Article

A Novel Deep Feature Learning Method Based on the Fused-Stacked AEs for Planetary Gear Fault Diagnosis

Xihui Chen [1], Aimin Ji [1,*] and Gang Cheng [2]

[1]   College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213022, China;
      chenxh@hhu.edu.cn
[2]   School of Mechatronic Engineering, China University of Mining and Technology, Xuzhou 221116, China;
      cumtchg@gmail.com
*   Correspondence: jiam@hhuc.edu.cn; Tel.: +86-135-0614-6597

check for
updates

Abstract: Planetary gear is the key component of the transmission system of electromechanical equipment for energy industry, and it is easy to damage, which affects the reliability and operation efficiency of electromechanical equipment of energy industry. Therefore, it is of great significance to extract the useful fault features and diagnose faults based on raw vibration signals. In this paper, a novel deep feature learning method based on the fused-stacked autoencoders (AEs) for planetary gear fault diagnosis was proposed. First, to improve the data learning ability and the robustness of feature extraction process of AE model, the sparse autoencoder (SAE) and the contractive autoencoder (CAE) were studied, respectively. Then, the quantum ant colony algorithm (QACA) was used to optimize the specific location and key parameters of SAEs and CAEs in deep learning architecture, and multiple SAEs and multiple CAEs were stacked alternately to form a novel deep learning architecture, which gave the deep learning architecture better data learning ability and robustness of feature extraction. The experimental results show that the proposed method can address the raw vibration signals of planetary gear. Compared with other deep learning architectures and shallow learning architecture, the proposed method has better diagnosis performance, and it is an effective method of deep feature learning and fault diagnosis.

Keywords: planetary gear; the fused-stacked AEs; deep feature learning; fault diagnosis

1. Introduction

With the rapid development of science and industrial technology, the electromechanical equipment for energy industry is becoming more and more efficient, reliable, and intelligent [1]. Due to the advantages of planetary gear, it has become a key component of the transmission system of electromechanical equipment for energy industry, such as the wind turbine, shearer, and so on. However, planetary gear generally works in harsh working conditions with heavy load, strong interference and high pollution, so faults often occur, which directly affect the transmission efficiency and even lead to catastrophic accidents of the electromechanical equipment in the energy industry [2,3]. Therefore, it is necessary to carry out the state monitoring and diagnostic analysis for planetary gear. The planetary gear composed of sun gear, planet gears, planet carrier, and inner ring gear has a more complex structure than fixed-axis gear, which provides a special form of motion [4]. The vibration signal measured on the shell of planetary gear is the superposition signal of the vibration excitation of multi-tooth meshing. Meanwhile, the special movement of planetary gear which causes its vibration signal is affected by the "passing effect" of planet gear and planet carrier and makes the transmission

paths among multi-tooth meshing points and the installation position of vibration sensors time-varying. (These factors give the measured vibration signal of planetary gear has the characteristics with stronger nonstationary, frequency modulation and amplitude modulation [5]).

To date, some scholars have performed related research on the fault diagnosis of planetary gear [6]. Chen [7] used a dual-tree complex wavelet transform to decompose the vibration signal of planetary gear, and original fault features were extracted based on the entropy features with multiple perspectives. Then, the optimized kernel Fisher discriminant analysis was used to extract the sensitive features. Finally, the fault recognition was achieved using the nearest neighborhood model. Feng [8] decomposed the complex vibration signal into a series of intrinsic mode functions (IMFs) by variational mode decomposition (VMD), and the sensitive IMFs containing the gear fault information were selected for further demodulation analysis. The fault diagnosis of planetary gear was detected according to the characteristics of demodulation spectrum. Liang [9] proposed a windowing and mapping strategy to extract the fault feature generated by the single cracked tooth of planetary gear. Cerrada [10] studied the feature selection method through attribute clustering. The irrelevant features can be removed and a diagnostic system with good performance can be obtained using the random forest classifier. However, the above fault diagnosis methods of planetary gear are based on the traditional fault diagnosis ideas, and they also have some shortcomings. The main shortcomings are reflected in the following aspects: (1) The traditional fault diagnosis methods need to extract the fault feature artificially using the relevant signal processing algorithm, and the feature extraction relies on human participation and experience, which creates a lack of automation; (2) the designed artificial fault feature extraction method is specially designed according to signal characteristics in advance, instead of acquiring fault features through active learning of the data, which is not universal enough to the object and conditions; (3) the traditional identification methods based on neural network used in fault diagnosis only have shallow architecture, but the shallow architecture is limited by the learning ability for the nonlinear features, and the advantages of neural network cannot be fully exploited [11]. Generally speaking, those methods require too much manual participation and depend on the complex signal process technologies. The feature extraction process lacks initiative and the effective features cannot be directly learned from the raw data. Aiming at the existing problems of feature extraction, deep learning provides an effective approach. It learns the deep features with abstract and expressive directly from raw data based on the bottom layer of deep learning architecture, and the features with more abstract and more expressive can be learned on the basis of the later layer. Finally, a highly complex architecture with multilayers can be automatically learned to express the deep features of raw data [12]. Deutsch [13] proposed a feature extraction method for rotating machine under variable operation condition based on deep learning, and the life prediction can be achieved on the deep features. Oh [14] studied a deep learning method based on deep belief network, which improved the adaptability of feature extraction process. Gao [15] built a deep belief network by stacking the restricted Boltzmann machines (RBMs), and the network was applied to the fault diagnosis of aircraft fuel system. In addition to the above methods, the idea of using AEs to build deep learning architecture has also been studied. Liu [16] built the deep learning architecture based on the stacking of multiple basic AEs, and the feature extraction and fault diagnosis of gearbox can be realized. Feng studied how to build the deep learning architecture with AEs and its improved model SAEs, respectively, and the experiments showed that the proposed methods are promising tools [17,18]. Thirukovalluru [19] built a deep learning architecture based on denoising autoencoder, and the bearing fault diagnosis of air compressor can be realized. The deep learning architecture can be constructed based on the AE model and its improved models, but different improved AE models have their own characteristics and single dominant focus, and the advantages of deep learning in feature extraction cannot be fully exploited. Therefore, constructing the novel deep learning architecture that can fully play to the advantages of different improved AE models and carrying out in-depth research on the application of deep learning in feature extraction are effective ways to achieve and promote the fault diagnosis performance of planetary gear.

This paper is structured as follows: In Section 2, the related theories of basic AE model are introduced. In Section 3, the modeling process of the novel deep feature learning method proposed in this paper is described in detail. In Section 4, the fault simulation experiment for planetary gear is conducted. In Section 5, the vibration signals of planetary gear are processed in detail by the proposed method and the experimental results are obtained. In Section 6, the further analysis and discussion of the experimental results are carried out. Finally, the conclusions are obtained in Section 7.

2. Basic Autoencoder (AE)

AE is an unsupervised machine learning structure, which is composed of an encoder and a decoder. The feature transform of the inputted data is conducted by encoder network, the inputted data is mapped from high-dimensional space to low-dimensional space, and the feature representation of the inputted data can be obtained. Then, the low-dimensional data is mapped to the high-dimensional space by decoder network, and the output-to-input reproduction process can be realized [20]. The structure of AE model is shown in Figure 1.
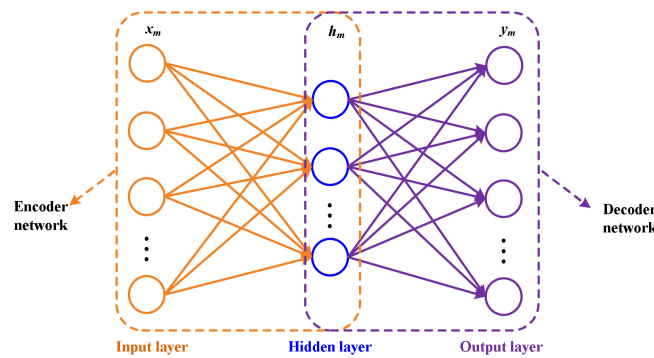

Figure 1. The structure of autoencoder (AE) model.

Assuming that there are M unlabeled training samples $x = \{x_1, x_2, \cdots, x_M\}$, and each sample has P observation vectors (for each sample x_m, $x_m = \{z_1, z_2, \cdots z_P\}$). The encoder process and decoder process of AE are as follows:

$$
\begin{aligned}
h_m &= f_\theta(x_m) = \sigma_1(W x_m + b) \\
y_m &= g_{\widetilde{\theta}}(h_m) = \sigma_2(\widetilde{W} h_m + d)
\end{aligned}
\tag{1}
$$

where f_θ is the encoder function and σ_1 is the activation function of encoder network. W is the weight matrix between input layer and hidden layer, b is the bias vector of encoder network, and the encoder parameter is $\theta = \{W, b\}$. The reconstructed values y_m can be obtained by decoder network, and g_θ is the decoder function, σ_2 is the activation function of decoder network. $\widetilde{\theta} = \{\widetilde{W}, d\}$ is the decoder parameter that is trained from AE. In order to make y_m equal to x_m, the minimize cost function $L(x, y)$ is used to describe the proximity between input and output, and the definition of $L(x, y)$ is as follows:

$$
L(x, y) = \frac{1}{M} \sum_{m=1}^{M} \|y_m - x_m\|^2
\tag{2}
$$

Generally speaking, a deep learning architecture can be formed by stacking multiple AEs. The corresponding training algorithm can be used to train the deep learning architecture composed of multiple AEs, so that it has the strong deep feature learning ability and can be used to learn and mine the feature information from a large number of raw data. However, the basic AE model still has some shortcomings. The deep learning architecture based on multiple AEs does not fully exploit the advantages of deep learning in feature extraction and its learning effect can be further improved [21].

3. The Proposed Method

The excellent fault feature extraction algorithm should be sensitive to faults to reflect the states of different fault types. Meanwhile, the robustness and adaptability of feature extraction process should be guaranteed. Therefore, a novel deep learning architecture for deep feature learning based on the fused-stacked AEs for planetary gear was proposed, and it composed of two types of improved structures of basic AE model: The SAE and CAE. SAE can effectively improve the data learning ability, and CAE can effectively guarantee the robustness of feature extraction. It is worth noting that for the deep learning architecture composed of multiple SAEs and multiple CAEs, the specific locations and parameter setting of each SAE and CAE play an important role in the effect of deep feature learning and fault diagnosis. Therefore, the specific locations and key parameters of each SAE and CAE in deep learning architecture were optimized using optimization algorithm to form a novel deep learning architecture based on the fused-stacked AEs. The deep learning architecture can achieve the best data learning ability and feature extraction robustness.

3.1. Sparse Autoencoder (SAE)

SAE is an improvement to the basic AE model. It relies on the idea of sparse coding, and the sparsity constraint is added into the AE model [22]. The average activation degree of hidden neurons is limited and the data learning ability of AE can be enhanced. Assuming that the activation value of the j-th neuron in the s-th hidden layer is $H_i^{(s)}(x)$ in the case of training samples x, the average activation degree of the s-th hidden layer with m neurons in the training samples can be expressed as follows:

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} [h_j^{(s)}(x_{(i)})] \tag{3}$$

Due to the sparsity requirement, it is hoped that most of the hidden neurons will not be activated, so the average activation degree $\hat{\rho}_j$ tends to a near zero constant ρ, and ρ is a sparsity parameter. Therefore, the additional penalty item is added to the cost function of AE to punish that $\hat{\rho}_j$ is not deviate from ρ. In this paper, the kullback-leibler (*KL*) divergence was used to achieve the purpose of punishment. The expression of the penalty item *PN* is as follows:

$$PN = \sum_{j=1}^{s_2} KL(\rho \parallel \rho_j) \tag{4}$$

where s_2 is the number of hidden neurons, $KL(\rho \parallel \rho_j)$ is the *KL* divergence, and the mathematical expression of *KL* divergence is as follows:

$$KL(\rho \parallel \rho_j) = \rho \ln \frac{\rho}{\rho_j} + (1 - \rho) \ln \frac{1 - \rho}{1 + \rho_j} \tag{5}$$

The penalty item is determined by the nature of *KL* divergence. If $\rho_j = \rho$, then $KL(\rho \parallel \rho_j) = 0$. Otherwise, the *KL* divergence value will gradually increase with ρ_j deviating from ρ. When the sparsity penalty item is added, the sparsity cost function of SAE can be expressed as follows:

$$J_{\text{SAE}} = L(x, y) + \beta \sum_{j=1}^{s_2} KL(\rho \parallel \rho_j) \tag{6}$$

where β is the weight of sparsity penalty item.

Therefore, the training of SAE can be realized by minimizing the sparsity cost function.

$$(W, b)_{\text{SAE}} = \min J_{\text{SAE}} \tag{7}$$

3.2. Contractive Autoencoder (CAE)

CAE is an improvement to the basic AE model, in which a contractive restriction is added into the AE model. The disturbance of training samples in all directions can be suppressed by restricting the Jacobian matrix of the output weight of hidden layer and the freedom of the feature representation can be reduced. The output data can be limited in a certain range of parameter space, and the robustness of feature extraction of AE model can be enhanced [23]. Therefore, a penalty term is added into the cost function of CAE to achieve the effect of local space contraction, which is the *F*-norm of Jacobian matrix of the output weight of hidden layer. The contractive cost function of CAE can be expressed as follows:

$$J_{\text{CAE}} = L(x, y) + \lambda \left\| J_f(x) \right\|_F^2 \tag{8}$$

where $L(x, y)$ is the basic cost function and $\left\| J_f(x) \right\|_F^2$ is the contractive penalty term. λ is penalty parameter, and its role is to adjust the proportion of contractive penalty item in the cost function.

The specific formula of the contractive penalty item can be expressed as follows:

$$\left\| J_f(x) \right\|_F^2 = \sum_{ij} \left(\frac{\partial h_j(x)}{\partial x_i} \right)^2$$

$$J_f(x) = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_1} \end{bmatrix} \tag{9}$$

where $J_f(x)$ is the Jacobian matrix of the output weight of hidden layer. If the penalty term has a small first-order derivative, the expression of hidden layer corresponding to training samples is smooth. Then, when the training samples change, the expression of hidden layer will not change greatly, which can achieve the purpose of insensitive to the changing of training samples [24].

Therefore, the training of CAE can be realized by minimizing the contractive cost function.

$$(W, b)_{\text{CAE}} = \min J_{\text{CAE}} \tag{10}$$

3.3. Quantum Ant Colony Algorithm (QACA)

In the proposed novel deep learning architecture composed of multiple SAEs and multiple CAEs, QACA was used to optimize the specific locations and key parameters of each SAE and CAE. QACA is a probabilistic search optimization algorithm which combines quantum computation and the ant colony algorithm. The ant colony algorithm is an optimization algorithm with good robustness and strong optimization ability, but it is easy to converge prematurely and can fall into local optimum. By combining quantum computation and the ant colony algorithm, the probability amplitude of quantum bits can be introduced to represent the current positions of ants and the pheromone coding can be completed [25]. The positions of ants are updated by quantum rotation gate, which gives the combined algorithm better population diversity, faster convergence speed, and global optimization ability.

3.3.1. Quantum Coding and Quantum Rotation Gate

In quantum coding, the quantum bits are used to describe the quantum states. Quantum bits have two possible states $|0\rangle$ and $|1\rangle$, which correspond to the classical bits 0 and 1, respectively. Quantum state $|\varphi\rangle$ can be expressed as $|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle$, and α and β are the probabilistic amplitudes of the quantum state $|\varphi\rangle$, and they satisfy $|\alpha|^2 + |\beta|^2 = 1$. The quantum state $|\varphi\rangle$ is an uncertain superposition state between $|0\rangle$ and $|1\rangle$. The quantum bits are used to encode the pheromones on the path traveled by ants, which is called as quantum pheromones. The quantum state $|\varphi\rangle$ can be converted to real number

pair $(\cos\theta, \sin\theta)$, and θ is the phase of quantum state $|\varphi\rangle$. Assuming that the number of quantum bits of individual X_i is n, and X_i can be expressed as follows:

$$X_i = \begin{bmatrix} \cos\theta_{i1} & \cos\theta_{i2} & \cdots & \cos\theta_{in} \\ \sin\theta_{i1} & \sin\theta_{i2} & \cdots & \sin\theta_{in} \end{bmatrix} \tag{11}$$

After incorporating quantum coding, the updating of the quantum pheromones on the path traveled by ants can be realized using quantum rotation gate. The formula is as follows:

$$\begin{bmatrix} \cos(\varphi_{ij}^{t+1}) \\ \sin(\varphi_{ij}^{t+1}) \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos(\varphi_{ij}^{t}) \\ \sin(\varphi_{ij}^{t}) \end{bmatrix} \tag{12}$$

where $\left[\cos(\varphi_{ij}^{t}), \sin(\varphi_{ij}^{t})\right]$ is the probabilistic amplitudes of quantum bits before quantum rotation gate processing and $\left[\cos(\varphi_{ij}^{t+1}), \sin(\varphi_{ij}^{t+1})\right]$ is the probabilistic amplitudes of quantum bits after quantum rotation gate processing.

3.3.2. Ant Transfer Rules and Transfer Probability

The ant transfer rules of the k-th ant in ant colony from node 1 to node 2 are as follows:

$$s = \begin{cases} \operatorname{argmax}\left\{[\tau_{ij}^{k}(t)]^{\alpha}[\eta_{ij}^{k}(t)]^{\beta}[\varepsilon_{ij}^{k}(t)]^{\gamma}\right\} & q \leq q_0 \\ \widetilde{s} & q > q_0 \end{cases} \tag{13}$$

where q is a random value uniformly distributed in the interior $[0, 1]$, and $q_0 (0 \leq q_0 \leq 1)$ is a constant. S is the set of all possible nodes that the k-th ant may arrive from node i. $\widetilde{s}$ is a target location selected according to the formula below:

$$p_{ij}^{k} = \frac{[\tau_{ij}^{k}(t)]^{\alpha}[\eta_{ij}^{k}(t)]^{\beta}[\varepsilon_{ij}^{k}(t)]^{\gamma}}{\sum\limits_{s\in allowed(i)}[\tau_{ij}^{k}(t)]^{\alpha}[\eta_{ij}^{k}(t)]^{\beta}[\varepsilon_{ij}^{k}(t)]^{\gamma}} \tag{14}$$

where p_{ij}^{k} is the transition probability of the k-th ant. $\tau_{ij}^{k}(t)$ is the pheromone concentration on the path from node i to node j in the t-th iteration. α is the pheromone heuristic factor, and $\eta_{ij}^{k}(t)$ is the distance heuristic information on the path from node i to node j in the t-th iteration, and its expression is $\eta_{ij}^{k}(t) = 1/d_{ij}$, d_{ij} is the distance between node i and node j. β is the distance heuristic factor, $\varepsilon_{ij}^{k}(t)$ is the consumption heuristic information on the path from node i to node j in the t-th iteration, and its expression is $\varepsilon_{ij}^{k}(t) = 1/E_{ij}$, E_{ij} is the energy consumption from node i to node j. γ is the heuristic factor for energy consumption [26].

3.3.3. Pheromone Updating

In the ant colony algorithm, the local updating of pheromone is conducted after each ant moves, and the global updating of the optimal path of all ants is conducted when all ants have completed one iteration. Assuming that the current node of ant is p_i, and the node after movement is p_j, the local updating rules of pheromone is as follows:

$$\tau(p_j) = (1-\rho_1)\cdot\tau(p_i) + \rho_1\cdot\Delta\tau_{ij} \tag{15}$$

where $\tau(p_j)$ is the pheromone concentration of the node after movement, and $\tau(p_i)$ is the pheromone concentration of the current node. $\rho_1(0 < \rho_1 < 1)$ is the local renewal volatilization coefficient of

pheromone, $\Delta\tau_{ij}$ is the pheromone concentration of each ant in the moving path in this iteration, and $\Delta\tau_{ij}$ can be expressed as follows:

$$\begin{cases} \Delta\tau_{ij} = \sum\limits_{k=1}^{n} \Delta\tau_{ij}^{k} \\ \Delta\tau_{ij}^{k} = Q/J_k \end{cases} \tag{16}$$

where $\Delta\tau_{ij}^{k}$ is the path of the k-th ant, Q is a constant, and J_k is the comprehensive cost of the k-th ant in this iteration [27].

The global updating of pheromone is carried out when all ants complete an iteration, and its updating rules are as follows:

$$\tau(p_j) = \begin{cases} (1-\rho_2)\cdot\tau(p_j) + \rho_2\cdot\Delta\tau_{best} & , & p_u = \widetilde{s} \\ (1-\rho_2)\cdot\tau(p_j) & , & p_u \neq \widetilde{s} \end{cases}$$
$$\Delta\tau_{best} = \begin{cases} Q/J_e, & \text{the optimal path from node } i \text{ to node } j \\ 0, & \text{else} \end{cases} \tag{17}$$

where ρ_2 is the global renewal volatilization coefficient of pheromone, Q is a constant, J_e is the comprehensive cost of the obtained optimal path in this iteration, and $\widetilde{s}$ is the current optimal solution.

3.4. The Proposed Method

A novel deep learning architecture based on the fused-stacked AEswas proposed, which combined the advantages of SAE and CAE. The QACA was used to optimize the specific locations and key parameters of multiple SAEs and multiple CAEs. Then, the deep learning architecture was stacked alternately by SAEs and CAEs, and each SAE or CAE formed a hidden layer of the obtained deep learning architecture in the most appropriate locations. The obtained deep learning architecture had the best data learning ability, and the extracted deep features had the best robustness. Meanwhile, determining the depth (the number of hidden layers) and breadth (the number of unit nodes of per hidden layer) of deep learning architecture is a key issue to be addressed. In this paper, the depth and breadth of deep learning architecture were determined by using the idea similar to literature [22,24]. First, the number of input nodes was determined by the length of input samples. Second, the number of hidden layers and the number of unit nodes of per hidden layer should be bigger, which can improve the deep feature learning ability of deep learning architecture. However, the extensive depth and breadth of deep learning architecture will inevitably cause the increase of model complexity, leading to a longer training time and even excessive learning. Third, in general, the number of unit nodes of latter hidden layer should be less than that of the previous hidden layer, which has the function of feature dimension reduction and data compression, and it is conducive to the final recognition; Fourth, the number of unit nodes of output layer was determined by the category number of training samples. In addition, the training process of deep learning architecture is also important. If the parameters of each hidden layer are initialized randomly according to general training method, the local optimal situation becomes serious as the number of hidden layers increases. Therefore, the greedy layer-wise pretraining combined with random gradient descent fine-tuning algorithm proposed by Hinton [28] was used. In this training algorithm, first, the training samples were regarded as unsupervised data, and each hidden layer (each SAE and each CAE) was trained separately according to Equations (3)–(7) or Equations (8)–(10). The training process started from the first hidden layer and was carried out separately layer by layer, allowing the pretraining process to be completed. On this basis, the training samples were set as the supervised data with type labels. All hidden layers in the deep learning architecture were treated as the training object, and the parameters of all hidden layers were fine-tuned using the random gradient descent until the model converges [22]. The structure of the proposed novel deep feature learning method is shown in Figure 2.
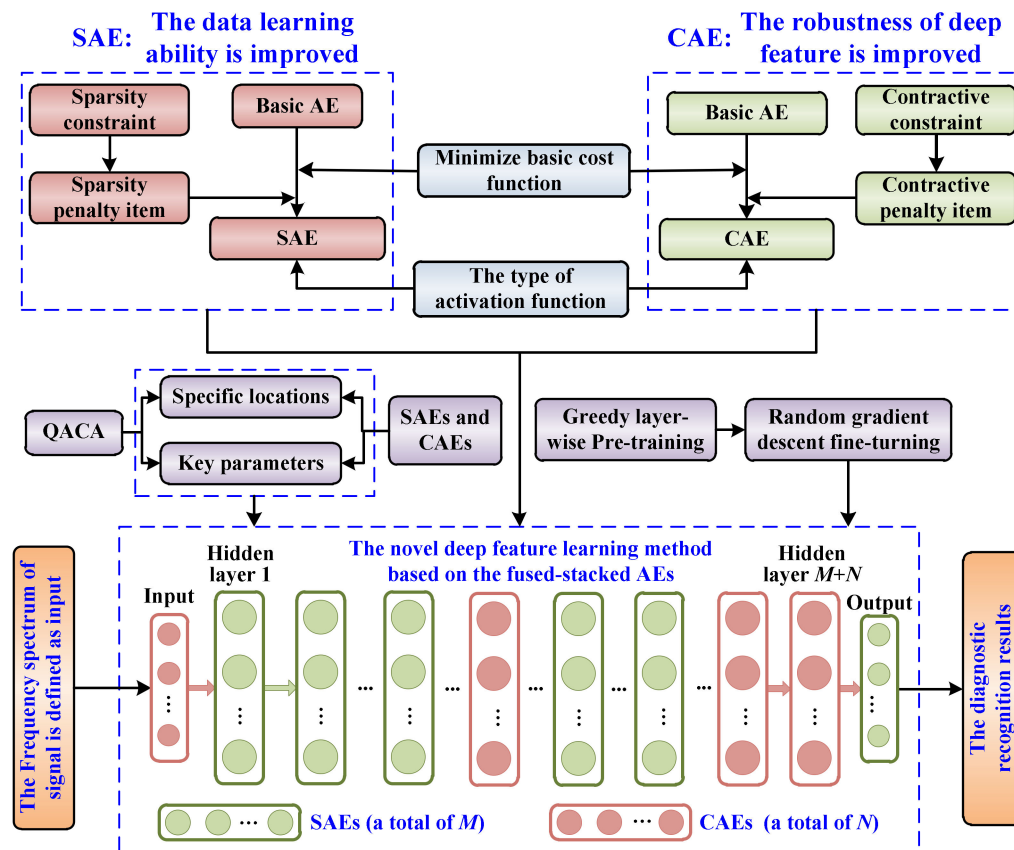
Figure 2. The structure of the proposed novel deep feature learning method.

4. Experiment Introduction

In order to validate the proposed method, the vibration signals of planetary gear with different states were collected on the fault simulation test bench. The fault simulation test bench is shown in Figure 3, and the basic parameters of its planetary gearbox are shown in Table 1. The acceleration vibration sensors installed on the shell of planetary gearbox were used to collect the vibration signals of planetary gear. In addition, in the experiment process, the different sun gear states of planetary gear were simulated, namely the normal gear, gear with one missing tooth, pitting gear, wear gear, broken gear, and cracked gear, as shown in Figure 4. Meanwhile, the setting key parameters in the experiment process are shown in Table 2. Under the setting conditions, the vibration signals of six types of planetary gear states were collected. For each planetary gear state, 320 groups of training samples and 100 groups of testing samples were intercepted and prepared, so there were 1920 groups of training samples and 600 groups of testing samples prepared for the following experimental analysis.

Table 1. Basic parameters of the planetary gearbox.

Structure of Planetary Gearbox	The First Stage Planetary Gear			The Second Stage Planetary Gear		
	Sun Gear	Planet Gear	Ring Gear	Sun Gear	Planet Gear	Ring Gear
Teeth	20	40	100	28	36	100

Table 2. The setting key parameters in the experiment process.

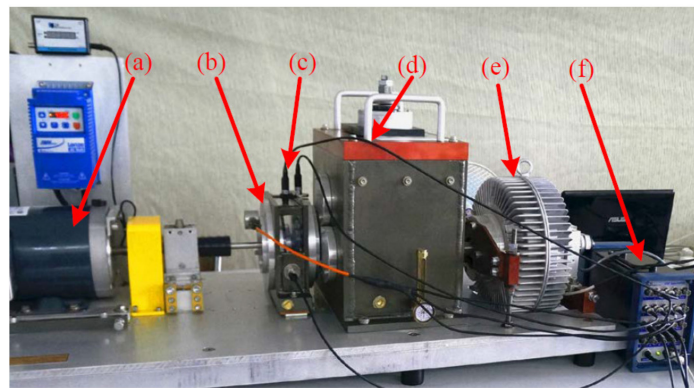| Motor Speed	Sampling Frequency	Load	Sample Length
3000 r/min	6400 Hz	13.5 Nm	3200

Figure 3. Fault experiment for planetary gear, (**a**) controllable motor, (**b**) planetary gearbox, (**c**) acceleration vibration sensors, (**d**) fixed-axis gearbox, (**c**) load system, (**f**) acquisition system.

Figure 4. Six types of planetary gear states.

5. Experimental Results and Analysis

The vibration signals of six types of planetary gear states are shown in Figure 5, and it can be seen from Figure 5 that there were no significant differences among the vibration signals in the time domain. To carry out the experimental analysis process of the proposed method, 1920 groups of training samples and 600 groups of testing samples were obtained on the fault simulation test bench shown in Figure 3 in advance. The training samples were used to train the proposed deep learning model, and the testing samples were used to test the performance of the trained model, so as to verify and analyze the effectiveness of the proposed deep feature learning method. Because the proposed method in this paper is a new form with improvement, there is no existing perfect code, commercial software, or toolbox to implement. Therefore, the code of the research method was self-created on the basis of relevant materials. When executing the proposed algorithm, the hardware platform adopted DELL T7820 tower server, its CPU had four cores and 3.6-G clock speed, and its memory had 64 GB. The software platform was jointly developed on the basis of Matlab and C# software.
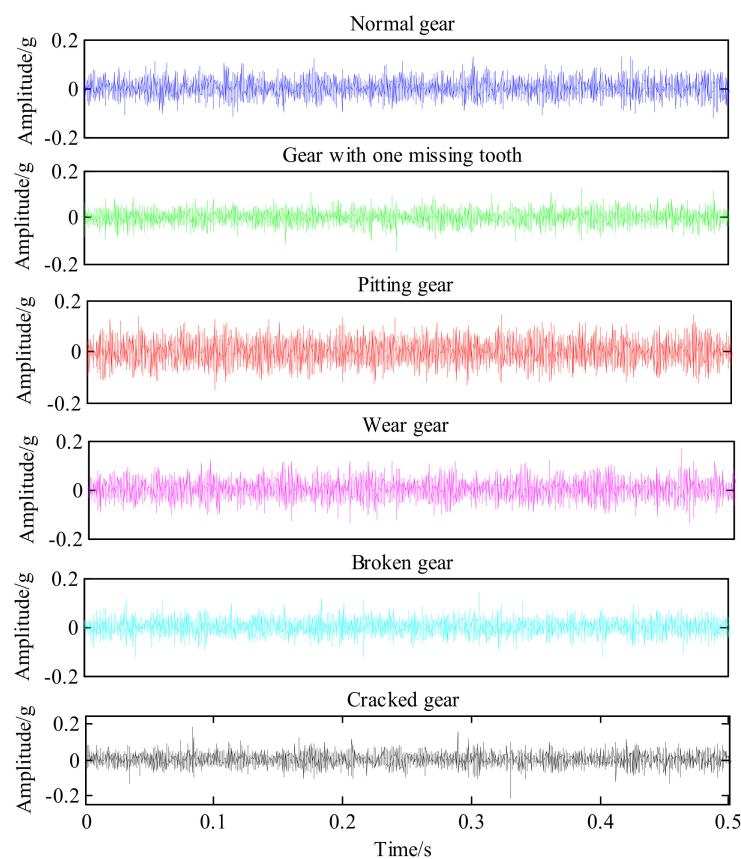
Figure 5. Vibration signals of six types of planetary gear states.

Next, the proposed method in this paper was verified and analyzed. According to the rules in Section 3.4, the length of the input sample was 3200, and the type number of planetary gear states is six. In this case study, the number of hidden layers of the deep learning architecture was selected as eight, and its depth and breadth were determined to be 3200-2600-2000-1600-1200-900-600-300-100-6. The greedy layer-wise pretraining combined with random gradient descent fine-tuning algorithm, was defined as the training method. The composition structure of each hidden layer (i.e., the specific locations of SAEs and CAEs) and their key parameters (the weight of sparsity penalty item β for SAE and the weight of contractive penalty item λ for CAE) were optimized by QACA, and the final diagnostic recognition rate was set as the optimization target in the training process. Finally, the trained deep learning architecture based on the fused-stacked AEs can be obtained. The number of ants in QACA was set as 20, and the number of iterations was set as 90. The optimization process using QACA is shown in Figure 6, and it can be seen that when the number of iterations reached 66, the optimization process tended to be stable and the diagnostic recognition rate for training samples reached 95.26%. It had the largest diagnostic recognition rate. Thus, well-trained deep learning architecture can be obtained. When the optimization was completed, the specific locations of SAEs and CAEs and their key parameters are shown in Table 3.
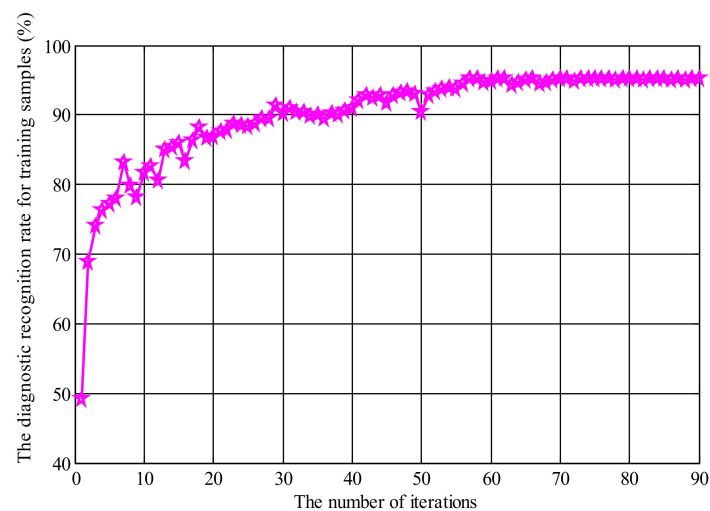
Figure 6. The optimization process using the quantum ant colony algorithm (QACA).

Table 3. The optimization results using the QACA.

Hidden Layers of Deep Learning Architecture	Hidden Layer 1	Hidden Layer 2	Hidden Layer 3	Hidden Layer 4	Hidden Layer 5	Hidden Layer 6	Hidden Layer 7	Hidden Layer 8
The specific location of SAEs and CAEs	SAE1	SAE2	SAE3	CAE1	SAE4	SAE5	CAE2	CAE3
The weight of sparsity penalty item β for SAE	$\beta_1 = 0.235$	$\beta_2 = 0.151$	$\beta_3 = 0.113$	×	$\beta_4 = 0.06$	$\beta_5 = 0.201$	×	×
The weight of contractive penalty item λ for CAE	×	×	×	$\lambda_1 = 0.105$	×	×	$\lambda_2 = 0.231$	$\lambda_3 = 0.217$

The testing samples were used to verify the trained deep learning architecture based on the fused-stacked AEs and a total of 600 testing samples. The diagnostic recognition rates of testing samples of different planetary gear state are shown in Table 4. It can be seen that the testing samples of each planetary gear state had good diagnostic recognition rate. The recognition rate of the normal gear was the highest, and it reached 100%. The recognition rate of the wear gear was the lowest, but it reached 90%. The average recognition rate had a good effect, and it reached 95.5%. Furthermore, in order to further prove the effectiveness and advantages of the proposed method in this paper, some applied methods in other related literatures were used for comparison, namely the deep learning architecture based on basic AEs, the deep learning architecture based on standard SAEs, the deep learning architecture based on standard CAEs, the deep learning architecture based on RBMs, and the shallow learning architecture based on BP neural network. During the process of comparison testing, the used training samples and testing samples were identical, and the diagnostic recognition rates of the testing samples using six methods are compared and shown in Figure 7.

Table 4. The diagnostic recognition rates of testing samples of different planetary gear states.

The Testing Samples for Different Planetary Gear States	Diagnostic Recognition Rate
Normal gear	100%
Gear with one missing tooth	97%
Pitting gear	94%
Wear gear	90%
Broken gear	97%
Cracked gear	95%
Average diagnostic recognition rate	95.5%

(**a**)

(**b**)

(**c**)

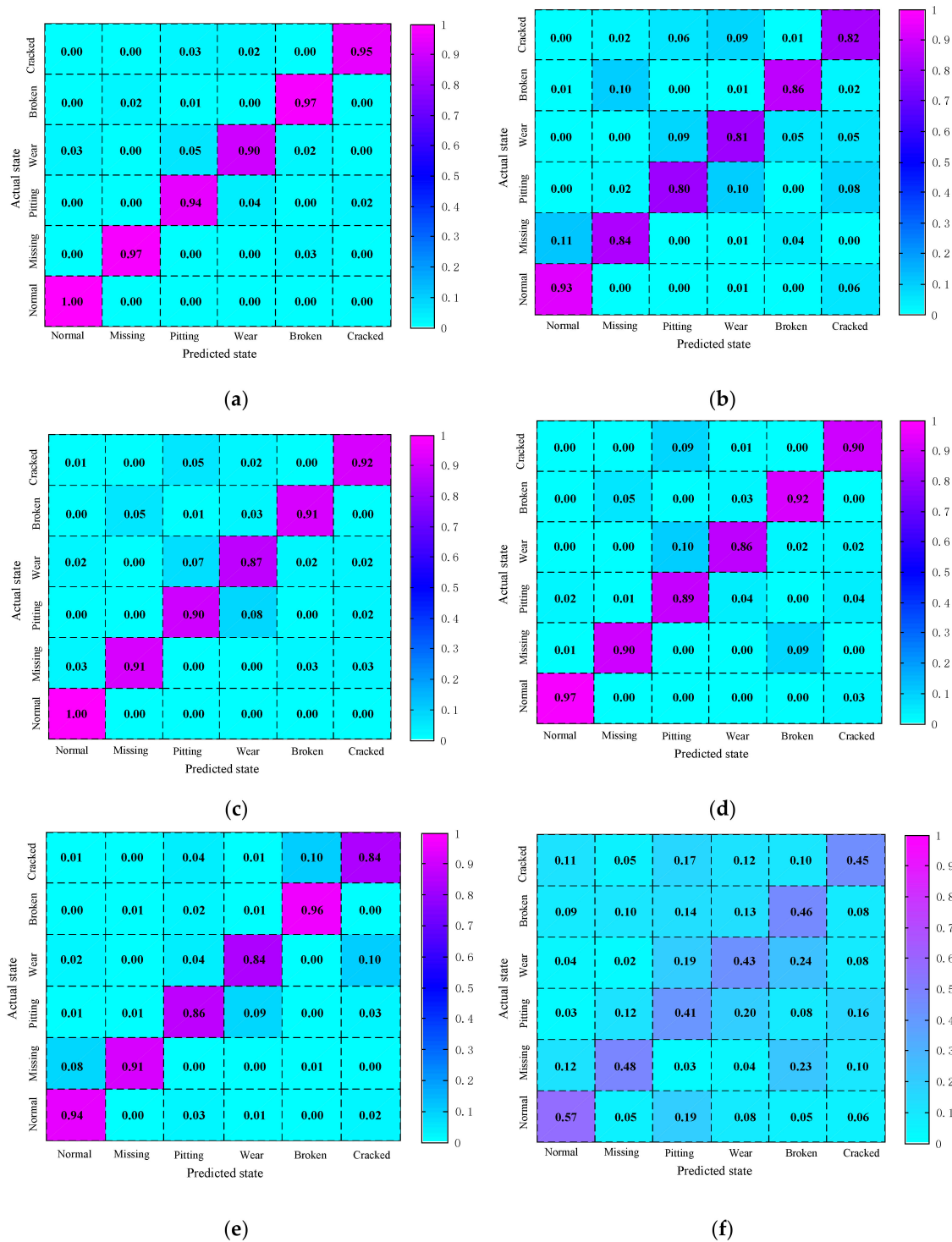(**d**)

(**e**)

(**f**)

Figure 7. The diagnostic recognition rates of the testing samples using six methods. (**a**) The deep learning architecture proposed in this paper; (**b**) the deep learning architecture based on basic autoencoders (AEs); (**c**) the deep learning architecture based on standard sparse autoencoders (SAEs); (**d**) the deep learning architecture based on standard contractive autoencoders (CAEs); (**e**) the deep learning architecture based on restricted Boltzmann machines (RBMs); (**f**) the shallow learning architecture based on back propogation (BP) neural network.

The average diagnostic recognition rate of each method can be further calculated according to Figure 7. The average diagnostic recognition rate of the proposed method is 95.5%, which is higher than

that of the deep learning architectures based on standard SAEs, the deep learning architecture based on standard CAEs, the deep learning architectures based on RBMs, the deep learning architectures based on basic AEs, and the shallow learning architecture based on BP neural network, which are 91.83%, 90.67%, 89.17%, 84.33%, and 46.67%, respectively. By comparison, it is obvious that the deep learning architectures have higher performance than the shallow learning architecture. This is because the valuable feature information can be obtained from raw data through the nonlinear transformation with multilayers in deep learning architecture, while the inadequate transformation processing in the shallow learning architecture is not enough to obtain the sufficient valuable information. In addition, both SAE and CAE are improved forms of the AE model, so the performance of the deep learning architecture based on basic AEs is lower than that of its improved form. For the analysis of the diagnostic recognition rate of each planetary gear state, it can be seen from the comparison that the proposed method had the best performance, and the diagnostic recognition results of the proposed method were all greater than that of other four deep learning architectures. This is because that the proposed deep learning architecture was obtained by the stacking of SAEs and CAEs, and the specific locations of multiple SAEs and multiple CAEs were optimized by QACA, which can provide the best data learning ability, and the robustness of feature extraction is also can be further maintained. Meanwhile, it can be seen from Figure 7 that there was confusion of state recognition among several planetary gear states, such as the pitting gear and wear gear state. This is mainly because that those fault states were relatively slight, and the signal feature information among those faults was similar.

6. Discussion

The advantages of the proposed method in the aspects of deep feature learning and the robustness of feature extraction are further discussed in the following section, and the deep feature learning process with layer-by-layer of deep learning architecture is deeply analyzed. Because the extracted features in each hidden layer were high-dimensional data (2600-2000-1600-1200-900-600-300-100), it was impossible to display the features completely, so the first two projected features in each hidden layer were selected for visualization. The first two projected deep features in each hidden layer of the proposed method are shown in Figure 8. In addition, in order to further illustrate the robustness of the deep feature learning process with layer-by-layer, the coefficient of variance (CV) was used to measure the discreteness of the first two projected deep features, which can eliminate the effects of the measurement scales and dimensions. The CVs of the first two projected deep features in each hidden layer are shown in Figures 9 and 10, respectively.
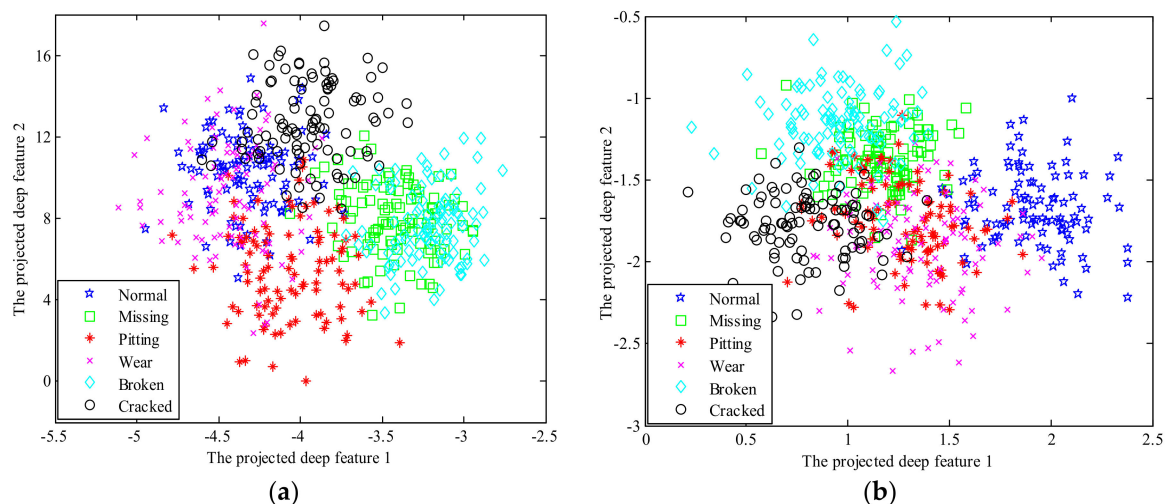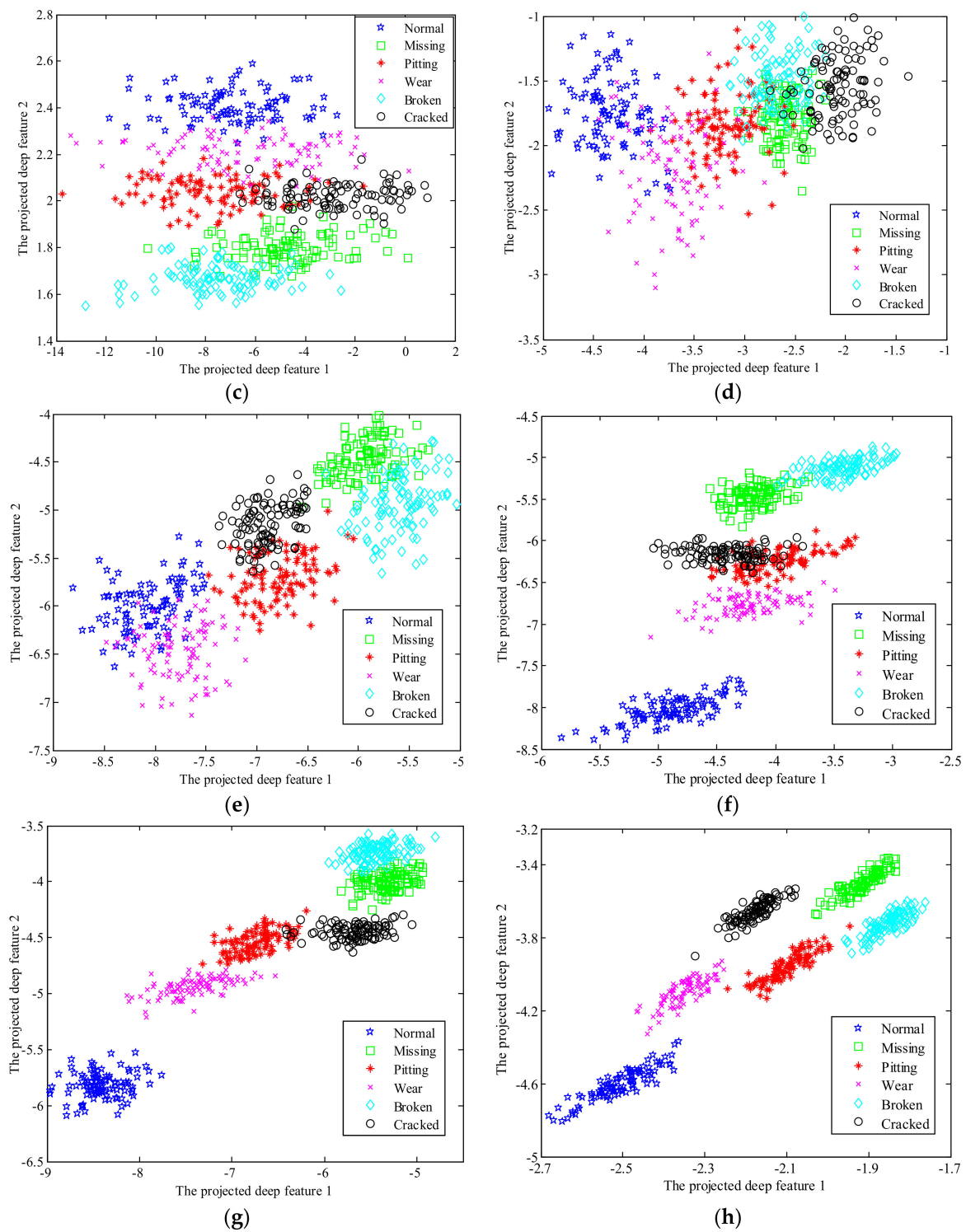

Figure 8. *Cont.*

Figure 8. The first two projected deep features in each hidden layer processed by the proposed method. (**a**) Hidden layer 1; (**b**) hidden layer 2; (**c**) hidden layer 3; (**d**) hidden layer 4; (**e**) hidden layer 5; (**f**) hidden layer 6; (**g**) hidden layer 7; (**h**) hidden layer 8.
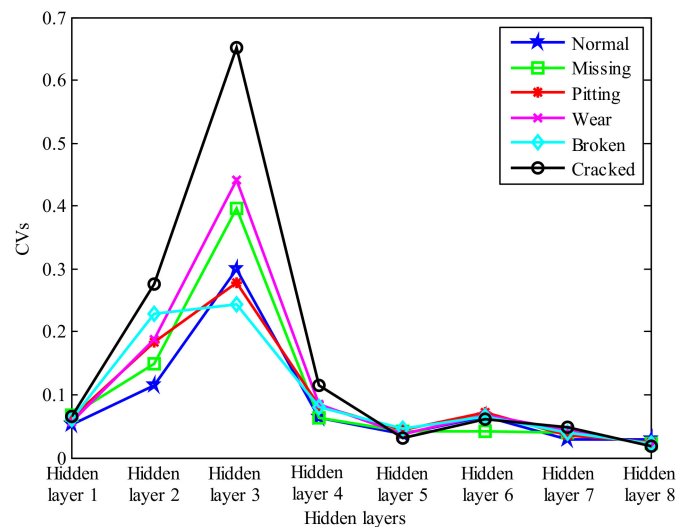
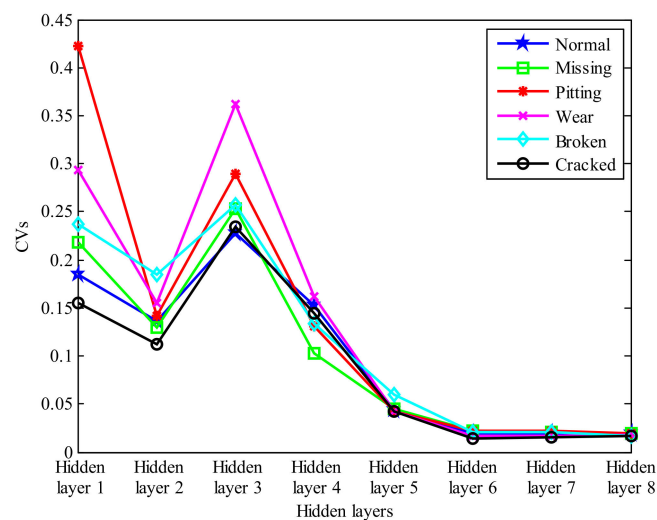**Figure 9.** The coefficients of variance (CVs) of the projected deep feature 1 in each hidden layer.

Figure 10. The CVs of the projected deep feature 2 in each hidden layer.

According to the analysis of Figures 8–10 and Section 5, we can summarize that: (1) In the proposed deep learning architecture based on the fused-stacked AEs, the lower hidden layers were mainly composed of SAEs, which focused on the feature learning ability from raw vibration signal. Because the input was the raw vibration signal without any processing, the extracted deep features of each planetary gear state were seriously confused. The higher hidden layers were mainly composed of CAEs, which focused on the distinguishability and robustness of deep feature extraction process based on the useful information learned from the previous layer. The established deep learning architecture makes full use of the powerful ability of complex nonlinear transformation of multiple hidden layers, and the deep feature layer-by-layer learning process can directly learn the effective deep features from raw vibration signal. (2) With the increase of the number of hidden layers, the distinguishability of the extracted deep features of each planetary gear state in each hidden layer was greatly improved. However, there was still the overlap among different planetary gear states on the same feature, such as the broken gear and the gear with one missing tooth on the projected deep feature 1, and the pitting gear and wear gear on the projected deep feature 2. The main reason is that these fault states are similar, which results in similar fault feature information and makes it difficult to distinguish them by single deep feature. Therefore, the final diagnosis of planetary gear states needs to be realized by combining multiple deep features. (3) The proposed deep learning architecture in this paper was

stacked alternately by SAEs and CAEs, and the specific locations and key parameters of each SAE and CAE were determined by QACA. In the first three hidden layers, the CVs of the extracted deep features were larger, and the discreteness of the extracted deep features of each planetary gear states was larger. The structure of hidden layer 4, determined by optimization algorithm, was CAE. The advantages of CAE in guaranteeing robustness and aggregation of feature extracting process caused the CVs of the extracted deep features in hidden layer 4 to decrease greatly, and the discreteness of the extracted deep features of each planetary gear state was smaller. The structures of the last few hidden layers were mainly CAEs, and the CVs of the extracted deep features tended to be constant. The aggregation and dispersion of the extracted deep features tended to be stable, and the deep feature learning process of the proposed deep learning architecture reached its optimal state. (4) The diagnostic recognition rate of the proposed deep feature learning method was higher than that of other methods, and this result is mainly attributed to the reasonable fusion of SAEs and CAEs. SAEs and CAEs with the most appropriate specific positions in deep learning architecture allowed each SAE and CAE to play to their own characteristics and advantages. The idea of fusing multiple SAEs and multiple CAEs to construct the deep learning architecture can be further developed and applied to other occasions. Generally, there are many basic models that can be used to build deep learning architecture. Because this paper only focused on the data learning ability and robustness of feature learning, SAE and CAE were used. But other basic models, such as the denoising autoencoder, variational autoencoder, restricted Boltzmann machine, and so on, can also be applied, and they have their own unique advantages. Therefore, for the specific problems and functional requirements, the appropriate basic models can be chosen and fused according to their own unique advantages, and the deep learning architecture can be constructed according to the ideas of this paper. Additional kinds of basic models can be fused to construct the deep learning architecture to provide better performance. This is also our next major research aim. The above analysis proves that the optimization algorithm, which was used to determine the specific locations of SAEs and CAEs, is feasible. Thus, the construction idea of the proposed novel deep learning architecture was also proved to be reasonable. The proposed novel deep learning architecture can directly process the raw vibration signals, and it is an effective and feasible method for deep feature learning and fault diagnosis of planetary gear.

7. Conclusions

In this paper, considering that the excellent feature extraction method should be sensitive to different faults of planetary gear, and that the robustness and adaptability of feature extraction process should be guaranteed, a novel deep feature learning method based on the fused-stacked AEs for planetary gear fault diagnosis was proposed. The proposed method composed of two types of improved structures of basic AE model: SAE and CAE, and the specific locations and key parameters of each SAE and CAE in deep learning architecture were optimized by QACA. A novel deep learning architecture based on the fused-stacked AEs can be established, which has both the powerful ability of SAE in data learning and the strong ability of CAE in guaranteeing the robustness of feature extraction. The experiments showed that the proposed method can directly process the raw vibration signals of planetary gear without any preprocessing, and the average diagnostic recognition rate for planetary gear reached 95.5%. Compared with other deep learning architectures and shallow learning architecture, the proposed method has better performance in deep feature learning and fault diagnosis.

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

References

1. Wang, F.T.; Liu, C.X.; Su, W.S.; Xue, Z.G.; Han, Q.K. Combined failure diagnosis of slewing bearings based on MCKD-CEEMD-ApEn. *Shock Vib.* **2018**, *2018*, 6321785. [CrossRef]
2. Fu, L.; Zhu, T.T.; Zhu, K.; Yang, Y.L. Condition monitoring for the roller bearings of wind turbines under variable working conditions based on the fisher score and permutation entropy. *Energies* **2019**, *12*, 3085. [CrossRef]
3. Lu, S.L.; He, Q.B.; Wang, J. A review of stochastic resonance in rotating machine fault detection. *Mech. Syst. Signal Process.* **2019**, *116*, 230–260. [CrossRef]
4. Zhao, C.; Feng, Z.P.; Wei, X.K.; Qin, Y. Sparse classification based on dictionary learning for planet bearing fault identification. *Expert Syst. Appl.* **2018**, *108*, 233–245. [CrossRef]
5. Chen, X.H.; Li, H.Y.; Cheng, G.; Peng, L.P. Study on planetary gear degradation state recognition method based on the features with multiple perspectives and LLTSA. *IEEE Access* **2019**, *7*, 7565–7576. [CrossRef]
6. Kandukuri, S.T.; Klausen, A.; Karimi, H.R.; Robbersmyr, K.G. A review of diagnostics and prognostics of low-speed machinery towards wind turbine farm-level health management. *Renew. Sust. Energy. Rev.* **2016**, *53*, 697–708. [CrossRef]
7. Chen, X.H.; Cheng, G.; Li, Y.; Peng, L.P. Fault diagnosis of planetary gear based on entropy feature fusion of DTCWT and OKFDA. *J. Vib. Control* **2018**, *24*, 5044–5061. [CrossRef]
8. Feng, Z.P.; Zhang, D.; Zuo, M.J. Planetary gearbox fault diagnosis via joint amplitude and frequency demodulation analysis based on variational mode decomposition. *Appl. Sci.* **2017**, *7*, 775. [CrossRef]
9. Liang, X.H.; Zuo, M.J.; Liu, L.B. A windowing and mapping strategy for gear tooth fault detection of a planetary gearbox. *Mech. Syst. Signal Process.* **2016**, *80*, 445–459. [CrossRef]
10. Cerrada, M.; Sanchez, R.V.; Pacheco, F. Hierarchical feature selection based on relative dependency for gear fault diagnosis. *Appl. Intell.* **2016**, *44*, 687–703. [CrossRef]
11. Wang, F.T.; Deng, G.; Ma, L.J. Convolutional neural network based on spiral arrangement of features and its application in bearing fault diagnosis. *IEEE Access* **2019**, *7*, 64092–64100. [CrossRef]
12. Lei, Y.G.; Jia, F.; Zhou, X.; Lin, J. Deep learning-based method for machinery health monitoring with big data. *J. Mech. Eng.* **2015**, *51*, 49–56. [CrossRef]
13. Deutsch, J.; He, D. Using deep learning-based approach to predict remaining useful life of rotating components. *IEEE Trans. Syst. Man Cybern.: Syst.* **2018**, *48*, 11–20. [CrossRef]
14. Oh, H.; Jung, J.H.; Jeon, B.C.; Youn, B.D. Scalable and unsupervised feature engineering using vibration-imaging and deep learning for rotor system diagnosis. *IEEE Trans. Ind. Electron.* **2018**, *65*, 3539–3549. [CrossRef]
15. Gao, Z.H.; Ma, C.B.; Song, D.; Liu, Y. Deep quantum inspired neural network with application to aircraft fuel system fault diagnosis. *Neurocomputing* **2017**, *238*, 13–23. [CrossRef]
16. Liu, G.F.; Bao, H.Q.; Han, B.K. A stacked autoencoder-based deep learning neural network for achieving gearbox fault diagnosis. *Math. Probl. Eng.* **2018**, *2018*, 5105709.
17. Feng, J.; Lei, Y.; Guo, L.; Lin, J.; Xing, S. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing* **2018**, *272*, 619–628.
18. Feng, J.; Lei, Y.G.; 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**, *72–73*, 303–315.
19. Thirukovalluru, R.; Dixit, S.; Sevakula, R.K.; Verma, N.K. A salour generating feature sets for fault diagnosis using denoising stacked auto-encoder. In Proceedings of the 2016 IEEE International Conference on Prognostics and Health Management (ICPHM), Ottawa, ON, Canada, 20–22 June 2016; pp. 1–7.
20. Martin, G.S.; Droguett, E.L.; Meruane, V. Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis. *Struct. Health Monit.* **2019**, *18*, 1092–1128. [CrossRef]
21. Udmale, S.S.; Singh, S.K.; Bhirud, S.G. A bearing data analysis based on kurtogram and deep learning sequence models. *Measurement* **2019**, *145*, 665–677. [CrossRef]
22. Shao, H.D.; Jiang, H.K.; Zhao, H.W.; Wang, F. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. *Mech. Syst. Signal Process.* **2017**, *95*, 187–204. [CrossRef]

23. Wu, E.Q.; Zhou, G.R.; Zhu, L.M.; Wei, C.F.; Ren, H.; Sheng, R.S.F. Rotated Sphere Haar Wavelet and Deep Contractive Auto-Encoder Network With Fuzzy Gaussian SVM for Pilot's Pupil Center Detection. *IEEE Trans. Cybern.* **2019**. [CrossRef]

24. Shao, H.D.; Jiang, H.K.; Wang, F.; Zhao, H.W. An enhancement deep feature fusion method for rotating machinery fault diagnosis. *Knowl.-Based Syst.* **2017**, *199*, 200–220. [CrossRef]

25. Xia, G.Q.; Han, Z.W.; Zhao, B.; Liu, C.Y.; Wang, X.W. Global Path Planning for Unmanned Surface Vehicle Based on Improved Quantum Ant Colony Algorithm. *Math. Probl. Eng.* **2019**, *2019*, 2902170. [CrossRef]

26. Mahdi, F.P.; Vasant, P.; Abdullah-Ai-Wadud, M.; Watada, J.; Kallimani, V. A quantum-inspired particle swarm optimization approach for environmental/economic power dispatch problem using cubic criterion function. *Int. Trans. Electr. Energy Syst.* **2018**, *28*, e2497. [CrossRef]

27. Li, J.J.; Xu, B.W.; Yang, Y.S.; Wu, H.F. Three-phase qubits-based quantum ant colony optimization algorithm for path planning of automated guided vehicles. *Int. J. Robot. Autom.* **2019**, *34*, 156–163. [CrossRef]

28. Hinton, G.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]