A novel bearing fault diagnosis method under small samples using time-frequency multi-scale convolution layer and hybrid attention mechanism module

IOP Publishing Journal Title
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

A novel bearing fault diagnosis method under small samples using time-frequency multi-scale convolution layer and hybrid attention mechanism module

JingsongXie[1], Mingqi Lin[1], BuyaoYang[2*], Zhibin Guo[1], Xingguo Jiang[1] , Tiantian Wang[2]

[1] School of Traffic and Transportation Engineering, Central South University, 410075, Changsha, P. R. China
[2]Mechanical and Vehicle Engineering, Hunan University, 410082, Changsha, P. R. China

Email: yangbuyaoyby@hnu.edu.cn

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Deep neural networks for bearing fault diagnosis have become the focus of research in recent years with its excellent feature extraction capability. However, the problem of diagnosis under small samples still needs to be solved in industrial applications, because bearings rarely work in the fault state in practice, resulting in the scarcity of fault data. To solve this problem, this paper proposes a new diagnosis model, a time-frequency multi-scale attention network, which structure allows the original signal and its transformed spectrum to be used as the input in parallel. A multi-scale convolutional layer is also designed to extract information from the signal at different scales to enhance the feature extraction capability of the network. In addition, a hybrid attention mechanism is added to integrate the redundant features and realize the complementarity between features. The experimental results of seven bearing diagnosis cases from two bearings show that the proposed method can achieve high diagnostic accuracy under small samples, which proves the superiority of the proposed method. The time domain signal and frequency domain signal were respectively used as input to train the model. By comparing the accuracy with the time-frequency combined signal as input, the superiority of the time-frequency domain signal as input is proved.

Keywords: Fault diagnosis; Hybrid attention mechanism Multi-scale convolution layer; Small samples

1. Introduction

Bearings are one of the important components in modern industrial equipment, widely used in various rotating machinery [1]. However, due to the long-term work of high speed, heavy load, and strong impact, bearings will cause wear, spalling, and other failures. If these failures are not handled well in time to restore the bearings to a healthy state, it will lead to the decline of the performance of mechanical equipment, and even lead to safety accidents, causing huge losses [2-4]. Based on such industrial needs, the field of mechanical fault diagnosis technology has achieved significant results in the past few decades. Fourier transform (FT), empirical mode decomposition (EMD), wavelet transform (WT), variational mode decomposition (VMD), and other signal processing methods have made a series of positive academic achievements and industrial applications when facing the problem of strong noise interference [5-8]. In recent years, with the development of artificial intelligence technologies such as machine learning (ML) and deep learning (DL), the intelligent fault diagnostic (IFD) methods of bearing failure have made great progress, which has become a popular means to solve the identification of bearing fault in the field of fault diagnosis [9-11]. However, in order to achieve the high ability of fault diagnosis, these traditional IFD methods often require a large number of labeled data that allows the models to be adequately trained.

Data is the basis for the implementation of intelligent fault diagnosis methods. Broadly speaking, there are three kinds of data that can be collected: simulation data, laboratory data, and engineering monitoring data. Although simulation data and laboratory data provide us with sufficient fault data, it is difficult to directly reflect the complex characteristics of actual machines. Meanwhile, among the actual engineering applications, the data with valid labels are very difficult to

obtain and few in number[12]. Therefore, this paper will specifically investigate the problem of few samples from the perspective of engineering applications.

Difficulties in obtaining the data lead to the fact that when facing the actual problems of industry, there is not enough labeled data to satisfy the data requirements of the model during training. Under the condition that only a few samples can be trained, the deep network cannot learn the most effective fault features, and it is easy to appear the phenomenon of over-fitting. The generation of over-fitting will reduce the model's generalization performance, resulting in a decrease in the accuracy of fault pattern recognition, and bringing great challenges for IFD methods.

The problem of fault diagnosis under small samples has attracted the attention of many researchers in recent years, and some methods have been proposed. These methods can be divided into two main categories according to different optimization objects: data-based methods (DBMs) and model-based methods (MBMs).

DBMs focus on reducing the scarcity of information under small samples to improve the model's generalization performance during the learning process, such as data augmentation (DA) and transfer learning (TL). Hu et al proposed a DA algorithm utilizing a resampling technique to simulate data under different rotating speeds and working loads, which can be regarded as a solution for both few-shot learnings as well as enhancing models' generalization ability [13].Pei et al proposed an enhanced few-shot Wasserstein auto-encoder (fs-WAE) motivated by optimal transport (OT) cost, promoting the diversity and authenticity of the generated samples [14]. Zhang et al proposed a deep learning-based synthetic over-sampling method, in which generative adversarial networks (GAN) was used to generate additional realistic fake samples and expand the available dataset afterward [15]. Zhou et al proposed another GAN-based method to synthesize fault instances, and an auxiliary loss of triplet form was introduced into the original loss function to enhance the quality of generated samples [16].

TL methods can generate additional beneficial knowledge by learning a source task similar to the target task and improving the model performance of the target task under a few samples. Chen et al proposed a hierarchy-guided transfer learning framework (HGTL) for fault recognition with few-shot samples, which extracted and transferred fault knowledge between similar tasks via transfer learning techniques [17]. Zhang et al pretrained the model by source domain samples, obtained a good feature encoder and fixed them, then fine-tuned the classifier module with a small amount of target domain data, which was a typical TL method [18]. Wu et al constructed a few-shot transfer learning method utilizing meta-learning for few-shot samples diagnosis in variable conditions, which transferred the knowledge form artificial fault bearings to natural fault bearings [19].

The focus of this article is MBMs, the purpose is to optimize the network structure to improve the feature extraction ability of the model and improve the results of the fault diagnosis. Ren et al proposed a capsule auto-encoder model, which extracted multiple meaningful feature capsules and fusion them by the dynamic routing algorithm, and reduced the dependence on the number of samples [20]. Zhang et al developed a siamese neural network model based on deep convolutional neural networks with wide first-layer kernels (WDCNN), which can acquire better feature representation [21]. Ye et al proposed a novel U-Net with CapsNet (UN-CN) to , which reduced the loss of features in the pooling process and ensured the integrity of the features to realize better results of fault diagnosis [22]. An et al proposed a few-shot fault diagnosis method for rolling bearing using local descriptors, which made full use of the lowly discriminative descriptors to improve the distinguishing ability [23]. In order to extract more effective and discriminative features, Lv et al introduced Squeeze-and-Excitation Networks (SENet) as an attention module which can enhance effective features and weaken invalid features [24]. Wang et al proposed a one-dimensional CNN with an attention mechanism (AM), which made CNN pay more attention to the interesting part of the fault signals to extract discriminative features [25]. Chen et al proposed a Transformer-based network with shifted windows, which used self-attention calculation in each non-overlapping window to improve the recognition accuracy of the model [26].

Although the above methods in the field of IFD under small samples have made a series of achievements, there are still some problems and challenges. On the one hand, the characteristics of the bearing failure have different scales. Some global features need to be detected from a larger scale perspective and some local changes require a small-scale perspective to find it in time. However, the deep neural network (DNN) represented by CNN always uses the same size convolution kernels for operation in each layer, which is inappropriate for the multi-scale features contained in the signal. The use of single-size kernels in each layer cannot extract the comprehensive fault features and affects the diagnostic performance of IFD models. On the other hand, the time domain is one of the angles of observing signals. The frequency domain can sufficiently express the periodic characteristics of rotating components such as bearings. Extracting appropriate features from a single signal domain is more difficult than the multi-signal domain, which has a greater challenge to the learning ability of the model.

To solve the above-mentioned problem, this study proposed a new time-frequency multi-scale attention network (TFMSAN) for bearing fault diagnosis under small samples. TFMSAN utilizes the time domain representation and frequency representation of the signal as the input, increasing the comprehensiveness of the signal. This kind of input makes the feature extraction ability of the model improve because the model can easier to learn effective features when the input is more comprehensive. In order to extract the multi-scale features from the input, a multi-scale parallel architecture is designed in the TFMSAN, which can perform the convolutional operation with different sizes of kernels at the same time. In addition, a hybrid attention framework (hybrid AM) has been constructed to integrate the multi-scale redundant features of the time and frequency domain in this

study. A hybrid AM includes both intra-domain and inter-domain attention mechanisms. The AM of features intra-domains and features between domains are added to the TFMSAN simultaneously, realizing the effective complementarity between different domains and ensuring the generalization ability of the TFMSAN. In general, the contributions of this paper are as follows:

1)An TFMSAN is proposed for handling the IFD problem under small samples.

2)Time domain and frequency domain representation of signals are utilized as the model's input to ensure the completeness of the information.

3)A multi-scale parallel architecture is designed to extract different scale features from signals.

4)A hybrid attention framework (AM) is constructed to integrate the redundant features and realize the complementarity between features.

The background and the principle of the proposed method are introduced in detail in Section 2, and then the structural framework of the proposed method is elaborated in Section 3. The experimental layout and the analysis of the results are described in section 4, while section 5 summarizes the results and the outlook for the future.

2. Background

2.1Convolution layer

The convolution neural network (CNN) was first proposed by LeCun et al in 1989 [27]. CNN is widely used in the fields of computer vision (CV) and natural language processing (NLP), because of its three characteristics of sparse interactions, parameter sharing, and equivariant representations, which greatly improve the network's ability to extract deep features [28]. Because the structure of CNN can automatically mine the deep abstract features of input data, it is also used in IFD recently. The convolution layer is the core part of the CNN, and gradual convolution can be performed between the input and the kernel. Assuming a 2D input I and a 2D kernel K, the process of convolution can be expressed as:

$$S(i,j)=(I*K)(i,j)=\sum_m\sum_n I(m,n)K(i-m,j-n) \quad (1)$$

Figure 1 shows the differences in convolution operations for different sizes of kernels. It can be seen that large convolution kernels have a larger sensory range and can capture features at lager scales, while small size kernels can find subtle features.

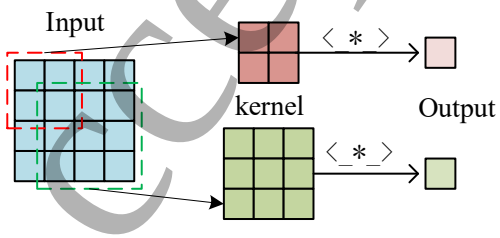


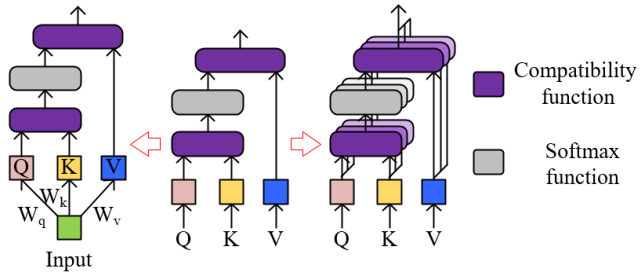
Figure 1 Convolution operations

2.2Attention Mechanism

Attention Mechanism has been formally proposed since 2014 [29], it has made great progress in the field of artificial intelligence, especially the field of NLP. The AM allows neural networks to pay more attention to the relevant information in the input and reduce the attention to unrelated information. Because of this advantage, AM also attracted the attention of many scholars in the field of fault diagnosis [30]. There are three core concepts in AM: Query, Key, and Value. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key [31]. In actual use, researchers usually do not calculate the compatibility of each query and key, but consolidate multiple queries in a matrix for calculation. AM can be expressed as formula 2, where Q is a matrix composed of some query, K means keys matrix, V means values matrix, and $\sqrt{d_k}$ is a scale factor. In this formula, the compatibility between the query and the key is to be calculated in the form of dot product.

$$\text{Attention}(Q,K,V)=\text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Unlike the lack of representation of information by a single AM, the multi-headed AM introduces multiple attention functions, enabling the model to discover interested information from multiple perspectives and obtain an extensive representation of information. Self-attention [32] is the variant of the AM, which autonomously generates the query, key, and value without relying on external information, allowing the model to notice correlations between different parts of the whole input. Self-attention is widely used in this study, to mine efficient fault features.



(a) Self-Attention (b) Attention (c) Multi-Head Attention

Figure 2 Attention mechanism and its variants

3. The proposed method

3.1Time-Frequency Multi-Scale Framework

The time domain signal provides an intuitive representation of the measurement results of the physical quantity, which accurately reflects the change in the physical quantity over time. The frequency domain provides an additional perspective to observe the signal and describes the frequency structure of the signal in detail. The frequency domain representation of a

signal can be obtained from the time domain representation by Fourier transform, as shown in equation 2.

$$X(j\omega)=\int_{-\infty}^{+\infty}x(t)e^{j\omega t}dt \tag{3}$$

It is necessary to link characteristics of the time domain and frequency domain and give a trade-off when analyzing the signal. Compared to using the time domain or frequency domain alone, using them together as an analysis object provides a more comprehensive and fuller understanding of the signal. This study constructed a time-frequency parallel architecture that uses the original time domain signal and the spectrum obtained by Fast Fourier Transform (FTT) as the network's input. The architecture can provide comprehensive time-frequency information to the network, making the network extract the sensitive fault features easier during the training process, although the features may be redundant.

Signals collected from mechanical equipment are usually complex and varied, consisting of a large number of different components and noise. The fault information of equipment is included in the multi-scale components, which makes it difficult to mine the appropriate fault feature with a single scale of convolution kernels. In this paper, a multi-scale kernel convolution network was established to obtain feature extraction results at multiple scales, by using simultaneous convolution operation between kernels of different sizes and the input signals. The multi-scale convolution process can be expressed as:

$$output=concat\{x*\omega_m\}=concat\left\{\sum_{i=1}^{k_m}x(i)\omega_m(i)\right\}, \tag{4}$$

$m=1,2,...,M$

where x is the input of the multi-scale convolution (MSConv) layer, ω_m is the m-th kernel, and M is the number of types of kernels with different sizes. In the MSConv layer, the input x convolutes with M kernels at first, and the results of convolution are concatenated into a whole tensor as the output of the MSConv layer.

In general, this paper proposed a Time-Frequency Multi-Scale Framework, namely TFMSF, which included time-frequency parallel architecture and MSConv layer, as shown in Figure 3. First, the original signal is transformed by FFT to obtain its spectrum, and then the time domain signal and spectrum are used as the input of the MSConv layer. There are three modules in MSConv: convolutional operations of different scales, batch normalization (BN) layer, and maximum pooling layer. Features from different domains and different scales, extracted by the MSConv layer, are combined into a feature vector finally in TFMSF. This framework can be used to extract the multi-scale information of the time domain and frequency domain from the original fault signal to ensure the completeness of the information and enhance the network's ability to extract the fault information under small samples.

3.2Time-Frequency Multi-Scale Attention Network

Although the network under the guidance of the TFMSF can mine multi-scale features of faults more comprehensively, the extracted features are often redundant because the information contained in the time and frequency domains is duplicated. In addition, when the mechanical equipment failure, some generated characteristics are discontinuous and periodic, not always present in signals. For example, when there is a single point defect in the outer ring of the bearing, the ball will pass the defect and produce impact vibration per turn. The traditional CNN pays the same attention to the data

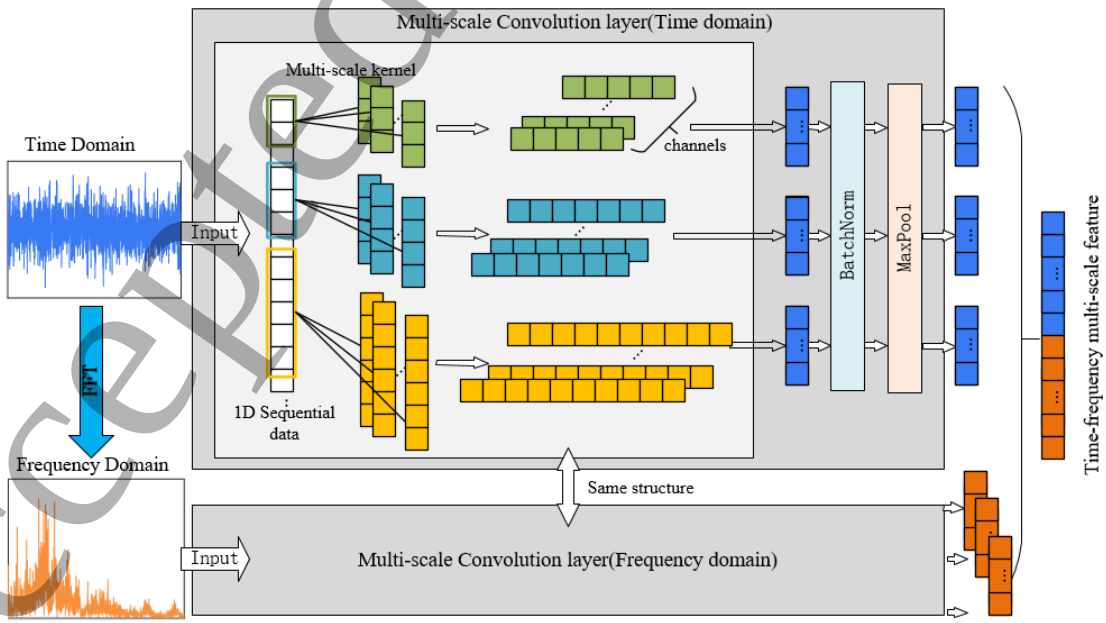


Figure 3 Time-Frequency Multi-Scale Framework

at different moments and lacks the ability to capture fault information segments, leading to the extraction of some features that are not related to the fault. The proposed TFMSAN introduced a hybrid attention mechanism, containing multiple self-attention modules. These self-attention modules can be divided into two categories in TFMSAN depending on the domain to which the input features belong. Intra-domain attention modules (Intra-AM) are used to achieve the localization of fault features in the time dimension, capturing the fault segments in the signal effectively, improving the extraction ability of discriminative features, and ignoring useless segments. Inter-domain attention modules (Inter-AM) are used to capture the intrinsic correlation between time-domain features and frequency-domain features, reducing the redundancy of features, achieving secondary selection and fusion of time-frequency multi-scale features, and improving the performance of features, as shown in Figure 4.

The original signal and its spectrum are first fed into the MSConv layer to obtain multi-scale features, after which intra-AM is used to mine the fault-sensitive features. After that, the obtained time-domain features and frequency-domain features are jointly used as the input of inter-AM to achieve fusion and enhancement in order to improve the generalization ability of the network. Finally, a classifier module made of two linear full connection layers identifies the fault classes based on the extracted multi-scale features. The classification error is calculated using the common cross-entropy loss function as shown in Equation 5, and the network is trained by updating the parameters with back-propagation techniques.

$$loss = -\left[y \log \hat{y} + (1-y) \log (1-\hat{y}) \right] \quad (5)$$

4. Experiment

In this section, seven bearing diagnostic experimental scenarios from two bearings are used to demonstrate the performance of the proposed TFMSAN.

4.1Experiment setting

4.1.1Parameters of the proposed TFMSAN

In TFMSAN, the convolutional layers are divided into three types according to the size of kernels,5×1, 9×1, and 16×1, respectively, and the number of convolution kernels is set to 32 × 1. To ensure that the same size output is obtained at different convolution scales, the step size in the convolution process is set to 1 and the padding is set to 2, 4, and 8, respectively. The number of head, embedded dimension of intra-AM is set to 2, 32, while the corresponding terms of the inter-AM is set to 4, 32. The classifier consists of two linearly connected layers with input size and output size of (1920, 300), (300, 4). The leaky rectified linear unit (LeakyReLU) activation was adopted for the whole network, and batch normalization was used for normalization. The full structural parameters of the network are shown in Table 1.

During the training process, 3 (or 5) samples of each fault type are randomly selected as the training set, and the remaining samples are used as the test set to satisfy the hypothesis of the IFD problem under small samples. Training is performed using a stochastic gradient descent (SGD)

Input	time domain samples				frequency domain samples			
MS Conv	Conv1D				Conv1D			
	Stride	1	1	1	Stride	1	1	1
	Kernel num	32	32	32	Kernel num	32	32	32
	Kernel size	5	9	16	Kernel size	5	9	16
	Batch Normalization				Batch Normalization			
	LeakyReLU (negative slope=0.5)				LeakyReLU (negative slope=0.5)			
	Max pooling (kernel size = 4)				Max pooling (kernel size=4)			
	Conv1D				Conv1D			
	Stride	2	2	2	Stride	2	2	2
	Kernel num	32	32	32	Kernel num	32	32	32
	Kernel size	5	9	16	Kernel size	5	9	16
	Batch Normalization				Batch Normalization			
	LeakyReLU (negative slope=0.5)				LeakyReLU (negative slope=0.5)			
	Max pooling (kernel size=4)				Max pooling (kernel size=4)			
Intra-AM	Num of heads:2, Embedded dimension: 32				Num of heads:2, Embedded dimension: 32			
MS Conv	Conv1D				Conv1D			
	Stride	4	4	4	Stride	4	4	4
	Kernel num	32	32	32	Kernel num	32	32	32
	Kernel size	5	9	16	Kernel size	5	9	16
	Batch Normalization				Batch Normalization			
	LeakyReLU (negative slope=0.5)				LeakyReLU (negative slope=0.5)			
	Max pooling (kernel size=4)				Max pooling (kernel size=4)			
	Conv1D				Conv1D			
	Stride	8	8	8	Stride	8	8	8
	Kernel num	32	32	32	Kernel num	32	32	32
	Kernel size	5	9	16	Kernel size	5	9	16
	Batch Normalization				Batch Normalization			
	LeakyReLU (negative slope=0.5)				LeakyReLU (negative slope=0.5)			
	Max pooling (kernel size=4)				Max pooling (kernel size=4)			
Inter-AM	Num of heads:4, Embedded dimension: 32							
Flatten layer	Flatten()							
Classifier	LeakyReLU(negative slope=0.3)							
	Linear(size of input sample=1920, size of output sample=300)							
	LeakyReLU(negative slope=0.3)							
	Linear(size of input sample=300, size of output sample=4)							
Result	Output							

Table 1 The structural parameters of the TFMSAN

*Conv1D (*in channels*, *out channels*, *kernel size*) means a 1D convolution layer with the number of channels in the input equals *in channels*, the number of channels produced by the convolution equals *out channels*, and the size of the convolving kernel equals *kernel size*.

optimizer with 1000 epochs per experiment. The initial learning rate (LR) is set to 0.001, and to ensure the convergence of the model, the learning rate decreases exponentially as the training progresses, as shown in the Equation 6:

$$lr = initial\_lr \cdot (\beta \cdot epoch + 1)^{\gamma} \quad (6)$$

Where *initial_lr* is the initial learning rate, the epoch is the completed training epochs. β and γ are two parameters that control the rate of LR's change and are empirically set to 0.01 and -0.75.

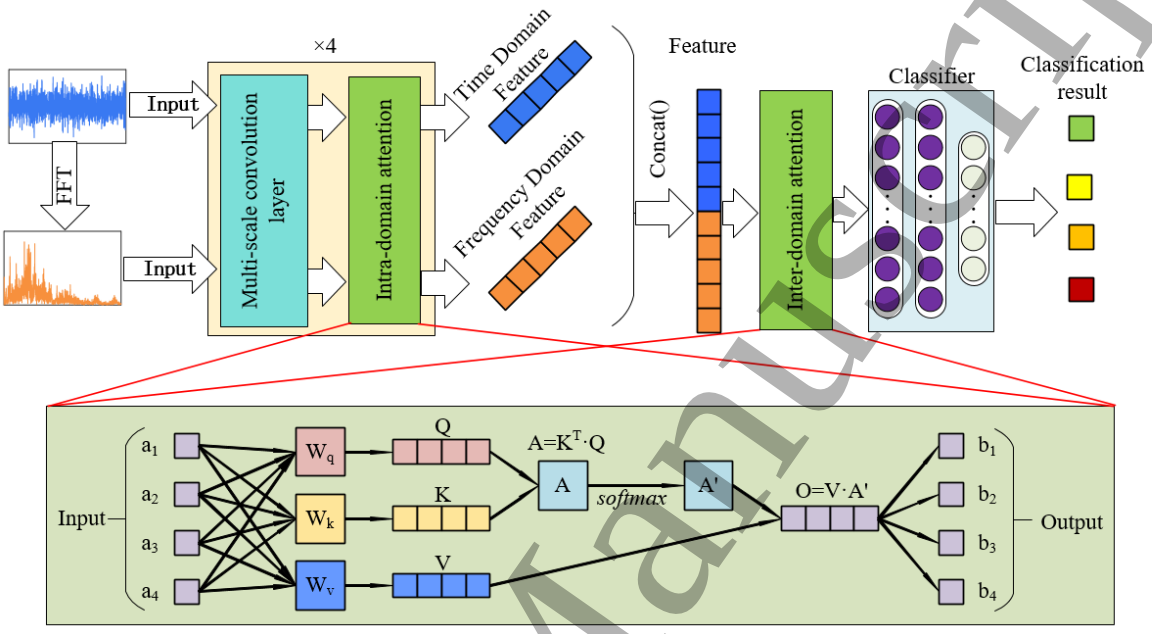


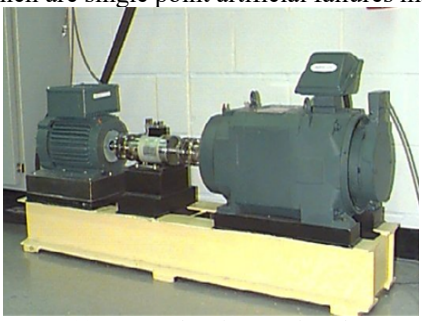
Figure 4 Time-Frequency Multi-Scale Attention Network

To validate the superiority of the proposed TFMSAN, the k-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), convolution neural network (CNN), and wide kernel CNN (WKCNN) were selected for comparison. Among them, KNN, SVM, and RF use the mature versions from the scikit-learn module.

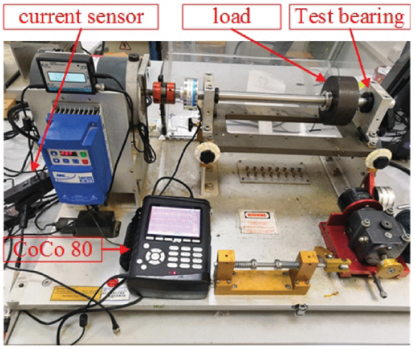
The CNN uses a convolutional structure similar to that of the TFMSAN, containing four convolution layers, but using only the original signal as the input. WKCNN utilizes a wider convolutional kernel to extract one-dimensional signal features more efficiently [33]. All these comparison methods were fine-tuned to achieve the best experimental results on the data set used in this paper.

4.1.2Dataset description

The data sets used in this study were collected from two different bearing failure simulation experimental benches to validate the proposed method. The data set used in this paper includes four data types, namely, bearing data under the condition of health and bearing data under the condition of inner ring failure, outer ring failure, and ball failure. The first data set is from the Case Western Reserve University(CWRU) Bearing Data Center. As shown in Figure 5(a), the bearing test bench includes a two-horsepower motor, a torque sensor, a power meter, and an electronic control system. The bearing type under test is 6205-2RS JEM SKF. Faults of bearing include inner ring failures, outer ring failures, and ball failures, which are single point artificial failures machined at



(a) CWRU bearing test bench



(b) SQ bearing test bench

Figure 5 The construction of bearing test benches

the corresponding locations respectively. The motor load includes 0-3 hp and the speed distribution is between 1720 and 1797 rpm. The vibration signal used in this paper is collected from the drive-side bearing measurement point with a sampling frequency of 12 kHz, for more detailed information see [34].

The second data set is from the Spectra Quest (SQ)test bench, as shown in Figure 5(b). It consists of a single-phase asynchronous motor as the power output, a heavy load of 5kg, and the test bearing on the right side of the rig. The type of bearing under test is ER16K. Inner ring failures, outer ring failures, and rolling element failures were manufactured in the test bearing manually, the same as the first data set. The speed is divided into three types: 300, 600, and 900rpm and the sampling frequency is 51.2kHz.

Table 2 Details of the experimental data

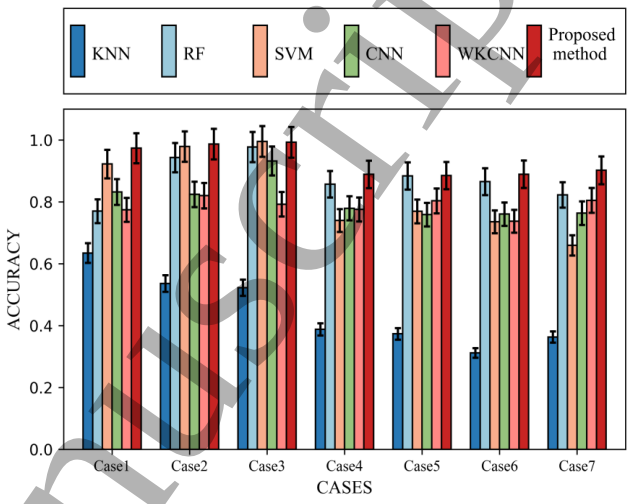
Cases	Test-rig	Category of Training/Test samples	Num of training samples (Each Category)	Num of Test samples	Load /HP	Speed /RPM
Case1	CWRU	Normal(N), Inner ring(I), Outer ring(O), Ball(B)	3/5	3/5	0	1797
Case2					1	1772
Case3					2	1750
Case4					3	1730
Case5	SQ			N(1203), I(876), O(876), B(876)	0	300
Case6					0	600
Case7				N(1204), I(876), O(876), B(876)	0	900

In order to verify the effectiveness of the proposed method under small samples, data from two bearing experimental benches were divided into seven cases depending on the speed and load. Each situation includes normal data of bearings, inner ring fault data, outer ring fault data, and rolling element fault data. All of the data were divided by time windows of length 2560 and steps 400 in this study. In each case, 3 (or 5) samples of each fault type were randomly selected as the training set, and the remaining samples are used as the test set to simulate small sample scenarios. The details of the experimental data are shown in Table 2.

4.2Experiment results

4.2.1Performance of models under 3 training samples

During the experiment, we tested 5 machine learning and deep learning algorithms such as RF, and CNN as controls, and finally obtained the experimental results of TFMSAN, and the comparison methods under seven cases are shown in Figure 6. To avoid inaccurate comparison results due to the specificity of the selected training samples, the final results show the mean and variance of 5 experiments. As can be seen from the figure, the proposed method achieves the best diagnostic results for these seven different bearing failure cases, which proves the effectiveness of the proposed method. In case 2 and case 3, there is no significant difference between SVM and the



*Proposed method refers to TFMSAN

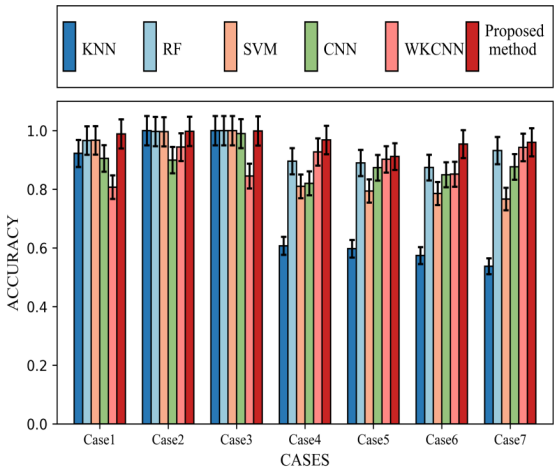
Figure 6 The diagnostic results of the different methods using 3 training samples

proposed method because they are both close to 100%, while in other cases the proposed method has a significantly better diagnostic effect than SVM, especially for SQ bearings. RF is the second-best performing method in cases 4, 5, 6, and 7 after the proposed method, presumably because the integration property in RF makes the model better resistant to overfitting, which can get better results under significant noise. The best diagnostic results were obtained in these cases, which also demonstrated the excellent interference resistance and generalization performance of the proposed method. CNN and WKCNN perform similarly in these cases, significantly lower than the proposed methods and traditional machine learning methods such as RF and SVM. We speculate that the reason for these phenomena is that CNN and WKCNN, based on deep neural networks and using raw signals as input to extract features, which lead to insufficient generalization of the learned features when it is difficult to obtain sufficient diagnostic knowledge in scenarios with small samples.

4.2.2Performance of models under 5 training samples

The diagnostic results of methods on seven different cases under the condition that training with 5 samples are shown in Figure 7. From the figure, we can find that in case 2 and case 3, KNN, RF, and SVM achieve similar results with the proposed method, which are all very close to 100%. In other cases, the proposed method achieves the best diagnostic results. Especially, in cases 4, 5, 6 and 7, the performance of the comparison methods declined more obviously, while the proposed method still achieved good diagnostic accuracy, which exceeded 90% in all cases, reflecting the excellent

feature extraction ability of the proposed method in different conditions.



*Proposed method refers to TFMSAN

Figure 7 The diagnostic results of the different methods using 5 training samples

Through the experimental results of the previous two different training samples, we can prove the superiority of TFMSAN. For this phenomenon, we believe it is because the input signals in the time domain and frequency domain are respectively input into the MSConv layer to obtain multi-scale features. Then, effective fault features can be mined through the intra-domain attention mechanism, and the time domain and frequency domain features are combined and the features are fused and enhanced through the inter-domain attention mechanism so that the classification can be made according to the multi-scale features. Thus, the proposed method has the highest accuracy.

4.2.3Performance of using time and frequency domain signals as input signals respectively

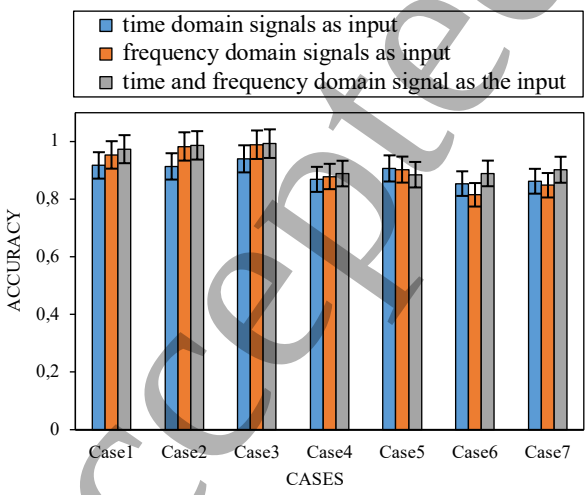


Figure 8 The diagnostic results for different input signals using 3 training samples

Table 3 Results of 3 different input signals under 3 training samples (%)

*Increase1=Time-frequency domain -Time domain

Cases	Time domain	Frequency domain	Time-frequency domain	Increase 1	Increase 2
Case1	91.7224	95.3618	97.3566	5.6342	1.9948
Case2	91.3842	98.3671	98.6770	7.2828	0.2999
Case3	94.1253	98.9136	99.2682	5.1429	0.3546
Case4	86.8847	87.8963	88.9000	2.0153	1.0037
Case5	90.6755	90.2734	88.5027	-2.1728	-1.7707
Case6	85.4123	81.5346	88.9437	3.5314	7.4091
Case7	86.2348	84.8491	90.2140	3.9792	5.3649
Mean				5.0826	2.9313

Increase2= Time-frequency domain - Frequency domain

Table 3 and Figure 8 show the experimental results of using the time domain signal, the frequency domain signal, and the combined time and frequency domain signal as the input signal respectively under 3 training samples. From the results, it can be seen that using time-frequency combined signals as input has an average improvement of 5.0826% in accuracy compared to using only time-domain signals as input, and at the same time, it has an average improvement of 2.9313% in accuracy compared to using only frequency-domain signals as input. This can prove that using time-frequency combined signals as input can provide multi-dimensional features for fault diagnosis, thereby improving the model's feature extraction ability under small sample conditions. According to this, we can get that time-frequency signals have good Local properties and adaptability to different scales, and can simultaneously characterize the time-domain characteristics and frequency-domain characteristics of signals.

4.2.4 Performance of the model with and without hybrid Attention Mechanism.

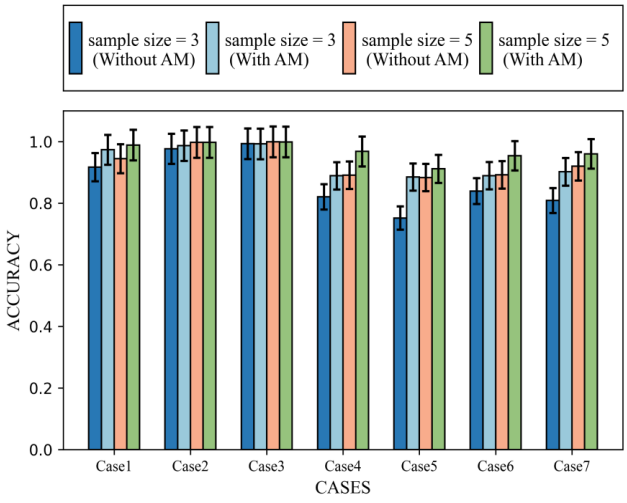


Figure 9 The diagnostic results for performance of hybrid AM

Table 4 Results with and without hybrid AM (%)

Cases	3 samples (Without AM)	3 samples (With AM)	increase	5 samples (Without AM)	5 samples (With AM)	increase
Case1	91.7461	97.3566	5.6105	94.4611	98.8966	4.4354
Case2	97.6505	98.6770	1.0266	99.7647	99.7572	-0.0076
Case3	99.3274	99.2682	-0.0592	99.9508	99.9093	-0.0415
Case4	82.1059	88.9000	6.7940	89.1127	96.8295	7.7169
Case5	75.1999	88.5027	13.302	88.3574	91.1689	2.8115
Case6	83.9316	88.9437	5.0120	89.2533	95.3990	6.1457
Case7	80.8851	90.2140	9.3289	91.9950	96.0255	4.0305
mean			5.8594			3.5844

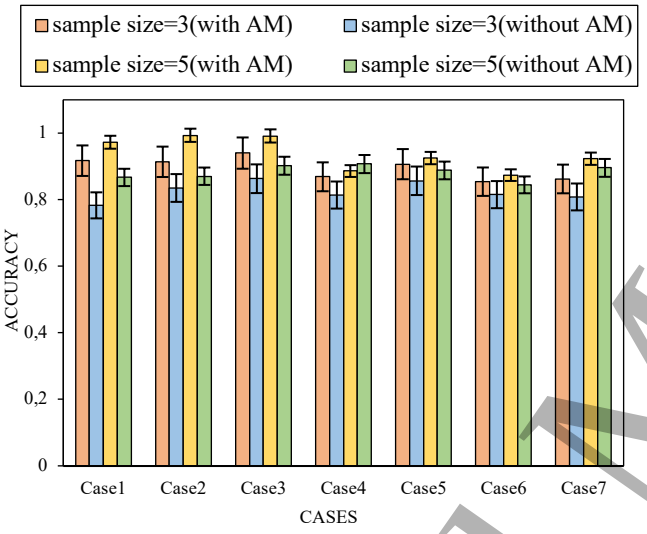


Figure 10 The diagnostic results of using time domain signals as input

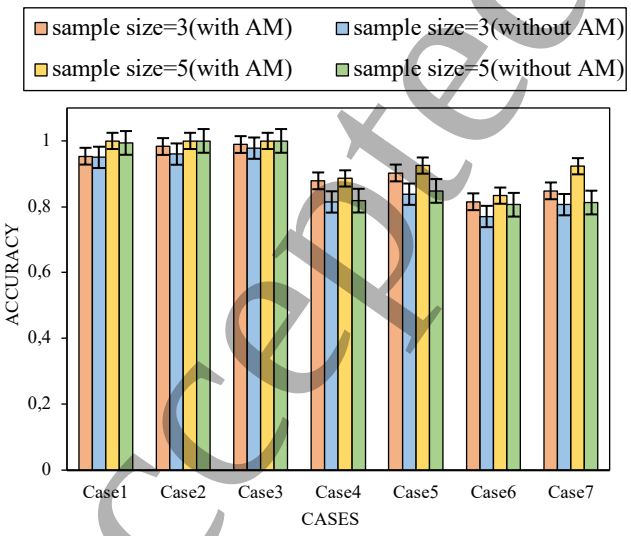


Figure 11 The diagnostic results of using frequency domain signals as input

Table 4 and Figure 9 show the experimental results with and without the addition of hybrid AM under the conditions of 3 training samples and 5 training samples. From them, we can get that adding AM can effectively improve the fault diagnosis results of the model. The average improvement is 5.8594% under three training samples and 3.5844% under five training samples. Such results demonstrate that AM can effectively improve the performance of the extracted features and enhance the generalization of the model under small samples.

Using only time domain or frequency domain signals as input, Figure 10 and Figure 11 show the experimental results with and without the addition of hybrid AM under the conditions of 3 training samples and 5 training samples. As can be seen from the two bar charts, no matter the time domain signal or frequency domain signal as input, the diagnostic accuracy of most models without mixed attention mechanism is lower than that of models with mixed attention mechanism. These results show that the hybrid AM can effectively improve the feature extraction performance for time domain, frequency domain, and time-frequency domain signals under appropriate conditions, so as to improve the fault diagnosis accuracy of the model under the condition of small samples.

5. Conclusion

In this paper, a new IFD model, TFMSAN, is proposed to solve the problem of poor generalization ability under small samples in real industrial environments. A neural network framework is constructed, namely TFMSF, to realize the parallel extraction of time- and frequency-domain multi-scale features to enhance the information extraction ability of the model. The framework uses multiple kernels of different sizes to build the MSConv layer to achieve different scales of convolution. In addition, two structurally identical branches in the framework are used to extract features in the time and frequency domains respectively. Meanwhile, a hybrid AM is introduced into the model to mine for more effective and focused features. Intra-AM and inter-AM are used for feature fusion of one domain and different domains, to integrate the redundant features and realize the complementarity between features. The experimental results prove the superiority of the proposed method, and the following three conclusions can be drawn.

1) The comparison results with the five other methods prove the superiority of TFMSAN for extracting efficient fault features, and show the effectiveness of the proposed method for fault diagnosis with small samples.

2) Experimental results in seven cases of bearing diagnosis experiments from two bearings demonstrate the reliability and generalization of the proposed method, as the proposed TFMSAN achieves the best or near best fault diagnosis accuracy in all these cases which is generally better than comparison methods.

3)By comparing the diagnostic accuracy of time-domain signal and frequency-domain signal as input with the

diagnostic results of time-frequency signal as input, it can be proved that the model can obtain more comprehensive and effective fault features when time-frequency signal is used as input.

4) The comparison results of the model with AM and the model without AM show the role of the hybrid AM, which can significantly improve the efficient representation and generalization performance of features.

In this study, it is assumed that all failure types have the same number of samples, but in the actual industry, the probability of occurrence of various failure types is not the same, resulting in an imbalance between the various types of samples. What's more, the proposed method still requires complete categories of data, and it is still impossible to effectively diagnose data sets that lack a certain fault category which is common in industry applications. How to make full use of the unbalanced small sample data to ensure the validity of the model is the focus of the next study.

6. Acknowledgements

Projects(P2021J036) were supported by the China National Railway Group Limited; Projects(No.2021JJ40765) were supported by the Natural Science Foundation of Hunan Province China; Projects(2020QNRC001) were supported by the Young Elite Scientists Sponsorship Program by CAST. The authors would like to thank the above funding agencies.

7. References

[1] Z. Wang, J. Zhou, W. Du, Y. Lei, and J. Wang, "Bearing fault diagnosis method based on adaptive maximum cyclostationarity blind deconvolution," *Mechanical Systems and Signal Processing,* vol. 162, p. 108018, 2022.

[2] J. Li *et al.*, "Research on rolling bearing fault diagnosis based on multi-dimensional feature extraction and evidence fusion theory," *Royal Society open science,* vol. 6, no. 2, p. 181488, 2019.

[3] H. Li, T. Liu, X. Wu, and Q. Chen, "An optimized VMD method and its applications in bearing fault diagnosis," *Measurement,* vol. 166, p. 108185, 2020.

[4] Y. Yu and C. Junsheng, "A roller bearing fault diagnosis method based on EMD energy entropy and ANN," *Journal of sound and vibration,* vol. 294, no. 1-2, pp. 269-277, 2006.

[5] J. Burriel-Valencia, R. Puche-Panadero, J. Martinez-Roman, A. Sapena-Bano, and M. Pineda-Sanchez, "Short-frequency Fourier transform for fault diagnosis of induction machines working in transient regime," *IEEE Transactions on Instrumentation and Measurement,* vol. 66, no. 3, pp. 432-440, 2017.

[6] D. Yu, J. Cheng, and Y. Yang, "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings," *Mechanical systems and signal processing,* vol. 19, no. 2, pp. 259-270, 2005.

[7] Z. Zhang, Y. Wang, and K. Wang, "Fault diagnosis and prognosis using wavelet packet decomposition, Fourier transform and artificial neural network," *Journal of Intelligent Manufacturing,* vol. 24, no. 6, pp. 1213-1227, 2013.

[8] S. Mohanty, K. K. Gupta, and K. S. Raju, "Comparative study between VMD and EMD in bearing fault diagnosis," in *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, 2014, pp. 1-6: IEEE.

[9] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mechanical systems and signal processing,* vol. 21, no. 6, pp. 2560-2574, 2007.

[10] F. Lv, C. Wen, Z. Bao, and M. Liu, "Fault diagnosis based on deep learning," in *2016 American control conference (ACC)*, 2016, pp. 6851-6856: IEEE.

[11] R. Surendran, O. I. Khalaf, and C. Andres, "Deep learning based intelligent industrial fault diagnosis model," *CMC-Computers, Materials & Continua,* vol. 70, no. 3, pp. 6323-6338, 2022.

[12] Wang S , Wang D , Kong D , et al. Few-Shot Rolling Bearing Fault Diagnosis with Metric-Based Meta Learning.[J]. Sensors, 2020(22).

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

[14] Z. Pei, H. Jiang, X. Li, J. Zhang, and S. Liu, "Data augmentation for rolling bearing fault diagnosis using an enhanced few-shot Wasserstein auto-encoder with meta-learning," *Measurement Science and Technology,* vol. 32, no. 8, p. 084007, 2021.

[15] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement,* vol. 152, p. 107377, 2020.

[16] Y. Zhuo and Z. Ge, "Auxiliary information-guided industrial data augmentation for any-shot fault learning and diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 17, no. 11, pp. 7535-7545, 2021.

[17] H. Chen, R. Liu, Z. Xie, Q. Hu, J. Dai, and J. Zhai, "Majorities help minorities: Hierarchical structure guided transfer learning for few-shot fault recognition," *Pattern Recognition,* vol. 123, p. 108383, 2022.

[18] Y. Zhang, S. Li, A. Zhang, C. Li, and L. Qiu, "A Novel Bearing Fault Diagnosis Method Based on Few-Shot Transfer Learning across Different Datasets," *Entropy,* vol. 24, no. 9, p. 1295, 2022.

[19] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Few-shot transfer learning for intelligent fault diagnosis of machine," *Measurement,* vol. 166, Dec 15 2020, Art. no. 108202.

[20] Z. Ren *et al.*, "A novel model with the ability of few-shot learning and quick updating for intelligent fault diagnosis," *Mechanical Systems and Signal Processing,* vol. 138, p. 106608, 2020.

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

[22] X. Ye, J. Yan, Y. Wang, J. Wang, and Y. Geng, "A novel U-Net and capsule network for few-shot high-voltage circuit breaker mechanical fault diagnosis," *Measurement,* vol. 199, p. 111527, 2022.

[23] L. An, F. Jia, B. Wang, J. Hou, J. Shen, and X. Song, "Few-shot fault diagnosis method for rolling bearing using local descriptors," in *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2022, pp. 1381-1386: IEEE.

[24] Q. Lv and Y. Song, "Few-shot learning combine attention mechanism-based defect detection in bar surface," *ISIJ International,* vol. 59, no. 6, pp. 1089-1097, 2019.

[25] Y. Wang, J. Yan, X. Ye, Q. Jing, J. Wang, and Y. Geng, "Few-shot transfer learning with attention mechanism for high-voltage circuit breaker fault diagnosis," *IEEE Transactions on Industry Applications,* vol. 58, no. 3, pp. 3353-3360, 2022.

[26] Z. Chen, J. Chen, S. Liu, Y. Feng, S. He, and E. Xu, "Multi-channel Calibrated Transformer with Shifted Windows for few-shot fault diagnosis under sharp speed variation," *ISA transactions,* 2022.

[27] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural computation,* vol. 1, no. 4, pp. 541-551, 1989.

[28] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, 2017, pp. 1-6: Ieee.

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[30] Z. Xu, C. Li, and Y. Yang, "Fault diagnosis of rolling bearings using an improved multi-scale convolutional neural network with feature attention mechanism," *ISA transactions,* vol. 110, pp. 379-393, 2021.

[31] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *Transactions*

of the Association for Computational Linguistics, vol. 4, pp. 259-272, 2016.
[32] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[33] X. Song, Y. Cong, Y. Song, Y. Chen, and P. Liang, "A bearing fault diagnosis model based on CNN with wide convolution kernels," *Journal of Ambient Intelligence and Humanized Computing,* vol. 13, no. 8, pp. 4041-4056, 2022.
[34] Case Western Reserve University Bearing Data Center, 2014

Jingsong Xie was born in Anren, Hunan, China, in 1989. He received the B.S. degree from the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China, in 2013, and the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, in 2018.
He joined the School of Traffic and Transportation Engineering, Central South University, Changsha, China, as a Lecturer. His research interests include fault diagnosis, machine learning, vibration analysis, and crack diagnosis.

Buyao Yang received the M.Sc dgree in mechanical engineering from Zhengzhou University, Zhengzhou, China. He is currently pursuing the D.E. degree in mechanical engineering with Hunan University, Changsha, China.His current research interests include rotating machine diagnosis, health management of High speed train bogie and transfer learning.

Tiantian Wang received the bachelor's and Ph.D. degrees from Beihang University, Beijing, China, in 2012 and 2018, respectively. He is currently a Vice Professor with Central South University and Hunan University. His current research interests include vehicle aerodynamics and vehicle structure, especially train/tunnel aerodynamics, and PHM for trains

