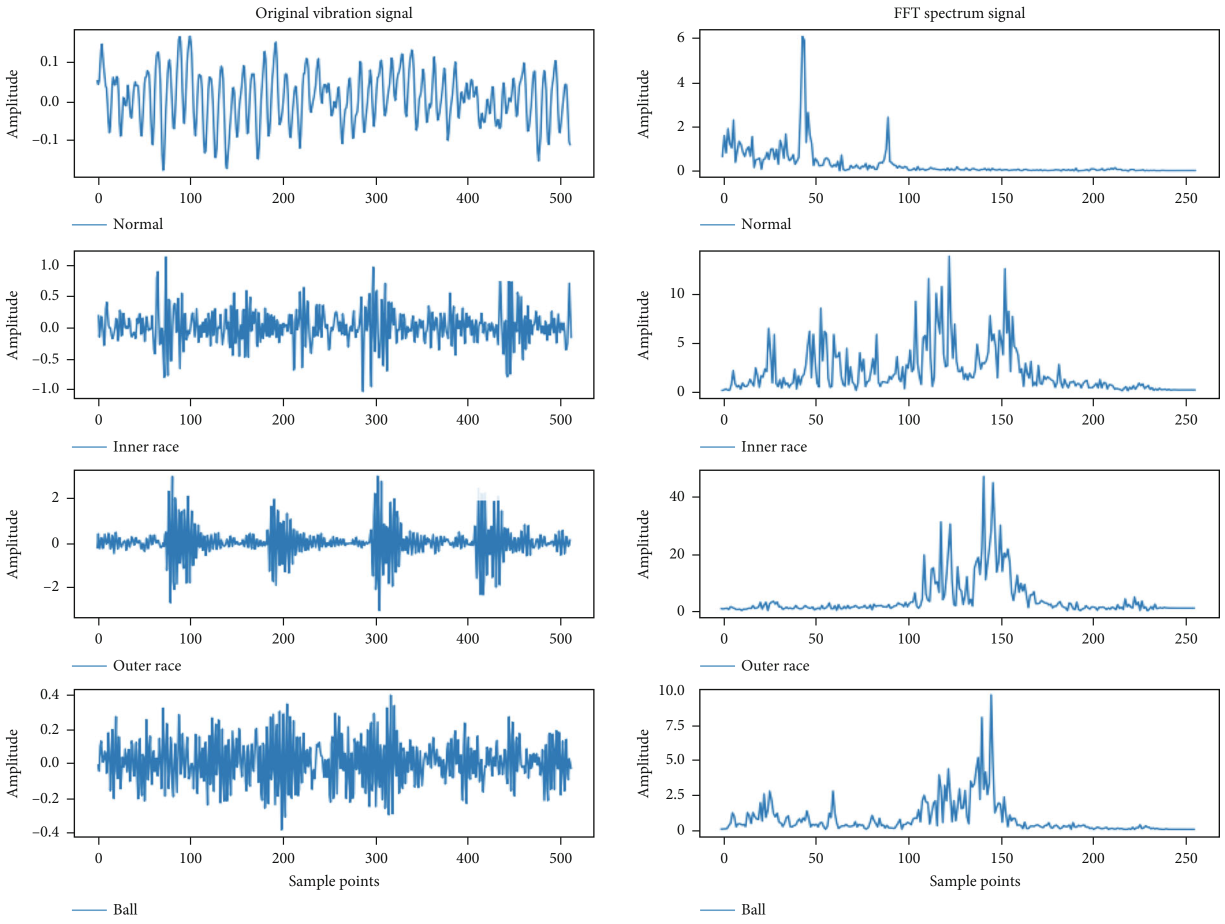


Figure 9: The visualization of original vibration and spectrum signals of CWRU bearing dataset in different health states.

data, and its mathematical form of updating the RA-CNN by the binary cross entropy loss function is shown in

$$L_{\mathrm{tr}(\mathbf{T}_i)}(f_\theta) = -\sum_{(x_i,y_i)\sim D_S} \left(y_i \log f_\theta(x_i) + (1-y_i)\log\left(1-f_\theta(x_i)\right)\right), \tag{11}$$

where (x_i, y_i) is the input vector and label pair. The training data is the support dataset $D_S(x_i, y_i)$ of each task T_i.

Then, the updated RA-CNN model $f(\theta')$ is used for outer loop optimization based on query set D_Q, and the total loss of all query set tasks takes the following form:

$$L_{\mathrm{tr}(\mathbf{T}_i)}(f_{\theta'}) = \sum_{(x_i,y_i)\sim D_Q} L_{\mathrm{tr}(\mathbf{T}_i)}\left(\theta' - \sum_{m=1}^{N} \alpha L_{\mathrm{tr}\left(\mathbf{T}_j\sim D_s\right)}(f_\theta)\right). \tag{12}$$

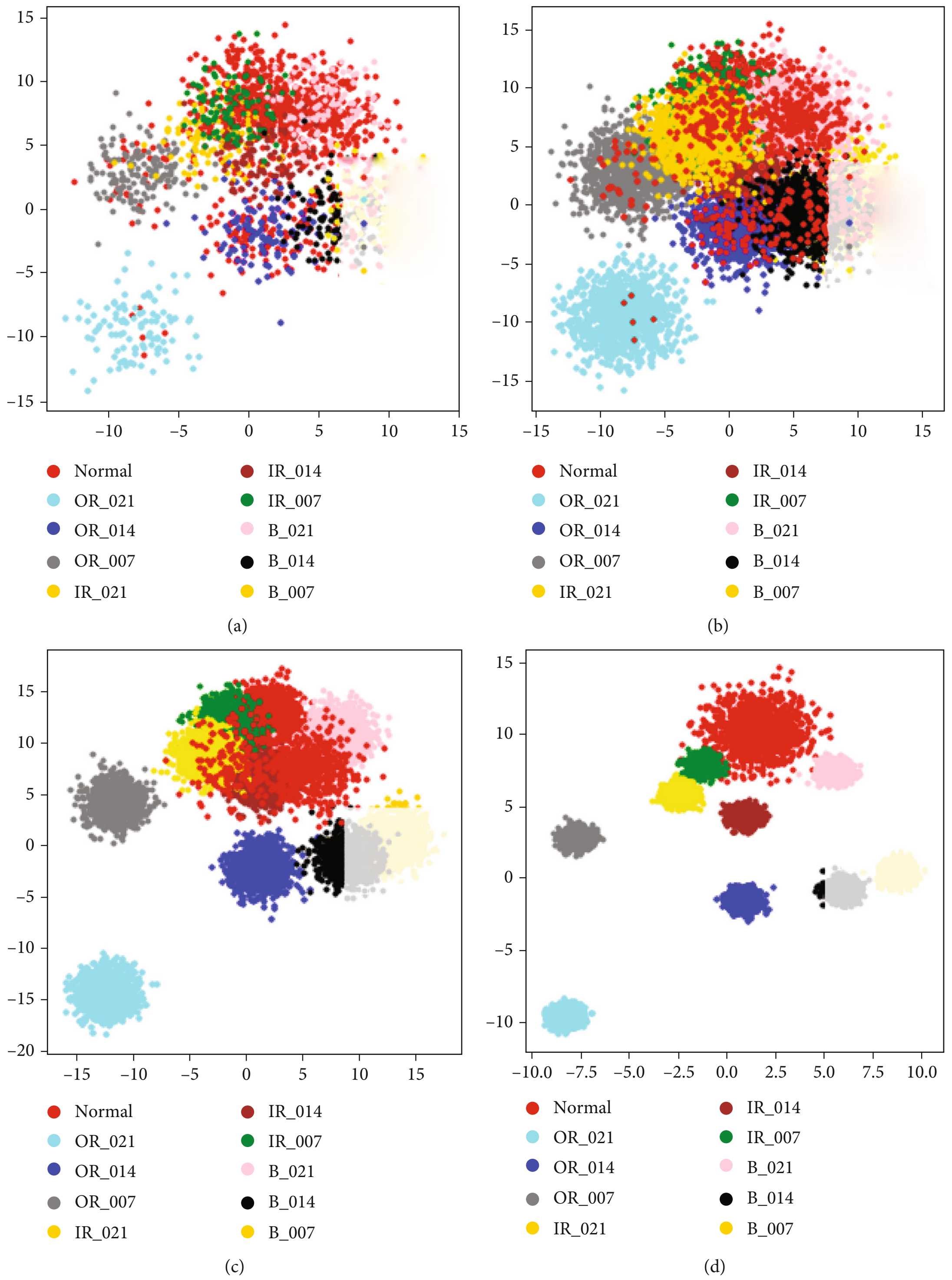
In order to accurately and rapidly classify bearings with unknown bearing data under finite samples and gradient step sizes, our metaobjective is to find the optimal model parameters in multiple tasks by minimizing total loss L_{MOFD}. Then, the metalearning parameters after inner loop optimization θ' are updated with the outer-loop learning rate β:

$$\theta^* \longleftarrow \theta' - \beta\nabla_{\theta'} L_{\mathrm{tr}(\mathbf{T}_i)}(f_{\theta'}). \tag{13}$$

In the metatesting stage, the RA-CNN model parameter θ^* can be fine-tuned based on the metatest support set to search for the best model parameter θ_{best} that can complete accurate fault classification on the query set. The pseudocode implementation of the proposed MOFD is outlined in Algorithm 2. The parameters of RA-CNN are shown in Table 1.

4. Experimental Analysis

This section takes the Case Western Reserve University (CWRU) bearing dataset as the experimental case to verify whether the MOFD model can effectively solve the bearing fault diagnosis problem under data imbalance. A 10-way diagnosis case is conducted on CWRU datasets. In the process of experimental setting, the relevant control variables are discussed in a certain range and finally compared with the current advanced related methods to determine the



(a)

(b)

(c)

(d)

FIGURE 10: Visualization with different oversampling techniques: (a) normal sample distribution; (b) SMOTE; (c) MWMOTE; (d) FSDA-SMOTE.

feasibility and progressiveness of the fault diagnosis model proposed in this paper.

4.1. Dataset Introduction. The CWRU dataset is a widely used bearing dataset for intelligent fault diagnosis for rotating machinery. As shown in Figure 8, the experimental equipment includes 1.5 kW motor, power tester, electronic controller, torque, and acceleration sensors, and single point damage failure on the inner, outer, and ball of the bearing is simulated, respectively, through EDM. The fault damage diameters are 0.18, 0.36, 0.54, and 0.71 inches, respectively. Vibration signals of normal bearings and damaged bearings with single point defects are collected at frequencies of 12 kHz or 48 kHz under four different motor loads. The experimental data collected

Table 3: The average accuracy (%) of imbalanced and rebalanced datasets by SVM.

Types	Imbalanced	Rebalanced (used data augmentation)				
		SMOTE	MWMOTE	LSTM	GAN	FSDA-SMOTE
Dataset A	81.54	86.23	85.34	88.17	**88.41**	86.37
Dataset B	83.31	87.14	88.18	90.67	90.58	**91.34**
Dataset C	88.07	89.75	90.33	95.50	94.12	**95.80**

Table 4: The parameter setting of STFT.

Window scale	Overlapping points	FFT points	Time–frequency resolution	Input image size
64	35	31	16 × 16	16 × 16(1)
128	103	31	16 × 16	16 × 16(2)
64	50	64	33 × 33	32 × 32(1)
128	116	63	32 × 32	32 × 32(2)
64	57	128	65 × 65	64 × 64(1)
128	122	128	65 × 65	64 × 64(2)
128	125	256	129 × 129	128 × 128

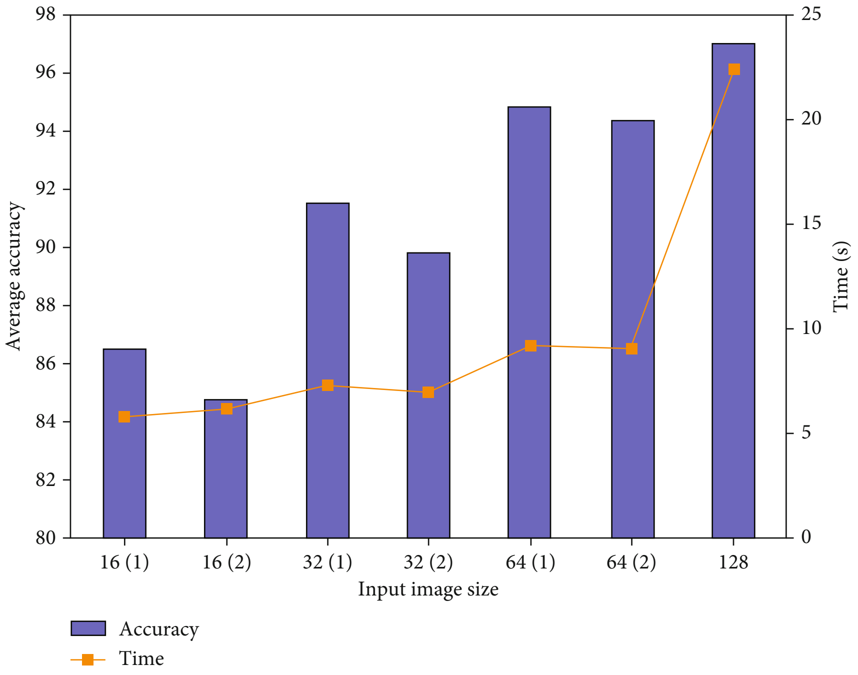
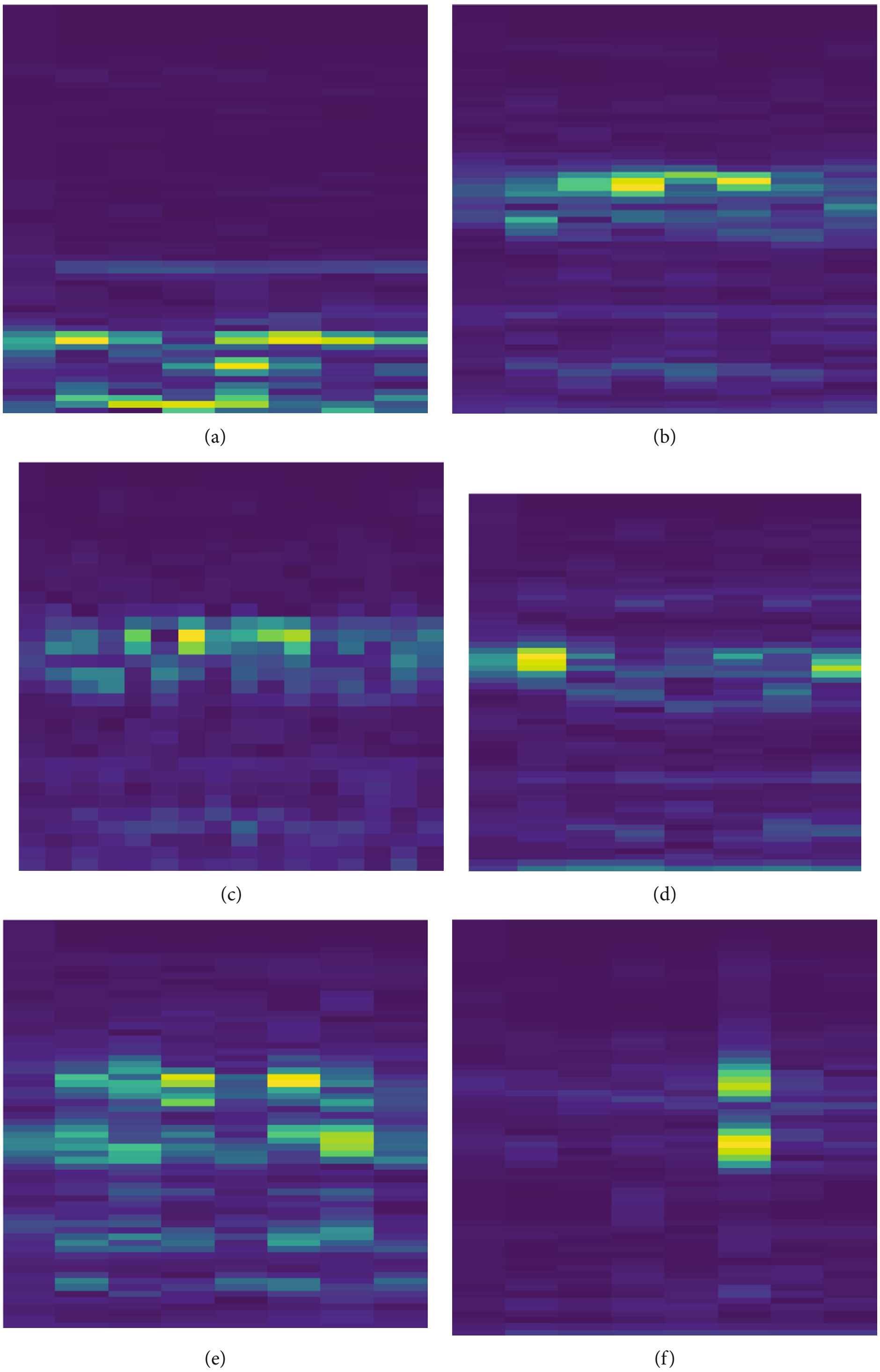


Figure 11: Average accuracy and time consumption with different input image size.

from the drive SKF-6205 and its sampling frequency is equivalent to 12 Khz. In Table 2, the bearing signal data is divided into 10 categories (one health state and 9 fault states) according to different fault diameters of bearings and fault states (normal, inner race, outer race, and ball).

4.2. Data Augmentation. As shown in Table 2, the imbalance ratio between each fault dataset and the normal dataset should be set at least less than 0.2 to simulate the experimental conditions of data imbalance. Dataset A is an unbalanced dataset with the imbalance ratio 0.05, which has 720 normal samples and 36 samples for each the other 9 fault categories; dataset B is an unbalanced dataset with the imbalance ratio 0.1, which has 720 normal samples and 72 samples for each the other 9 fault categories; and dataset C is an unbalanced dataset with the imbalance ratio 0.2, which has 720 normal samples and 144 samples for each other 9 fault categories. FSDA-SMOTE uses imbalanced datasets A, B, and C to enhance data and obtain rebalanced datasets A$'$, B$'$ and C$'$, the number of samples for every category is 720. In addition,



(a) (b)

(c) (d)

(e) (f)

Figure 12: Continued.

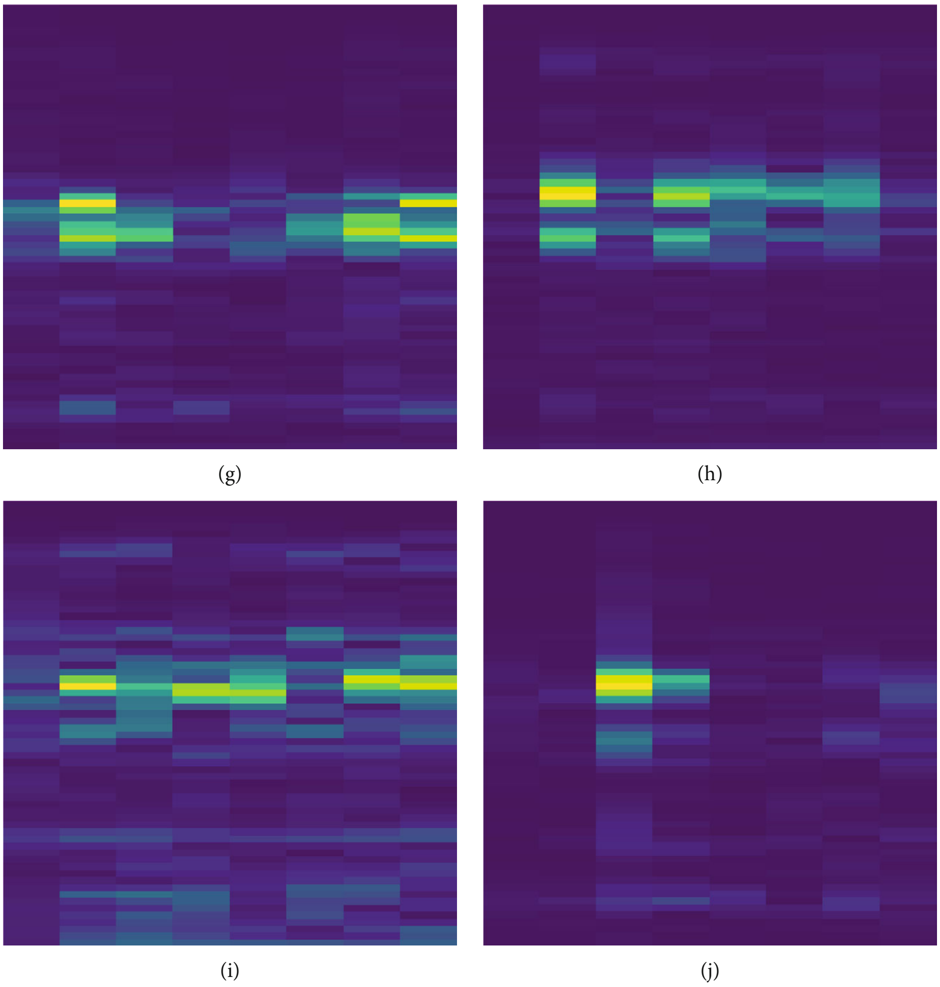
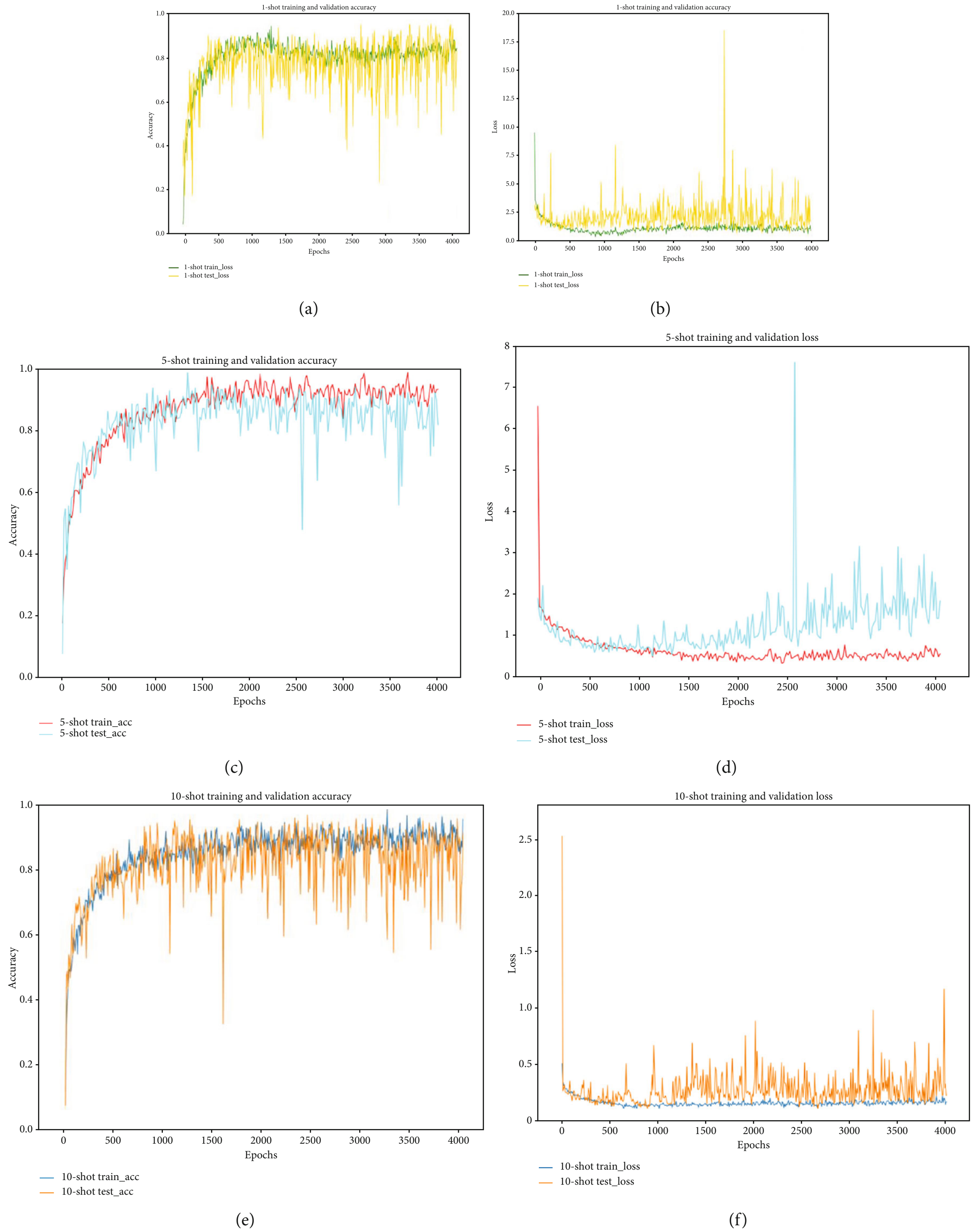


Figure 12: Time-frequency images of CWRU bearing dataset under different fault states: (a) normal; (b) B_007; (c) B_014; (d) B_021; (e) IR_007; (f) IR_014; (g) IR_021; (h) OR_007; (i) OR_014; (j) OR_021.

each category has a test set with 720 samples to verify the trained fault diagnosis model. Dataset D is a test set with 100 samples in each category and a total of 1000 samples.

In the stage of sampling data, the original signal data is segmented in an average order. There is no overlap between these segments, and the length of segments is 512. The frequency spectrum signal is obtained by short-time Fourier transform of the fragment, which has more important physical information and contains more useful fault diagnosis information. In Figure 9, the spectrum signals under different fault states with the fault diameter of 0.007 inch are introduced into the spectrum space and sample fault spectrum signals. Among them, the left part of Figure 9, respectively, represents the original bearing vibration information under different fault states with the speed of 1797 RPM, states with the speed of 1797 RPM, and the right part of Figure 9, respectively, represents the spectrum signals after short-time Fourier transform under different fault states. In order to test the effectiveness of FSDA-SMOTE algorithm on limited imbalanced bearing fault diagnosis, several other data augmentation methods are compared in the paper, including 2 oversampling algorithms (SMOTE, MWMOTE) and 2 network generation models (LSTM [15], GAN [31]). That is, the FSDA-SMOTE algorithm in data augmentation stage of Figure 2 is replaced with the other data augmentation methods. The parameters of oversampling algorithms are set as follows: data augmentation algorithms made the size of the fault instances sampling the same as the normal instances; the k nearest neighbor parameters of SMOTE and MWMOTE is 5. MWMOTE and FSDA-SMOTE's number of clusters is 9. Additionally, in order to highlight the differences in the spatial distribution of new samples generated by different sampling algorithms, dataset A in Table 2 is taken as an example with visualized application T-SNE of each sampling algorithm.

Figure 10 shows the projection distribution of dataset A and its synthetic samples in two-dimensional space, which is visible through the sample density areas under different color labels to further visually illustrate the distribution differences between different sample categories. However, some samples also infiltrate into each other outside the sample dense area,



(a) (b) (c) (d) (e) (f)

Figure 13: The metalearning process of MOFD with different shots.

leading to fuzzy boundary between classes and even abnormal data. At this time, it is often necessary to consider whether to treat them as discrete points. The processing result of different sampling algorithms for dataset A is shown in Figure 10. Figure 10(a) shows the original distribution of various samples in the feature space of unbalanced dataset A, and the data distribution of each category is chaotic, so it is impossible to set the boundary between classes well to separate each category sample. Figure 10(b) uses SMOTE algorithm to enhance the minority fault data of dataset A.

However, the synthesizing data based on the traditional SMOTE is easy to receive the interference of noise data and affect the real distribution of failure data, thus reducing the quality of the synthesized sample. In Figure 10(c), although the MWMOTE can optimize minority boundaries by denoising and assigning information weight to minority failure

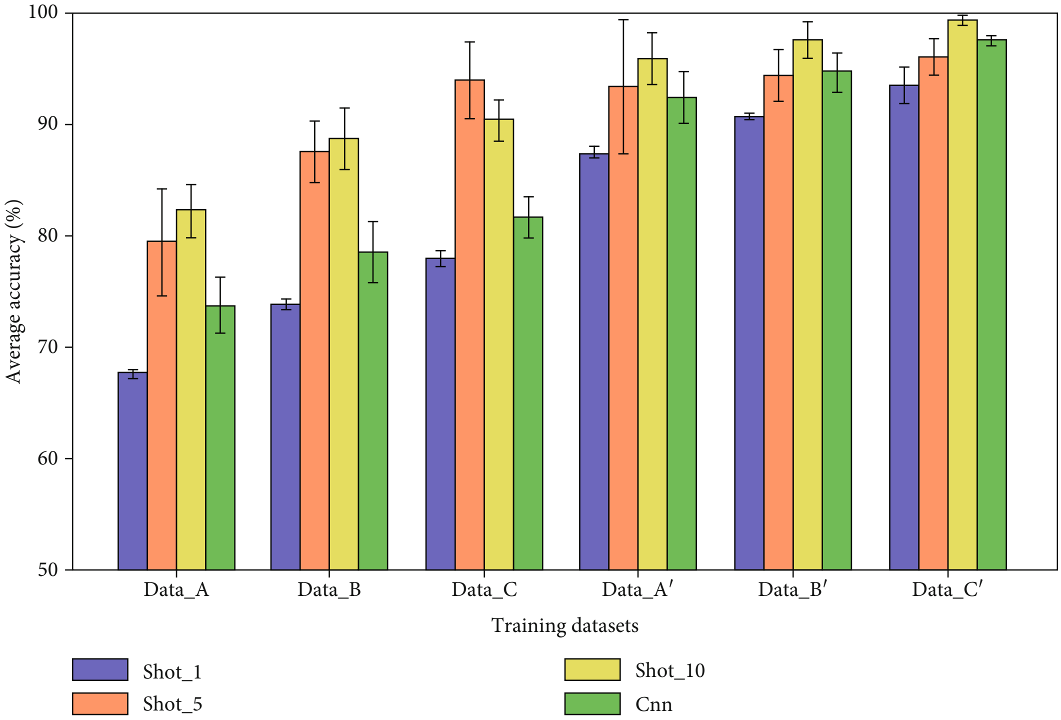


FIGURE 14: The compared results of metalearning (1\5\10-shot) and baseline CNN with CWRU datasets.

TABLE 5: The accuracy of fault diagnosis models on CWRU dataset A.

Types	Imbalanced	Rebalanced (used data augmentation)				
	/	SMOTE	MWMOTE	LSTM	GAN	FSDA-SMOTE
VMD-SVM	68.57	77.32	83.12	81.17	82.39	86.54
SDAE-DNN	75.96	80.13	86.26	87.67	90.58	92.34
CNN	73.74	80.48	85.71	82.36	87.22	91.03
Res-Net	76.92	80.32	83.72	79.90	88.27	90.26
Instance-TL	80.14	82.75	86.56	90.53	83.17	89.08
Siamese network	83.65	88.19	91.14	87.07	90.34	92.16
Proposed	85.76	89.33	90.03	80.41	91.12	**95.71**

samples, the difference in weight allocation will also result in partial fault sample clusters (such as IR_007, B_021) being ignored, thus failing to correctly learn the fault samples within the group boundary. Figure 10(d) uses the FSDA-SMOTE method proposed in this paper. In this method, the similarity between synthetic samples and original samples is enhanced by defining the spatial density of samples in minority fault samples cluster. Due to the spatial density of the sample generated in Figure 10(d), the sample synthesis range is closer to the center of the minority sample cluster. The boundary between classes becomes very clear, which is beneficial to improve the reduction degree of synthetic samples.

As shown in Table 3, unbalanced data A, B and C are balanced with different imbalance ratios by using different data augmentation methods and rebalanced data A′, B′, and C′ are obtained. SVM, as a classification model with outstanding stability, is introduced to carry out fault diagnosis and classification experiments on unbalanced and rebalanced datasets, respectively. The experiment is repeated for 10 times, and the average accuracy of 10 times fault classification is taken as the final judgment result. It can be seen from Table 3 that the unbalanced training data will obviously affect the model's classification performance, and the influence of such problems can be effectively reduced by expanding the fault data (the minority of the training data) through data augmentation method. Specifically, FSDA-SMOTE generated rebalanced dataset A′ is about 4.83%, 8.03%, and 7.73% higher than imbalanced dataset A; SMOTE is about 4.69%, 3.83%, and 1.68% higher; MWMOTE increases by 3.8%, 4.87%, and 2.26%. LSTM increases by 1.94%, 4.44%, and 7.43%, and GAN increases by 6.87%, 7.37%, and 4.32%. The above experimental data prove that the fault identification ability of the rebalanced dataset obtained by data augmentation is better than the unbalanced dataset, which indicates the feasibility and effectiveness of data augmentation method for solving the problem of low accuracy of fault identification under data imbalance. In addition, the average accuracy of rebalanced datasets with different imbalanced ratios is also different in bearing fault diagnosis by using the same data augmentation method. From dataset A to dataset C, with the gradual increase of imbalance ratio, the average accuracy of the generated complex balanced dataset is higher. From this conclusion, it can be inferred that increasing the number of training data is conducive to better learning fault information of the diagnostic model, and it also proves the necessity of data enhancement in the case of data imbalance. Finally, FSDA-SMOTE gives better average accuracy on three imbalance ratios than the other data augmentation methods. In summary, the experiment results prove the feasibility of FSDA-SMOTE synthetic sample and the reliability of sample quality based on the CWRU's dataset.

4.3. Data Preprocessing. As shown in Figure 3, the rebalanced dataset is converted into a one-dimensional time-domain signal into a spectrogram image through the STFT. In the process of STFT, different window scales, overlapping points, and Fourier points are set to obtain different time-frequency resolutions. To facilitate pooling in the CNN, the last row and last column are removed when the time-frequency resolution is odd. The rebalanced dataset A' based on FSDA-SMOTE is selected as the training dataset for RA-CNN classifier testing to obtain the image input size at the optimal time-frequency resolution. The RA-CNN input image sizes are listed in Table 4.

In order to avoid contingency, 20 experiments are repeated to obtain the average test accuracy and time consumption. The diagnosis results are shown in Figure 11. The test accuracy is greater than 85%, which indicates that feature extraction based on 2D time-frequency image input can extract deeper fault characterization features. When the input image size is 64×64, the fault identification average accuracy is over 94%. However, with the increase of image scale, the time consumed for fault diagnosis also increases significantly. Considering the test accuracy and time cost comprehensively, the window function of scale 128 and the time-frequency image scale of 64×64 are the default parameters of STFT. The time-frequency images of bearing vibration signals under 10 different health states after short-time-Fourier transform are shown in Figure 12.

4.4. Experiment Setting. Metalearning experiments apply a novel convolutional neural network model RA-CNN with 4 convolutional layers as the backbone structure to train a metalearning classifier for fault diagnosis under data imbalance, which also contains the residual learning module and the temporal and spatial attention module. During the process of metalearning experiment, Adam optimizer is used (learning rate $\alpha = 0.01$) for inner loop optimization and SGD with momentum (momentum = 0.9) for outer optimization. The metalearning rate β used for metaupdate step (locate on line 10 of Algorithm 2) has an initial value of 0.1 and changes dynamically as the metaiteration increases. The number of subtasks N of $D_{\text{meta\_train}}$ and $D_{\text{meta\_test}}$ are 800 and 200, respectively. The adaptation steps (m) for metalearning are 50, and all the metalearning experiments are allowed to run for 4000 metaepochs before fine-tuning. In order to facilitate a fair comparison with other fault diagnosis models, the network conditions and hyperparameters used in the experiment should not be changed as much as possible when it is not necessary. The input image is resized to 64×64 and normalized the pixel value between [0, 1]. The early stop strategy is used to avoid overfitting and obtain best model. In order to avoid occasionality that an experiment might occur, we repeated 20 experiments for each verification process, and the average results were as close to the real diagnosis result of the model as possible. All the models proposed in this paper are tested in the same operating environment; the specific contents are as follows: Intel Xeon(R) CPU E3-1231 V3 @ 3.40 GHz, NVIDIA GeForce RTX2080Ti, Ubuntu18.04, Python3.6, CUDA 11.2, and TensorFlow 2.7.

4.5. Experimental Result Analysis. In order to further demonstrate the feasibility and superiority of the MOFD's fault classification ability under data imbalance, six baseline models are proposed to compare the fault diagnosis results with the MOFD method, which includes VMD-SVM [7], SDAE-DNN [8], CNN [10], Res-Net [12], Instance-TL [35], and Siamese network [41]. Considering the differences between traditional model and metalearning model in data training methods, MOFD and Siamese networks based on 10-way classification tasks are proposed, and different data augmentation techniques are used for 1-shot, 5-shot, 10-shot learning. So, the accuracy and loss curves of 1, 5, and 10 shots are drawn in dataset A, and the variation trend of MOFD during metatraining and self-testing is observed metatesting. In the first 2000 iterations, the accuracy of metatraining and metatesting improves significantly, then slows down, and finally, approaches 100%. Figure 13(a) shows that the MLFD model is well trained without overfitting. Figure 13(b) shows the loss reduction process of 1-shot, and the model tends to converge after 1200 iterations. Figures 13(c) and 13(d) illustrate the changes of 5-shot and in Figures 13(e) and 13(f) is 10-shot, whose process are similar to 1-shot, but the initial loss of 10-shot is a lower value than others, because the 10-shot can obtain more bearing fault information to train the model better than 1-shot or 5-shot. Therefore, we compared the fault classification results of CNN, MOFD1, MOFD5, and MOFD10 under different metatask sets, respectively, to verify the influence of the amount of metatask set data on model fault feature learning. Figure 14 shows that the accuracy of MOFD 10-shot performed better than MOFD1-shot in all 6 experiments under variable datasets, and it is worth emphasizing that FSDA-SMOTE data augmentation is used to solve the dataset imbalanced in all of six experiments.

After determining the training method of metalearning, it is concluded that the model based on metalearning strategy can better identify and classify faults by comparing MOFD and baseline CNN without metalearning, and the results are shown in Figure 14. It is obvious that the diagnostic performance of MOFD on six datasets is very good, but the performance of CNN will fluctuate according to the degree of data imbalance. Especially under the experiment with dataset A, the accuracy of the CNN is only 73.74%, which is 9.52% lower than MOFD10-shot. This is because the imbalance of datasets will shift the classification boundary and change the model's perception of the original data distribution. MOFD has the ability to obtain prior knowledge of general bearing faults through the known fault data and apply them to fault diagnosis of data imbalanced. It is difficult for baseline CNN to effectively obtain enough fault features for bearing fault diagnosis under data imbalance based on its simple network structure and limited samples.

In the end, the seven methods proposed above will be compared with each other through the CWRU bearing dataset. Table 5 shows the fault diagnosis results of each fault diagnosis model under different data augmentation methods. The fault diagnosis models used uniformly dataset A with the imbalanced ratio 0.05 as the training data to magnify the internal imbalance of the dataset.

As shown in Table 5, the average accuracy of almost all models on balanced datasets is higher than that on unbalanced datasets. The transfer learning model based on instances (instance-TL) has the least improvement (2.61%), and the VMD-SVM model has the most improvement (17.97%). In addition, the experimental results show that among these data augmentation methods, FSDA-SMOTE generated the fault data effectively improves the diagnostic performance of models and gets the best or above average in the comparison of multiple methods and models. Therefore, the MOFD model based on metalearning framework proposed can effectively solve the problem of bearing fault diagnosis under data imbalance.

5. Conclusion

In this paper, a method based on metalearning framework is proposed to describe the fault diagnosis problem of data imbalance as a few-shot learning problem based on image classification. Based on the time-series correlation of bearing vibration signals and the time-invariance of frequency domain features, time-frequency transform technology STFT is used to learn the deep features of bearing signals from the image level, and FSDA-SMOTE data augmentation method clusters minority samples to define samples' density in the class and generates high quality fault data to enhance the diversity of training data. The effectiveness of this method is verified by comparing with several data augmentation techniques. In addition, a metaoptimization fault diagnosis method (MOFD) for k-way N-shot metalearning training method is proposed. Compared with the traditional use of CNN as a metalearner, MOFD uses RA-CNN based on residual learning and time-space attention mechanism for fault feature extraction, which not only avoids overfitting due to limited fault training data but also improves the ability of metalearner network to extract key fault features. Experimental results based on public CWRU datasets show that the MOFD method can significantly improve the data imbalance problem, and RA-CNN can obtain more representative fault feature information under the limited data of metatask. Besides, it has better fault classification performance and good robustness and generalization. Therefore, the proposed method has a certain reference value in solving the bearing fault diagnosis problem under data imbalance.

Data Availability

The dataset is available upon request. Experimental data of the paper uses the bearing datasets from Case Western Reserve University (CWRU) (https://engineering.case.edu/bearingdatacenter/welcome).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is funded by the Natural Science Foundation of Hunan Province (nos. 2022JJ50002, 2020JJ6086, and 2021JJ50049)

References

[1] M. Ma, C. Sun, and X. F. Chen, "Deep coupling autoencoder for fault diagnosis with multimodal sensory data," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 3, pp. 1137–1145, 2019.

[2] Y. J. Wang, F. M. Sun, and X. H. Li, "Compound dimensionality reduction based multi-dynamic kernel principal component analysis monitoring method for batch process with large-scale data sets," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 1, pp. 471–480, 2020.

[3] X. Wang, Y. Qin, Y. Wang, S. Xiang, and H. Chen, "_ReLTanh_ : an activation function with vanishing gradient resistance for SAE-based DNNs and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 363, pp. 88–98, 2019.

[4] S. S. Udmale and S. K. Singh, "A mechanical data analysis using kurtogram and extreme learning machine," *Neural Computing and Applications*, vol. 32, no. 8, pp. 3789–3801, 2020.

[5] B. Cai, K. Hao, Z. Wang et al., "Data-driven early fault diagnostic methodology of permanent magnet synchronous motor," *Expert Systems with Applications*, vol. 177, pp. 4157–4174, 2021.

[6] Y. Xue, D. Y. Dou, and J. G. Yang, "Multi-fault diagnosis of rotating machinery based on deep convolution neural network and support vector machine," *Measurement*, vol. 156, no. 7, p. 107571, 2020.

[7] H. Li, Y. Zhao, Y. Zhang, and E. Zio, "Remaining useful life prediction using multi-scale deep convolutional neural network," *Applied Soft Computing Journal*, vol. 89, p. 106113, 2020.

[8] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: a review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, 2020.

[9] B. P. Cai, X. T. Sun, J. X. Wang et al., "Fault detection and diagnostic method of diesel engine by combining rule-based algorithm and BNs/BPNNs," *Journal of Manufacturing Systems*, vol. 57, pp. 148–157, 2020.

[10] M. Ye, X. Yan, and M. Jia, "Rolling bearing fault diagnosis based on VMD-MPE and PSO-SVM," *Entropy*, vol. 23, no. 6, p. 762, 2021.

[11] C. Lu, Z. Y. Wang, W. L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Processing*, vol. 130, pp. 377–388, 2017.

[12] H. Li, J. Huang, and S. Ji, "Bearing fault diagnosis with a feature fusion method based on an ensemble convolutional neural network and deep neural network," *Sensors*, vol. 19, no. 9, p. 2034, 2019.

[13] S. Li, W. Yang, A. Zhang et al., "A novel method of bearing fault diagnosis in time-frequency graphs using InceptionResnet and deformable convolution networks," *IEEE Access*, vol. 8, pp. 92743–92753, 2020.

[14] D. Peng, Z. Liu, H. Wang, Y. Qin, and L. Jia, "A novel deeper one-dimensional CNN with residual learning for fault diagnosis of wheelset bearings in high-speed trains," *IEEE Access*, vol. 7, pp. 10278–10293, 2018.

[15] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2017.

[16] P. Yao, S. Yang, and P. Li, "Fault diagnosis based on RseNet-LSTM for industrial process," in *2021 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 728–732, Chongqing, China, 2021.

[17] K. N. Ravikumar, A. Yadav, H. Kumar, K. V. Gangadharan, and A. V. Narasimhadhan, "Gearbox fault diagnosis based on multi-scale deep residual learning and stacked LSTM model," *Measurement*, vol. 186, p. 110099, 2021.

[18] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the conference on computer vision and pattern recognition (CVPR)*, pp. 3156–3164, Honolulu, USA, 2017.

[19] R. Ye, W. Wang, Y. Ren, and K. Zhang, "Bearing fault detection based on convolutional self-attention mechanism," in *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pp. 869–873, Xiamen, China, 2020.

[20] H. Men, H. Yuan, Y. Shi, M. Liu, and M. Liu, "A residual network with attention module for hyperspectral information of recognition to trace the origin of rice," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 263, p. 120155, 2021.

[21] A. Zhang, S. Li, Y. Cui, W. Yang, and J. Hu, "Limited data rolling bearing fault diagnosis with few-shot learning," *IEEE Access*, vol. 7, pp. 110895–110904, 2019.

[22] N. U. Maulidevi and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3413–3423, 2022.

[23] C. Zhang, K. Peng, and J. Dong, "A full-condition monitoring method for nonstationary dynamic chemical processes with cointegration and slow feature analysis," *AICHE Journal*, vol. 64, no. 5, pp. 1662–1681, 2018.

[24] W. Yu and C. Zhao, "Sparse exponential discriminant analysis and its application to fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5931–5940, 2018.

[25] B. P. Cai, Z. D. Wang, H. M. Zhu et al., "Artificial intelligence enhanced two-stage hybrid fault prognosis methodology of PMSM," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7262–7273, 2022.

[26] Z. Chai and C. Zhao, "Enhanced random forest with concurrent analysis of static and dynamic nodes for industrial fault classification," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 54–66, 2020.

[27] T. Hu, T. Tang, R. Lin, M. Chen, S. Han, and J. Wu, "A simple data augmentation algorithm and a self-adaptive convolutional architecture for few-shot fault diagnosis under different working conditions," *Measurement*, vol. 156, article 107539, 2020.

[28] X. Li, W. Zhang, Q. Ding, and J. Q. Sun, "Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation," *Journal of Intelligent Manufacturing*, vol. 31, no. 2, pp. 433–452, 2020.

[29] Y. Zhang, X. Li, L. Gao, L. Wang, and L. Wen, "Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning," *Journal of Manufacturing Systems*, vol. 48, pp. 34–50, 2018.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[31] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New imbalanced fault diagnosis framework based on Cluster-MWMOTE and MFO-optimized LS-SVM using limited and complex bearing data," *Engineering Applications of Artificial Intelligence*, vol. 96, article 103966, 2020.

[32] H. Yi, Q. Jiang, X. Yan, and B. Wang, "Imbalanced classification based on minority clustering synthetic minority oversampling technique with wind turbine fault detection application," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 5867–5875, 2021.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[34] B. Zhao and Q. Yuan, "Improved generative adversarial network for vibration-based fault diagnosis with imbalanced data," *Measurement*, vol. 169, article 108522, 2021.

[35] Y. O. Lee, J. Jo, and J. Hwang, "Application of deep neural network and generative adversarial network to industrial maintenance: a case study of induction motor fault detection," in *2017 IEEE international conference on big data*, pp. 3248–3253, Boston, MA, USA, 2017.

[36] W. Mao, Y. Liu, L. Ding, and Y. Li, "Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: a comparative study," *IEEE Access*, vol. 7, pp. 9515–9530, 2019.

[37] J. Liu, C. Zhang, and X. Jiang, "Imbalanced fault diagnosis of rolling bearing using improved MsR-GAN and feature enhancement-driven CapsNet," *Mechanical Systems and Signal Processing*, vol. 168, article 108664, 2022.

[38] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," https://arxiv.org/abs/1701.04862.

[39] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions on systems, man, and cybernetics: systems*, vol. 49, no. 1, pp. 136–144, 2019.

[40] Z. Wu, H. Jiang, K. Zhao, and X. Li, "An adaptive deep transfer learning method for bearing fault diagnosis," *Measurement*, vol. 151, article 107227, 2017.

[41] Q. Sun, Y. Liu, T. S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 403–412, Hawaii, USA, 2017.

[42] C. Li, S. Li, A. Zhang, Q. He, Z. Liao, and J. Hu, "Meta-learning for few-shot bearing fault diagnosis under complex working conditions," *Neurocomputing*, vol. 439, pp. 197–211, 2021.

[43] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, Sydney, Australia, 2017.

[44] C. Yu, Y. Ning, Y. Qin, W. Su, and X. Zhao, "Multi-label fault diagnosis of rolling bearing based on meta-learning," *Neural Computing and Applications*, vol. 33, no. 10, pp. 5393–5407, 2021.

[45] C. Wang and Z. Xu, "An intelligent fault diagnosis model based on deep neural network for few-shot fault diagnosis," *Neurocomputing*, vol. 456, pp. 550–562, 2021.

[46] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3637–3645, 2018.

[47] Y. Chen, Y. Hong, J. Long, Z. Yang, Y. Huang, and C. Li, "Few shot learning for novel fault diagnosis with an improved prototypical network," in *2021 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, Nanjing, China, 2021.