IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2020, accepted May 11, 2020, date of current version June 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2996713

A Novel Data-Driven Fault Feature Separation Method and Its Application on Intelligent Fault Diagnosis Under Variable Working Conditions

SHUNMING LI, ZENGHUI AN, AND JIANTAO LU

College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Corresponding author: Shunming Li (smli@nuaa.edu.cn)

This work was supported in part by the Major National Science and Technology under Project 2017-IV-0008-0045, in part by the National Natural Science Foundation of China under Grant 51675262 and Grant 51975276, in part by the Advance Research Field Fund Project of China under Grant 61400040304, in part by the Key Laboratory of Ministry of Industry and Information Technology under Grant KL2019N001, and in part by the AECC Hunan Aviation Powerplant Research Institute under Grant KF7-QT-8025.

ABSTRACT As mechanical fault diagnosis enters the era of big data, the traditional fault diagnosis methods under variable working condition are difficult to be applied because of the massive computation cost and excessive reliance on human labor. For the application of intelligent fault diagnosis under variable working conditions, the crucial difficulty is that the variable speed or load can cause smearing and skewing of classable feature. It is the key to break the predicaments by extracting the features which are irrelevant to the working conditions and contain fault information. This paper propose a new intelligent fault diagnosis framework under variable working conditions called **D**ata-driven **F**ault feature **S**eparation **M**ethod (DFSM) which can eliminate the working condition features from all the information and employ the rest fault information for diagnosis. In our DFSM, classification loss ensures the basic classified ability, first. Second, uncorrelation loss increases the discrepancy between fault features and working condition features. Then, reference loss guides the working condition encoder only extracting the working condition features. Finally, autoencoder loss ensures that all the information has been extracted. It should be noticed that our DFSM is trained only using the dataset under a certain working condition and can diagnose faults with high accuracy under variable even time-varying working condition, which means that it is easy to be applied. The experimental results of rolling bearing dataset show that the proposed DFSM can not only break the limitation of existing methods, but also achieve a superior performance comparing with related method. Besides, through the visual understanding, the proposed DFSM is certified that it is able to eliminate the working condition information and extract precise classable feature for fault diagnosis.

INDEX TERMS Data-driven, feature separation, intelligent fault diagnosis, variable working conditions.

I. INTRODUCTION

Rolling bearing is one of the most important mechanical components and is widely applied in the rotary machineries of aircrafts, wind turbines, automobiles, and so forth [1], [2]. However, bearings are vulnerable because of the harsh working environment, such as humidity, high temperature, and variable load [3], which can lead to a catastrophic failure of the entire mechanical system, and consequently heavy investment and productivity losses [4], [5]. Therefore, the fault diagnosis of bearing is very important for ensuring a high performance transmission. However, in general, rotary machinery often runs at variable speed because of different working conditions or time-varying working condition like power and load, which causes difficulty on fault diagnosis [6]. When the working condition is variable, such fluctuation and changing condition may cause smearing and skewing of classable feature like spectrum, which indicates that these fault features will no longer be observable and detected.

For the time-varying working condition, plenty of time-frequency analyzing methods have employed in existent report. For instance, Chen and Feng [7] improved polynomial chirplet transform (PCT) by iterative algorithm, and achieved merits of fine time-frequency resolution and cross term free

The associate editor coordinating the review of this manuscript and approving it for publication was Amir Masoud Rahmani.

nature, and validated the method by lab experiments on a real world 4 kW induction motor driven planetary gearbox test rig. Wang *et al.* [8] used the reference signal from a current signal measured from the stator of the generator for vibration order tracking. Huang *et al.* [9] verified that nuisance attribute projection (NAP) can effectively eliminate the effect of nuisance attributes through projection and the information of fault pattern is retained in the features. Jiang *et al.* [10] proposed time-frequency ridge fusion and logarithm transformation to track the targeted ridge curve reliably, and improved the diagnosis accuracy. It can be found that plenty of studies have been conducted on fault diagnosis under time-varying working condition. But the amount of data collected has grown in an exponential manner with the development of modern machinery fault detection systems [11], [12]. This means that fault diagnosis enters the era of big data [13], which makes these studies suffer three deficiencies: 1) Lots of the actual effort goes into the design of time-frequency feature extracting. So the existing methods largely depend on much prior knowledge about signal processing techniques. 2) Even though the correct and clear time-frequency features are obtained, it is still difficult to diagnose the fault, since the depending on diagnostic expertise. Besides, the signal is corrupted by heavy background noise, which affects the diagnosis result. 3) Most existing methods require massive computation cost, which reduces efficiency of fault diagnosis. Therefore, it is a tough work to process big data.

Machine learning which is regarded as a data-driven model, depends less human knowledge and is quite suitable to deal with mechanical big data, has drawn wide attention in fault diagnosis field, such as Artificial Neural Networks (ANN) [14], [15], Autoencoders (AE) [16], [17], Restricted Boltzmann Machine (RBM) [18], Convolutional Neural Networks (CNN) [19], [20], Sparse Filtering [21], [22] and k-Nearest Neighbor [23]. For the fault diagnosis under different working conditions, deep transfer learning, as a branch of deep learning, has been employed. The application can be described as follows: take the roller bearing fault diagnosis problem as an example, model is trained with the labeled dataset under working condition A and unlabeled dataset under working condition B, and then the actual application in fault diagnosis is to recognize the test data collected under working condition B. For example, Lu *et al.* [24] presented a deep model based on domain adaptation method for machine fault diagnosis. A gearbox dataset collected under different working conditions was used to test the performance of the proposed method. Wen *et al.* [25] set up a new deep transfer learning method for fault diagnosis. The validation dataset was acquired from a bearing testbed operating under different loading conditions. They both get the satisfying testing accuracies under their experiment conditions. But according to the above description, necessary condition for the transfer learning based methods is the dataset of different fault samples under working condition B. However, mechanical fault datasets, similar to medical datasets [26], genomics [27] and financial datasets [28], are also very limited since the vast majority of samples are normal samples. When the machine is employed under another working condition, the collected samples must be imbalanced [29], [30] and generally lack plenty of categories. This condition will lead to the change of distribution and further result in the negative transfer, which troubles the existing methods.

As we can see from the two types of methods, the fundamental goal is consistent. The key is to extract the features which are irrelevant to the working conditions and contain fault information. For fault diagnosis under time-varying working condition, the existing methods focus on extracting the time-invariant feature like angular domain feature or NAP. For intelligent fault diagnosis methods under different working conditions, the goal is to extract the cross-domain feature which is consistent in the source domain and target domain. The fundamental motivation of all the mentioned methods is that fault information is irrelevant to the working conditions [31] and the present methods all focus on extracting fault feature. But, in this paper, we focus more on eliminating the working condition information instead of extracting fault feature. In practical application, we have the labeled dataset under a certain working condition in general. According to the theory, samples of the dataset must contain the fault information, and the fault information is irrelevant to the working conditions. Therefore, intelligent fault diagnosis method under variable working conditions only trained with the labeled dataset under a certain working condition can be realized, so long as the method can design a model which can extract the fault features and eliminate the working condition information, which is the idea of our method. Inspired by this theory, we proposed a data-driven intelligent fault diagnosis method under variable working conditions in this paper, which is called **D**ata-driven **F**ault feature **S**eparation **M**ethod (DFSM). The main contributions of our work can be summarized as follows.

1) For intelligent fault diagnosis method under variable working conditions, different from the present methods, we proposed a new idea that we can indirectly extract fault features by eliminating the working condition information.

2) To ensure the feature more stable, kernel method is introduced for calculating the distribution distances and vector distances.

3) Through resetting the fault features belonging normal health condition to zero, we ensure the working condition encoder can only extract the working condition features.

4) DFSM can extract all the information of training samples through two encoders with the help of autoencoder optimization object.

This paper is organized as follows. In Section II, several theory backgrounds are described. The framework of DFSM is detailed in Section III. In Section IV, the diagnosis case of bearing dataset under variable rotational speed is studied to test the effectiveness of DFSM. Finally, main conclusions are given in Section V.

II. THEORY BACKGROUND

A. CONVOLUTIONAL OPERATION AND POLLING LAYER

Comparing with feature extracting method of inner product like fully-connected layer, convolutional operation has shift invariant, i.e., it can extract the certain fault information with only one convolution kernel no matter when the fault happens, which is suitable for processing vibration signal. In this paper, we employ the one-dimensional convolutional operation since vibration signal is one-dimensional. Concretely, the output $\boldsymbol{c}$ of one-dimensional convolutional layer is obtained by

$$\boldsymbol{c} = f(\boldsymbol{w}^* \boldsymbol{u} + b) \tag{1}$$

where, * represents convolution, $\boldsymbol{w} \in \boldsymbol{R}^l$ is referred to as the convolution kernel, b is the corresponding bias, $\boldsymbol{u}$ is the input of convolutional layer and f represents the activation function.

The stride of introduced convolutional operation is 1 and the dimensions of $\boldsymbol{c}$ and $\boldsymbol{u}$ are same, because we want to gain more information in the process of feature extracting. Each convolutional layer is connected with a pooling layer which is conducted to reduce the dimension of convolution features. The max pooling function is utilized in the paper, which returns the maximum value within a certain sub-region as follows

$$\boldsymbol{p}_{[i]} = \max\left(\boldsymbol{c}_{[(i-1)\times s:i\times s]}\right) \tag{2}$$

where, s is the pooling length and the subscript $[i]$ represents the ith point of pooling output $\boldsymbol{p}$. In the same way, $[(i-1)\times s: i \times s]$ represents the sub-region between $(i-1)\times s$th point and $i \times s$th point of convolution feature $\boldsymbol{c}$.

B. AUTOENCODER

As one of the widely used deep learning techniques, AE has captured increasing attention in the field of fault diagnosis [16], [17], [32]. Our DFSM considers decomposing and reconstructing the vibration signal so we adopt the structure of AE. Generalization of AE is detailed in this section.

AE includes encoder and decoder. We can abstract the encoder into a map ψ_e: $\mathcal{X} \rightarrow \mathcal{S}$, where, $\mathcal{X}$ and $\mathcal{S}$ are data space and hidden feature space, respectively. Therefore, the hidden layer feature $\boldsymbol{u}_i$ can be obtained by

$$\boldsymbol{u}_i = \psi_e(\boldsymbol{x}_i; \theta_e) \tag{3}$$

where, $\boldsymbol{x}_i$ is the input of encoder and θ_e is the parameters of ψ_e. Then, the hidden layer feature $\boldsymbol{u}_i$ is used to reconstruct the input $\boldsymbol{x}_i$ through decoder. Similarly, we abstract the decoder into a map ψ_d: $\mathcal{S} \rightarrow \mathcal{O}$ with parameters θ_d, where, $\mathcal{O}$ denotes output space. Thus the decoder process is defined as follows

$$\hat{\boldsymbol{x}}_i = \psi_d(\boldsymbol{u}_i; \theta_d) \tag{4}$$

AE want to reduce the errors between input and output. Therefore, the loss function is

$$L_{ae}(\theta_e, \theta_d; \boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\right\|^2 \tag{5}$$

where, $X = \{x_i\}_{i=1}^n$ is the training data set and $||\cdot||$ denotes *l2-norm* or called *Frobenius norm*.

C. MAXIMUM MEAN DISCREPANCY BETWEEN TWO DISTRIBUTIONS

Maximum Mean Discrepancy (MMD) is a measure of the difference between two distributions from their samples. It is an effective criterion that compares distributions. Given two distributions $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ on activation vector space $\mathcal{V}$, MMD is defined as

$$\mathrm{MMD}\left(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}\right) = \sup_{\sigma \in \mathcal{F}} \left(\mathcal{E}_{\boldsymbol{v}\sim\mathcal{D}^{(1)}}[\sigma(\boldsymbol{v})] - \mathcal{E}_{\boldsymbol{v}\sim\mathcal{D}^{(2)}}[\sigma(\boldsymbol{v})]\right) \tag{6}$$

where, $\mathcal{F}$ is a class of functions f: $\mathcal{V} \rightarrow \mathcal{H}$. $\mathcal{H}$ denotes Reproducing Kernel Hilbert Space (RKHS) [33]. Based on the fact that σ is in the unit ball in a universal RKHS, one may rewrite the empirical estimate of MMD as

$$D\left(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)}\right) = \left\| \frac{1}{n_1}\sum_{i=1}^{n_1} \phi\left(\boldsymbol{v}_i^{(1)}\right) - \frac{1}{n_2}\sum_{j=1}^{n_2} \phi\left(\boldsymbol{v}_j^{(2)}\right) \right\|_{\mathcal{H}} \tag{7}$$

where, $\boldsymbol{V}^{(1)} = \{v_i^{(1)}\}_{i=1}^{n_1}$ and $\boldsymbol{V}^{(2)} = \{v_j^{(2)}\}_{j=1}^{n_2}$ are feature sets on activation vector space V. The superscripts (1) and (2) represent that the samples follow distributions $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, respectively. Besides, $\varphi(\cdot) : \mathcal{V} \rightarrow \mathcal{H}$ is referred to a feature space map.

To calculate the distribution distance of high-level learned features between different distributions, we employ kernel method. Then, the practical computation of MMD is written as,

$$\begin{aligned} D_k\left(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)}\right) &= [\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \left\langle \phi\left(\boldsymbol{v}_i^{(1)}\right), \phi\left(\boldsymbol{v}_j^{(1)}\right)\right\rangle \\ &\quad + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} \left\langle \phi\left(\boldsymbol{v}_i^{(2)}\right), \phi\left(\boldsymbol{v}_j^{(2)}\right)\right\rangle \\ &\quad - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \left\langle \phi\left(\boldsymbol{v}_i^{(1)}\right), \phi\left(\boldsymbol{v}_j^{(2)}\right)\right\rangle]^{\frac{1}{2}} \\ &= [\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} k\left(\boldsymbol{v}_i^{(1)}, \boldsymbol{v}_j^{(1)}\right) \\ &\quad + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} k\left(\boldsymbol{v}_i^{(2)}, \boldsymbol{v}_j^{(2)}\right) \\ &\quad - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} k\left(\boldsymbol{v}_i^{(1)}, \boldsymbol{v}_j^{(2)}\right)]^{\frac{1}{2}} \end{aligned} \tag{8}$$

where, $D_k(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)})$ is the unbiased estimation of $D(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)})$. $k(^{\bullet,\bullet})$ is a kernel function, which could compute the inner product in a higher dimensional space, i.e., $k(\boldsymbol{x}, y) =< \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y})>$. For further improving the stability of learned feature, we adopt multiple kernel variant of MMD

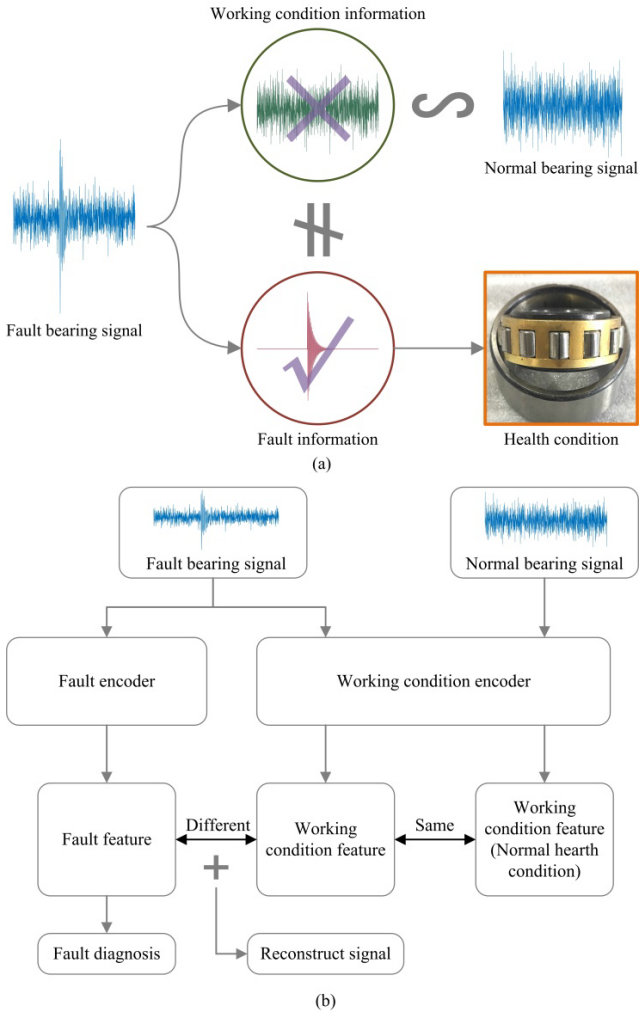

FIGURE 1. Motivation of DFSM: (a) Fundamental theory, (b) Sketch of DFSM.

(MK-MMD) in this paper which is obtained by

$$D_{MK}\left(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)}\right) = \sum_{k \in K} D_k\left(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)}\right) \tag{9}$$

where, K is the set of kernel functions.

III. PROPOSED METHOD

The proposed DFSM is based on CNN and makes reference to the idea of AE to extract the fault features and eliminate the working condition information. In this section, we mainly describe the details of our DFSM, including the main idea, model structure and training strategy.

A. MOTIVATION OF DFSM

For better illustration of DFSM, we detail our motivation first. Fig.1 is given to make it easier to understand. As can be seen from Fig.1(a), fault bearing signal contains two primary information: most is the working condition information that is useless for fault diagnosis and the rest is fault information which is difficult to extract. Therefore, the present methods focus on the representation of fault information, which is the paramount classifiable feature for detecting the health condition. According to the previous description in section I, fault information is irrelevant to the working condition. We consider that fault information is different from the working condition information. If we can completely extract the working condition information, we can obtain the corresponding fault information. So the point is to find the reference of working condition information. It should be noticed that the normal bearing signal only contains the working condition information. The working condition feature of fault bearing signal and normal bearing signal must be similar. Thus, the motivation of DFSM is to eliminate the working condition information from fault bearing signal with reference of normal bearing signal and extract the fault information for detecting the health condition.

We make a brief introduction of DFSM framework combining with the sketch as shown in Fig.1(b). We consider designing two encoders, one decoder and one classifier. One encoder is fault encoder ψ_{fe} which can extract the fault feature and another is working condition encoder ψ_{we} for extracting the working condition feature. Fault feature is used as the input of classifier ψ_c for fault diagnosis. We want to increase the discrepancies between the working condition feature and fault feature and reduce the differences of working condition features extracted from fault bearing signal and normal bearing signal. Besides, working condition feature and fault feature are combined as the inputs of decoder.

B. DFSM MODEL

This section details the realization method of our DFSM. DFSM is constructed by combining CNN with AE and redefining the original objective function of them. With the help of these losses, DFSM learns the precise fault feature without working condition information. For describing the method conveniently, the illustration of method is shown in Fig.2.

There are C health conditions, and the vibration signals of machines are obtained under different health conditions. These signals compose the training data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in R^{mx}$ is the ith sample containing m_x vibration data points and y_i is the health condition label. We define a fault encoder ψ_{fe}, a working condition encoder ψ_{we}, a classifier ψ_c and a decoder ψ_d with parameters θ_{fe}, θ_{we}, θ_c and θ_d, respectively. Their forms are shown in Fig.2. It should be noticed that the original vibration signal in time domain is used as the input of our method.

The forms of two encoders are same, which contain three convolutional layers, three max pooling layers and one fully-connected layer as shown in Fig.2, where, $l20s1$ denotes that the length of convolution kernel is 20 and the stride of convolutional operation is 1, 32/ReLU is the channel number and activation function. For max pooling layers, $l4s4$ denotes that the pooling length is 4 and the stride is 4. For fully-connected layer, m_h/ReLU is the output dimension and activation function. The form of decoder contains

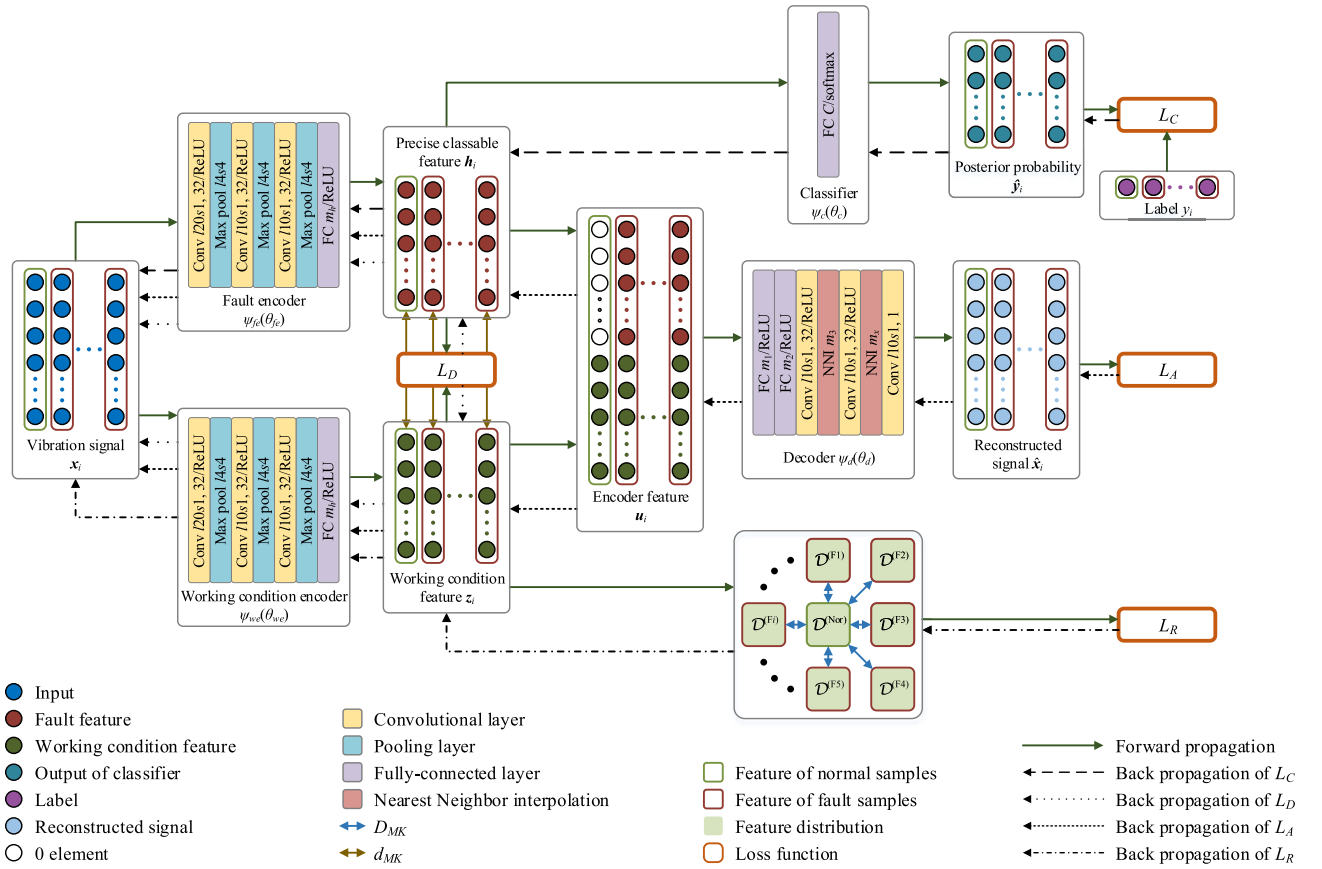

FIGURE 2. Structure illustration of the proposed method.

three convolutional layers, two fully-connected layers and two nearest neighbor interpolation operations (NNI) as shown in Fig.2, where, NNI m_x means that the feature dimensions are extended to m_x using nearest neighbor interpolation operation. The classifier contains only one fully-connected layer. By employing softmax regression, the classifier can estimate the health conditions of machines and give the posterior probabilities of categories as follows,

$$\hat{\boldsymbol{y}}_i = \text{softmax}\left(\boldsymbol{o}_i\right) = \frac{e^{\boldsymbol{o}_i}}{\sum_{j=1}^{C} e^{\boldsymbol{o}_{i[j]}}} \tag{10}$$

where, $\hat{\boldsymbol{y}}_i \in R^C$ is the posterior probability vector of ith sample and $\boldsymbol{o}_i$ is the output of fully-connected layer without activation function.

C. CLASSIFICATION LOSS

The model should ensure the basic ability of classification. Therefore, the first optimization object of the proposed model is to minimize the health condition classification error on the source domain labeled data. The posterior probability vector $\hat{\boldsymbol{y}}_i$ of ith sample $\boldsymbol{x}_i$ can be obtained by

$$\boldsymbol{h}_i = \psi_{fe}\left(\boldsymbol{x}_i; \theta_{fe}\right) \tag{11}$$

$$\hat{\boldsymbol{y}}_i = \psi_c\left(\boldsymbol{h}_i; \theta_c\right) \tag{12}$$

where, $\boldsymbol{h}_i \in R^{mh}$ is the precise classable feature. Then, the desired objective function can be defined as a standard softmax regression loss,

$$L_C\left(\theta_{fe}, \theta_c; \boldsymbol{x}_i, y_i\right) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1\left\{y_i = c\right\} \log \hat{\boldsymbol{y}}_{i[c]} \tag{13}$$

where, $1\{\bullet\}$ denotes the indicator function returning 1 if the condition is true, and 0 otherwise.

D. UNCORRELATION LOSS

Fault information is irrelevant to the working condition. Therefore, the fault feature and working condition feature should be uncorrelated. We want to decrease the inner product between the working condition feature $\boldsymbol{z}_i$ extracted by ψ_{we} and precise classable feature $\boldsymbol{h}_i$ extracted by ψ_{fe}. Then the uncorrelation loss is defined as follows,

$$L_D\left(\theta_{fe}, \theta_{we}; \boldsymbol{x}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left\langle \boldsymbol{h}_i, \boldsymbol{z}_i \right\rangle \tag{14}$$

where, $\boldsymbol{z}_i = \psi_{we}\left(\boldsymbol{x}_i; \theta_{we}\right)$ is the working condition feature.

E. REFERENCE LOSS

Reference is quite important for our DFSM. If lack it the two encoders can only extract two different features instead of

fault feature and working condition feature. As mentioned before, normal health condition samples are suited reference. Normal health condition samples contain all the working condition information, so they can be regarded as the samples eliminating fault information. Thus features extracted from fault samples through working condition encoder should be similar with features extracted from normal samples. It should be noticed that this similarity is between the whole distributions of different categories. Therefore, the reference loss can be defined using MK-MMD,

$$L_R\left(\theta_{we}; \boldsymbol{x}_i, y_i\right) = \sum_{i=1}^{C-1} D_{MK}\left(\boldsymbol{Z}^{(Nor)}, \boldsymbol{Z}^{(Fi)}\right) \tag{15}$$

where, $\boldsymbol{Z}^{(\mathrm{Nor})} = \{z_a^{(\mathrm{Nor})}\}_{a=1}^{n_{\mathrm{Nor}}}$ and $\boldsymbol{Z}^{(\mathrm{F}i)} = \{z_b^{(Fi)}\}_{b=1}^{n_{\mathrm{F}i}}$ are feature sets. The superscripts (Nor) and (Fi) represent that the samples follow distributions $\mathcal{D}^{(\mathrm{Nor})}$ and $\mathcal{D}^{(\mathrm{F}i)}$, i.e., the labels of samples are normal health condition and ith fault, respectively.

Minimizing the reference loss means the features extracted through working condition encoder are both similar to the reference without fault information.

F. AUTOENCODER LOSS

It should be noticed that there is an important concealment condition: the two encoders can extract all the information of samples and the working condition encoder only extract working condition feature. Therefore, the precise classable feature $\boldsymbol{h}_i$ and working condition feature $\boldsymbol{z}_i$ should contain all the sample information, and thus can be the sufficient condition for reconstructing the vibration signal. Besides, we want to ensure $\boldsymbol{z}_i$ containing working condition information as much as possible, so long as ψ_{we} can encoder more working condition information, i.e. for normal health condition samples, we can almost reconstruct the vibration signal only using working condition feature $\boldsymbol{z}_i$. So we consider using autoencoder loss and design the decoder.

First, the input of decoder should be constructed combining $\boldsymbol{h}_i$ and $\boldsymbol{z}_i$. We change the elements of fault features $\boldsymbol{h}_i$ which belong to normal samples to zero and then obtain $\hat{\boldsymbol{h}}_i$ because we want to reconstruct the normal vibration signal only using working condition information. So the encoder feature $\boldsymbol{u}_i$ is

$$\boldsymbol{u}_i = [\hat{\boldsymbol{h}}_i, \boldsymbol{z}_i]^T \tag{16}$$

Then the reconstructed sample can be obtained through decoder ψ_d

$$\hat{\boldsymbol{x}}_i = \psi_d\left(\boldsymbol{u}_i; \theta_d\right) \tag{17}$$

Finally, the desired objective function can be defined as a standard autoencoder loss,

$$L_A\left(\theta_{fe}, \theta_{we}, \theta_d; \boldsymbol{x}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\right\|^2 \tag{18}$$

Minimizing the autoencoder loss means that the two encoders can extract all the information from samples, besides, encoder ψ_{we} can extracted the entire working condition feature.

G. OPTIMIZATION OBJECT AND ALGORITHM

By integrating (13), (14), (15) and (18) together, the final objective function of DFSM is written as

$$L\left(\theta; \boldsymbol{x}_i, y_i\right) = L_C + \alpha L_D + \beta L_R + \gamma L_A \tag{19}$$

where, α, β and γ are both greater than zero and control the tradeoff among optimization objects, $\theta = [\theta_{fe}, \theta_{we}, \theta_c, \theta_d]$ is parameters.

According to equation (19) and Fig.2, it is convenient to train the proposed method by stochastic gradient descent (SGD) algorithm. Therefore the loss function equation (19) is rewritten as follows,

$$\begin{aligned} L\left(\theta_{fe}^*, \theta_{we}^*, \theta_c^*, \theta_d^*\right) = \min_{\theta_{fe},\theta_{we},\theta_c,\theta_d} & L_C\left(\theta_{fe}, \theta_c\right) + \alpha L_D\left(\theta_{fe}, \theta_{we}\right) \\ & + \beta L_R\left(\theta_{we}\right) + \gamma L_A\left(\theta_{fe}, \theta_{we}, \theta_d\right) \end{aligned} \tag{20}$$

Based on the equation (22) and SGD algorithm, the parameters θ are updated as follows,

$$\theta_{fe} \leftarrow \theta_{fe} - \varepsilon\left(\frac{\partial L_C}{\partial \theta_{fe}} + \alpha\frac{\partial L_D}{\partial \theta_{fe}} + \gamma\frac{\partial L_A}{\partial \theta_{fe}}\right) \tag{21}$$

$$\theta_{we} \leftarrow \theta_{we} - \varepsilon\left(\alpha\frac{\partial L_D}{\partial \theta_{we}} + \beta\frac{\partial L_R}{\partial \theta_{we}} + \gamma\frac{\partial L_A}{\partial \theta_{we}}\right) \tag{22}$$

$$\theta_c \leftarrow \theta_c - \varepsilon\frac{\partial L_C}{\partial \theta_c} \tag{23}$$

$$\theta_d \leftarrow \theta_d - \varepsilon\gamma\frac{\partial L_A}{\partial \theta_d} \tag{24}$$

where, ε is the learning rate.

When the training process is completed, we can obtain a fault encoder ψ_{fe} precisely extracting the classable feature, a working condition encoder ψ_{we} encoding all the working condition information, a classifier ψ_c which can give the posterior probability of categories and a decoder ψ_d reconstructing samples. When we use testing samples to verify the proposed DFSM, only the fault encoder and the classifier are employed for fault diagnosis.

IV. CASE STUDY: FAULT DIAGNOSIS OF ROLLING BEARING USING DSFM

This section shows the ability of proposed DFSM for dealing with different rotational speed samples. The DFSM model only trained with the samples under a certain rotational speed is employed to classify the samples under other rotational speeds and time-varying speed. We investigate the results and certify the idea of DFSM.

A. DATA DESCRIPTION

As shown in Fig. 3, the vibration signals of rolling bearings were collected from a test bench which consisted of an induction motor, four supporting pillow blocks, a tachometer and a coupling. The tested bearing is installed in the farthest supporting pillow block from the induction motor.

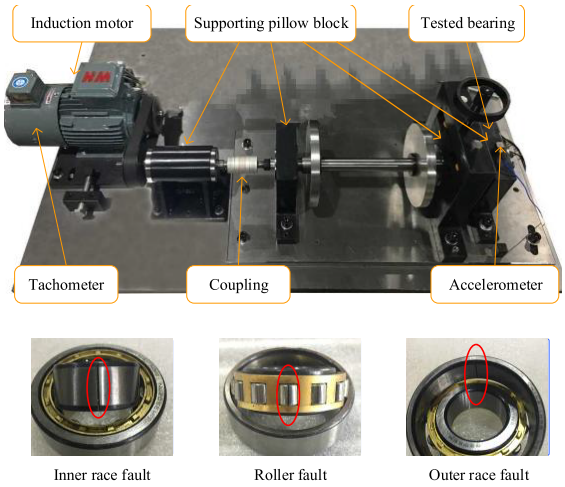

FIGURE 3. The arrangement of bearing test bench and bearing with fault.

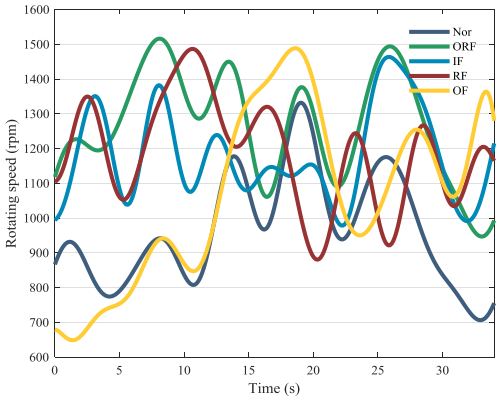

FIGURE 4. Rotational speeds of varying speed dataset.

An accelerometer is mounted on the supporting pillow block of tested bearing to measure the vibration signals. The sampling frequency is 25.6 kHz. The bearing dataset vibration signals are collected under five different health conditions which include normal condition (Nor), inner race fault (IF), roller fault (RF), outer race fault (OF) and concurrent faults in the outer race and roller (ORF).

The training dataset is composed of the labeled signals under rotational speed of 1500 rpm, which is a time-invariant working condition. There are 1000 signals for each health condition and each signal contains 1200 data points. For testing the method, two types of testing datasets severally composed by time-invariant rotational speed signals and time-varying rotational speed signals are built. The time-invariant rotational speed signals under three drive motor speeds (900rpm, 1000rpm, and 1300rpm) are used to build the time-invariant rotational speed testing dataset (TI1, TI2 and TI3), which has the large discrepancy of rotational speed comparing with the training dataset. For testing the robustness of proposed method, the signals under large speed oscillation compose the time-varying speed dataset TV which has the large discrepancy of angular acceleration comparing with the training dataset. The rotational speeds are ruleless as shown in Fig. 4 and the time-frequency distributions are indistinct according to Fig.5. It should be noticed that almost all the rotational speeds of TV are under 1500 rpm, which are incompatible with the training dataset. We randomly sample segments with 1200 data points from all the testing signals as testing samples to compose the TV dataset of 10000 samples and TI dataset of 30000 samples.

B. DIAGNOSIS OF DFSM

The form of kernel function should be selected first. According to Ref. [34], the Gaussian radial basis function (RBF), i.e., $k_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2/2s^2)$, maps the original features to an infinite-dimensional space, and it has been well studied and proven to make distance useful in practice [35], where s is the standard deviation. Different s corresponds to different infinite-dimensional spaces. Therefore, the proposed method uses different RBF and sums them to calculate the distances for ensuring the features stabilized in different infinite-dimensional spaces. We use RBF whose mid-value is m and times between two s. To solve the problem of selecting penalty parameters, m is the mean of $||\boldsymbol{x}_1 - \boldsymbol{x}_2||$ on the training data. For example, if $m = 1$ and the number of kernel is 5, the final standard deviations are [0.25, 0.5, 1, 2, 4].

The structure parameters of the DFSM is $[m_h, m_1, m_2, m_3] = [100, 600, 800, 1000]$. In the beginning of optimization, the error is quite big and could decrease with the epoch. So the learning rate should decrease in the process of optimization. Based on Ref. [36], the learning rate and training step of SGD are $0.01/(1+10\times q)^{0.75}$ and 500, respectively, where q is the training progress linearly changing from 0 to 1. α, β and γ are equal to 1. For training our DFSM better, we adopt batch normalization (BN) to every convolutional layer. But batch normalization has the capability of generalization, we design a batch normalized CNN with the same structure to ψ_{fe} and ψ_c for certifying the efficacy of DFSM. It should be noticed that 15 trails are carried out for each experiment in the following studies to reduce the effects of the randomness.

The diagnosis results are shown in Fig.6. In this figure, the testing accuracy is the minimum value of 15 trials, and the positive error bars show the range of testing accuracy. It can be seen that the testing accuracies of different testing datasets are over 98.54%, which means that the proposed DFSM is able to diagnosis the fault under different working conditions even time-varying working condition only trained with a certain working condition and the accuracies are quite high. For CNN with BN, the highest testing accuracy range is 96.48%∼96.80% of testing dataset TI3 because the testing working condition is nearest to the condition of training dataset's. The testing accuracies are lower when the rotational speed goes lower. This means that the generalization ability of BN is limited. The comparing results of the two methods certify that the structure of DFSM is effective. Besides, we have investigated the parameters α, β and γ and find that the

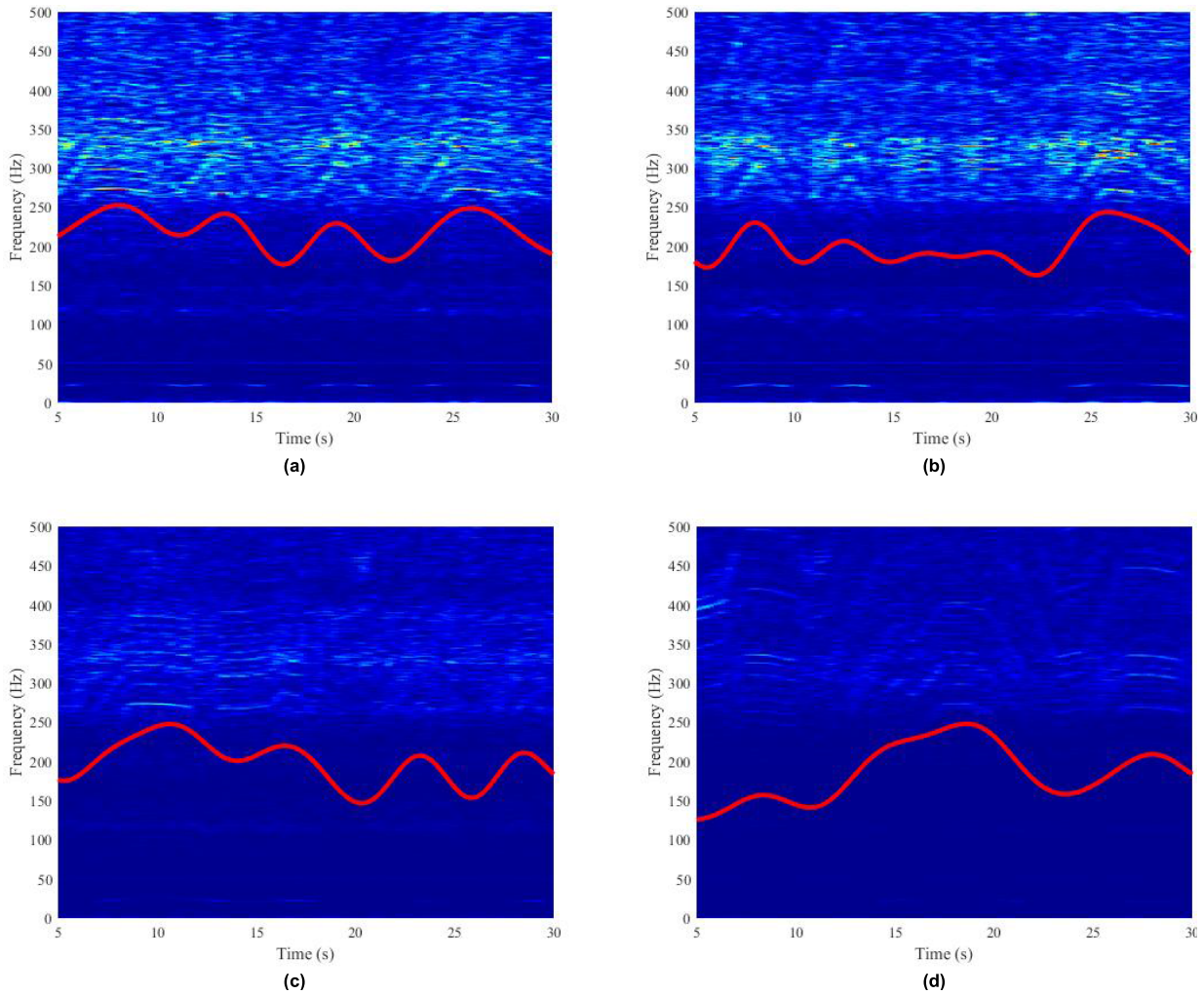

FIGURE 5. Time-frequency distribution of fault signals and fitted curves of tenfold rotational speed frequencies: (a) ORF (b) IF(c) RF (d) OF.

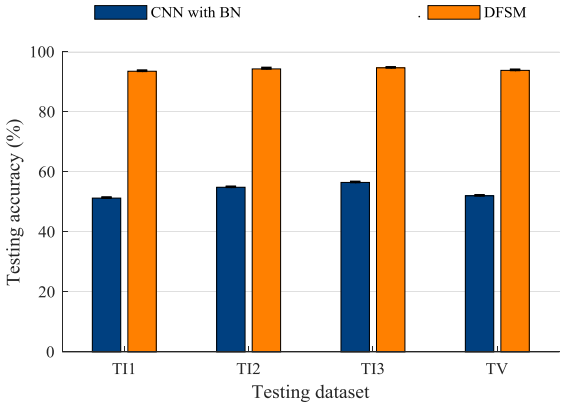

FIGURE 6. Diagnosis results of different testing dataset.

selecting ranges of them are relatively large (around 1∼100) but it is best to select them on the same order of magnitude. For bearing fault diagnosis, the recommended values of them are both 1.

C. EQUATIONS VERIFICATION OF DFSM

To study the feature extracted from the two encoders obviously, t-distributed stochastic neighbor embedding (t-SNE) [37] is used. This technique makes it possible to embed these 100-D vectors in a 2-D image in such a way that the vectors are close together in the 100-D space are also close together in the 2-D plot [38]. For each category and testing dataset, we randomly select 100 samples and then extract the precise classable features and working condition features for reducing dimensions. The t-SNE visualization are displayed in Fig.7(a)∼(b). As we can see from the t-SNE visualization, most features of the same health condition are gathered in the corresponding cluster and most features of different health conditions are separated. Meanwhile we can find that features learned by DFSM exhibit tight health condition class clustering while mixing the feature distribution between different testing datasets, which means that the proposed DFSM only trained by a certain working condition is able to extract the fault features under different working condition. Working condition features extracted through ψ_w are consistent for

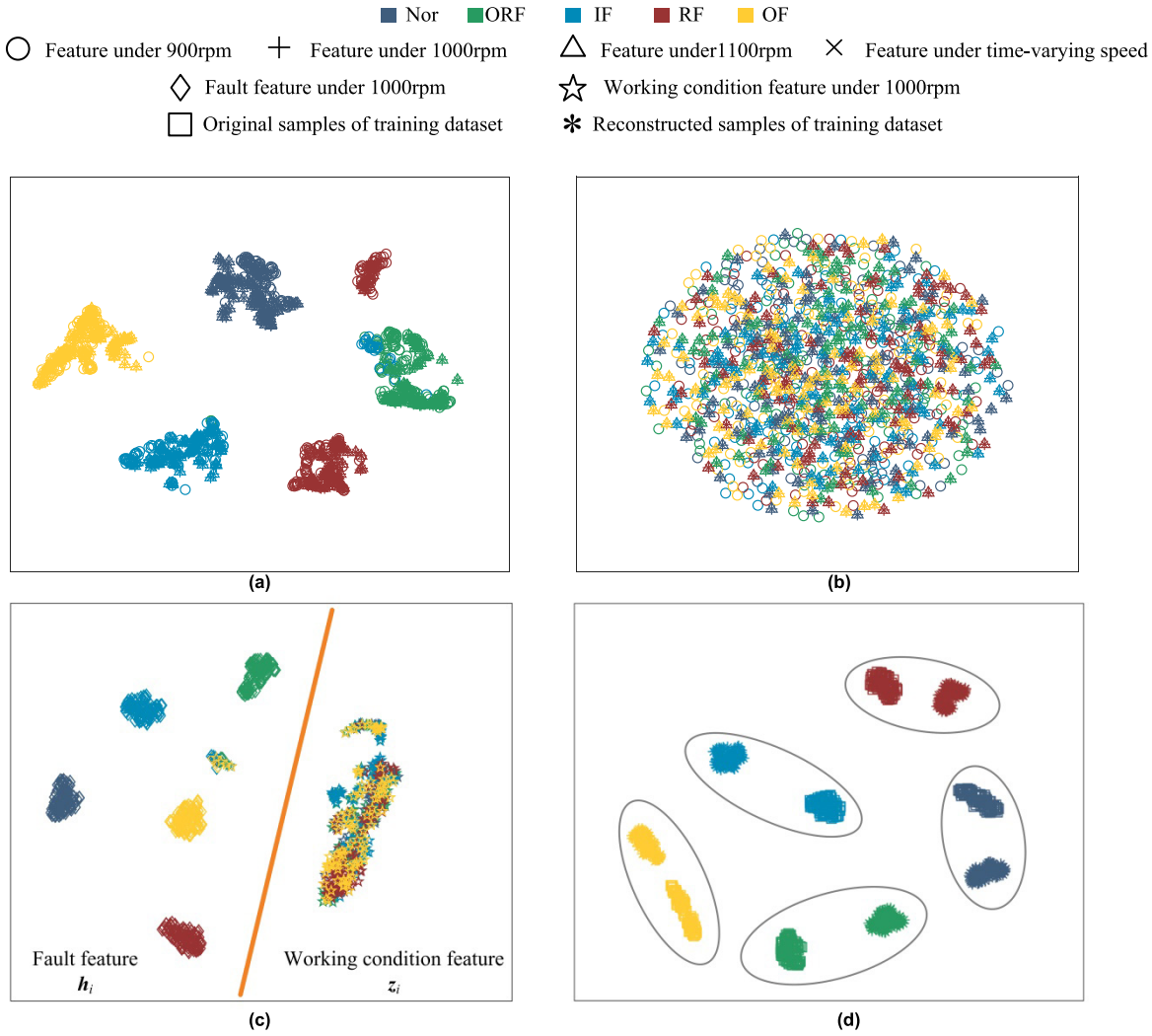

FIGURE 7. The t-SNE visualization of (a) precise classable feature, (b) working condition feature, (c) two different features under 1000 rpm and (d) original samples and reconstructed samples.

every health condition and the t-SNE visualization is difficult to classify as shown in Fig.7(b). Besides, for verifying that the two encoder (fault encoder and working condition encoder) extract different information, the fault feature $\boldsymbol{h}_i$ and working condition feature $\boldsymbol{z}_i$ under 1000 rpm are investigated. We also use t-SNE to reduce the dimension. The results are shown in Fig.7(c). As can be seen from the figure, we can distinguish the two feature just using a linear boundary, which means that the fault feature $\boldsymbol{h}_i$ and working condition feature $\boldsymbol{z}_i$ comprise different information.

However, the different precise classable features and the same working condition features cannot certify that we realize idea of DFSM because if the two kinds of features cannot be used to reconstruct the original samples, the fault features is not regarded as the features eliminating all the working condition information. For certify the idea of DFSM, the training dataset is processed: first, for training samples and reconstructed samples, Fast Fourier Transform (FFT) is applied to every signal sample to obtain the corresponding frequency amplitude; second for obtaining more stable spectrum, we average the frequency amplitudes of every 10 samples and obtain 500 average spectra of samples and 500 average spectra of corresponding reconstructed samples; then, we extract 14 handcrafted features, i.e., mean, root mean square, kurtosis, variance, crest factor, wave factor, and eight energy ratios of wavelet package transform of every spectra; finally, we adopt t-SNE to reduce dimension and display the principal components in Fig.7(d). In this figure, the reconstructed samples are different from the original samples, but the difference is not large. There are always the corresponding reconstructed samples close to the original samples, which indicates that the information of them are similar. Therefore, we consider that the two encoders can extract all the information of training samples. It should be noticed that the distance between the original samples and the reconstructed samples under normal health condition is the shortest. This means the information of them is basically consistent. For samples under normal health condition, when we reconstruct

TABLE 1. Classification comparison of the rolling bearing dataset.

Method	Training dataset	Testing dataset	Testing accuracy	Training time
CNN	Labeled dataset under load 900rpm	Dataset under 1100rpm	81.24%	16 min
TCA [39]	Labeled dataset under 900rpm + Unlabeled dataset under 1100rpm	Dataset under 1100rpm	87.34%	-
JDA [40]	Labeled dataset under 900rpm + Unlabeled dataset under 1100rpm	Dataset under 1100rpm	91.42%	-
DAFD [24]	Labeled dataset under 900rpm + Unlabeled dataset under 1100rpm	Dataset under 1100rpm	94.73%	14 min
DTL [25]	Labeled dataset under 900rpm + Unlabeled dataset under 1100rpm	Dataset under 1100rpm	95.85%	19 min
	Labeled dataset under 1100rpm + Unlabeled dataset under 900rpm	Dataset under 900rpm	89.77	18 min
Our DFSM	**Only labeled dataset under load 900rpm**	Dataset under 1100rpm	**95.97%**	21 min
	Only labeled dataset under load 1100rpm	Dataset under 900rpm	**93.81%**	20 min

the samples, only the features extracted through working condition encoder ψ_w is used, which indicates that the working condition encoder ψ_w is able to extract all the working condition information of samples.

Through the visual understanding, the proposed DFSM is certified that it is able to eliminate the working condition information and extract precise classable feature for fault diagnosis.

D. COMPARING WITH RELATED WORK

To show the effectiveness of DFSM, we compare it with the methods in related work using the same rolling bearing dataset.

For our DFSM, we want to study the importance of DFSM structure, so we do not use BN. The learning rate and training step are 0.001 and 150, respectively and other model sets are same to section 4. We design a CNN with the same structure to ψ_{fe} and ψ_c for certifying the efficacy of DFSM.

Transfer **C**omponent **A**nalysis (TCA) [39] is the representative work of searching the feature subspace by marginal probability adaptation in the domain adaptation field. **J**oint **D**istribution **A**daptation (JDA) [40] combines marginal probability adaptation with conditional probability adaptation. For the two methods, spectra of samples are the inputs, the kernel functions are both RBF kernel with standard deviation 1 and output dimension is 100. Then a SVM classifier is trained on the labeled source data to classify the unlabeled target data.

Deep neural network for domain **A**daptation in **F**ault **D**iagnosis (DAFD) [24] and **D**eep **T**ransfer **L**earning (DTL) [25] are the successful methods based on transfer learning (or called domain adaptation) for fault intelligent diagnosis. The reporting results are quite valuable for comparison. It should be noticed that authors used single layer network in the paper of DAFD.

Besides, we investigate the training time of the methods trained by backpropagation. The computation platform is a PC with an Intel I5cpu and 8G RAM. The Classification comparison of the rolling bearing dataset is displayed in Table 1.

According to the calculated results in Table 1, we can observe that the accuracies of DFSM reach 95.97% and 93.81% under two different experiments. This result outperforms other listed methods. Specifically, the accuracy of CNN is 81.24%, which is 14.73% lower than that of DFSM. This comparing result means that the structure of DFSM is effective. DAFD obtains best accuracy 94.73% among the single layer network, which is 1.24% lower than our method. For DTL, the diagnosis accuracies are both lower than DFSM, especially when the testing dataset is under load 0hp. Besides, it should be noticed that our DFSM is trained only using one dataset under a certain working condition. But, training DFSM cost a little more time comparing with related work. The reason is that DFSM contains extra working condition information encoder and decoder which should be trained.

V. CONCLUSION

For intelligent fault diagnosis under variable working condition, the key is to extract the features which are irrelevant to the working conditions and contain fault information. The fundamental theory is that fault information is irrelevant to the working conditions. Inspired by this theory, we consider indirectly extracting fault features by eliminating the working condition information with the reference of normal health condition samples, and then proposed four loss functions to realize our DFSM. The experimental results of two rolling bearing datasets show that the proposed DFSM can not only break the limitation of existing methods, but also achieve a superior performance comparing with related methods. It should be noticed that our DFSM is trained only using the dataset under a certain working condition and can diagnose faults with high accuracy under variable even time-varying working condition, so it is easy to be applied. Through the visual understanding, the proposed DFSM is certified that the proposed optimization objects can realize the idea of DFSM. Besides, as we perform this study, we find that it is best to adopt BN because the model is quite complex and a little difficult to train. Hence, we will do research on more simple framework in the future study.

REFERENCES

[1] D. Abboud, M. Elbadaoui, W. A. Smith, and R. B. Randall, "Advanced bearing diagnostics: A comparative study of two powerful approaches," *Mech. Syst. Signal Process.*, vol. 114, pp. 604–627, Jan. 2019.

[2] 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 Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2019.

[3] X. Yan, M. Jia, W. Zhang, and L. Zhu, "Fault diagnosis of rolling element bearing using a new optimal scale morphology analysis method," *ISA Trans.*, vol. 73, pp. 165–180, Feb. 2018.

[4] S. Lu, Q. He, and J. Zhao, "Bearing fault diagnosis of a permanent magnet synchronous motor via a fast and online order analysis method in an embedded system," *Mech. Syst. Signal Process.*, vol. 113, pp. 36–49, Dec. 2018.

[5] L. Saidi, J. B. Ali, and F. Fnaiech, "Application of higher order spectral features and support vector machines for bearing faults classification," *ISA Trans.*, vol. 54, pp. 193–206, Jan. 2015.
[6] S. Lu and X. Wang, "A new methodology to estimate the rotating phase of a BLDC motor with its application in variable-speed bearing fault diagnosis," *IEEE Trans. Power Electron.*, vol. 33, no. 4, pp. 3399–3410, Apr. 2018.
[7] X. Chen and Z. Feng, "Time-frequency space vector modulus analysis of motor current for planetary gearbox fault diagnosis under variable speed conditions," *Mech. Syst. Signal Process.*, vol. 121, pp. 636–654, Apr. 2019.
[8] J. Wang, Y. Peng, and W. Qiao, "Current-aided order tracking of vibration signals for bearing fault diagnosis of direct-drive wind turbines," *IEEE Trans. Ind. Electron.*, vol. 63, no. 10, pp. 6336–6346, Oct. 2016.
[9] W. Huang, J. Cheng, and Y. Yang, "Rolling bearing fault diagnosis and performance degradation assessment under variable operation conditions based on nuisance attribute projection," *Mech. Syst. Signal Process.*, vol. 114, pp. 165–188, Jan. 2019.
[10] X. Jiang, S. Li, and Q. Wang, "A study on defect identification of planetary gearbox under large speed oscillation," *Math. Problems Eng.*, vol. 2016, pp. 1–18, Jan. 2016.
[11] Z. Wang, Z. Han, F. Gu, J. X. Gu, and S. Ning, "A novel procedure for diagnosing multiple faults in rotating machinery," *ISA Trans.*, vol. 55, pp. 18–208, Mar. 2015.
[12] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.
[13] J. K. Sinha and K. Elbhbah, "A future possibility of vibration based condition monitoring of rotating machines," *Mech. Syst. Signal Process.*, vol. 34, nos. 1–2, pp. 231–240, Jan. 2013.
[14] B. A. Paya, I. Esat, and M. Badi, "Artificial neural networks based fault diagnosis of rotating machinery using wavelet transforms as a preprocessor," *Mech. Syst. Signal Process.*, vol. 11, no. 5, pp. 751–765, 1997.
[15] J. Rafiee, F. Arvani, A. Harifi, and M. H. Sadeghi, "Intelligent condition monitoring of a gearbox using artificial neural network," *Mech. Syst. Signal Process.*, vol. 21, no. 4, pp. 1746–1754, May 2007.
[16] F. Jia, Y. 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] H. Shao, H. Jiang, Y. Lin, and X. Li, "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders," *Mech. Syst. Signal Process.*, vol. 102, pp. 278–297, Mar. 2018.
[18] L. Liao, W. Jin, and R. Pavel, "Enhanced restricted Boltzmann machine with prognosability regularization for prognostics and health assessment," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7076–7083, Nov. 2016.
[19] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, pp. 490–502, Nov. 2016.
[20] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016.
[21] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.
[22] Z. An, S. Li, J. Wang, W. Qian, and Q. Wu, "An intelligent fault diagnosis approach considering the elimination of the weight matrix multicorrelation," *Appl. Sci.*, vol. 8, no. 6, p. 906, Jun. 2018.
[23] Y. Lei and M. J. Zuo, "Gear crack level identification based on weighted k nearest neighbor classification algorithm," *Mech. Syst. Signal Process.*, vol. 23, no. 5, pp. 1535–1547, Jul. 2009.
[24] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.
[25] 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.
[26] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks," 2017, *arXiv:1712.01636*. [Online]. Available: http://arxiv.org/abs/1712.01636
[27] F. Pouladi, H. Salehinejad, and A. M. Gilani, "Recurrent neural networks for sequential phenotype prediction in genomics," in *Proc. Int. Conf. Develop. E-Syst. Eng. (DeSE)*, Dec. 2015, pp. 225–230.
[28] H. Salehinejad and S. Rahnamayan, "Customer shopping pattern prediction: A recurrent neural network approach," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Athens, Greece, Dec. 2016, pp. 1–6.
[29] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 281–288, Feb. 2009.
[30] F. Jia, Y. Lei, N. Lu, and S. Xing, "Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization," *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, Sep. 2018.
[31] X. Jia, M. Zhao, Y. Di, Q. Yang, and J. Lee, "Assessment of data suitability for machine prognosis using maximum mean discrepancy," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5872–5881, Jul. 2018.
[32] J. Wang, S. Li, Z. An, X. Jiang, W. Qian, and S. Ji, "Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis of machines," *Neurocomputing*, vol. 329, pp. 53–65, Feb. 2019.
[33] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. 49–57, Jul. 2006.
[34] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," Dec. 2014, *arXiv:1412.3474*. [Online]. Available: http://arxiv.org/abs/1412.3474
[35] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.
[36] Y. Ganin, E. Ustinovan, H. Ajakann, P. Germainn, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, 2016.
[37] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[38] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.
[39] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[40] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2200–2207.

SHUNMING LI received the Ph.D. degree in mechanics from Xi'an Jiaotong University, China, in 1988. He is currently a Professor with the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include noise and vibration analysis and control, signal processing, machine fault diagnosis, sensing and measurement technology, and intelligent vehicles.

ZENGHUI AN received the B.S. and M.S. degrees from the University of Jinan, Jinan, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include mechanical fault diagnosis and deep learning.

JIANTAO LU received the Ph.D. degree from the School of Mechanical Engineering, Xi'an Jiaotong University, China, in 2019. He is currently a Lecturer with the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include mechanical fault diagnosis and underdetermined and nonlinear blind source separation.