IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 June 2024, accepted 5 August 2024, date of publication 16 August 2024, date of current version 27 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444828

RESEARCH ARTICLE

Dual-Path Fault Diagnosis of Small Sample for Mechanical Systems Based on Multiple Attention Mechanisms

XIN LI, MEILING ZHANG, AND HUBO GUO
School of Information Engineering, Shenyang University, Shenyang 110044, China

Corresponding author: Xin Li (li_xin@syu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62073205.

ABSTRACT For fault diagnosis, it is important to effectively leverage the inherent characteristics of small datasets, but it is rarely considered in many the existing deep learning approaches. To this end, a dual-path network model based on multiple attention mechanisms is proposed in this work. The proposed model enriches the features of small sample by combining one-dimensional (1-D) frequency signals with two-dimensional (2-D) time-frequency images. A 1-D attention mechanism is applied in the 1-D frequency extraction path to focus classification on sensitive frequency information, while an improved global attention mechanism is added to the 2-D feature extraction path, which refines the key features and reduces interference resulting from noise in the image data. Moreover, the stability of the learning process is enhanced through the application of a combinatorial loss function composed of label smoothing regularization and gradient harmonizing mechanism loss functions. Finally, the diagnostic performance of the proposed method is validated using two different public fault diagnosis datasets in comparison with the state-of-the-art methods.

INDEX TERMS Vibration signal, attention mechanism, fault diagnosis, wavelet transform, small sample.

I. INTRODUCTION

The real-time diagnosis of bearing faults in mechanical systems based on vibration signals is imperative for ensuring safety and high production in industry [1], [2]. However, the increasing complexity and variability of modern working conditions have in-creased the frequency and the range of faults encountered by mechanical equipment. These conditions have necessitated the development of fault diagnosis methods with an increasing degree of intelligence and sophistication.

In recent years, data-driven methods relying on large-scale historical data in conjunction with artificial intelligence techniques, such as artificial neural networks (ANNs) [3], random forest (RF) classifiers [4], and support vector machine (SVM) [5], have been rapidly developed and now include many successful applications. Huo et al. [6] developed a hybrid technology with SVM for identifying faults of rolling bearings. Nevertheless, these conventional artificial intelligence methods rely on complicated processes for the manual feature extraction and reduction, and their performances are subject to the quality of the features considered and the complexity of the data. Effectively extracting hidden high-dimensional features becomes very difficult when the data are complex, nonlinear, or involve several dimensions. These issues represent significant limitations in practical applications.

These issues associated with conventional data-driven fault diagnosis methodologies have been addressed via a variety of deep learning technologies [7], [8], [9], [10], [11], such as convolutional neural networks (CNNs) [12], [13], [14] and generative adversarial networks [15], Generative adversarial networks [16], Deep belief networks [17], Recurrent neural networks [18], etc. Deep learning-based approaches can be classified into two types according to whether the input signal is a one-dimensional (1-D) signal, or a two-dimensional (2-D) signal. In terms of 1-D vibration signals, Zhao et al. [19] applied an enhanced gated recurrent neural network

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

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

to recognize faults of rolling bearings for one-dimensional input. Zhao et al. [20] designed a novel deep network with soft thresholding for improving the anti-noise property of feature extraction. Miao et al. [21] presented a sparse representation layer in a convolutional structure to suppress hidden data noise and learn 1-D high quality features directly. The use of 2-D image-based data transforms the signal classification problem into one of image classification. For example, Ding and He [22] employed 2-D wavelet packet energy images in the fault diagnosis model to improve the utilization of feature information. In addition, attention mechanisms have generated considerable interest at present for conducting fault diagnosis in deep learning frameworks [23]. For example, Zhang et al. [24] proposed a capsule network with attention mechanism that extracted dual-scale features from 2-D images. Huang et al. [25] developed a multi-scale CNN method combining channel attention that enhanced diagnostic performance by focusing on sensitive features in the learning process. Jia et al. [26] proposed an improved CNN model to reduce noise in the 1-D vibration signals employed in fault diagnosis.

The increasing range of faults encountered by mechanical systems in recent years has also generated considerable interest in the need for conducting mechanical fault diagnosis using small training datasets [26]. A number of methods have been commonly applied for this purpose, such as data augmentation [27], transfer learning [28], and meta-learning [29]. In general, these methods address the sparsity of available data pertaining to a specific fault condition in different ways. One common type of approach seeks to generate new training data based on the small dataset. In contrast, methods like transfer learning are applicable in a setting where only small datasets are available for a specific condition of interest but considerable data exists for a related condition. Accordingly, Feng et al. [30] constructed a gradient-penalized GAN model based on a multi-module learning strategy to produce failure signals of various states [31]. Yang et al. [32] developed an effective trainable network based on transfer learning. Han et al. [33] applied a meta-learning model for identifying roller bearing faults under variable speed conditions. Another method for addressing the issues associated with small datasets that has demonstrated excellent performance involves the use of modeling. Liu et al. [34] used a novel fault classification technique that applied an attention mechanism to effectively model fused spatiotemporal features of small vibration data. However, despite the significant advances in deep learning-based fault diagnosis, the ability to make good use of the inherent features of small datasets requires further development.

The present work addresses this issue by proposing a systematic method denoted as dual-path fault diagnosis based on multiple attention mechanisms (DPMAM) that not only fuses spatiotemporal features and small vibration data effectively, but also applies separate attention mechanisms to focus the classification network onto the most sensitive information available. Moreover, the DPMAM method performs well under variable speed and field noise conditions. The primary contributions of this work can be described as follows.

(1) The dual-path network structure enriches the features of small samples by combining 1-D frequency signal inputs in one path, denoted as the 1-D feature extraction module, which applies the fast Fourier transform (FFT) to generate feature information, and 2-D time-frequency image inputs in the other path, denoted as the 2-D feature extraction module, which applies the wavelet transform (WT) to generate feature information based on wavelet time-frequency (WTF) images. Finally, the two features are concatenated to a 1-D feature sequence.

(2) An improved 1-D attention mechanism is applied in the 1-D feature extraction module to focus the classification network on sensitive frequency information. Mean-while, an improved global attention mechanism (GAM) is applied in the 2-D feature extraction module for refining the key features in the time-frequency images and reducing interference arising from noise. The proposed attention mechanisms are im-proved relative to previously proposed mechanisms by adding a multilayer perceptron.

(3) A novel combinatorial loss function is applied when training the proposed dual-path network. The stability of the learning process is enhanced by applying a gradient harmonizing mechanism (GHM) loss function, which is always applied in the field of image recognition for addressing sample imbalance. Furthermore, label smoothing regularization (LSR) is used instead of the actual class labels, and an LSR loss function is applied to enhance learning capability and alleviate over-fitting.

The remainder of this paper is organized as follows. Section II introduces the theoretical foundation. Section III presents a detailed description of the proposed method. The performance of the proposed method is validated in Section IV in comparison with the performances of other state-of-the-art methods based on two public datasets. Finally, the conclusions of the study are presented in Section V.

II. THEORY OF CONVOLUTIONAL NEURAL NETWORK

A standard CNN architecture is composed of a convolution (Conv) layer, pooling layer, fully connected (FC) layer, and a softmax classifier layer [35]. The convolution and pooling layers are employed for extracting features, the FC layer maps the feature space calculated by the pooling layer to a sample label space, and the softmax layer converts the prediction vector into probability values based on the softmax activation function.

A convolution layer is described as

$$x_j^l = \sum\nolimits_i x_i^{l-1} k_{ij}^l + b_j^l \tag{1}$$

where x_j^l is the j-th feature graph output of the l-th layer, x_i^{l-1} is the i-th feature graph of the $(l-1)$-th layer, k_{ij}^l is the convolution kernel between the i-th input feature graph and the j-th feature graph, and b_j^l is an offset term.

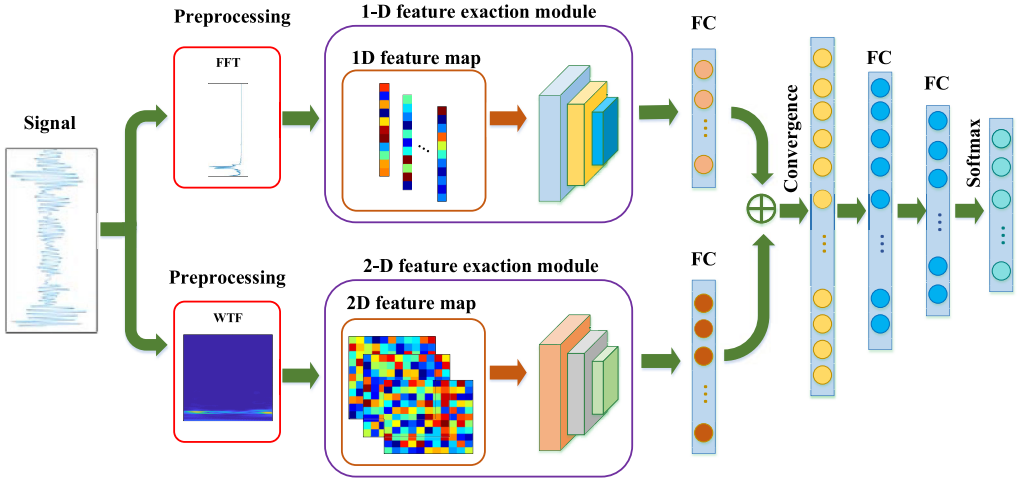

FIGURE 1. Schematic illustrating the overall process of the proposed DPMAM model.

A pooling layer applies a downsampling operation to the output of the Conv layer to achieve dimensionality reduction while retaining the main original characteristics. Meanwhile, it provides a network structure that is not prone to overfitting. The output of a maximum pooling (MaxPool) layer $p^{h(m,t)}$ is obtained as

$$p^{h(m,t)} = \max_{(n-1)g<t<ng} \left\{ a^{h(m,t)} \right\} n = 1, 2, \cdots . \tag{2}$$

where $a^{h(m,t)}$ is the activation value of the t-th neuron of the m-th feature graph in the h-th layer, g is the width of the pooling region, and n is the n-th pooling kernel.

The FC layer achieves the classification task. Therefore, the output $\hat{y}$ of the FC layer is described as

$$\hat{y} = f\,(Wx + b) \tag{3}$$

where W is the weight matrix, x is the input vector, b is a bias vector, and $f\,(\cdot)$ is an activation function, which is the softmax activation function in the present work.

The softmax activation function returns a normalized probability $\hat{y}_i$ of a prediction with label i, and is defined as follows:

$$\hat{y}_i = \frac{e^{(W_i x + b_i)}}{\sum_{k=1}^{C} e^{(W_k x + b_k)}} \tag{4}$$

where C is the number of classes.

III. PROPOSED METHOD

A. OVERVIEW

The overall architecture of the DPMAM model is illustrated schematically in Figure 1. As can be seen, a raw 1-D vibration signal is firstly input into the network model, which branches into two different paths to extract the 1-D and 2-D features of the input signal using the FFT and WT, respectively. The 1-D feature maps are processed by the attention convolution and pooling layers in the 1-D feature extraction module, and the 2-D feature maps (i.e., time-frequency images) are processed by the attention convolution and pooling layers in the 2-D feature extraction module. Here, a large convolution kernel can improve the anti-noise robustness of 2-D feature extraction. After passing through the respective FC layers, the features are fused through bitwise convergence, and the signal is classified finally by a two-layer FC network with a softmax layer. The details of the model are addressed in the following subsections.

Finally, the WTF image is obtained after combining the frequency sequence with the raw time sequence.

B. DATA PREPROCESSING

In the first data preprocessing step, the raw vibration signal is segmented into $\boldsymbol{N}$ samples, $\boldsymbol{y(n)}$ with a length of $\boldsymbol{k}$. Then, in the 1-D path, $\boldsymbol{y(n)}$ is preprocessed by the Fast Fourier transform (FFT) in the first 1-D path to generate feature information.

In the 2-D path, $\boldsymbol{y(n)}$ is also preprocessed by the WT in the 2-D path to produce WTF images as feature information [32]. Cmor wavelet is selected as the basis function [32].

C. DUAL-PATH NETWORK

1) 1-D FEATURE EXTRACTION MODULE

The 1-D feature extraction module with the 1-D attention mechanism is illustrated schematically in Figure 2. Firstly, the input vector z_O is transformed into a characteristic vector z by a 5×1 Conv function F_1 and a MaxPool function F_{MP1} with a 1-D batch normalization (BN) function as follows:

$$z = F_{MP1}[F_1(z_O)] \tag{5}$$

The mechanism applied in the proposed 1-D feature extraction module is based on a previously proposed attention mechanism [36]. Here, a global max pooling (GMP) layer is used to obtain the crucial pulses z_{GMP} from z as follows:

$$z_{\mathrm{GMP}} = \max_{0 \le j < d} z(1, j) \tag{6}$$

Moreover, z_{GMP} and z are then concatenated and input into a 1×1 Conv function F_1. This yields the following intermediary matrix f_{im}:

$$f_{im} = \delta\left(F_1\left[cat(z, z_{GMP}\right]\right) \tag{7}$$

after being subjected to the Acon-C activation function $\delta(g)$ for addressing data nonlinearity, which is defined as

$$\delta(x) = (p_1 - p_2)z\sigma[\beta(p_1 - p_2)x] + p_2 z \tag{8}$$

Here, β, p_1, and p_2 are activation function parameters, where $\beta = p_1 = 1$ and $p_2 = 0$, and σ denotes the sigmoid function. Then, f_{im} is separated into pulses z', which are retained, and other pulses that are discarded according to the previously proposed method [35]. The retained pulses are then input into another 1×1 Conv function F_2 and a sigmoid function σ to obtain the following signal:

$$g = \sigma[F_2(f_{im}^{z'})] \tag{9}$$

Finally, after re-weighting, inputting into a 7×1 Conv function F_3 and a MaxPool function F_{MP2}, and reshaping, the final output y_c is obtained as follows:

$$y_c = F_{MP2}[F_3(z \otimes g)] \tag{10}$$

where $\otimes$ represents element-wise product.

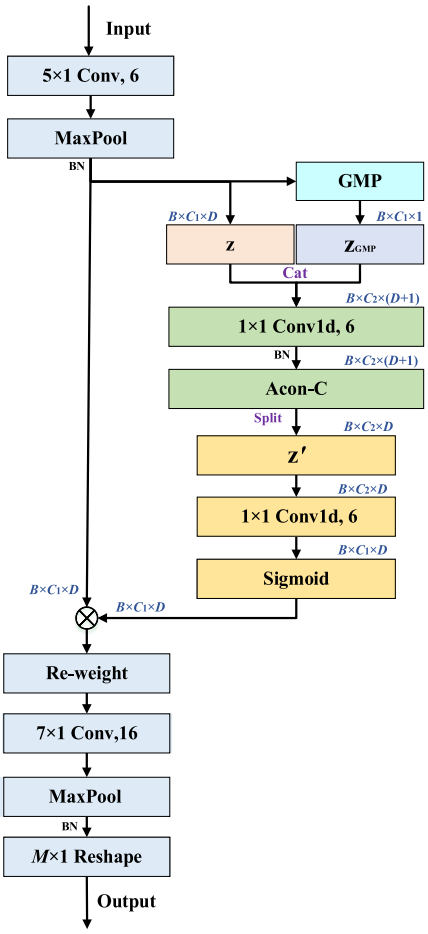

FIGURE 2. Schematic illustrating the architecture of the 1-D feature extraction module with the 1-D attention mechanism.

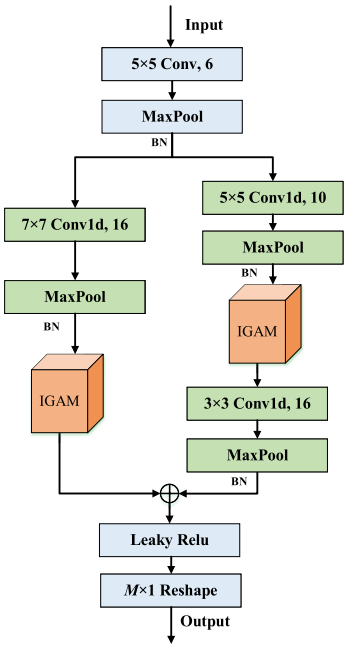

FIGURE 3. Schematic illustrating the architecture of the 2-D feature extraction module with the 2-D attention mechanism.

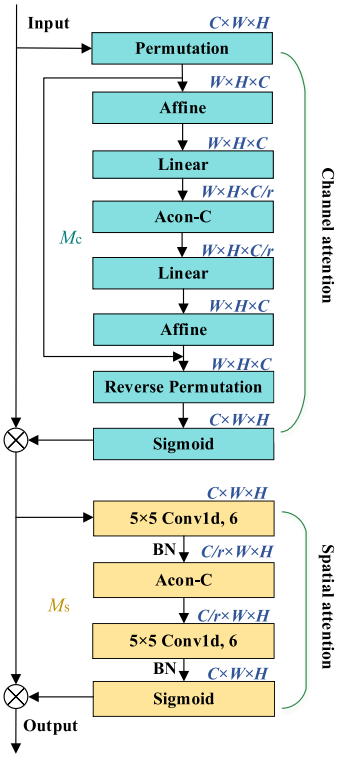

FIGURE 4. Schematic illustrating the architecture of the GAM module.

2) 2-D FEATURE EXTRACTION MODULE

The architecture of the 2-D feature extraction module with the corresponding attention mechanism applied herein is similar to that of a residual block [37], and is illustrated schematically in Figure 3. As can be seen, two convolution attention channels are applied to characterize adequate

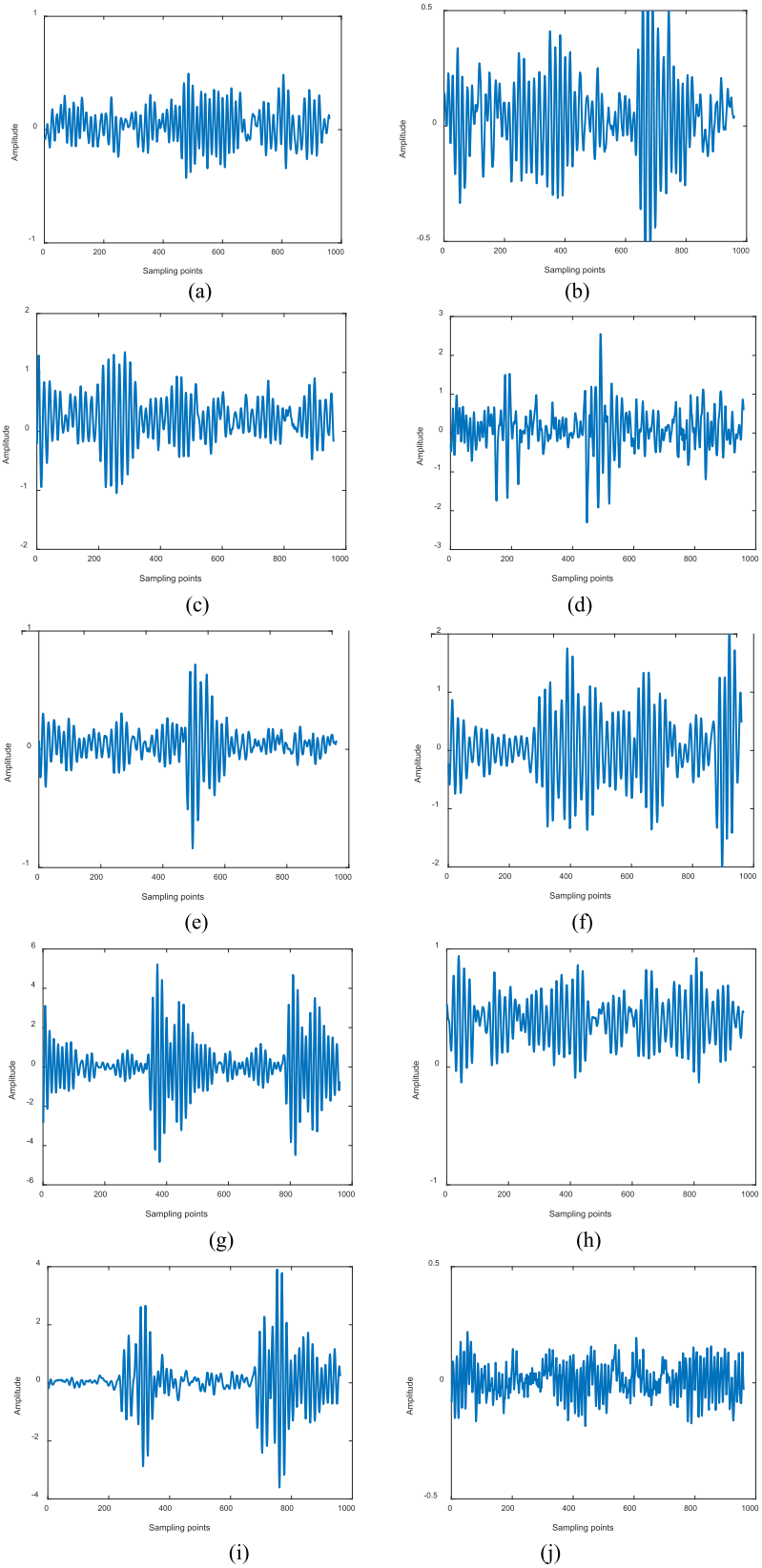

FIGURE 5. Representative vibration signals associated with different bearing conditions in the CWRU dataset: (a) 0.007 inch ball fault; (b) 0.014 inch ball fault; (c) 0.021 inch ball fault; (d) 0.007 inch inner race fault; (e) 0.014 inch inner race fault; (f) 0.021 inch inner race fault; (g) 0.007 inch outer race fault; (h) 0.014 inch outer race fault; (i) 0.021 inch outer race fault; (j) normal race condition.

information describing the complicated features in WTF images. A preprocessed WTF image firstly passes through a Conv layer with a convolution kernel of 5 and a MaxPool layer in the one convolution channel, while a larger convolution kernel of 7 is applied in the other convolution channel to remove noise. The outputs of the MaxPool layers in both channels are processed via improved global attention mechanisms (IGAMs). In contrast to a standard convolutional block attention module (CBAM), which considers both the spatial dimension and channel dimension without losing cross-dimensional information by ignoring interactions between the spatial and channel dimensions, a GAM preserves the information to ensure the capture of key signal characteristics in all dimensions. Finally, the outputs of the two channels are concatenated and subjected to a Leaky-Relu function for overcoming activation function sparsity.

As illustrated schematically in Figure 4, IGAM module consists of a channel attention sub-module and a spatial attention sub-module, which both use the same reduction ratio r. The mechanism is used to multiply the corresponding output with the input as

$$\begin{cases} I' = M_C(I) \otimes I \\ I'' = M_S(I') \otimes I' \end{cases} \tag{11}$$

Here, the WTF image $I \in R^{C\times W\times H}$, where C, W, and H represent the number, width, and height of channels, respectively, M_C and M_S are the respective channel and spatial attention maps, and $\otimes$ represents the multiplication of corresponding elements. Accordingly, I' is the image resulting from the channel attention sub-module and I'' is the image resulting from the spatial attention sub-module.

The channel attention sub-module uses three-dimensional (3-D) arrangements to retain 3-D information. The corresponding feature matrix $x_o$passes through a permutation layer and an affine layer, and then passes through a five-layer multilayer perceptron (MLP) to magnify cross-dimensional dependencies. Here, the following affine function:

$$Aff_{\alpha,\beta}(x_o) = Diag(\alpha) \cdot x_o + \beta \tag{12}$$

where α and β are learnable weight parameters, and the function Diag$(\cdot)$ creates a diagonal matrix, is adopted in this module rather than batch normalization because an affine function typically requires only a very short reasoning time and has no reliance on batch statistics. Finally, the MLP model applies an Acon-C activation function.

Spatial attention sub-module uses two convolution layers for spatial information fusion to focus on spatial information. The number of channels is reduced with a convolution to reduce the amount of computation. After Acon-C activation and another convolution operation, the number of channels is then increased to keep invariant. Finally, it is output by a sigmoid function.

D. LOSS FUNCTION

A loss function aides in model training by providing a consistent measure of the difference between the values predicted by a model and the true values. Under the condition of K classes, the GHM classification loss is defined as follows:

$$L_{GHM} = \sum_{k=1}^{K} \frac{L_{CE}(p_k, p_k^*)}{GD(g_k)} \tag{13}$$

Here, $L_{\mathrm{CE}}(g)$ is cross entropy loss function, where p_k is the distribution of the predicted values and p_k^* is the distribution of the actual values for the k-th class, and $GD(g_k)$ is the gradient density function, where g_k is gradient norm of the k-th class. The LSR loss is expressed in conjunction with a smoothing coefficient ε as follows:

$$L_{LSR} = (1-\varepsilon)L_{CE} + \varepsilon U_k \tag{14}$$

$$U_k = -\frac{\sum_{k=1}^{K} log(p_k)}{K} \tag{15}$$

Hence, the combinatorial loss function L_{Com} applied herein is a weighted sum of L_{GHM} and L_{LSR}, and is defined according to a weight ω as follows:

$$L_{Com} = \omega L_{GHM} + (1-\omega)L_{LSR} \tag{16}$$

The value $\omega = 0.3$ was applied herein based on the results of various comparison experiments.

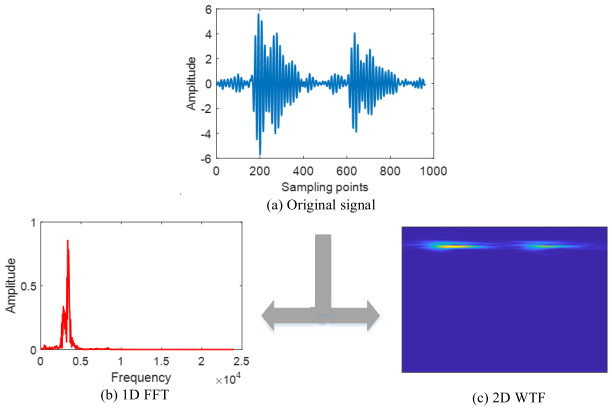

FIGURE 6. Representative data preprocessing result for the CWRU dataset in Case I.

IV. CASE STUDIES

The two public fault diagnosis datasets were employed to evaluate the fault diagnosis performance of the proposed DPMAM method including the Case Western Reserve University (CWRU) rolling bearing dataset (48k Drive End) [38], which is a benchmark constant-speed dataset widely employed in bearing fault diagnosis, and the Xi'an Jiaotong University (XJTU) variable speed Spectral Quest machinery fault dataset (VSQ) [39], which includes samples with variable speeds and field noise. Accordingly, the XJTU-VSQ dataset is useful for evaluating the anti-noise performance of a fault diagnosis method. These two datasets were respectively applied in Cases I and II. Model training was implemented using Pytorch 1.10 and Python 3.8 on a personal computer running Windows 10 with an Intel Core i7-8700 CPU, 16 GB

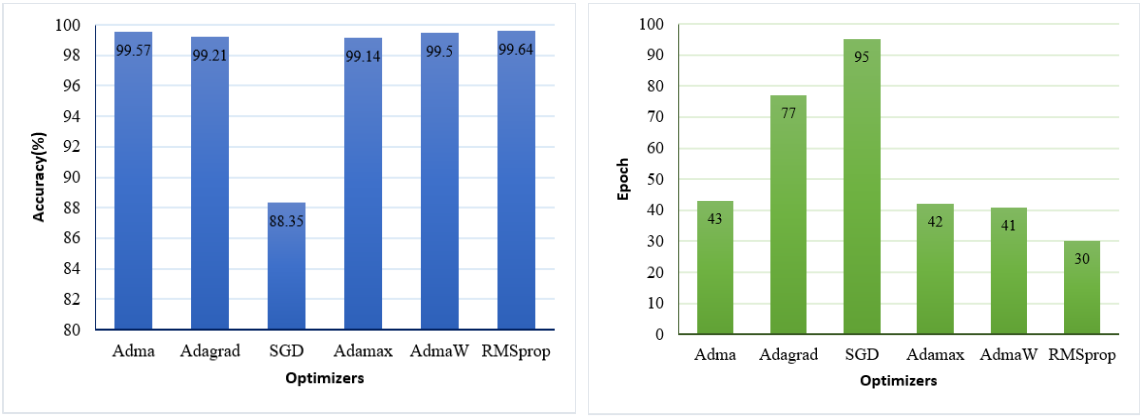

FIGURE 7. Performance of different optimizers: (a) accuracy; (b) stopping epoch.

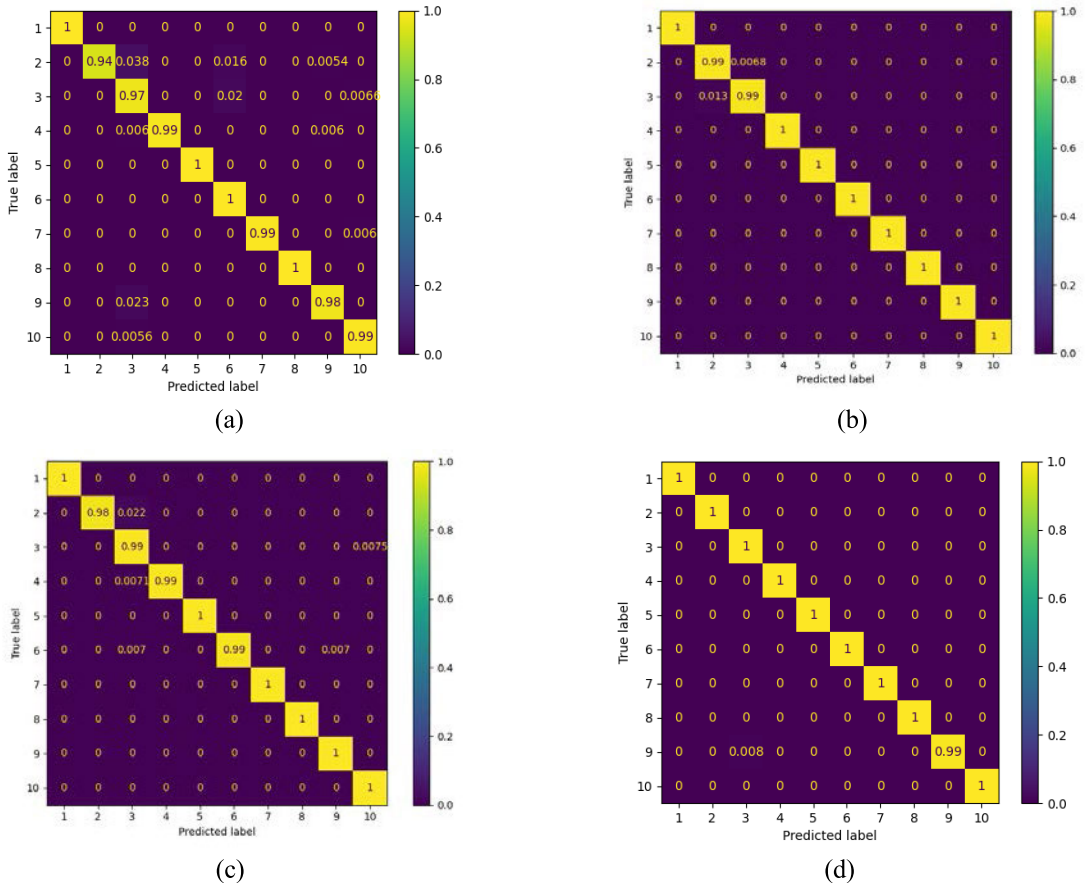

FIGURE 8. Confusion matrixes pertaining to the classification results obtained with the four SRs extracted from the 0 HP data subset in Case I: (a) SR1; (b) SR2; (c) SR3; (d) SR4.

of RAM, and an RTX 2060 GPU. In addition, the optimization performances of six commonly applied optimizers were compared, including Adma, Adagrad, SGD, Adamax, AdmaW, RMSprop, for representative data in Case I.

Each condition consisted of 200 randomly selected samples for each experiment and each sample was composed of 960 data points. The order of the samples was shuffled prior to each experiment to guarantee a random arrangement of training and testing samples. Each experiment was repeated 20 times in succession. The impact of data sparsity was evaluated using four set ratios (SRs) of training samples, validation samples, and testing samples consisting of 2:1:17,

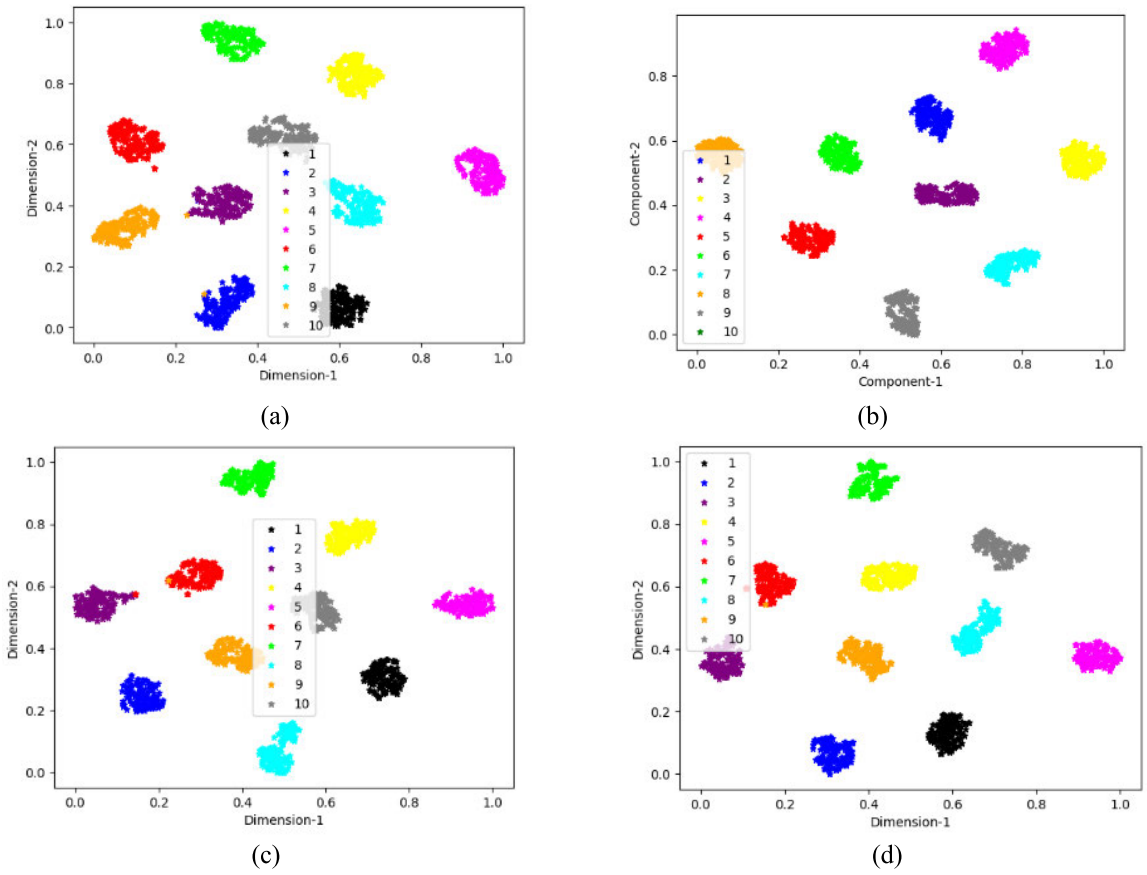

FIGURE 9. Visualizations of two-dimension feature distributions obtained with the four SRs extracted from the 0 HP data subset in Case I: (a) SR1; (b) SR2; (c) SR3; (d) SR4.

3:1:16, 2:1:7, and 3:1:6 under each condition. The random distribution of training samples among classes in these different SRs can be evaluated according to the sample imbalance ratio R_{im}, which is defined as

$$R_{im} = \frac{Max(N_c) - Min(N_c)}{Avg(N_c)} \tag{17}$$

where $N_c = \{n_{tra}^1, n_{tra}^2, \cdots, n_{tra}^{10}$, is the number of samples n_{tra} pertaining to each fault class in the training set, the function Max(·) selects the maximum n_{tra} value, the function Min(·) selects the minimum n_{tra} value, Avg(·) is the average function, and $R_{\text{im}} \in [0, 2)$. When $R_{\text{im}} = 0$, the number of training samples in each condition is equal, and the degree of sample imbalance among classes increases with increasing R_{im}. The DPMAM parameters included a reduction ratio of $r = 4$ applied to the attention mechanisms and a maximum training epoch of 200. Decreasing accuracy under continuous training was avoided by defining the actual training epoch in accordance with an early stopping method using a patience value of 10 [36]. All other parameter values are discussed in the following subsection. Comparison experiments were executed with four existing models without data augmentation, including 1DCNN [40] and MK-ResCNN [41], which apply standardized 1-D frequency signals as inputs, and Letnet-5 [42] and ResNet-18 [37], which apply 2-D WTF images as inputs.

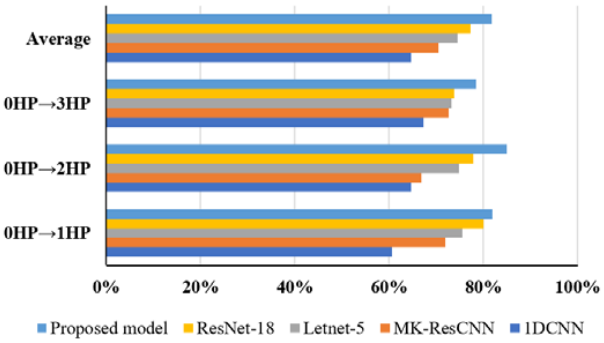

FIGURE 10. Classification accuracies obtained by the various methods considered for the different load data subsets in Case I when trained using the 0 HP data subset.

A. CASE I: CWRU DATASET

1) DATA DESCRIPTION AND PROCESSING

The vibration signals in the 48 k Drive End dataset were sampled from accelerometers mounted on the fan end and the drive end of a 2-HP motor housing. The dataset included four

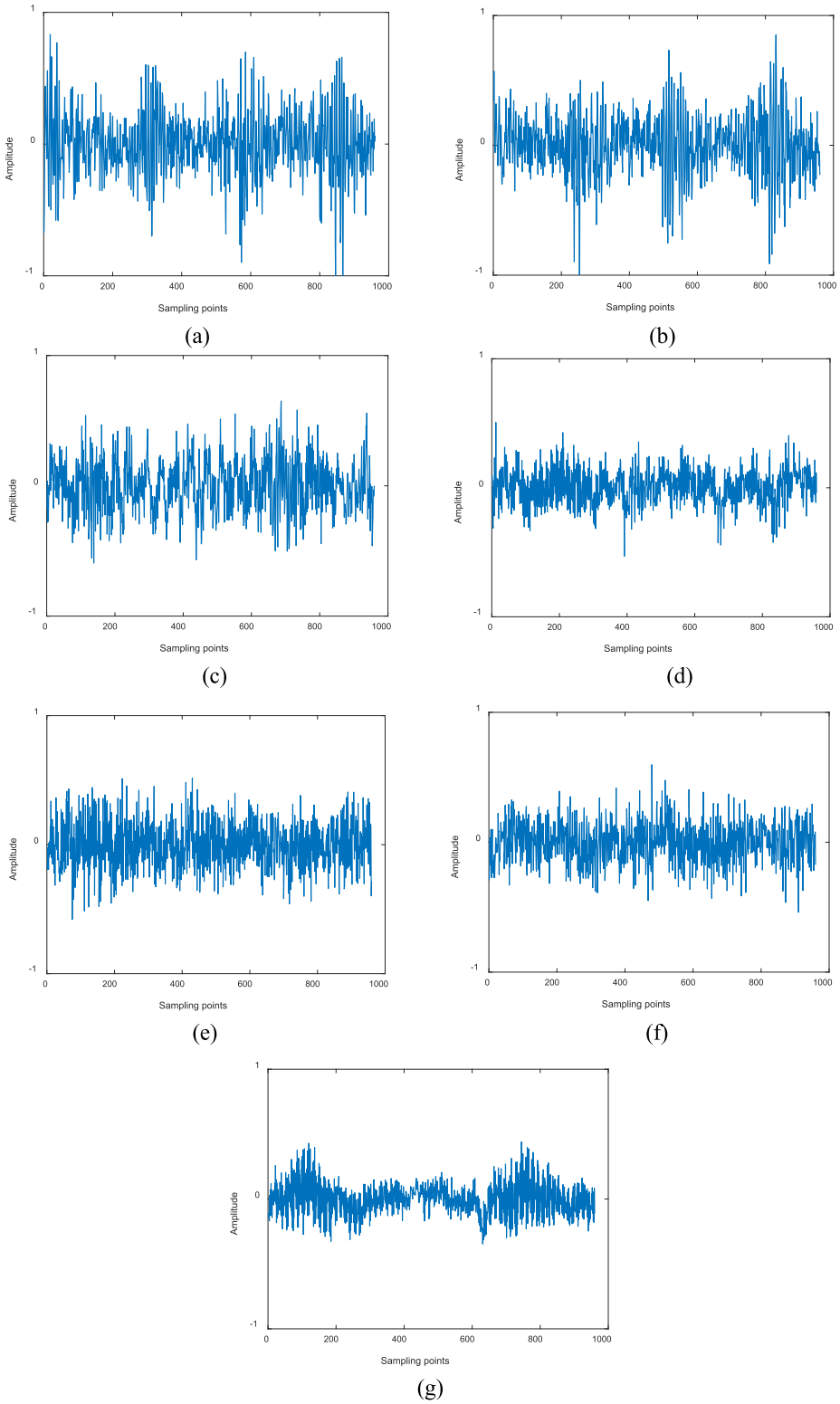

FIGURE 11. Representative vibration signals associated with different bearing conditions in the XJTU dataset: (a) minor inner race fault; (b) medium inner race fault; (c) severe inner race fault; (d) minor outer race fault; (e) medium outer race fault; (f) severe outer race fault; (g) normal race condition.

data subsets (0 HP, 1 HP, 2 HP, 3 HP) with different loads of 0 kW, 0.735 kW, 1.471 kW, and 2.206 kW, respectively. In all experiments, the models were trained with samples selected from the 0 HP data subset. The fault conditions include ball, inner race, and outer race faults, and each fault category includes three fault sizes of 0.007 inches, 0.014 inches,

and 0.021 inches. Representative vibration signals associated with the 10 different signal classes are presented in Figure 5, and include 9 fault conditions (Figure 5(a)–(i)) and 1 normal condition (Figure 5(j)). The SRs and R_{im} values pertaining to the data samples employed in Case I are listed in Table 1. The scale is set as 256. In addition, a representative result of preprocessing, including the raw vibration signal, and the corresponding 1-D FFT data and 2-D WTF image, are presented in Figure 6.

TABLE 1. Set ratios and sample imbalance ratios Rim of the four SRs employing different proportions of training, validation, and testing samples in Case I.

Index	Set ratio	R_{im}
SR1	2:1:17	0.70
SR2	3:1:16	0.55
SR3	2:1:7	0.60
SR4	3:1:6	0.48

2) TRAINING PERFORMANCE ANALYSIS

The training performance of the proposed model is significantly sensitive to the batchsize and the applied optimizer. Therefore, we first investigated the training performance obtained with SR3 data extracted from the 0 HP data subset conducted using the RMSprop optimizer with different batchsizes, and the actual training epoch reached under the early stopping method, the evaluation loss, evaluation accuracy, and total training time are listed in Table 2 for different batchsizes. As can be seen, the training time and number of training epochs decrease with increasing batchsize up to a batchsize of 64, and then increase with further increase in the batchsize. This increase in the training time arises because the larger batchsize tends to decrease the generalization ability of model. Therefore, a balance is obtained with a batchsize of 64, which was applied in all subsequent experiments. Moreover, as shown in Figure 7, the RMSprop optimizer was found to achieve the highest accuracy of all optimizers considered at the lowest stopping epoch of 30. Therefore, the RMSprop optimizer was employed for training in all subsequent experiments. It should be noted that the batchsize and optimizer results obtained with all other SR datasets were similar to those obtained with SR3.

3) IMPACT OF SET RATIO ON PREDICTION PERFORMANCE

The confusion matrixes pertaining to the classification results obtained by the trained DPMAM model under the four SRs extracted from the 0 HP data subset in Case I are presented in Figure 8. As can be seen, the different fault conditions are identified with an accuracy close to 100%, except that the 0.014-inch ball fault (Class 2) is identified at a relatively lower accuracy level at SR1 and SR3. It may be because that the physical phenomena is raised in this specific case, such as resonance. The features in the frequency domain are also not obvious. The classification accuracy of the method can

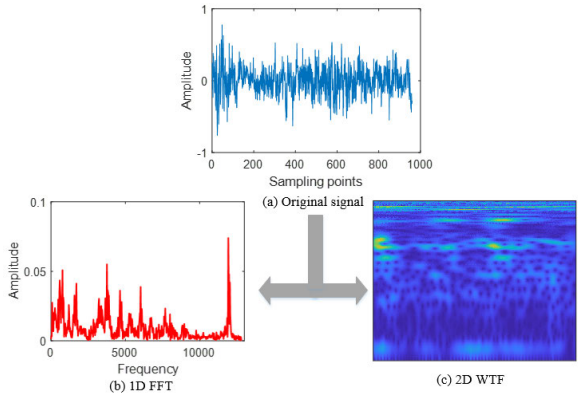

FIGURE 12. Representative data preprocessing result for the XJTU dataset in Case II.

be affected by SR. The visualizations presented in Figure 9 of 2-D feature distributions obtained with the four SRs in Case I were obtained by applying a nonlinear unsupervised dimension reduction technique based on t-distributed random neighborhood embedding. As can be seen, each class is generally separated successfully.

TABLE 2. Training results obtained with different batchsizes in Case I with SR3 data extracted from the 0 HP data subset.

Batchsize	Early stopping	Evaluation loss	Evaluation accuracy (%)	Training time (s)
8	37	0.0504	100	39.51
16	34	0.0263	100	20.23
32	31	0.0141	100	12.77
64	30	0.0082	100	11.34
128	35	0.0042	100	11.80

4) PERFORMANCE COMPARISONS WITH OTHER METHODS

The classification results obtained by the various methods considered under the four SRs extracted from the 0 HP data subset in Case I are listed in Table 3. As can be seen, the classification accuracies generally increase with decreasing R_{im}, as expected with increasing sample imbalance. However, the classification accuracies of the proposed DPMAM method outperform that of all other methods considered, particularly at high R_{im}. Thus, the proposed method has better fault detection performance for small-sample conditions. It should be noted that the classification performance of ResNet-18 is close to that of the DPMAM method, even at high R_{im}. However, the training time required by ResNet-18 was uniformly greater than that required by the DPMAM model.

The generalization ability of these methods was evaluated by analyzing their classification performance obtained for the model trained with samples extracted from the 0 HP data subset when applied to testing samples extracted from the 1 HP, 2 HP, and 3 HP data subsets. The classification accuracies obtained by the various methods considered for the different load data subsets in Case I are presented in

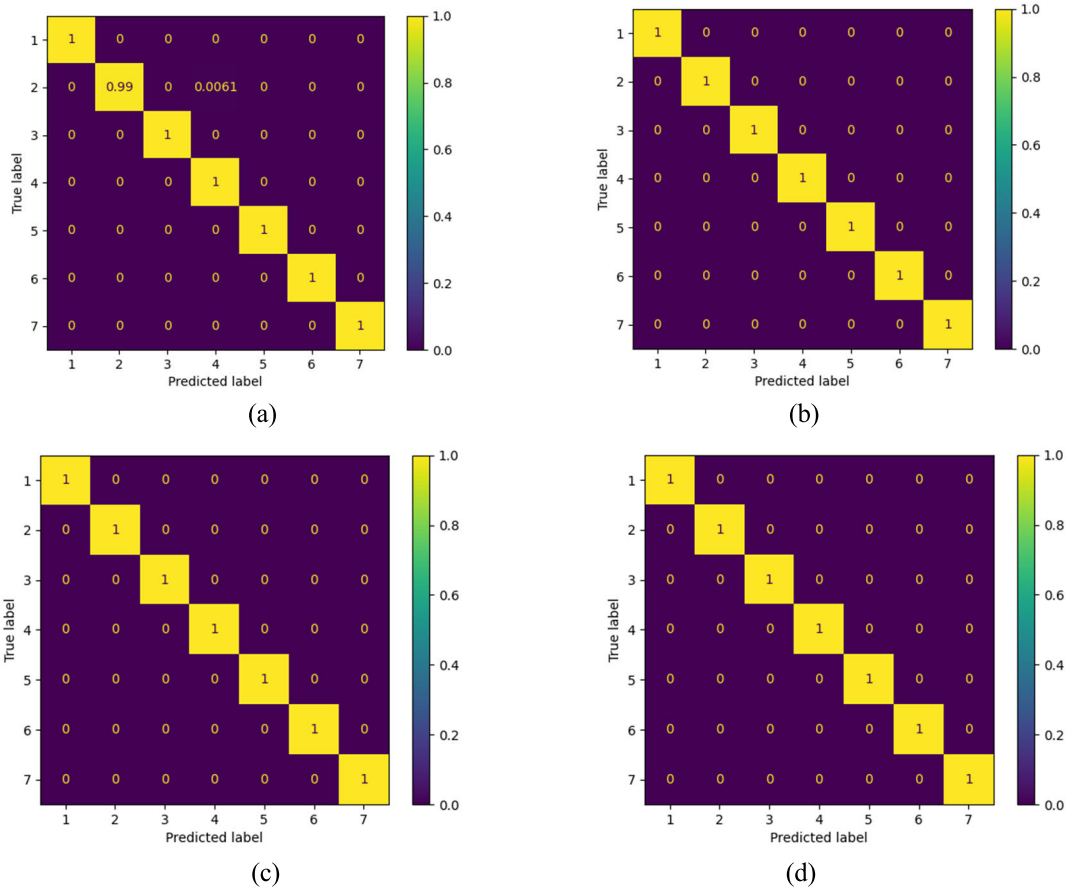

FIGURE 13. Confusion matrixes pertaining to the classification results obtained with the four SRs in Case II: (a) SR1; (b) SR2; (c) SR3; (d) SR4.

TABLE 3. Classification results obtained by the various methods considered for the 0 HP data subset in Case I.

Method	Average classification accuracies under different SR (%)			
	SR1	SR2	SR3	SR4
1DCNN	83.05	90.80	96.35	97.00
MK-ResCNN	86.27	91.21	96.64	98.08
Letnet-5	91.94	96.28	97.17	99.16
ResNet-18	98.05	98.17	99.06	99.83
DPMAM	**98.64**	**99.79**	**99.64**	**99.91**

Figure 10. As can be seen, the proposed DPMAM method provides superior classification performance under all cases. Moreover, its superiority is particularly dominant when the load of the testing data samples differs greatly from the load of the training data samples (i.e., for 2 HP and 3 HP testing data samples).

B. CASE II: XJTU DATASET

1) DATA DESCRIPTION AND PROCESSING

The vibration signals in the VSQ dataset were sampled under continuous varying speeds, and collected with a sampling frequency of 25.6 kHz. The fault conditions in the VSQ dataset include six bearing faults with minor, medium, and severe levels of wear for the outer race and inner race. Representative vibration signals associated with the 6 fault conditions and 1 normal condition are presented in Figure 11(a)–(f) and Figure 11(g), respectively. The SRs and R_{im} values pertaining to the data samples employed in Case II are listed in Table 4. In addition, a representative result of prepocessing, including the raw vibration signal, and the corresponding 1-D FFT data and 2-D WTF image, are presented in Figure 12. In contrast to what was observed in Figure 6, the results presented in Figure 12(b) and (c) clearly demonstrate that the signal includes a sizeable level of field noise disturbance.

TABLE 4. Set ratios and sample imbalance ratios Rim of the four SRs employing different proportions of training, validation, and testing samples in Case II.

SR Index	Set ratio	R_{im}
SR1	2:1:17	0.55
SR2	3:1:16	0.48
SR3	2:1:7	0.37
SR4	3:1:6	0.35

2) IMPACT OF SET RATIO ON PREDICTION PERFORMANCE

The confusion matrixes pertaining to the classification results obtained by the trained DPMAM model under the four

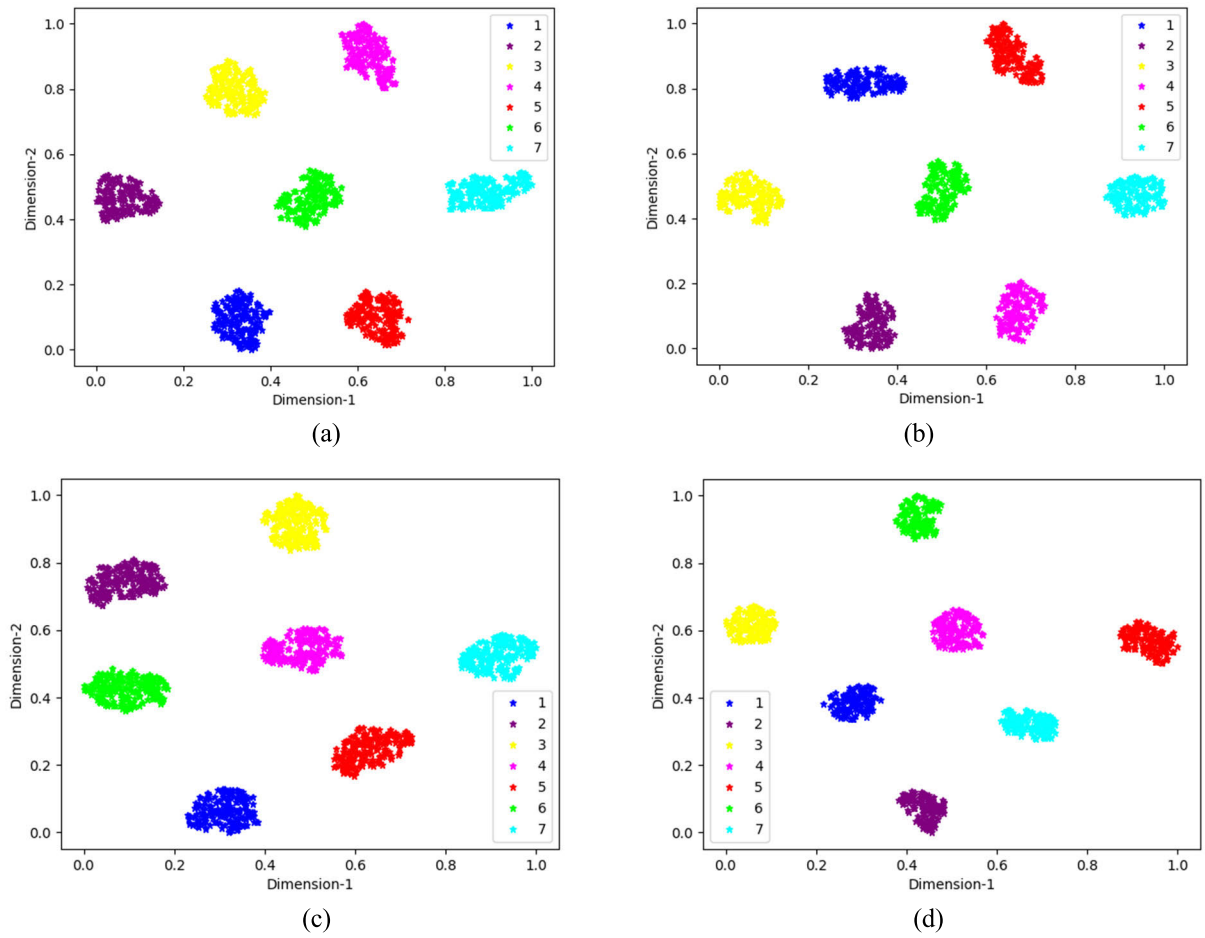

FIGURE 14. Visualizations of two-dimension feature distributions obtained with the four SRs in Case II: (a) SR1; (b) SR2; (c) SR3; (d) SR4.

SRs extracted from the XJTU-VSQ dataset in Case II are presented in Figure 13. As can be seen, the different fault conditions are identified with an accuracy greater than 99.90%. In particular, all normal samples (Class 4) were well recognized. As shown in Figure 14, the 2-D feature distributions demonstrate that the DPMAM model provides satisfactory fault classification performance.

3) PERFORMANCE COMPARISONS WITH OTHER METHODS

The classification accuracies obtained by the various methods considered under the four SRs extracted from the XJTU-VSQ dataset in Case II are listed in Table 5. Note that the classification accuracies generally increase with decreasing R_{im}. Moreover, the proposed DPMAM method again outperforms all other methods considered, particularly at high R_{im}. The results also demonstrate that the accuracy of the two fault diagnosis methods using 2-D WTF images as inputs were much greater than those applying 1-D signals as inputs. This differed markedly from the results obtained in Case I (Table 3), which can be attributed to the particular benefit of employing WTF images under conditions of high signal noise. Accordingly, the results demonstrate that the proposed DPMAM method provides particularly superior classification performance under variable speeds and high signal noise conditions compared to the other methods considered.

TABLE 5. Classification accuracies of the various methods compared in Case II.

Method	Average accuracies with different set ratios (%)			
	SR1	SR2	SR3	SR4
1DCNN	89.28	89.78	93.26	98.21
MK-ResCNN	86.38	93.12	93.16	97.62
Letnet-5	97.56	98.36	99.40	98.73
ResNet-18	98.99	99.64	99.40	99.52
DPMAM	**99.91**	**100.00**	**100.00**	**100.00**

V. CONCLUSION

The present work addressed current limitations in the ability of deep learning methods to make full use of the inherent characteristics of small vibration datasets by proposing a dual-path model based on multiple attention mechanisms. The proposed model combines 1-D frequency signals with 2-D time-frequency images to enrich the features of small

datasets. Then, the classification task is focused on sensitive frequency information by applying a 1-D attention mechanism in the 1-D frequency extraction path, while a global attention mechanism is incorporated into to the 2-D feature extraction path to refine the key features and reduce interference due to noise in the image data. Moreover, leaky ReLu and Acon-C activation functions are introduced in the model to further enhance the generalization and robustness of the classification process. The excellent classification performance of the proposed method was demonstrated in conjunction with two different public fault diagnosis datasets based on comparisons with the performances of 1DCNN, MK-ResCNN, Letnet-5, and ResNet-18 fault diagnosis methods. The results clearly demonstrated that the proposed method provides more accurate classification performance than the other methods considered when subject to a small training dataset.

NOMENCLATURE

ACRONYMS

ANN	Artificial neural network.
RF	Random forest.
SVM	Support vector machine.
CNN	Convolutional neural network.
GNN	Generative adversarial network.
DPMAM	Multiple attention mechanism.
FFT	Fast Fourier transform.
WT	Wavelet transform.
WTF	Wavelet time-frequency.
GAM	Global attention mechanism.
GHM	Gradient harmonizing mechanism.
LSR	Label smoothing regularization.
FC	Fully connected.
IGAM	Improved global attention mechanism.
CBAM	Convolutional block attention module.
SR	Set ratio.

REFERENCES

[1] 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," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587, doi: 10.1016/j.ymssp.2019.106587.

[2] Y. Feng, J. Chen, J. Xie, T. Zhang, H. Lv, and T. Pan, "Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107646, doi: 10.1016/j.knosys.2021.107646.

[3] D. Li, Z. Cai, B. Qin, and L. Deng, "Signal frequency domain analysis and sensor fault diagnosis based on artificial intelligence," *Comput. Commun.*, vol. 160, pp. 71–80, Jul. 2020, doi: 10.1016/j.comcom.2020.05.034.

[4] M. Cerrada, G. Zurita, D. Cabrera, R.-V. Sánchez, M. Artés, and C. Li, "Fault diagnosis in spur gears based on genetic algorithm and random forest," *Mech. Syst. Signal Process.*, vols. 70–71, pp. 87–103, Mar. 2016, doi: 10.1016/j.ymssp.2015.08.030.

[5] R. Hao, Z. Peng, Z. Feng, and F. Chu, "Application of support vector machine based on pattern spectrum entropy in fault diagnostics of rolling element bearings," *Meas. Sci. Technol.*, vol. 22, no. 4, Apr. 2011, Art. no. 045708, doi: 10.1088/0957-0233/22/4/045708.

[6] Z. Huo, Y. Zhang, L. Shu, and M. Gallimore, "A new bearing fault diagnosis method based on Fine-to-Coarse multiscale permutation entropy, Laplacian score and SVM," *IEEE Access*, vol. 7, pp. 17050–17066, 2019, doi: 10.1109/ACCESS.2019.2893497.

[7] Y. Huo, S. Gang, and C. Guan, "FCIHMRT: Feature cross-layer interaction hybrid method based on Res2Net and transformer for remote sensing scene classification," *Electronics*, vol. 12, no. 20, p. 4362, Nov. 2023, doi: 10.3390/electronics12204362.

[8] T. Yang, N. Sun, and Y. Fang, "Neuroadaptive control for complicated underactuated systems with simultaneous output and velocity constraints exerted on both actuated and unactuated states," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 1–11, Aug. 2021, doi: 10.1109/TNNLS.2021.3115960.

[9] T. Yang, N. Sun, H. Chen, and Y. Fang, "Adaptive optimal motion control of uncertain underactuated mechatronic systems with actuator constraints," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 1, pp. 210–222, Feb. 2023, doi: 10.1109/TMECH.2022.3192002.

[10] H. Yang, X. Li, and W. Zhang, "Interpretability of deep convolutional neural networks on rolling bearing fault diagnosis," *Meas. Sci. Technol.*, vol. 33, no. 5, May 2022, Art. no. 055005, doi: 10.1088/1361-6501/ac41a5.

[11] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, and X. Huang, "A review of the application of deep learning in intelligent fault diagnosis of rotating machinery," *Measurement*, vol. 206, Jan. 2023, Art. no. 112346, doi: 10.1016/j.measurement.2022.112346.

[12] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vibrat.*, vol. 377, pp. 331–345, Sep. 2016, doi: 10.1016/j.jsv.2016.05.027.

[13] Z. Chen, K. Gryllias, and W. Li, "Mechanical fault diagnosis using convolutional neural networks and extreme learning machine," *Mech. Syst. Signal Process.*, vol. 133, Nov. 2019, Art. no. 106272, doi: 10.1016/j.ymssp.2019.106272.

[14] W. Yu and C. Zhao, "Broad convolutional neural network based industrial process fault diagnosis with incremental learning capability," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5081–5091, Jun. 2020, doi: 10.1109/TIE.2019.2931255.

[15] T. Pan, J. Chen, T. Zhang, S. Liu, S. He, and H. Lv, "Generative adversarial network in mechanical fault diagnosis under small sample: A systematic review on applications and future perspectives," *ISA Trans.*, vol. 128, pp. 1–10, Sep. 2022, doi: 10.1016/j.isatra.2021.11.040.

[16] Y. Huo, D. Guan, and L. Dong, "Intelligent fault diagnosis of unbalanced samples using optimized generative adversarial network," *Appl. Sci.*, vol. 14, no. 11, p. 4927, Jun. 2024, doi: 10.3390/app14114927.

[17] C. Che, H. Wang, X. Ni, and Q. Fu, "Domain adaptive deep belief network for rolling bearing fault diagnosis," *Comput. Ind. Eng.*, vol. 143, May 2020, Art. no. 106427, doi: 10.1016/j.cie.2020.106427.

[18] X. Kong, X. Li, Q. Zhou, Z. Hu, and C. Shi, "Attention recurrent autoencoder hybrid model for early fault diagnosis of rotating machinery," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021, doi: 10.1109/TIM.2021.3051948.

[19] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018, doi: 10.1109/TIE.2017.2733438.

[20] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020, doi: 10.1109/TII.2019.2943898.

[21] M. Miao, Y. Sun, and J. Yu, "Deep sparse representation network for feature learning of vibration signals and its application in gearbox fault diagnosis," *Knowledge-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108116, doi: 10.1016/j.knosys.2022.108116.

[22] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 8, pp. 1926–1935, Aug. 2017, doi: 10.1109/TIM.2017.2674738.

[23] H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, "Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application," *Measurement*, vol. 199, Aug. 2022, Art. no. 111594, doi: 10.1016/j.measurement.2022.111594.

[24] Q. Zhang, J. Li, W. Ding, Z. Ye, and Z. Meng, "Mechanical fault intelligent diagnosis using attention-based dual-scale feature fusion capsule network," *Measurement*, vol. 207, Feb. 2023, Art. no. 112345, doi: 10.1016/j.measurement.2022.112345.

[25] Y.-J. Huang, A.-H. Liao, D.-Y. Hu, W. Shi, and S.-B. Zheng, "Multi-scale convolutional network with channel attention mechanism for rolling bearing fault diagnosis," *Measurement*, vol. 203, Nov. 2022, Art. no. 111935, doi: 10.1016/j.measurement.2022.111935.

[26] L. Jia, T. W. S. Chow, and Y. Yuan, "GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105794, doi: 10.1016/j.engappai.2022.105794.
[27] 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, Jun. 2021, doi: 10.1016/j.neucom.2021.01.099.
[28] Y. Liu, H. Jiang, C. Liu, W. Yang, and W. Sun, "Data-augmented wavelet capsule generative adversarial network for rolling bearing fault diagnosis," *Knowl.-Based Syst.*, vol. 252, Sep. 2022, Art. no. 109439, doi: 10.1016/j.knosys.2022.109439.
[29] 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, Oct. 2021, doi: 10.1016/j.neucom.2020.11.070.
[30] Y. Feng, J. Chen, T. Zhang, S. He, E. Xu, and Z. Zhou, "Semi-supervised meta-learning networks with squeeze-and-excitation attention for few-shot fault diagnosis," *ISA Trans.*, vol. 120, pp. 383–401, Jan. 2022, doi: 10.1016/j.isatra.2021.03.013.
[31] T. Zhang, J. Chen, F. Li, T. Pan, and S. He, "A small sample focused intelligent fault diagnosis scheme of machines via multimodules learning with gradient penalized generative adversarial networks," *IEEE Trans. Ind. Electron.*, vol. 68, no. 10, pp. 10130–10141, Oct. 2021, doi: 10.1109/TIE.2020.3028821.
[32] X. Yang, B. Liu, L. Xiang, A. Hu, and Y. Xu, "A novel intelligent fault diagnosis method of rolling bearings with small samples," *Measurement*, vol. 203, Nov. 2022, Art. no. 111899, doi: 10.1016/j.measurement.2022.111899.
[33] T. Han, C. Liu, R. Wu, and D. Jiang, "Deep transfer learning with limited data for machinery fault diagnosis," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107150, doi: 10.1016/j.asoc.2021.107150.
[34] S. Liu, J. Chen, S. He, Z. Shi, and Z. Zhou, "Subspace network with shared representation learning for intelligent fault diagnosis of machine under speed transient conditions with few samples," *ISA Trans.*, vol. 128, pp. 531–544, Sep. 2022, doi: 10.1016/j.isatra.2021.10.025.
[35] L. Jing, M. Zhao, P. Li, and X. Xu, "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox," *Measurement*, vol. 111, pp. 1–10, Dec. 2017, doi: 10.1016/j.measurement.2017.07.017.
[36] X. Zhang, C. He, Y. Lu, B. Chen, L. Zhu, and L. Zhang, "Fault diagnosis for small samples based on attention mechanism," *Measurement*, vol. 187, Jan. 2022, Art. no. 110242, doi: 10.1016/j.measurement.2021.110242.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[38] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015, doi: 10.1016/j.ymssp.2015.04.021.
[39] Z. Shi, J. Chen, Y. Zi, and Z. Zhou, "A novel multitask adversarial network via redundant lifting for multicomponent intelligent fault detection under sharp speed variation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021, doi: 10.1109/TIM.2021.3055821.
[40] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016, doi: 10.1109/TIE.2016.2582729.
[41] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3797–3806, Jun. 2020, doi: 10.1109/TII.2019.2941868.
[42] R. Bai, Q. Xu, Z. Meng, L. Cao, K. Xing, and F. Fan, "Rolling bearing fault diagnosis based on multi-channel convolution neural network and multi-scale clipping fusion data augmentation," *Measurement*, vol. 184, Nov. 2021, Art. no. 109885, doi: 10.1016/j.measurement.2021.109885.

XIN LI was born in Shenyang, Liaoning, China, in 1982. He received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from Northeast University, Shenyang, in 2004 and 2009, respectively.

From 2009 to 2011, he was a Postdoctoral Researcher with Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang. Since 2011, he has been an Associate Professor with Shenyang University, Shenyang. His research interest includes application of intelligent technology.

MEILING ZHANG received the B.E. degree in automation from Shenyang University, Shenyang, China, in 2021. She is currently a Graduate Student majoring in control science and engineering with Shenyang University.

Her research interests include deep learning and person re-identification.

HUBO GUO received the B.E. degree in communications engineering from Shenyang University, Shenyang, China, in 2021. He is currently a Graduate Student majoring in control science and engineering with Shenyang University.

His research interests include deep learning and person re-identification.