IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2019, accepted August 13, 2019, date of publication August 21, 2019, date of current version August 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936243

Improved Deep Transfer Auto-Encoder for Fault Diagnosis of Gearbox Under Variable Working Conditions With Small Training Samples

ZHIYI HE[1,2], HAIDONG SHAO[1,2], XIAOYANG ZHANG[3], JUNSHENG CHENG[1,2], AND YU YANG[1,2]

[1]State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha 410082, China
[2]College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China
[3]Xi'an Aeronautics Computing Technique Research Institute, Xi'an 710065, China

Corresponding author: Haidong Shao (hdshao@hnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51575168 and Grant 51875183, and in part by the Fundamental Research Funds for the Central Universities under Grant 531118010335.

ABSTRACT It is considerable to solve practical fault diagnosis task of gearbox under variable working conditions by introducing sufficient auxiliary data. For this purpose, a new approach called improved deep transfer auto-encoder is proposed for intelligent diagnosis of gearbox faults under variable working conditions with small training samples. First, multi-wavelet is employed as activation function for effectively learning useful features hidden in the non-stationary vibration data. Second, correntropy is used to modify the cost function to enhance the reconstruction quality. Third, pre-train an improved deep auto-encoder using sufficient auxiliary data in the source domain, and transfer its parameters to the target model. Finally, the improved deep transfer should be fine-tuned by small training samples in the target domain to adapt to the characteristics of the rest testing data. The proposed approach is used to analyze two sets of experimental vibration data collected from gearbox under variable working conditions. The results show that the proposed approach can accurately diagnose different faults of gearbox even the working conditions have significant changes, which is superior to the existing methods.

INDEX TERMS Improved deep transfer auto-encoder, gearbox fault diagnosis, variable working conditions, multi-wavelet activation function, modified cost function.

I. INTRODUCTION

Due to great loading capacity, large reduction ratio, high transmission efficiency and other prominent advantages, gearbox has a very wide application in aircraft engine, wind turbine and high speed railway. Different types of fault will occur in gearbox after long-term working under the conditions with high temperature, high speed, heavy loading and strong impact, which may lead to safety accidents [1]–[3]. Therefore, gearbox fault diagnosis has become an important part in the field of intelligent maintenance and health management.

Artificial intelligence has attracted increasingly attention in recent years for enhancing automation monitoring and inference capabilities of industrial equipment [4]. For gearbox health monitoring, despite intelligent diagnosis research has made gratifying progress [5]–[11], still the following problems have not been well solved. (1) The raw vibration signals collected from gearbox are always nonlinear and non-stationary with a lot of background noise. In addition, different fault locations and fault severities lead to the diversity of fault types, which have put forward high requirements for signal pre-processing and feature extraction [2]. (2) In consideration of economic cost and human labor, it is hard and unrealistic to obtain enough fault data in engineering practice, which will result in terrible lack of training samples for intelligent diagnosis model [12]. (3) The complexity of working conditions (variable speeds and variable loadings) may lead to significant distribution differences between the training and testing samples, meaning that the intelligent diagnosis model trained by the vibration data collected under a certain working condition is usually not suitable for other cases [13]. Thus, to automatically learn the characteristic information hidden in the raw data and realize accurate fault

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

identification under different working conditions, new skills are urgently needed to improve the existing intelligent diagnosis methods.

Due the powerful and automatic feature learning ability, deep learning has become a highly concerned intelligent method for machinery fault diagnosis in the past few years [14]–[24]. However, the successful construction of intelligent diagnosis models designed with deep structures is still inseparable from sufficient training data [13]. Moreover, the training data and testing data should meet the demand of the same distribution. Transfer learning is another great breakthrough in artificial intelligence area, which aims to solve the tasks between different but related domains based on the existed knowledge [25]. By means of transfer learning and deep learning, distribution differences between the training data (Source domain) and test data (Target domain) can be allowed to some degree. To date, transfer learning has made several academic achievements in the conventional pattern recognition fields [26]–[28]. For intelligent fault diagnosis of rotating machinery, some researches have begun to explore the application of transfer learning in the last three years. Zhang *et al.* [29] combined transfer learning and neural network for bearing fault identification under changeable working conditions. Wen *et al.* [30] proposed transfer diagnosis approach using sparse auto-encoder for classifying different fault types of bearing under variable working conditions. Qian *et al.* [31] constructed transfer learning network based on high-order Kullback-Leibler divergence to achieved intelligent fault diagnosis of gearbox and bearing under variant working conditions.

Through literature review, it can be seen that transfer learning has shown some potential to overcome distribution difference problem of gearbox fault data collected from different operating conditions. However, in the transfer diagnosis cases mentioned above, the change range of rotating speed is very small, thereby making the distribution differences not serious. However, the rotating speed and working load of gearbox usually change greatly in practical engineering [32], leading to significant differences of data samples. Thus, it is of practical importance to build better deep transfer models to achieve gearbox fault diagnosis under obvious changes in working conditions.

In this paper, a new approach based on improved deep transfer auto-encoder is proposed to diagnose different gearbox faults under variable working conditions with small training samples. First, multi-wavelet is employed as activation function for effectively learning useful features hidden in the non-stationary vibration data. Second, correntropy is used to modify the cost function to enhance the reconstruction quality. Then, pre-train an improved deep auto-encoder using sufficient auxiliary data in the source domain, and transfer its parameters to the target model. Finally, the improved deep transfer should be fine-tuned by small training samples in the target domain to adapt to the characteristics of the rest testing samples. The proposed approach is used to analyze two sets of experimental vibration data collected from gearbox under variable working conditions. The results show that the proposed approach can accurately diagnose different health conditions of gearbox even the working conditions have significant changes, which is better than the existing methods.

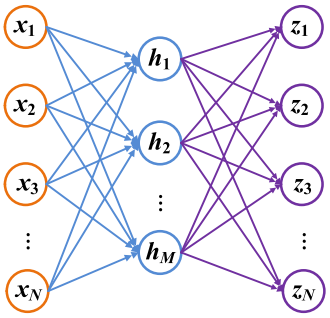

FIGURE 1. The structure of a basic auto-encoder (AE) model.

The rest of this paper is arranged as follows. Section II shortly introduces basic auto-encoder theory. The proposed approach is described in Section III. Transfer fault diagnosis cases under variable working conditions are designed to verify the superiority of the proposed approach in Section IV. Section V gives the final conclusions and future work.

II. BRIEF INTRODUCTION OF BASIC AUTO-ENCODER

As shown in Figure 1, auto-encoder (AE) is composed of an encoder and a decoder, which has become a popular base model for constructing various deep architectures due to its good capability for unsupervised feature learning. The encoder tries to obtain the representative feature representation of input data, and the decoder aims to recover input from the representation [33]. Some important formulas of the basic AE model are presented as following:

$$\mathbf{h} = s_g\left(\mathbf{w}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right) \tag{1}$$

$$\mathbf{z} = s_f\left(\mathbf{w}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}\right) \tag{2}$$

in which $\mathbf{x} \in \Re^N$ denotes an input data sample, $\mathbf{h} \in \Re^M$ denotes the feature representation, $\mathbf{z} \in \Re^N$ denotes the output, $\left\{\mathbf{w}^{(1)}, \mathbf{b}^{(1)}, \mathbf{w}^{(2)}, \mathbf{b}^{(2)}\right\}$ represents the parameter set, including the weights $\mathbf{w}^{(1)}$, $\mathbf{w}^{(2)}$ and biases $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$ in different layers, s_g denotes the activation function of encoder, usually selected as Sigmoid, s_f denotes the activation function of decoder, which is selected according to the specific normalized range of the input data.

The training purpose of AE model is to adjust the parameters to keep the output as close as possible to the input. The most widely used cost functions is expressed as [3]

$$C^{\mathrm{Tra}} = \frac{1}{2}\sum_{i=1}^{N}(z_i - x_i)^2 + \frac{\lambda}{2}\sum_{L=1}^{2}\sum_{i=1}^{s_L}\sum_{j=1}^{s_{L+1}}\left(W_{ij}^{(L)}\right)^2 + r\left(\sum_{j=1}^{M}\mu\log\frac{\mu}{\hat{\mu}_j} + (1-\mu)\log\frac{1-\mu}{1-\hat{\mu}_j}\right) \tag{3}$$

where x_i and z_i are the ith dimension elements of $\mathbf{x}$ and $\mathbf{z}$, respectively, $r$denotes sparsity penalty coefficient, μ denotes sparsity coefficient, $\hat{\mu}_j$ denotes the average activation value for the jth hidden node, and λ denotes weight decay coefficient.

III. THE PROPOSED METHOD

A. MULTI-WAVELET ACTIVATION FUNCTION

Generally, the activation function employed in hidden layer of basic AE is Sigmoid or Tanh, their main problems are computational complexity and gradient vanishing, which will result in low-efficiency weight updating. Rectified linear unit (ReLU) is fast and can avoid gradient vanishing [13]. However, the non-zero centered output and neuron dying problem will degrade the training performance [34]. What is more important, the raw vibration signals collected from gearbox are always non-stationary with complex noise, researches have investigated that neural networks designed with conventional activation functions usually fail to achieve exact mapping between multiple output patterns and non-stationary input data [35].

Wavelet neural network (WNN) has good time-frequency localization property and zoom characteristic. Compared with conventional neural networks, the superiority of WNN for analyzing non-stationary signals has been verified in lots of classification and regression cases. Multi-wavelet neural network (MWNN) is the extension of WNN, which has faster convergence and better characteristics in the approximation of non-stationary signals [36]–[38]. To date, few researches have reported about multi-wavelet activation functions applied in deep learning field, therefore, it is worth trying to design novel deep learning models using multi-wavelet to solve task.

Currently, there have been developed some multi-wavelets with excellent properties, such as GHM multi-wavelet, CL multi-wavelet and SA4 multi-wavelet. However, the scaling functions of these multi-wavelets have no explicit expressions, which will greatly increase the difficulty of calculating and updating the parameters of auto-encoder model. The scaling functions of the multi-wavelet developed by Plonka and Strela not only have very approximate expressions, but also hold some good properties such as orthogonality, regularity, symmetry and compact support [38], which are good choices to be used as activation functions of auto-encoder to improve the analysis performance for non-stationary vibration signal collected from gearbox. In this paper, the two scaling functions of the multi-wavelet are given in Figure 2, and defined as follows

$$\varphi_1(t) = \begin{cases} -2t^3 + 3t^2 & t \in [0, 1) \\ (2-t)^2(2t-1) & t \in [1, 2] \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\varphi_2(t) = \begin{cases} -t^2(3t-3) & t \in [0, 1) \\ -(2-t)^2(3t-3) & t \in [1, 2] \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

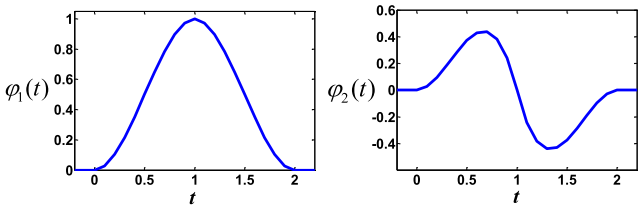

FIGURE 2. The waveforms of two multi-wavelet scaling functions.

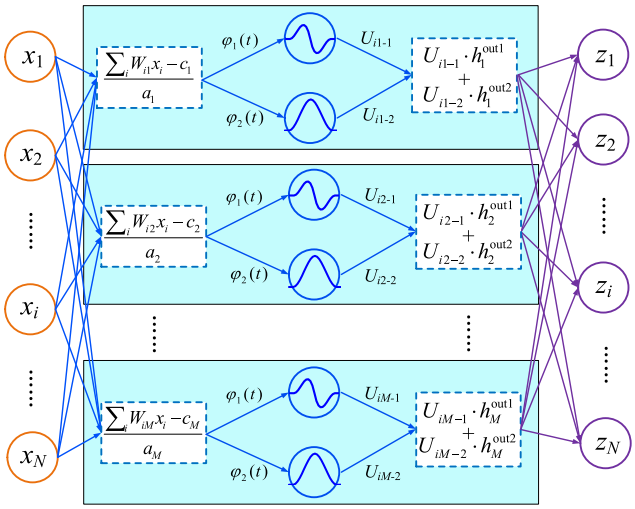

FIGURE 3. The structure of improved AE designed with multi-wavelet activation function.

The structure of improved AE model designed with multi-wavelet activation function can be seen in Figure 3. Based on the two multi-wavelet scaling functions, for an input data sample $\mathbf{x}$, the output expression for the hidden node is

$$h_j^{\text{out1}} = \begin{cases} -2\tau^3 + 3\tau^2 & \tau \in [0, 1) \\ 0 & \text{otherwise} \\ (2-\tau)^2(2\tau-1) & \tau \in [1, 2] \end{cases} \tag{6}$$

$$h_j^{\text{out2}} = \begin{cases} -\tau^2(3\tau-3) & \tau \in [0, 1) \\ 0 & \text{otherwise} \\ -(2-\tau)^2(3\tau-3) & \tau \in [1, 2] \end{cases} \tag{7}$$

with

$$\tau = \frac{\sum_{i=1}^{N} W_{ij}x_i - c_j}{a_j} \tag{8}$$

$$h_j^{\text{out}} = U_{ij-1} \cdot h_j^{\text{out1}} + U_{ij-2} \cdot h_j^{\text{out2}} \tag{9}$$

in which h_j^{out1} and h_j^{out2} refer to the output portions of $\varphi_1(t)$ and $\varphi_2(t)$ for hidden node j, respectively, and h_j^{out} is the final output of multi-wavelet functions, x_i is the ith dimension element of $\mathbf{x}$, W_{ij} is weight between hidden node j and input node i, U_{ij-1} and U_{ij-2} are weights between hidden node j and output node i based on $\varphi_1(t)$ and $\varphi_2(t)$, respectively, a_j and c_j are scale factor and shift factor, respectively.

The activation function of output layer is selected as Tanh, and then the reconstructed output can be calculated as

$$z_i = \text{Tanh}\left(\sum_{j=1}^{M}\left(U_{ij-1}\cdot h_j^{\text{out1}} + U_{ij-2}\cdot h_j^{\text{out2}}\right)\right) \quad (10)$$

$$\text{Tanh}(x) = (e^x - e^{-x})/(e^x + e^{-x}) \quad (11)$$

where z_i refers to the ith dimension element of $\mathbf{z}$, and M refers to the number of hidden nodes.

B. MODIFIED COST FUNCTION

The widely used cost function of basic AE model given in (3) is sensitive to learn features from non-stationary signals with complex noise [3]. Correntropy, a robust measure criterion [39], focuses on local similarity between two random vectors, which has shown advantages for dealing with complex signals with noise. Here, correntropy is used to modify the cost function to further reduce the reconstruction error, defined as

$$C_{corr} = \frac{1}{\sqrt{2\pi}\sigma}\sum_{i=1}^{N}\exp\left(-\frac{(x_i - z_i)^2}{2\sigma^2}\right) \quad (12)$$

in which $\sigma$is the kernel size of Gaussian kernel. To avoid over-fitting, a weight decay term is usually suggested to add into the cost function as well, and finally the cost function is modified as

$$\begin{aligned} C^{\text{Mod}} = &-\frac{1}{\sqrt{2\pi}\sigma}\sum_{i=1}^{N}\exp\left(-\frac{(x_i - z_i)^2}{2\sigma^2}\right) \\ &+ r\sum_{j=1}^{M}\left(\mu\log\frac{\mu}{\hat{\mu}_j} + (1-\mu)\log\frac{1-\mu}{1-\hat{\mu}_j}\right) \\ &+ \frac{\lambda}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{M}\left((W_{ij})^2 + (U_{ij-1})^2 + (U_{ij2})^2\right)\right) \end{aligned} \quad (13)$$

The weight parameters of the improved AE can be adjusted through iterative stochastic gradient descent by minimizing the modified cost function in (13), listed as follows

$$W_{ij} = W_{ij} - \eta\left(\partial C^{\text{Mod}}/\partial W_{ij}\right) \quad (14)$$

$$U_{ij-1} = U_{ij-1} - \eta\left(\partial C^{\text{Mod}}/\partial U_{ij-1}\right) \quad (15)$$

$$U_{ij-2} = U_{ij-2} - \eta\left(\partial C^{\text{Mod}}/\partial U_{ij-2}\right) \quad (16)$$

where η refers to the learning rate.

C. IMPROVED DEEP TRANSFER AUTO-ENCODER

It is necessary to add the depth of the improved AE model and introduce softmax classifier into the highest level, so as to refine the quality of the learned features and achieve classification ability meanwhile. Specifically, each individual improved AE model is pre-trained in unsupervised way through minimizing the modified loss function, then the learned features of previous improved AE model are fed into the input layer of the next improved AE model. The feature representations given by the last improved AE model are used as the input vector for softmax classifier. Figure 4 shows layer-by-layer construction process of the improved deep auto-encoder model with three improved AEs, in which *Feature* I, II, III are learned from *Improved AE* 1, 2, 3, respectively. More details about the construction of deep auto-encoders can be seen in [16].

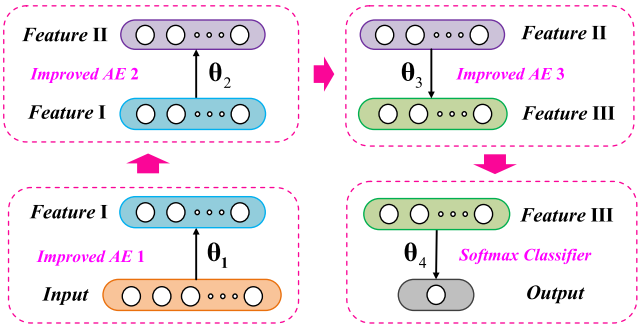

FIGURE 4. The construction of improved DAE with three improved AEs.

Improved deep transfer auto-encoder combines improved deep auto-encoder and the idea of parameter transfer. The specific process is described as follows. (1) Train an improved deep auto-encoder (contains softmax classifier) denoted as Deep model $^{(\mathbf{S})}$ using the training samples in the source domain. (2) The excellent performance of the trained Deep model $^{(\mathbf{S})}$ is tested by the testing samples in the source domain. (3) Design another improved deep auto-encoder model denoted as Deep model $^{(\mathbf{T})}$ with the completely same structure as Deep model $^{(\mathbf{S})}$. (4) Transfer the existing parameter knowledge of Deep model $^{(\mathbf{S})}$ to initialize Deep model $^{(\mathbf{T})}$, i.e., $\mathrm{W}^{(\mathbf{S})} = \mathrm{W}^{(\mathbf{T})}$, $\mathrm{U}^{(\mathbf{S})} = \mathrm{U}^{(\mathbf{T})}$. (5) Fine-tune Deep model $^{(\mathbf{T})}$ using small training samples from target domain to adapt to the characteristics of the remaining testing data. By now, the construction of improved deep auto-encoder has been successfully implemented, which can be used for transfer diagnosis of gearbox faults under variable working conditions, and the flowchart is given in Figure 5.

IV. CASE STUDY

CASE 1: TRANSFER DIAGNOSIS BETWEEN DIFFERENT WORKING CONDITIONS

A. EXPERIMENTAL GEARBOX DATA DESCRIPTION

In this case study, gearbox fault data provided by PHM 2009 Data Challenge is used to test the feasibility of the proposed approach [40]. Four spur gears are installed into the gearbox for simulating different health conditions, shown in Figure 6. Vibration data are collected at 66.67 kHz sampling frequency under five kinds of shaft speeds (30, 35, 40, 45 and 50Hz) and two kinds of loadings (High and Low). Some abbreviation rules are used for simplicity, i.e., 30L means the data form working condition with 30Hz (1800rpm) shaft speed and low loading, 50H means 50Hz (3000rpm) and high loading.

The vibration data from the input shaft is used in this case study. The source domain dataset is created by the collected

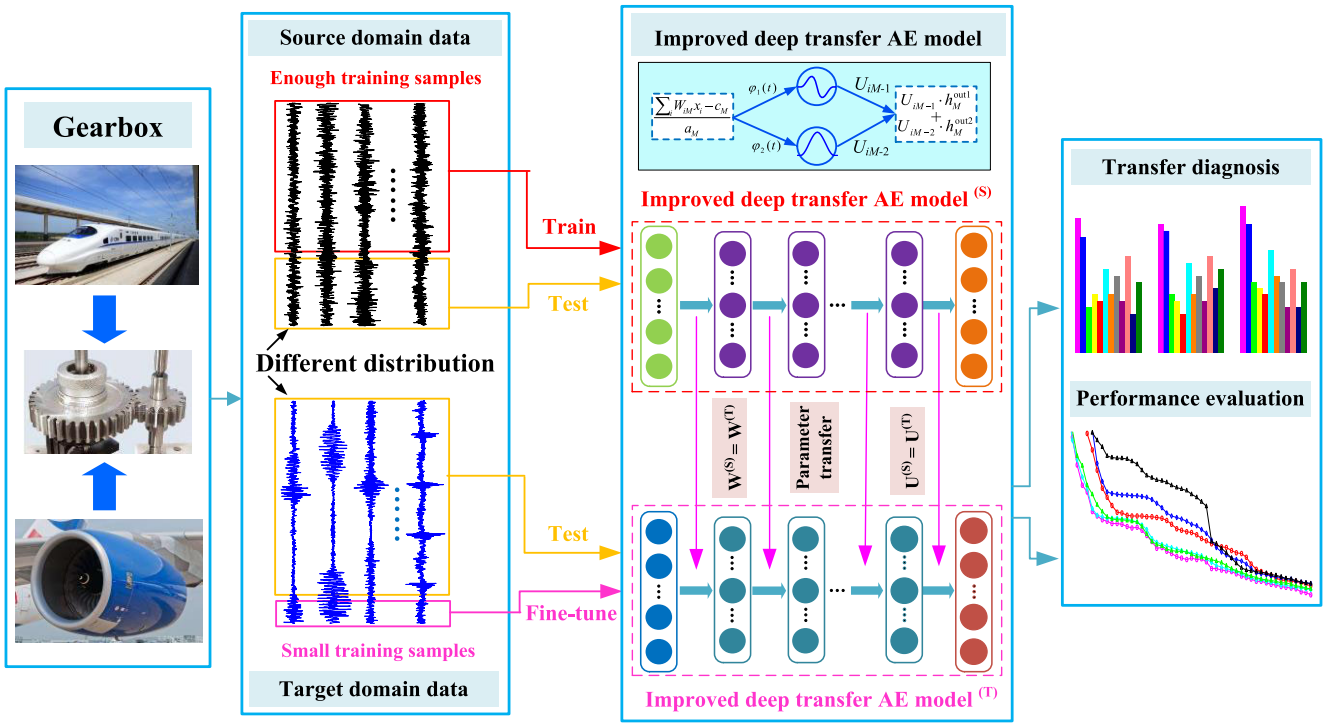

FIGURE 5. The framework of the proposed approach.

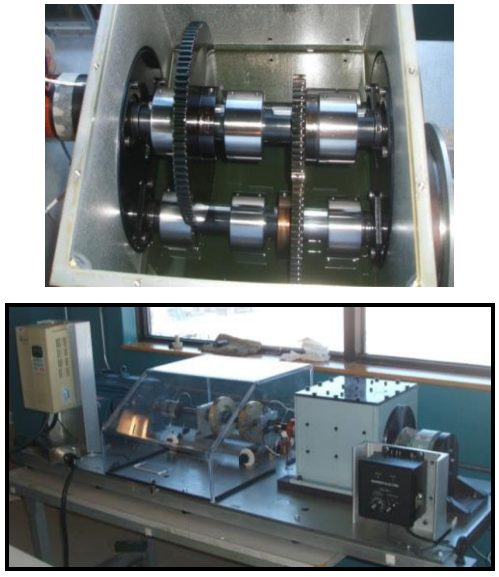

FIGURE 6. Gearbox fault test rig provided by PHM 2009 Data Challenge.

vibration data under 30L, and the target domain data is from 50H. Eight gearbox health conditions are created under different working conditions, including one normal condition and seven types of combined faults conditions, listed in Table I.

Each health condition from source domain has 145 samples consists of 120 training samples, while each target domain data only contains 10 training (Fine-tune) samples. Each sample refers to a signal segment including 6000 sampling points with 70% (4200 points) overlap. The details about the source domain and target domain can be seen in Table II. Eight kinds of Data samples (After removing the mean) are plotted in Figure 7. It can be found that there seems to be little similarity between the data samples from source domain and target domain, meaning that obvious changes of working conditions will lead to serious distribution discrepancy.

B. COMPARISONS WITH OTHER DEEP LEARNING METHODS WITHOUT TRANSFER STRATEGY

In order to verify the superiority of transfer learning strategy, some existing deep learning techniques are used for comparisons, including basic DAE (deep auto-encoder with Sigmoid), DBN (deep belief network) and CNN (convolutional neural network). The following two things should be noted:

- The proposed method is firstly trained by 120 training samples from source domain, and then fine-tuned by 10 training sample from target domain. After that, it is used for analyzing the rest 20 testing samples.
- For all the comparative methods, the training and testing samples are both from target domain (without parameter transfer). The numbers of training samples are 10, 40 and 100, respectively, while the numbers of testing samples are always set as 20.

A total of 10 repeated validations are carried out to examine the accuracy and stability meanwhile. For each method, the input is the normalized form of the raw vibration data (6000-dimensional). It can be seen from Table III that the average testing accuracy given by the proposed method is 93.06% (1489/1600). The average accuracies of the nine comparative methods are 41.81%, 74.06%, 82.44%, 42.31%, 70.19%, 80.25%, 39.13%, 71.25%, and 89.31%, respectively,

TABLE 1. Descriptions of eight types of gearbox health conditions.

Health conditions of gearbox	Gear			Bearing			Shaft	
	32T	48T	80T	IS :IS	ID :IS	OS :IS	Input	Output
Condition 1	Good	Good	Good	Good	Good	Good	Good	Good
Condition 2	**Chipped**	**Eccentric**	Good	Good	Good	Good	Good	Good
Condition 3	Good	**Eccentric**	Good	Good	Good	Good	Good	Good
Condition 4	Good	**Eccentric**	**Broken**	**Ball**	Good	Good	Good	Good
Condition 5	**Chipped**	**Eccentric**	**Broken**	**Inner**	**Ball**	**Outer**	Good	Good
Condition 6	Good	Good	**Broken**	**Inner**	**Ball**	**Outer**	**Imbalance**	Good
Condition 7	Good	Good	Good	**Inner**	Good	Good	Good	**Sheared**
Condition 8	Good	Good	Good	Good	**Ball**	**Outer**	**Imbalance**	Good

Remarks: IS=Input Shaft; ID=Idler Shaft; OS=Output Shaft; **:**IS=Input Side.

TABLE 2. Details about the source domain dataset and target domain dataset.

Transfer datasets	Rotating speed	Loading	Health condition of gearbox	Total size of samples	Total size of training samples
Source domain dataset	**1800rpm (30Hz)**	**Low**	Condition 1- Condition 8	1160 (120+25)*8	960 (120*8)
Target domain dataset	**3000rpm (50Hz)**	**High**	Condition 1- Condition 8	240 (**10**+20)*8	**80 (10*8)**

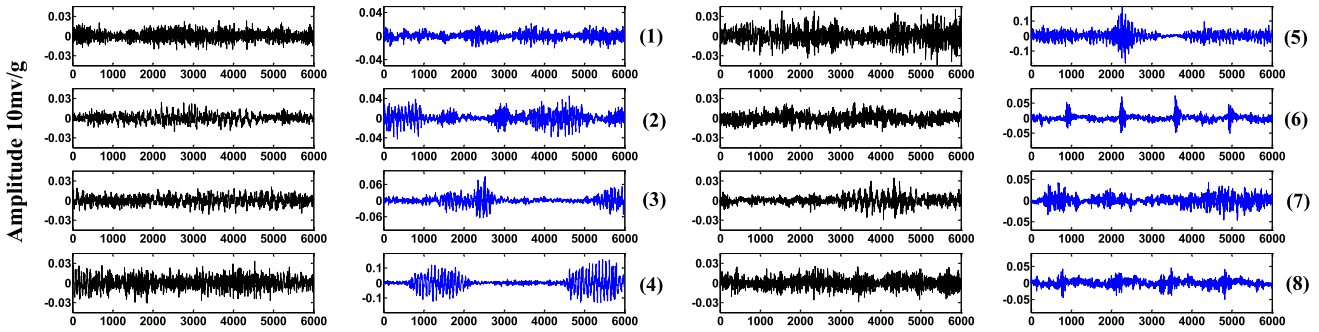

FIGURE 7. The vibration waveforms of the data samples from the eight gearbox health conditions: Source domain (black) and target domain (blue).(1-8 mean condition 1–condition 8.

which are lower than the proposed method. Besides, the standard deviation given by the proposed approach is 0.6215, and it is smaller than all the comparative approaches, meaning that the proposed approach holds better stability. Through the comparison results, the following two conclusions can be drawn. (1) The diagnosis performance of deep learning techniques depends heavily on sample size, without large amounts of training samples, deep learning techniques usually fail to show satisfactory results. (2) The proposed method is more effective than other deep learning techniques without transfer learning strategy. The main reason is that train a good deep neural network from scratch is difficult and time-consuming because lots of weights and biases are randomly initialized. In the proposed method, the number of adjusted parameters can be greatly reduced and reasonable initialization can be achieved through transferring parameters of model pre-trained by the source domain data to the target model. To enable the target model to further adapt to the characteristics of the testing samples in the target domain, small target training samples are then used to fine-tune the well pre-trained model.

The specific structure of the proposed method are 6000-2000-800-150-8, meaning that 2000, 800 and 150 nodes exist in the first, second and third hidden layers, respectively, which is determined by experimentation with a simple idea similar to [16]. The iteration numbers in the pre-trained process and fine-tuning process are 60 and 25, respectively. parameters r,μ,λ, σ are 4, 0.08, 0.002 and 0.5, respectively. Most of these parameters are decided through experimentations. The structure of basic DAE is also 6000-2000-800-150-8, iteration numbers in the pre-trained process and fine-tuning process are both set as 100,

TABLE 3. Diagnosis results of different methods.

Diagnosis methods	Number of source training samples	Number of target training samples	Diagnosis results
Proposed	**120**	**10**	**93.06% ± 0.6215**
Basic DAE	0	10	41.81% ± 2.2172
Basic DAE	0	40	74.06% ± 1.5574
Basic DAE	0	100	82.44% ± 1.0391
Basic DBN	0	10	42.31% ± 2.3308
Basic DBN	0	40	70.19% ± 1.8087
Basic DBN	0	100	80.25% ± 1.3455
Basic CNN	0	10	39.13% ± 2.0453
Basic CNN	0	40	71.25% ± 1.2905
Basic CNN	0	100	89.31% ± 0.8019

Diagnosis results: Average testing accuracy ± standard deviation.

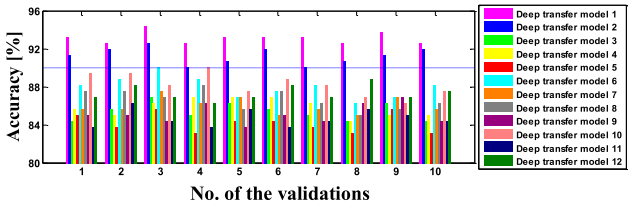

FIGURE 8. Statistical diagnosis results of different deep transfer models.

TABLE 4. The comparison results among different deep transfer models.

Deep transfer models	Average accuracy	Standard deviation
Deep transfer model 1	**93.06% (1489/1600)**	**0.6215**
Deep transfer model 2	91.19% (1459/1600)	0.8565
Deep transfer model 3	85.38% (1366/1600)	0.8937
Deep transfer model 4	85.81% (1373/1600)	0.9340
Deep transfer model 5	84.19% (1347/1600)	0.9794
Deep transfer model 6	87.94% (1407/1600)	1.1044
Deep transfer model 7	86.00% (1376/1600)	0.8437
Deep transfer model 8	86.63% (1386/1600)	1.0291
Deep transfer model 9	85.13% (1362/1600)	1.0121
Deep transfer model 10	88.19% (1411/1600)	1.1950
Deep transfer model 11	84.69% (1355/1600)	0.8961
Deep transfer model 12	87.31% (1397/1600)	0.7823

parameters r,μ,λ are selected as 4, 0.08 and 0.002, respectively. The structure of basic DBN is 6000-2000-800-150-8, learning rate, iteration number and momentum are set as 0.15, 100 and 0.85, respectively. The structure of basic CNN called LeNet-5 consists of an input layer, two convolutional layers, two pooling layers and an output layer [41]. No further skills are used for improving the basic DAE, DBN and CNN.

C. COMPARISONS WITH OTHER DEEP TRANSFER MODELS

To the verify the superiority of the proposed deep transfer model (DAE: Multi-W & C^{Mod}), nine kinds of DAE transfer models and two kinds of DBN transfer models are used for comparison, including transfer model 2 (DAE: Multi-W & C^{Tra}), transfer model 3 (DAE: Sigmoid & C^{Tra}), transfer model 4 (DAE: Tanh & C^{Tra}), transfer model 5 (DAE: ReLU & C^{Tra}), transfer model 6 (DAE: Morlet & C^{Tra}), transfer model 7 (DAE: Sigmoid & C^{Mod}), transfer model 8 (DAE: Tanh & C^{Mod}), transfer model 9 (DAE: ReLU & C^{Mod}), transfer model 10 (DAE: Morlet & C^{Mod}), transfer model 11 (Basic DBN) and transfer model 12 (Gaussian DBN) [42]. The statistical results of 10 validations are given in Figure 8 and Table IV. The average accuracy on the testing samples of the proposed approach is 93.06% (1489/1600), which is higher than other 11 kinds of deep transfer models. For the third validation as example, the confusion matrix is shown in Figure 9. Thus, the proposed deep transfer model has higher accuracy and better stability than other deep transfer models faced with the same transfer diagnosis task.

The good performance of the proposed deep transfer model mainly benefits from the replaced multi-wavelet activation function and modified cost function. The superiority of the modified cost function can be repeated proved by comparing deep transfer models 1 & 2, deep transfer models 7 & 8, and deep transfer models 9 & 10, because these three pairs are all designed with the same activation function while different cost functions. The feasibility of multi-wavelet activation function can be tested through comparing deep transfer models 1, 7, 8, 9 and 10, because they are constructed with the modified cost function while different activation functions. Take the third validation as an example, Figure 10 is the reconstruction error curves of these models in pre-trained process. It can be observed that the reconstruction error given by deep transfer model 1 is smaller and provides faster convergence than others.

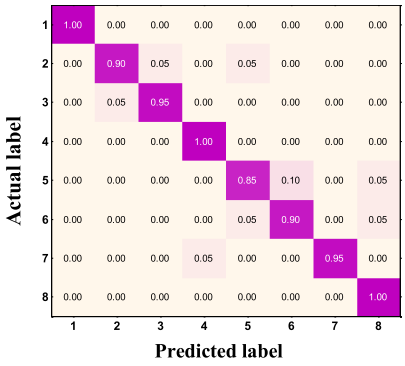

FIGURE 9. Confusion matrix of the proposed method for the third validation.

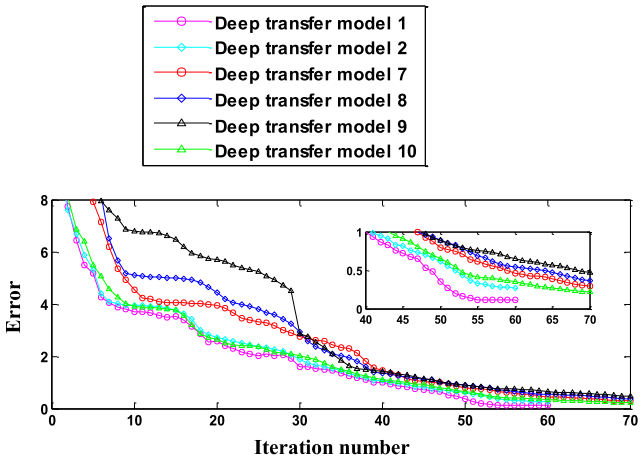

FIGURE 10. The reconstruction error curves of different deep transfer models in pre-trained process.

D. CONSIDERATION FOR TIME-DELAY

As mentioned before, each sample refers to a signal segment including 6000 sampling points, and two consecutive samples have 70% (4200 points) overlap, meaning that the first sample is [1, 6000] (from the 1st data points to the 6000th data points), the second is [1801, 7200], and so on. In order to fully test the proposed method, the time-delay problem of time-series data is considered. The following two things should be noted:

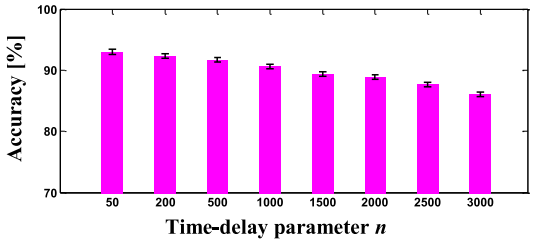

FIGURE 11. The diagnosis results of the proposed method under different time-delay parameters.

- For source samples, the first sample still is [1, 6000], the second is [1801, 7200], and so on. Each condition has 145 samples consists of 120 training samples.
- For target samples, the first sample is [1+n, 6000+n], the second is [1801+n, 7200+n], and so on. The time-delay parameter n ranges from 50 to 3000. Each condition has 10 samples for fine-tuning and 20 samples for testing.

Figure 11 shows the diagnosis results (average accuracy) of the proposed method under different time-delay parameters (50 to 3000). From Figure 12, it can be seen that the average testing accuracy of the proposed method become smaller (from 93% to 86%) with increase of time-delay degrees. The reason is that there exist distribution differences between training samples and testing samples, and time-delay problem will lead to larger differences. However, even time-delay parameter reach to 3000, the result given by the proposed method is still higher than 86%, because small training samples in the target domain are used to fine-tune the pre-trained model to adapt to the characteristics of the rest testing samples.

CASE 2: TRANSFER DIAGNOSIS FROM CONSTANT SPEED TO VARIABLE SPEEDS

The transfer diagnosis task in CASE 1 actually belongs to piecewise variable working conditions, however, in practical engineering, dynamic working regimes are more common. Thus, the feasibility of the proposed method for transfer fault diagnosis from constant speed to variable speeds is considered in CASE 2. The experimental device is shown in Figure 12, mainly consists of motor, tested gear (37 teeth) and bearing (SKF 6307). Four types of vibration data from gear and bearing are collected with sampling frequency of 8192 Hz, including tooth breakage, tooth breakage & outer race fault, tooth breakage & inner race fault and tooth breakage & ball fault.

Due to lack of engineering data, here, the collected vibration data under constant speed (600rpm) is treated as lab data, and the data from variable speeds is simulated as industrial on-site data. The details of variable speeds for four fault conditions are shown in Figure 13. It can be seen that the changing patterns of the rotating speeds are different.

Each fault condition in the source domain has 158 samples consists of 140 training samples, while in target domain each fault condition contains only 30 training (fine-tuning)

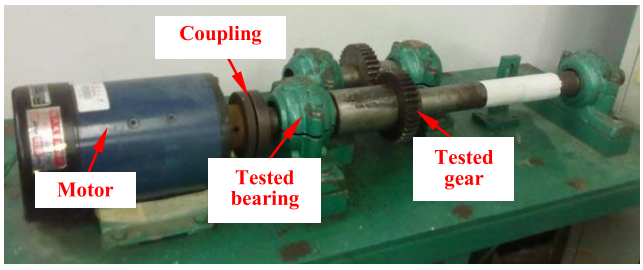

FIGURE 12. Experimental setup for simulating faults of gear and bearing.

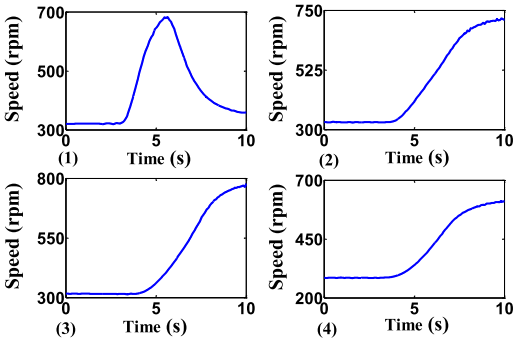

FIGURE 13. The details of variable rotating speeds for four fault conditions: (1) Tooth breakage, (2) Tooth breakage & Outer race fault, (3) Tooth breakage & Inner race fault, (4) Tooth breakage & Ball fault.

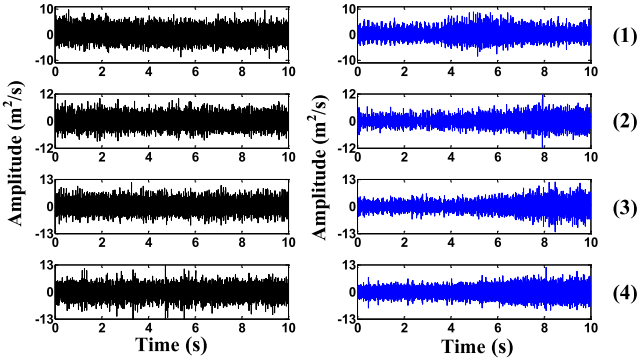

FIGURE 14. The collected vibration signals of four fault conditions: (1) Tooth breakage, (2) tooth breakage & outer race fault, (3) tooth breakage & inner race fault, (4) tooth breakage & ball fault. (source: black, target: blue).

samples and 40 testing samples. It should be noted that for each fault condition, the 30 target training samples are selected from three different positions (beginning, middle and end) in the collected signal. Each sample consists of 1024 data points with 50% overlap. Raw signals of the four fault conditions in this case are plotted in Figure 14. From Figure 14, it can be seen that target domain samples show strong non-stationary characteristics due to fluctuation of the rotating speeds during data acquisition process, leading to significant differences from source domain.

In this case study, fast Fourier transform, a well-known signal processing technique, is applied to acquire frequency spectrum (512-dimensional) of each data sample, so as to reduce the differences among different data samples caused by changeable speeds. Take three samples

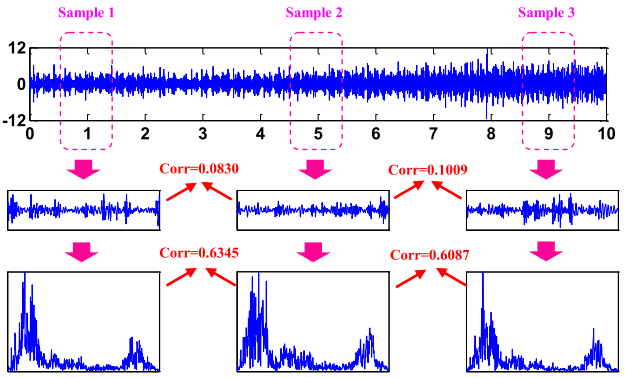

FIGURE 15. Frequency spectrums of three samples from different speeds.

TABLE 5. The comparison results in case 2.

Deep transfer models	Average accuracy	
	Raw data	Frequency spectrum
Deep transfer model 1	**86.13%**	**90.56%**
Deep transfer model 2	83.44%	88.81%
Deep transfer model 7	80.25%	85.31%
Deep transfer model 8	80.56%	85.50%
Deep transfer model 9	80.38%	85.13%
Deep transfer model 10	83.25%	86.44%

(Samples 1, 2, and 3) from Condition 2 (Tooth breakage & Outer race fault) of target domain as examples, calculate the correlation coefficients between Samples 1 & 2 and Samples 2 & 3, respectively, as shown in Figure 15. It can be seen that their frequency spectrums hold more similarities than the raw data.

A total of 10 repeated validations are carried out to compare the diagnosis performance among deep transfer models 1, 2, 7, 8, 9 and 10, as listed in Table V. It can be seen that: (1) the average testing accuracies of all the methods based on frequency spectrum are higher than the raw data. (2) The proposed method on frequency spectrum reaches to 90.56% (1449/1600, 1600=4∗40∗10), and it is higher than the accuracies using the five comparative methods, which are 88.81%, 85.31%, 85.50%, 85.13% and 86.44%, respectively. (3) Although the proposed method gives the best diagnosis results, it cannot be compared with the CASE 1. In order to further improve the accuracy, more advanced techniques should be introduced, such as nuisance attribute projection and order tracking. As a summary, with the help of multi-wavelet, modified cost function and parameter transfer idea, the diagnosis knowledge learned from constant speed working conditions can be transferred to variable speeds to some degree.

V. CONCLUSIONS

In this paper, multi-wavelet activation function and modified cost function are used to enhance the deep auto-encoder. Based on improved deep auto-encoder and parameter transfer, improved deep transfer auto-encoder is proposed to diagnose gearbox faults under variable working conditions with small training samples.

Two sets of experimental vibration data of gearbox are used to validate the superiority of the proposed approach. The results demonstrate that the proposed method can accurately diagnosis different faults of gearbox, even the working conditions have significant changes, which is better than the existing methods. Deep transfer learning is able to solve hard tasks from largely different domains, which has big potential to be applied in engineering practice. Despite this paper preliminarily explores the applications of piecewise variable working conditions and simple dynamic working regimes, more complex and practical cases fails to be considered. The authors will continue to study this meaningful issue by analyzing more practical industrial on-site datasets and focus on the real applications of deep transfer learning in the future.

REFERENCES

[1] Y. Wang, B. Tang, L. Meng, and B. Hou, "Adaptive estimation of instantaneous angular speed for wind turbine planetary gearbox fault detection," *IEEE Access*, vol. 7, pp. 49974–49984, 2019.
[2] L. Jing, M. Zhao, P. Li, and X. Xu, "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox," *Measurement*, vol. 111, pp. 1–10, Dec. 2017.
[3] H. Shao, H. Jiang, H. Zhao, and F. Wang, "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 95, pp. 187–204, Oct. 2017.
[4] 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. Inf.*, vol. 13, no. 3, pp. 1310–1320, Jun. 2017.
[5] J.-H. Zhong, J. Zhang, J. Liang, and H. Wang, "Multi-fault rapid diagnosis for wind turbine gearbox using sparse Bayesian extreme learning machine," *IEEE Access*, vol. 7, pp. 773–781, 2018.
[6] Ł. Jedliński and J. Jonak, "Early fault detection in gearboxes based on support vector machines and multilayer perceptron with a continuous wavelet transform," *Appl. Soft Comput.*, vol. 30, pp. 636–641, May 2015.
[7] D. Dabrowski, "Condition monitoring of planetary gearbox by hardware implementation of artificial neural networks," *Measurement*, vol. 91, pp. 295–308, Sep. 2016.
[8] T. Wang, Q. Han, F. Chu, and Z. Feng, "Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review," *Mech. Syst. Signal Process.*, vol. 126, pp. 662–685, Jul. 2019.
[9] X. Yan and M. Jia, "Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection," *Knowl.-Based Syst.*, vol. 163, pp. 450–471, Jan. 2019.
[10] M. Cerrada, C. Li, R.-V. Sánchez, F. Pacheco, D. Cabrera, and J. V. de Oliveira, "A fuzzy transition based approach for fault severity prediction in helical gearboxes," *Fuzzy Sets Syst.*, vol. 337, pp. 52–73, Apr. 2018.
[11] J. Qu, Z. Zhang, and T. Gong, "A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion," *Neurocomputing*, vol. 171, pp. 837–853, Jan. 2016.
[12] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.
[13] 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.
[14] 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.
[15] H. Shao, H. Jiang, F. Wang, and H. Zhao, "An enhancement deep feature fusion method for rotating machinery fault diagnosis," *Knowl.-Based Syst.*, vol. 119, pp. 200–220, Mar. 2017.
[16] F. Jia, Y. G. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 303–315, May 2016.

[17] M. Ma, C. Sun, and X. Chen, "Deep coupling autoencoder for fault diagnosis with multimodal sensory data," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1137–1145, Mar. 2018.
[18] H. Hu, B. Tang, X. Gong, W. Wei, and H. Wang, "Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2106–2116, Aug. 2017.
[19] G. Jiang, H. He, P. Xie, and Y. Tang, "Stacked multilevel-denoising autoencoders: A new representation learning approach for wind turbine gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 9, pp. 2391–2402, Sep. 2017.
[20] F. Cheng, J. Wang, L. Qu, and W. Qiao, "Rotor-current-based fault diagnosis for DFIG wind turbine drivetrain gearboxes using frequency analysis and a deep classifier," *IEEE Trans. Ind. Appl.*, vol. 54, no. 2, pp. 1062–1071, Mar./Apr. 2018.
[21] Y. Qi, C. Shen, D. Wang, J. Shi, X. Jiang, and Z. Zhu, "Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery," *IEEE Access*, vol. 5, pp. 15066–15079, 2017.
[22] W. Sun, R. Zhao, R. Yan, S. Shao, and X. Chen, "Convolutional discriminative feature learning for induction motor fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1350–1359, Jun. 2017.
[23] S. Haidong, J. Hongkai, L. Xingqiu, and W. Shuaipeng, "Intelligent fault diagnosis of rolling bearing using deep wavelet auto-encoder with extreme learning machine," *Knowl.-Based Syst.*, vol. 140, pp. 1–14, Jan. 2018.
[24] M. H. Zhao, M. Kang, B. Tang, and M. Pecht, "Multiple wavelet coefficients fusion in deep residual networks for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4696–4706, Jun. 2019.
[25] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[26] S. M. Salaken, A. Khosravi, T. Nguyen, and S. Nahavandi, "Seeded transfer learning for regression problems with deep learning," *Expert Syst. Appl.*, vol. 115, pp. 565–577, Jan. 2019.
[27] S. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.
[28] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, May 2018.
[29] 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.
[30] 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.
[31] W. Qian, S. Li, and J. Wang, "A new transfer learning method and its application on rotating machine fault diagnosis under variant working conditions," *IEEE Access*, vol. 6, pp. 69907–69917, 2018.
[32] I. Vamsi, G. R. Sabareesh, and P. K. Penumakala, "Comparison of condition monitoring techniques in assessing fault severity for a wind turbine gearbox under non-stationary loading," *Mech. Syst. Signal Process.*, vol. 124, pp. 1–20, Jun. 2019.
[33] H. Shao, H. Jiang, H. Zhang, and T. Liang, "Electric locomotive bearing fault diagnosis using a novel convolutional deep belief network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 3, pp. 2727–2736, Mar. 2018.
[34] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying ReLU and initialization: Theory and numerical examples," 2019, *arXiv:1903.06733*. [Online]. Available: https://arxiv.org/abs/1903.06733
[35] A. Subasi, A. Alkan, E. Koklukaya, and M. K. Kiymik, "Wavelet neural network classification of EEG signals by using AR model with MLE preprocessing," *Neural Netw.*, vol. 18, no. 7, pp. 985–997, Sep. 2005.
[36] L. Jiao, J. Pan, and Y. Fang, "Multiwavelet neural network and its approximation properties," *IEEE Trans. Neural Netw.*, vol. 12, no. 5, pp. 1060–1066, Sep. 2001.
[37] G. Plonka and V. Strela, "Construction of multiscaling functions with approximation and symmetry," *SIAM J. Math. Anal.*, vol. 29, no. 2, pp. 481–510, Mar. 1998.
[38] H. Dai, H. Zhang, and W. Wang, "A multiwavelet neural network-based response surface method for structural reliability analysis," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 30, pp. 151–162, Feb. 2015.
[39] R. He, B.-G. Hu, W.-S. Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1485–1494, Jun. 2011.
[40] J. Xie, L. Zhang, L. Duan, and J. Wang, "On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis," in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, Jun. 2016, pp. 1–6.
[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[42] Z. Liu, Z. Jia, C.-M. Vong, S. Bu, and J. Han, "Capturing high-discriminative fault features for electronics-rich analog system via deep learning," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1213–1226, Jun. 2017.

ZHIYI HE received the M.S. degree in mechanical engineering from Hunan University, Changsha, China, in 2016, where he is currently pursuing the Ph.D. degree in mechanical engineering with the College of Mechanical and Vehicle Engineering.

His research interests include rotating machinery fault diagnosis, artificial intelligence, and machine learning applications.

HAIDONG SHAO received the Ph.D. degree in vehicle operation engineering from Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently an Assistant Professor with the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China.

His current research interests include fault diagnosis, intelligent prognosis, and information fusion.

XIAOYANG ZHANG received the B.Sc. degree from Northwestern Polytechnical University, Xi'an, China, in 2013, and the M.Sc. degree from The University of Nottingham, Nottingham, U.K., in 2018. He is currently an Assistant Engineer with the Aeronautics Computing Technique Research Institute, Aviation Corporation of China.

His current research interests include aeronautics computers and electromechanical computers.

JUNSHENG CHENG received the Ph.D. degree in manufacturing engineering and automation from Hunan University, Changsha, China, in 2005, where he is currently a Professor with the College of Mechanical and Vehicle Engineering.

His main research interests include mechanical fault diagnosis, dynamics signal processing, and vibration and noise control.

YU YANG received the Ph.D. degree in mechanical engineering from the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China, in 2005, where she is currently a Professor with the College of Mechanical and Vehicle Engineering.

Her research interests include pattern recognition, digital signal processing, and machine fault diagnosis.