ORIGINAL ARTICLE

Open Access

An Intelligent Fault Diagnosis Method of Multi-Scale Deep Feature Fusion Based on Information Entropy

Zhiwu Shang[1,2*] , Wanxiang Li[1,2], Maosheng Gao[1,2], Xia Liu[1,2] and Yan Yu[1,2]

Abstract

For a single-structure deep learning fault diagnosis model, its disadvantages are an insufficient feature extraction and weak fault classification capability. This paper proposes a multi-scale deep feature fusion intelligent fault diagnosis method based on information entropy. First, a normal autoencoder, denoising autoencoder, sparse autoencoder, and contractive autoencoder are used in parallel to construct a multi-scale deep neural network feature extraction structure. A deep feature fusion strategy based on information entropy is proposed to obtain low-dimensional features and ensure the robustness of the model and the quality of deep features. Finally, the advantage of the deep belief network probability model is used as the fault classifier to identify the faults. The effectiveness of the proposed method was verified by a gearbox test-bed. Experimental results show that, compared with traditional and existing intelligent fault diagnosis methods, the proposed method can obtain representative information and features from the raw data with higher classification accuracy.

Keywords: Fault diagnosis, Feature fusion, Information entropy, Deep autoencoder, Deep belief network

1 Introduction

With the development of machine learning, including artificial neural networks (ANNs), support vector machines (SVMs), random forest (RF), and other algorithms, research on intelligent fault diagnosis that combines shallow learning with a fault diagnosis has gradually emerged. Compared with a traditional fault diagnosis, an intelligent fault diagnosis significantly improves recognition accuracy and efficiency. However, an intelligent fault diagnosis based on shallow learning has certain limitations. According to the literature [1–8], an excellent diagnostic performance depends directly on the quality of the extracted features. This limitation indirectly leads us to a significant amount of energy on tedious feature

extraction and feature selection. This results in a low efficiency and weak generalization of the fault diagnosis.

As a sub-problem of machine learning, deep learning overcomes the limitations of traditional machine learning [9]. Deep learning can learn effective feature expressions from raw data through unsupervised learning to avoid feature extraction and feature selection, which not relies on signal processing technology and fault diagnosis knowledge. Therefore, deep learning has attracted increasing attention and has been applied in various fields. At present, the deep learning model mainly includes the following models: deep belief network (DBN), deep autoencoders, convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

In recent years, deep learning models have been widely used for fault diagnosis. Wang et al. [10] proposed an enhanced intelligent diagnosis method based on multi-sensor data fusion and improved deep convolution neural network models, which enhances the sample feature learning, expands the discrepancy of different fault

*Correspondence: shangzhiwu@126.com
[1] School of Mechanical Engineering, Tiangong University, Tianjin 300387, China
Full list of author information is available at the end of the article

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 2 of 16

features, and achieves higher prediction accuracy. In addition, Jia et al. [11] obtained frequency domain data through a Fourier transform of the vibration signal of the planetary gearbox and then input it into a deep autoencoder for fault identification. Zhang et al. [12] proposed using a stacked sparse autoencoder for the fault diagnosis of robust oxide fuel cell systems. Yu [13] used particle swarm optimization to improve the coding stage of the stack denoising autoencoder (SDAE) such that it can evolve the structure and parameters of the SDAEs simultaneously. This method can realize manifold regularization learning and feature selection, and realize a gearbox fault diagnosis. Shao et al. [14] proposed the use of a wavelet function for the activation function of an autoencoder to construct a deep wavelet autoencoder and enhance the feature extraction of the original signal, and then combined it with an extreme learning machine to realize an intelligent fault diagnosis of the rolling bearings. Shang et al. [15] extracted the features of the time domain, frequency domain, and time-frequency domain from vibration signals as the input of the DBN model. They then recognized the fault severity of the rolling bearings. Guo et al. [16] proposed a hierarchical learning rate-adaptive deep convolution neural network (ADCNN) based on a CNN model for fault pattern recognition and fault severity assessment. Compared with the traditional CNN model, an ADCNN can automatically select an appropriate learning rate and higher accuracy. Yang et al. [17] used a CNN, gated recurrent units, and an attention mechanism to construct a deep neural network to monitor and diagnose the bearing condition. Furthermore, Zhang et al. [18] proposed a CNN model for training intervention, which improved the anti-noise ability and regional adaptability of the model by using the dropout rule and minimal batch training. This method also effectively solves the problem of bearing fault identification under noisy environments and different working loads.

According to previous analyses [10–18], the current research on the deep neural network in the field of fault diagnosis has mainly focused on the performance of a single-structure model. However, because the collected signals often contain noise, a fault diagnosis of the gearbox using a single neural network model has low accuracy, poor stability, and low generalization ability. As a new machine learning technology, ensemble learning achieves a better learning effect than a single learner by combining multiple learners. It can effectively solve the shortcomings of a single-structure deep-learning model. Shao et al. [19] used different activation functions to construct multiple autoencoders and combined them with ensemble learning to design a combination strategy to achieve a fault diagnosis of the rolling bearings. It was

proved that this method achieves higher accuracy and stability than a single deep autoencoder. Chen et al. [20] used a feature-level multi-mode fusion method to extract features from vibration signals, and combined it with deep learning to realize a rolling bearing fault diagnosis. Xiao [21] designed a combination strategy of ensemble learning for traffic accident detection, which combines a KNN with an SVM to obtain a better final output and improve the robustness of the model. In addition, Wu et al. [22] proposed a multi-view fault diagnosis structure based on variational pattern decomposition and multiscale convolutional neural networks. The structure can adaptively extract signal characteristics from different angles and realize the fault diagnosis of high-speed trains. Zheng et al. [23] used a feature extraction method applying composite multi-scale fuzzy entropy and combined it with an ensemble support vector machine (ESVM) to detect and diagnose rolling bearings. In addition, most feature fusion methods adopt simple and easy-to-understand voting and averaging methods. However, these methods treat all feature extraction models with the same weight, ignoring the differences in the individual models [19]. This disadvantage weakens features with more significant contributions and strengthens the features with smaller contributions, resulting in more irrelevant information in the fused features and insufficient representation capability, thereby affecting the subsequent fault identification. As a quantitative index to measure the information content of the system, information entropy can be used as a criterion for parameter selection [24]. Therefore, this study uses different autoencoders and combines with information entropy to design a novel feature fusion strategy, and builds a multi-scale feature extraction structure, which enhances the feature learning ability of the raw signal and improves the accuracy of fault diagnosis and stability.

In this paper, a novel intelligent fault diagnosis method is proposed, which is a multi-scale deep feature fusion method based on information entropy. The proposed method was mainly divided into three steps. First, autoencoders with different working principles were stacked to form multiple deep neural networks. A multi-scale feature extraction structure is then constructed using a deep neural network to enhance the ability to extract deep features. Second, based on information entropy, a feature fusion strategy is designed to obtain low-dimensional and high-quality deep features. This strategy ensures that the fused features have excellent robustness and representativeness. Finally, the fused feature is input into the DBN classifier to identify a fault. The vibration signal of the gearbox was analyzed experimentally using the proposed method. The experimental results show that the proposed method overcomes the

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 3 of 16

shortcomings of feeble stability, weak generalization, and low recognition accuracy of single-structure deep neural network model. This method is more effective than existing intelligent fault diagnosis methods.

The basic framework of this study is as follows: The theoretical background of the proposed method is briefly introduced in Section 2. In Section 3, the basic flow of the proposed method for fault diagnosis is described in detail. In Section 4, by collecting the vibration signals of the gearbox, the performance of the fault diagnosis method proposed in this study is analyzed and compared with shallow learning and standard deep learning models. Finally, important conclusions are presented in Section 5.

2 Theoretical Background of the Proposed Method

Because of unsupervised learning and layer-by-layer learning, autoencoders are widely used in pattern recognition, speech recognition, feature extraction, and fault diagnosis. In recent years, fault diagnosis based on autoencoders has become a hot topic for researchers. This section briefly introduces the working principles of a normal autoencoder, other autoencoders, and a DBN classifier.

2.1 Normal Autoencoder

Inspired by the structure of a DBN, Bengio et al. [25] proposed a normal autoencoder (NAE). A NAE has three layers of neurons: an input layer, a hidden layer, and an output layer. The purpose of a three-layer network is to encode the n-dimensional input vector $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]$ into a p-dimensional expression $\boldsymbol{h} = [h_1, h_2, \cdots, h_p]$ to reconstruct the input vector in the expression, as shown in Figure 1. The function of the coding layer is to form the hidden layer $\boldsymbol{h}$ by the operation of the input vector $\boldsymbol{x}$ and weight matrix $\boldsymbol{W}$. The decoding layer reconstructs the output vector $\hat{\boldsymbol{x}} = [\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n]$ by decoding $\boldsymbol{h}$. During the process of fault diagnosis, the primary purpose of the NAE is to minimize the reconstruction error between the output vector $\hat{\boldsymbol{x}}$ and the original input vector $\boldsymbol{x}$. The hidden layer can obtain a fine decoding feature $\boldsymbol{h}$ to extract the features of the input vector.

The forward training of the NAE includes two steps: an encoder and a decoder. The encoding process maps the original data to the hidden layer:

$$\boldsymbol{h} = sigm(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}). \tag{1}$$

In the decoding process, the hidden vector $\boldsymbol{h}$ is used to reconstruct the input vector:

$$\hat{\boldsymbol{x}} = sigm(\boldsymbol{W}'\boldsymbol{h} + \boldsymbol{b}'), \tag{2}$$

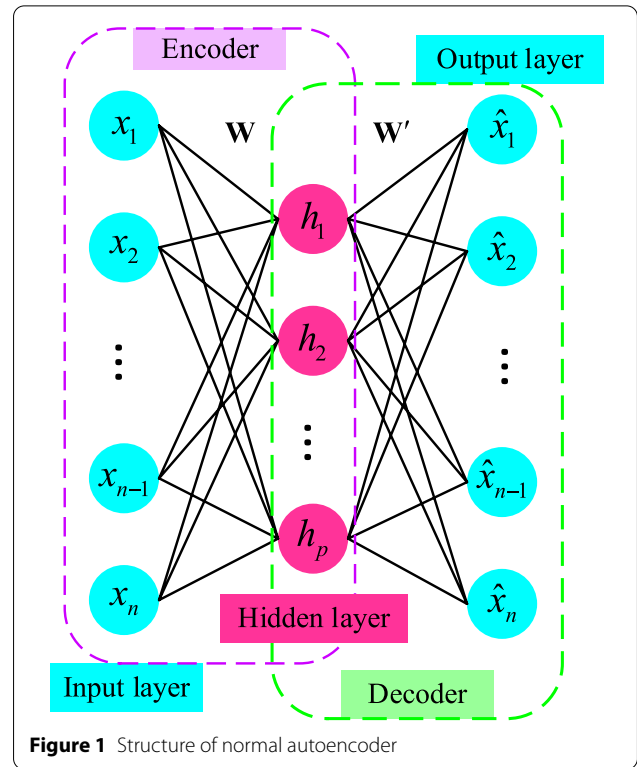**Figure 1** Structure of normal autoencoder

where $\boldsymbol{W}$ represents the weight matrix of the input layer to the hidden layer, $\boldsymbol{W}'$ is the weight matrix of the hidden layer to the input layer, $\boldsymbol{b}$ and $\boldsymbol{b}'$ represent the offset vectors of the coding and decoding layers, respectively. In addition, $sigm()$ is a nonlinear activation function.

The NAE needs to update the parameters $\boldsymbol{W}, \boldsymbol{W}', \boldsymbol{b}$, and $\boldsymbol{b}'$ to minimize the reconstruction error. The reconstruction error is defined as follows:

$$J_{NAE}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = L(\boldsymbol{x}, \hat{\boldsymbol{x}})$$
$$= -\sum_{i=1}^{n} [x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i)], \tag{3}$$

where $L(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is the loss function used to measure the difference between the n-dimensional input vector $\boldsymbol{x}$ and the output vector $\hat{\boldsymbol{x}}$, and x_i and $\hat{x}_i$ are the ith elements of the input vector and the output vector, respectively. The gradient descent algorithm [26] is commonly used to minimize the loss function to update the network parameters.

2.2 Other Forms of Autoencoder

A sparse autoencoder (SAE) adds a sparse penalty term to the loss function to ensure that the extracted features have a sparse response. It usually uses the KL distance to introduce a sparse penalty term to constrain the loss function [27]. The loss function is defined as follows:

$$J_{SAE}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = L(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \beta \sum_{j=1}^{p} KL(\rho||\hat{\rho}_j), \quad (4)$$

where β is the parameter of sparse penalty constraints, p is the number of hidden neurons, and $KL(\cdot)$ is the Kullback–Leibler divergence to improve the sparsity of hidden layer features.

A denoising autoencoder (DAE) allows the hidden layer to learn more robust deep features by adding random noise to the input vector. Gaussian noise is typically added to the input vector $\boldsymbol{x}$ to construct the damaged input vector $\tilde{\boldsymbol{x}}$. The loss function of the DAE is defined as follows:

$$\tilde{\boldsymbol{h}} = sigm(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b}), \quad (5)$$

$$\hat{\boldsymbol{x}} = sigm(\boldsymbol{W}'\tilde{\boldsymbol{h}} + \boldsymbol{b}'), \quad (6)$$

$$\begin{aligned} J_{DAE}(\boldsymbol{x}, \hat{\boldsymbol{x}}) &= L(\boldsymbol{x}, \hat{\boldsymbol{x}}) \\ &= -\sum_{i=1}^{n} [x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i)], \end{aligned} \quad (7)$$

where $\hat{\boldsymbol{x}}$ represents the reconstructed input vector, and x_i and $\hat{x}_i$ are the ith elements of the input and output vectors, respectively.

The contractive autoencoder (CAE) is a variant of the NAE that learns the robustness by adding the Jacobian matrix $J_{\boldsymbol{h}}(\boldsymbol{x})$ of the input and output of the hidden layer to the loss function. The loss function of the contractive autoencoder can be expressed as follows:

$$J_{CAE}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = L(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda||J_{\boldsymbol{h}}(\boldsymbol{x})||_F^2, \quad (8)$$

where λ is the regularization coefficient of the CAE, which controls the hyperparameter of the regularization strength.

The network structures of SAE, DAE, and CAE were the same as those of NAE. Likewise, they all used the gradient descent algorithm to minimize the loss function to achieve the purpose of learning the model parameters.

2.3 DBN Classifier

The DBN classifier is a deep neural network formed by a stacked restricted boltzmann machine (RBM). Its structure is shown in Figure 2 and consists of the first layer (data input) and the second layer (hidden layer 1). The second layer (hidden layer 1) and the third layer (hidden layer 2) constitute RBM 2. The third layer (hidden layer 2) and fourth layer (hidden layer 3) constitute three RBM stacks of RBM 3. The RBM consists of two layers of neurons (visual layer v and hidden layer h), and neurons of the same layer are independent of each other [28].

The training process of the DBN is divided into pre-training and fine-tuning stages. First, the parameters of the model are initialized to the optimal value through the layer-by-layer pre-training of the RBM learning rules. The parameters are then fine-tuned using a back-propagation algorithm according to the expected label. In the pre-training stage, the RBM of each layer was trained from the bottom to the top. Assuming that the RBM of the first layer has been trained, the conditional probability of the hidden variable is as follows:

$$p(\boldsymbol{h}^{(i)}|\boldsymbol{h}^{(i-1)}) = sigm(\boldsymbol{b}^{(i)} + \boldsymbol{W}^{(i)}\boldsymbol{h}^{(i-1)}), \ 1 \le i \le (l-1), \quad (9)$$
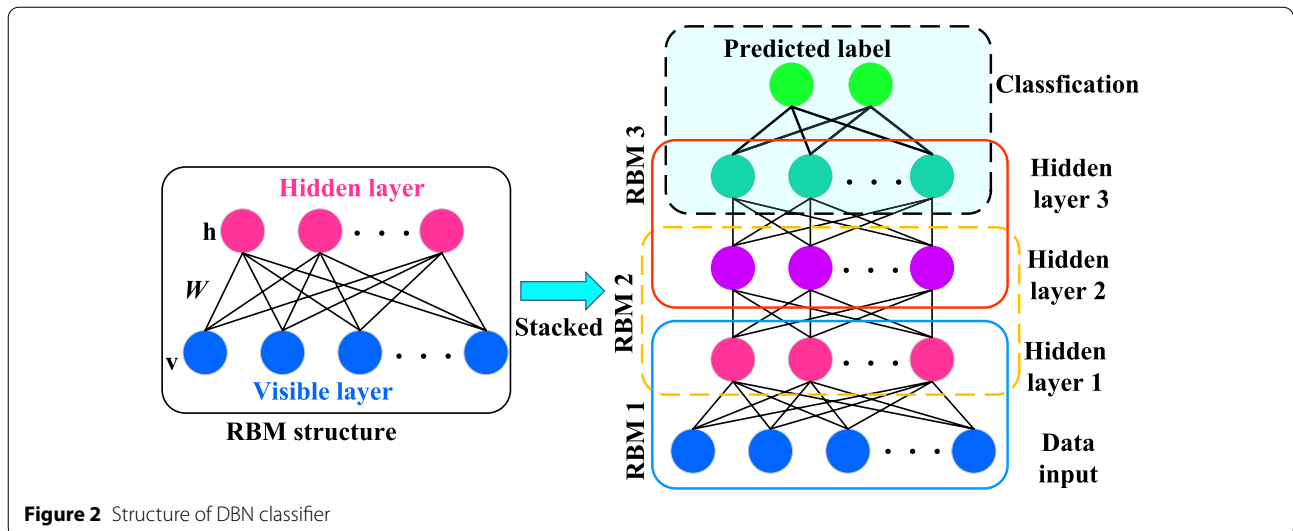
Figure 2 Structure of DBN classifier

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 5 of 16

where $\boldsymbol{b}^{(i)}$ and $\boldsymbol{W}^{(i)}$ are the bias and weight of the ith layer RBM, respectively, and when $i = 0$, $\boldsymbol{h}^{(0)} = \boldsymbol{v}$ is the input raw data.

After pre-training, the back-propagation algorithm is used to fine-tune the network according to the expected label, such that the model parameters reach the optimal solution.

3 Proposed Method

The primary content of this section is divided into three parts: the construction of the feature extraction structure of a multi-scale deep neural network, the design of the feature fusion strategy based on information entropy, and the implementation process of the proposed method.

3.1 Feature Extraction Structure of Multi-Scale Deep Neural Network

Owing to the simple structure of a single deep neural network (DNN), the stability and generalization ability of gearbox fault diagnosis are poor. To overcome the weaknesses mentioned above, this paper proposes a feature extraction structure for a multi-scale deep neural network based on different properties of the autoencoder. As shown in Figure 3, the hidden layer of the autoencoder with different characteristics extracts the deep features of the raw data. The hidden layer units are then superimposed to form a deep neural network in order of training. Finally, multiple deep neural network models were combined in parallel to form a multi-scale deep feature extraction structure.

As shown in Figure 3, each deep autoencoder contains a plurality of hidden layers and the last layer is used as the output of the deep features. The multi-scale deep feature extraction structure is constructed based on the difference in the feature extraction ability of the autoencoder with different characteristics. Compared with a single deep neural network structure, this structure can extract all features of the vibration signal to the greatest extent possible. In this study, we utilize NAE, DAE, SAE, and CAE to create a deep normal autoencoder (DNAE), deep denoising autoencoder (DDAE), deep sparse autoencoder (DSAE), and deep contractive autoencoder (DCAE) in a stack form. These deep neural networks are combined in parallel to construct a multi-scale deep feature extraction structure to achieve feature extraction.

3.2 Design of Feature Fusion Strategy

A multi-scale deep neural network feature extraction structure was constructed to extract the features. The next step is to design a feature fusion strategy to obtain the fused features. Data fusion is generally divided into three levels: data, feature, and decision fusion. In different stages of fusion, the majority voting method and the average method are easy to understand and are widely used. However, as the main disadvantage of majority voting and averaging methods, all individual models have the same weight and are treated equally [19]. This weakens the feature with larger contribution and enhances the features with smaller contribution, which results in more irrelevant information in the fused feature and insufficient performance, affecting the subsequent fault
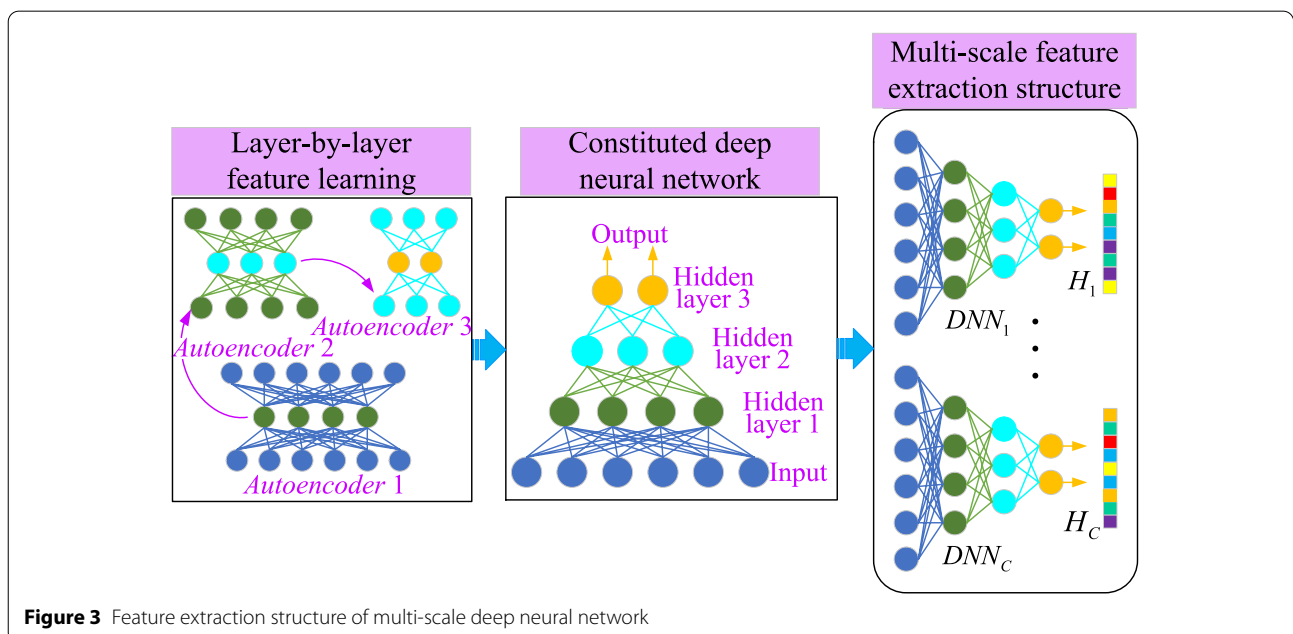

Figure 3 Feature extraction structure of multi-scale deep neural network

Shang *et al. Chin. J. Mech. Eng.* (2021) 34:58

Page 6 of 16

recognition. Information entropy can be used as a criterion for parameter selection as a quantitative index of information content in a system [24]. Therefore, a new weight allocation method based on information entropy is proposed in this study to effectively fuse the extracted deep features.

In this study, based on information entropy, different entropy weights are allocated according to the accuracy of each model, which avoids redundant information contained in fusion features, enhances the expression ability of deep features, and improves the quality of such features. This combination strategy includes the following three points. (1) Evaluation matrix A is constructed according to the exact values corresponding to different fault types in each DNN model. (2) The entropy weight of each DNN model is calculated for the fusion feature according to evaluation matrix A. (3) Fusion features are calculated according to the feature fusion formula. A flow chart of the designed feature fusion strategy is shown in Figure 4, and the detailed steps are described as follows:

Step 1: Assuming that the number of DNNs in the multi-scale feature extraction model is C and the type of fault is d, the training samples $X = [x_1, x_2, \cdots, x_n]$

and the training labels $Y = [y_1, y_2, \cdots, y_n]$ are input into the multi-scale feature extraction model for feature learning, and the evaluation matrix $A \in \Re^{d \times C}$ is calculated, where is x_i the sample data and y_i is the label data corresponding to x_i:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1C} \\ A_{21} & A_{22} & \cdots & A_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ A_{d1} & A_{d2} & \cdots & A_{dC} \end{bmatrix}, \tag{10}$$

where A_{ij} represents the accuracy of the ith $(i = 1, 2, \cdots, d)$ fault corresponding to the jth $(j = 1, 2, \cdots, C)$ DNN model.

Step 2: According to A, the information entropy of the jth DNN model is defined as follows:

$$D_j = -\frac{1}{\ln d} \sum_{i=1}^{d} A_{ij} \ln A_{ij}, j = 1, 2, \cdots C, \tag{11}$$

Based on information entropy, the entropy weight of the jth DNN is defined as follows:
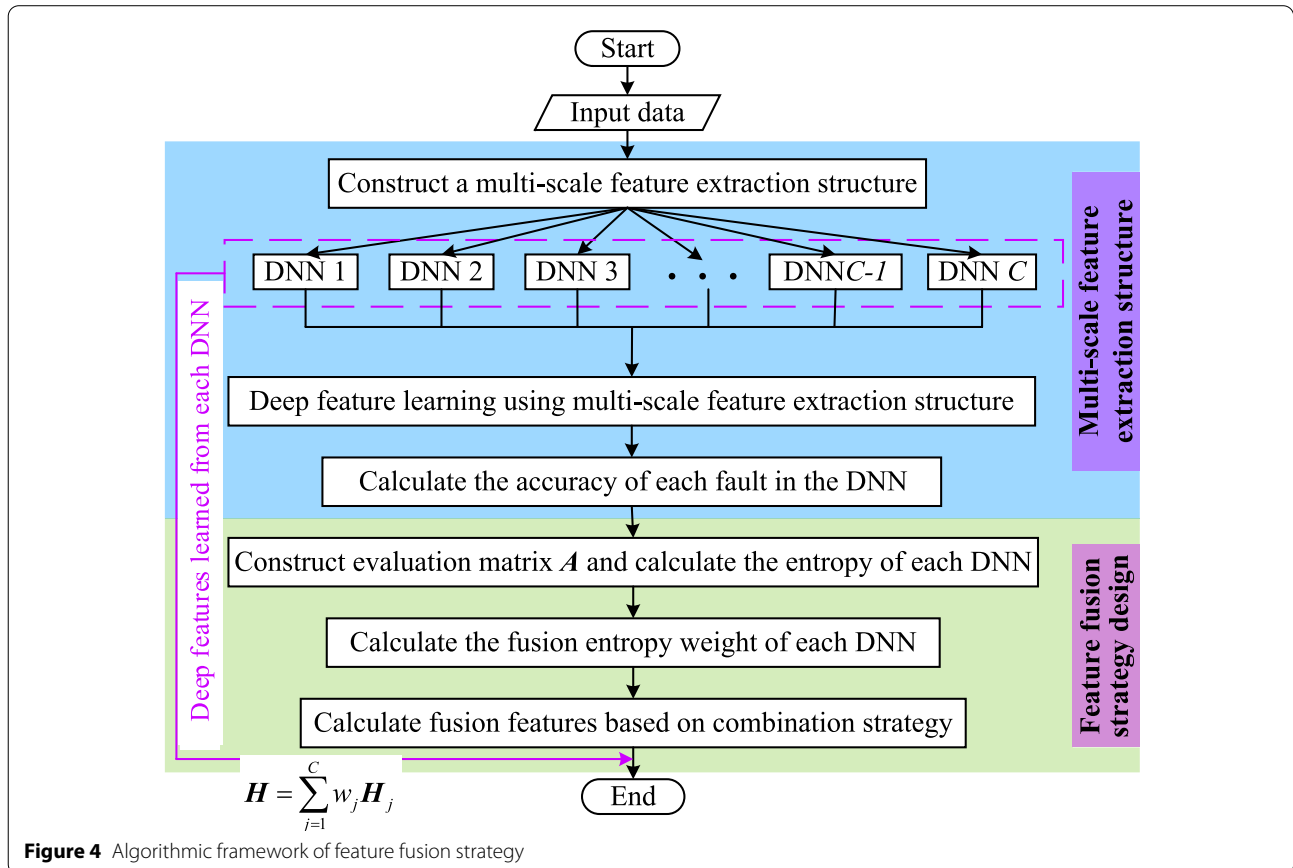

Figure 4 Algorithmic framework of feature fusion strategy

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 7 of 16

$$w_j = \frac{1 - D_j}{C - \sum\limits_{j=1}^{C} D_j}, j = 1, 2, \cdots, C, \tag{12}$$

where w_j satisfies $0 \le w_j \le 1$, $w_1 + w_2 + \cdots + w_C = 1$.

Step 3: Calculate the fused feature **H**:

$$\boldsymbol{H} = \sum_{j=1}^{C} w_j \boldsymbol{H}_j, \tag{13}$$

where $\boldsymbol{H}_j$ is the deep feature learned by the *j*th DNN model.

After the above three steps, the fused feature is input into the DBN classifier to complete the fault diagnosis of the gearbox.

3.3 General Procedure of the Proposed Method

In this paper, a multi-scale deep feature fusion method based on information entropy is proposed for the intelligent fault diagnosis of a gearbox. The general framework of the proposed method is shown in Figure 5, and the general steps are as follows.

> Step 1: The acceleration sensor is installed on the experimental device, and the signal acquisition device collects the vibration signal of the gearbox. The vibration signals were then divided into training and testing samples.
> Step 2: NAE, DAE, SAE, and CAE are stacked to generate DNAE, DDAE, DSAE, and DCAE, respectively, and then construct a multi-scale feature extraction structure and a feature fusion strategy based on information entropy.
> Step 3: The multi-scale deep neural network feature extraction structure is used to learn the deep features of the training samples, and obtain the fused features according to the proposed feature fusion strategy.
> Step 4: The DBN classifier is trained using the fused feature and training labels to obtain the trained DBN classifier.
> Step 5: The testing samples are used to verify the effectiveness of the proposed method.

4 Experimental Verification and Discussion
4.1 Data Description

In this experiment, the data of the gearbox are collected using the test-bed shown in Figure 6. The test-bed consists of a three-phase 3-hp motor, a two-stage planetary gearbox, a two-stage fixed shaft gearbox supported by rolling bearings and a programmable magnetic brake.

The frequency of the motor was 30 Hz, and the sampling frequency of the acceleration sensor was 3 kHz. One end of the accelerometer was installed in the vertical radial direction of the base of the fixed axle gear box, and the other end was connected to the acquisition device.

The detailed parameters of the faulty working conditions of the fixed shaft gearbox are shown in Table 1, and the locations of the faults in the gearbox are shown in Figure 7. During this experiment, six working conditions are considered. The time-domain and frequency-domain waveforms of the vibration signals (the first 3000 data points) under six working conditions were acquired, as shown in Figure 8. Figures 8(a)–(f) represent the time-domain and frequency-domain diagrams of the normal signal, gear hub crack signal, broken teeth signal, compound fault 1, compound fault 2, and compound fault 3, respectively. Each working condition contained 300 samples, and the vibration signal of each sample was composed of 300 consecutive sampling data points. To verify the accuracy and reliability of the proposed method, the training samples and test samples of each working condition were divided into 70% and 30% respectively.

4.2 Results and Analysis of Fault Diagnosis

To verify the effectiveness and advancement of the proposed method, the same dataset was validated using BPNN [2], Softmax classifier [3], SVM [4], and RF [6] fault diagnosis methods based on the shallow learning method. In addition, DNAE [29], DDAE [30], DSAE [31], DCAE [32] and CNN [33] models are also used to diagnose the faults of fixed axle gearboxes. The following points need further explanation.

1) The proposed method only needs to segment the collected vibration signals, and there are no feature extraction technologies for processing the vibration signals.
2) The inputs of DNAE, DDAE, DSAE and DCAE belong to the same dataset as the input of the proposed method, and the input of the CNN is 400-dimensional sample.
3) The BPNN, SVM, RF, and Softmax classifier have only one form of input. That is, 28 features were extracted using signal processing technology, including 10 time-domain features, 10 frequency-domain features, and 8 time-frequency domain features. The detailed parameters of these 28 features are referenced in Refs. [34] and [15].

In addition, to ensure the reliability of the experimental results of the proposed method, 10 experiments were conducted on the dataset of the fixed shaft gearbox. The average test accuracy and standard deviation of the
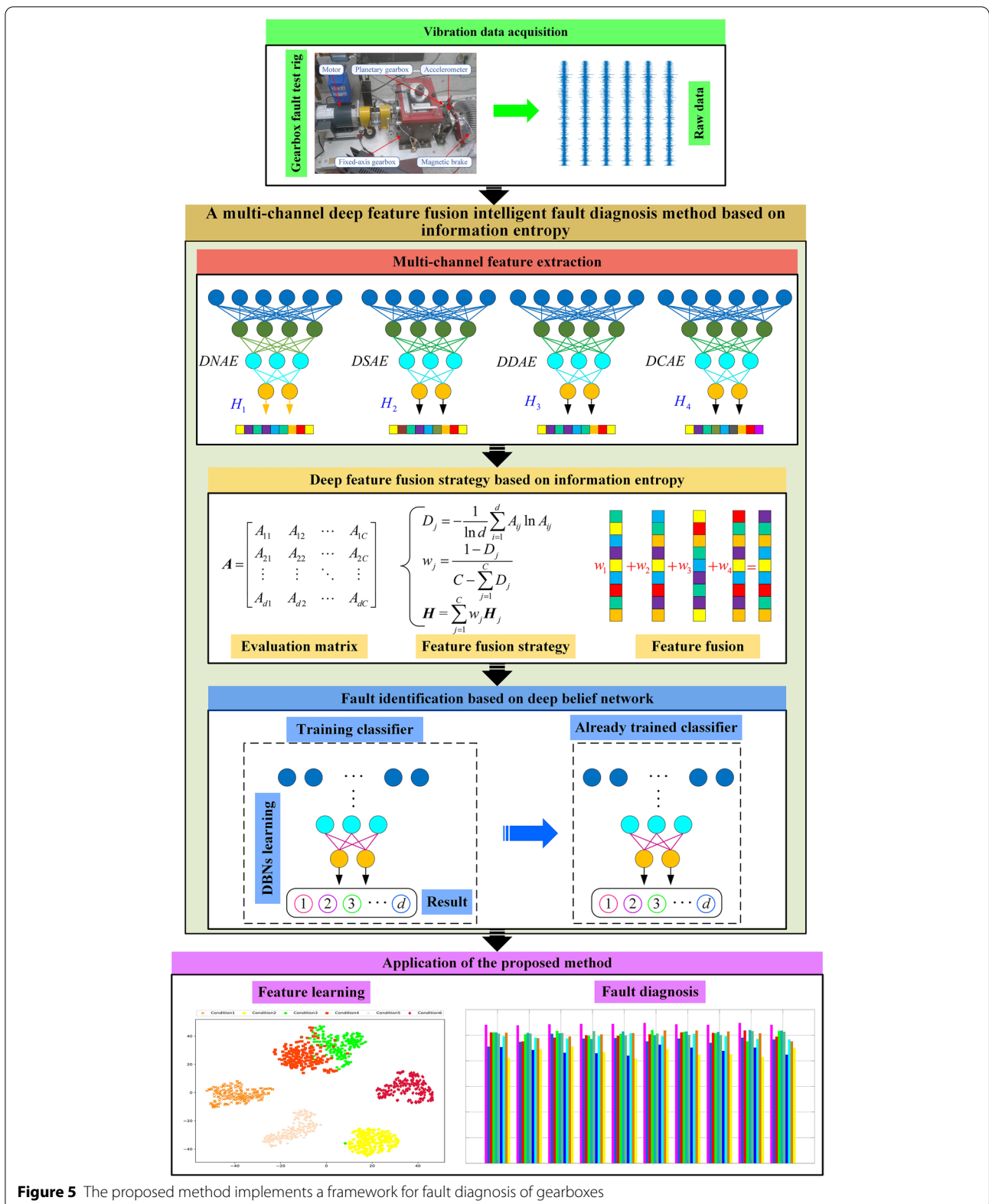
Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 8 of 16

Figure 5 The proposed method implements a framework for fault diagnosis of gearboxes

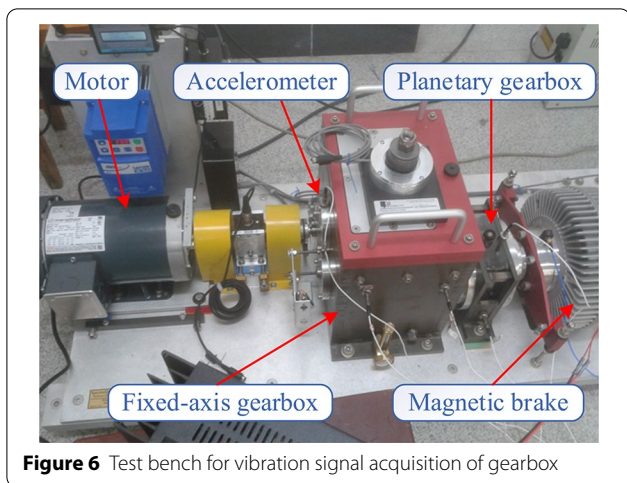Shang *et al. Chin. J. Mech. Eng.* (2021) 34:58

Page 9 of 16

Figure 6 Test bench for vibration signal acquisition of gearbox

Table 1 Detailed description of the six gearbox conditions

Fixed-axis gearbox operating conditions	Size of training/testing samples	Condition label
Normal	210/90	1
Gear hub crack	210/90	2
Gear broken teeth	210/90	3
Compound faults 1	210/90	4
Compound faults 2	210/90	5
Compound faults 3	210/90	6

Notes: Compound faults 1, wheel cracks, gear wear, and rolling bearing ball damage; compound faults 2, wheel cracks, broken teeth and rolling bearing ball damage; compound faults 3, gear wear, broken teeth and rolling bearing ball damage

proposed method and other fault-diagnosis methods are listed in Table 2. It can be concluded that compared with other methods, the proposed method achieves higher testing accuracy (94.31%) and lower standard variance (0.3187). Compared with the shallow learning of the BPNN, SVM, RF, and Softmax classifier, the average accuracy of the proposed method is higher than 84.07%, 89.39%, 90.40%, and 83.14%, respectively. Therefore, this method can directly extract fault features from vibration signals for a fault diagnosis, eliminating the tedious process of manual participation in feature extraction. Compared with the standard deep learning model, the accuracy of the proposed method is also higher than 88.19% for DNAE, 90.13% for DDAE, 90.69% for DSAE, 90.94% for DCAE and 90.74% for the CNN. Moreover, the standard deviation of the diagnostic results of the proposed method in the test samples is 0.3187, which is much lower than the values of 1.3593, 1.2359, 1.2823, 1.0287, 0.6171, 1.4211, 0.8097, 1.3824, and 1.7112 of the Methods 2–10, respectively. The proposed method can improve the recognition accuracy and enhance the stability of the recognition when a fault diagnosis is carried out for the vibration signals of the gearboxes.

Figure 9 shows the detailed diagnosis results of the test samples verified by different methods during 10 trials, which are shown in an intuitive form. The accuracy of the 10 experimental tests of the proposed method are 94.07%, 93.86%, 94.26%, 94.44%, 94.44%, 94.81%, 94.26%, 94.07%, 94.81%, and 94.07%, respectively. In addition, the time cost of the proposed method was compared with that of DNAE, DDAE, DSAE and DCAE, as shown in Figure 10. In Figure 10, the time cost of DNAE, DSAE and DCAE are approximately equal, and the time cost of DDAE is greater than that of DNAE, DSAE and DCAE. Moreover, the time cost of the proposed method is larger than that of DNAE, DDAE, DSAE and DCAE. However, with the rapid development of computers, the cost gap of the proposed method is narrow. The test accuracy of the proposed method is significantly higher than that of the other fault diagnosis methods. Table 3 lists the main parameters of the proposed method. The four deep
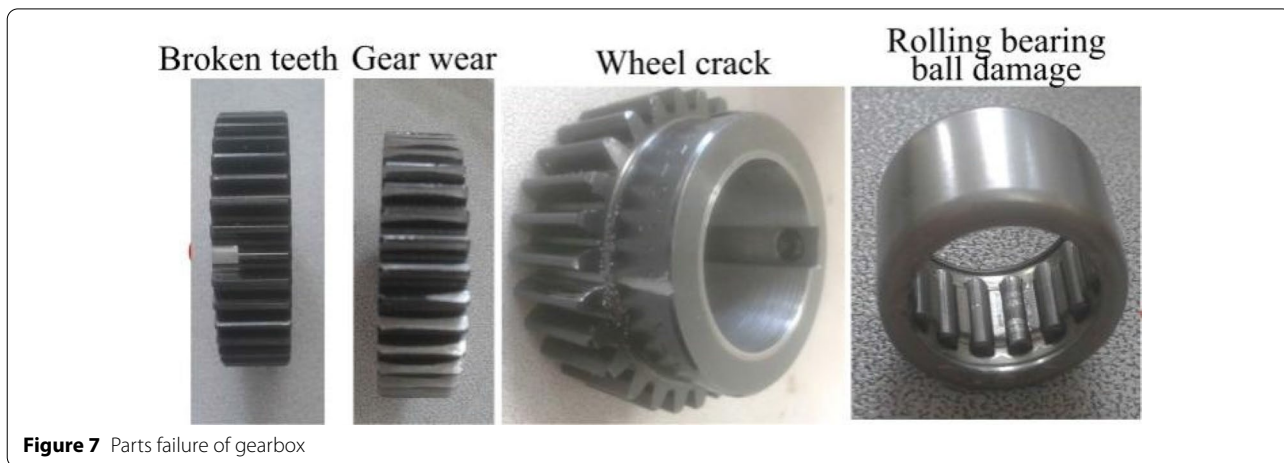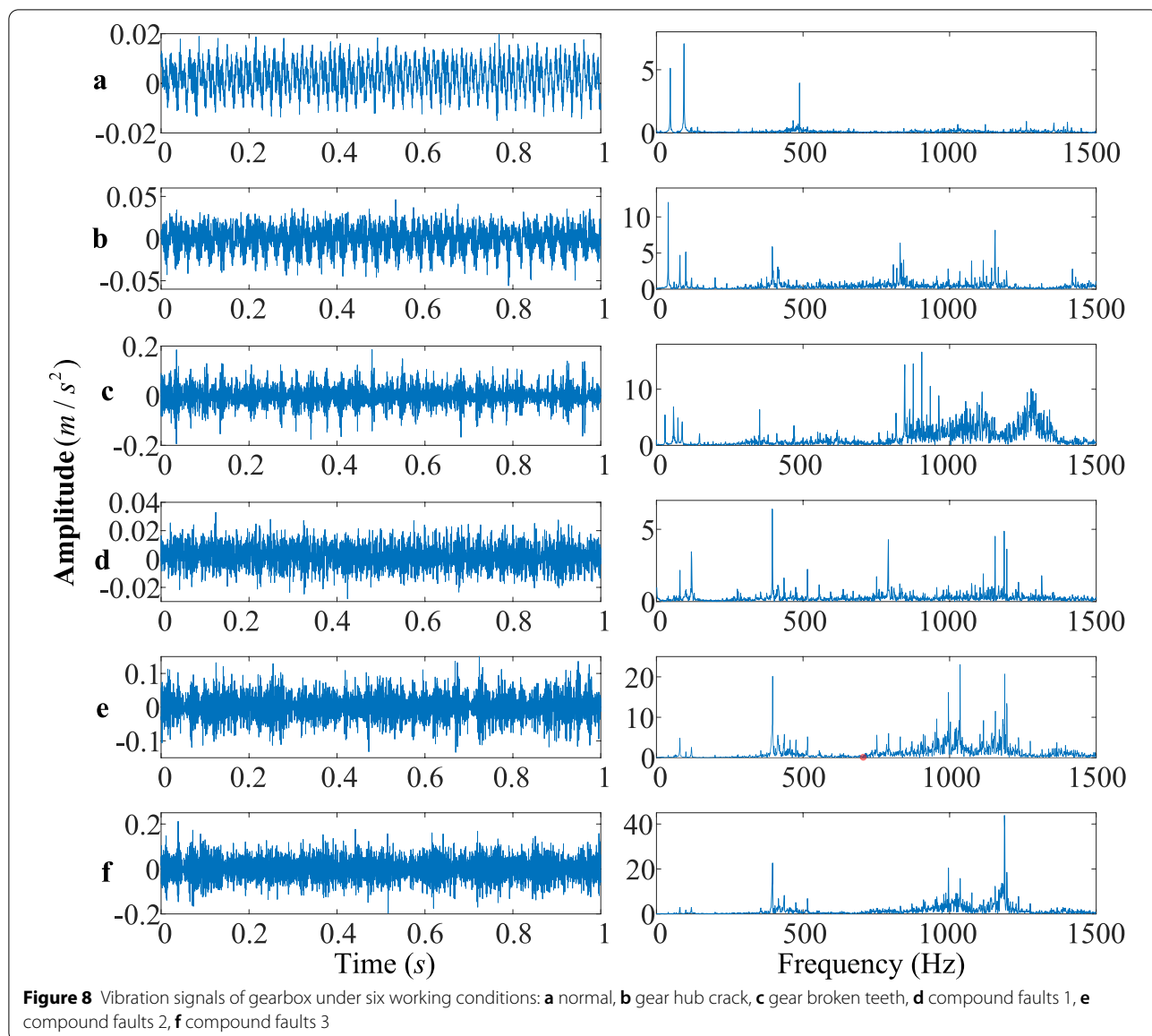

Figure 7 Parts failure of gearbox

Figure 8 Vibration signals of gearbox under six working conditions: **a** normal, **b** gear hub crack, **c** gear broken teeth, **d** compound faults 1, **e** compound faults 2, **f** compound faults 3

autoencoder architectures of the multi-scale deep feature extraction structure were 300-200-100-80. The feature learning among them is independent and does not experiences interference. The model structure of the DBN classifier is 80-40-40-40-6, which includes the input layer, hidden layers, and output layer.

The main parameters of the other methods are described as follows: (1) For Method 2 (DNAE), the structure is 300-200-100-80. The learning rate is 0.01, and the number of pre-training iterations for each NAE is 500. (2) For Method 3 (DDAE), the structure is 300-200-100-80, and the learning rate and noise loss coefficient are 0.017 and 0.1, respectively. In addition, the number of pre-training iterations per DAE is 500. (3) For Method

4 (DSAE), the structure is 300-200-100-80. The learning rate, sparse penalty constraints, and sparsity parameters are 0.016, 0.1, and 0.15, respectively. In addition, the number of pre-training iterations per SAE is 500. (4) For Method 5 (DCAE), the structure is 300-200-100-80, and the learning rate and regularization coefficient are 0.025 and 0.05, respectively. The number of pre-training iterations for each CAE is 500. (5) For Method 6 (CNN), the structure of the CNN consists of an input layer, two convolutional layers, two pooling layers, and a fully connected layer. The size of the input layer is 20 × 20, and the number of convolution kernels of the first and second convolution layers are 3 and 4, respectively. In addition, the step size of the two pooling layers is set to 2, and the

Shang *et al. Chin. J. Mech. Eng.* (2021) 34:58

Page 11 of 16

Table 2 Diagnostic results of different methods from 10 experiments

Methods	Input dimensions	Average testing accuracy (%)	Standard deviation of testing accuracy
Method 1	300	94.31	0.3187
Method 2	300	88.19	1.3593
Method 3	300	90.13	1.2359
Method 4	300	90.69	1.2823
Method 5	300	90.94	1.0287
Method 6	400	90.74	0.6171
Method 7	28	84.07	1.4211
Method 8	28	89.39	0.8097
Method 9	28	90.40	1.3824
Method 10	28	83.14	1.7112

Notes: Method 1, the proposed method with raw data; Method 2, DNAE with raw data; Method 3, DDAE with raw data; Method 4, DSAE with raw data; Method 5, DCAE with raw data; Method 6, CNN with raw data; Method 7, BP with 28 features; Method 8, SVM with 28 features; Method 9, RF with 28 features; Method 10, Softmax classifier with 28 features.

learning rate and the numbers of iterations are 0.01 and 500, respectively. (6) For Method 7 (BPNN), the structure is 28-40-6, the learning rate is 0.15, and the number of iterations is 1000. (7) For Method 8 (SVM), the type of kernel function is a Gaussian function, and the penalty coefficient of the loss function is 0.54. (8) For Method 9 (RF), the number of trees is 400, the maximum depth of the trees is 70, the minimum number of samples required for splitting the internal nodes is 70, and the minimum number of samples required for the leaf nodes is 80. (9) For Method 10 (Softmax classifier), the learning rate is 0.25 and the number of iterations is 1000.

The multi-class confusion matrix is a method for measuring the performance of deep learning models. In the first experiment, Figures 11 and 12 show the multiclass confusion matrix of the test set used in the proposed method and the other deep autoencoders, respectively. The horizontal axis in Figure 11 represents the prediction label of the fault, the vertical axis represents the actual label of the fault, and the diagonal element indicates that the probability prediction value is equal to the real value. The color bar on the right corresponds to the value of the multiclass confusion matrix. The multiclass confusion matrix can visually express the accuracy of the label predictions and the actual labels. Compared with Figure 12, Figure 11 shows that the prediction accuracy of the proposed method in label 1 is significantly improved to 0.92. Similarly, the prediction accuracy of labels 3 and 4 is slightly improved, and the labels 2, 5, and 6 both reach the optimal solution. Therefore, in gearbox fault diagnosis, the ability to extract deep features using the proposed method is higher than that of single-structure deep autoencoders and helps to improve the classification accuracy.

4.3 Visual Comparison of Deep Features
Principal component analysis (PCA) is a conventional algorithm used to reduce the number of data dimensions. It maps high-dimensional data into a low-dimensional space by transforming the matrix. The PCA visualizes high-dimensional data by giving each high-dimensional sample a position with two or three coordinates. The deep features are visualized through the PCA, which further illustrates the effectiveness of the proposed method by infusing deep features and identifying faults. A deep
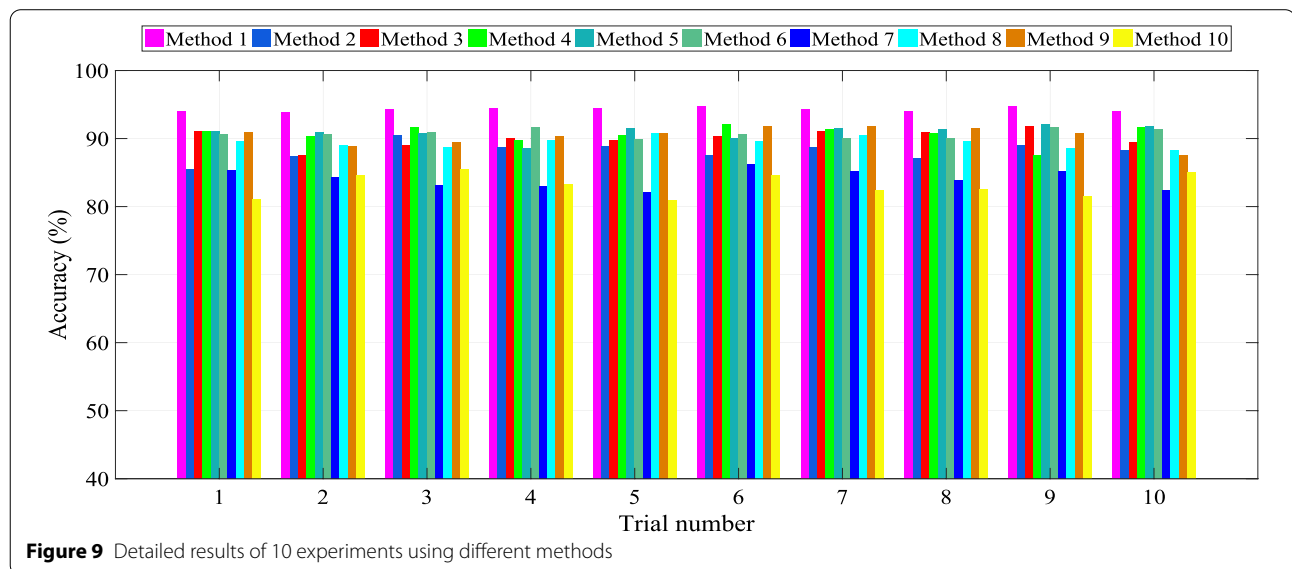

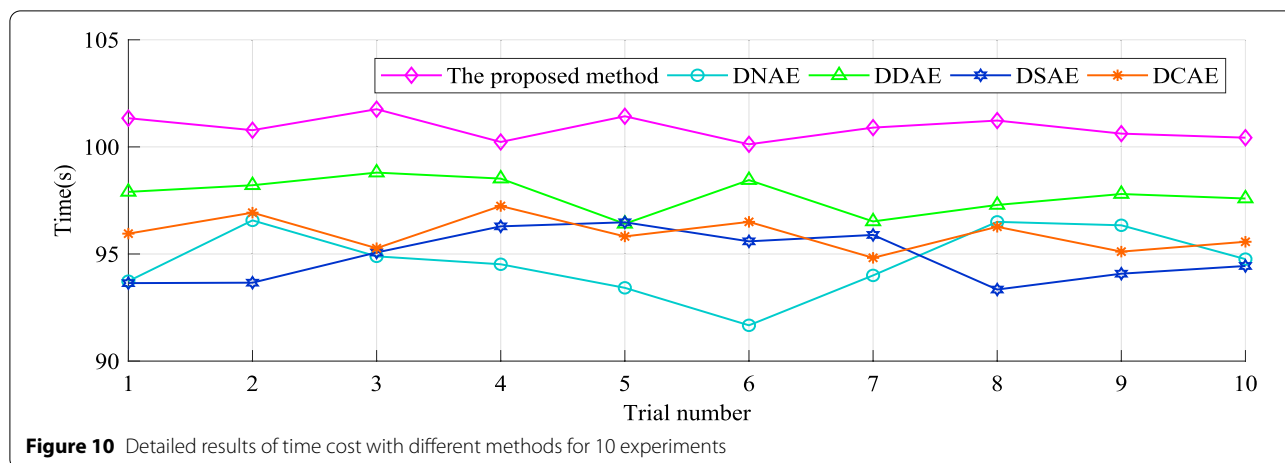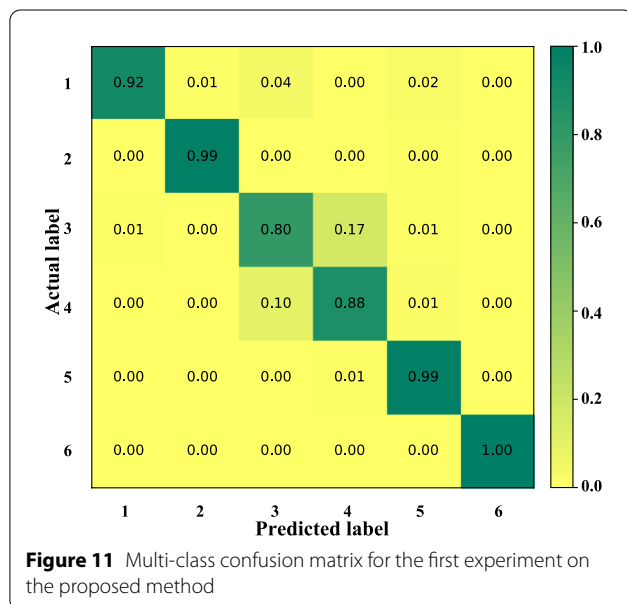**Figure 9** Detailed results of 10 experiments using different methods

Shang *et al. Chin. J. Mech. Eng.* (2021) 34:58

Page 12 of 16

Figure 10 Detailed results of time cost with different methods for 10 experiments

Table 3 The main parameters of the proposed method

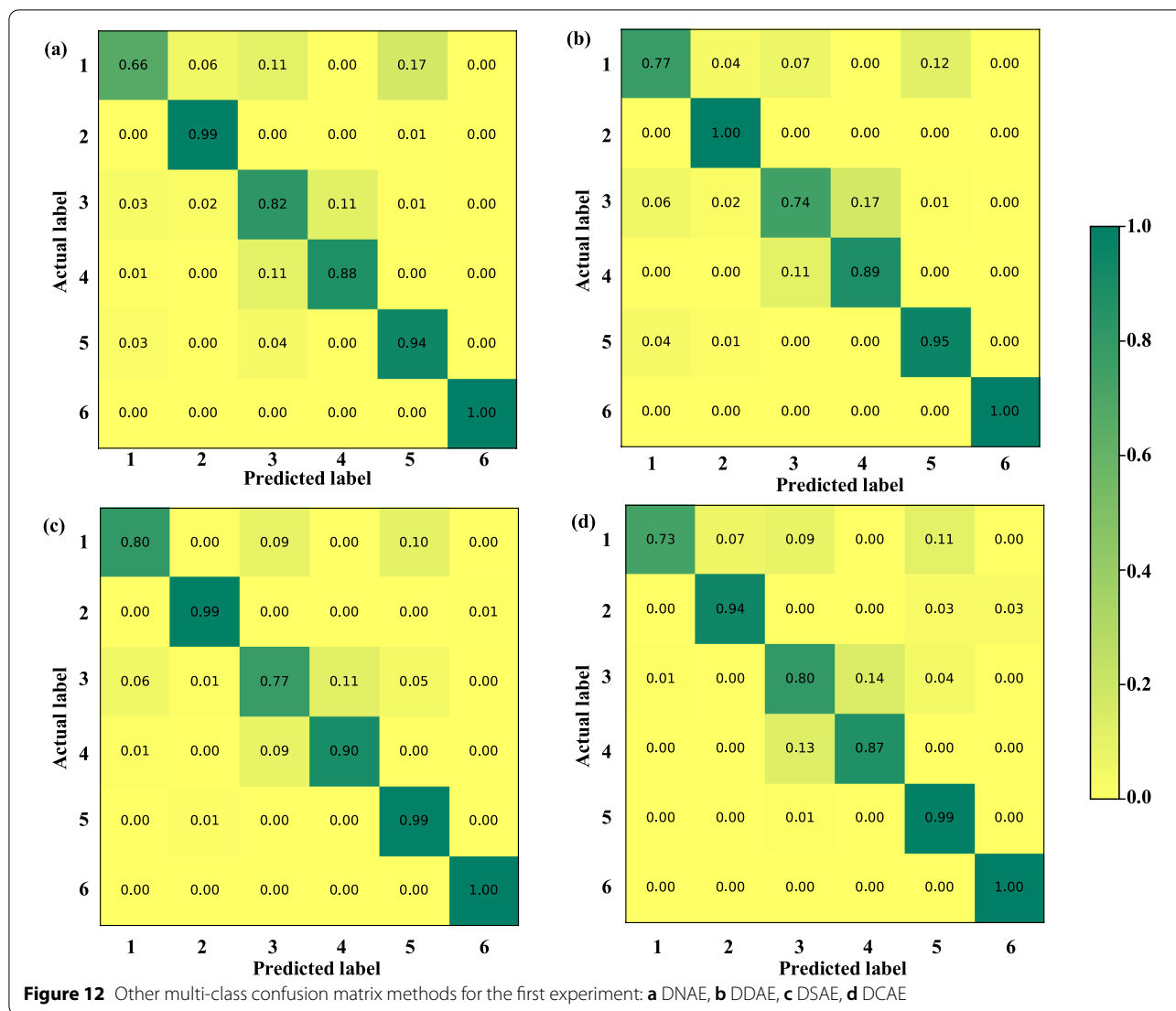| Description	Value
The number of hidden layers in each DAE	3
The number of input layers	300
The number of first hidden layers	200
The number of second hidden layers	100
The number of third hidden layers	80
Learning rate of four deep autoencoders	0.025
Number of iterations of four deep autoencoders	500
Penalty parameter of sparse autoencoder	0.2
Corruption parameters of denoising autoencoder	0.15
Regularization coefficient of contractive autoencoder	0.1
Learning rate of DBN classifier	0.02
Momentum of DBN	0.15

Figure 11 Multi-class confusion matrix for the first experiment on the proposed method

feature is the output value of the third hidden layer of the autoencoder. As shown in Figure 13, the deep features (80 dimensions) extracted from the third hidden layer were mapped into two-dimensional and three-dimensional coordinate systems after PCA dimensionality reduction. Among them, PCA1, PCA2, and PCA3 represent the first three main components of the deep features after PCA dimensionality reduction, which correspond to the x-axis, y-axis, and z-axis of the coordinate systems, respectively. The legend corresponds to the conditional labels listed in Table 1.

As shown in Figures 13(a), (b), (c), and (d), in the two-dimensional coordinate system, most boundaries of the features extracted using the DNAE, DDAE, DSAE, and DCAE models are clearly distinguishable. However, a small part of the boundary still overlaps and is difficult to distinguish, which directly increases the difficulty of fault identification. Moreover, in the three-dimensional coordinate system, most of the deep features have been separated, but there is still a small amount of overlap between the boundaries among the different features. This phenomenon shows that the features extracted using the DNAE, DDAE, DSAE and DCAE models have redundant information. Figure 13(e) shows the feature visualization results of the proposed method. Compared with Figure 13(a), (b), (c), and (d), the fault feature boundary of the proposed method in the two-dimensional coordinate system is clearer, and the fault features are completely separated in the three-dimensional coordinate system. Furthermore, the same type of fault feature aggregation effect was shown to be excellent. Therefore, the comparison results indicate that the proposed method can efficiently reduce the amount of redundant information of the deep features and improve the quality of the features.

In summary, compared with single-structure diagnosis model, the proposed method constructs the multi-scale

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 13 of 16

Figure 12 Other multi-class confusion matrix methods for the first experiment: **a** DNAE, **b** DDAE, **c** DSAE, **d** DCAE

deep neural network feature extraction structure by combining NAE, DAE, SAE, and CAE with different characteristics in parallel, and then applies deep feature fusion strategy based on information entropy. It solves the problems of weak feature extraction ability of single-structure deep learning models, poor stability of diagnosis models, and low accuracy of diagnosis. In addition, comparison experiments and feature visualization prove that the proposed method has higher recognition accuracy and better stability than traditional and existing intelligent fault diagnosis methods.

5 Conclusions
To address the critical issues regarding the improvement of the feature extraction and classification accuracy in single-structure deep learning, a multi-scale deep feature fusion intelligent fault diagnosis method based on

information entropy was proposed. In this study, NAE, DAE, SAE, and CAE with different characteristics were used to construct a multi-scale deep neural network feature extraction structure in parallel to enhance the ability of deep feature extraction. In addition, an entropy weight deep feature fusion strategy designed based on information entropy to capture the representative and robust features was described.

To verify the effectiveness of the proposed method, in the parallel axis gearbox fault diagnosis experiment, compared with the shallow learning model and the existing deep learning model, the fault accuracy of the proposed method is improved to 94.31% and its standard deviation is reduced to 0.3187. In addition, the proposed method avoids the tedious process of manual extraction of fault features, effectively and automatically extracts valuable fault features directly from
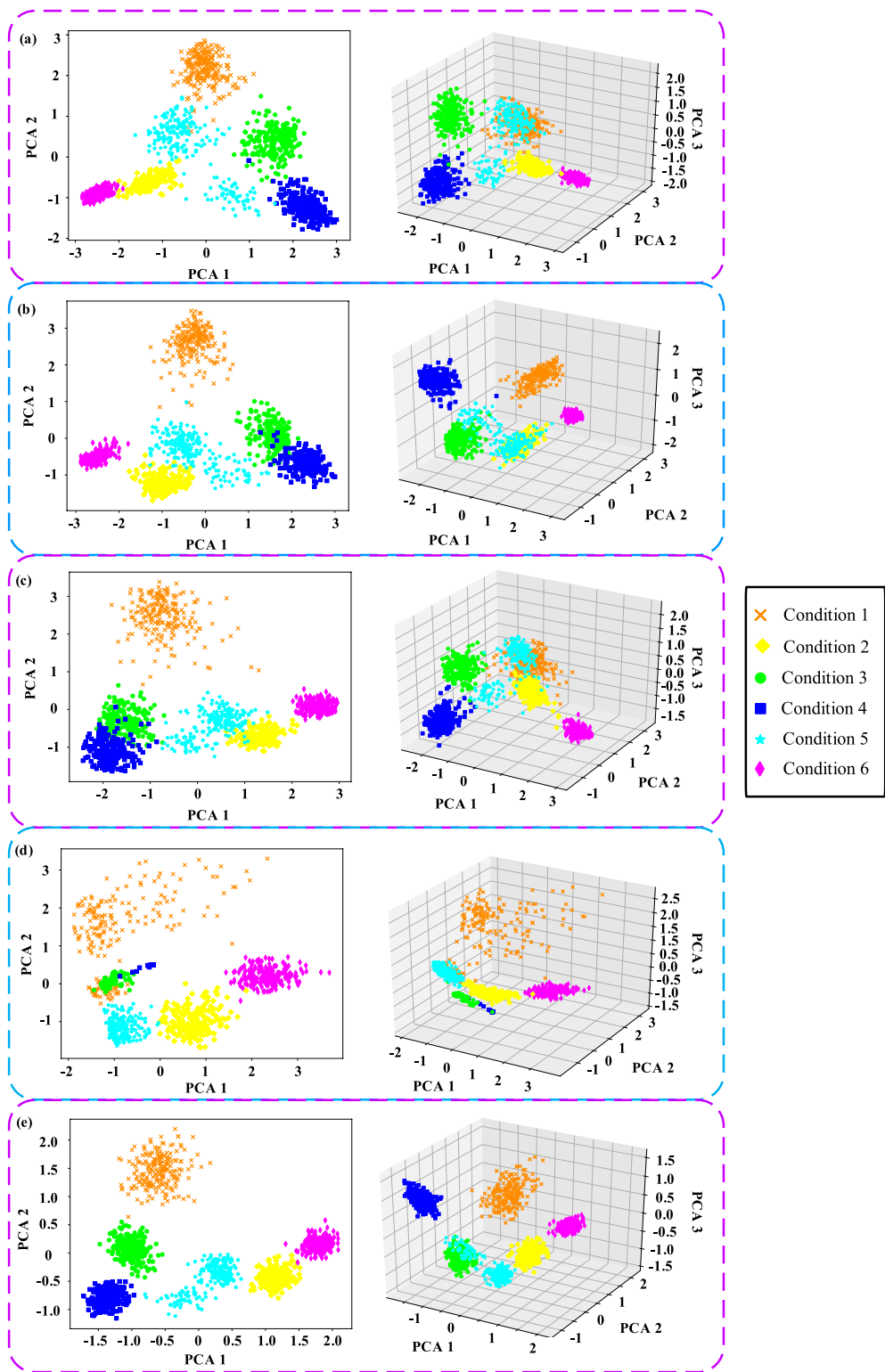
Figure 13 Feature visualization of hidden layer: **a** DNAE, **b** DDAE, **c** DSAE, **d** DCAE, **e** The proposed method

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 15 of 16

vibration signals, and improves the accuracy of fault identification. Moreover, through the comparative analysis of the multi-class confusion matrix and feature visualization, it is verified that the proposed method can improve the quality of deep features and capture robust fault features. In conclusion, the method proposed in this paper achieves high-quality feature mining capability and higher fault recognition accuracy. Considering that the selection of model hyper-parameters depends on the experiment conducted and artificial experience, in the future study, we will introduce intelligent optimization algorithms into deep neural networks to build smarter fault diagnosis methods.

Authors' Information
Zhiwu Shang, born in 1977, is a professor at *School of Mechanical Engineering, Tiangong University, China*. He received his Ph.D. degree in Mechanical Engineering from *Tianjin University*, *China*. He has published 50 journal papers in the fields of fault diagnosis and product development.

Wanxiang Li, born in 1993, is currently a doctoral candidate in mechanical engineering at *Tiangong University*, *China*. His research interests include deep learning, machine fault diagnosis and prognostics.

Maosheng Gao, born in 1993, is currently a doctoral candidate in mechanical engineering at *Tiangong University*, *China*. His research interests include mechanical signal processing, mechanical dynamics and fault diagnosis.

Xia Liu, born in 1994, is currently a master's candidate in mechanical engineering at *Tiangong University*, *China*. Her research interests include fault diagnosis and product development.

Yan Yu, born in 1995, is currently a master's candidate in Mechanical Engineering at *Tiangong University*, *China*. Her research interests include fault diagnosis and product development.

Author Details
[1]School of Mechanical Engineering, Tiangong University, Tianjin 300387, China. [2]Tianjin Modern Electromechanical Equipment Technology Key Laboratory, Tianjin 300387, China.

References

[1] B A Jaouher, F Nader, S Lotfi, et al. Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. *Applied Acoustics*, 2015, 89: 16-27.

[2] J M Li, X F Yao, X D Wang, et al. Multiscale local features learning based on BP neural network for rolling bearing intelligent fault diagnosis. *Measurement*, 2020, 153: 107419.

[3] K Adem, S Kiliçarslan, O Comert. Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications*, 2019, 115: 557-564.

[4] F Mei, N Liu, H Y Miao, et al. On-line fault diagnosis model for locomotive traction inverter based on wavelet transform and support vector machine. *Microelectronics Reliability*, 2018, 88-90: 1274-1280.

[5] Y G Lei, Z J He, Y Y Zi. EEMD method and WNN for fault diagnosis of locomotive roller bearings. *Expert Systems with Applications*, 2011, 38 (6): 7334-7341.

[6] Q Hu, X S Si, Q H Zhang, et al. A rotating machinery fault diagnosis method based on multi-scale dimensionless indicators and random forests. *Mechanical Systems and Signal Processing*, 2020, 139: 106609.

[7] Z W Shang, X Liu, W X Li, et al. A rolling bearing fault diagnosis method based on fastDTW and an AGBDBN. *Insight*, 2020, 62: 457-463.

[8] F Shen, C Chen, J W Xu, et al. A fast multi-tasking solution: NMF-theoretic co-clustering for gear fault diagnosis under variable working conditions. *Chinese Journal of Mechanical Engineering*, 2020, 33: 16. https://doi.org/10.1186/s10033-020-00437-3.

[9] Y G Lei, F Jia, J Lin, et al. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Transactions on Industrial Electronics*, 2016, 63(5): 3137-3147.

[10] H Wang, S Li, L Song, et al. An enhanced intelligent diagnosis method based on multi-sensor image fusion via improved deep learning network. *IEEE Transactions on Instrumentation and Measurement*, 2020, 69(6): 2648-2657.

[11] F Jia, Y G Lei, J Lin, et al. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mechanical Systems and Signal Processing*, 2016, 72-73: 303-315.

[12] Z H Zhang, S H Li, Y W Xiao, et al. Intelligent simultaneous fault diagnosis for solid oxide fuel cell system based on deep learning. *Applied Energy*, 2019, 233: 930-942.

[13] J B Yu. Evolutionary manifold regularized stacked denoising autoencoders for gearbox fault diagnosis. *Knowledge-Based Systems*, 2019, 178: 111-122.

[14] H D Shao, H K Jiang, X Q Li, et al. Intelligent fault diagnosis of rolling bearing using deep wavelet auto-encoder with extreme learning machine. *Knowledge-Based Systems*, 2018, 140: 1-14.

[15] Z W Shang, X X Liao, R Geng, et al. Fault diagnosis method of rolling bearing based on deep belief network. *Journal of Mechanical Science and Technology*, 2018, 32(11): 5139-5145.

[16] X J Guo, L Chen, C Q Shen. Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. *Measurement*, 2016, 93: 490-502.

[17] Z B Yang, J P Zhang, Z B Zhao, et al. Interpreting network knowledge with attention mechanism for bearing fault diagnosis. *Applied Soft Computing*, 2020, 97: 106829.

[18] W Zhang, G H Li, G L Peng, et al. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing*, 2018, 100: 439-453.

[19] H D Shao, H K Jiang, Y Lin, et al. A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders. *Mechanical Systems and Signal Processing*, 2018, 102: 278-297.

[20] C C Che, H W Wang, X M Ni, et al. Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis. *Measurement*, 2020, 108655.

[21] J L Xiao. SVM and KNN ensemble learning for traffic incident detection. *Physica A: Statistical Mechanics and its Applications*, 2018, 517: 29-35.

[22] Y P Wu, W D Jin, J X Ren, et al. A multi-perspective architecture for high-speed train fault diagnosis based on variational mode decomposition and enhanced multi-scale structure. *Applied Intelligence*, 2019, 49(11): 3923-3937.

[23] J D Zheng, H Y Pan, J S Cheng. Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble

Shang *et al. Chin. J. Mech. Eng.*   (2021) 34:58

Page 16 of 16

support vector machines. *Mechanical Systems and Signal Processing*, 2017, 85: 746-759.

[24] B Duan, Z Y Li, P W Gu, et al. Evaluation of battery inconsistency based on information entropy. *Journal of Energy Storage*, 2018, 16: 160-166.

[25] Y Bengio, A Courville, P Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, 35(8): 1798-1828.

[26] D E Rumelhart, G E Hinton, R J Williams. Learning representations by back-propagating errors. *Nature*, 1986, 323(6088): 533-536.

[27] P Vincent, H Larochelle, I Lajoie, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 2010, 11(12): 3371-3408.

[28] P Tamilselvan, P F Wang. Failure diagnosis using deep belief learning based health state classification. *Reliability Engineering & System Safety*, 2013, 115: 124-135.

[29] G F Liu, H Q Bao, B K Han. A stacked autoencoder-based deep neural network for achieving gearbox fault diagnosis. *Mathematical Problems in Engineering*, 2018, 2018: 5105709.

[30] Z L Chen, Z N Li. Fault diagnosis method of rotating machinery based on stacked denoising autoencoder. *Journal of Intelligent & Fuzzy Systems*, 2018, 34(6): 3443-3449.

[31] M Sohaib, J M Kim. Reliable fault diagnosis of rotary machine bearings using a stacked sparse autoencoder-based deep neural network. *Shock and Vibration*, 2018, 2018: 2919637.

[32] C Q Shen, Y M Qi, J Wang, et al. An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder. *Engineering Applications of Artificial Intelligence*, 2018, 76: 170-184.

[33] D K Appana, A Prosvirin, J M Kim. Reliable fault diagnosis of bearings with varying rotational using envelope spectrum and convolution neural networks. *Soft Computing*, 2018, 22(20): 6719-6729.

[34] J X Qu, Z S Zhang, T Gong. A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion. *Neurocomputing*, 2016, 171: 837-853.