IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 March 2023, accepted 30 March 2023, date of publication 3 April 2023, date of current version 10 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264276

RESEARCH ARTICLE

Bearing Fault Diagnosis Based on Mel Frequency Cepstrum Coefficient and Deformable Space-Frequency Attention Network

YUNJI ZHAO[1], NANNAN ZHANG[2], ZHIHAO ZHANG[2], AND XIAOZHUO XU[1], (Member, IEEE)

[1]Henan International Joint Laboratory of Direct Drive and Control of Intelligent Equipment, School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454003, China
[2]Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454003, China

Corresponding author: Xiaozhuo Xu (xxz@hpu.edu.cn)

This work was supported in part by the Foundation of the National Natural Science Foundation of China under Grant 61973105, Grant 61573130, and Grant 52177039; in part by the Fundamental Research Funds for the Universities of Henan Province under Grant NSFRF200504; and in part by the Key Technologies R&D Program of Henan Province of China under Grant 212102210145, Grant 212102210197, and Grant 222102220016.

ABSTRACT The main bearing is the core component of gas-fired generator, and its reliability directly affects the stability of the whole system. Therefore, it is of great significance to study the fault diagnosis of the main bearing of gas-fired generator. In the bearing fault diagnosis based on vibration signal, how to extract the signature features of fault effectively is the key to achieving accurate fault diagnosis. Based on extracting the signature features of faults, how to classify the fault features efficiently is another key to achieving accurate fault diagnosis. Based on this, we propose a bearing fault diagnosis method based on Mel frequency cepstrum coefficient (MFCC) and deformable space-frequency attention network (DSFAN). In view of the inconsistent feature distribution of different types of faults, the MFCC algorithm is introduced to preprocess the original fault signals and extract their signature features. Then, the network model DSFAN is constructed based on the space-frequency feature attention mechanism (SFFAM). DSFAN can extract the global constraint features and distributed constraint features of fault signals and realize bearing fault diagnosis. To make full use of classification information, the data processed by MFCC is constructed into a three-dimensional data cube as the input of DSFAN. Finally, the validity of the proposed method MFCC-DSFAN is verified on CWRU, XJTU, and gas-fired generator data sets. The experimental results show the excellent performance of MFCC-DSFAN for fault diagnosis and prove the effectiveness of the attention module in feature extraction.

INDEX TERMS Frequency attention, space attention, deformable convolution networks, Mel frequency cepstrum coefficient, fault diagnosis.

I. INTRODUCTION

Bearing is a basic and important mechanical component in gas-fired generators. Once the bearing fails, and the fault is not diagnosed and dealt with in time, the degree of bearing damage will gradually increase over time, which will affect the normal operation of gas-fired generator [1]. Therefore, it is of great significance to study bearing fault diagnosis and gradually improve the diagnosis efficiency for maintaining the stable operation of gas-fired generators. In the bearing fault diagnosis based on vibration, temperature, and acoustic signals, vibration signals containing rich equipment operating status information are the most widely used [2]. Fault feature extraction and classification are two key steps of bearing fault diagnosis based on vibration signal [3].

How to extract the signature features of faults from fault signals is the key to achieving accurate fault diagnosis [4], [5]. Therefore, researchers have done a lot of research

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

and proposed many fault feature extraction methods. Time domain feature analysis is the earliest feature extraction method, which is intuitive and accurate. However, due to the complexity of the environment, the collected signals are nonlinear and unstable, which leads to an increase in the amount of computation. Therefore, the Fourier transform, which can convert signals from the time domain to the frequency domain, has gradually attracted attention. It can decompose complex signals into simple signal superposition, which is easier to analyze. For example, Zhao et al. [6] use Fourier transform to process fault signals and extract signal spectrum features for fault diagnosis. Compared with the time-domain analysis method, the frequency-domain analysis method can extract more complex and iconic fault features, which is helpful for accurate fault diagnosis. In addition, there are some frequency domain features such as envelope spectrum and high-order spectrum [7], [8]. However, in the process of signal conversion from the time domain to the frequency domain, it is difficult to obtain the location of each frequecncy signal in the time domain. Single-dimension analysis in the time domain or frequency domain can not fully reflect the signal characteristics. Therefore, feature extraction methods based on time-frequency domain analysis are gradually developed. For example, Zhang et al. [9] proposed a feature extraction method based on empirical wavelet transform, which can extract more iconic fault features from complex signals, improve its resistance to noise, and finally obtain better diagnostic results. All the above methods are committed to extracting the signature features of fault signals, which is very important for accurate fault diagnosis.

Based on signature feature extraction, how to classify faults effectively is another key to achieving accurate fault diagnosis. Recent years, the bearing fault diagnosis method based on deep learning can adaptively extract deep fault features and realize fault classification, which has been widely used in bearing fault diagnosis. Convolutional neural network (CNN) has become one of the most popular classification network models in bearing fault diagnosis due to its powerful nonlinear feature extraction ability. Xia et al. [10] used a convolutional neural network to achieve bearing fault diagnosis. However, the traditional convolutional neural network often needs a large number of training samples for training to obtain higher diagnostic accuracy, which will limit its application in fault diagnosis [11], [12], [13], [14]. Therefore, Huang et al. [15] added multi-scale learning on the basis of CNN and proposed a multi-scale cascade convolutional neural network model. This model can integrate multi-scale information from original vibration signals and extract more abundant fault features. Compared with the traditional CNN, it can achieve higher fault diagnosis efficiency under normal or noise conditions with fewer samples. However, it is worth noting that the original fault signal contains not only frequency distribution information but also spatial information [16], [17], [18]. Due to the inherent structure of CNN, it tends to ignore the spatial constraint information between fault data, which cannot extract complete fault features [19], [20], [21].

Based on the above analysis, inspired by the attention mechanism of feature in [22] and the law of human visual attention, a bearing fault diagnosis method based on Mel frequency cepstrum coefficient (MFCC) and deformable space-frequency attention network (DSFAN) was proposed. First, MFCC was introduced to process the original signal and extract the signature features of the fault signal. Then to fully extract the space and frequency information of the signal data, we further process the data and construct three-dimensional data cubes. In terms of feature extraction, DSFAN was constructed based on the space-frequency feature attention mechanism (SFFAM), which can extract deep spatial frequency features from fault data. We first design a frequency attention module (FeAM) to learn the three-dimensional data cubes, which can extract more important frequency distribution features and reduce the interference of useless information to fault diagnosis. In addition, we hope to pay more attention to the categories with the same label as the center category or those useful for center classification, and less attention to the categories with different labels or those useless for classification. Therefore, a space attention module (SaAM) is designed to learn the importance of surrounding categories and give them appropriate attention. In order to further extract spatial information between categories and refine the extracted space-frequency features, we constructed a deformable convolution block (DeCB) inspired by the deformable convolution network. The performance of the proposed method MFCC-DSFAN is tested on three bearing datasets, including Case Western Reserve University (CWRU) bearing dataset, Xi'an Jiaotong University (XJTU) bearing dataset, and the experimental dataset. The experimental results show that the proposed method can achieve good diagnostic results. The contribution of this paper is mainly analyzed from the following three aspects.

1) In view of the inconsistency of feature distribution of different types of fault signal data, this paper introduces MFCC, a classical signal processing method in speech recognition, into fault signal processing. It constructs Mel filter banks to extract the signature features of fault signals.

2) In order to extract the constraint features of fault signals, a space-frequency feature attention network structure was proposed in this paper, and the data after MFCC processing was constructed into a three-dimensional data cube as the input data of the network. The network structure can extract the global and distributed constraint features of the fault signal, and realize the fault diagnosis of the main bearing of gas-fired generator.

3) We have verified the effectiveness of MFCC-DSFAN on three bearing datasets, which will provide some ideas for the problems of feature extraction in other fields.

TABLE 1. Abbreviation.

Full names	Abbreviations
Mel frequency cepstrum coefficient	MFCC
Deformable space-frequency attention network	DSFAN
Space-frequency feature attention mechanism	SFFAM
Convolutional neural network	CNN
Frequency attention module	FeAM
Space attention module	SaAM
Deformable convolution block	DeCB
Case Western Reserve University	CWRU
Xi'an Jiaotong University	XJTU
Discrete Fourier transform	DFT
Discrete cosine transform	DCT
Deformable convolution network	DCN
Rectified linear function	ReLU
Rolling element fault	RF
Inner race fault	IF
Outer race 3 o'clock fault	OF@3
Outer race 6 o'clock fault	OF@6
Outer race 12 o'clock fault	OF@12
Residual network	ResNet
CNN-gcForest hybrid model	CNN-gcForest
EWDCNN-LSTM hybrid method	NHDLM
Improved residual dense networks	IRDN
Hybrid multimodal fusion with deep learning	HMF-DL
Composite fault diagnosis based on ACMD, Gini Index Fusion, and AO-LSTM	ACM-GIF-AO-LSTM

The rest of this paper is described in the following way. In Section II, we introduce the work related to MFCC-DSFAN, such as deformable convolution networks. In Section III, we illustrate the method of this paper in detail. In Section IV, we show the setting and results of the experiment and analyze the results. Conclusions are drawn in Section V. Table1 lists the full names of all abbreviations in the paper.

II. RELATED WORK

A. MEL FREQUENCY CEPSTRUM COEFFICIENT

Mel frequency cepstrum coefficient (MFCC) is proposed based on the human auditory feature, which converts the signal to a frequency domain and filters it. MFCC can use the nonlinear relationship between Mel frequency and Hz frequency to obtain the Hz spectrum features with stronger robustness. Its effectiveness in voice recognition has been proved in [23]. Subsequently, MFCC is widely used in many directions of voice recognition. Compared with other kinds of acoustic features, MFCC can obtain better recognition results [24], [25]. As we all know, due to the complexity and diversity of the working environment of mechanical equipment, the collected signals are nonlinear, unstable, and affected by noise, which is disadvantageous to the realization of accurate fault diagnosis [26]. In view of the good performance of MFCC in feature extraction of acoustic signals, we use it to process the collected original fault signals and extract the signature features of the signal.

The processing steps of MFCC include pre-emphasis, framing, windowing, discrete Fourier transform(DFT), Mel band-pass filter, and discrete cosine transform(DCT). The first step is to pre-emphasis the original fault signal, which is equivalent to a high-pass filter. Pre-emphasis processing can not only amplify the high-frequency part of the signal, increase the signal-to-noise ratio of the high-frequency part, and make the spectrum of the signal flatter, but also avoid the numerical problems in the follow-up work, especially in Fourier transform. The second step is to segment the pre-emphasized signal by frame, and then apply Hamming window to each frame signal. The third step is to perform a discrete Fourier transform on each frame signal to obtain the frequency spectrum, and then calculate the power spectrum. The fourth step is to construct a group of Mel filters to make the power spectrum smoother and eliminate the effect of harmonics. The conversion between Mel frequency and actual frequency is shown in Eq.(1).

$$f_{Mel} = 2595\log_{10}(1 + \frac{f}{700}) \tag{1}$$

where f and f_{Mel} represents actual frequency and Mel frequency respectively. We know that the filter banks obtained in step 4 are overlapping. Therefore, the correlation between the power spectrum obtained by different filters is very strong, which will bring some trouble to the machine learning algorithm. We use discrete cosine transform to eliminate the correlation between power spectrums and obtain MFCC.

B. DEFORMED CONVOLUTION NETWORKS

How to adapt to the spatial transformation of targets is a very key problem in the field of visual recognition. To solve the above problems, two new modules, including deformable convolution and deformable RoI pooling, are proposed in [27], and the constructed network is called a deformable convolution network (DCN). Compared with convolutional neural network (CNN), DCN has a stronger ability for geometric transformation modeling and obtains better results in visual recognition, such as target tracking and image recognition.

In fault diagnosis, the original fault signal contains rich spatial information. We hope to extract domain spatial information useful for center category classification. Inspired by the application of DCN in visual recognition, this paper introduces deformable convolution module in DCN to further extract the spatial features between categories. Deformable convolution adds 2D offsets based on a standard convolution kernel. It can adaptively adjust the sampling position according to the characteristics of the data through offset learning, so as to capture richer features that are more conducive to classification. We assume that the input feature map is x and the regularized grid is R. The outputs of standard convolution and deformable convolution can be obtained by Eq.(2) and Eq.(3), respectively.

$$p(b_0) = \sum_{b_q \in R} w(b_q) \cdot x(b_0 + b_q) \tag{2}$$

$$p(b_0) = \sum_{b_q \in R} w(b_q) \cdot x(b_0 + b_q + \Delta b_q) \tag{3}$$

where b_0 represents every point on the feature map x; b_q enumerates the locations in R. The weight and offset of b_q are $w(b_q)$ and Δb_q, respectively. They are obtained by training.

The offset learning of deformable convolution involves a small number of parameters and calculations, and the parameters can be trained by backpropagation. The module can easily replace the standard convolution.

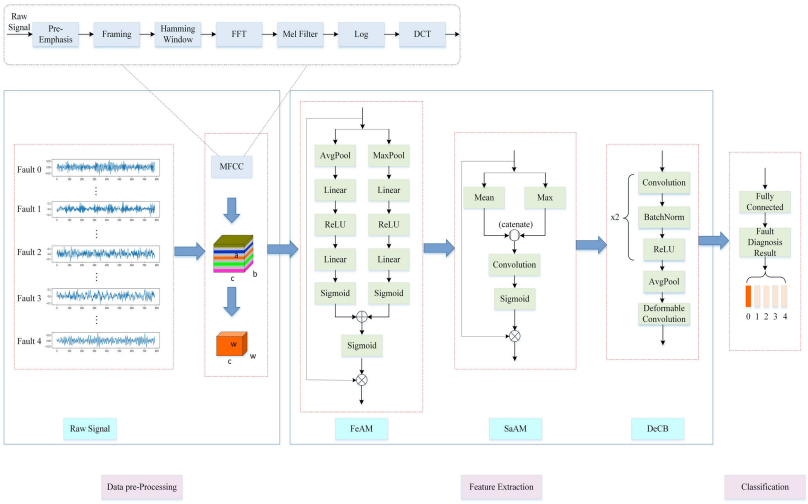


FIGURE 1. The framework of the proposed method.

III. METHODOLOGY

A. OVERVIEW OF THE PROPOSED METHODOLOGY

Through data preprocessing, redundant information of data is reduced and three-dimensional data cubes are constructed, which are more conducive to extracting useful classification features in the follow-up work. A deformable space-frequency attention network (DSFAN) is constructed to extract the space-frequency feature from the preprocessed signal data pertinently and adaptively to ensure the efficiency of diagnosis. Fig.1 shows the framework of the proposed method.

Mel frequency cepstrum coefficient (MFCC) is first used to preprocess the original one-dimensional signal, which can extract the fault signature features from the signals. The feature data obtained after MFCC processing are still one-dimensional and labeled with specific labels. To fully extract the frequency and space information of the feature data, we further process the data and construct three-dimensional data cubes. Suppose that there are $\mathcal{N}$ signal data $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\mathcal{N}\} \in \mathbb{R}^{1\times C}$ and $\mathcal{K}$ types of bearing states $\mathbb{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_\mathcal{K}\}$. Through MFCC processing, we can get $\mathcal{N}$ labeled data $\mathbb{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_\mathcal{N}\} \in \mathbb{R}^{1\times c}$. According to different categories, construct all the data into a bearing dataset block $\mathcal{H} \in \mathcal{R}^{a\times b\times c}$, where $a\times b\times c$ represents the spatial dimension of the constructed dataset block. In this block, the interval between different categories of data is 10, as shown in the gray part of the data block in Fig.1. In these intervals, the values corresponding to all data points are set to 0. Then, in the data block, we can construct three-dimensional data cubes $\mathbb{P} = [\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_\mathcal{N}] \in \mathbb{R}^{\omega\times\omega\times c}$ centered at the category data in $\mathbb{D}$, where $\omega \times \omega \times c$ represents the size of each cube. To make the label data of the edge also serve as the center and build a data cube, we fill the dataset block to obtain a new block $\mathcal{H}_1 \in \mathcal{R}^{(a+2\omega)\times(b+2\omega)\times c}$, where $(a+2\omega)\times(b+2\omega)\times c$ represents the spatial dimension of the new block, in which the filling data is all set to 0. The labels of the constructed data cubes are consistent with that of the central data.

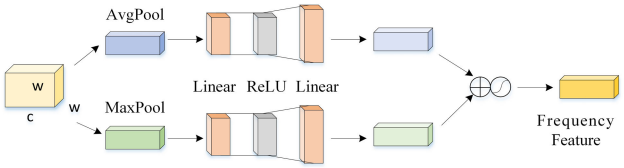


FIGURE 2. The structure of FeAM.

Data cubes are taken as the input of the DSFAN model, which can extract the space-frequency features that are helpful to the classification of the center category. The constructed frequency attention module (FeAM) is used to learn the importance of the neighborhood category data in the data block to the center category classification and give a certain weight, which can further reduce the influence of useless information. This process can extract sufficient frequency distribution features from the input three-dimensional data block. After FeAM, a spatial attention module (SaAM) is designed to adaptively learn and extract neighborhood spatial features that are useful for center category classification. A deformable convolution block can extract more detailed spatial information between different categories and refine the extracted frequency and space features. The fully connected layer is used to integrate the features obtained from the previous layer and realizes the classification of the center category of data blocks. The details of the network model are described in the following subsections.

B. FREQUENCY ATTENTION MODULE

The purpose of constructing FeAM is to adaptively focus more attention on the frequency-domain data that help to extract important frequency distribution features by learning the input data. In order to achieve this goal, we are required

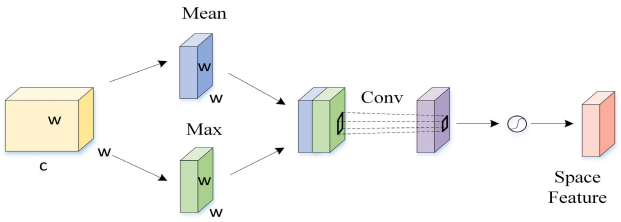


FIGURE 3. The structure of SaAM.

TABLE 2. Setting of FeAM.

Layers	Kernel number	Size	Stride	Padding
AvgPool2d	/	/	/	/
MaxPool2d	/	/	/	/
FC1	100	1×1	/	/
FC2	800	1×1	/	/

to design a feature mapping function. All frequency-domain data are adaptively assigned appropriate weights by this mapping function and a weight vector about frequency distribution features is obtained. In the process of mapping, FeAM also needs to pay attention to the relationship between different frequency-domain data, which is to avoid the loss of feature information useful for classification as much as possible. Fig.2 shows the structure of FeAM. Table2 shows the configuration of the different types of layers in FeAM. The two pooling layers are adaptive average pooling and adaptive max pooling.

To obtain the global frequency distribution features $f^{avg} \in \mathbb{R}^{1\times1\times c}$, global adaptive average pooling is used for each frequency domain feature on the $a \times b$ spatial dimension [28]. The f^{avg} of each element in the direction of c is calculated by Eq.(4).

$$f_c^{avg} = \frac{1}{a \times b} \sum_{i=1}^{a} \sum_{j=1}^{b} v_c(i, j) \tag{4}$$

where $v_c(i, j)$ is the value at point (i, j) on the v_c channel. In addition to average pooling, inspired by [29], we consider the complementary effect of global max pooling on average pooling in extracting global information. Therefore, we also use adaptive max pooling to extract global feature information from input data. Like adaptive average pooling, global adaptive max pooling operates on each frequency domain feature on $a \times b$ spatial dimension. f^{max} is calculated by Eq.(5).

$$f_c^{\max} = \max(v_c(i, j)) \tag{5}$$

To obtain data features with better expressive, limit the complexity of the model and facilitate migration applications, we introduce two fully connected layers after the pooling layers. The first fully connected layer reduces the dimension of f obtained from the pooling layer by adjusting P_1 parameters. Then, the rectified linear function (ReLU) is used to reduce the interdependence of parameters. The second fully connected layer increases the dimension of f obtained from the ReLU by adjusting P_2 parameters and then uses the sigmoid function to enhance the recognition of features. To reduce the complexity of the model and the training time, we share the parameters of two fully connected layers. The results y^{avg} and y^{max} of the two pooling branches are calculated by Eq.(6). Then, y^{avg} and y^{max} are added by Eq.(7).

$$y = F_1(f, P) = \psi\left(g(f, P)\right) = \psi\left(P_2(\phi(P_1 f))\right) \tag{6}$$

$$y = y^{avg} + y^{\max} \tag{7}$$

where ψ and ϕ represent sigmoid function and ReLU, respectively. Then we use y to rescale the input features v to get the final output of FeAM.

$$u_c = F_2(v_c, y_c) = y_c v_c \tag{8}$$

where $u = [u_1, u_2, \ldots, u_c]$ and $F(v_c, y_c)$ represent the frequency-wise multiplication of the scalar y_c and the feature map $v_c \in \mathbb{R}^{a\times b}$.

C. SPACE ATTENTION MODULE

By learning the importance of surrounding categories to the classification of the center category, SaAM can enhance the attention to the categories with the same label as the center category and reduce the attention to the categories with different labels from the center category. Therefore, the features obtained by SaAM should be consistent with the input block in height and weight. If the category at a certain position in the neighborhood has the same label as the center category, the value of this position is set to 1, otherwise, it is zero. Fig.3 shows the structure of SaAM. Table3 shows the parameters of the convolutional layer in SaAM.

As can be seen from Fig.3, SaAM first takes the mean and max values on the channel dimension of the input feature block. The results after pooling are calculated by Eq.(9) and Eq.(10).

$$h_{i,j}^{avg} = \frac{1}{c} \sum_{\eta=1}^{c} u_c(i, j) \tag{9}$$

$$h_{i,j}^{\max} = \max(u_c) \tag{10}$$

where $u_c(i, j)$ is the value at point (i, j) on the u_c channel. Unlike FeAM, in SaAM, the results after pooling are spliced horizontally, then used as the input of the convolution layer, and finally passed through a sigmoid function.

$$h = \psi([h^{avg}, h^{\max}] * P_3) \tag{11}$$

where ψ represents sigmoid function and $*$ refer to convolution operation. Then we use h to rescale the input features u_c to get the final output of SaAM.

$$u' = F_3(u, h) = hu \tag{12}$$

where $u' = [u'_{1,1}, u'_{1,2}, \ldots, u'_{i,j}]$ and $F(u, h)$ represent the frequency-wise multiplication of the scalar h and the feature map $u_{i,j} \in \mathbb{R}^{1\times1\times c}$.

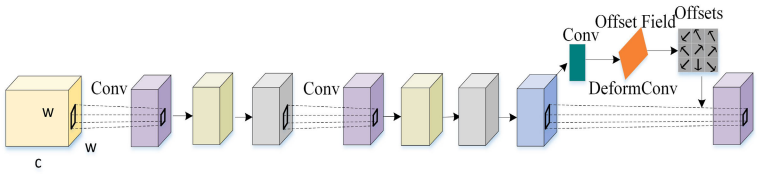


FIGURE 4. The structure of DeCB.

TABLE 3. Setting of SaAM.

Layer	Kernel number	Size	Stride	Padding
Conv2d	1	3×3	1	1

TABLE 4. Setting of DeCB.

Layers	Kernel number	Size	Stride	Padding
Conv2d	8	3×3	1	1
AvgPool2d	/	3×3	1	1
DeformConv2d	8	3×3	1	1

D. DEFORMABLE CONVOLUTION BLOCK

In view of the excellent ability of deformable convolution to extract spatial constraint information, we construct DeCB to further extract spatial information between categories and refine the extracted space-frequency features. Fig.4 shows the structure of DeCB. Table4 lists the related parameters of DeCB.

As can be seen from Fig.4, DeCB is composed of two regular convolutional layers, a deformable convolution layer, two batch normalizations, an average pooling layer, and a fully connected layer. The following equations show the calculation process of the module.

$$u^{m+1} = \phi\left(\varphi(u^m * P^{m+1} + d^{m+1})\right) \tag{13}$$

$$u^{m+2} = \phi\left(\varphi(u^{m+1} * P^{m+2} + d^{m+2})\right) \tag{14}$$

$$u^{m+3}_{s_0} = \sum_{s_n \in R} w(s_n) \cdot u^{m+2}(s_0 + s_n + \Delta s_n) \tag{15}$$

$$u^{m+4} = F_4(P^{m+3}, u^{m+3}) = P^{m+3} u^{m+3} \tag{16}$$

where φ represents batch normalization. u^m, d^{m+2} and $*$ refer to the output result of the m layer, the bias of the $m+1$ layer, and convolution operation respectively. s_0 represents every point on the feature map u^{m+3}; s_n enumerates the locations in R. The weight and offset of s_n are $w(s_n)$ and Δs_n, respectively. They are obtained by training.

IV. EXPERIMENTS

In this section, we first verify the algorithm on the CWRU and XJTU data sets. Four factors affecting the performance of the proposed model are tested and analyzed on CWRU bearing dataset and XJTU bearing dataset. The model is configured based on the results of parameter experiments and compared with several classical deep learning-based fault diagnosis algorithms on two public datasets. The influence of each module in the network model on the diagnosis performance of the model is analyzed. Then the proposed model is further tested on the gas-fired generator data set.

TABLE 5. Numbers of training and testing samples for the CWRU dataset.

Class labels	Fault types	Train	Test
0	RF	80	720
1	IF	80	720
2	OF@3	80	720
3	OF@6	80	720
4	OF@12	80	720

All experiments run on a Windows system with an Intel(R) Core(TM) I7-11800H processor, 16.0 GB of memory, and an NVIDIA GeForce RTX 3060. In addition, we use PyTorch as the deep learning framework and Python as the programming language.

A. ALGORITHM VERIFICATION

1) DATA SET DESCRIPTION

CWRU dataset: CWRU bearing data set is obtained from Case Western Reserve University bearing data center [30]. This data set is mainly for two different bearings, including drive end bearing SKF 6205 and fan end bearing SKF 6203. Five types of faults are set according to fault locations. Under the sampling frequency of 12kHz and 48kHz, the fault data are collected from the fan end, drive end, and base end. Due to the difference in fault diameters and motor speeds, each fault category often contains a large amount of fault data. The acceleration data used in this paper are the fault data of the drive end bearing collected at the 12kHz sampling frequency. There are five fault types: rolling element fault (RF), inner race fault (IF), outer race 3 o'clock fault (OF@3), outer race 6 o'clock fault (OF@6), and outer race 12 o'clock fault (OF@12), including three fault diameters: 0.1778mm, 0.3556mm, and 0.5334mm.

The CWUR data set used in this paper contains 80 training samples and 720 testing samples for each fault type. Each sample data is a data cube with a size of 11 × 11×1000. This paper only uses one channel signal collected from the sensor on the fan end. To obtain more sample data, signal data with the size of 1 × 8000 are selected repeatedly from the collected channel signals at a certain step size to prepare

TABLE 6. Numbers of training and testing samples for the XJTU dataset.

Class labels	Fault location	Bearing lifetime	Train	Test
0	Outer race	2 h 3 min	70	630
1	Outer race	2 h 41 min	70	630
2	Outer race	2 h 38 min	70	630
3	Cage	2 h 2 min	70	630
4	Outer race, Inner race	52 min	70	630

the sample dataset. MFCC is adapted to process the signal data size of 1×8000 and get the data with the size of 1×1000. According to different types, all the data processed by MFCC are constructed into a $90 \times 90 \times 1000$ dataset block. In the dataset block, the interval between different categories of data is 10. In these intervals, the values corresponding to all data points are set to 0. Because the total number of samples in each category is less than 900, there are some blanks in each type of data block. We set the data values of all blank points to 0. Then, the data cubes with the size of $11 \times 11 \times 1000$ are constructed with each labeled data as the center in this $90 \times 90 \times 1000$ dataset block. To make the label data of the edge also serve as the center and build a data cube, the $90 \times 90 \times 1000$ dataset block is filled to obtain a dataset block with the size of $100 \times 100 \times 1000$. Finally, 4000 sample data cubes with the size of $11 \times 11 \times 1000$ can be obtained. Then 10% of the labeled data cubes are taken as training samples and the rest as testing samples. The details of training samples and testing samples of each fault category are shown in Table5.

XJTU dataset: XJTU bearing data set is obtained from the Joint Laboratory of mechanical equipment health monitoring [31]. The test bearing of the data set is LDK UER204 rolling bearing. There are three working conditions (2100 r/min and 12 kN; 2250 r/min and 11 kN; 2400 r/min and 10 kN) and five bearing faults under each working condition. This data set contains two channels of data, including horizontal vibration signal and vertical vibration signal, which are collected by sensors fixed in the horizontal and vertical directions of the test bearing. In the experiment, we consider five types of faults under the 12KN working conditions and use the horizontal vibration signal in fault data. XJTU data sets are made in the same way as CWRU data sets. 1×8000 signal data is selected repeatedly from the horizontal vibration signal with a certain step size, and then MFCC is used to process it. According to different types, all the data processed by MFCC are constructed into a $90 \times 90 \times 1000$ dataset block. To make the label data of the edge also serve as the center and build a data cube, the $90 \times 90 \times 1000$ dataset block is filled to obtain a dataset block with the size of $100 \times 100 \times 1000$. Then, the data cubes with the size of $11 \times 11 \times 1000$ are constructed with each labeled data as the center in this $100 \times 100 \times 1000$ dataset block. Finally, 3500 sample data cubes with the size of $11 \times 11 \times 1000$ can be obtained. Then 10% of the labeled data cubes are taken as training samples and the rest as testing samples. The details of training samples and testing samples of each fault category are shown in Table6.

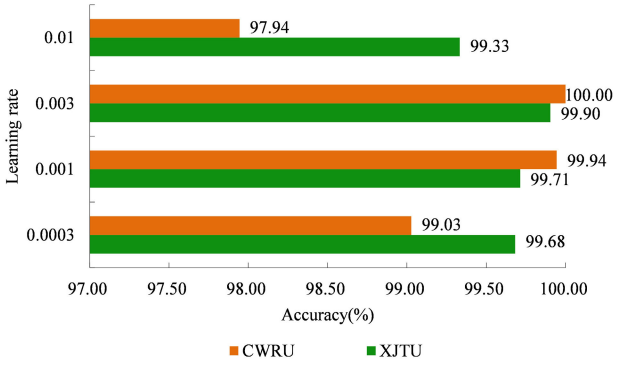


FIGURE 5. Accuracy of the proposed method under different learning rates on the CWRU and XJTU datasets.

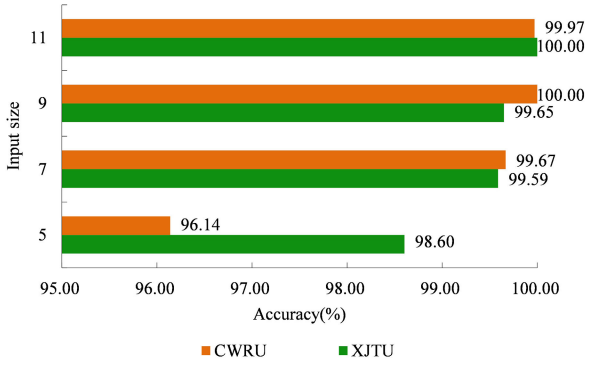


FIGURE 6. Accuracy of the proposed method under different sizes of spatial input on the CWRU and XJTU datasets.

On the CWRU and XJTU data sets, we randomly select 10% of the samples as the training samples and the remaining samples as the testing samples. The batch size of all experiments is set to 16.

2) PARAMETER SETTING

In the process of model training and testing, selecting appropriate parameters is very important to improve the diagnostic performance of the model after training. In this section, we analyze the influence of some key parameters, including the learning rate, the size of spatial input, and the scale of training samples. These parameters that need to be set manually are also called hyper-parameters. The batch size is set to 16. In each parameter experiment, we set 50 epochs and then select the model with the best classification result on the testing set for comparison.

Learning rates: The learning rate controls the learning process in the training model. Choosing the appropriate learning rate on different data sets can speed up the convergence of the model to the local minimum. Therefore, we use the parameter sweep method to select the appropriate learning rate from {0.01, 0.003, 0.001, 0.0003} on each data set. The experimental results are shown in Fig.5. According to the test

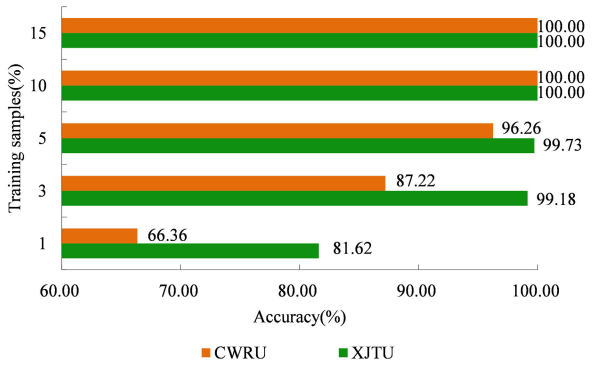


FIGURE 7. Accuracy of the proposed method under different scales of training samples on the CWRU and XJTU datasets.

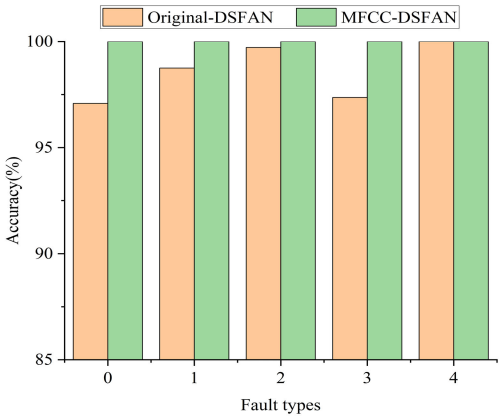


FIGURE 8. The classification accuracy of all categories.

TABLE 7. Accuracy under different input data.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
Original-DSFAN	98.38	2.621	1.692
MFCC-DSFAN	100	0	1.618

TABLE 8. Parameters of different network models.

Models	Deformable	Kernel size	Stride	Padding
CNN	False	3×3	1	1
DCN	True	3×3	1	1
ResNet	False	3×3	1	1

TABLE 9. Accuracy of different models.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
MFCC-CNN	97.17	1.636	0.884
MFCC-DCN	97.33	2.223	1.887
MFCC-ResNet	99.18	1.364	1.800
MFCC-DSFAN	100	0	1.618

results under different learning rates shown in Fig.5, we set the learning rate of CWRU and XJTU data sets to 0.003 in subsequent experiments.

Size of spatial input: The size of spatial input affects the amount of spatial information and then affects the final recognition result. Choosing the appropriate space input size can extract useful spatial information for classification and improve the efficiency of fault identification. Therefore, we set several spatial input sizes on two data sets. The results are shown in Fig.6. The input size is 11×11 on both CWRU and XJTU data sets.

Scale of training samples: Fig.7 shows the classification results of the method in this paper under different training sample proportions. As shown in Fig.7, the classification accuracy gradually improves with the increase of training samples on two data sets. When the proportion of training samples is 10%, our method can achieve 100% classification accuracy on CWRU and XJTU data sets.

3) PERFORMANCE ANALYSIS AND COMPARISONS

The method proposed in this paper mainly includes two parts: one is to pre-process the original data by Mel frequency cepstrum coefficient (MFCC); the other is to extract deep fault features and realize classification by using the deformable space-frequency attention network (DSFAN) constructed in this paper. The network model of this method is configured according to the results of the hyper-parameter experiments. To evaluate the performance of this method, we tested this method on CWRU and XJTU data sets and compared it with other methods. These methods include using data not processed by MFCC, using different network structures, and several advanced algorithms.

On the CWRU dataset, we first analyze the impact of different input data on the performance of the method proposed in this paper. The data processed by MFCC and the original data are respectively input into DSFAN for feature extraction and classification. To avoid accidental phenomena, we repeated the experiments ten times. The classification results are shown in Table7. The results are presented in the form of average±standard deviation of ten experiments. As can be seen from Table7, the average accuracy of MFCC-DSFAN reaches 100%, which is 1.62% higher than that of Original-DSFAN. Compared with Original-DSFAN, MFCC-DSFAN can obtain better diagnosis results, which proves the superiority of the data preprocessing method based on MFCC in extracting signature features of fault data and increasing the discrimination between different categories of data. Combined with the DSFAN model, MFCC-DSFAN can accurately identify the bearing fault status in a shorter time. Fig.8 shows the classification results of all categories. We can see that MFCC-DSFAN can accurately classify each category.

In order to verify the performance of our proposed DSFAN model, we compare it with several common deep neural networks in fault diagnosis, including CNN, deformable convolution network(DCN), and residual network(ResNet). CNN

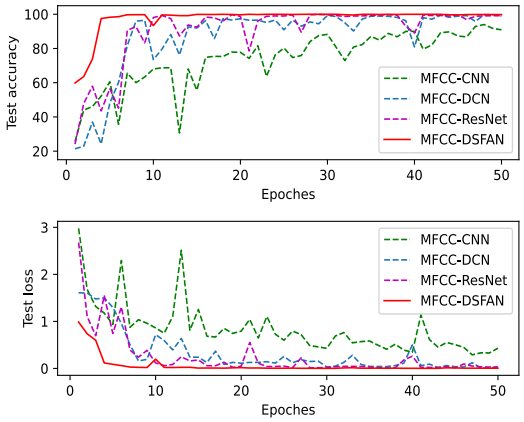


FIGURE 9. Accuracy and loss convergence with training epochs on the CWRU dataset.

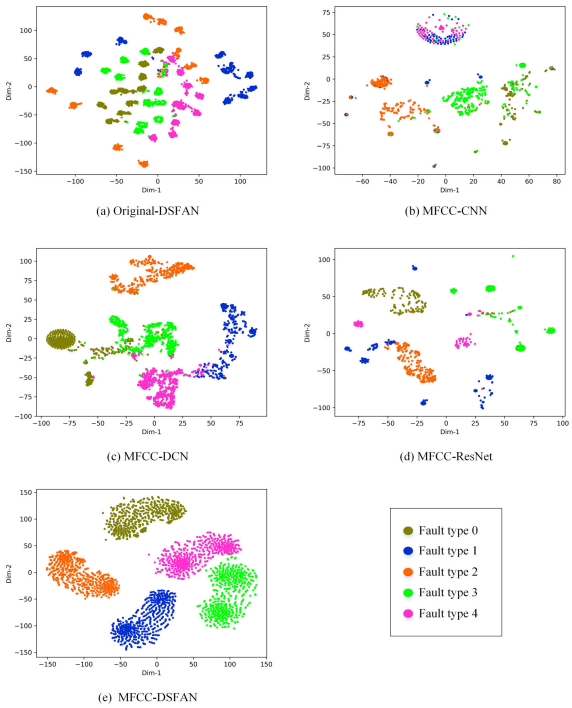


FIGURE 10. T-SNE feature visualization of different methods on the CWRU dataset.

TABLE 10. Performance comparison of different algorithms on the CWRU dataset.

Methods	CNN-gcForest	NHDLM	IRDN	MFCC-DSFAN
Accuracy(%)	99.79	98	99.24	100

model includes two convolution layers, two max pooling layers, one average pooling layer, and two fully connected layers. DCN consists of two ordinary convolution layers, two deformable convolution layers, one average pooling layer, and one fully connected layer. The 34-layer ResNet model has one max pooling layer, one average pooling layer, and one fully connected layer. In comparison experiments, the learning rates of three fault diagnosis methods based on convolutional neural network, deformable convolutional network, and residual network are 0.03, 0.01, and 0.05, respectively. Batch size and training epochs are 16 and 50, respectively. The structural parameters of the network models, including kernel size, stride, and padding, are listed in Table8. We mainly discuss the performance of network models. Therefore, the inputs of network models are all the data processed by MFCC.

Table9 shows the classification results of different network models. It can be seen from Table9 that the average accuracy of MFCC-DCN is 0.16% higher than that of MFCC-CNN. It can be analyzed that compared with CNN, DCN can learn more abundant spatial constraint features from data cubes and has better recognition ability. The spatial constraint information of data can indeed contribute to improving diagnosis accuracy, and the time-frequency features of fault data are also important. Therefore, compared with DCN, the proposed DSFAN with frequency attention and space attention modules has a 2.67% higher diagnosis accuracy and more stable performance. Moreover, DSFAN requires a shorter processing time and has higher efficiency than DCN and ResNet. These results demonstrate the effectiveness of the proposed DSFAN model in extracting deep fault features and fault recognition. Fig.9 shows the accuracy and loss convergence of different models over 50 training epochs on the CWRU data set. From Fig.9, we can see that the convergence is achieved in about 20 epochs, which proves the fast convergence of DSFAN. The experimental results show that the DSFAN model which can fully extract space-frequency features has the most stable classification performance and the highest classification efficiency.

In order to intuitively analyze the ability of all methods to extract classification features, we use T-SNE to visualize the features extracted by all methods. Feature visualization is shown in Fig.10. It can be seen from Fig.10 that the method proposed in this paper can obtain better clustering results. The spatial differentiation between different categories of feature data is great and the boundary is smooth because DSFAN uses the attention module to learn the frequency domain distribution information and spatial constraint relationship from data cubes.

The above experimental results show that the fault diagnosis method combined with MFCC and DSFAN proposed in this paper can obtain good classification performance. In order to further evaluate the performance of the method, we compare it with the state-of-art deep learning-based fault diagnosis methods. These methods include CNN-gcForest hybrid model(CNN-gcForest) [32], EWDCNN-LSTM hybrid method(NHDLM) [33], improved residual dense networks(IRDN) [34]. Table10 shows the statistical results. As can be seen from Table10, compared with

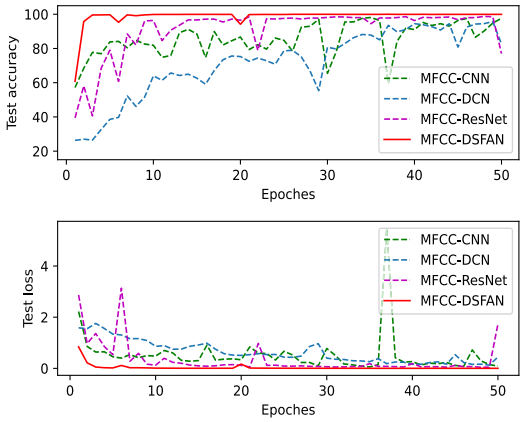


FIGURE 11. Accuracy and loss convergence with training epochs on the XJTU dataset.

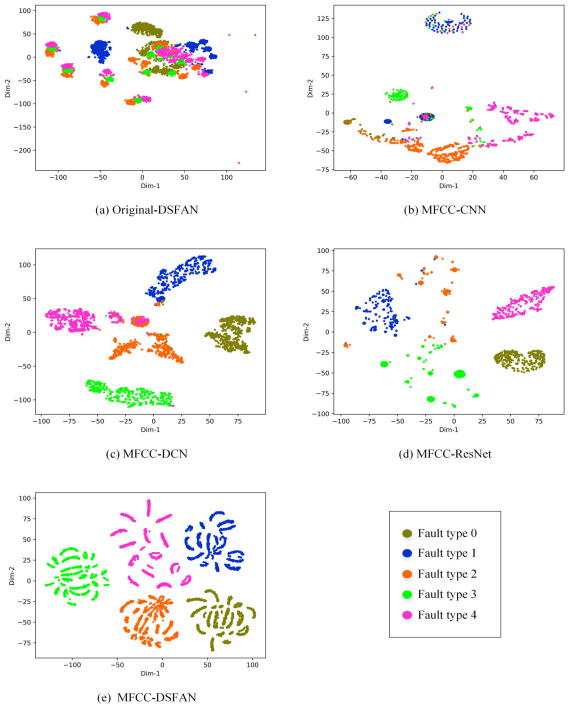


FIGURE 12. T-SNE feature visualization of different methods on the XJTU dataset.

the current advanced methods, the diagnosis results of the method proposed in this paper can achieve the highest classification accuracy.

XJTU bearing data set is used to further test the performance of the method proposed in this paper. On the XJTU dataset, we also consider the influence of input data and network model and compare the method proposed in this paper with other methods. Table11 shows the statistical results. It can be seen from Table11 that the fault diagnosis method combined with MFCC and DSFAN model proposed in this paper can obtain the highest average recognition accuracy on the XJTU data set, which is much higher than

TABLE 11. Accuracy of different methods.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
Original-DSFAN	96.95	1.687	1.356
MFCC-CNN	97.43	1.198	0.786
MFCC-DCN	96.94	2.521	1.663
MFCC-ResNet	99.47	0.418	1.530
MFCC-DSFAN	100	0	1.675

TABLE 12. Performance comparison of different algorithms on the XJTU dataset.

Methods	HMF-DL	ACM-GIF-AO-LSTM	IRDN	MFCC-DSFAN
Accuracy(%)	99.57	98.67	97.37	100

the method of using original data as DSFAN input and the methods of inputting MFCC processed data into other models, indicating that the proposed method MFCC-DSFAN has good performance in fault feature extraction and recognition. Fig.11 shows the accuracy and loss convergence of different models over 50 training epochs on the XJTU data set. From Fig.11, we can see that the convergence is achieved in about 30 epochs. Compared with other methods, the proposed method DSFAN has a faster convergence speed and more stable performance. The T-SNE feature visualization is shown in Fig.12.

On the XJTU data set, to further test the performance of the method in this paper, we compare it with the state-of-art deep learning-based fault diagnosis methods. These methods include the hybrid multimodal fusion with deep learning(HMF-DL) [35], composite fault diagnosis based on ACMD, Gini Index Fusion, and AO-LSTM(ACM-GIF-AO-LSTM) [36], improved residual dense networks(IRDN) [34]. The comparison results are shown in Table12. As can be seen from Table12, the average classification accuracy of our method is higher than that of other advanced methods.

4) ABLATION

In the above experiments, our DSFAN model shows good performance in fault classification. In this section, we analyze the impact of the attention modules on the performance of the DSFAN model. Firstly, we construct a new network model according to the presence or absence of the attention modules. Secondly, we retain the frequency attention module or the space attention module. Finally, we exchange the positions of the two attention modules, which the space attention module being placed in front of the frequency attention module. For the convenience of comparison, we take the data processed by MFCC as the input of all network models. The comparison results on CWRU and XJTU data sets are shown in Table13 and Table14 respectively.

It can be seen from Table13 and Table14 the fault recognition accuracies of the models with attention modules are higher than that of the model without attention modules, which shows that the attention modules can select the useful data and conduce to the extraction of classification features.

TABLE 13. The impact of attention modules on the performance of the proposed network model on the CWRU dataset.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
No attention	97.03	1.137	1.417
Frequency attention	98.79	0.391	1.559
Space attention	98.21	1.884	1.543
Space attention-Frequency attention	98.49	0.444	1.618
Frequency attention-Space attention	100	0	1.618

TABLE 14. The impact of attention modules on the performance of the proposed network model on the XJTU dataset.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
No attention	97.33	2.758	1.220
Frequency attention	98.97	0.399	1.727
Space attention	98.74	0.675	1.328
Space attention-Frequency attention	98.94	0.529	1.727
Frequency attention-Space attention	100	0	1.675

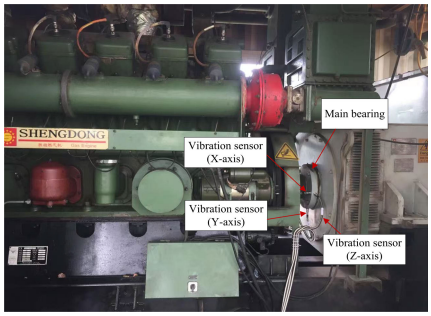


FIGURE 13. Gas-fired generator bearing test rig.

The frequency distribution feature extracted by the frequency attention module is the essential feature of fault data, which is very important for fault diagnosis. For data with more useless information, the frequency attention module can often obtain better classification results than the space attention module. However, it is not easy to accurately identify each type of fault only based on the frequency distribution features of fault data. Proper adjustment of the spatial position of the two attention modules can obtain higher fault recognition accuracy than other models. FeAM is first used to extract frequency distribution features from the constructed three-dimensional data cube, and then the SaAM is used to extract spatial constraint features, which can obtain more accurate diagnosis results. Therefore, the proposed method MFCC-DSFAN extracts the spatial constraint features between different types of data based on preserving the essential information of fault data to obtain the deep space-frequency features. Finally, the extracted space-frequency features are refined by DeCB.

B. ENGINEERING VERIFICATION

To verify the reliability of MFCC-DSFAN, experiments are carried out on the gas-fired generator dataset. The experimental parameters are set according to the parameter experiment, where the learning rate is 0.003, the network input size is 11×11, and the proportion of training samples is 10%.

TABLE 15. Numbers of training and testing samples for the gas-fired generator dataset.

Class labels	Fault level	Train	Test
0	Normal	60	540
1	First-level	84	756
2	Second-level	60	540
3	Third-level	66	594
4	Fourth-level	30	270

TABLE 16. Classification results on the gas-fired generator dataset.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
MFCC-CNN	96.64	2.727	0.387
MFCC-DCN	98.01	0.804	1.297
MFCC-ResNet	98.99	0.479	1.410
MFCC-DSFAN	100	0	1.607

1) DATA SET DESCRIPTION

Gas-fired generator dataset: The bearing data is from the gas-fired generator main bearing fault test rig. The test rig is shown in Fig.13. Three vibration sensors are arranged along the X, Y, and Z directions on the main bearing components of the gas-fired generator. The vibration sensors are used to sample the vibration signal of the main bearing of the gas-fired generator at a sampling frequency of 8KHz and a sampling period of 12 seconds.

The duration of bearing from fault occurs to complete damage is short. Fig. 14 visualizes the degradation process of bearing through the RMS curve. In order to facilitate fault maintenance and ensure the normal operation of the generator set, the working conditions of the bearing are divided into five categories according to the degree of damage: normal, first-level fault, second-level fault, third-level fault, and fourth-level fault. It can be seen from Fig. 14 that the first, second, third, and fourth level fault data correspond to the sample data of RMS curves 0-70, 73-126, 130-187, and 189-216 respectively. When the fault reaches the four-level, the bearing is

TABLE 17. The impact of attention modules on the performance of the proposed network model on the gas-fired generator dataset.

Methods	Average accuracy(%)	Standard deviation(%)	Time(s)
No attention	97.76	0.328	1.071
Frequency attention	98.72	0.193	1.517
Space attention	97.79	0.643	1.350
Space attention-Frequency attention	97.60	1.217	1.566
Frequency attention-Space attention	100	0	1.607

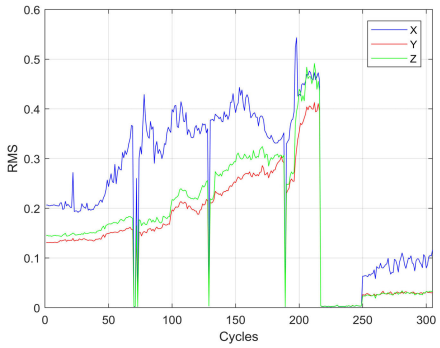


FIGURE 14. The visualization of fault stages.

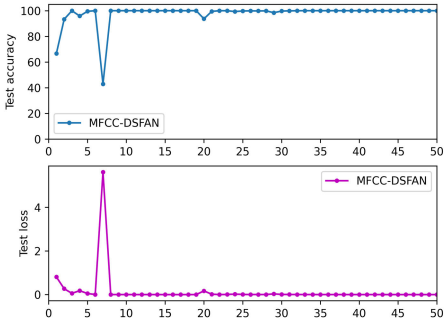


FIGURE 15. Accuracy and loss convergence with training epochs on the gas-fired generator dataset.

seriously damaged and needs to be replaced in time. It can be seen from the RMS curve that the bearing is in normal working condition at the stage of 250-300.

The vibration signals in the X direction are used as the experimental data. Taking the normal state data as an example, 1 × 1000 data can be obtained by taking the vibration signal data with the size of 1 × 8000 as input and using MFCC for processing. Other types of data are processed in the same way as normal state data. According to different types, all the data processed by MFCC are constructed into a 90 × 90×1000 dataset block. To make the label data of the edge also serve as the center and build a data cube, the 90 × 90×1000 dataset block is filled to obtain a dataset block with the size of 100 × 100×1000. Then, the data cubes with the size of 11 × 11×1000 are constructed with each labeled data as the center in this 100 × 100×1000 dataset block. Finally, we can obtain 3000 sample data cubes with the size of 11 × 11×1000. Then 10% of the labeled data cubes are taken as training samples and the rest as testing samples. The details of training samples and testing samples of each fault category are shown in Table15.

2) EXPERIMENTAL RESULT

The classification results are shown in Table16. It can be seen from Table16 that the average recognition accuracy of ten experimental results of the proposed method MFCC-DSFAN can reach 100%, and the standard deviation is 0. Compared with other algorithms based on deep learning, MFCC-DSFAN has excellent and stable diagnostic performance. Fig. 15 shows the accuracy and loss convergence of different models over 50 training epochs on the gas-fired generator dataset. From Fig. 15, we can see that the convergence is achieved in about 30 epochs. In addition, we also analyzed the effect of each module. The experimental results are shown in Table17. It can be seen from Table17 that the model based on the frequency attention-space attention module can still obtain the best diagnostic results on the gas-fired generator dataset.

V. CONCLUSION

In order to extract the signature features of different types of fault data and realize fault diagnosis, this paper proposes a bearing fault diagnosis method based on Mel frequency cepstrum coefficient (MFCC) and deformable space-frequency attention network (DSFAN). First, MFCC is used to preprocess the original fault signal and extract the signature features of different categories of data. Second, the space-frequency attention network is designed to extract frequency distribution constraint features and the global constraint features, and realize fault diagnosis. Finally, we test the performance of our method on CWRU, XJTU, and gas-fired generator datasets. The results are summarized as follows: 1) Compared with the original data, the fault data processed by MFCC is easier to be identified; 2) Compared with CNN, DCN, and other network models, DSFAN can obtain the best diagnosis results. The fault diagnosis accuracy of DSFAN can reach 100% and maintain the minimum standard deviation; 3) Compared with other fault diagnosis algorithms based on deep learning, MFCC-DSFAN still shows the best diagnosis performance; 4) The performance analysis of the attention modules shows that the model DSFAN constructed based on the frequency attention-space attention module has better

and more stable performance than other models. Most fault diagnosis methods are offline diagnosis, which is unfavorable for the timely detection of bearing faults. Therefore, we will take the realization of bearing fault online diagnosis as the research focus in future work.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments on this work.

DECLARATIONS

No potential conflicts of interest are reported by the authors.

REFERENCES

[1] L. Xu, S. Chatterton, and P. Pennacchi, "Rolling element bearing diagnosis based on singular value decomposition and composite squared envelope spectrum," *Mech. Syst. Signal Process.*, vol. 148, Feb. 2021, Art. no. 107174.

[2] X. Tang, B. Hu, and H. Wen, "Fault diagnosis of hydraulic generator bearing by VMD-based feature extraction and classification," *Iranian J. Sci. Technol., Trans. Electr. Eng.*, vol. 45, no. 4, pp. 1227–1237, Dec. 2021.

[3] S. Gao, Z. Pei, Y. Zhang, and T. Li, "Bearing fault diagnosis based on adaptive convolutional neural network with Nesterov momentum," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9268–9276, Apr. 2021.

[4] W. Q. Song, H. Liu, and E. Zio, "Long-range dependence and heavy tail characteristics for remaining useful life prediction in rolling bearing degradation," *Appl. Math. Model.*, vol. 102, pp. 268–284, Feb. 2022.

[5] Q. Liu, J. Zhang, J. Liu, and Z. Yang, "Feature extraction and classification algorithm, which one is more essential? An experimental study on a specific task of vibration signal diagnosis," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 6, pp. 1685–1696, Jun. 2022.

[6] D. Zhao, W. Cheng, R. X. Gao, R. Yan, and P. Wang, "Generalized Vold–Kalman filtering for nonstationary compound faults feature extraction of bearing and gear," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 401–410, Feb. 2020.

[7] T. S. Duan, Z. Q. Liao, T. F. Li, H. H. Tang, and P. Chen, "Bearing fault diagnosis based on state-space principal component tracking filter algorithm," *IEEE Access*, vol. 9, pp. 158784–158795, 2021.

[8] H. Shao, J. Lin, L. Zhang, and M. Wei, "Compound fault diagnosis for a rolling bearing using adaptive DTCWPT with higher order spectra," *Qual. Eng.*, vol. 32, no. 3, pp. 342–353, Jul. 2020.

[9] K. Zhang, C. Ma, Y. Xu, P. Chen, and J. Du, "Feature extraction method based on adaptive and concise empirical wavelet transform and its applications in bearing fault diagnosis," *Measurement*, vol. 172, no. 5, Feb. 2021, Art. no. 108976.

[10] M. Xia, T. Li, L. Xu, L. Liu, and C. W. De Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, Feb. 2018.

[11] X. Kong, B. Cai, Y. Liu, H. Zhu, C. Yang, C. Gao, Y. Liu, Z. Liu, and R. Ji, "Fault diagnosis methodology of redundant closed-loop feedback control systems: Subsea blowout preventer system as a case study," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 53, no. 3, pp. 1618–1629, Mar. 2023.

[12] Y. Guan, Z. Meng, D. Sun, J. Liu, and F. Fan, "2MNet: Multi-sensor and multi-scale model toward accurate fault diagnosis of rolling bearing," *Rel. Eng. Syst. Saf.*, vol. 216, Dec. 2021, Art. no. 108017.

[13] P. Liu, Y. H. Liu, B. P. Cai, X. L. Wu, K. Wang, X. X. Wei, and C. Xin, "A dynamic Bayesian network based methodology for fault diagnosis of subsea Christmas tree," *Appl. Ocean Res.*, vol. 94, pp. 101–110, Jan. 2020.

[14] K. Zhang, J. Wang, H. Shi, X. Zhang, and Y. Tang, "A fault diagnosis method based on improved convolutional neural network for bearings under variable working conditions," *Measurement*, vol. 182, no. 1, Sep. 2021, Art. no. 109749.

[15] W. Huang, J. Cheng, Y. Yang, and G. Guo, "An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis," *Neurocomputing*, vol. 359, no. 3, pp. 77–92, Sep. 2019.

[16] D. C. Li, M. Zhang, T. B. Kang, B. Li, H. B. Xiang, K. S. Wang, Z. L. Pei, X. Y. Tang, and P. Wang, "Fault diagnosis of rotating machinery based on dual convolutional-capsule network (DC-CN)," *Measurement*, vol. 187, Jan. 2022, Art. no. 110258.

[17] Z. Ye and J. Yu, "AKSNet: A novel convolutional neural network with adaptive kernel width and sparse regularization for machinery fault diagnosis," *J. Manuf. Syst.*, vol. 59, no. 2, pp. 467–480, Apr. 2021.

[18] J. C. Ma, J. A. Shang, X. Zhao, and P. Zhong, "Bayes-DCGRU with Bayesian optimization for rolling bearing fault diagnosis," *Appl. Intell.*, vol. 52, pp. 11172–11183, Jan. 2022.

[19] X. W. Xu, Z. R. Tao, W. W. Ming, Q. L. An, and M. Chen, "Intelligent monitoring and diagnostics using a novel integrated model based on deep learning and multi-sensor feature fusion," *ISA Trans.*, vol. 110, pp. 379–393, Apr. 2021.

[20] Z. Wang, Q. Liu, H. Chen, and X. Chu, "A deformable CNN-DLSTM based transfer learning method for fault diagnosis of rolling bearing under multiple working conditions," *Int. J. Prod. Res.*, vol. 59, no. 16, pp. 4811–4825, Aug. 2021.

[21] J.-R. Jiang, J.-E. Lee, and Y.-M. Zeng, "Time series multiple channel convolutional neural network with attention-based long short-term memory for predicting bearing remaining useful life," *Sensors*, vol. 20, no. 1, p. 166, Dec. 2019.

[22] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[23] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[24] S. E. Kucukbay and M. Sert, "Audio-based event detection in office live environments using optimized MFCC-SVM approach," in *Proc. IEEE 9th Int. Conf. Semantic Comput.*, Feb. 2015, pp. 6979–6983.

[25] D. Sharma and I. Ali, "A modified MFCC feature extraction technique for robust speaker recognition," in *Proc. Int. Conf. Adv. Comput., Commun. Inform.*, Aug. 2015, pp. 1052–1057.

[26] H. Liu, W. Song, Y. Zhang, and A. Kudreyko, "Generalized Cauchy degradation model with long-range dependence and maximum Lyapunov exponent for remaining useful life," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, Mar. 2021.

[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, May 2017, pp. 764–773.

[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Apr. 2020.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2018, p. 17.

[30] 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.

[31] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020.

[32] Y. Xu, Z. Li, S. Wang, W. Li, T. Sarkodie-Gyan, and S. Feng, "A hybrid deep-learning model for fault diagnosis of rolling bearings," *Measurement*, vol. 169, no. 6, Feb. 2021, Art. no. 108502.

[33] Y. Gao, C. H. Kim, and J.-M. Kim, "A novel hybrid deep learning method for fault diagnosis of rotating machinery based on extended WDCNN and long short-term memory," *Sensors*, vol. 21, no. 19, p. 6614, Oct. 2021.

[34] J. Sun, J. Wen, C. Yuan, Z. Liu, and Q. Xiao, "Bearing fault diagnosis based on multiple transformation domain fusion and improved residual dense networks," *IEEE Sensors J.*, vol. 22, no. 2, pp. 1541–1551, Jan. 2022.

[35] C. Che, H. Wang, X. Ni, and R. Lin, "Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis," *Measurement*, vol. 173, no. 7, Mar. 2021, Art. no. 108655.
[36] J. Ma and X. Wang, "Compound fault diagnosis of rolling bearing based on ACMD, Gini index fusion and AO-LSTM," *Symmetry*, vol. 13, no. 12, p. 2386, Dec. 2021.

YUNJI ZHAO received the Ph.D. degree from the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China, in 2012.

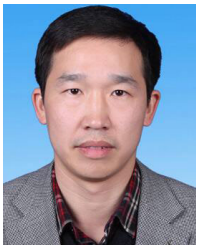
He is currently an Associate Professor with the School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China. His current research interests include pattern recognition, image processing, and artificial intelligence.

NANNAN ZHANG received the B.E. degree from the School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China, in 2020, where she is currently pursuing the M.Sc. degree.

Her research interests include pattern recognition, image processing, and target detection.

ZHIHAO ZHANG received the B.E. degree in electronic and information engineering from Xinyang Normal University, in 2019. He is currently pursuing the M.Sc. degree with the School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China.

His research interests include pattern recognition, image processing, and target detection.



XIAOZHUO XU (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Henan Polytechnic University, Jiaozuo, China, in 2003, 2006, and 2016, respectively.

He is currently an Associate Professor with the School of Electrical Engineering and Automation, Henan Polytechnic University. He is the author of three provincial awards, more than 40 articles, and more than 20 inventions. His research interests include electrical machine design and intelligent control, linear motor systems and its applications, and electromechanical fault diagnosis.