Unsupervised Mechanical Fault Feature Learning Based on Consistency Inference-Constrained Sparse Filtering

RAN WANG*, (Member, IEEE), FENGKAI LIU, XIONG HU, AND JIN CHEN

1College of Logistics Engineering, Shanghai Maritime University, Shanghai 201306, China
2State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai 200240, China
Corresponding author: Ran Wang (ranwang@shmtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51505277, and in part by the National Key Research and Development Program of China under Grant 2019YFB2004600.

ABSTRACT In machinery fault diagnosis, a large amount of monitoring data is often unlabeled, while the number of labeled data is limited. Therefore, learning effective features from massive unlabeled data is a challenging issue for machinery fault diagnosis. In this paper, a simple unsupervised feature learning method, consistency inference-constrained sparse filtering (CICSF), is proposed to learn mechanical fault features with enhanced clustering performance for fault diagnosis. Firstly, inspired by the data augmentation strategy, consistency inference of latent representations for time series (CILRTS) is derived, which infers that training data instances segmented from the same time series should possess consistent latent feature representations. Then, CILRTS is integrated into sparse filtering (SF) as an additional constraint in the latent feature space. The developed CICSF method can optimize the inter-class sparsity and intra-class similarity of the feature distribution simultaneously. Thus, it can learn more effective features from massive unlabeled data. Finally, based on CICSF, a semi-supervised machinery fault diagnosis method is developed. After unsupervised feature learning by CICSF, a softmax regression classifier is trained with limited labeled data to realize machinery fault diagnosis. Experimental results on bearing and gearbox datasets verify the effectiveness of the proposed method. Moreover, comparisons with standard SF and several auto-encoder (AE) variants validate its superiority in unsupervised feature learning and fault diagnosis using limited labeled data.

INDEX TERMS Unsupervised feature learning, machinery fault diagnosis, consistent inference of latent representations for time series, sparse filtering, auto-encoder.

I. INTRODUCTION

In modern industries, machinery becomes more automatic and sophisticated than ever before, which requires a higher level of reliability. Therefore, it is necessary to conduct machinery condition monitoring and fault diagnosis. In the past decades, the rapid development of machine learning techniques has accelerated the progress of machinery intelligent fault diagnosis. Various traditional machine learning methods have been successfully applied in machinery fault diagnosis, such as the artificial neural network (ANN), support vector machine (SVM), hidden Markov model (HMM), and so on [1]–[4]. These methods generally have two key steps: feature extraction and fault classification.

Features extracted from the monitoring data can significantly affect the effectiveness of the fault diagnosis results. However, in these methods, feature extraction is often implemented manually according to specific fault diagnosis tasks. In real-world scenarios, it is difficult to choose proper feature extraction methods for machinery fault diagnosis. The optimal feature set often varies from case to case in different applications. So, adaptive feature learning is highly desirable in machinery intelligent fault diagnosis.

As a state-of-the-art machine learning technique, deep learning (DL) can learn hierarchical features automatically from large-scale data instead of manual feature extraction. In recent years, various DL models, e.g., convolutional neural network (CNN), residual neural network (ResNet), have been applied to intelligent fault diagnosis of bearings and planet gearboxes [5]–[11]. For example, Peng et al. [6] proposed...
a novel deep learning method based on one-dimensional ResNet for high-speed bearing fault diagnosis. Jia et al. [12] proposed a deep normalized CNN to deal with data imbalance in the process of fault classification. Most of these deep learning-based diagnosis methods are supervised learning methods, where large-scale labeled data are needed to train deep models. Data labels can provide explicit guidance in supervised feature learning. However, in practical industry, it is difficult to obtain massive labeled data from mechanical equipment, because machines are not allowed to operate in the event of failure. Moreover, manual annotation of large-scale machinery faulty datasets is time-consuming, laborious, and expensive.

To overcome this difficulty, unsupervised feature learning methods can be employed to learn latent features from massive unlabeled data. In recent years, several unsupervised feature learning methods have been applied in machinery fault diagnosis, for example, restricted Boltzmann machine (RBM) [13], [14], deep belief network (DBN) [15], [16], auto-encoder (AE), stacked AE (SAE) [17], deep AE (DAE) [18]–[20], sparse DAE [21]–[23], convolutional AE (CAE) [24], and other AE variants [25], [26]. For example, Yang et al. [13] employed an energy-based model, stacked RBM, to capture the system-wide patterns and applied it in wind turbine condition monitoring. Chen et al. [14] proposed an unsupervised feature learning method based on convolutional RBM model for bearing fault diagnosis. Tang et al. [15] proposed an adaptive learning rate DBN and applied it in rotating machinery fault diagnosis. Jia et al. [17] proposed SAE-constructed deep neural networks (DNNs) for machinery fault diagnosis. In this method, machinery fault diagnosis was conducted by firstly pretraining DNNs with massive unlabeled data, and then fine-tuning the model with limited labeled data for classification. Yu [25] proposed a manifold regularized stacked denoising autoencoders algorithm for gearbox fault diagnosis. In this algorithm, gearbox fault diagnosis was performed by an unsupervised feature learning phase and a supervised fine-tuning phase. Most of these unsupervised methods, i.e., DBN, AE and its variants (AEs), adopt the encoder-decoder architecture, which attempt to learn low-dimensional latent representations by reconstructing the unlabeled inputs.

Although DBN and AEs have achieved some success in machinery fault diagnosis, these methods still exhibit some problems. Firstly, DBN and AEs have poor ability to learn effective features for fault diagnosis purpose. By minimizing the reconstruction error, DBN and AEs encourage the encoder to learn the principal components of the inputs. However, the reconstruction accuracy may not be an ideal metric for learning effective features for machinery fault diagnosis. Usually, DBN and AEs are just used to pretrain the DNN model, and the final performance of learned features relies greatly on the succeedent supervised fine-tuning process. Moreover, both DBN and AEs are fully-connected deep neural networks with complex structures. Meanwhile, these methods largely attempt to learn models to accurately approximate the data distribution, which would complicate the learning algorithms. Therefore, these methods often require proper tuning of a large number of hyperparameters.

Sparse filtering (SF) is also a prevalent unsupervised feature learning method, which was proposed by Ngiam et al. [27]. SF is just a two-layer neural network, which is much simpler than DBN and AEs. Therefore, SF can overcome the parameter tuning difficulty and converge easily to an optimal solution [28]. Meanwhile, SF attempts to learn discriminative features by optimizing the sparsity of latent features, instead of learning the principal components of the input data [29]. Since SF is simple and efficient, Lei et al. [30] introduced SF introduced it into machinery fault diagnosis and verified its effectiveness on the famous benchmark bearing dataset. Zhang et al. [31] proposed an unsupervised learning method called general normalized sparse filtering (GNSF), which leverages adjustable normalization parameters to guarantee accuracy. Qian et al. [32] proposed a transfer SF with high-order Kullback-Leibler (KL) divergence to learn discriminative and sharing features of different domains, and then applied it in rotating machinery fault diagnosis under variant working conditions. Even though SF has performed well in these studies, effective feature learning is still a challenging problem. The optimization objective of sparsity encourages SF to learn discriminative features from unlabeled data. However, effective features for machinery fault diagnosis should have excellent clustering performance. They should possess both inter-class variance and intra-class compactness properties, which means that effective features should be discriminative between different classes and gathering within the same class. With just the sparsity constraint, SF could not always learn satisfactory features and achieve desirable machinery fault diagnosis results.

In order to enhance the unsupervised feature learning ability and obtain better diagnosis results, a simple unsupervised feature learning method based on consistency inference-constrained sparse filtering (CICSF) is proposed and applied in machinery fault diagnosis. In CICSF, a natural inference CILRTS is proposed and imposed on SF as an additional constraint, which infers that training data segments generated from the same monitoring time series should possess consistent latent representations in the feature space. CICSF can optimize the inter-class sparsity and intra-class similarity of latent features simultaneously, and thus learn more effective features for machinery fault diagnosis. Then, based on CICSF, a semi-supervised machinery fault diagnosis method is finally developed. After unsupervised feature learning, machinery fault diagnosis is finally realized by a softmax regression classifier trained with limited labeled data. Thanks to the effective features learned by CICSF, the proposed diagnosis method can obtain satisfactory diagnosis results using limited labeled data.

The main contributions of this paper are summarized as follows:
(1) A natural inference CILRTS is proposed to utilize the inherent similarity of unlabeled training instances as latent guidance for unsupervised feature learning.

(2) An unsupervised feature learning method CICSF is derived by introducing CILRTS as an additional constraint to SF, which can optimize the inter-class sparsity and intra-class similarity of latent features simultaneously. The developed CICSF can enhance the unsupervised feature learning performance of SF.

(3) Based on CICSF, a semi-supervised machinery fault diagnosis method is developed, which is mainly consisted of three stages: preprocessing, CICSF-based unsupervised feature learning, and supervised fault classification with limited labeled data. Experimental results on bearing and gearbox datasets validate that the proposed method can learn more effective features from massive unlabeled data and obtain satisfactory diagnosis results with limited labeled data.

The rest of this article is organized as follows. In Section II, fundamental theories of standard SF and softmax regression are briefly introduced. In Section III, the principle of CILRTS and the proposed CICSF-based fault diagnosis method are described in detail. Then, two experimental cases on a bearing dataset and a gearbox dataset are presented to validate the effectiveness and superiority of the proposed method in Section IV and Section V, respectively. Finally, conclusions are drawn in Section VI.

II. THEORIES OF SPARSE FILTERING AND SOFTMAX REGRESSION

A. SPARSE FILTERING

SF is a simple and efficient unsupervised feature learning method, which optimizes exclusively for sparsity in the feature distribution instead of explicitly modeling the data distribution. SF only focuses on three key properties of the feature distribution, i.e., population sparsity, lifetime sparsity, and high dispersal [27]. Population sparsity means that only a few features can be activated for each sample. Lifetime sparsity means that features should be discriminative to distinguish samples. And high dispersal implies that features should have uniform activity distribution. Through optimizing for these three properties, SF attempts to learn desirable latent features from unlabeled data.

As shown in Fig. 1, SF can be viewed as an unsupervised two-layer network. The inputs of the input layer are collected from unlabeled data. The training dataset is denoted as \( \{ x^i \}_{i=1}^M \), where \( x^i \in \mathbb{R}^{N_{in} \times 1} \) represents the ith sample containing \( N_{in} \) data points, and \( M \) is the number of samples. SF can map the input samples onto their features \( f^i \in \mathbb{R}^{N_{out} \times 1} \) through the weight matrix \( W \in \mathbb{R}^{N_{out} \times N_{in}} \), where \( N_{out} \) is the number of samples. \( N_{in} \) is the number of features. The purpose of training SF is to obtain the weight matrix \( W \).

Consider the situation that SF computes linear features for every input sample. The feature vector of the input training sample \( x^i \) can be expressed as

\[
f^i = Wx^i. \tag{1}
\]

Feature vectors of all the training samples can construct a feature matrix \( f \in \mathbb{R}^{N_{out} \times M} \), whose ith column is \( f^i \). In the training of SF, firstly, we need to normalize the feature matrix by rows. Each row (feature) is normalized to be equally active by dividing each feature by its \( l_2 \)-norm across all the samples.

\[
\hat{f}_j = f_j / \| f_j \|_2, \tag{2}
\]

where \( f_j \) denotes the jth row of the feature matrix \( F \).

Then, the feature matrix is further normalized by columns, so that features of every sample can lie on the unit \( l_2 \)-ball.

\[
\hat{j}^i = \hat{j} / \| \hat{j} \|_2, \tag{3}
\]

Finally, the normalized features are optimized for sparsity using the \( l_1 \)-norm. The training loss function of SF can be written as:

\[
L_{SF}(\hat{f}) = \sum_{i=1}^{M} \| \hat{j}^i \|_1. \tag{4}
\]

So, the weight matrix \( W \) can be solved by minimizing the loss function \( L_{SF}(\hat{f}) \).

\[
\min_{W} \sum_{i=1}^{M} \| \hat{j}^i \|_1. \tag{5}
\]

It should be noted that in Equation (4), \( l_1 \)-norm is adopted to measure the sparsity of the normalized features. Actually, there are also some other sparsity metrics, for example, \( l_0 \)-norm and \( l_1/l_2 \)-norm [33]. Anyway, \( l_1 \)-norm is the most commonly-used sparsity measure. So, the term \( \| \hat{j}^i \|_1 \) loss function in Equation (4) can measure the population sparsity of normalized features of the ith sample. Therefore, by optimizing the sparsity of normalized features, SF tends to learn discriminative features from the input samples.

Once the optimal weight matrix \( W \) is obtained, the feature vector of any input sample can be calculated by Equation (1). In addition, SF can be extended to a nonlinear mapping using an activation function [27]. In this way, Equation (1) can be extended as

\[
f^i = \varphi(Wx^i), \tag{6}
\]

where \( \varphi(\cdot) \) represents the activation function. More details of SF can be found in reference [27].
B. SOFTMAX REGRESSION

Softmax regression is usually adopted in the last layer of DNNs for multiclass classification [9], [30]. Thus, soft regression is utilized as the machinery fault classifier after unsupervised feature learning.

Suppose that an input feature set \( \{ f^i \}_{i=1}^{N} \) is given with its label set \( \{ y^i \}_{i=1}^{N} \), where \( N \) is the number of labeled training samples, \( f^i \in \mathbb{R}^{N \times 1}, y^i \in \{ 1, 2, \ldots, L \} \), and \( L \) is the number of class labels. For each input sample \( f^i \), softmax regression tries to estimate the probability of the input sample belonging to each label. Concretely, the probability of \( f^i \) for label \( l \) \((l = 1, 2, \ldots, L)\) can be calculated as

\[
p_l^i = \frac{\exp(\theta_l^T f^i)}{\sum_{k=1}^{K} \exp(\theta_k^T f^i)} \tag{7}
\]

where \( \theta_l^T \) is the transpose of the \( l \)-th vector of the parameter matrix of the model, i.e., \( \theta = [\theta_1, \theta_2, \ldots, \theta_L]^T \).

Then, the soft regression model can be trained by minimizing the cost function below.

\[
J(\theta) = -\frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} \{ y^i = k \} \log(p_k^i) \tag{8}
\]

where \( 1\{\cdot\} \) represents the indicator function, which returns 1 if the condition is true, and 0 otherwise, and \( \theta_{l,n} \) is the \( n \)-th element of \( \theta_l \). By minimizing the cost function, the first term of the cost function aims to minimize the prediction errors between the predicted labels and actual labels. And the second term is the weight decay term, which aims to improve the generalization performance.

After training, the soft regression model can estimate the probability of any test sample belonging to each label. Finally, the label with the largest probability is determined as the predicted label.

III. THE PROPOSED METHOD

The main objective of this work is to enhance the unsupervised mechanical fault feature learning ability and obtain satisfactory fault diagnosis results when limited labeled data are provided. In this section, CILRTS is firstly derived, which aims to learn consistent latent representations from the unlabeled samples within the same class. Then, to improve the feature learning performance, the unsupervised feature learning method CICSF is developed by imposing CILRTS to SF as an additional constraint. Based on CICSF, a semi-supervised machinery fault diagnosis method is finally proposed, which is mainly implemented by unsupervised feature learning based on CICSF and supervised fault classification based on softmax regression. Procedures of the whole semi-supervised fault diagnosis method are presented in detail.

A. CILRTS

In machinery fault diagnosis, while constructing the training dataset, data augmentation has been widely performed by the overlapped strategy with a sliding window. Data augmentation is a label-preserving process, which can ensure the feature distributions of the augmented samples and the original seed samples are consistent. It is extremely probable that training data instances which are segmented through the sliding window from a relatively short period of monitoring time series belong to the same machinery condition. Inspired by this principle, a simple and natural inference can be proposed that data segments which are generated from the same monitoring time series should possess consistent latent representations in the feature space. This inference is termed as consistent inference of latent representation for time series (CILRTS).

As illustrated in Fig. 2, suppose that two training samples are vibration time series acquired under two different health conditions, which are separately denoted as fault 1 and fault 2. Every training sample is divided into two segments using a sliding window. Next, all the segments are projected into a latent feature space. So, latent representations of two segments from fault 1 (triangles) are naturally consistent in the latent space. Similarly, latent representations of two segments from fault 2 (pentagons) should be also consistent. The consistency can be measured by the distance between latent representations in the learned feature space.

CILRTS can be utilized as a guideline in unsupervised feature learning by minimizing the distance of learned features. Mathematical descriptions of CILRTS are given below.

For a time-domain signal \( s^k = \{ s_1, s_2, s_3, \ldots, s_K \} \), the overlapped data augmentation strategy is employed to divide the time series \( s^k \) into a segment group \( x^k = \{ x_i^k \}_{i=1}^{M} \), where \( s_i \) is a data point of the time series and \( x_i^k \) represents the \( i \)-th example generated from the seed signal \( s^k \). In supervised learning, classification function \( g(\cdot) \) is used to map data point \( x^k \) to label \( y^k \), which can be expressed as

\[
y^k = g(f(x_i^k)) = g(f(x_i^k)), \tag{9}
\]

where \( f(\cdot) \) is a map function from raw data to feature space. However, in unsupervised learning, data label \( y^k \) is unknown, and a function \( f(\cdot) \) is expected to map data points \( x_i^k \) to latent representation space.

According to CILRTS, since \( x_i^k \) and \( x_j^k \) are generated from the same time series \( s^k \), their feature representations should satisfy the following equation:

\[
f(x_i^k) = f(x_j^k). \tag{10}
\]
Due to the presence of ambient noise and other interferences, the equation is too ideal to achieve. So, it can be optimized through measuring the similarity of $x_i^k$ and $x_j^k$ in terms of distance $f(x_i^k, f(x_j^k))$. Then, the similarity of all training examples can be optimized by the following loss function:

$$L_D = \sum_{k=1}^{K} \sum_{i=1}^{M} \text{distance}(f(x_i^k), f(x_j^k))$$  \hspace{1cm} (11)

where $f(x_i^k)$ represents the expected latent representation of $x_i^k$, and the distance function $d(f(x_i^k), f(x_j^k))$ measures the similarity between $f(x_i^k)$ and $f(x_j^k)$ in the latent space. Both Euclidean distance and Kullback-Leibler (KL) divergence can be adopted as the distance function. In this paper, the Euclidean distance is adopted.

In order to simplify the calculation process, $f(x_i^k)$ can be used as a reference representation to replace the unknown expected latent representation. In this way, Equation (11) can be simplified as

$$L_D = \sum_{k=1}^{K} \sum_{i=2}^{M} \text{distance}(f(x_i^k), f(x_j^k)).$$ \hspace{1cm} (12)

Smaller loss value in Equation (12) indicates higher similarity. By minimizing the loss function $L_D$ in Equation (12), CILRTS encourages an unsupervised neural network to learn the similarity of samples within the same class, and thus improve the clustering performance of learned features.

B. CICSF-BASED MECHANICAL FAULT FEATURE LEARNING AND DIAGNOSIS

A machinery fault diagnosis method is developed based on the improved unsupervised feature learning method CICSF which introduces CILRTS into the standard SF. Just like other intelligent fault diagnosis methods, the developed fault diagnosis method contains two processes: offline training and online fault diagnosis. The illustration of the training process of the proposed CICSF-based feature learning and fault diagnosis method is displayed in Fig. 3. As shown in Fig. 3, the proposed machinery fault diagnosis method is mainly consisted of three stages: preprocessing, CICSF-based unsupervised feature learning, and fault classification. Procedures of these stages are described in detail as follows.

1) PREPROCESSING

In the training process, the whole training dataset $X$ contains two kinds of data: massive unlabeled samples and limited labeled samples.

The whole training dataset is utilized to train the unsupervised feature learning model CICSF without the involvement of data labels. In this situation, the training dataset without labels can be denoted as $X = \{x_i^k\}_{k=1}^{N_1}$, which has $N_1$ unlabeled samples. Suppose that the input dimension of the feature learning model CICSF is $N_{in}$, and the output dimension $N_{out}$. In order to prepare for CICSF training, each sample $x^k$ is segmented into a group of training instances $\{x_i^k\}_{i=1}^{M}$ by the overlapped strategy with a sliding window, where $x_i^k \in \mathbb{R}^{N_{out} \times 1}$ represents the $i^{th}$ segment generated from the seed sample $x^k$, and $M$ is the number of segments obtained from the same sample. So, the unlabeled training segment set can be denoted as $S_1 = \{x_i^k\}_{i=1}^{M} \times \{y_i^k\}_{k=1}^{N_1}$. The total number of segments in $S_1$ is $N_{S1}$, and $N_{S1} = M \times N_1$.

The labeled training subset in the whole training dataset is utilized to train the softmax regression model. Suppose the labeled subset is denoted as $X_2 = \{x_i^k, y_i^k\}_{k=1}^{N_2}$, where $y_i^k$ is the corresponding health condition label of $x_i^k$, and the number of labeled samples $N_2$ is often much smaller than $N_1$. After preprocessing, the labeled training segment set $S_2 = \{\{x_i^k\}_{i=1}^{M}, y_i^k\}_{k=1}^{N_2}$ can be obtained.

2) UNSUPERVISED FEATURE LEARNING BY CICSF

In machinery fault diagnosis, effective features should have a clustering distribution in the latent space, which possess similarity within the same class and discrimination between different classes. To improve the feature learning ability, CILRTS is introduced into SF as an additional constraint. The developed unsupervised feature learning method is called as consistency inference-constrained SF (CICSF).

In the training process, the preprocessed unlabeled training segment set $S_1$ is fed into the CICSF model. Firstly, the weight matrix $W$ is initialized by a normal distribution. Then, $W$ is updated iteratively. In this process, sigmoid function is adopted as the nonlinear activation function in Equation (6). The feature matrix $F$ of the input training dataset can be calculated as:

$$F_1 = \text{sigmoid}(WS_1) = \frac{1}{1 + e^{-WS_1}},$$ \hspace{1cm} (13)

where $F_1 \in \mathbb{R}^{N_{out} \times N_{S1}}$. The feature matrix $F$ is normalized by Equations (2) and (3). During optimization, the similarity loss $L_D$ of CILRTS is fused into the original sparsity loss $L_{SF}$ of SF as an additional constraint. The complete loss function of CICSF can be written as:

$$L_1 = L_{SF} + \lambda L_D,$$ \hspace{1cm} (14)

where $L_{SF}$ is defined in Equation (4), $L_D$ is defined in Equation (12), and $\lambda$ is a tradeoff parameter. By minimizing
the loss function $L_1$, the optimal weight matrix $W$ can be obtained, and the CICSF feature learning model is constructed consequently. The minimization problem can be solved with the L-BFGS algorithm. In this way, the CICSF model can optimize both the inter-class sparsity and intra-class similarity of latent features simultaneously.

Then, with the optimized weight matrix $W \in \mathbb{R}^{N_{out} \times N_{in}}$, the segments group $\{x^i_k\}_{k=1}^M$ of every training sample $x_k$ in $S_2$ can be mapped into a feature set $\{f^i_k\}_{k=1}^M$. In order to obtain more robust features, a feature averaging procedure can be adopted. The learned feature vector $f^i_k$ of $x_k$ is calculated by averaging the learned features of $M$ segments, which can be written as

$$f^i_k = \frac{1}{M} \sum_{i=1}^M f^i_k.$$  

(15)

It should be noted that in Equation (15), $M$ can also be set to other values. For example, in order to increase the number of learned features, $M/2$ or other smaller values can be adopted as the averaging time.

After feature averaging, the learned feature matrix of the labeled training subset can be obtained as $F_2 = \{f^i_k\}_{i=1}^{N_2}$, where $F_2 \in \mathbb{R}^{N_{out} \times N_2}$. $F_2$ is prepared for training the softmax regression classifier in the next step.

3) MACHINERY FAULT DIAGNOSIS

After unsupervised feature learning, machinery fault diagnosis is conducted by softmax regression. In the training process, softmax classifier is trained with the learned features of the labeled training samples. Once the feature set of $S_2$, namely $F_2 = \{f^i_k\}_{i=1}^{N_2}$, is obtained, $F_2$ can be combined with the corresponding label set $\{y^i\}_{i=1}^{N_2}$ to train softmax regression. Through minimizing the cost function in Equation (8), the soft regression model can be trained.

Finally, in the testing process, the testing dataset $Z = \{z^i\}_{i=1}^{N_{test}}$ is segmented following the preprocessing step, where $N_{test}$ is the number of testing samples. Then the preprocessed testing segments are fed into the trained CICSF-based unsupervised feature learning model. After feature averaging, the learned feature set $F_3 = \{f^i_k\}_{i=1}^{N_{test}}$ can be obtained. In the end, using $F_3$ as the input of the trained softmax regression classifier, the probability of which label the feature $f^i_k$ belongs to, can be calculated by Equation (7). Consequently, health conditions of the testing samples can be evaluated.

IV. CASE STUDY I: BEARING FAULT DIAGNOSIS WITH THE PROPOSED METHOD

A. DATA DESCRIPTION

In order to verify the effectiveness of the proposed method, the motor bearing vibration signals provided by Case Western Reserve University (CWRU)[34] are analyzed in this section. As shown in Fig. 4, the test rig mainly consists of a 2 HP motor, a torque sensor, dynamometer, and control electronics.

Vibration signals were collected from the drive end of the test-bed by the acceleration sensor under four different bearing conditions, including normal condition (NC), outer race fault (OF), inner race fault (IF), and rolling ball fault (BF). And each of the three fault types had three different severity levels, i.e., 0.18mm, 0.36mm, and 0.53 mm. So, the whole dataset contains ten health conditions under four different loads (0, 1, 2, 3 hp). The sampling frequency is 12 kHz. For brevity, the outer race fault with severity level 0.18mm is abbreviated as OF18 in this paper. The other nine health conditions are also denoted in this way. There are 110 samples for each health condition under one load. So, the whole dataset contains 4400 samples, and each sample contains 2000 data points.

In the preprocessing step of the proposed method, 25% of samples in the whole dataset are utilized as the training dataset, which are selected randomly and evenly from every health condition under each load. The rest of the samples are used for testing. In the training dataset, just limited samples are utilized as labeled samples, while the majority of training samples are utilized as unlabeled data. Then, every sample is randomly divided into 20 data segments with a sliding window. Each segment contains 800 data points. Therefore, there are 22000 segments in the training dataset.

B. PARAMETER SELECTION OF THE PROPOSED METHOD

There are mainly three tunable hyperparameters in the proposed method, i.e., the input dimension $N_{in}$ and the output dimension $N_{out}$ of the CICSF model, and the tradeoff parameter $\lambda$ in Equation (14).

Firstly, the input dimension $N_{in}$ is investigated, which is equal to the length of the data segments. In order to analyze the influence of the input dimension, $N_{in}$ varies from 200 to 1000. Meanwhile, let $N_{out} = N_{in}/4$, and the tradeoff parameter $\lambda$ is set to 0.03. The diagnosis results using different output dimensions are shown in Fig. 5. It can be seen that with the increase of the input dimension $N_{in}$, the training and testing accuracies increase gradually. When $N_{in}$ reaches to 800, the training and testing accuracies are above 99% and then keep steady. Anyway, given limited samples, when $N_{in}$ is too large, the number of data instances will be small. Meanwhile, larger input dimension will increase the dimension of the weight matrix $W$, which would cost more computing time. For these reasons, in this experiment, the input dimension of the CICSF model is set to 800.

FIGURE 4. Bearing test rig of CWRU.
Secondly, the output dimension $N_{\text{out}}$ is investigated. To this end, $N_{\text{out}}$ varies from 100 to 300 with an increment of 50, while other parameters are kept to be constant. The tradeoff parameter $\lambda$ is still set to 0.03. The diagnosis results using different output dimensions are shown in Fig. 6. It can be seen that all the training accuracies and testing accuracies are over 98%. When $N_{\text{out}}=100$, the training and testing accuracies are relatively low. As $N_{\text{out}}$ increases, the training and testing accuracies become higher gradually and then keep steady after $N_{\text{out}}$ is larger than 200. In addition, the larger $N_{\text{out}}$ is, the more parameters the weight matrix $W$ will contain. So, in order to strike a balance between the diagnosis accuracy and the computation cost, the output dimension $N_{\text{out}}$ is chosen as 200 in this paper.

Secondly, the tradeoff parameter $\lambda$ in the loss function of CICSF is investigated. In this investigation, $N_{\text{out}}$ is set as 200. The diagnosis results using different tradeoff parameters are displayed in Fig. 7. When $\lambda$ is too large, for example, $\lambda = 0.5$, the diagnosis accuracy is very low. That’s because large $\lambda$ makes the CICSF model difficult to optimize for sparsity. In contrast, when $\lambda$ is too small, the effect of the additional CILRTS constraint is too small. Then, the proposed CICSF model would degrade into standard SF. As shown in Fig. 7, when $\lambda =0.03$, the diagnosis accuracy becomes the highest value. So, the tradeoff parameter between the sparsity loss and the similarity loss is set as 0.03.

C. FAULT DIAGNOSIS RESULTS OF THE PROPOSED METHOD

After preprocessing, two steps of the proposed method are performed, i.e., unsupervised feature learning and fault classification. While updating the weight matrix $W$ of the CICSF model, the loss values of the training dataset and the testing dataset are calculated as the number of iterations increases. Fig. 8 shows the training and testing loss curves. It can be seen that as the number of iterations increase, the training and testing loss curves decrease gradually and tend to converge when the epoch number reaches 100. So, the number of iterations in the training of CICSF model is set to 100. Other important parameters are selected according to the above section, i.e., $N_{\text{out}}=200$, $\lambda = 0.03$. It should be noted that in order to reduce the effect of randomness, ten trials are conducted for all the following experiments.

Then, the softmax regression model is trained with a subset of labeled samples, which are selected randomly from the whole training dataset. When the number of labeled samples is 100, the fault diagnosis accuracies in the ten trials are all over 98%, and the average accuracy is 98.97%. Moreover, the standard deviation of the diagnosis accuracies is 0.058, which means that the diagnosis results of the
proposed method are stable. To show the detailed diagnosis results of the proposed method, the confusion matrix of the average testing accuracies in ten trials is shown in Fig. 9.

As shown in Fig. 9, most of the bearing health conditions are detected correctly with the proposed method. Just a few of testing samples of OF54 are misclassified as IF54 or BF18. These results demonstrate that the proposed method can realize bearing fault diagnosis effectively with a combination of limited labeled data and sufficient unlabeled data.

D. COMPARISONS WITH OTHER RELATED METHODS

In order to verify the superiority of the proposed fault diagnosis method based on CICSF, the proposed method is compared with other related methods using the same bearing dataset. The related unsupervised learning methods used for comparison include standard SF [31], deep auto-encoder (DAE), sparse deep auto-encoder (SDAE) [20], and convolutional auto-encoder (CAE) [23]. The structural parameters, main hyperparameters, and optimization algorithms used in the proposed method and the above comparative methods are briefly summarized in Table 1.

In Table 1, structural parameters ‘800-200’ of the proposed method and SF indicate the input and output dimensions of the basic SF feature learning network. A simple off-the-shelf minimizer L-BFGS is utilized in these two methods. Structural Parameters of DAE and SDAE denote the number of layers in the networks and the dimension of every layer. For the last method CAE, the structural parameters in the convolutional layer indicate the size of the convolutional kernels and the number of kernels. The padding method chosen in TensorFlow in the convolutional layer is ‘SAME’. Moreover, max pooling is adopted in the pooling layers. The activation function used in layers conv1~deconv5 is Relu, and sigmoid function is employed in the deconv6 layer. In DAE, SDAE, and CAE, Adam gradient optimization algorithm is adopted to adjust the weights of the network, and the learning rates used in these methods are also listed. In Table 1, just the parameters in the feature learning networks are listed.

For fault classification, softmax regression models are adopted in all of these methods. The output dimensions of the softmax regression models are all set to 10, which is the same as the number of bearing health conditions. So, parameters of softmax classifiers in these methods are omitted for brevity in Table 1. It should be noted that in order to provide a reliable comparison, other parameters in these methods are also adjusted to be optimal values.

Comparisons between these methods are mainly focused on two aspects: the unsupervised feature learning performance, and the fault diagnosis results, which are detailed as follows.

1) COMPARISONS OF THE UNSUPERVISED FEATURE LEARNING PERFORMANCE

Unsupervised feature learning is a vital step for machinery fault diagnosis. To a large extent, the diagnosis result is dependent on the effectiveness of the learned features. Therefore, the unsupervised feature learning performance of the proposed method is firstly compared with those of other related methods.

Besides, as stated above, the feature learning ability of AEs relies greatly on the supervised fine-tuning process. In order to analyze this problem, in DAE, SDAE, and CAE, two conditions are considered, i.e., whether supervised fine-tuning is utilized or not. Thus, eight methods are utilized in this comparison, which are described in Table 2. For brevity,
TABLE 2. Descriptions of the methods used for comparison.

| Method     | Description             | Usage of data labels |
|------------|-------------------------|----------------------|
| 1          | The proposed CICSF      | Not used             |
| 2          | SF                      | Not used             |
| 3          | DAE                     | Used in fine-tuning  |
| 4          | DAE without FT          | Not used             |
| 5          | SDAE                    | Used in fine-tuning  |
| 6          | SDAE without FT         | Not used             |
| 7          | CAE                     | Used                 |
| 8          | CAE without FT          | Not used             |

In the following descriptions of these methods, ‘fine-tuning’ is abbreviated as ‘FT’.

In methods 3, 5, 7, AEs are pretrained with unlabeled data, and then softmax classifier is added to the network for fine-tuning the whole network with labeled training data. It should be noted that all the 1100 samples are utilized as labeled samples in the fine-tuning process. After global fine-tuning, the learned features are utilized for comparison. In this sense, methods 3, 5, 7 are not strict unsupervised feature learning methods. In methods 4, 6, 8, only the pretraining process is adopted without fine-tuning. In this experiment, 25% of the samples in the whole dataset, i.e., 1100 samples, are utilized to train the feature learning models, without the involvement of data labels.

In order to illustrate the performance of feature learning intuitively, the learned features are visualized by t-distributed stochastic neighbor embedding (t-SNE) [35]. Latent features learned by every method are mapped into a two-dimensional scatter diagram by t-SNE. The feature visualization results of the above eight methods are shown in Fig. 10. The color legends 1-10 indicate labels of the ten bearing health conditions, i.e., NC, OF18, OF36, OF53, IF18, IF36, IF53, BF18, BF36, and BF53. The x-axis and y-axis represent the first and second feature mapped feature dimensions by t-SNE.

Firstly, the proposed CICSF method is compared with standard SF. In Fig. 10(a), it can be seen that the latent features learned by CICSF are obviously distributed into ten clusters. This means that the CICSF model can learn discriminative features from the unlabeled training samples. By contrast, as shown in Fig. 10(b), the distribution of features learned by standard SF can also be roughly divided into 10 clusters, but some feature points would be separated into the wrong clusters. So, the proposed CICSF method improves the unsupervised feature learning performance of standard SF.

Secondly, according to the comparisons between Fig. 10(c) and (d), Fig. 10(e) and (f), Fig. 10(g) and (h), it can be found that the feature learning performance of AEs is dependent on the global fine-tuning process. As shown in Fig. 10(d), (f), (h), without fine-tuning, the feature visualization results of methods 4, 6, 8 are in a mess, and only features under the normal condition (yellow) can be separated in these figures. These comparison results show that without the aid of labeled data in fine-tuning, AEs cannot learn effective features from unlabeled data for classification tasks.

Thirdly, the feature visualization result of the proposed CICSF method is also better than methods 3-8. Although learned features of methods 3, 5, 7 show acceptable clustering quality in Fig. 10(c), (e), and (g), labeled data are required for fine-tuning in these methods.

In conclusion, from the above comparisons, it can be seen that the unsupervised feature learning performance of the proposed CICSF method outperforms those of other commonly used unsupervised feature learning methods, including the standard SF and AEs.

FIGURE 10. The feature visualization results of different methods using the bearing dataset.
does not need to be updated. In fault diagnosis methods based on DAE, SDAE, and CAE, the whole network should be tuned using the labeled samples.

As stated above, 1100 unlabeled samples have been adopted to train the feature learning model. And the rest of samples are used as the testing dataset. The purpose of the experiment in this section is to compare the diagnosis performances of the methods with limited labeled samples. Provided that the number of labeled samples in the whole training dataset is 100, the average testing accuracies in 10 trials and their standard deviations of the aforementioned diagnosis methods are listed in Table 3.

| Method | Description | No. of labeled samples | Testing accuracy (%) | Standard deviation |
|--------|-------------|------------------------|----------------------|-------------------|
| 1      | CICSF       | 100                    | 98.97                | 0.058             |
| 2      | SF          | 100                    | 96.53                | 0.109             |
| 3      | DAE         | 100                    | 87.11                | 0.947             |
| 4      | SDAE        | 100                    | 88.37                | 0.818             |
| 5      | CAE         | 100                    | 89.74                | 0.706             |

In Table 3, the proposed method obtains the highest average testing accuracy of 98.97%, which enhances the accuracy of standard SF by over 2%. Meanwhile, the standard deviation of accuracies in 10 trials is also the lowest one. It should be noted that the testing accuracy of the standard SF is lower than the accuracy obtained in Ref. [30]. The reason is that just 100 labeled samples are utilized to train the softmax regression classifier in this experiment, which is just about 1/10 of the whole training dataset. While in reference [30], all the samples in the training dataset are used as labeled training samples. As to AEs, the diagnosis result of CAE is a little better than DAE and SDAE. However, it can be seen that the diagnosis accuracies of DAE, SDAE, and CAE are all below 90%, and the standard deviations are all much higher than the proposed method and standard SF. The lower diagnosis accuracies of AEs in this experiment are also due to the unsatisfactory feature learning ability when just limited number of labeled training samples are provided. These results also prove that the effectiveness of AEs greatly relies on the fine-tuning process, where sufficient labeled samples are required. On the whole, the proposed method is superior to the standard SF and AEs, when the number of labeled samples is small.

In order to analyze the influence of the number of labeled samples, diagnosis results of different methods with various numbers of labeled training samples are shown in Fig. 11. The sizes of the labeled training subset are set as 50, 100, 150, 200, 300, respectively.

It can be seen that the fault diagnosis accuracies increase as the number of labeled training samples increase. When the number of labeled samples reaches to 300, the diagnosis accuracies of the proposed CICSF-based method and standard SF tend to be close to 100%. In contrast, diagnosis accuracies of DAE, SDAE, and CAE are still about 3% lower than the proposed method, which implies that more labeled training samples are needed. However, in practical applications, it is difficult to obtain massive labeled data, and most of the monitoring data are unlabeled. When the number of labeled samples is very small, for example, 50, the diagnosis accuracy of the proposed method is much higher than those of other methods. The proposed method can still achieve an accuracy of 94.5%, while other methods, especially AEs, cannot diagnosis the bearing health conditions accurately. That is because in the unsupervised feature learning stage, the proposed CICSF feature learning model can learn more effective features from massive unlabeled data. As a result, even if just limited labeled samples are provided, the proposed method can still obtain acceptable diagnosis results.

V. CASE STUDY II: GEARBOX FAULT DIAGNOSIS WITH THE PROPOSED METHOD

A. DATA DESCRIPTION

To further validate the effectiveness of the proposed method, a gearbox vibration dataset [36] is investigated. Vibration signals were collected from the drivetrain dynamic simulator (DDS) shown in Fig. 12. The rotating speed is 1200rpm.

There are five different health conditions in the gearbox dataset, including normal condition (NC), chipped, miss, root, surface. The description of these five health conditions and the dataset are listed in Table 4. Under every health condition, there are 880 samples, and each sample contains...
TABLE 4. Description of the gearbox dataset.

| Label | Health condition | NO. of samples | Description of the health condition |
|-------|------------------|----------------|-------------------------------------|
| 1     | NC               | 880            | Normal condition                    |
| 2     | Chipped          | 880            | Crack occurs in the gear feet       |
| 3     | Miss             | 880            | Missing one of feet in the gear     |
| 4     | Root             | 880            | Crack occurs in the root of gear    |
| 5     | Surface          | 880            | Wear occurs in the surface of gear  |

2000 data points. Then, in the preprocessing step, each sample is randomly divided into 20 data segments by the overlapped strategy, and each training segment contains 800 data points. So, there are 4400 samples in the entire dataset. And 25% of these samples in every health condition are randomly selected as the training dataset. The rest of samples are utilized as the testing dataset.

B. FAULT DIAGNOSIS RESULTS AND COMPARISONS

In this experiment, the parameter analysis results are similar to those in the previous bearing fault diagnosis case, which are omitted for brevity. So, parameter configurations of the proposed method and other related methods are just the same as those in Table 1.

Firstly, the feature learning results of the proposed CICSF method and other related methods in Table 2 are visualized and compared in Fig. 13. The color legends 1-5 denote the labels of the five health conditions in Table 4, i.e., NC, Chipped, Miss, Root, and Surface, respectively. The x-axis and y-axis represent the first and second feature mapped feature dimensions by t-SNE.

It can be seen that mapped features of both CICSF and standard SF can be clearly divided into five clusters. Since there are only five health conditions, the feature learning and fault diagnosis tasks are relatively simple. As a result, the difference between features learned by CICSF and SF is not easy to distinguish directly through the feature visualization results. The feature visualization result of CAE in Fig. 13(g) is also acceptable. However, labeled samples are involved in this method. Anyway, the clustering performance of learned features of CICSF in Fig. 13(a) is still much better than DAE, SDAE and CAE without FT. Meanwhile, the feature visualization results of DAE, SDAE and CAE after fine-tuning shown in Fig. 13(c), (e), (g) are superior to those learned by these methods without fine-tuning. It should be noted that in Fig. 13(c), (e), and (g), all the 1100 samples in the training dataset are utilized as labeled samples in the fine-tuning processes in DAE, SDAE and CAE. Since sufficient labeled samples are utilized in fine-tuning, features learned by AEs can be divided into ten categories, especially in Fig. 13(g). By contrast, in Fig. 13(d), (f), (h), just mapped features of NC (orange) and Root (red) can be separated, while features of other three similar gear defect conditions are mixed together. These comparisons demonstrate that the feature learning performance of AEs greatly relies on the fine-tuning process, where sufficient labeled data are required. In contrast, the proposed CICSF unsupervised feature learning method can learn discriminative features from unlabeled data.

Secondly, 100 samples are randomly selected as labeled samples from the total 1100 samples in the training dataset. These 100 labeled samples are utilized to train the softmax regression model. Then, 3300 testing samples are fed into the trained networks. The average testing accuracies and standard deviations in 10 trials using different methods are given in Table 5. It can be seen that although just 100 labeled samples are provided, the proposed method can still achieve 99.3% testing accuracy, which is 2% higher than standard SF. In comparison, diagnosis accuracies of DAE, SDAE, and CAE are much lower than the proposed method, which is because of the small number of labeled training samples used in this experiment. Meanwhile, the standard deviation of the proposed method is also the lowest one, which indicates the stability of the proposed method.
TABLE 5. Gearbox fault diagnosis results using different methods.

| Method | Description | No. of labeled samples | Testing accuracy (%) | Standard deviation |
|--------|-------------|------------------------|----------------------|--------------------|
| 1      | CICSF       | 100                    | 99.30                | 0.013              |
| 2      | SF          | 100                    | 97.30                | 0.109              |
| 3      | DAE         | 100                    | 89.16                | 0.253              |
| 4      | SDAE        | 100                    | 88.39                | 0.144              |
| 5      | CAE         | 100                    | 90.72                | 0.150              |

To analyze the influence of the number of labeled samples, different numbers of labeled training samples are provided in the above methods. Gearbox fault diagnosis results of the above methods with various numbers of labeled samples are shown in Fig. 14.

![figure14.png](image-url)

FIGURE 14. Gearbox fault diagnosis results of different methods with various numbers of labeled training samples.

It can be seen that, as the number of labeled training samples increases, the fault diagnosis accuracies of different methods increase in the beginning and tend to be stable in the end. In general, due to the excellent unsupervised feature learning ability, the proposed CICSF-based diagnosis method can achieve higher diagnosis accuracies than the standard SF, DAE, SDAE, and CAE, especially when the number of labeled samples is less than 200. In addition, compared with the proposed CICSF-based method and standard SF, AEs need much more labeled samples to obtain comparative diagnosis results.

These aforementioned diagnosis results and comparisons validate the effectiveness and superiority of the proposed method in unsupervised feature learning and fault diagnosis, in case that limited labeled samples and massive unlabeled samples are provided.

VI. CONCLUSION

In machinery fault diagnosis, it is often difficult to acquire a large number of labeled monitoring data, while massive unlabeled data is often available. Under this condition, it is important to extract effective features from unlabeled data for machinery fault diagnosis. In this paper, a simple and effective unsupervised feature learning method CICSF is proposed, which enhances the feature learning performance of SF by introducing an additional constraint termed as CILRTS. CICSF can both optimize the intra-class similarity and inter-class sparsity of the feature distribution, and thus learn more effective features for classification tasks. Based on CICSF, a three-stage machinery fault diagnosis method is finally developed, which aims to obtain satisfactory diagnosis results using limited labeled samples.

Experimental results on a bearing dataset and a gearbox dataset both verify the effectiveness of the proposed method. Furthermore, comparisons with other related methods, including SF, DAE, SDAE, and CAE, demonstrate that the proposed method can improve the clustering performance of learned features and obtain higher diagnosis accuracy with limited labeled samples. Therefore, the proposed method could be a promising approach for machinery fault diagnosis, especially when the number of labeled samples is small while massive unlabeled data is available. Meanwhile, the proposed CICSF unsupervised learning method also has the potential to be applied in clustering tasks. However, there is still some work to do to further optimize it. For example, in this study, the number of unlabeled samples under every health condition is assumed to be almost equal. In the future, the problem of imbalanced fault classification should be furtherly studied.

ACKNOWLEDGMENT

The authors are very grateful to Prof. Ruqiang Yan and his team for their support of the gearbox vibration dataset.

REFERENCES

[1] P. Gangsar and R. Tiwari, “A support vector machine based fault diagnostics of induction motors for practical situation of multi-sensor limited data case,” Measurement, vol. 135, pp. 694–711, Mar. 2019.
[2] H. Jiang, J. Chen, and G. Dong, “Hidden Markov model and nuisance attribute projection based bearing performance degradation assessment,” Mech. Syst. Signal Process., vol. 72–73, pp. 184–205, May 2016.
[3] R. Wang, C. Wang, J. Li, and W. Lu, “An intelligent fault diagnosis method of rolling element bearing based on acoustic imaging and Gabor wavelet transform,” in Proc. 25th Int. Congr. Sound Vib., Hiroshima Calling (ICSV), Hiroshima, Japan: IIAV, vol. 5, Jul. 2018, pp. 2684–2691.
[4] H. Jiang, J. Yuan, Q. Zhao, H. Yan, S. Wang, and Y. Shao, “A robust performance degradation modeling approach based on Student’s t-HMM and nuisance attribute projection,” IEEE Access, vol. 8, pp. 49629–49644, 2020.
[5] L. Wen, X. Li, L. Gao, and Y. Zhang, “A new convolutional neural network-based data-driven fault diagnosis method,” IEEE Trans. Ind. Electron., vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
[6] D. Peng, Z. Liu, H. Wang, Y. Qin, and L. Jia, “A novel deeper one-dimensional CNN with residual learning for fault diagnosis of wheelset bearings in high-speed trains,” IEEE Access, vol. 7, pp. 10278–10293, 2019.
[7] S. Wang and J. Xiang, “A minimum entropy deconvolution-enhanced convolutional neural networks for fault diagnosis of axial piston pumps,” Soft Comput., vol. 24, no. 4, pp. 2983–2997, Feb. 2020.
[8] H. D. Shao, J. S. Cheng, H. K. Jiang, Y. Yang, and Z. T. Wu, “Enhanced deep gated recurrent unit and complex wavelet packet energy moment entropy for early fault prognosis of bearings,” Knowl.-Based Syst., vol. 188, p. 14, Jan. 2020, Art. no. 105022.
[9] Z. Chen, A. Mauricio, W. Li, and K. Gryllias, “A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks,” Mech. Syst. Signal Process., vol. 140, Jun. 2020, Art. no. 106683.
[10] R. Wang, F. Liu, F. Hou, W. Jiang, Q. Hou, and L. Yu, “A non-contact fault diagnosis method for rolling bearings based on acoustic imaging and convolutional neural networks,” IEEE Access, vol. 8, pp. 132761–132774, 2020.
[11] H. Shao, M. Xia, G. Han, Y. Zhang, and J. Wan, “Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer CNN and thermal images,” IEEE Trans. Ind. Informat., early access, Jun. 30, 2020, doi: 10.1109/TII.2020.3005965.

[12] 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.

[13] W. Yang, C. Liu, and D. Jiang, “An unsupervised spatiotemporal graphical modeling approach for wind turbine condition monitoring,” Renew. Energy, vol. 127, pp. 230–241, Nov. 2018.

[14] L. Chen, G. Xu, Y. Wang, and J. Wang, “Detection of weak transient signals based on unsupervised learning for bearing fault diagnosis,” Neurocomputing, vol. 314, pp. 445–457, Nov. 2018.

[15] S. Zhang, C. Shen, D. Wang, S. Li, W. Huang, and Z. Zhu, “Adaptive deep feature learning network with nesterov momentum and its application to rotating machinery fault diagnosis,” Neurocomputing, vol. 305, pp. 1–14, Aug. 2018.

[16] W. Qian, S. Li, J. Wang, and Q. Wu, “A novel supervised sparse feature auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples,” IEEE Access, vol. 7, pp. 115368–115377, 2019.

[17] H. Zhu, J. Cheng, C. Zhang, J. Wu, and X. Shao, “Stacked pruning sparse denoising autoencoder based intelligent fault diagnosis of rolling bearings,” Appl. Soft Comput., vol. 88, Mar. 2020, Art. no. 106060.

[18] M. Sun, H. Wang, P. Liu, S. Huang, and P. Fan, “A sparse stacked denoising autoencoder with optimized transfer learning applied to the fault diagnosis of rolling bearings,” Measurement, vol. 146, pp. 305–314, Nov. 2019.

[19] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 818–833.

[20] J.-B. Yu, “Evolutionary manifold regularized stacked denoising autoencoders for gearbox fault diagnosis,” Knowl.-Based Syst., vol. 178, pp. 111–122, Aug. 2019.

[21] 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,” Eng. Appl. Artif. Intell., vol. 76, pp. 170–184, Nov. 2018.

[22] J. Ngiam, P. W. Koh, Z. Chen, S. Bhaskar, and A. Y. Ng, “Sparse filtering,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1125–1133.

[23] F. M. Zennaro and K. Chen, “Towards understanding sparse filtering: A theoretical perspective,” Neural Netw., vol. 98, pp. 154–177, Feb. 2018.

[24] W. Qian, S. Li, J. Wang, and Q. Wu, “A novel supervised sparse feature extraction method and its application on rotating machinery fault diagnosis,” Neurocomputing, vol. 320, pp. 129–140, Dec. 2018.

[25] Y. Lei, F. Jia, J. Lin, S. Xing, and S. Xing, “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.

[26] Z. Zhang, S. Li, J. Wang, Y. Xin, and Z. An, “General normalized sparse filtering: A novel unsupervised learning method for rotating machinery fault diagnosis,” Mech. Syst. Signal Process., vol. 124, pp. 596–612, Jun. 2019.

[27] W. Qian, S. Li, and J. Wang, “A new transfer learning method and its application on rotating machinery fault diagnosis under variant working conditions,” IEEE Access, vol. 6, pp. 69907–69917, Aug. 2018.

[28] N. Hurley and S. Rickard, “Comparing measures of sparsity,” IEEE Trans. Inf. Theory, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[29] Case Western Reserve University (CWRU) Bearing Data Center. Accessed: Sep. 2019. [Online]. Available: https://csegroups.case.edu/bearingdatacenter/pages/download-data-file

[30] L. Chen, G. Xu, Y. Wang, and J. Wang, “Detection of weak transient signals based on unsupervised learning for bearing fault diagnosis,” Neurocomputing, vol. 314, pp. 445–457, Nov. 2018.

[31] S. Zhang, C. Shen, D. Wang, S. Li, W. Huang, and Z. Zhu, “Adaptive deep feature learning network with nesterov momentum and its application to rotating machinery fault diagnosis,” Neurocomputing, vol. 305, pp. 1–14, Aug. 2018.

[32] W. Qian, S. Li, J. Wang, and Q. Wu, “A novel supervised sparse feature auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples,” IEEE Access, vol. 7, pp. 115368–115377, 2019.

[33] S. Zhang, C. Shen, D. Wang, S. Li, W. Huang, and Z. Zhu, “Adaptive deep feature learning network with nesterov momentum and its application to rotating machinery fault diagnosis,” Neurocomputing, vol. 305, pp. 1–14, Aug. 2018.

[34] W. Qian, S. Li, J. Wang, and Q. Wu, “A novel supervised sparse feature auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples,” IEEE Access, vol. 7, pp. 115368–115377, 2019.

[35] L. van der Maaten and G. Hinton, “Visualizing high-dimensional data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 2, pp. 2579–2605, 2008.

[36] S. Shao, S. McAleer, R. Yan, and P. Baldi, “Highly accurate machine fault diagnosis using deep transfer learning,” IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2446–2455, Apr. 2019.