




Article

# A Novel Stacked Auto Encoders Sparse Filter Rotating Component Comprehensive Diagnosis Network for Extracting Domain Invariant Features

Rui Ding <sup>1</sup>, Shunming Li <sup>1</sup>, Jiantao Lu <sup>1,\*</sup> , Kun Xu <sup>1</sup>  and Jinrui Wang <sup>2</sup> 

<sup>1</sup> College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China; Jerry.DingRui@gmail.com (R.D.); smli@nuaa.edu.cn (S.L.); kunxu.nuaa@gmail.com (K.X.)

<sup>2</sup> College of Mechanical and Electronic Engineering, Shandong University of Science Technology, Qingdao 266000, China; wangjinrui@sdust.edu.cn

\* Correspondence: lujt@nuaa.edu.cn

Received: 14 July 2020; Accepted: 28 August 2020; Published: 2 September 2020



**Abstract:** In recent years, the method of deep learning has been widely used in the field of fault diagnosis of mechanical equipment due to its strong feature extraction and other advantages such as high efficiency, portability, and so on. However, at present, most kinds of intelligent fault diagnosis algorithms mainly focus on the diagnosis of a single fault component, and few intelligent diagnosis models can simultaneously carry out comprehensive fault diagnosis for a rotating system composed of a shaft, bearing, gear, and so on. In order to solve this problem, a novel stacked auto encoders sparse filter rotating component comprehensive diagnosis network (SAFC) was proposed to extract domain invariant features of various health conditions at different speeds. The model clusters domain invariant features at different speeds through the self-coding network, and then classifies fault types of various parts through sparse filtering. The SAFC model was validated by the vibration data collected, and the results show that this model has higher diagnostic performance than other models.

**Keywords:** deep learning; stacked auto encoders; sparse filtering; domain invariant feature; multiple components

## 1. Introduction

With the coming of the IoT (Internet of Things) era, the reliability of mechanical equipment is more and more demanding [1]. The rotating components of mechanical equipment such as shafts, bearings, and gears are prone to shaft cracks, bearing wear, and gear breakage, which have great safety hazards [2]. In order to ensure the safe and reliable operation of equipment, accurate and efficient fault diagnosis technology is particularly important. As a powerful tool for processing big data, the deep learning [3,4] model has been successfully applied to mechanical intelligent fault diagnosis.

In the field of intelligent fault diagnosis of rotating parts, many scholars have studied related diagnosis by using various deep learning methods. Deng et al. [5] used the features extracted from the time domain, frequency domain, and time–frequency domain of vibration signals as deep Boltzmann machine (DBM) inputs to realize the accurate classification of seven different bearing faults. Ding et al. [6] adopted the wavelet packet energy map as the input signal and designed the depth convolutional neural networks (CNN) for bearing fault diagnosis. Meanwhile, in order to better display the multi-layer representation of the neural network, a multi-scale layer was added behind the final convolutional layer as the connection between the output of the final convolutional layer and the previous pooling layer. Li et al. [7] first implemented wavelet packet transform (WPT) on the original vibration signal, and then input the extracted statistical characteristics into a two-layer DBM

to realize the fault diagnosis of the gear box, and the classification rate of 11 operating modes reached 97.68%. Liu et al. [8] proposed a method based on variational mode decomposition (VMD) and CNN for feature extraction and fault diagnosis of planetary gear for local weak feature information, and realized the identification and classification of planetary gear weak fault state, where VMD based partition extraction method was better than EEMD, which could achieve 100% total CNN recognition rate with less training time (14 times). Chen et al. [9] not only adopted the statistical features in the time domain and frequency domain, but also added the test values of load and speed to form a set of feature vectors as the input of the deep belief network (DBN), realizing the fault diagnosis of the gear box.

Even more exciting, stacked auto encoders (SAE) [10], as a widely used deep learning algorithm, has attracted attention in the field of fault diagnosis. Feng Jia et al. [11] established a five-layer SAE and used frequency domain signals as input to diagnose the rotating bearing faults under different loads and speeds, with the diagnosis accuracy reaching almost 100%. Ngiam et al. [12] proposed an unsupervised feature learning method: sparse filtering (SF). This algorithm only focuses on optimizing the sparse performance of learning features while ignoring the distribution of learning data. Ya Guo Lei et al. [13] introduced sparse filtering into mechanical intelligent fault diagnosis. First, sparse filtering was used to extract unsupervised features of the time-domain signals of bearings, and then softmax was used as a classifier to realize fault diagnosis and good results were obtained.

For various fault diagnosis components, the main intelligent diagnosis algorithm was to detect the fault of the bearing and gear, respectively. Chen et al. [14] selected the method of adding Gaussian noise to the training data for data enhancement, and then input it into SAE for bearing fault diagnosis. Li Heng et al. [15] obtained spectrum samples by short-time Fourier transform (STFT) operation of the bearing vibration signals, and then input them into CNN for network training. The results showed that this method had high identification accuracy for different bearing fault types and could improve the robustness of this method by increasing the type and number of fault samples. Wang Weifeng et al. [16] proposed a fault diagnosis method based on Bi LSTM (binary long short term memory) network in order to solve problems such as large amounts of data and difficulty in extracting features in gear fault diagnosis. The results showed that the Bi LSTM network method for gear fault diagnosis was better than the CNN and LSTM network.

From what has been discussed above, it can be seen that the method of deep learning has been widely used in the field of fault diagnosis of mechanical equipment due to its strong feature extraction and other advantages such as its high efficiency, portability, and so on [17]. However, at present, all kinds of intelligent fault diagnosis algorithms mainly focus on the diagnosis of a single fault component, and few intelligent diagnosis models can simultaneously carry out comprehensive fault diagnosis for a rotating system composed of a shaft, bearing, gear, and so on. In order to solve this problem, a novel stacked auto encoders sparse filter rotating component comprehensive diagnosis model (SAFC) was proposed to extract domain invariant features of various health conditions at different speeds and classify them.

The proposed model can be used to extract domain invariant features of various health conditions at different speeds. Using the spectrum of fault signal as input, the deep neural network (DNN) was established through SAE, and the back propagation (BP) algorithm was used to fine-adjust the network. Meanwhile, the batch standardization (BN) algorithm was introduced into DNN to realize the rapid training of intelligent fault diagnosis model of rotating parts. After the clustering and dimension reduction of the data, the data was input into the sparse filtering model, and the unsupervised feature extraction of the input signal was carried out by using the sparse filtering. Finally, softmax was used as the classifier to realize fault diagnosis. The SAFC model was validated by the vibration data collected, and the results show that the proposed model has high diagnostic performance.

The main innovations are as follows:

- (1) A novel SAFC model that can extract and classify fault signal domain invariant features is proposed.
- (2) The proposed method is able to process multiple rotating parts at the same time and has the ability to process large amounts of data.

- (3) The domain invariant features of the same fault at different speeds can be extracted and clustered.

The organizational structure of the rest of this paper is as follows. Section 2 briefly introduces the SAE and SF theories, respectively; Section 3 introduces the architecture and training methods of SAFC model; Section 4 verifies the superiority of the SAFC model through two groups of experiments; Section 5 is the conclusion.

## 2. Introduction to Theory

### 2.1. Stacked Auto Encoders

The basic component unit of a SAE is an automatic encoder (AE) structure, and an AE is a symmetrical three-layer neural network consisting of an input layer, hidden layer, and output layer. The structure is shown in Figure 1. The learning process includes encoding and decoding: encoding is used to map input signals into hidden layer expression, while decoding is a reconstruction of hidden layer expression [18]. In the process of reconstruction, a reconstruction error is set to minimize the reconstruction error through iterative training to obtain the best hidden layer data expression.

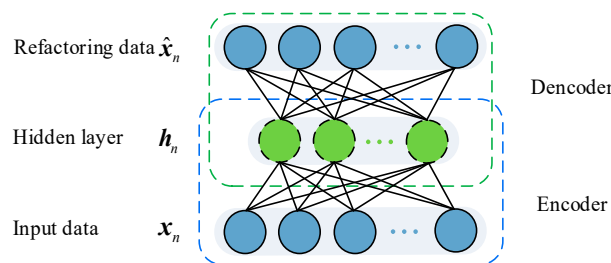


Figure 1. The automatic encoder AE architecture.

To build an AE, three things need to be done: building an encoder, building a decoder, and setting up a loss function measuring the information lost due to compression. The process of encoding and decoding usually adopts a parametric equation, and the parameter optimization process can be realized by minimizing the loss function.

If the input is a tableless dataset  $\{x_n\}_{n=1}^N$ , in which  $x_n \in R^{m \times l}$ ,  $h_n$  is the hidden layer encoding vectors, and  $\hat{x}_n$  is the output layer decoding vector. Therefore, the encoding process is as follows:

$$h_n = f(Wx_n + b_1) \tag{1}$$

where  $f(\cdot)$  is the coded activation function [19];  $W$  is the coded weight matrix; and  $b_1$  is the offset vector.

The decoding process is:

$$\hat{x}_n = g(W^T h_n + b_2) \tag{2}$$

where  $g(\cdot)$  is the coded activation function;  $W^T$  is the coded weight matrix; and  $b_2$  is the offset vector.

The network parameter set is optimized by minimizing the reconstruction error:

$$O(\Theta) = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^n L(x_i, \hat{x}_i) \tag{3}$$

where  $L$  is loss function:  $L(x_i, \hat{x}_i) = \|x_i - \hat{x}_i\|^2$ .

AE was superimposed layer by layer, that is, the first hidden layer was the second input, thus forming the DNN model structure based on SAE, as shown in Figure 2. The constituted DNN first carries out the layer by layer training of the forward propagation realization network, and realizes the network weight updating and parameter fine-tuning by the BP algorithm. When given an input signal, the input layer and the first hidden layer of DNN are regarded as the coding network of the first AE.

The parameters obtained after the first AE training are then used to initialize the second hidden layer of DNN. The calculation process of the first coding vector of the input signal is as follows:

$$h_n^1 = f(x_n) \tag{4}$$

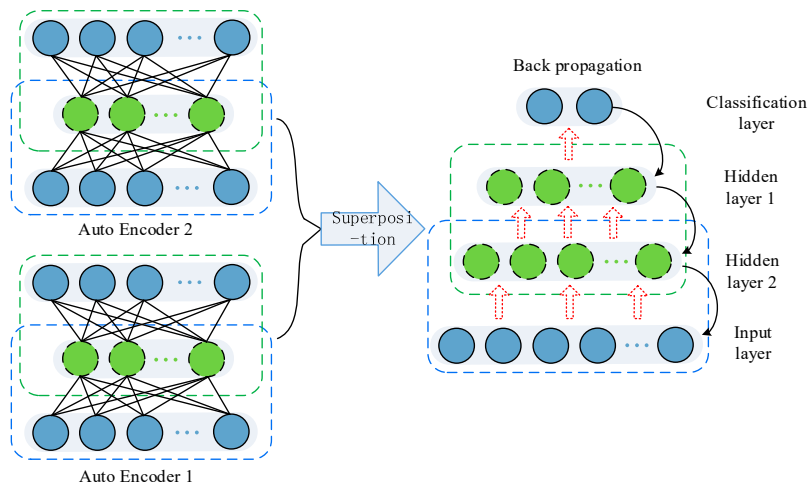


Figure 2. The stacked auto encoders SAE architecture.

Then, the encoding vector was used as input data to train the second AE, that is, the first hidden layer and the second hidden layer of the DNN were used as the encoding network of the second AE. Accordingly, the second hidden layer of DNN is initialized by the second AE, step by step until the last AE completes the training:

$$h_n^M = f(h_n^{M-1}) \tag{5}$$

After pre-training the superimposed AE, the softmax classification layer containing sample tags was used as the DNN output layer, followed by SAE, and then the BP algorithm was used to reverse fine-tune the DNN parameters layer by layer. Therefore, the DNN output calculated from the input signal was:

$$y_n = f(h_n^M) \tag{6}$$

### 2.2. Batch Normalization

The batch normalization can reparameterize almost all DNN in an elegant way. The process can be applied to each activation layer without parameter adjustment. It can be standardized in an independent way for each row of a matrix with zero mean and unit variance. Suppose an n-dimensional input  $x = (x_1 \dots x_n)$  in order to improve training and reduce the internal covariant transfer problem, the batch standardization algorithm adopts two necessary simplification steps:

First, normalize each unit:

$$\hat{x}_i = \frac{x_i - E[x_i]}{\sqrt{\text{Var}[x_i]}} \tag{7}$$

where  $E[x_i]$  is the mean value of each unit and  $\sqrt{\text{Var}[x_i]}$  is the standard deviation. Then, translate and scale the normalized value:

$$f_i = \gamma_i \hat{x}_i + \beta_i \tag{8}$$

where  $\gamma_i$  represents the scaling parameter and  $\beta_i$  represents the translation parameter.



Second, assume that there are  $m$  values in  $x_i$ , namely the minimum batch is  $\phi = \{x_{1\dots m}\}$ ,  $\hat{x}_{1\dots m}$  represents the normalized value, and the corresponding linear transformation is  $y_{1\dots m}$ . Therefore, the batch standardization transformation is  $\text{BN}_{\gamma,\beta} : x_{1\dots m} \rightarrow y_{1\dots m}$ , namely:

$$E[x_\phi] = \frac{1}{m} \sum_{j=1}^m x_j \tag{9}$$

$$\text{Var}[x_\phi] = \frac{1}{m} \sum_{j=1}^m (x_j - E[x_\phi])^2 \tag{10}$$

$$\hat{x}_j = \frac{x_j - E[x_\phi]}{\sqrt{\text{Var}[x_\phi] + \varepsilon}} \tag{11}$$

$$\hat{y}_j = \gamma \hat{x}_j + \beta \tag{12}$$

where  $\varepsilon$  is an imposed constant to avoid undefined gradient problems of  $\sqrt{s}$  when  $s = 0$ .

Furthermore, gradient loss  $\ell$  is back propagated in batch standardized training:

$$\frac{\partial \ell}{\partial \hat{x}_j} = \frac{\partial \ell}{\partial y_j} \gamma \tag{13}$$

$$\frac{\partial \ell}{\partial \text{Var}[x_\phi]} = \sum_{j=1}^m \frac{\partial \ell}{\partial \hat{x}_j} (x_j - E[x_\phi]) \cdot \left[-\frac{1}{2} (\text{Var}[x_\phi] + \varepsilon)^{-\frac{3}{2}}\right] \tag{14}$$

$$\frac{\partial \ell}{\partial E[x_\phi]} = \sum_{j=1}^m \frac{\partial \ell}{\partial \hat{x}_j} \frac{-1}{\sqrt{\text{Var}[x_\phi] + \varepsilon}} \tag{15}$$

$$\frac{\partial \ell}{\partial x_j} = \frac{\partial \ell}{\partial \hat{x}_j} \frac{-1}{\sqrt{\text{Var}[x_\phi] + \varepsilon}} + \frac{\partial \ell}{\partial \text{Var}[x_\phi]} \cdot \frac{2(x_j - E[x_\phi])}{m} + \frac{\partial \ell}{\partial E[x_\phi]} \cdot \frac{1}{m} \tag{16}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{j=1}^m \frac{\partial \ell}{\partial y_j} \cdot \hat{x}_j \tag{17}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{j=1}^m \frac{\partial \ell}{\partial y_j} \tag{18}$$

Batch normalized transformation introduces standardized activation into each layer of the network to ensure that all layers of the network can continuously learn the input distribution to reduce the problem of internal covariate transfer, so an easy initial environment can be established at the beginning of network training to speed up network training.

Applying the batch normalization to each activation layer of the SAE. Batch normalization transformations are applied to normalization  $Wx_n + b_1$  in Equation (1). The bias is considered to be subtracted in the subsequent mean subtraction process, so the bias was not used here. Therefore, Equation (1) can be changed to:

$$h_n = f(\text{BN}(Wx_n)) \tag{19}$$

In addition, the saturation problem can lead to the disappearance of the gradient by using linear rectifier unit function (rectified linear units, ReLU) to solve.

$$\text{ReLU}(x) = \max(x, 0) \tag{20}$$

The ReLU function can make it easier for the network to learn optimization during training. Compared with the traditional sigmoid and tanh functions, its advantages are as follows:

- (1) Due to the nonlinear characteristics of sigmoid and tanh functions, both of them carry out exponential operations, leading to an increase in the calculation amount when taking the derivative of back propagation, while the ReLU function can save a lot of calculation time.
- (2) When sigmoid and tanh functions carry out back propagation in deep network, it is very easy to generate a gradient disappearance problem, that is, when approaching the saturation zone, the derivative approaches 0, transformation slows down, and information is easily lost. Therefore, deep network training cannot be realized.
- (3) The ReLU function can set part of the output to 0, making the model more sparse, thus reducing the dependency relationship between network parameters and reducing the overfitting phenomenon.

### 2.3. Sparse Filtering

As a simple and effective unsupervised feature learning algorithm, the sparse filtering algorithm only needs one parameter adjustment, namely the number of features to be learned [20]. Its working principle is mainly to bypass the estimation of data distribution and optimize only a simple cost function, the L2 norm, to directly analyze the distribution of optimized features.

The structure of sparse filtering is a two-layer neural network including input layer, output layer, and weight matrix. The input is the collected signal and the output is the learned feature. The collected signals are divided into many identical samples to form a training set  $\{x^i\}_{i=1}^M$ , where  $x^i \in \mathfrak{R}^{N \times 1}$  is a sample containing  $N$  data points and  $M$  is the number of samples. The sample is mapped to the eigenvector  $f^i \in \mathfrak{R}^{L \times 1}$  by the usual-value matrix  $W \in \mathfrak{R}^{N \times L}$ , and  $f_j^i$  was used to represent the  $j$ th eigenvalue of the  $i$ th sample of the eigenmatrix. The features learned are linear:

$$f_j^i = W_j^T x^i \tag{21}$$

First, all row vectors of the eigenmatrix are normalized, then all column vectors of the eigenmatrix are normalized, and finally the sum of the absolute values of all elements of the matrix is obtained. The details are as follows:

First, each feature is normalized to an equal activation value, that is, the L2 norm of each feature divided by all the samples in which it is located:

$$\tilde{f}_j = f_j / \|f_j\|_2 \tag{22}$$

Then, the characteristics of each sample are normalized, so that the characteristics of all samples fall on the unit sphere of L2 norm:

$$\hat{f}^i = \tilde{f}^i / \|\tilde{f}^i\|_2 \tag{23}$$

Finally, the L1 norm penalty is used to carry out the sparsity constraint to realize all the normalized features before optimization. Assuming that a dataset has  $M$  samples, the objective function of sparse filtering is:

$$\underset{W}{\text{minimize}} \sum_{i=1}^M \|f^{\hat{i}}\|_1 = \sum_{i=1}^M \left\| \frac{\tilde{f}^i}{\|\tilde{f}^i\|_2} \right\| \tag{24}$$

### 3. Proposed Smart Diagnosis Method

The SAFC model is composed of SAE and SF, where SAE has five layers of network and SF has two layers of network, as shown in Figure 3. The specific steps of the extracting features are as follows:

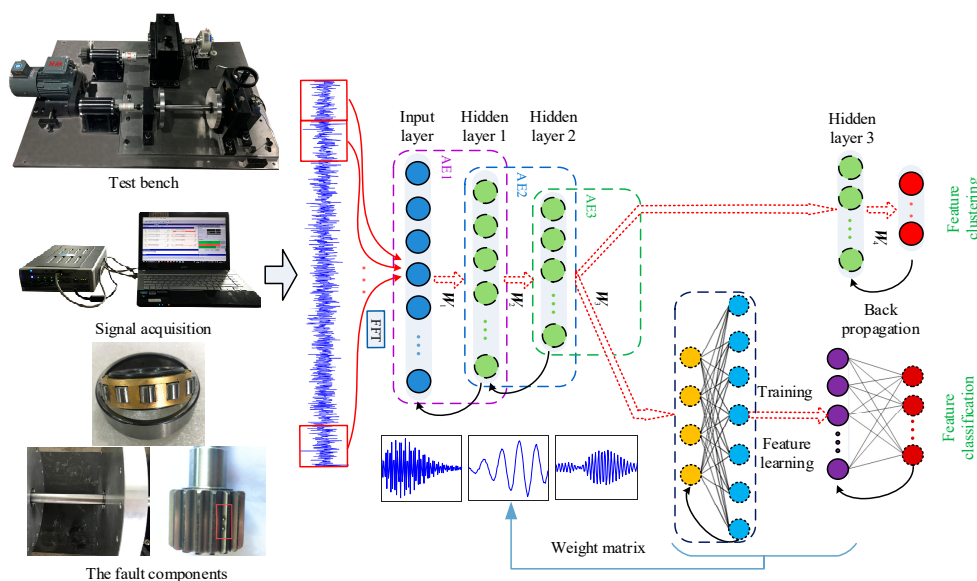
- (1) All the collected time domain signals are converted into frequency domain signal through FFT. According to different health conditions, the spectrum samples of vibration signals at different speeds are composed into a training set  $\{X^i, l^i\}_{i=1}^M$ , where  $M$  is the number of samples,  $X^i \in \mathfrak{R}^{N \times 1}$  means that the  $i$ th sample contains  $N$  Fourier coefficients, and  $l^i$  is the health label of the  $i$ th sample.

- (2) The no label training set  $\{X^i\}_{i=1}^M$  was used to pre-train the SAE layer by layer. The pre-training process was to superimpose  $N$  AE into  $N$  hidden layers, and the characteristics learned in the previous layer were used as the input of the next layer.
- (3) BP algorithm combined with label data  $\{X^i, l^i\}_{i=1}^M$  was used to realize weight updating and parameter fine-tuning of SAE, and back propagation training was carried out by minimizing the error between extracting features and health labels.
- (4) BN was applied to each activation layer of SAE, and the characteristic difference at different speeds was reduced by its translation and scaling.
- (5) The features initially extracted after dimension reduction through two-layer SAE were taken as new training samples and a new training sample set  $\{x^i, l^i\}_{i=1}^m$  was formed.  $m$  is the number of samples [21],  $x^j \in R^{N_{in} \times 1}$  means the  $i$ th sample contains  $N_{in}$  data points and  $l_i$  is the label of the sample.
- (6) The sample set was formed into a matrix form  $T \in R^{N_{in} \times M}$ , and then input into the sparse filtering model for training. The weight matrix  $W$  is obtained through the minimization formula (24).
- (7) The learning feature  $f^i \in R^{N_{out} \times 1}$  can be calculated by  $W$  and  $x_i$ . ReLU was used as the activation function here to extend the sparse filtering to nonlinear mapping. Therefore, Equation (20) is extended to:

$$f_j^i = \text{ReLU}(W_j^T x^i) \tag{25}$$

- (8) When the learning feature set  $\{f^i\}_{i=1}^M$  is obtained, it is combined with the tag set  $\{l^i\}_{i=1}^M$  and input into the softmax regression classifier for training. Then, the remaining samples are used as test samples to test the accuracy of the proposed method [22].

The testing process is as follows. First, each test sample is converted into spectrum after FFT transformation, and all spectrum samples are combined into matrix form as input; second dimension reduction and preliminary feature extraction are carried out in superposition self-coding to cluster the same type of signals with different speeds. Then, the trained sparse filter model is used to study the characteristics of the test samples. Finally, the learned features are combined with the real fault type label, and then input to the softmax regression model that has been trained for classification, thus obtaining the test accuracy rate.



**Figure 3.** The overall architecture of the sparse filter rotating component comprehensive diagnosis model (SAFC) model.

#### 4. Experimental Verification

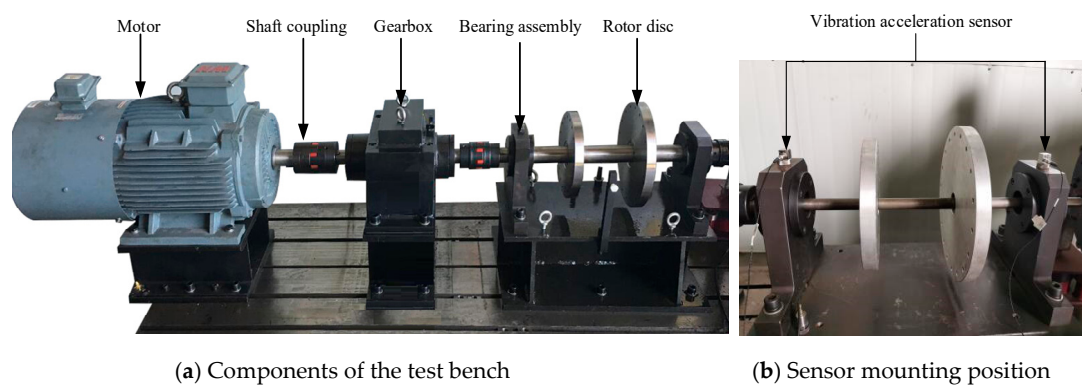
To verify the feasibility of the proposed method, a rotating parts fault diagnosis bench was designed based on two groups of tests. The first group classified the test data according to different parts, while the second group classified the test data on the basis of different fault locations.

##### 4.1. Case Study I: Depending on Different Fault Components

##### 4.1.1. Test Equipment and Data Introduction

As shown in Figure 4, the designed experimental equipment mainly included the motor, coupling, planetary gear box, bearing seat, rotor, and so on. Aside from the normal fault-free vibration signal, six different fault types were set respectively for different fault positions of the shaft, bearing, and gear. Depending on the location and depth of the shaft crack, the six fault types were three cracks on the left side of the larger rotor (1.2 mm, 2.4 mm, and 3.6 mm) and three cracks on the right side of the smaller rotor (1.2 mm, 2.4 mm, and 3.6 mm), as shown in Figure 5. The six faults were referred simply to as L1, L2, L3, R1, R2, and R3, where L represents left, and R represents right. According to the location and depth of the bearing fault, the six fault types were three inner ring faults (0.2 mm, 0.6 mm, and 1.2 mm) and three outer ring faults (0.2 mm, 0.6 mm, and 1.2 mm), as shown in Figure 6. These six faults were also simply referred to as I1, I2, I3, O1, O2, and O3, where I represents inner, and O represents outer. On the basis of the location and depth of the gear failure, the six fault types were three kinds of sun wheel failures and three kinds of planetary wheel failures: crack, pitting, and wear, as is shown in Figure 7. In the same way, these six faults were simply referred to as P1, P2, P3, S1, S2, and S3, where P represents planetary, and S represents sun. Therefore, there were three groups of seven health conditions each, and a total of 21 sets of test data needed to be processed. However, the 21 sets of test data to be processed actually only represent 19 health conditions since the failure-free data in the three sets was the same.

The data settings are shown in Table 1. During the test, the motor speed was respectively set as 900 r/min, 1100 r/min, 1300 r/min, and 1500 r/min. The output speed was three times that of the motor speed since the speed ratio of planetary gearbox was 1:3. It is necessary to collect test data at all speeds for the all 19 cases above-mentioned. Therefore, 78 sets of test data needed to be collected. For the collection, a three-way vibration acceleration sensor (PCB315A) was installed on the surface of the bearing seat. Vibration signals on the upper surface of the bearing seat were collected by the LMS data acquisition instrument, and the sampling frequency was set at 12.8 KHz. A total of 200 samples were collected for each health condition, and 600 data points were taken for each sample at each rotation speed. By stitching together signals at different speeds for the same health condition, 200 samples of each health condition were obtained, each with 2400 data points.



**Figure 4.** Rotating components fault diagnosis test bench.

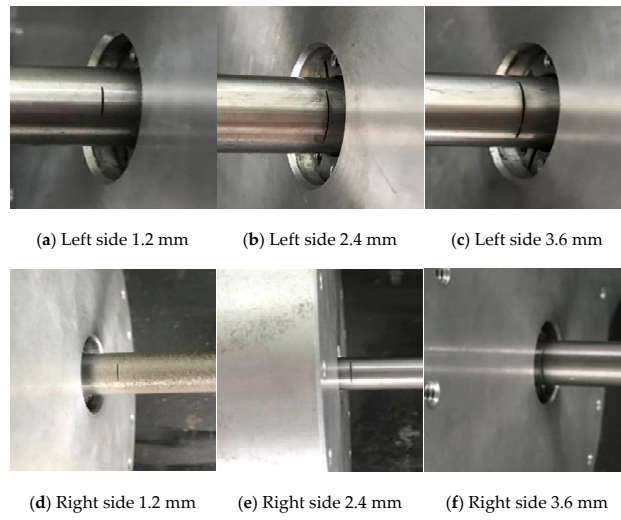


Figure 5. Shaft crack.

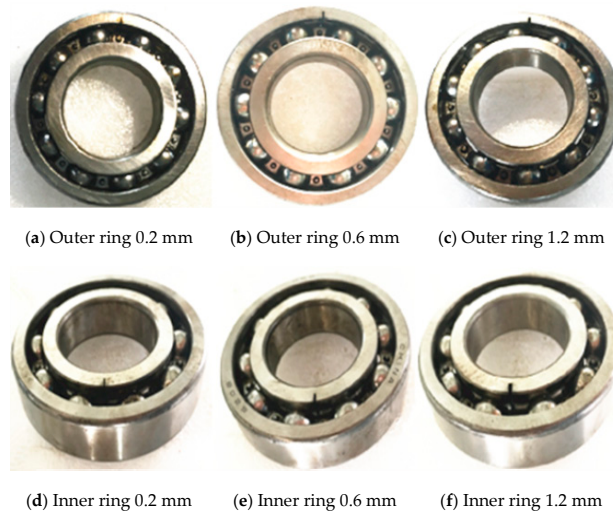


Figure 6. Bearing fault.

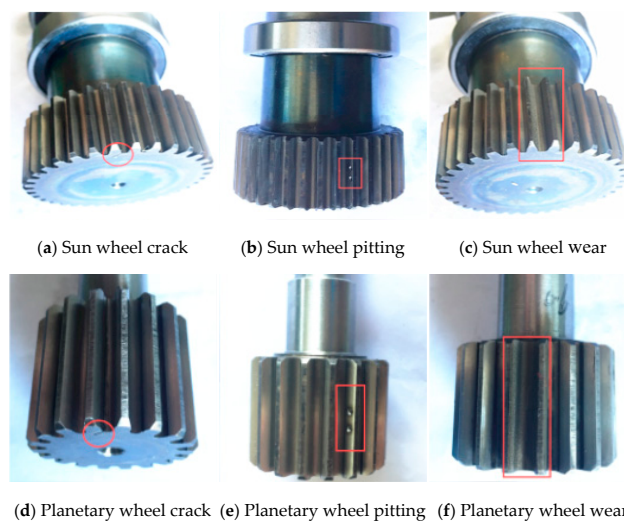
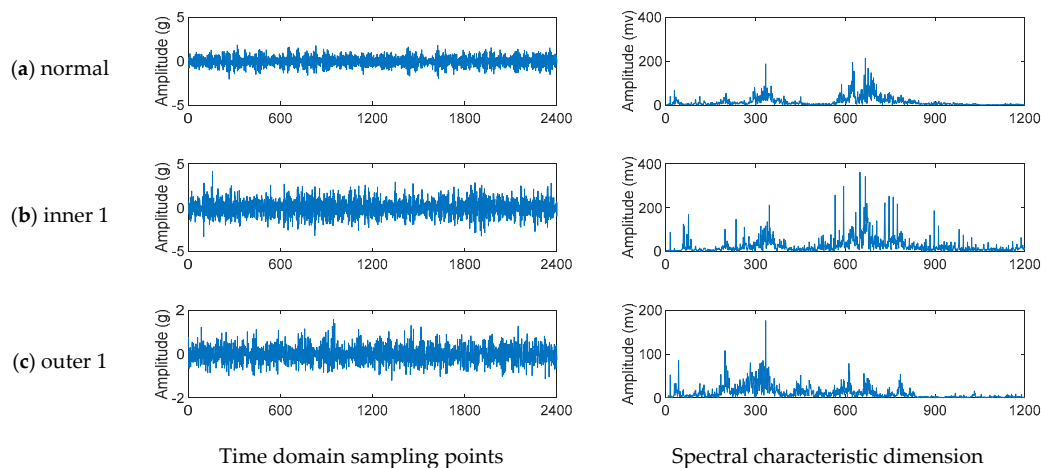


Figure 7. Bearing fault.

**Table 1.** Description of datasets.

Group Number		1	2	3	4	5	6	7
Shaft	Fault location	Normal	Crack on left side of the larger rotor			Crack on right side of the smaller rotor		
	abbreviation	N	L1	L2	L3	R1	R2	R3
	Depth (mm)	0	1.2	2.4	3.6	1.2	2.4	3.6
Bearing	Fault location	Normal	Failure of bearing inner ring			Failure of bearing outer ring		
	abbreviation	N	I1	I2	I3	O1	O2	O3
	Depth (mm)	0	0.2	0.6	1.2	0.2	0.6	1.2
Gear	Fault location	Normal	Failure of planetary wheel			Failure of sun wheel		
	abbreviation	N	P1	P2	P3	S1	S2	S3
	Fault type	none	crack	pitting	wear	crack	pitting	wear

The 2400 data points of each sample were processed by FFT to obtain 1200 Fourier coefficients. The reason why frequency spectrum was chosen was due to the time shifting characteristics of the time-domain samples. It is difficult to ensure the consistent position of the fault feature points in different samples of the same fault, which produces great difficulty in feature extraction. Additionally, the length of each signal input to the classifier will affect the accuracy rate. The spectrum information can not only easily obtain richer information of different health conditions from the test signal by FFT, but its more regular characteristic distribution also overcomes the time-shift characteristic of the time-domain signal. Taking the bearing data, for example, the corresponding signal wave-forms in the time domain and frequency domain are shown in Figure 8. It can be seen that it is difficult to distinguish the different fault types from both the time domain and frequency domain.



**Figure 8.** Time and frequency domain wave-forms of bearing signals with different health conditions.

#### 4.1.2. Parameter Selection

The number of neurons in the input layer of the SAFC model was the same as the sample dimension, which was 200, the dimension of the sample spectrum. In the three hidden layers, the number of neurons was set as 600, 200, and 100. The number of neurons in the output layer was the same as the total number of health conditions. There was a total number of 20 training iterations for each layer. The learning rate was determined to be 1 E-4, while batch size was determined by parameter selection. After the data was output through the SAE network, the second hidden layer of SAE was used as the input layer of SF. The number of neurons in the input layer was also the same as the sample dimension. The number of neurons in the output layer was determined by parameter selection [23]. The test accuracy of the SAE part was taken as clustering accuracy, and the test accuracy of SF part as classification accuracy.

First of all, the proportion of training samples in the SAE part was selected for research. Gear data was selected for testing, and the training batch size was tentatively set at 10. Meanwhile, the percentage of training samples in SF was set at 10%, and the output dimension was set at 200. In order to eliminate



the effect of randomness, each group of tests were carried out 20 times. The diagnostic accuracy is shown in Figure 9. The results show that with the increase in the training sample size, the diagnostic accuracy increased continuously, and the time spent increased almost linearly. When the percentage of SAE training samples increased to 20%, the clustering accuracy reached  $99.98\% \pm 0.03\%$ , and the classification accuracy reached  $99.22 \pm 0.11\%$ , after which both the two accuracy rates remained almost unchanged. Therefore, after weighing the accuracy of classification against the time spent, 20% was used as a percentage of the training sample in the following experiment.

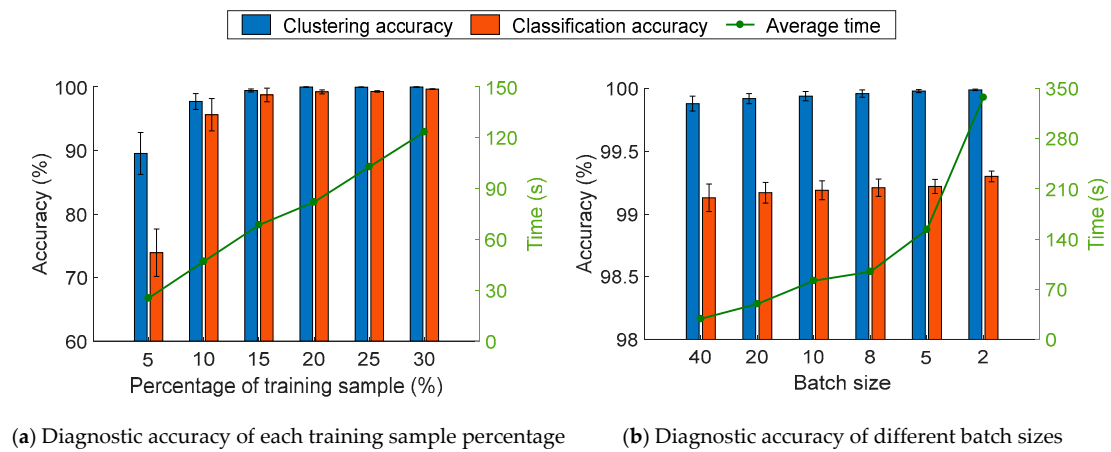


Figure 9. SAE parameter selection.

Then, after considering the selection of SAE training batch size, the diagnostic results of different batch sizes are as shown in Figure 10. It can be seen that when the batch size was reduced, the clustering accuracy and classification accuracy were slightly improved, but the average training time increased sharply with the decrease in batch size. When the batch size was 5, the training time was up to 153.27 s, which processed only one-third of the data that needed to be processed. When the batch size was 40, the time taken was 28.74 s, which was far less than 82.01 s when the batch size was 10, but the accuracy and variance were almost the same. Therefore, 40 was selected as the batch size.

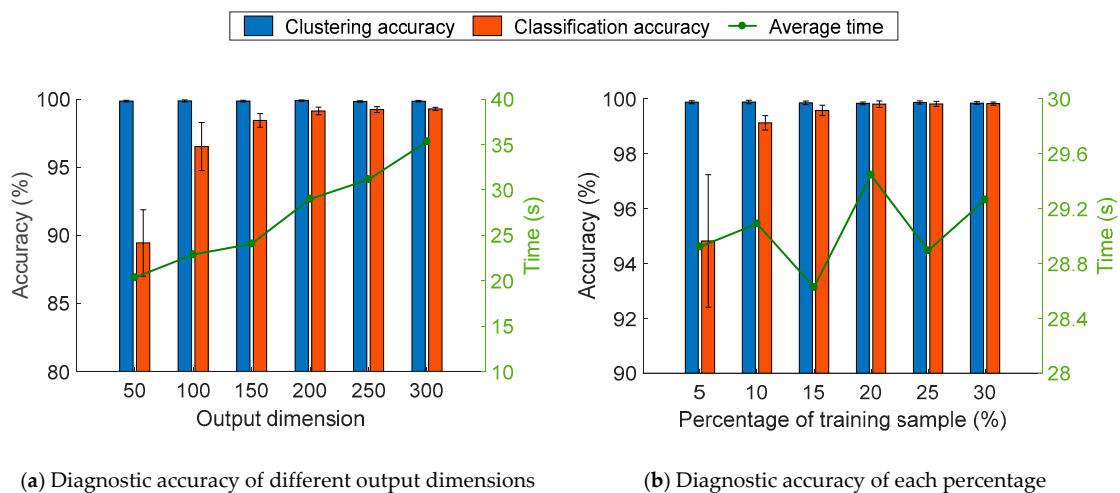


Figure 10. SF parameter selection.

The parameters of the SF part were also taken into consideration after confirming the parameters of the SAE part. In order to study the effect of output dimension, 10% samples were still selected for testing. The diagnostic accuracy is shown in the figure. The results show that with the increase in output dimension, the diagnostic accuracy increases and the corresponding standard deviation decreases.

The results show that with the increase of the output dimension, the diagnostic accuracy increases and the corresponding standard deviation decreases, while the average training time increases. Therefore, a trade-off was made between classification accuracy and time spent. Considering that the increase in accuracy was no longer significant after the output dimension reached 200, 200 was selected as the number of neurons in the output layer.

The diagnostic results with different proportions of training samples in the SF part are shown in Figure 10. It can be seen that with the increase in sample percentage, the diagnostic accuracy was constantly improved and the standard deviation was reduced. When the sample percentage was 10%, the classification accuracy reached  $99.13\% \pm 0.15\%$ . However, as the percentage of training samples increased further, the diagnostic accuracy hardly improved, and the diagnostic time fluctuated only within a small range. Therefore, 10% was chosen as the proportion of training sample.

#### 4.1.3. The Results of the Diagnosis

The first 20% of the sample was selected in the SAE section, and after the dimension reduction, the first 10% of the sample was selected in the SF section. Using the same training dataset to train the three models, the diagnostic accuracy of the data of the three parts is shown in Figure 11. It can be seen that SAFC, which contained both methods, was more accurate and stable than SAE or SF alone. In the process of diagnosis, the diagnostic accuracy of the three methods for gear fault was the lowest among the three parts, so took gear as an example. The average accuracy and time of the three methods were calculated, the results of which showed that the SAFC accuracy was above 99.13%, and the standard deviation was no more than 0.11 while the mean accuracy rate of SAE and SF was only 86.37% and 93.80%, and the standard deviation exceeded 3.10 and 1.05. This shows that the proposed method had higher accuracy and stability. The average training time of the sample was also calculated, and the specific calculation results are shown in Table 2. It can be seen that the proposed method SAFC is not only more accurate than SF, but also faster in diagnosis, which benefits from the rapid speed of SAE in training data and its dimension reduction of the data. It can also be seen that although SAFC is not as fast as SAE in diagnostic speed, its accuracy is far better than SAE, which is due to the powerful feature extraction capability of SF.

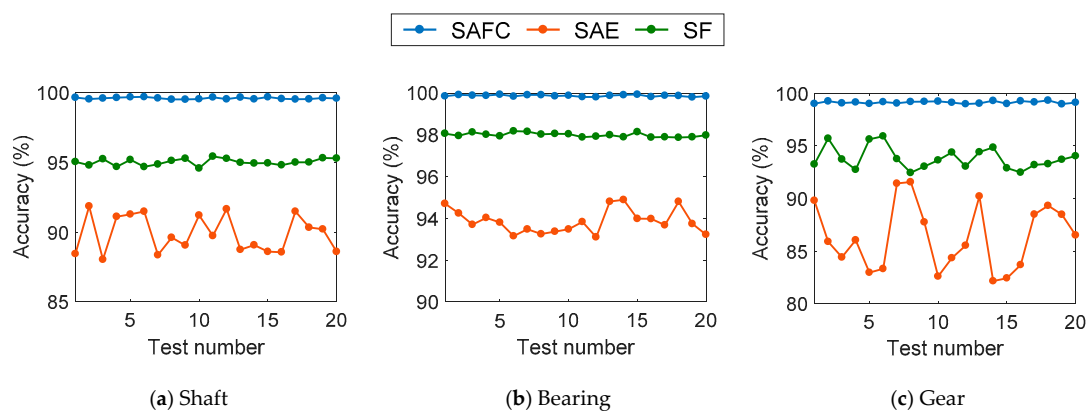


Figure 11. Diagnostic accuracy of the three methods for each component.

Table 2. Time and accuracy of diagnosis.

	SAFC		SAE		SF	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
Shaft	$99.61 \pm 0.06$	$28.85 \pm 0.65$	$89.86 \pm 1.33$	$17.63 \pm 0.29$	$95.03 \pm 0.24$	$42.27 \pm 1.05$
Bearing	$99.87 \pm 0.04$	$28.23 \pm 0.67$	$93.86 \pm 0.57$	$17.58 \pm 0.24$	$97.99 \pm 0.10$	$42.93 \pm 1.10$
Gear	$99.13 \pm 0.11$	$28.39 \pm 0.48$	$86.37 \pm 3.10$	$17.33 \pm 0.25$	$93.80 \pm 1.05$	$42.46 \pm 1.24$

In order to prove the classification effect of the proposed method better, the three methods were used to extract features from the data of three parts, and the t-SNE dimension reduction algorithm was used to transform the obtained feature vectors into two-dimensional views.

The results of feature mapping of the three methods are respectively shown in Figures 12–14 [24]. It can be seen that in the proposed method, characteristic samples of the same health status were clustered in the corresponding clusters, and test samples of different health status were well separated. Only one sample of Sun 3 was divided into the cluster of planet 2 in gear failure. However, in SAE, not only in the bearing data, a sample of inner ring 2 was not well aggregated, but also in the shaft data, three groups of crack characteristics on the left side of the larger rotor were not very well separated. In addition, in the process of gear fault diagnosis, four groups of characteristics were staggered, while in the SF, although the characteristics of different health conditions were generally separated, uniform health conditions at different speeds were also separated, which obviously did not meet the classification requirements.

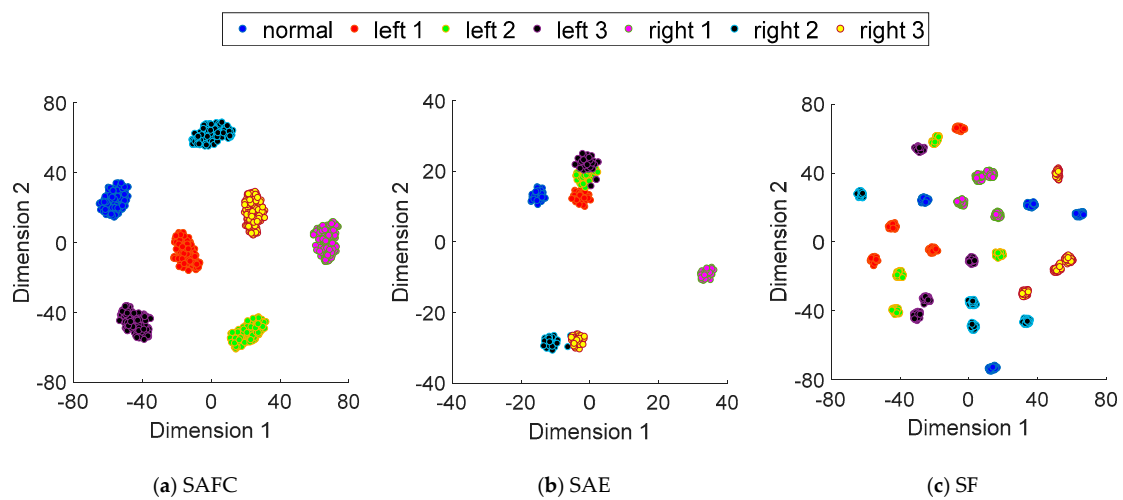


Figure 12. Signal characteristic dimension reduction of shaft crack using the three methods.

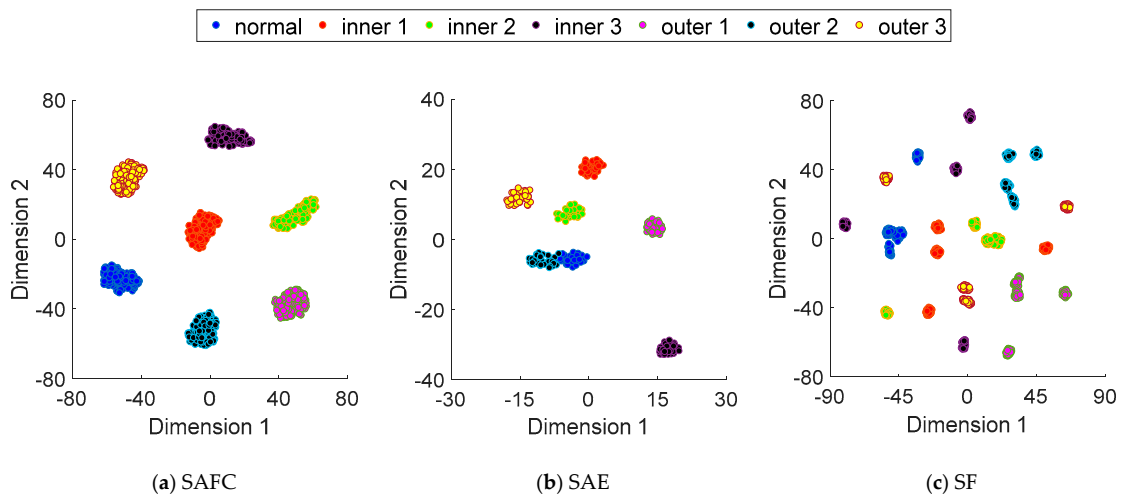


Figure 13. Signal characteristic dimension reduction of bearing fault using the three methods.

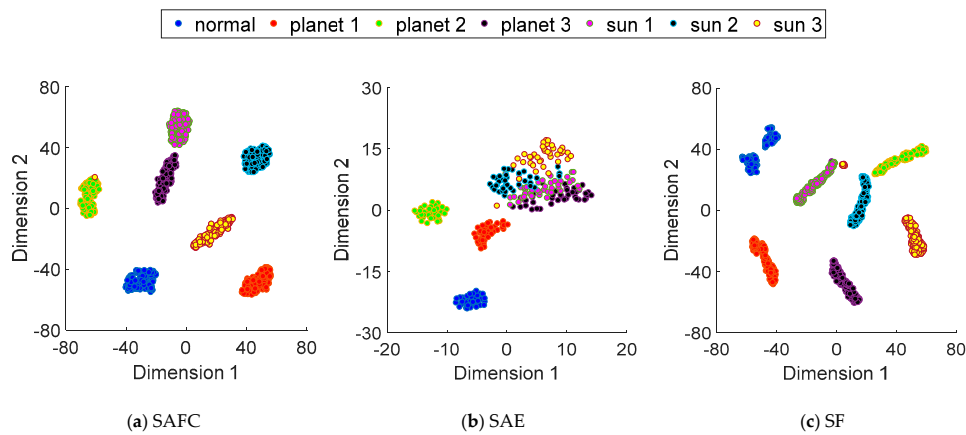


Figure 14. Signal characteristic dimension reduction of gear fault using the three methods.

Figures 15–17 show the visualization of the feature distribution of the three components under the three methods. It can be seen that in SAFC, the characteristics of different fault types were different with obvious differences, while in SAE, the difference was relatively fuzzy. In SF, subdivision occurred in each of the same characteristics, that is, the same health condition showed different characteristics, which was consistent with the results in t-SNE. In conclusion, the proposed method can well extract domain invariant features and classify them.

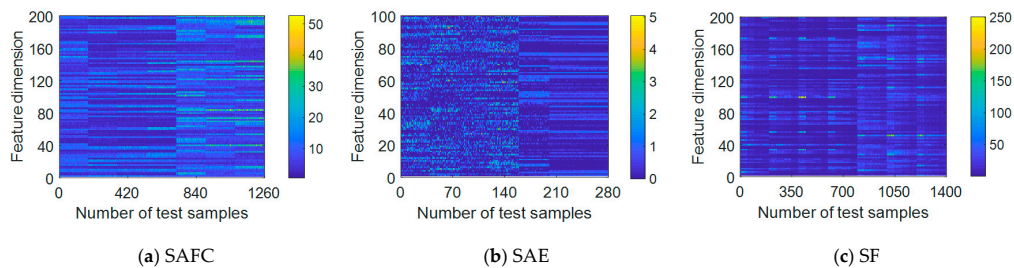


Figure 15. Signal characteristic of shaft crack using the three methods.

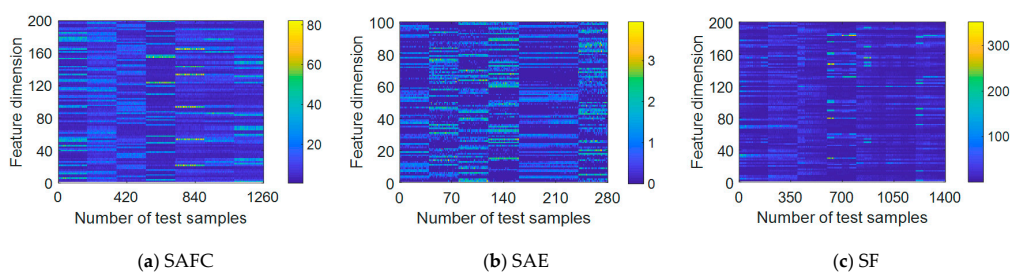


Figure 16. Signal characteristic of bearing fault using the three methods.

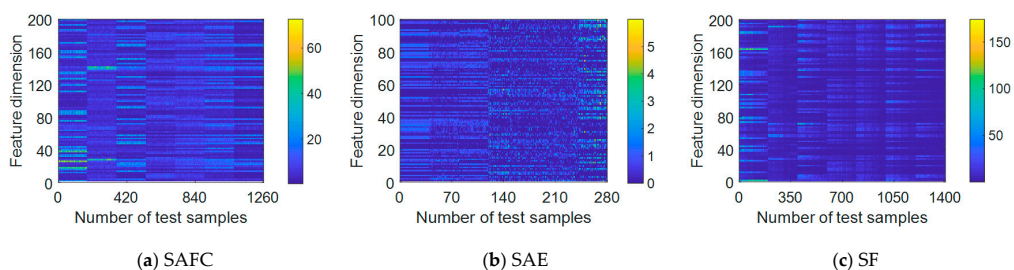


Figure 17. Signal characteristic of gear fault using the three methods.

#### 4.2. Case Study II: According to Different Fault Location

##### 4.2.1. Data Introduction

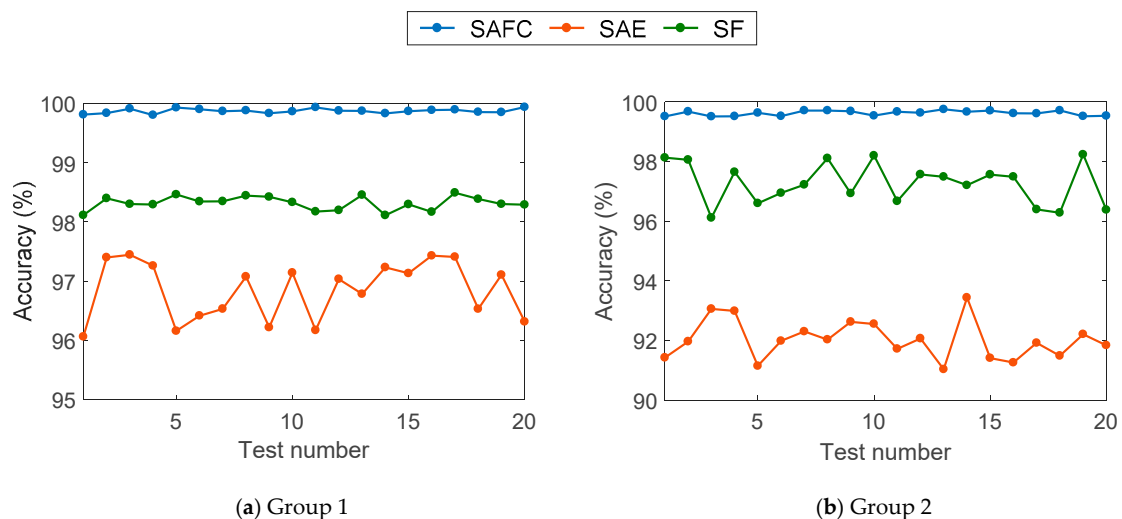
To further prove the effectiveness of the proposed method in distinguishing different fault parts, the previously measured test data were regrouped according to the different fault locations. The general idea is that different degrees of failure of one of the two fault locations of each part at different speeds were combined with normal trouble-free data, which resulted in two datasets (Table 3) with 10 health conditions each. For example, we can make a group of three sets of crack data on the right side of the smaller rotor, three sets of fault data of the inner ring of the bearing, and three sets of fault data of the planet wheel and normal fault-free data. The remaining three sets of crack data on the left side of the larger rotor, three sets of failure data of the bearing outer ring, three sets of failure data of the sun wheel and normal trouble-free data are made into the other group. Again, the normal failure-free data is Shared by both sets of data, so there are still 19 different health conditions. The selection of parameters is the same as before.

**Table 3.** Description of datasets.

Group Number		1	2	3	4	5	6	7	8	9	10
Group 1	location	Normal	Right side of the smaller rotor			Bearing inner race			planet gear		
	acronym	N	R1	R2	R3	I1	I2	I3	P1	P2	P3
Group 2	location	Normal	Left side of the larger rotor			Bearing outer race			sun gear		
	acronym	N	L1	L2	L3	O1	O2	O3	S1	S2	S3

##### 4.2.2. The Result of the Diagnosis

The diagnostic accuracy of the three methods for the two sets of data is shown in Figure 18. It can be seen that SAFC is more accurate and stable than SAE and SF. The specific calculation results are shown in the Table 4. In the diagnosis of the two sets of data, the diagnostic accuracy is generally higher than the diagnosis of each component, which is because the characteristics of different components have more difference than the characteristics of different positions of the same component. Similarly as before, the diagnostic time of SAFC is between SAE and SF, but SAFC still has the highest diagnostic accuracy.

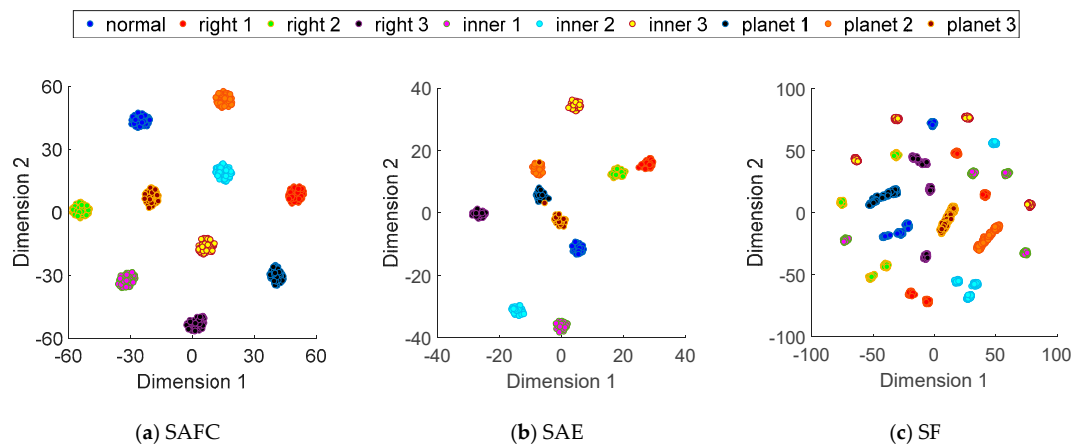


**Figure 18.** Diagnostic accuracy of three methods for two groups of data.

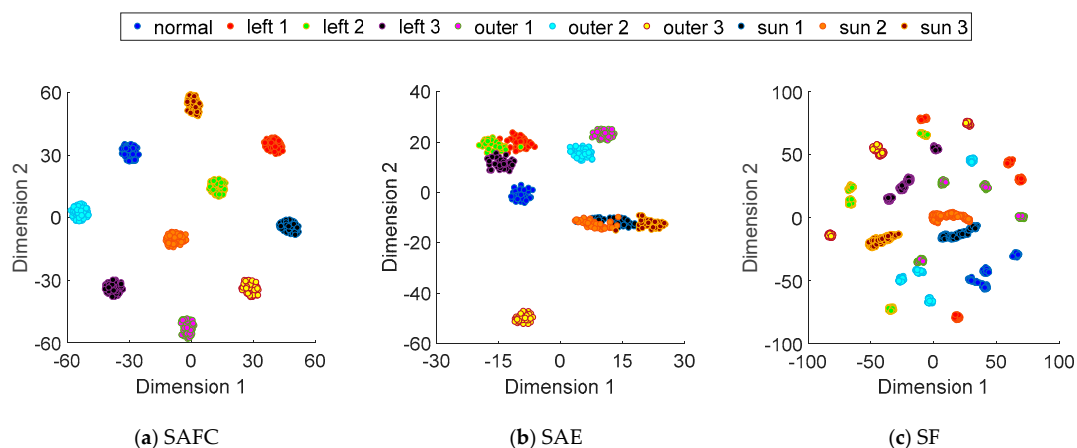
**Table 4.** Time and accuracy of diagnosis.

	SAFC		SAE		SF	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
Group 1	99.91 ± 0.04	43.35 ± 1.46	96.84 ± 0.49	25.16 ± 0.76	98.32 ± 0.12	62.37 ± 2.57
Group 2	99.62 ± 0.08	43.64 ± 1.81	92.03 ± 0.66	25.84 ± 0.89	97.27 ± 0.60	62.49 ± 2.67

Again, these three methods were used to extract features from the two sets of data, and t-SNE dimension reduction was used to obtain a two-dimensional view. The feature mapping results of the three methods are shown in Figures 19 and 20. It can be seen that in the proposed method, characteristic samples of the same health status are clustered in the corresponding clusters, and test samples of different health status are well separated. However, in SAE, not only in the group 1, some samples of planet wheel 3 were not well aggregated, but also in the group 2, some groups of characteristics were not very well separated. While in the SF, although the characteristics of different health conditions were generally separated, uniform health conditions at different speeds were also separated, which obviously did not meet the classification requirements.



**Figure 19.** Signal feature dimension reduction for group 1 using the three methods.



**Figure 20.** Signal feature dimension reduction for group 2 using the three methods.

Figures 21 and 22 show the visualization of the feature distribution of the two groups under the three methods. It can be seen that in SAFC, the characteristics of different fault types were different with obvious differences, while in SAE, the difference was relatively fuzzy. In SF, subdivision occurred in each of the same characteristics, that is, the same health condition showed different characteristics,



which was consistent with the results in t-SNE. In conclusion, the proposed method can well extract domain invariant features and classify them.

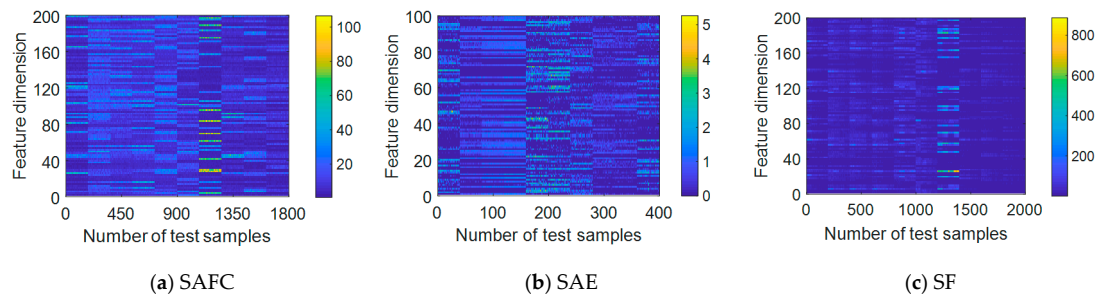


Figure 21. Signal characteristics for group 1.

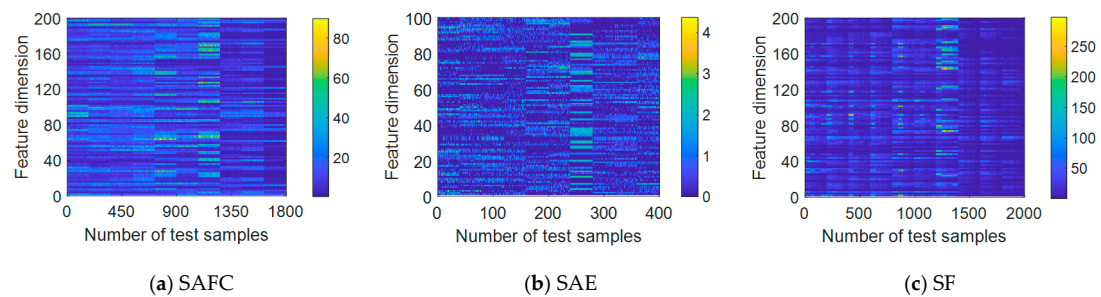


Figure 22. Signal characteristics for group 2.

4.3. A Comparison with the Other Two Methods

In the literature [25,26], the deep learning methods of BNAE and L1/2-SF are provided respectively, which also have the function of feature extraction. With the previous dataset in Section 4.2, the three models were tested separately, and the proposed method was compared with the two methods, as shown in Figure 23. It can be seen that the proposed SAFC is superior to the compared method in feature extraction and classification, which is also presented in the obfuscation matrix, as shown in Figure 24, which again proves the feasibility of the proposed method.

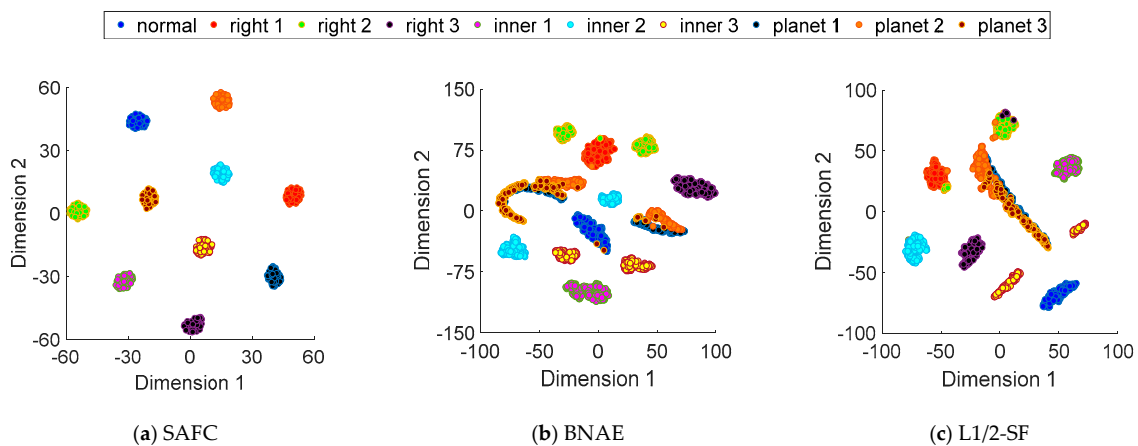
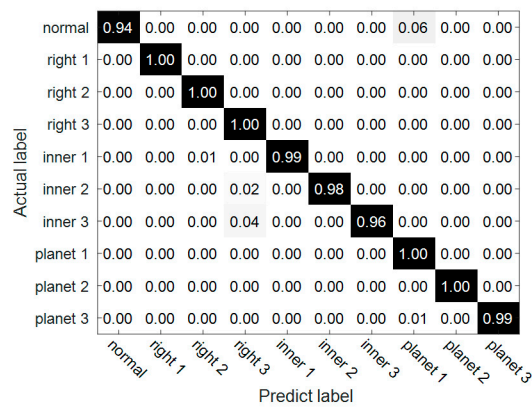
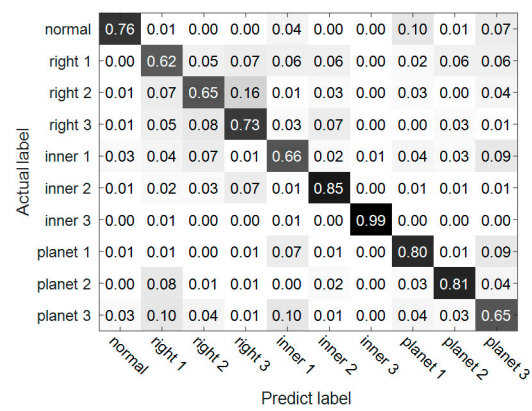


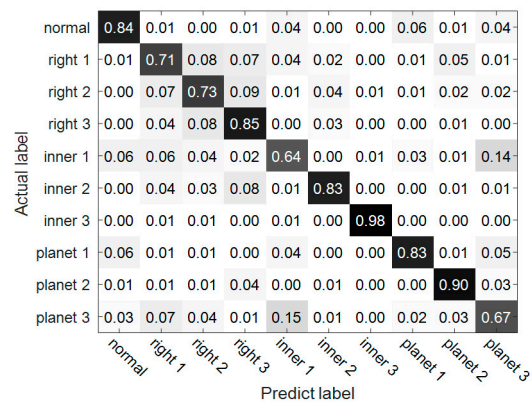
Figure 23. Signal feature dimension reduction using the three methods.



(a) SAFC



(b) BNAE



(c) L1/2-SF

Figure 24. Training sample distribution and mixed matrix comparison.

### 5. Conclusions

A SAFC architecture was proposed to solve the problem of fault diagnosis. This model has a strong feature extraction capability in which SAE is used to reduce and cluster the data, while SF is used to extract and classify the features. A good diagnostic performance is demonstrated under different workloads.

The results in Section 4 show that SAFC can not only extract domain invariant features at different speeds, but can also well classify the extracted features according to different parts and fault locations.

SAFC had a higher diagnostic accuracy and stability than SAE or SF alone. When dealing with large data volumes, SAFC is able to extract domain invariant features quickly thanks to the rapid training data capability of SAE. When classifying extracted features, SAFC has a high diagnostic accuracy as a benefit of the strong feature extraction capability of SF.

In future work, we will further study the classification of different parts and consider the diagnostic impact of sample imbalance.

**Author Contributions:** Conceptualization, R.D. and K.X.; methodology R.D. and J.W.; software, R.D.; validation, R.D. and J.L.; formal analysis, R.D. and S.L.; investigation S.L.; resources S.L.; data curation K.X.; writing—original draft preparation R.D.; writing—review and editing R.D.; visualization R.D.; supervision J.L.; project administration S.L.; funding acquisition R.D., J.L. and S.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Major national Science and Technology Projects (2017-IV-0008-0045), the Advance Research Field Fund Project of China (61400040304), and the National Natural Science Foundation of China (51975276).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lei, Y. *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery*; Xi'an Jiaotong University Press: Xi'an, China, 2017.
2. Shen, C. Research on Fault Diagnosis and Prediction Methods of Key Components of Rotating Machinery Equipment. Ph.D. Thesis, University of Science and Technology of China, Changsha, China, 2014.
3. Hinton, G.; Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
4. Luciano, L.; Hamza, A.B. Deep learning with geodesic moments for 3D shape classification. *Pattern Recognit. Lett.* **2018**, *105*, 182–190.
5. Deng, S.; Cheng, Z.; Li, C.; Yao, X.; Chen, Z.; Sanchez, R.-V. Rolling Bearing Fault Diagnosis Based on Deep Boltzmann Machines. In Proceedings of the IEEE Prognostics and System Health Management Conference, Chengdu, China, 19–21 October 2017; pp. 1–6.
6. Ding, X.; He, Q. Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis. *IEEE Trans. IM* **2017**, *66*, 1926–1935. [[CrossRef](#)]
7. Li, C.; Sanchez, R.; Zurita, G.; Cerrada, M.; Cabrera, D.; Vásquez, R. Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. *Mech. Syst. Sign. Process.* **2016**, *76*, 283–293. [[CrossRef](#)]
8. Liu, C.; Cheng, G.; Chen, X.; Pang, Y. Planetary Gears Feature Extraction and Fault Diagnosis Method Based on VMD and CNN. *Sensors* **2018**, *18*, 1523. [[CrossRef](#)]
9. Chen, Z.; Li, C.; Sánchez, R. Multi-layer neural network with deep belief network for gearbox fault diagnosis. *J. Vibroeng.* **2015**, *17*, 2379–2392.
10. Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–127. [[CrossRef](#)]
11. Jia, F.; Lei, Y.; Lin, J.; Zhou, X.; Lu, N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **2016**, *72*, 303–315. [[CrossRef](#)]
12. Ngiam, J.; Pang, W.K.; Chen, Z.; Bhaskar, S.A.; Koh, P.W.; Ng, A.Y. Sparse Filtering. In Proceedings of the International Conference on Neural Information Processing Systems, Granada, Spain, 12–17 December 2011; pp. 1125–1133.
13. Lei, Y.; Jia, F.; Lin, J.; Xing, S.; Ding, S.X. An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. *IEEE Trans. Ind. Electron.* **2016**, *63*, 3137–3147. [[CrossRef](#)]
14. Chen, R.; Chen, S.; He, M.; He, D. Rolling bearing fault severity identification using deep sparse auto-encoder network with noise added sample expansion. *Proc. Inst. Mech. Eng. Part O. J. Risk Reliability* **2017**, *231*, 666–679. [[CrossRef](#)]
15. Li, H.; Zhang, H.; Qin, X.R.; Zhou, P. Bearing fault diagnosis method based on SHORT-time Fourier Transform and convolutional neural network. *Vib. Impact* **2018**, *37*, 124–131.

16. Wang, W.; Qiu, X.; Sun, J.; Wang, F. Fault diagnosis method of Gear based on double-layer Long-length memory network. *J. Acad. Armored Force Eng.* **2018**, *2*, 81–86.
17. Xu, K.; Li, S.; Wang, J.; An, Z.; Qian, W.; Ma, H. A novel convolutional transfer feature discrimination network for imbalanced fault diagnosis under variable rotational speed. *Meas. Sci. Technol.* **2019**, *30*, 10. [[CrossRef](#)]
18. Han, B.; Wang, X.; Ji, S.; Zhang, G.; Jia, S.; He, J. Data-enhanced Stacked Autoencoders for Insufficient Fault Classification of Machinery and Its Understanding Via Visualization. *IEEE Access* **2020**, *8*, 67790–67798. [[CrossRef](#)]
19. Chakraborty, S. Learning representation for multi-view data analysis: Models and applications. *Comp. Rev.* **2019**, *60*, 284.
20. Han, Y.; Lee, K. Detecting fingering of overblown flute sound using sparse feature learning. *EURASIP J. Audio Speech Music Proc.* **2016**, *2016*, 1–10. [[CrossRef](#)]
21. Zhang, Z.; Li, S.; Wang, J.; Xin, Y.; An, Z. General normalized sparse filtering: A novel unsupervised learning method for rotating machinery fault diagnosis. *Mech. Syst. Signal Proc.* **2019**, *124*, 596–612. [[CrossRef](#)]
22. Lei, Y. Individual intelligent method-based fault diagnosis. *Intell. Fault Diagn. Remain. Useful Life Predict. Rotating Mach.* **2017**, *10*, 167–174.
23. Zhang, Z.; Chen, H.; Li, S.; Wang, J. A new method for intelligent fault diagnosis based on time-frequency feature extraction and Softmax regression. *J. Central S. Univ.* **2019**, *26*, 1607–1618. [[CrossRef](#)]
24. George, S.; Hamza, A.B. Fault Detection Using Robust Multivariate Control Chart. *Expert Syst. Appl.* **2009**, *36*, 5888–5894.
25. Jinrui, W.; Shunming, L.; Zenghui, A.; Xingxing, J.; Weiwei, Q.; Shanshan, J. Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis of machines. *Neurocomputing* **2019**, *329*, 53–65.
26. Jinrui, W.; Shunming, L.; Han, B.; An, Z.; Xin, Y.; Qian, W.; Wu, Q. Construction of a batch-normalized autoencoder network and its application in mechanical intelligent fault diagnosis. *Meas. Sci. Technol.* **2019**, *30*, 015106.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).