Limited Data Rolling Bearing Fault Diagnosis With Few-Shot Learning

ANSI ZHANG[1,2], **SHAOBO LI**[1,3,4], **YUXIN CUI**[2], **WANLI YANG**[3],
RONGZHI DONG[3], **AND JIANJUN HU**[2,3]

[1]Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University, Guiyang 550025, China
[2]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
[3]School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
[4]Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

Corresponding authors: Shaobo Li (lishaobo@gzu.edu.cn) and Jianjun Hu (jianjunh@cse.sc.edu)

ABSTRACT This paper focuses on bearing fault diagnosis with limited training data. A major challenge in fault diagnosis is the infeasibility of obtaining sufficient training samples for every fault type under all working conditions. Recently deep learning based fault diagnosis methods have achieved promising results. However, most of these methods require large amount of training data. In this study, we propose a deep neural network based few-shot learning approach for rolling bearing fault diagnosis with limited data. Our model is based on the siamese neural network, which learns by exploiting sample pairs of the same or different categories. Experimental results over the standard Case Western Reserve University (CWRU) bearing fault diagnosis benchmark dataset showed that our few-shot learning approach is more effective in fault diagnosis with limited data availability. When tested over different noise environments with minimal amount of training data, the performance of our few-shot learning model surpasses the one of the baseline with reasonable noise level. When evaluated over test sets with new fault types or new working conditions, few-shot models work better than the baseline trained with all fault types. All our models and datasets in this study are open sourced and can be downloaded from https://mekhub.cn/as/fault_diagnosis_with_few-shot_learning/.

INDEX TERMS Deep learning, few-shot learning, bearing fault diagnosis, limited data.

I. INTRODUCTION

Fault diagnosis is widely applied in diverse areas such as manufacturing, aerospace, automotive, power generation, and transportation [1]–[4]. Recently, intelligent fault diagnosis techniques with deep learning have attracted a lot of attention due to their avoidance of dependency on the time-consuming and unreliable human analysis and increased efficiency in fault diagnosis [5], [6]. However, most of these techniques require a large amount of training data. In real-world fault diagnosis, the signals of the same faults often bear large difference between different working conditions, leading to a major challenge in fault diagnosis: it is often impossible to obtain sufficient samples to make the classifier robust for every fault types. This situation can arise for several reasons: (1) industry systems are not allowed to run into faulty states

due to the consequences, especially for critical systems and failures; (2) most electro-mechanical failures occur slowly and follow a degradation path such that failure degradation of a system might take months or even years, which makes it difficult to collect related datasets [7]. (3) working conditions of mechanical systems are very complicated and frequently change from time to time according to production requirements. It is unrealistic to collect and label enough training samples [8]. (4) especially in real-world applications, fault categories and working conditions are usually unbalanced. It is thus difficult to collect enough samples for every fault type under different working conditions.

There are some studies about limited data fault diagnosis. Hang *et al.* [9] proposed the principal component analysis (PCA) and applied it in the field of high-dimensional unbalanced fault diagnosis data. Duan *et al.* [10] applied a new support vector data description method for machinery fault diagnosis with unbalanced datasets. In recent years,

The associate editor coordinating the review of this article and approving it for publication was Li Zhang.

deep learning has achieved impressive results in application areas such as computer vision, image and video processing, speech recognition, and natural language processing [11]. Deep learning methods have also been applied to fault diagnosis and obtained state-of-the-art results [6] using techniques such as auto-encoders (AE) [12]–[15], restricted boltzmann machine (RBM) [16]–[19], convolutional neural networks (CNNs) [20]–[24], recurrent neural networks (RNNs) [25]–[27], transfer learning based neural networks [28]–[31]. and generative adversary networks GANs e.g. Sun *et al.* [12] applied sparse auto-encoder (SAE) neural network and achieve superior performance for feature learning and classification in the field of induction motor fault diagnosis. Li *et al.* [16] applied two-layer deep Boltzmann machines (DBMs) to learn deep representations of the statistical parameters for gearbox fault diagnosis. Zhang *et al.* [22] proposed CNN with Training Interference (TICNN) for bearing fault diagnosis whose input is raw time-series signal. Zhao *et al.* [27] designed a deep learning model named Convolutional Bi-directional Long Short-Term Memory Networks (CBLSTM) by combining CNN and LSTM. To address the varying working conditions issues in fault diagnosis, Shao *et al.* [14] proposed a transfer learning based approach for fault diagnosis. Lu *et al.* [30] applied a novel deep neural network model with domain adaptation for fault diagnosis. Yang *et al.* [32] proposed a feature-based transfer neural network (FTNN). To address the unbalanced data in fault diagnosis, Cabrera *et al.* [33] use GANs model to assess the data distribution for every minority faulty mode to synthetically increase its size.

The ability of deep neural networks to learn low-level and high-level features from abundant datasets has been well known and has been widely exploited in fault diagnosis [5], [6]. However, except transfer learning based methods which can address the time-changing working conditions issue and GANs methods which can address the unbalanced data issue, most of these intelligent fault diagnosis methods based on deep neural networks have not addressed one of the major challenges: limited fault samples for one or more fault types under all working conditions.

Recently, few-shot learning based deep neural networks have made great progress in addressing the data scarcity issue [34]–[37] and has become an exciting field of machine learning. Few-shot learning was first addressed in the 1980s [38]. Fei-Fei *et al.* [34] developed a variational Bayesian framework for one-shot image classification using the premise that previously learned classes could be leveraged to help forecast future ones when very few examples are available from a given category. Wolf *et al.* [35] chose to focus on a metric learning approach using a standard bag of features representation to learn a similarity kernel for image classification of insects. Wu *et al.* [36] address one-shot learning in the context of path planning algorithms for robotic actuation. Koch *et al.* [37] proposed siamese neural networks for one-shot image recognition. Vinyals *et al.* [39] applied matching networks for one-shot learning by employing the
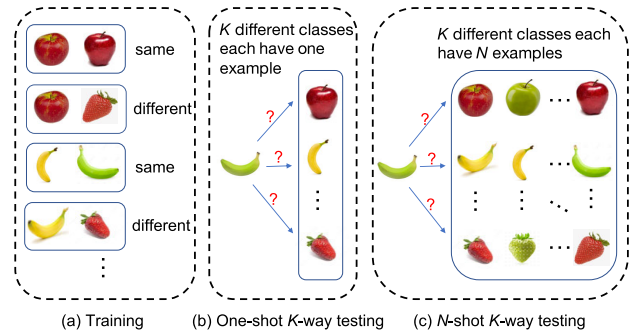

FIGURE 1. Few-shot learning general strategy.

ideas from metric learning based on deep neural features and from recent advances that augment neural networks with external memories. Altae-Tran *et al.* [40] applied one-Shot learning for limited data drug discovery. Snell *et al.* [41] proposed prototypical networks for few-shot learning by learning a metric space in which classification can be performed by computing distances to prototype representations of each class. Qiao *et al.* [42] proposed few-shot image recognition by Predicting Parameters from activations. Zhang *et al.* [43] proposed a conceptually simple and general framework called MetaGAN for few-shot learning problems. Different from semi-supervised few-shot learning, their algorithms can deal with semi-supervision at both sample-level and task-level. However, despite the success of one-shot and few-shoting learning in other applications, to our knowledge, these methods have not been applied to solving the critical sample scarcity issue in rolling bearing fault diagnosis.

In this paper, we proposed a few-shot learning neural network approach for rolling bearing fault diagnosis with limited data. Our contribution in this paper includes:

(1) We proposed a few-shot learning approach for bearing fault diagnosis with limited data, which is achieved by developing a siamese neural network model based on deep convolutional neural networks with wide first-layer kernels (WDCNN) [8].

(2) We demonstrated for the first time that few-shot learning based diagnosis models can boost the performance of fault diagnosis by making full use of the same or different class sample pairs and recognize the test sample from the classes that have only a single or few samples. For example, in small training set with 90 training samples, our method can achieve an accuracy of 92.56% compared to 80.36% without using few-shot learning strategy.

(3) As the number of training samples increases, the test performance does not monotonically increase when the test data set has a significant difference from training data set.

This rest of the paper is organized as follows: Section II describes the few-shot learning based fault diagnosis algorithm. Section III presents the experiments, results and discussion. Section IV concludes the paper.
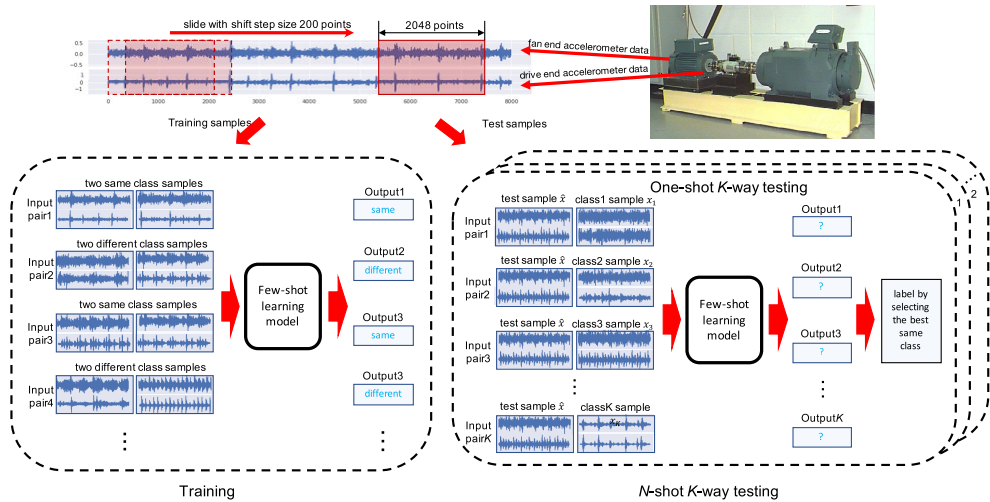
FIGURE 2. Flowchart of the few-shot learning based fault diagnosis.

II. METHODS

A. FEW-SHOT LEARNING GENERAL STRATEGY

The few-shot learning general strategy is shown in Figure 1. It is based on multiple applications of one-shot learning. First, it trains a model with a collection of sample pairs with the same or different categories. Input is a sample pair with the same or different classes (x_1^i, x_2^i). Output is the probability $p(x_1^i, x_2^i)$ that two input samples are the same. Unlike traditional classification, the performance of few-shot learning is typically measured by N-shot K-way testing as shown in Figure 1(c).

In one-shot K-way testing, it is given a test sample $\hat{x}$ to classify and a support set S as shown in Equation 1, which contains K samples and each sample has a distinct label y.

$$S = \{(x_1, y_1), \ldots, (x_K, y_K)\} \tag{1}$$

And then it classifies the test sample according to the most similar sample in the support set as shown in Equation 2.

$$C(\hat{x}, S) = \underset{c}{\mathrm{argmax}}(P(\hat{x}, x_c)), x_c \in S \tag{2}$$

In N-shot K-way testing, the model is given a support set consisting of K different classes each having N samples $(S_1, \ldots, S_N)$. And then the model has to determine which of the support set classes the test sample should belong to as shown in Equation 3.

$$C(\hat{x}, (S_1, \ldots, S_N)) = \underset{c}{\mathrm{argmax}}(\sum_{n=1}^{N} P(\hat{x}, x_{cn})), x_{cn} \in S_n \tag{3}$$

B. FEW-SHOT LEARNING FRAMEWORK FOR BEARING FAULT DIAGNOSIS

The few-shot learning for bearing fault diagnosis is based on multiple applications of one-shot learning. As shown in Figure 2, there are three steps in few-shot learning: data preparation (top), model training (left), and model testing (right). The details of data preparation will be introduced
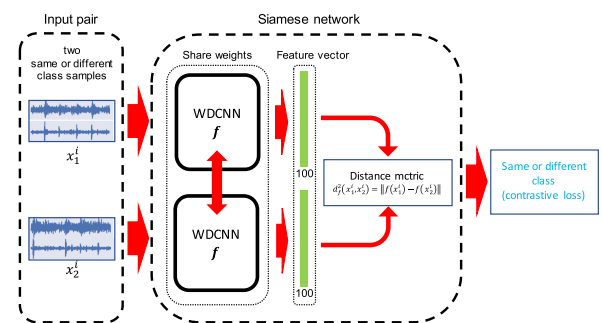

FIGURE 3. Few-shot learning model based on CNNs (WDCNN).

in Section III. The input of model training is the collection of samples pair with the same or different category labels. The output of model training is a probability distance to judge if the pair are of the same or different category. The details of model testing will be explained in below subsection.

Figure 3 shows our few-shot learning model for bearing fault diagnosis, which is a siamese neural network [37] based on deep convolutional neural networks with wide first-layer kernels (WDCNN) [8]. In this network model, two identical WDCNN networks are set up with the same network architecture and shared weights. The WDCNN network architecture as detailed in Table 1 is setup the same as [8]. It first uses wide kernels to extract features, and then uses small kernels to acquire better feature representation. This design strategy is due to the fact that designing a model with all small kernels is unrealistic and small kernels at the first layer can be easily disturbed by high frequency noise common in industrial environments.

1) INPUT

The input data is a sample pair of the same or different classes. We directly use the deep convolutional neural network (WDCNN) to extract features from the raw vibration signals. This is called end-to-end deep learning approach.

TABLE 1. Detail of few-shot learning model based on CNNs (WDCNN).

No.	Layer Type	Kernel Size/Stride	Kernel Number	Output Size (Width x Depth)	Padding
1	Convolution1	64 x 1/16 x 1	16	128 x 16	same
2	Pooling1	2 x 1/2 x 1	16	64 x 16	valid
3	Convolution2	3 x 1/1 x 1	32	64 x 32	same
4	Pooling2	2 x 1/2 x 1	32	32 x 32	valid
5	Convolution3	3 x 1/1 x 1	64	32 x 64	same
6	Pooling3	2 x 1/2 x 1	64	16 x 64	valid
7	Convolution4	3 x 1/1 x 1	64	16 x 64	same
8	Pooling4	2 x 1/2 x 1	64	16 x 64	valid
9	Convolution5	3 x 1/1 x 1	64	6 x 64	valid
10	Pooling5	2 x 1/2 x 1	64	3 x 64	valid
11	Fully-connected	100	1	100x1	

2) DISTANCE METRIC

Let M indicate the minibatch size and i, the ith minibatch. The twins sub-networks are optimised based on the distance metric between their outputs, which is calculated by Equation (4) where f is the WDCNN neural network.

$$d_f^2(x_1^i, x_2^i) = \left\| f(x_1^i) - f(x_2^i) \right\| \tag{4}$$

3) OUTPUT

The output is the "distance" of the feature vector outputs from the WDCNN twins in terms of whether their outputs are considered quite similar versus being quite dissimilar. It is obtained by Equation (5) which represents the probability that two input samples are the same, where $sigm$ is the sigmoid function and FC is a dense fully connected layer.

$$P(x_1^i, x_2^i) = sigm(FC(d_f^2(x_1^i, x_2^i))) \tag{5}$$

4) LOSS FUNCTION

Let $t_j = y(x_{1j}^i, x_{2j}^i)$ be a length-M vector which contains the labels for the minibatch. We let $t_j^i = 1$ whenever x_{1j}^i and x_{2j}^i are from the same fault class and $t_j^i = 0$ otherwise, where j is the jth sample pair from ith minibatch. The loss function is a regularized cross-entropy as:

$$L(x_1^i, x_2^i, t^i) = t^i log(P(x_1^i, x_2^i)) + (1 - t^i)log(1 - P(x_1^i, x_2^i))$$
$$+ \lambda^T |W|^2 \tag{6}$$

5) OPTIMIZATION

The network is optimized by Adam optimizer, which computes individual adaptive learning rates for each parameter. The parameters are updated by:

$$m_w^{(T+1)} = \beta_1 m_w^{(T)} + (1 - \beta_1)\nabla_w L^{(T)}$$
$$v_w^{(T+1)} = \beta_2 v_w^{(T)} + (1 - \beta_2)(\nabla_w L^{(T)})^2$$
$$\hat{m}_w = \frac{m_w^{(T+1)}}{1 - (\beta_1)^{T+1}}$$
$$\hat{v}_w = \frac{v_w^{(T+1)}}{1 - (\beta_2)^{T+1}}$$
$$w^{(T+1)} = w^{(T)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon} \tag{7}$$

where $w_{(T)}$ means the parameters at epoch T, $L_{(t)}$ is the loss function, β_1 and β_2 are the forgetting factors for the first

TABLE 2. Description of rolling bearing datasets.

Fault Location		None	Ball			Inner Race			Outer Race			Load
Fault Diameter (inch)		0	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021	
Fault Labels		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Dataset A	Train	660	660	660	660	660	660	660	660	660	660	1
	Test	25	25	25	25	25	25	25	25	25	25	
Dataset B	Train	660	660	660	660	660	660	660	660	660	660	2
	Test	25	25	25	25	25	25	25	25	25	25	
Dataset C	Train	660	660	660	660	660	660	660	660	660	660	3
	Test	25	25	25	25	25	25	25	25	25	25	
Dataset D	Train	1980	1980	1980	1980	1980	1980	1980	1980	1980	1980	1,2,3
	Test	75	75	75	75	75	75	75	75	75	75	

and second moments of gradients respectively, and m and v are moving averages.

6) TRAINING

The few-shot learning training for fault diagnosis are shown on the left in Figure 2. The input is a sample pair with the same or different clasess (x_1^i, x_2^i), where i indexes the ith minibatch. Output is the probability distance $p(x_1^i, x_2^i)$ as detailed in Equation 5, then the loss is calculated by Equation 6 and the model is optimised by Equation 7.

7) N-SHOT K-WAY TESTING

We use multiple one-shot K-way testings to simulate N-shot K-way Testing. As for five-shot N-way testing task, it is the same as N-shot K-way testing task. We repeat one-shot K-way testing five times as the five-shot data support set while each time the data support set S is randomly selected from the training data. After five times of one-shot N-way testing, we get five probability vector $(P_1, P_2, P_3, P_4, P_5)$ and calculate the maximum sum of probability of the same label by Equation 3.

III. EXPERIMENTS AND RESULTS

To verify the performance of our few-shot learning algorithm for limited data fault diagnosis, we selected 12k drive end bearing fault data in the Case Western Reserve University (CWRU) Bearing Datasets [44], [45] as the original experiments data. As shown in Table 2, there are four types of the bearing fault location: normal, ball fault, inner race fault, and outer race fault. Each fault type contains three types: 0.007 inches, 0.014 inches, and 0.021 inches respectively, so we have ten types fault labels in total. Each fault label contains three type loads of 1, 2 and 3 hp (motor speeds of 1772, 1750 and 1730 RPM respectively).

In experiments, each sample is extracted from two vibration signals as shown in Figure 2. We use half of the vibration signal to generate training samples and the rest to generate test samples. The train samples are generated by the sliding window of size 2048 points sliding with 80 points shift step. The test samples are generated by the same size sliding window sliding without overlap. Datasets A, B and C are under different work conditions with loads of 1, 2 and 3 hp respectively, and each contains 660 training samples and 25 test samples. Datasets D is under the three loads of work conditions and contains 1980 training samples and 75 testing

TABLE 3. Details of three methods in our benchmark study.

Method	Brief Introduction	Training Input	Training Output	Testing
Baseline	deep convolutional neural networks with wide first-layer kernels, detailed in Table 1	one sample	fault label	same as training
One-shot	a siamese neural network with two WDCNNs, detailed in Figure 3	two same or different sample	distance metric squashed into [0,1] (same or different)	one-shot K-way (different from taining detaied in the right of Figure 2)
Five-shot	simulating by repeating one-shot five times	same as one-shot	same as one-shot	five-shot K-way

samples in total. In the following experiments, comparing with the baseline WDCNN method, we will discuss the effect of the proposed few-shot learning method to address the limited data fault diagnosis challenge. For following experiments, the detail compares baseline WDCNN and few-shot (One-shot and Five-shot) methods are shown in Table 3. It should be noted that the A, B, C, and D datasets are used to generate the data set in different experiments.

A. EFFECT OF THE NUMBER OF TRAINING SAMPLES ON PERFORMANCE

In this experiment, we will evaluate the effect of proposed few-shot learning method to address the first two challenges in limited data fault diagnosis: 1) industry systems are not allowed to run into faulty states due to the consequences, especially for critical systems and failures; 2) most electro-mechanical failures occur slowly and follow a degradation path such that failure degradation of a system might take months or even years [7]. We conducted a series of comparison experiments by setting the test dataset in Dataset D (Table 2) as the test set and randomly selecting 60, 90, 120, 200, 300, 600, 900, 1500, 6000, 19800 samples respectively from the training samples of the whole dataset D. For every training set, the support vector machines (SVM) method uses whole training sets to fit models. We searched the proper parameters for the SVM algorithm. Other methods use 60% samples as the training set and the rest samples as the validation set. Then we evaluated the effect of the sample number on the performance of each training model. For each training set size, we repeated the sample selection process five times to generate five different training sets to deal with the bias of randomly selected small training sets. For each such random training sample set, we repeated the algorithm training and testing experiment four times to deal with the randomness of the algorithms. Together, for every series of experiments, we repeated twenty times in total.

First we compared the performances of deep learning algorithms and conventional machine learning algoithm, the SVM. As shown in Figure 4, it is clear that the accuracies of deep learning methods are much higher than those of SVM. For training sets with 60 samples, the WDCNN, One-shot, and Five-shot algorithms achieved accuracies of 73.97%, 79.33%, and 82.80% respectively, which are all sharply higher than 18.93%, the accuracy of SVM. For training sets with 90 samples, the WDCNN, One-shot, and Five-shot algorithms achieved accuracies of 77.39%, 88.41%, and 91.37% respectively, which are all sharply higher than
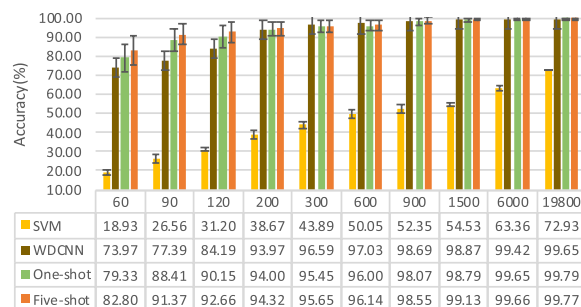

	60	90	120	200	300	600	900	1500	6000	19800
SVM	18.93	26.56	31.20	38.67	43.89	50.05	52.35	54.53	63.36	72.93
WDCNN	73.97	77.39	84.19	93.97	96.59	97.03	98.69	98.87	99.42	99.65
One-shot	79.33	88.41	90.15	94.00	95.45	96.00	98.07	98.79	99.65	99.79
Five-shot	82.80	91.37	92.66	94.32	95.65	96.14	98.55	99.13	99.66	99.77

FIGURE 4. Diagnosis results of the proposed few-shot (one-shot and five-shot) learning using different numbers of training samples, compared with WDCNN.

26.56%, the accuracy of SVM. We also found that the accuracies of SVM, WDCNN and few-shot learning all increase while their standard deviations decline with increasing number of training samples. Next, we check whether few-shot learning algorithm performs better than standard WDCNN in experiments with small number of training samples (e.g. with 60, 90 and 120 samples). For all three training sample sizes, our few-shot model performs better with an average of 9% higher in accuracy than the WDCNN model. Especially when the training set size is set to 90, the few-shot model gets 13% higher in accuracy than the WDCNN model. When the number of training samples is further increased to 200, 300, and 600, results in Figure 4 shows that the performance of few-shot is slightly worse than that of WDCNN. The accuracies of both algorithms are actually very close and are both higher than 94%. When the number of training samples is even increased to 900 or more, the performances of these two algorithms become almost the same and the accuracies of both algorithms are higher than 98%. These performance comparisons show that our proposed few-shot learning algorithm enjoys the much better performance when trained with limited datasets without losing too much when there are aboundant training samples. Furthermore, it is also observed that the accuracies of five-shot learning are consistently better than the accuracies of one-shot learning, as shown in the third and fourth row in Figure 4 table.

To better understand the effect of few-shot learning in limited data diagnosis, Figure 5(a),(b) show feature visualization via t-SNE: the last hidden fully-connected layer visualization which trained by the limited data with 90 training samples. Figure 5(c),(d) show confusion matrix results for 90 training samples. From Figure 5(a),(b), it is clear that the features in the one-shot model are much more divisible than in WDCNN
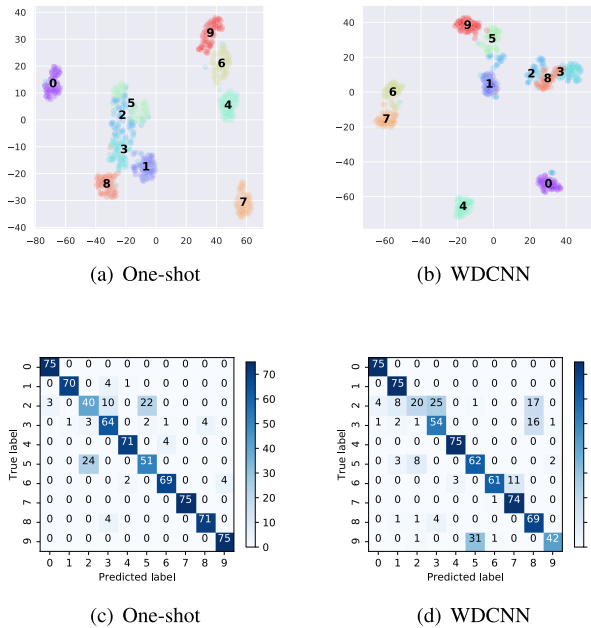
(a) One-shot

(b) WDCNN

(c) One-shot

(d) WDCNN

FIGURE 5. (a),(b) are feature visualization via t-SNE: the last hidden fully-connected layer trained by 90 training samples. (c),(d) are confusion matrix for 90 training samples results.

TABLE 4. Performance under noise environment.

model	exp	SNR (dB)								
		-4	-2	0	2	4	6	8	10	None
Five-shot	60	31.83	47.64	63.39	73.91	78.39	80.69	81.74	82.35	82.80
	90	34.70	50.08	66.53	79.93	86.44	89.24	90.26	90.95	91.37
	120	31.87	49.04	67.08	80.28	88.12	91.05	91.69	92.29	92.66
	200	29.91	45.52	61.32	76.55	85.69	90.79	92.70	93.72	94.32
	300	29.30	44.25	62.13	77.88	86.63	91.33	93.63	94.55	95.65
	600	28.75	45.77	65.63	80.39	88.63	92.49	94.34	95.22	96.14
	900	38.95	54.09	72.59	87.59	94.20	96.92	97.87	98.21	98.55
	1500	35.05	49.61	66.33	82.17	92.65	96.48	98.10	98.81	99.13
	6000	37.87	55.23	69.82	83.69	94.57	98.46	99.49	99.62	99.66
	19800	41.25	57.54	71.38	85.95	95.21	98.93	99.57	99.67	99.77
One-shot	60	31.36	46.61	60.70	70.52	74.24	77.33	78.02	78.29	79.33
	90	33.65	49.35	65.26	77.63	83.71	86.18	87.85	88.29	88.41
	120	30.82	47.80	65.65	79.51	85.63	88.52	89.08	89.42	90.15
	200	29.83	45.01	60.84	76.11	85.22	90.21	92.51	93.47	94.00
	300	29.23	44.14	61.84	77.61	86.48	91.05	93.41	94.25	95.45
	600	28.72	45.45	65.30	80.10	88.38	92.40	94.30	95.04	96.00
	900	38.45	53.69	72.48	87.39	93.87	96.22	97.48	97.77	98.07
	1500	34.73	48.85	66.16	81.80	92.03	96.03	97.80	98.31	98.79
	6000	37.77	54.99	69.97	83.50	94.47	98.41	99.41	99.62	99.65
	19800	41.23	57.48	71.34	85.75	95.16	98.88	99.57	99.67	99.79
WDCNN	60	31.43	45.13	58.11	66.07	70.12	72.56	72.82	73.26	73.97
	90	39.78	56.64	67.12	71.86	74.74	75.84	76.75	77.09	77.39
	120	37.27	56.02	70.84	77.21	80.06	81.65	82.81	83.70	84.19
	200	42.42	64.59	82.87	89.13	92.02	93.20	93.84	94.10	93.97
	300	44.57	71.68	87.39	93.31	94.98	96.06	96.41	96.61	96.59
	600	35.17	59.19	78.21	89.62	93.95	95.67	96.61	96.89	97.03
	900	47.77	68.85	87.95	95.09	97.25	98.00	98.29	98.50	98.69
	1500	37.81	63.99	90.49	97.35	98.63	98.79	98.78	98.92	98.87
	6000	34.51	64.12	91.43	97.87	99.00	99.31	99.30	99.40	99.42
	19800	45.06	63.63	83.30	94.97	98.91	99.55	99.61	99.65	99.65

(a) 90 training samples

(b) 300 training samples

(c) 6000 training samples

(d) 19800 training samples
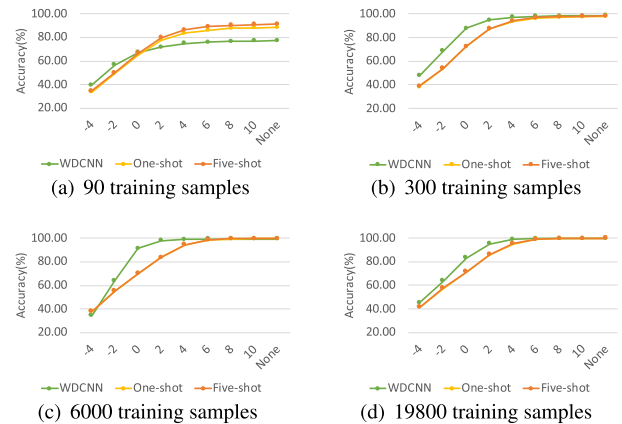
FIGURE 6. Comparing results of the different number of training data under the noise environment.

model, especially the 2,3,8 categories features. So in Figure 5(c),(d) we can find that the 2,3,8 categories in the one-shot model are easier to diagnose than in WDCNN model.

B. PERFORMANCE UNDER NOISE ENVIRONMENT

In this experiment, we will evaluate effect of the proposed few-shot learning method to address the third challenge in limited data fault diagnosis: working conditions of mechanical systems are very complicated and change many times from time to time according to production requirement. It is unrealistic to collect and label enough training samples [8]. We discuss the performance under noise environment to simulate the change of working conditions in datasets D. Signal-to-noise ratio (SNR) is defined as the ratio of signal power to the noise power, often expressed in decibels detailed as follows:

$$\text{SNR}_{\text{dB}} = 10 \log_{10}(P_{signal}/P_{noise}) \qquad (8)$$

where P_{signal} and P_{noise} are the power of the signal and the noise, respectively. In this case, the same as [8], the models are trained with the original data provided by CWRU, then tested with added different SNR white Gaussian test samples. The different SNR ranges from −4 dB to 10 dB. The smaller SNR value is, the stronger power of noise is.

Table 4 shows the results of diagnosing noise signal by the proposed few-shot (one-shot and five-shot) learning models and the compared WDCNN model. The greener the background color, the better the result is. It is clear that the accuracy increases and is greener as the noise gets weak, e.g., the average accuracy is only near 40% when the SNR is −4 dB in 19800 training data samples, while the accuracy

surges to above 99% when SNR is 6 dB. And the accuracy increase as the data number for training gets more significant when the noise is not very strong.

As shown in Figure 6, to check the effect of the data number for training under the noise environment, we compare results with different training data number. From results, we easily found that five-shot is better than one-shot. For a small number of training data, when the features between training and test set are similar, the few-shot performance is better than the WDCNN. e.g., as shown in Figure 6(a), when the SNR equals to −4 dB as the features between training and test set have a big difference, the few-shot is inferior to the WDCNN. And when the SNR equals to 0 dB as the features between training and test set have some difference, few-shot and WDCNN performance are same. However, when SNR equals to 10 dB as the features between training and test set are similar, the few-shot is much better than the WDCNN, higher over 10%. Besides, when there has sufficient training data, the few-shot is better than the WDCNN. e.g., as shown in
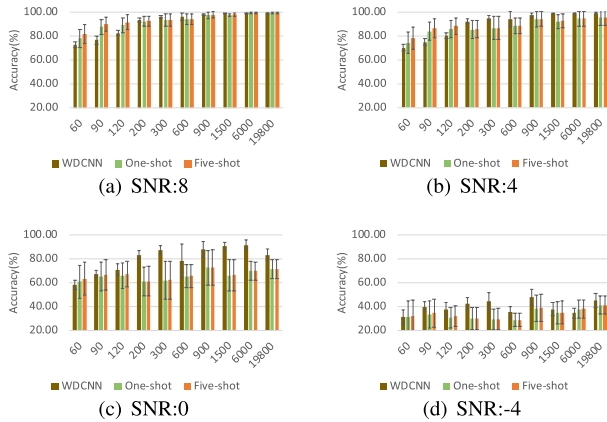
FIGURE 7. Comparing results of different noise environments under a series of number of training samples.

FIGURE 8. Comparing performance under unseen categories.

TABLE 5. Scenario setting for domain adaptation.

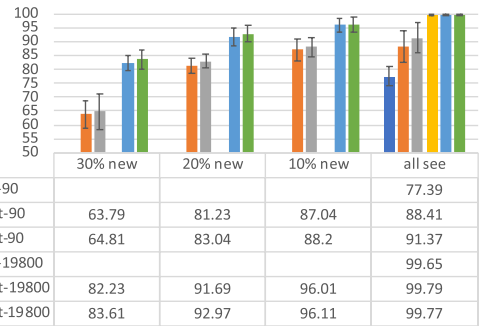| Domain types	Source domain	Target domain	
Description	labeled signals under one single load	unlabeled signals under new load	
Domain details	Training set A	Test set B	Test set C
	Training set B	Test set A	Test set C
	Training set C	Test set A	Test set B
Target	Diagnose unlabeled vibration signals in new target domain		

Figure 6(b), the WDCNN is better than few-shot as the training data size increases. However, as shown in Figure 6(c), when the training data size increases to a certain amount, the gap between few-shot and WDCNN is decreasing. And as shown in Figure 6(d), when there is sufficient training data, the few-shot is better than the WDCNN.

In Figure 7, we compare the test results of three models under different noise environments. In Figure 7(a),(b) with 8 dB and 4 dB SNR respectively, there is less difference between the training set and the test set. The test results become better as the number of training samples increases. However, in Figure 7(c),(d) with 0 dB and −4 dB SNR respectively, there is large difference between the training set and the test set. The test performance does not monotonically increase as the number of training samples increases. Because the model trained by a small number of training samples is mainly considered to be under-fitting and the test accuracy increases as the number of training samples increases. Then when the number of training samples increases with the larger difference between training and test sets, the trained model performs well in the training set, but it may easily show overfitting on significantly different test sets which will cause test accuracy decrease.

As mentioned above, as the number of training samples increases, the test performance does not monotonically increase when the test data set has a significant difference from the training data set. Therefore, the appropriate number of training samples could get the best results on significant different test sets.

C. PERFORMANCE UNDER NEW CATEGORIES
In this experiment, we will evaluate effect of the proposed few-shot learning method to address the fourth challenge in limited data fault diagnosis: especially in real-world applications, fault categories and working conditions are usually unbalanced. It is thus difficult to collect enough samples for every fault type under different working conditions. We mainly pay attention to unbalanced fault categories. When new categories appear, traditional deep learning methods

need to retrain to deal with the new categories diagnosis. However, different from the traditional deep learning methods, the few-shot learning model can be directly used in the new category diagnosis by just giving a few new category samples. We train the WDCNN model from all categories and the few-shot model from 30% randomly new categories to all see categories in datasets D. We repeated each such experiment ten times to deal with the randomness of the algorithms.

The results are shown in Figure 8. The accuracy of the few-shot raises with the decrease of number of new categories. The accuracy of five-shot is better than the accuracy of one-shot. For the limited data with 90 training samples, the accuracy of few-shot is higher than the WDCNN model when number of new categories are equal or under 20%. Besides, for enough training samples, the accuracy of few-shot can get over 90% performance when number of new categories are equal or under 20%. Therefore, few-shot learning could enhance performance well under a few new categories. Thus, when new categories appear, the few-shot learning model could be directly used in the new category diagnosis and get better performance.

D. PERFORMANCE UNDER NEW WORK CONDITIONS
In this experiment, we will evaluate effect of the proposed few-shot learning method to address the fourth challenge in limited data fault diagnosis. We mainly pay attention in unbalanced working condition. When new categories appear, we would like to evaluate how well the few-shot performance. We use domain adaptation to simulate new work conditions. The description of scenario setting for domain adaptation is illustrated in Table 5, which the training or test set A, B, and C same as Dataset A, B, and C in Table 2.

As shown in Figure 9, for the limited data with 90 training samples, the average accuracy of few-shot learning method in the six scenarios performs better than the WDCNN method.
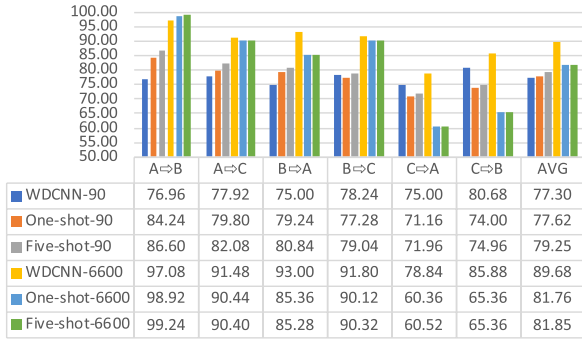
FIGURE 9. The six new working condition domain shifts results on the Datasets A, B and C.

	A⇒B	A⇒C	B⇒A	B⇒C	C⇒A	C⇒B	AVG
WDCNN-90	76.96	77.92	75.00	78.24	75.00	80.68	77.30
One-shot-90	84.24	79.80	79.24	77.28	71.16	74.00	77.62
Five-shot-90	86.60	82.08	80.84	79.04	71.96	74.96	79.25
WDCNN-6600	97.08	91.48	93.00	91.80	78.84	85.88	89.68
One-shot-6600	98.92	90.44	85.36	90.12	60.36	65.36	81.76
Five-shot-6600	99.24	90.40	85.28	90.32	60.52	65.36	81.85

FIGURE 9. The six new working condition domain shifts results on the Datasets A, B and C.

TABLE 6. Complexity of tree datasets.

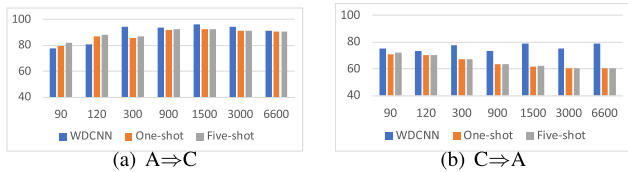| Dataset	A	B	C
Speed	1772 rpm	1750 rpm	1730 rpm
Complexity	high	normal	low

(a) A⇒C

(b) C⇒A

FIGURE 10. Comparing results of the different noise environment under a series of training data number.

However, with enough training samples, the average accuracy of the WDCNN method in the six scenarios is more better than the few-shot learning method. Besides, both these methods perform poorly when training with set C and testing in set A. As shown in Table 6, as set A's speed is higher than set B, we think that set A is more complicated than set B, same as set B and set C. Therefore, when training with lower speed set C and testing in higher speed set A, there has lots features change. Thus it causes the poor performance. It can also be confirmed in Figure 10 results. As shown in Figure 10(a), when training with the complicated set A and testing in set C, the features change a little. Thus the test performance increases as the number of training samples increase. However, as shown in Figure 10(b), when training with the set C and test in the complicated set A, the features change a lot. Thus the test performance does not monotonically increase as the number of training samples increase.

The few-shot learning method could not be better when training with limited data to enough data and the WDCNN method could. However, the average accuracy of WDCNN method not better than the WDCNN trained with 200 labeled data (Figure 4). We think that it is unreasonable to use the model trained only from one load condition and directly used in a new load, and the effect is not satisfactory. Because the model is trained only for one load situation, it is difficult for the model to learn the knowledge of the changes in the category features caused by load changes. Thus that may cause poor performance under new loads. So we designed the new scenario setting for domain adaptation illustrated

TABLE 7. New scenario setting for domain adaptation.

Domain types	Source domain	Target domain
Description	labeled signals under two load	unlabeled signals under new load
Domain details	Training set AB Training set AC Training set BC	Test set C Test set B Test set A
Target	Diagnose unlabeled vibration signals in new target domain	

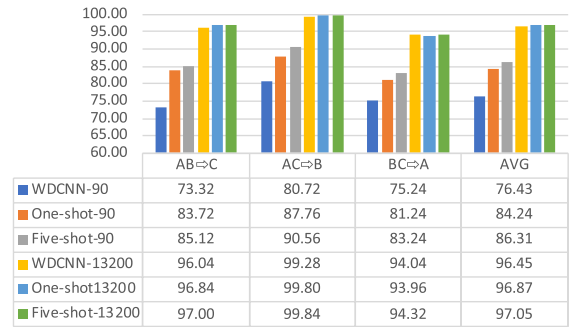|	AB⇒C	AC⇒B	BC⇒A	AVG
WDCNN-90	73.32	80.72	75.24	76.43
One-shot-90	83.72	87.76	81.24	84.24
Five-shot-90	85.12	90.56	83.24	86.31
WDCNN-13200	96.04	99.28	94.04	96.45
One-shot13200	96.84	99.80	93.96	96.87
Five-shot-13200	97.00	99.84	94.32	97.05

FIGURE 11. The new three new working condition domain shifts results on the Datasets A, B and C.

in Table 7, by training with the labeled signals under two load and testing in unlabeled signals under new load.

As shown in Figure 11, for the limited data with 90 training samples, the average accuracy of these methods in the new scenario setting is better but not significate than in the old scenario setting. But when there are enough samples which allow to effectively learn the knowledge of the changes in the category features caused by load changes, the average accuracy of both methods perform much better than in the old scenario setting. For the few-shot method, it is higher over than 15%. For WDCNN method, it is higher over than 7%. Besides, in the new scenario setting, we find the few-shot performs better than the WDCNN.

IV. CONCLUSION

This paper presents a few-shot learning approach for rolling bearing fault diagnosis with limited data. Our algorithm addresses one of the major challenges in limited data fault diagnosis: the difficulty of obtaining sufficient numbers of samples in data-driven fault diagnosis. Our few-shot fault diagnosis model is based on the siamese neural network for one-shot learning. It works by exploiting sample pair of the same or different categories, measuring the ''distance'' of two WDCNN twins feature vector outputs in terms of whether their outputs are considered quite similar versus dissimilar.

Our method was validated on the CWRU Datasets with extensive experiments by comparing its performance with baseline (the popular WDCNN fault diagnosis model). The experimental results showed that few-shot learning is effective for fault diagnosis with both limited data or sufficient data. By comparing testing results under different noise environments, we found that large difference between the training and test sets may cause the test performance to not monotonically increase as the number of training samples increases. When evaluated over test sets

with new fault types or new working conditions, few-shot models work better than the baseline trained with all fault types. Furthermore, to facilitate open research in the field of bearing fault diagnosis, all our models in this work and datasets are open sourced and can be downloaded at the URL (https://mekhub.cn/as/fault_diagnosis_with_few-shot_learning/).

REFERENCES

[1] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques-Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.

[2] S. Nandi, H. A. Toliyat, and X. Li, "Condition monitoring and fault diagnosis of electrical motors—A review," *IEEE Trans. Energy Convers.*, vol. 20, no. 4, pp. 719–729, Dec. 2005.

[3] J. Chen and R. J. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*, vol. 3. Boston, MA, USA: Springer, 2012.

[4] Y. Lei, J. Lin, M. J. Zuo, and Z. He, "Condition monitoring and fault diagnosis of planetary gearboxes: A review," *Measurement*, vol. 48, pp. 292–305, Feb. 2014.

[5] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.

[6] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[7] N. Gebraeel, A. Elwany, and J. Pan, "Residual life predictions in the absence of prior degradation knowledge," *IEEE Trans. Rel.*, vol. 58, no. 1, pp. 106–117, Mar. 2009.

[8] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017.

[9] Q. Hang, J. Yang, and L. Xing, "Diagnosis of rolling bearing based on classification for high dimensional unbalanced data," *IEEE Access*, vol. 7, pp. 79159–79172, 2019.

[10] L. Duan, M. Xie, T. Bai, and J. Wang, "A new support vector data description method for machinery fault diagnosis with unbalanced datasets," *Expert Syst. Appl.*, vol. 64, pp. 239–246, Dec. 2016.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[12] W. Sun, S. Shaoa, R. Zhaob, R. Yana, X. Zhangc, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, Jul. 2016.

[13] Z. Chen and W. Li, "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1693–1702, Jul. 2017.

[14] H. Shao, H. Jiang, Y. Lin, and X. Li, "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders," *Mech. Syst. Signal Process.*, vol. 102, pp. 278–297, Mar. 2018.

[15] T. Junbo, L. Weining, A. Juneng, and W. Xueqian, "Fault diagnosis method study in roller bearing based on wavelet transform and stacked auto-encoder," in *Proc. 27th Chin. Control Decis. Conf. (CCDC)*, May 2015, pp. 4608–4613.

[16] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mech. Syst. Signal Process.*, vols. 76–77, pp. 283–293, Aug. 2016.

[17] X. Wang, J. Huang, G. Ren, and D. Wang, "A hydraulic fault diagnosis method based on sliding-window spectrum feature and deep belief network," *J. Vibroeng.*, vol. 19, no. 6, pp. 4272–4284, Sep. 2017.

[18] Z. Gao, C. Ma, D. Song, and Y. Liu, "Deep quantum inspired neural network with application to aircraft fuel system fault diagnosis," *Neurocomputing*, vol. 238, pp. 13–23, May 2017.

[19] S. Shao, W. Sun, P. Wang, R. X. Gao, and R. Yan, "Learning features from vibration signals for induction motor fault diagnosis," in *Proc. Int. Symp. Flexible Automat. (ISFA)*, Aug. 2016, pp. 71–76.

[20] R. Liu, G. Meng, B. Yang, C. Sun, and X. Chen, "Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1310–1320, Jun. 2017.

[21] S. Li, G. Liu, X. Tang, J. Lu, and J. Hu, "An ensemble deep convolutional neural network model with improved D-S evidence fusion for bearing fault diagnosis," *Sensors*, vol. 17, no. 8, p. 1729, 2017.

[22] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.

[23] C. Yuanhang, P. Gaoliang, X. Chaohao, Z. Wei, L. Chuanhao, and L. Shaohui, "ACDIN: Bridging the gap between artificial and real bearing damages for bearing fault diagnosis," *Neurocomputing*, vol. 294, pp. 61–71, Jun. 2018.

[24] Y. Yao, H. Wang, S. Li, Z. Liu, G. Gui, Y. Dan, and J. Hu, "End-to-end convolutional neural network model for gear fault diagnosis based on sound signals," *Appl. Sci.*, vol. 8, no. 9, p. 1584, 2018.

[25] M. Yuan, Y. Wu, and L. Lin, "Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network," in *Proc. IEEE Int. Conf. Aircr. Utility Syst. (AUS)*, Oct. 2016, pp. 135–140.

[26] R. Zhao, R. Yan, J. Wang, and K. Mao, "Learning to monitor machine health with convolutional Bi-directional LSTM networks," *Sensors*, vol. 17, no. 2, p. 273, 2017.

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

[28] R. Zhang, H. Tao, L. Wu, and Y. Guan, "Transfer learning with neural networks for bearing fault diagnosis in changing working conditions," *IEEE Access*, vol. 5, pp. 14347–14357, 2017.

[29] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[30] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.

[31] A. Zhang, H. Wang, S. Li, Y. Cui, Z. Liu, G. Yang, and J. Hu, "Transfer learning with deep recurrent neural networks for remaining useful life estimation," *Appl. Sci.*, vol. 8, no. 12, p. 2416, Nov. 2018.

[32] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, May 2019.

[33] D. Cabrera, F. Sancho, J. Long, R.-V. Sánchez, S. Zhang, M. Cerrada, and C. Li, "Generative adversarial networks selection approach for extremely imbalanced fault diagnosis of reciprocating machinery," *IEEE Access*, vol. 7, pp. 70643–70653, 2019.

[34] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[35] L. Wolf, T. Hassner, and Y. Taigman, "The one-shot similarity kernel," in *Proc. 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 897–902.

[36] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from RGBD images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 7–12.

[37] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015, pp. 1–30.

[38] G. Sussman and K. Yip, "Sparse representations for fast, one-shot learning," in *Proc. Nat. Conf. Artif. Intell.*, Orlando, FL, USA, 1997, pp. 1–29.

[39] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3630–3638.

[40] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS Central Sci.*, vol. 3, no. 4, pp. 283–293, 2017.

[41] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. NIPS*, 2017, pp. 4077–4087.

[42] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7229–7238.

[43] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. New York, NY, USA: Curran Associates, 2018, pp. 2365–2374.

[44] X. Lou and K. A. Loparo, "Bearing fault diagnosis based on wavelet transform and fuzzy inference," *Mech. Syst. Signal Process.*, vol. 18, no. 5, pp. 1077–1095, Sep. 2004.

[45] 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.*, vol. 64, pp. 100–131, Dec. 2015.
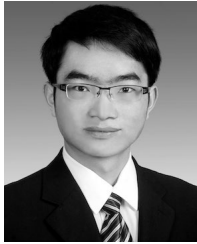
ANSI ZHANG received the B.S. degree in mechanical engineering from Shanghai University, Shanghai, China, in 2014. From 2017 to 2018, he was a visiting Ph.D. Student with the University of South Carolina, Columbia, USA. He is currently pursuing the Ph.D. degree with the Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, China.

SHAOBO LI received the Ph.D. degree in computer software and theory from the Chinese Academy of Sciences, China, in 2003.

From 2007 to 2015, he was the Vice Director of the Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University (GZU), China. He is currently a Professor with the School of Mechanical Engineering, GZU. He is also a part-time Doctoral Tutor with the Chinese Academy of Sciences. He has been the Dean of the School of Mechanical Engineering, GZU, since 2015. He has published more than 200 papers in major journals and international conferences. His current research interests include big data of manufacturing and intelligent manufacturing. His research has been supported by the National Science Foundation of China and the National High-Tech R&D Program (863 Program). He received honors and awards from New Century Excellent Talents in the University of Ministry of Education of China, Excellent Expert and Innovative Talent of Guizhou Province, the Group Leader of Manufacturing Informatization, and the Alliance Vice Chairman of intelligent manufacturing industry in Guizhou Province.

YUXIN CUI received the B.E. degree in bioinformatics and the M.E. degree in computer science and technology from the Harbin Institute of Technology, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree in computer science with the University of South Carolina, Columbia, USA.

His research interests include machine learning and computer vision.

WANLI YANG received the B.S. degree in mechanical engineering from Guizhou University, Guiyang, China, in 2016, where he is currently pursuing the M.S. degree with the School of Mechanical Engineering.

His current research interests include machinery condition monitoring, intelligent fault diagnostics, and prognostics of rotating machinery.

RONGZHI DONG received the B.S. degree in engineering mechanics from China Agricultural University, in 2017. She is currently pursuing the M.S. degree with the School of Mechanical Engineering, Guizhou University, China.

Her research interests include machine learning, deep learning, data mining, and material informatics.

JIANJUN HU received the B.S. and M.S. degrees in mechanical engineering from the Wuhan University of Technology, China, in 1995 and 1998, respectively, and the Ph.D. degree in computer science from Michigan State University, in 2004, with a focus on machine learning and evolutionary computation.

He was a Postdoctoral Fellow in bioinformatics with Purdue University and the University of Southern California, from 2004 to 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA. He is also a Visiting Professor with the School of Mechanical Engineering, Guizhou University, Guiyang, China. His research interests include machine learning, deep learning, data mining, evolutionary computation, fault diagnosis, bioinformatics, and material informatics.

- - -