Hindawi
Shock and Vibration
Volume 2020, Article ID 8884179, 13 pages
https://doi.org/10.1155/2020/8884179

Research Article

Deep Transfer Learning-Based Fault Diagnosis for Gearbox under Complex Working Conditions

Zitong Wan,[1,2] **Rui Yang**,[3,4] **and Mengjie Huang**[1]

[1] *Design School, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*
[2] *Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3BX, UK*
[3] *School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*
[4] *Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

Correspondence should be addressed to Rui Yang; r.yang@xjtlu.edu.cn

Received 4 August 2020; Revised 23 September 2020; Accepted 28 October 2020; Published 16 November 2020

Academic Editor: Changqing Shen

In the large amount of available data, information insensitive to faults in historical data interferes in gear fault feature extraction. Furthermore, as most of the fault diagnosis models are learned from offline data collected under single/fixed working condition only, this may cause unsatisfactory performance for complex working conditions (including multiple and unknown working conditions) if not properly dealt with. This paper proposes a transfer learning-based fault diagnosis method of gear faults to reduce the negative effects of the abovementioned problems. In the proposed method, a cohesion evaluation method is applied to select sensitive features to the task with a transfer learning-based sparse autoencoder to transfer the knowledge learnt under single working condition to complex working conditions. The experimental results on wind turbine drivetrain diagnostics simulator show that the proposed method is effective in complex working conditions and the achieved results are better than those of traditional algorithms.

1. Introduction

With the extensive application of technology in industrial production, fault diagnosis is playing an increasingly important role. In production, the occurrence of abnormal accidents can be avoided and economic losses and casualties can be reduced through timely detection of equipment fault [1]. Data-driven fault diagnosis has high accuracy for practical complex system diagnosis tasks such as gear faults in rotating machinery due to its complex structure which is hard to carry out mathematical modeling [2, 3]. It consists of three main directions: signal processing, statistical analysis, and artificial intelligence-based methods [4]. In signal processing methods, the signals are analyzed by several techniques to extract fault features, such as wavelet filter and singular spectrum analysis [5, 6]. Statistical analysis methods utilize the statistical methods such as principal component analysis and partial least square methods to analyze the historical data [7, 8]. Artificial intelligence-based methods apply different artificial intelligence techniques in fault diagnosis, such as neural network, support vector machine, and fuzzy logic [9–11].

Among fault diagnosis methods-based on artificial intelligence, deep learning method has been broadly applied in detecting abnormal situations. Deep learning methods such as recurrent neural network and convolution neural network are widely studied and applied in the field of fault diagnosis in industrial systems due to their self-learning and adaptivity [12, 13].

Sparse autoencoder (SAE) is a type of neural network that can learn features from unlabeled data, which was proposed based on autoencoder (AE) in 2006 [14]. AE takes the input information as its learning target to extract features and reduce dimensionality through encoding and decoding [15, 16]. In fault diagnosis, AE is trained to extract the features of input data, which is not suitable when the distribution of testing data is different from training data [17, 18]. To fortify the adaptability and flexibility of the network model, the concept of transfer learning is proposed to apply the knowledge learned in a pretrained model into novel tasks [19, 20].

With the above literature review on artificial intelligence-based fault diagnosis methods, there are still two main difficulties in fault diagnosis of practical complex systems such as rotating machinery:

(1) *Insensitive Information.* Insensitive information can be described as the components caused by irrelevant variables in original signals [21, 22]. Liu combined a 1D autoencoder and convolutional neural network in detecting faults of rotating machinery under noisy environment [23]. Wang adopted conditional variational neural networks to extract the features of planetary gearbox under noisy environment [24]. Zhang proposed a deep learning model based on convolutional neural network with wide first-layer kernels for fault diagnosis to withstand interference information [25]. The research works in literature review did not consider the effect of insensitive information, such as the features that contribute little or eventually have a negative interference fault diagnosis performance. The other problem with these approaches is that they did not consider the performances of the proposed methods under different working conditions, which is discussed afterwards.

(2) *Complex Working Conditions.* In the actual production, system operational parameters give rise to complex working conditions such as multiple working conditions and even unknown working conditions. The model trained under single working condition is not able to effectively adapt to complex conditions on this occasion [26]. Moreover, serious distribution discrepancy can be observed between training data and testing data when the structures of two data sets are different [27]. To solve this problem, many approaches have been proposed by researchers. Wang discussed domain adaptation for different conditions in transfer learning for gearing fault diagnosis [28]. Hasan adopted transfer learning and convolutional neural network in bearing to make sure that the model is adaptable in different working conditions [29]. Qian proposed a new transfer learning method to detect faults of rotating machine under variant working conditions [30]. However, these previously mentioned research works have limitations: first, there is no discussion or analysis on both multiple working conditions and unknown working conditions; second, the handling of insensitive information which can affect the fault diagnosis performance is not comprehensive.

Based on the above literature review, there is no research work at present that investigates the effects of both two difficulties simultaneously. The contributions of this paper are listed as follows: (1) the problem with both complex working conditions and insensitive information is investigated; (2) a deep transfer learning-based fault diagnosis method with sensitive features selection and the combination of SAE based on transfer learning is proposed for the abovementioned problems. To reduce the difficulty of signal analysis under complex conditions, transfer learning is applied to adjust the accuracy of the model under such conditions. Transfer learning refers to applying the prior knowledge learned from one task to a different but related task, which was first proposed at the 1995 NIPS-95 seminar on "Learning to Learn" [31]. Transfer learning reduces the cost of model construction and data requirement when there is difference between source and target data, which is applied in different fields, such as data mining, image recognition, language translation, fault diagnosis, and fault diagnosis [32–35].

The rest of this paper is organized as follows. In Section 2, the proposed algorithm is introduced in detail, including sensitive features selection, SAE, and MMD. In Section 3, hardware experiments are conducted on wind turbine drivetrain diagnostics simulator to show the effectiveness of the proposed method for five fault types of gear. Conclusion is made in Section 4.

2. The Proposed Method

In order to increase the accuracy of fault diagnosis under complex working conditions, a transfer learning-based fault diagnosis method using sensitive features selection is proposed in this section. The relevant methods and algorithm details are explained in the following four subsections.

2.1. Sensitive Features Selection. The signal properties in time and frequency domain including amplitude, probability distribution, and energy change when the fault occurs. In rotating machinery, usually 11 time domain and 13 frequency domain characteristic parameters are analyzed for fault diagnosis [36]. However, the large number of characteristic parameters incurs the following two problems: (1) fault features may not be accurately extracted due to the random components in the signal; (2) large dimension data enhance the modeling difficulty [37]. In this paper, cohesion evaluation is applied to select the sensitive features, which can reserve sensitive features and remove insensitive features by evaluating the cohesion of each feature [38].

Suppose that there is a feature set containing H categories, with $q_{m,h,j}$ denoting j-th feature of m-th sample in the h categories as shown in

$$\left\{q_{m,h,j}, \quad m = 1, 2, \ldots, M_h; h = 1, 2, \ldots H; j = 1, 2, \ldots, J\right\}, \tag{1}$$

where M_h represents the number of samples h and J is the number of features in each category. Table 1 lists the steps to compute the cohesion factor γ_j, which reflects feature sensitivity from the intercategory and intracategory cohesion. Cohesion indicates the relationship among categories based on standard deviation, which reflects the details of the difference in the overall data distribution. In Table 1, difference of intracategory distance and difference of intercategory distance are computed based on the average

Table 1: Cohesion evaluation.

Step	Process parameter	Equation
1	Intracategory distance	$d_{h,j} = ((\sum_{m,r=1}^{M_h} \lvert q_{m,h,j} - q_{r,m,j}\rvert)/(M_h(M_h-1)))$
2	Average intracategory distance	$d_j^{\text{inner}} = (1/H)\sum_{H=1}^{H} d_{H,j}$
3	Difference of intracategory distance	$e_j^{\text{inner}} = (\max(d_{h,j})/\min(d_{c,j}))$
4	Average of each feature	$u_{h,j} = (1/M_h)\sum_{m=1}^{M_h} q_{m,h,j}$
5	Average intercategory distance	$d_j^{\text{outer}} = ((\sum_{h,c=1}^{S} \lvert u_{h,j} - u_{c,j}\rvert)/H(H-1))$
6	Difference between categories	$e_j^{\text{outer}} = (\max(\lvert u_{h,j} - u_{c,j}\rvert)/\min(\lvert u_{k,j} - u_{z,j}\rvert))$
7	Distance weighting factor	$\mathrm{d}w_j = (1/((e_j^{\text{inner}}/\max(e_j^{\text{inner}})) + (e_j^{\text{outer}}/\max(e_j^{\text{outer}}))))$
8	Distance evaluation factor	$\alpha_j = (\mathrm{d}w_j d_j^{\text{outer}}/d_j^{\text{inner}})$
9	Intracategory standard deviation	$\sigma_{h,j} = \sqrt{(\sum_{m=1}^{M_h} (q_{m,h,j} - u_{h,j})^2/M_h - 1)}$
10	Average intracategory standard deviation	$clt_j^{\text{inner}} = (1/H)\sum_{h=1}^{H} \sigma_{h,j}$
11	Difference of intracategory standard deviation	$f_j^{\text{inner}} = (\max(clt_{h,j})/\min(clt_{c,j}))$
12	Distance of each feature	$cd_{n,r,s,j} = \lvert q_{n,s,j} - q_{r,s,j}\rvert$
13	Quadratic sum of feature distance	$qs = \sum_{m,r=1}^{M_s} (cd_{m,r,h,j} - d_{h,j})^2$
14	Standard deviation of feature distance (intracategory cohesion)	$stc_{h,j} = \sqrt{(qs/(M_h(M_h-1)-2))}$
15	Average intercategory cohesion difference	$clt_j^{\text{outer}} = ((\sum_{h,c=1}^{H} \lvert stc_{h,j} - stc_{c,j}\rvert)/H(H-1))$
16	Imparity measure of intercategory cohesion difference	$f_j^{\text{outer}} = (\max(\lvert stc_{h,j} - stc_{c,j}\rvert)/\min(\lvert stc_{k,j} - stc_{z,j}\rvert))$
17	Cohesion weighting factor	$cw_j = (1/((f_j^{\text{inner}}/\max(f_j^{\text{inner}})) + (f_j^{\text{outer}}/\max(f_j^{\text{outer}}))))$
18	Cohesion factor	$\gamma_j = (cw_j clt_j^{\text{outer}}/clt_j^{\text{inner}})$

distance from step 1 to step 6. Distance factor of every category is computed by intercategory and intracategory ratios using step 7 and step 8. Average intracategory standard deviation is calculated to gain the difference from step 9 to step 11. Next, average intercategory cohesion difference is defined and computed from step 12 to step 16. Finally, in order to obtain cohesion factor, a weighting factor is defined in step 17 to measure cohesion difference which is similar to the distance factor. It is used to evaluate the cohesion of each category, which can be distinguished if the cohesion difference between intracategory and intercategory is large.

There are two problems mainly causing the difficulty in accurate sensitive feature extraction: (1) large intracategory distance: in this situation, sensitive features sorted by distance evaluation factor α_j only are not accurate and some sensitive features can be discarded as insensitive features since large intracategory distance d_j^{inner} reduces the priority of sensitive features; (2) large intracategory cohesion difference: in this premise, there is an overlap between categories if the intercategory cohesion difference is small, resulting in inaccurate selection of sensitive features. The two problems mentioned above can be solved by combining distance and cohesion evaluations, which prevents one of them from producing excessive effect on the result.

The sensitive factor η_j is computed in the following equation, which is combined with distance evaluation factor and cohesion factor:

$$\eta_j = \beta_j \frac{clt_j^{\text{outer}} + coe * d_j^{\text{outer}}}{clt_j^{\text{inner}} + coe * d_j^{\text{inner}}}, \tag{2}$$

where *coe* is a coefficient to modulate the proportion of distance and cohesion evaluation and β_j is the sensitivity weighting coefficient. According to the distance evaluation factor α_j of step 8 and cohesion factor γ_j of step 18 in Table 1, the sensitivity weighting coefficient β_j is represented in

$$\beta_j = \frac{1}{\left(f_j^{\text{inner}}/\max\left(f_j^{\text{inner}}\right)\right) + \left(f_j^{\text{outer}}/\max\left(f_j^{\text{outer}}\right)\right) + \left(e_j^{\text{inner}}/\max\left(e_j^{\text{inner}}\right)\right) + \left(e_j^{\text{outer}}/\max\left(e_j^{\text{outer}}\right)\right)}. \tag{3}$$

The sensitive factor η_j reflects the influence degree of different features in categories. The sensitivities of features are sorted from large to small according to the value of η_j. Through sensitivity factor, features that are sensitive to classification are retained, while the insensitive information is discarded without figuring out the type of the features. This preprocessing reduces the complexity of subsequent computation and can help improve the classification accuracy.

2.2. Transfer Learning. Transfer learning is type of a learning mode applying prior knowledge learnt in a task in solving related but different tasks. The prior knowledge including data features and labels can assist the analysis of a related task when it is difficult to process directly due to data collection difficulty, high modeling cost, and long training time. In transfer learning, a domain refers to a data set and its probability distribution. Particularly, the domain containing prior knowledge is called source domain, and the domain containing unknown knowledge is called target domain [39]. The aim of transfer learning is to learn the target task with the help of the knowledge of source task, such as features, parameters, and labels.

Transfer learning is effective when there is connection between the source domain and the target domain. So far, the most studied scenario in transfer learning is to reduce the difference between source data and the target data with the same tasks [40]. In this case, transfer learning maintains the reusability of the model by reducing the distribution differences between data sets. With the deep research of transfer learning, a few studies start to work on the scenario that the tasks of source and target domain are different with the same data set [41]. In this paper, the data collected in single working condition and complex working conditions are shown as source data and target data, respectively. Transfer learning reserves the reusability of the model trained by data of single working condition through measuring and reducing the distribution difference between the source data and target data.

2.3. Transfer Learning-Based Sparse Autoencoder. Sparse autoencoder is developed from autoencoder, which is an unsupervised learning network with an encoder and a decoder. As shown in Figure 1, the encoder reduces the dimension of the input data for feature extraction purpose, and the decoder recombines the encoded information and restores the encoded information to the original data [42, 43]. SAE improves the ability of feature extraction by adding a sparsity limitation to the neurons in the hidden layers. In this paper, a three-layer SAE is applied as the network model, and the structure of SAE is introduced.

After the preprocessing, a m × n data set can be represented as $X = \{x_{ij}\}$, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. m is the number of samples, and n is the dimension of each sample. The source features obtained by the encoder are denoted by ξ, and the output of the decoder $\widehat{X}$ is close to X. The parameter set is represented by $\theta = \{W^e, B^e, W^d, B^d\}$, where W^e and W^d are weights of encoder and decoder, respectively, and B^e and B^d are bias of encoder and decoder, respectively. Based on the above introduction, the value of the features ξ and the output of the decoder $\widehat{X}$ are shown in the following equations:

$$\xi = \sigma\left(W^e X + B^e\right), \tag{4}$$

$$\widehat{X} = \sigma\left(W^d \xi + B^d\right), \tag{5}$$

where σ is the activation function sigmoid, whose formula is shown in

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{6}$$

To restrict the number of active nodes in hidden layer, sparsity penalty factor Kullback–Leibler (KL) divergence is measured. The average activation value of k-th node in hidden layer, ρ_k, is calculated in

$$\rho_k = \frac{1}{n} \sum_{i=1}^{n} \left[\sigma(x_i)\left(w_{ik}^e x_i + b_k^e\right)\right], \tag{7}$$

where w_{ik}^e and b_k^e belong to W^e and B^e. By using relative entropy and (7), KL is represented in

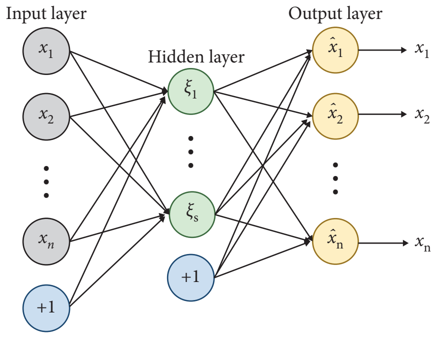


Figure 1: The structure of autoencoder.

$$KL\rho \| \rho_k = \rho \log \frac{\rho}{\rho_k} + (1 - \rho) \log \frac{1 - \rho}{1 - \rho_k}, \tag{8}$$

where ρ is a predefined sparse parameter. To achieve the sparsity of the active values in hidden layer, the value of sparsity parameter ρ should be close to 0. For this purpose, it is necessary to adjust the value of ρ_{k} until $\rho = \rho_{\mathrm{k}}$ to ensure that KL reaches its minimum value, which is close to 0. Therefore, the cost function of SAE $J_{\mathrm{SAE}}(\theta)$ can be expressed as

$$J_{\mathrm{SAE}}(\theta) = \sum_{i=1}^{n} L(x_i, \widehat{x}_i) + \frac{\alpha}{2} \sum_{i=1}^{n} \left(\left(W^e\right)^2 + \left(W^d\right)^2\right) + \beta KL\left(\rho \| \rho_k\right), \tag{9}$$

where $\widehat{x}_i$ is the output, L is the loss function of SAE, W^e and W^d are the weight of encoder and decoder, and α and β are the weight parameters. Through minimizing $J_{\mathrm{SAE}}(\theta)$, the features of data obtained offline such as single working condition can be extracted as a priori knowledge for transfer learning.

2.4. Maximum Mean Discrepancy. Maximum mean discrepancy is a distance for measuring the difference of probability distribution between two data sets X and Y, which is widely applied in transfer learning [44]. When the probability distributions of X and Y are different, it is not appropriate to apply the same classification model to achieve satisfactory performance [45]. In the problem discussed in this paper, large probability distribution difference leads to the fact that the model trained with data obtained under single working condition is not applicable to complex working conditions. The accuracy of the model can be improved by minimizing MMD with a transformation function to minimize the distance between the transformed feature sets which are obtained in sensitive feature selection. Suppose that the probability distribution of data sets X and Y is p and q, respectively; the expression of MMD is as follows:

$$D_H(X, Y) = \sup_{\Phi \in H} \left(E_{X \sim p}[\Phi(x)] - E_{Y \sim q}[\Phi(y)]\right), \tag{10}$$

where H represents Reproducing Kernel Hilbert Space (RKHS). RKHS is a complete inner product space which can transfer the data set that is not linearly separable to high dimensional space via mapping [46]. Equation (10) represents the upper bound of the mapping of probability distribution between two data sets in RKHS. For the convenience of computation, the square of MMD $\widehat{D}_H^2(X,Y)$ is applied, whose formula is listed in

$$\widehat{D}_H^2(X,Y) = \left\| \frac{1}{N_X}\sum_{i=1}^{N_X}\Phi(x_i) - \frac{1}{N_Y}\sum_{i=1}^{N_Y}\Phi(y_i)\right\|_H^2, \tag{11}$$

where N_X and N_Y are the sample numbers of X and Y. The smaller the value of MMD is, the smaller the probability distribution discrepancy between the two data sets is.

2.5. The Proposed Algorithm. In this subsection, an improved algorithm is proposed to transfer the knowledge in single working condition to make it available in complex working conditions, whose architecture is plotted in Figure 2. In data collection, the data are separated into two parts: data in single working condition (called source data) and data in complex working conditions (called target data) which consist of multiple and unknown working conditions. Before feature extraction, sensitive features are selected through cohesion evaluation to constitute a sensitivity parameter set as input data. In the network, after the training of source parameter set is completed, the parameters of the SAE are reused for learning the target labels. To apply the knowledge learnt from source data to the target task through transfer learning, the distance of the source features and the target features is minimized by MMD. The trained target features are classified by a softmax classifier to obtain the target labels.

The expected effect of the proposed algorithm is explained via illustration in Figure 3. In this figure, it is assumed that source data and target data have large probability distribution difference and there are three fault types in both sets: feature 1 and feature 2 are sensitive to fault 1 and fault 2, respectively, while feature 3 is insensitive to fault 1 or fault 2. The red patterns represent the incorrectly classified samples. The classification result without sensitive feature selection and transfer learning is shown in Figure 3(a): part of sensitive features is classified incorrectly because of large probability distribution; feature 3 is kept and dispersed into two fault types without sensitive features selection, which interferes with the accurate description of the faults. Figure 3(b) shows the classification result after insensitive feature 3 is discarded, but large probability distribution difference results in inaccurate classification of feature 1 and feature 2. In Figure 3(c), after minimizing the probability distribution difference between the two data sets, feature 1 and feature 2 are classified correctly, while feature 3 is reserved as useless information. Although the methods in Figures 3(b) and 3(c) make improvements to some extent, they fail to solve all the problems shown in Figure 3(a). The effectiveness of the proposed algorithm is shown in Figure 3(d): insensitive feature 3 is discarded by sensitive feature selection and feature 1 and feature 2 are correctly classified after transfer learning. The drawbacks of the methods in Figures 3(b) and 3(c) motivate this proposed method.

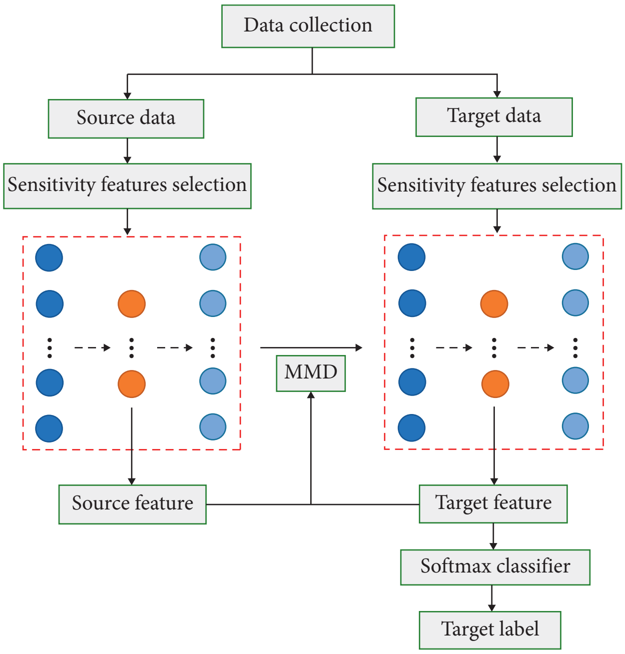


Figure 2: The architecture of the proposed network.

The fault diagnosis process of the proposed algorithm is divided into 7 steps, as shown in Table 2: sensitive feature selection from step 1 to step 3, network training of source data in step 4, and network adaptation by transfer learning from step 5 to step 7. The detailed information of each part is explained as follows:

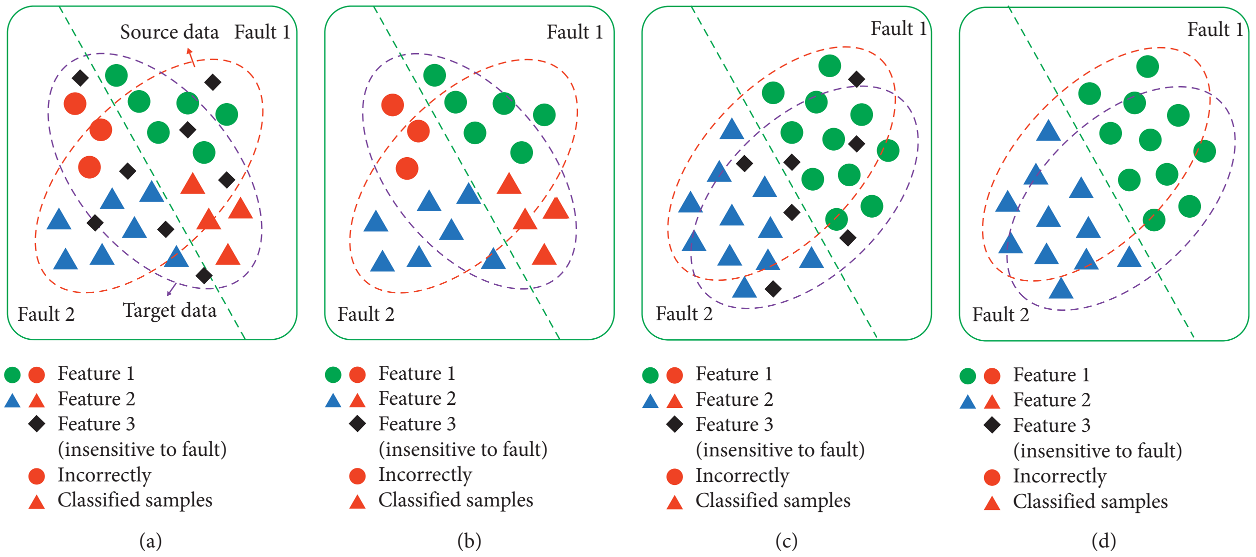
(1) Sensitive feature selection: first, training data set D^s and testing data set D^c are collected under single and complex working conditions of rotating machinery; second, features are computed and sorted by cohesion evaluation shown in Table 1; third, sensitive features are chosen according to sensitive factor η_j in (2), which are reserved to constitute sensitivity parameter set as input data.

(2) Network training of source data: the total cost function of the proposed algorithm $J(\theta)$ that is made up with $J_{\text{SAE}}(\theta)$ and MMD is shown in (12) as

$$J(\theta) = \sum_{i=1}^{n} L(x_i,\hat{x}_i) + \frac{\alpha}{2}\sum_{i=1}^{n}\left(\left(W^e\right)^2 + \left(W^d\right)^2\right) + \beta KL\left(\rho\|\rho_k\right) + \tau\widehat{D}_H^2\left(D^{tr},D^{ts}\right), \tag{12}$$

where α, β, and τ are weighted parameters, the first three terms are the cost function of SAE, and the last term is the square of MMD. In the training of D^s only, MMD coefficient parameter τ is set to 0, and the network is trained to gain the updated model parameter set θ and data features.

(3) Network adaptation by transfer learning: the performance of the model obtained is validated by



(a) (b) (c) (d)

Figure 3: Fault diagnosis: (a) without sensitive features selection and transfer learning, (b) with sensitive features selection, without transfer learning, (c) without sensitive features selection, with transfer learning, and (d) with sensitive features selection and transfer learning.

Table 2: Fault diagnosis process of the proposed algorithm.

Step	Description
1	Collect the signal under single working condition as training data set D^s; collect the signal under complex working conditions as testing data set D^c.
2	Extract feature data set of time domain and frequency domain in Table 1 as preparation.
3	Calculate sensitive factor η_j in (2) to keep these features in which the value of η_j is large. These parameters constitute sensitivity parameter set as input data.
4	Let $\tau = 0$ in (12); train network to gain suitable parameter set θ and the source features.
5	Assign τ suitable values in (12) to validate the network by target data set D^c until minimizing the cost function in (12) by comparing the distance between the target features and source features, using θ from step 4 as initial parameters.
6	After step 5 is done, record the parameters and features of testing.
7	Send the features into classifier to gain the fault types.

training D^c. In the adaptation stage, to minimize the probability distribution difference between D^s and D^c, the value of MMD is minimized by giving coefficient parameter τ different values. The model parameters and features under complex working conditions are obtained to classify the fault types. In this paper, softmax classifier is chosen to solve this multiclass task, which maps the features obtained above to another vector of $(0, 1)$, and the probability closest to 1 is selected to estimate the output of classifier.

3. Experiment and Result Analysis

3.1. Experiment Setup. In this paper, wind turbine drivetrain diagnostics simulator (WTDS) is used to collect the experiment data to verify the effectiveness of the proposed scheme. WTDS consists of a planetary gearbox, a fixed axis gearbox, a magnetic brake, a motor, and 4 sensors as illustrated in Figure 4. Figure 5 shows the working diagram of WTDS, in which the speed and load of WTDS are controlled by computer to change the working conditions through speed and brake controllers. Data are collected to the computer by four sensors, including two vibration sensors, a pressure sensor, and a torque sensor, indicated in Figure 5.

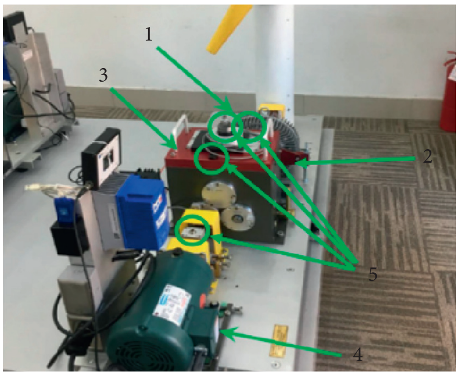


Figure 4: Appearance of WTDS: 1-magnetic brake, 2-planetary gearbox, 3-fixed axis gearbox, 4-motor, and 5-sensors.

In this research, experiments are conducted in WTDS with five different types of gears, respectively: normal gear,

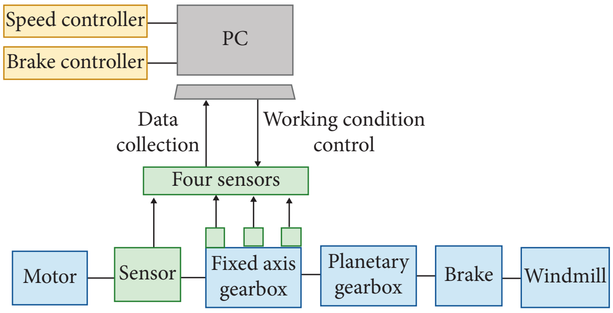


Figure 5: Illustration of WTDS.

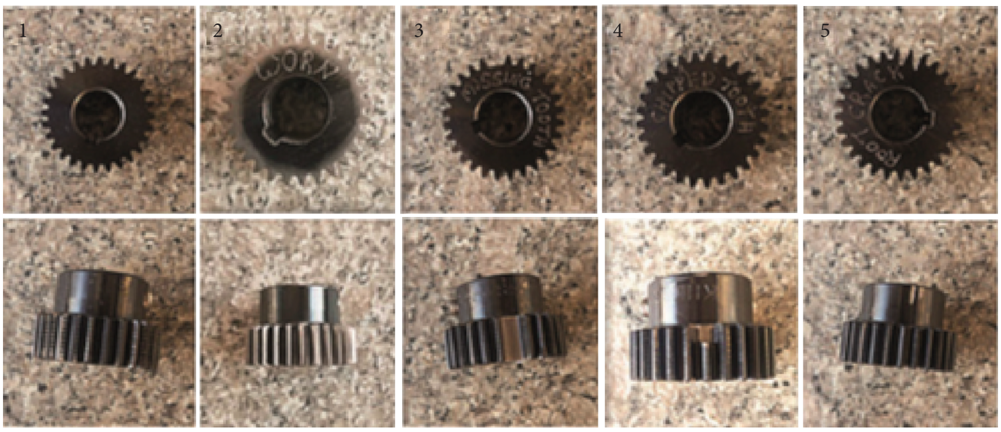


Figure 6: Five types of gears.

Table 3: 9 Working conditions.

Motor frequency (Hz)	Load voltage
6	3 V
6	5 V
6	8 V
10	3 V
10	5 V
10	8 V
14	3 V
14	5 V
14	8 V

surface worn, missing tooth, chipped tooth, and root crack as shown in Figure 6. 9 working conditions of WTDS under different load voltages and rotating speeds are considered as shown in Table 3.

3.2. Data Collection. Signals from four sensors are collected with sampling frequency of 5120 Hz and sampling time of 6.4 s. To ensure that the data are more effective, the original signals are the average value of four sensors in 3 experiments. After sensitive feature selection, both numbers of the training data and the testing data in every group consist of 80 samples. Figure 7 illustrates the vibration signals in time domain and frequency domain, respectively.

The results under complex working conditions, consisting of multiple and unknown working conditions, are analyzed separately in this research. 9 data sets groups are set in Table 4; each contains a training set and a testing set. The details of multiple and unknown working conditions are explained: (1) multiple working conditions from group 1 to group 6: to simulate the collection of multiple working conditions data, the testing data are composed of five pieces, which are from the five working conditions except the condition in training data. The data with load voltages of 5 V and 8 V are selected as 6 training sets, and data under other five working conditions shown in Table 3 (excluding working conditions in training sets) are randomly mixed as testing sets under multiple working conditions labeled from multi-A to multi-F. For example, in group 1, the testing set multi-A is the mixture of the data under 6 Hz–8 V, 10 Hz–5 V, 10 Hz–8 V, 14 Hz–5 V, and 14 Hz–8 V without 6 Hz–5 V, because 6 Hz–5 V is the condition of the training set of group 1. (2) Unknown working conditions from group 7 to group 9: data with load voltage 8 V and 3 V are selected as training sets and testing sets, respectively, labeled from single-G to single-I. In the three groups, the working conditions in testing sets are totally different from those in training sets to ensure the unknown of the testing sets. To observe the performance of transfer learning in unknown working conditions, the two voltages with a large difference are selected.

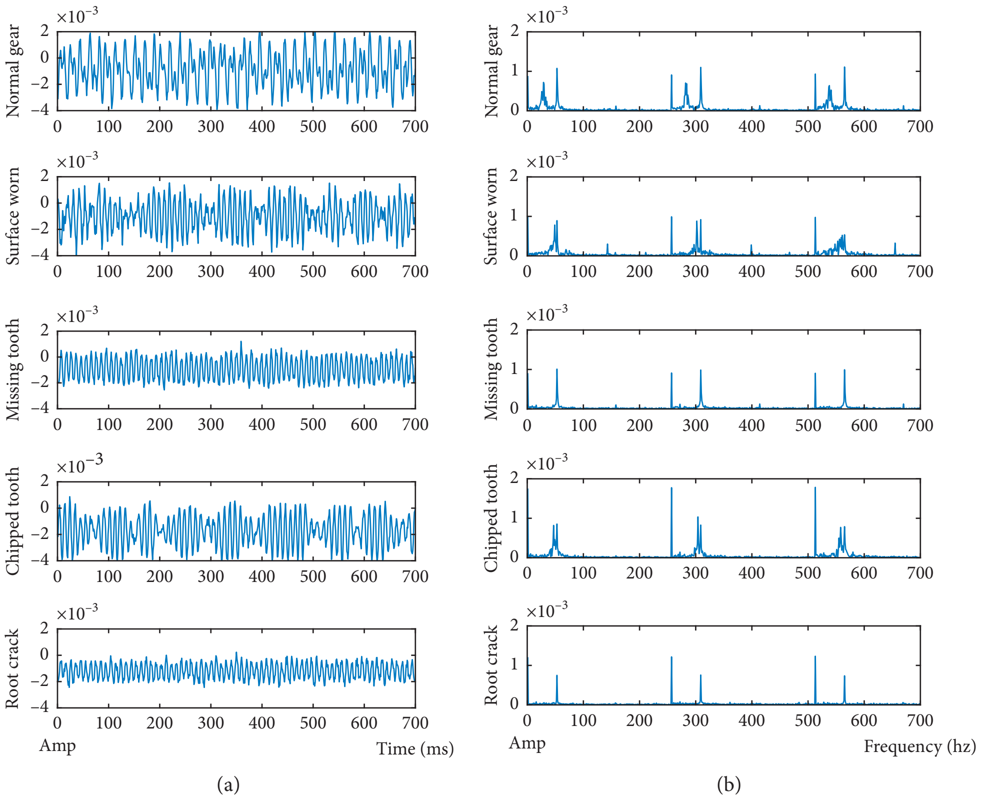


Figure 7: Vibration signals. (a) Vibration signal in time domain. (b) Vibration signal in frequency domain.

Table 4: Data sets groups.

Group	Training set	Testing set
1	6 Hz–5 V	Multi-A
2	6 Hz–8 V	Multi-B
3	10 Hz–5 V	Multi-C
4	10 Hz–8 V	Multi-D
5	14 Hz–5 V	Multi-E
6	14 Hz–8 V	Multi-F
7	14 Hz–8 V	Single-G (6 Hz–3 V)
8	6 Hz–8 V	Single-H (10 Hz–3 V)
9	10 Hz–8 V	Single-I (14 Hz–3 V)

3.3. Experimental Results. Before experimental results analysis, the architecture of SAE is shown: the value of η_j is set as 80 to obtain the sensitive feature set. Accordingly, the number of input layers and output layers of SAE is 80, and the number of hidden layers is 60 with sparsity limitation of 0.3. To observe the effects of KL divergence and transfer learning in the cost function in (12), the values of β and τ are varied to adjust KL and MMD terms and the value of α is predefined as a constant ($\alpha = 0.001$). The range of KL and MMD weight parameter τ is set by experimentation, which affects the results positively: β is searched from $\{1, 2, 3, 4, 5\}$ and τ is searched from $\{0, 1, 5, 10, 15, 20\}$ to test the performance of the proposed algorithm. Particularly, $\tau = 0$ implies no domain adaptation in cost function of (12). The rest of this subsection discusses the performance of transfer learning, the classification results under multiple working conditions and unknown working conditions, and the performance of the proposed algorithm comparing with other feature extraction methods.

3.3.1. Performance of Transfer Learning. To observe the influence by transfer learning, the parameter β is fixed at 3 and only parameter τ is changed from 0 to 20. After testing the model with different τ values, the classification accuracies of the six data set groups are shown in Figure 8 and the MMD variation curve is shown in Figure 9.

From Figures 8 and 9, it is apparent that the classification accuracies of all the nine data sets can be improved with MMD term and the corresponding MMD values can be reduced to be around 0.1. When $\tau = 0$, the network trained under single working condition has poor adaptability to complex working conditions, with the classification accuracies of nine data sets fluctuating around 85% which are represented in the blue line in Figure 8. With transfer learning by using MMD, all the classification accuracies of nine data sets are improved significantly with values fluctuating around 96%. This result indicates that transfer learning has positive effect on both multiple and unknown working conditions.

3.3.2. Classification Results. The classification results of multiple and unknown working conditions are illustrated in this part, which are listed in Table 4, which are shown in

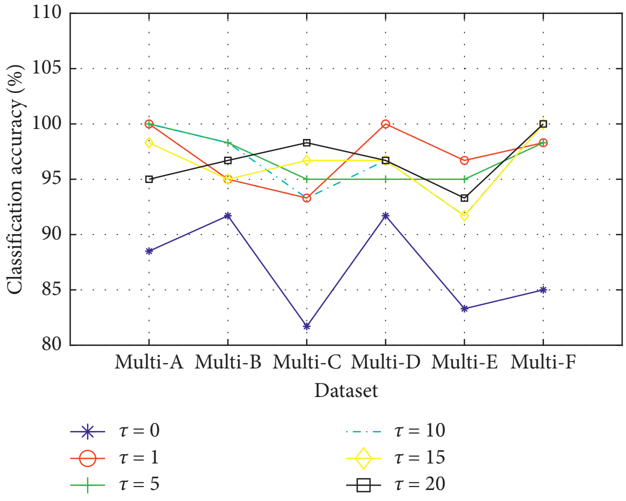


Figure 8: Effect of τ on classification accuracy.

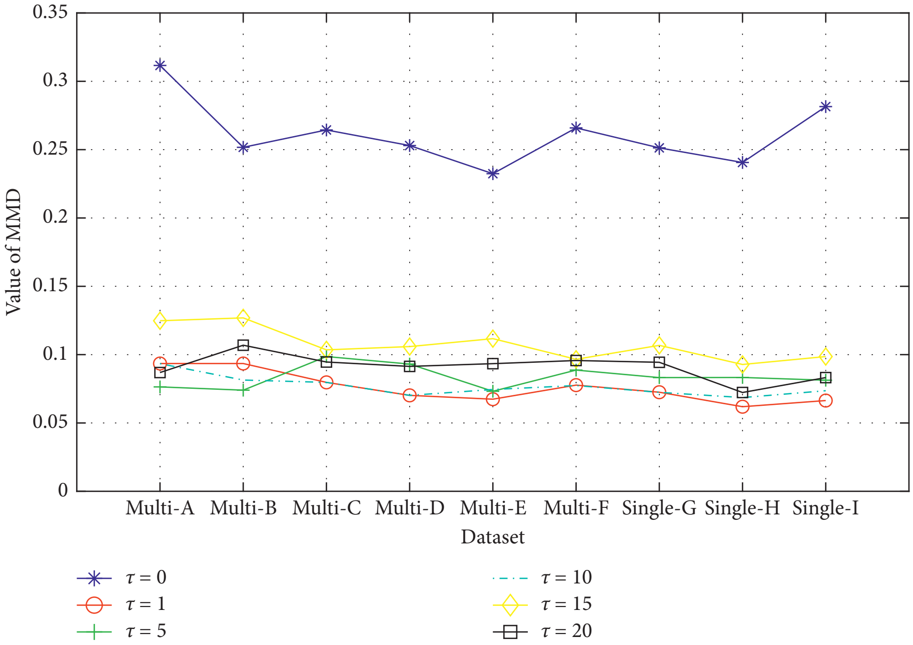


Figure 9: MMD variation curve.

Tables 5 and 6. The combinations of two parameters β and τ exert an intricate effect on the result.

To observe the effect of the proposed method on experimental results when KL divergence changes, the average results of multiple working conditions are displayed in Table 5 with different combinations of β and τ of group 1 to group 6. It can be observed that the highest classification accuracy can reach 99.17% (when $\beta = 2$ and $\tau = 20$) and the classification accuracies are all below 90% without domain adaptation (when $\tau = 0$). From the analysis above, the proposed method suggests a significant improvement under multiple working conditions after reducing the probability distribution difference between training and testing sets.

The average results of group 7 to group 9 are shown in Table 6, which represent the unknown working conditions. In the first row of Table 6, although the classification accuracy reaches 91.70% when $\beta = 5$ without domain adaptation, the results fluctuate wildly with the lowest being 83.33%. After domain adaptation, all the classification accuracies are higher than 95% with the highest classification accuracy reaching 100% when $\beta = 2$ and $\tau = 15$. The above analysis indicates that after reducing the probability distribution difference between unknown data set and training data set, the trained network is adaptable to the unknown working conditions.

3.3.3. *Comparison with Different Feature Diagnosis Methods.*

Data-driven fault diagnosis methods contain three main directions: artificial intelligence-based methods, statistical analysis, and signal processing, statistical analysis. To

Table 5: Classification accuracy for parameters combinations in multiple working conditions.

Accuracy	$\beta = 1$ (%)	$\beta = 2$ (%)	$\beta = 3$ (%)	$\beta = 4$ (%)	$\beta = 5$ (%)
$\tau = 0$	86.67	87.23	86.95	87.50	87.52
$\tau = 1$	98.05	97.22	97.22	96.95	97.23
$\tau = 5$	96.10	97.78	96.67	97.32	96.93
$\tau = 10$	96.40	95.53	96.40	96.67	97.23
$\tau = 15$	95.82	96.40	96.67	97.50	95.82
$\tau = 20$	97.23	99.17	96.93	97.77	97.52

Table 6: Classification accuracy for parameters combinations in unknown working conditions.

Accuracy	$\beta = 1$ (%)	$\beta = 2$ (%)	$\beta = 3$ (%)	$\beta = 4$ (%)	$\beta = 5$ (%)
$\tau = 0$	87.77	83.33	87.77	89.43	91.70
$\tau = 1$	97.23	97.80	96.70	97.80	98.90
$\tau = 5$	97.23	97.80	97.23	97.77	98.87
$\tau = 10$	97.77	98.90	97.80	98.33	97.23
$\tau = 15$	97.80	100.00	96.13	98.33	97.53
$\tau = 20$	98.33	96.70	95.57	97.53	97.80

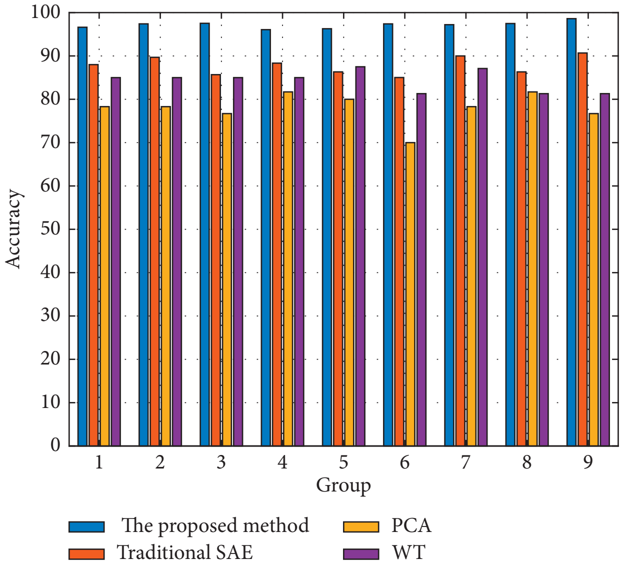
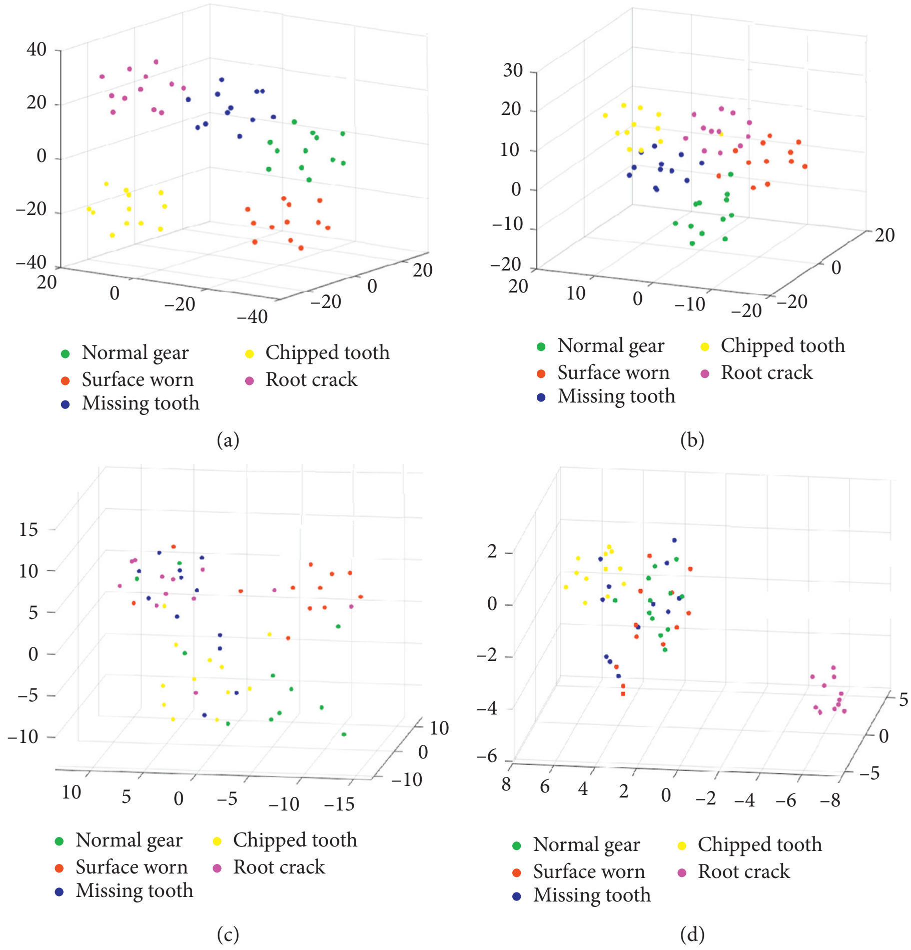


Figure 10: Diagnosis results of the 9 groups with 4 methods.

explore the performance of the proposed method, it is compared with the following three data-driven methods: traditional SAE, principal component analysis (PCA), and wavelet transform (WT), which are representative of the three directions mentioned above. Figure 10 shows the classification results of the nine experimental groups and the visualization results are shown in Figure 11.

As shown in Figure 10, the average accuracies of PCA float around 80%, the lowest of which even reaches 70% in group 6. Both SAE and WT perform better, most fluctuating between 80% and 90%, with the highest of traditional SAE over 90% in group 9. In contrast, the results of the proposed method are all over 95%, which is overall precise to other methods in this figure.

Figure 11 displays the distribution of classification result for five types of gears of data set multi-D (when $\beta = 3$) of the four methods. Figure 11(a) shows the classification result of the proposed method, where all the five types of gears are classified and clustered with clear boundaries and no overlap. Figure 11(b) shows the classification result of the traditional SAE, where the boundaries of the four faulty gear types are unclear. The result of PCA is shown in Figure 11(c), where the surface worn gears are classified, but the other four types are not separated. In Figure 11(d), the result of WT shows that the root crack gears are classified well and the small part of missing tooth is wrongly classified to chipped tooth. Similar to PCA, the other three types are not separated. From Figures 10 and 11, it is clear that the proposed method performs better than the other three methods in both multiple and unknown working conditions.



(a)

(b)

(c)

(d)

Figure 11: Classification results visualization. (a) Classification results of the proposed method. (b) Classification result of traditional SAE. (c) Classification results of PCA. (d) Classification results of WT.

4. Conclusion

In this paper, to investigate the fault diagnosis problem under complex working conditions, a fault diagnosis method for gearbox based on transfer learning is introduced. The proposed method selects sensitive features to decrease the adverse impact of insensitive information and transfers the knowledge learnt under single working condition to complex working conditions through transfer learning. To verify the performance of the model in complex working conditions, experiments are carried out on wind turbine drivetrain diagnostics simulator, which simulates five fault types of gear. Results are compared with traditional SAE, PCA, and WT, which indicate that the classification accuracy is significantly improved after sensitive feature selection and transfer learning. The future work of current research can be extended to other working conditions and data sets.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest in the work reported in this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61603223), Key Program Special Fund in XJTLU (KSF-E-34), and Research Development Fund of XJTLU (RDF-18-02-30).

References

[1] C. Wen, F. Lv, Z. Bao, and M. Liu, “A review of data driven-based incipient fault diagnosis,” *ACTA Automatica SINICA*, vol. 42, no. 9, pp. 1285–1299, 2016.

[2] W. Du, J. Zhou, Z. Wang, R. Li, and J. Wang, “Application of improved singular spectrum decomposition method for composite fault diagnosis of gear boxes,” *Sensors*, vol. 18, no. 11, p. 3804, 2018.

[3] X. Jiang, C. Shen, J. Shi, and Z. Zhu, "Initial center frequency-guided VMD for fault diagnosis of rotating machines," *Journal of Sound and Vibration*, vol. 435, pp. 36–55, 2018.

[4] J. Li, D. Zhou, X. Si et al., "Review of incipient fault diagnosis methods," *Control Theory & Applications*, vol. 29, no. 12, pp. 1517–1529, 2012.

[5] M. Kordestani, M. F. Samadi, M. Saif, and K. Khorasani, "A new fault diagnosis of multifunctional spoiler system using integrated artificial neural network and discrete wavelet transform methods," *IEEE Sensors Journal*, vol. 18, no. 12, pp. 4990–5001, 2018.

[6] S. Krishnannair, C. Aldrich, and G. T. Jemwa, "Detecting faults in process systems with singular spectrum analysis," *Chemical Engineering Research and Design*, vol. 113, pp. 151–168, 2016.

[7] R. Sharifi and R. Langari, "Nonlinear sensor fault diagnosis using mixture of probabilistic PCA models," *Mechanical Systems and Signal Processing*, vol. 85, no. 15, pp. 638–650, 2017.

[8] R. Ghimire, C. Zhang, and K. R. Pattipati, "A rough set-theory-based fault-diagnosis method for an electric power-steering system," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 5, pp. 2042–2053, 2018.

[9] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, no. 93, pp. 490–502, 2016.

[10] K. H. Hui, M. H. Lim, M. S. Leong, and S. M. Al-Obaidi, "Dempster-shafer evidence theory for multi-bearing faults diagnosis," *Engineering Applications of Artificial Intelligence*, vol. 57, pp. 160–170, 2017.

[11] R. M. Monaro, J. C. M. Vieira, D. V. Coury, and O. P. Malik, "A novel method based on fuzzy logic and data mining for synchronous generator digital protection," *IEEE Transactions on Power Delivery*, vol. 30, no. 3, pp. 1487–1495, 2015.

[12] T. Bruin, K. Verbert, and R. Babuska, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 523–533, 2017.

[13] S. S. Udmale, S. K. Singh, R. Singh, and A. K. Sangaiah, "Multi-fault bearing classification using sensors and convnet-based transfer learning approach," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1433–1444, 2020.

[14] Y. Bengio, P. Lamblin, and D. Popovici, "Greedy layer-wise training of deep networks," in *Proceeding of the 19th International Conference on Neural Information Processing Systems*, Doha, Qatar, 2006.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[16] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2019.

[17] X. Li, W. Zhang, and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, no. 8, pp. 77–95, 2018.

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

[19] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2019.

[20] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: a new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2019.

[21] D. Huang, L. Ke, B. Mi et al., "A new incipient fault diagnosis method combining improved RLS and LMD algorithm for rolling bearings with strong background noise," *IEEE Access*, vol. 6, pp. 26 001–026 010, 2018.

[22] H. Azami, M. Rostaghi, D. Abasolo, and J. Escudero, "Refined composite multiscale dispersion entropy and its application to biomedical signals," *IEEE Transactions on Bio-Medical Engineering*, vol. 64, no. 12, pp. 2872–2879, 2017.

[23] X. Liu, Q. Zhou, J. Zhao, H. Shen, and X. Xiong, "Fault diagnosis of rotating machinery under noisy environment conditions based on a 1-D convolutional autoencoder and 1-D convolutional neural network," *Sensors*, vol. 19, no. 4, pp. 972–991, 2019.

[24] Y.-r. Wang, Q. Jin, G.-D. Sun, and C.-F. Sun, "Planetary gearbox fault feature learning using conditional variational neural networks under noise environment," *Knowledge-Based Systems*, vol. 163, pp. 438–449, 2019.

[25] X. Li, J. Wang, and H. Wang, "Sparsity-oriented nonconvex nonseparable regularization for rolling bearing compound fault under noisy environment," *Shock and Vibration*, vol. 2020, Article ID 8823102, 19 pages, 2020.

[26] Y. Zhang, Z. Ren, and S. Zhou, "A new deep convolutional domain adaptation network for bearing fault diagnosis under different working conditions," *Shock and Vibration*, vol. 2020, Article ID 8850976, 14 pages, 2020.

[27] M. J. Hasan, M. M. M. Islam, and J.-M. Kim, "Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions," *Measurement*, vol. 138, pp. 620–631, 2019.

[28] Q. Wang, G. Michau, and O. Fink, "Domain adaptive transfer learning for fault diagnosis," *IEEE Access*, vol. 7, pp. 47 663–47 673, 2019.

[29] M. J. Hasan and J.-M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning," *Applied Sciences*, vol. 8, no. 12, pp. 2357–2371, 2018.

[30] 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. 69 907–69 917, 2018.

[31] S. Thrun and L. Pratt, *Learning to Learn: Introduction and Overview*, pp. 3–17, Springer, Berlin, Germany, 1998.

[32] Y. Peng and B.-L. Lu, "Discriminative manifold extreme learning machine and applications to image and EEG signal classification," *Neurocomputing*, vol. 174, pp. 265–277, 2016.

[33] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2416–2425, 2019.

[34] M. Johnson, M. Schuster, and Q. V. Le, "Google's multilingual neural machine translation system: enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[35] Y. Chen, J. Wang, M. Huang, and H. Yu, "Cross-position activity recognition with stratified transfer learning," *Pervasive and Mobile Computing*, vol. 57, pp. 1–13, 2019.

[36] Y. Lei, Z. He, and Y. Zi, "Fault diagnosis based on novel hybrid intelligent model," *Chinese Journal of Mechanical Engineering*, vol. 44, no. 7, pp. 112–117, 2008.

[37] C. Engelhardt, M. Baker, A. Mouron, and H. Vold, "Separation of sine and random components from vibration measurements," in *Topics in Modal Analysis II, Conference Proceedings of the Society for Experimental Mechanics Series*, pp. 339–350, Springer, Berlin, Germany, 2012.

[38] Q. Lu, R. Yang, M. Zhong, and Y. Wang, "An improved fault diagnosis method of rotating machinery using sensitive features and RLS-BP neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1585–1593, 2020.

[39] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, 2020 (in press).

[40] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proceedings of the International Conference on Artificial Neural Networks*, pp. 270–279, Munich, Germany, September 2019.

[41] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[42] M. Xia, T. Li, L. Liu, L. Xu, and C. W. de Silva, "Intelligent fault diagnosis approach with unsupervised feature learning by stacked denoising autoencoder," *IET Science, Measurement & Technology*, vol. 11, no. 6, pp. 687–695, 2017.

[43] C. Shen, Y. Qi, J. Wang, G. Cai, and Z. Zhu, "An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 170–184, 2018.

[44] X. Jia, M. Zhao, Y. Di, Q. Yang, and J. Lee, "Assessment of data suitability for machine prognosis using maximum mean discrepancy," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5872–5881, 2018.

[45] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 3, pp. 2296–2305, March 2017.

[46] O. Taouali, I. Elaissi, and H. Messaoud, "Dimensionality reduction of RKHS model parameters," *ISA Transactions*, vol. 57, pp. 205–210, 2015.