Mechanical Systems and Signal Processing 102 (2018) 278–297

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders

Shao Haidong, Jiang Hongkai *, Lin Ying, Li Xingqiu

School of Aeronautics, Northwestern Polytechnical University, 710072 Xi'an, China

ARTICLE INFO

Article history:
Received 23 June 2017
Received in revised form 14 August 2017
Accepted 16 September 2017

Keywords:
Intelligent fault diagnosis
Rolling bearings
Ensemble deep auto-encoders
Activation functions
Combination strategy

ABSTRACT

Automatic and accurate identification of rolling bearings fault categories, especially for the fault severities and fault orientations, is still a major challenge in rotating machinery fault diagnosis. In this paper, a novel method called ensemble deep auto-encoders (EDAEs) is proposed for intelligent fault diagnosis of rolling bearings. Firstly, different activation functions are employed as the hidden functions to design a series of auto-encoders (AEs) with different characteristics. Secondly, EDAEs are constructed with various auto-encoders for unsupervised feature learning from the measured vibration signals. Finally, a combination strategy is designed to ensure accurate and stable diagnosis results. The proposed method is applied to analyze the experimental bearing vibration signals. The results confirm that the proposed method can get rid of the dependence on manual feature extraction and overcome the limitations of individual deep learning models, which is more effective than the existing intelligent diagnosis methods.

1. Introduction

Rotating machinery equipment plays an irreplaceable role in modern industry [1]. Rolling bearing is the most significant component in the rotating machinery, which directly influences its performance and operation [2]. Therefore, automatic, accurate and robust identification of rolling bearings operating conditions has become increasingly important.

Different from the conventional diagnosis methods based on various signal processing techniques, intelligent diagnosis aims to effectively analyze massive collected data and automatically provide diagnosis results, which has become a new trend in the field of equipment condition monitoring [3]. Artificial neural network (ANN) and support vector machine (SVM), two typical intelligent methods, have become the most popular for machinery fault diagnosis in the past decades [4]. According to Refs. [5–7], three main steps are the necessity for almost all traditional intelligent diagnosis methods: feature extraction, feature selection and pattern recognition, which will result in two inherent shortcomings: (1) The vibration signals collected from bearings are always very complex and non-stationary with heavy background noises [8]. What's worse, different fault types, fault severities and fault orientations further increase the challenge of diagnosis. Therefore, various advanced signal processing techniques must be well mastered for fault feature extraction. (2) In order to fully describe the characteristics of bearings under complex conditions, a lot of features in time domain, frequency domain and time-frequency domain have to be extracted for designing high-dimensional feature vector [9]. In most cases, there exist some features have nothing to do with diagnosis target or have high correlation with others [10]. Therefore, feature selection must

* Corresponding author.
E-mail address: jianghk@nwpu.edu.cn (H. Jiang).

https://doi.org/10.1016/j.ymssp.2017.09.026
0888-3270/

be carried out to ensure satisfactory results. However, it is a blind, subjective and time-consuming task to select the most sensitive features in different diagnosis issues without enough prior knowledge [11]. Actually, feature selection usually relies heavily on engineering experience. The inherent shortcomings mentioned above arise from the shallow architectures employed in the traditional intelligent diagnosis methods [11]. Consequently, there is an urgent need to develop deep structures for unsupervised feature learning and intelligent fault diagnosis of rolling bearing.

Deep learning is a great breakthrough in artificial intelligence (AI) field, which has the potential to overcome the inherent shortcomings of the traditional intelligent diagnosis methods [12]. The biggest success of deep learning methods is that they can automatically learn the representative features from the raw data [13]. In order words, deep learning methods can greatly get rid of the reliance on signal processing techniques and hand-engineered features. Due to the powerful feature learning ability, deep learning attracts more and more attention and has been gradually applied in mechanical fault diagnosis field in the last three years [14–16]. However, current deep learning models mainly focus on the performance research of the individual models. Because of the complexity of the collected vibration signals and even the unbalance between the normal samples and fault samples [17], there still exist some problems when using the individual deep learning models for bearing intelligent fault diagnosis, such as single network structure and low generalization ability. Ensemble learning is a new technique, which uses multiple individual learners and a certain combination strategy to get better results than each individual learner. Recently, a lot of ensemble learning methods have been applied for machinery fault diagnosis, such as random forest (RF), Boosting and ensemble SVMs [18–21]. However, their performance still depends heavily on signal processing techniques and manual feature extraction. Considering that the research of deep learning with ensemble learning is still in its infancy, thus, it is meaningful to develop ensemble deep learning models which can take full advantages of deep learning and ensemble learning.

In this paper, a novel method called ensemble deep auto-encoders (EDAEs) is developed for bearing intelligent fault diagnosis. The proposed method can be divided into three main steps: Firstly, different activation functions are employed as the hidden functions to design a series of auto-encoders (AEs) with different characteristics. Secondly, EDAEs are constructed with various auto-encoders for unsupervised feature learning from the measured vibration signals. Finally, a combination strategy is designed to ensure accurate and stable diagnosis results. The proposed method is applied to analyze the experimental bearing vibration signals. The results confirm that the proposed method can get rid of the dependence on manual feature extraction and overcome the limitations of individual deep learning models, which is more effective than the existing intelligent methods.

The rest of this paper is organized as follows. The theory of standard deep auto-encoder is briefly introduced in Section 2. In Section 3, the proposed method is described in detail. In Section 4, the experimental results are analyzed and discussed. Finally, conclusions are given in Section 5.

2. Standard deep auto-encoder

Deep auto-encoder (DAE), deep belief network (DBN) and convolutional neural network (CNN) are three popular deep learning models [22]. Different from the DBN and CNN, DAE is a purely unsupervised feature learning model, which can be trained more effectively and easily [3].

DAE is constructed with several auto-encoders (AEs), and each AE is a three-layer neural network. The AE aims to minimize the reconstruction error between the input data and output data. The structures of a standard AE and DAE are shown in Fig. 1. For an unlabeled and m-dimensional training sample $\mathbf{x} = [x_1, x_2, \ldots, x_m] \in \Re^{1\times m}$, the first step of AE training is to trans-

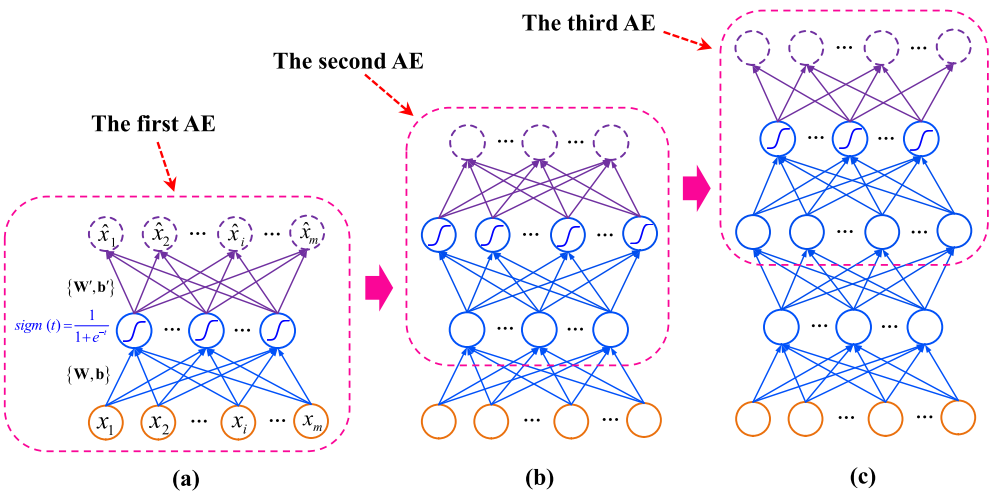

Fig. 1. The structures of standard AE and DAE. (a) Standard AE, (b) a DAE stacked with two AEs, (c) a DAE stacked with three AEs.

form the input data $\mathbf{x}$ into a hidden representation (also called a hidden feature vector) $\mathbf{h} = [h_1, h_2, \ldots, h_p] \in \Re^{1\times p}$ through the activation function [3]

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

$$sigm(t) = 1/(1 + e^{-t}) \tag{2}$$

where $\mathbf{W}$ is the weight matrix, $\mathbf{b}$ is a bias vector, and $\theta = \{\mathbf{W}, \mathbf{b}\}$ is the parameter set between the input layer and hidden layer. $sigm(\cdot)$ is Sigmoid function, which is usually employed as the activation function of the standard AE.

Then, the hidden vector $\mathbf{h}$ is mapped back into a reconstruction vector $\hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m]$ as follows

$$\hat{\mathbf{x}} = sigm(\mathbf{W}'\mathbf{h} + \mathbf{b}') \tag{3}$$

where $\theta' = \{\mathbf{W}', \mathbf{b}'\}$ is the parameter set between the hidden layer and output layer.

The AE training aims to optimize the parameter set $\boldsymbol{\theta} = \{\theta, \theta'\} = \{\mathbf{W}, \mathbf{b}, \mathbf{W}', \mathbf{b}'\}$ minimizing the reconstruction error. To data, two kinds of cost functions have been developed for measuring the reconstruction error of auto-encoders between the input vector and the reconstruction vector, which are the conventional mean square error cost function and the newly developed cross-entropy cost function, respectively. Compared with the mean square error cost function, the cross-entropy cost function shows faster convergence speed and stronger global optimization capability [23,24]. For one unlabeled and m-dimensional training sample, the cross-entropy cost function is defined as

$$J_{AE}(\boldsymbol{\theta}) = L(\mathbf{x}, \hat{\mathbf{x}}) = -\sum_{i=1}^{m}[x_i \log(\hat{x}_i) + (1 - x_i)\log(1 - \hat{x}_i)] \tag{4}$$

where $L(\mathbf{x}, \hat{\mathbf{x}})$ is the loss function which measures the difference between the input vector $\mathbf{x}$ and the reconstruction vector $\hat{\mathbf{x}}$. x_i is the ith dimension input of the unlabeled training sample, and $\hat{x}_i$ is the ith dimension reconstructed output.

3. The proposed method

In this paper, we propose a novel method called ensemble deep auto-encoders (EDAEs) for the intelligent fault diagnosis of bearings. This method includes three parts: ensemble deep auto-encoders construction, combination strategy design and the general procedures of the proposed method.

3.1. Ensemble deep auto-encoders construction

Due to the simplicity of an individual network structure and the difficulty in parameters selection, the individual deep auto-encoder will probably show low generalization ability in dealing with the diverse, massive and complex vibration data collected from rolling bearings. In order to overcome the limitations of the individual deep auto-encoders and enhance the generalization performance, the ensemble of multiple deep auto-encoders is a good choice. The activation function is used to provide nonlinear modeling ability, which has a significant impact on the performance of the neural network [25–27]. In addition, neural networks differing in activation functions usually show different characteristics and complementary learning behaviors. In this paper, a series of deep auto-encoders designed with different activation functions are used to construct EDAEs.

For an unlabeled training sample $\mathbf{x} = [x_1, x_2, \ldots, x_m] \in \Re^{1\times m}$, the input data $\mathbf{x}$ is transformed into $\mathbf{h} = [h_1, h_2, \ldots, h_p] \in \Re^{1\times p}$ through different activation functions

$$\mathbf{h} = \psi_{Act}(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{5}$$

where $\psi_{Act}(\cdot)$ represents the activation functions of hidden layer.

In this paper, 15 different activation functions are used as the hidden functions to design their corresponding auto-encoders with different characteristics. The equations and derivatives of the 15 activation functions are listed in Table 1, and the waveforms are shown in Fig. 2. Overall, the 15 activation functions can be divided into two categories: exponential operation and non-exponential operation. The functions based on exponential operation such as Sigmoid and TanH have been widely applied in different kinds of neural networks in the past decades, however, their main shortcomings are the high computational cost and gradient vanishing problem [28]. The non-exponential functions such as Rectified linear unit (ReLU), Parameteric rectified linear unit (PReLU) and Exponential linear unit (ELU) are the recently developed activation functions in deep learning field [29,30], which are fast and avoid the vanishing gradient problem. However, the outputs of these new functions are not zero-centered, and they may be not stable sometimes.

Considering the advantages of sparse inputs in learning useful features [4], in this paper, the regularization cross-entropy cost function of the different kinds of auto-encoders for one unlabeled and m-dimensional training sample can be rewritten as

$$J_{AE}(\boldsymbol{\theta}) = -\sum_{i=1}^{m}[x_i \log(\hat{x}_i) + (1 - x_i)\log(1 - \hat{x}_i)] + \beta\left(\sum_{j=1}^{p}\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}\right) \tag{6}$$

Table 1
Equations and derivatives of 15 activation functions.

Function names	Equations	Derivatives
Identity	$f(x) = x$	$f'(x) = 1$
ArcTan	$f(x) = (\tan(x))^{-1}$	$f'(x) = 1/(x^2+1)$
TanH	$f(x) = 2/(1+e^{-2x}) - 1$	$f'(x) = 1 - (f(x))^2$
Sinusoid	$f(x) = \sin(x)$	$f'(x) = \cos(x)$
Softsign	$f(x) = x/(1+\lvert x\rvert)$	$f'(x) = 1/(1+\lvert x\rvert)^2$
Rectified linear unit (ReLU)	$f(x) = \begin{cases} 0(x<0) \\ x(x \geqslant 0) \end{cases}$	$f'(x) = \begin{cases} 0(x<0) \\ 1(x \geqslant 0) \end{cases}$
Leaky rectified linear unit (Leaky ReLU)	$f(x) = \begin{cases} 0.01x(x<0) \\ x(x \geqslant 0) \end{cases}$	$f'(x) = \begin{cases} 0.01(x<0) \\ 1(x \geqslant 0) \end{cases}$
Parameteric rectified linear unit (PReLU)	$f(\alpha,x) = \begin{cases} \alpha x(x<0) \\ x(x \geqslant 0) \end{cases}$	$f'(\alpha,x) = \begin{cases} \alpha(x<0) \\ 1(x \geqslant 0) \end{cases}$
Exponential linear unit (ELU)	$f(\alpha,x) = \begin{cases} \alpha(e^x-1)(x<0) \\ x(x \geqslant 0) \end{cases}$	$f'(\alpha,x) = \begin{cases} f(\alpha,x)+\alpha(x<0) \\ 1(x \geqslant 0) \end{cases}$
SoftPlus	$f(x) = \ln(1+e^x)$	$f'(x) = 1/(1+e^{-x})$
Bent identity	$f(x) = (\sqrt{x^2+1}-1)/2 + x$	$f'(x) = x/(2\sqrt{x^2+1}) + 1$
SoftExponential	$f(\alpha,x) = (e^{\alpha x}-1)/\alpha + \alpha\alpha > 0$	$f'(\alpha,x) = e^{\alpha x}\alpha > 0$
Sinc	$f(x) = \begin{cases} 1(x=0) \\ \sin(x)/x(x \neq 0) \end{cases}$	$f(x) = \begin{cases} 0(x=0) \\ \cos(x)/x - \sin(x)/x^2(x \neq 0) \end{cases}$
Sigmoid	$f(x) = 1/(1+e^{-x})$	$f'(x) = f(x)(1-f(x))$
Gaussian	$f(x) = e^{-x^2}$	$f'(x) = -2xe^{-x^2}$

where β is the sparse penalty factor, $\hat{\rho}_j$ is the average activation value of hidden unit j, and ρ is a sparse parameter. m and p are the dimensions of the input vector and hidden vector, respectively.

In order to learn the deep features of the input data, each DAE should be constructed with several trained auto-encoders. In addition, in order to further improve the diagnosis results and generalization ability, EDAEs are constructed with all kinds of DAEs. The construction steps of each DAE are the same, which is achieved through the successive training of each individual auto-encoder (from lowest to highest) [31]. Fig. 3 shows the layer-by-layer construction process of an individual DAE with three auto-encoders. Firstly, the collected vibration data (input data) is used to train the first auto-encoder (***Auto-encoder* 1**), and ***Feature* I** (low-level hidden features) can be learned. Then, ***Feature* I** is used as the input of the second auto-encoder (***Auto-encoder* 2**) to acquire ***Feature* II** (higher-level hidden features). The training process is continued for the third auto-encoder (***Auto-encoder* 3**) and ***Feature* III** (the highest-level hidden features). Finally, the learned highest-level features are fed into a *Softmax* classifier for fault pattern recognition.

3.2. Combination strategy design

Having constructed the EDAEs, and the next step is to design a combination strategy for reporting the final diagnosis results. Among the different combination strategies, majority voting is convenience and intelligible, which has been widely applied in different ensemble learning methods. However, the major disadvantage of majority voting strategy is that all the individual models have same weights and are treated equally [32,33].

In this paper, a new and simple combination strategy is designed based on the majority voting, selective ensemble and weight assignment. This new combination strategy mainly contains four key points: (1) Determine a threshold for the accepted diagnosis accuracy, only the individual DAEs with accuracies exceeding the threshold are considered, while the others are removed. (2) Assign different weights to those DAEs met the requirements based on their corresponding accuracy values. (3) Report the combined diagnosis result of each sample based on the weight scores. (4) In order to maintain the stability of the combined diagnosis results, repeated trials are carried out. The flowchart of the designed combination strategy is shown in Fig. 4, and the detailed steps are described as follows:

Step 1: The training sample set $\mathbf{x} = \{\mathbf{x}_j, L_j\}_{j=1}^{N} (\mathbf{x}_j \in \Re^m, L_j \in \Re^1 \in \{1,2,\ldots,C\})$ is fed into $\mathrm{DAE}_i (i = 1,2,\ldots K)$ for fault pattern classification, respectively, and then their corresponding accuracies $\mathrm{acc}_i (i = 1,2,\ldots K)$ can be acquired, where $\mathbf{x}_j$ is the sample data, L_j is the actual label of $\mathbf{x}_j$, N is the number of samples, C is the number of classes and K is the number of individual DAE (in this paper, K = 15).

Step 2: Decide the accepted level $Threshold \in [0.75, 0.95]$, and compare each acc_i with *Threshold*. Only these $\mathrm{acc}_i \geqslant Threshold$ are taken into consideration, and then $\tilde{K}$ DAEs with their corresponding accuracies met the requirements are selected. The $\tilde{K}$ DAEs are expressed as $\mathrm{MDAE}_i (i = 1,2,\ldots \tilde{K})$ and the $\tilde{K}$ accuracies $Acc_i (i = 1,2,\ldots \tilde{K}, 1 \leqslant \tilde{K} \leqslant K)$.

Step 3: Assign different weights to those DAEs met the requirements

$$w_i = \frac{Acc_i}{\sum_{i=1}^{\tilde{K}} Acc_i} \left(\sum_{i=1}^{\tilde{K}} w_i = 1 \right) \tag{7}$$

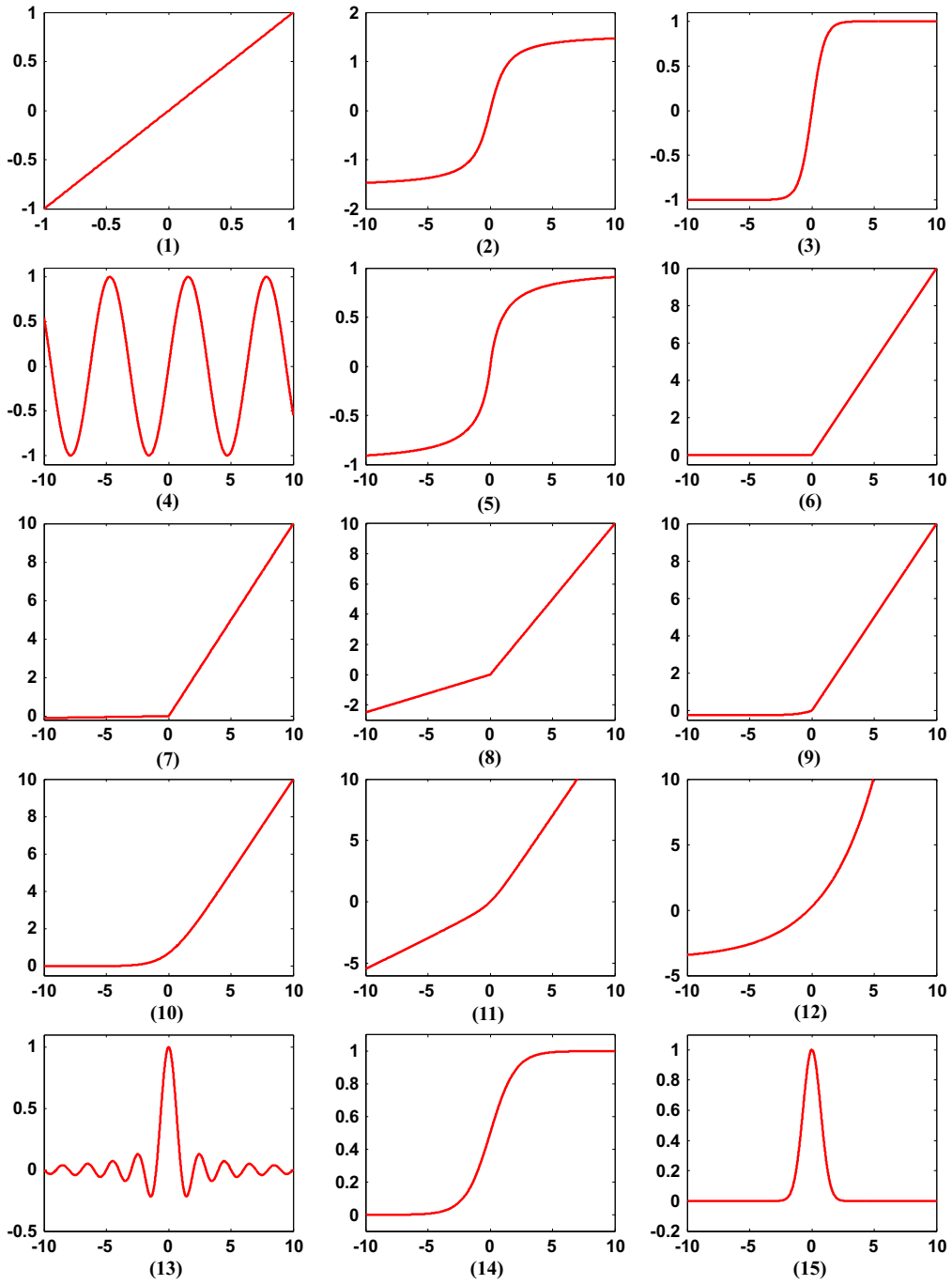

Fig. 2. The waveforms of the 15 kinds of activation functions. (1) Identity; (2) ArcTan; (3) TanH; (4) Sinusoid; (5) Softsign; (6) ReLU; (7) Leaky ReLU; (8) PReLU ($\alpha = 0.25$); (9) ELU ($\alpha = 0.25$); (10) SoftPlus; (11) Bent identity; (12) SoftExponential ($\alpha = 0.25$); (13) Sinc; (14) Sigmoid; (15) Gaussian.

The scores that sample $\mathbf{x}_j$ belongs to class c are calculated as

$$Score\_c\_\mathbf{x}_j = \sum_{i=1}^{\tilde{K}} w_i \cdot P(\mathrm{MDAE}_i(\mathbf{x}_j), c), c = 1, 2, \ldots, C \tag{8}$$

with

$$P(\mathrm{MDAE}_i(\mathbf{x}_j), c) = \begin{cases} 1, if\ \ MDAE_i(\mathbf{x}_j) = c \\ 0, if\ \ MDAE_i(\mathbf{x}_j) \neq c \end{cases} \tag{9}$$

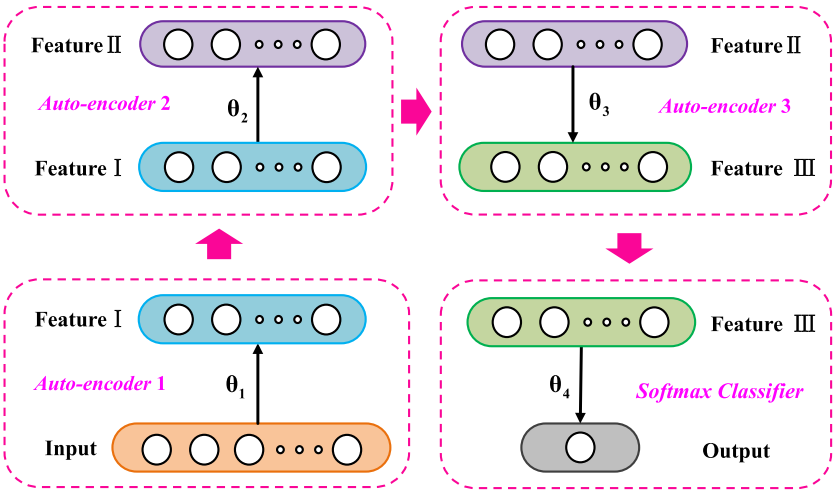

Fig. 3. The construction of DAE with three auto-encoders.

where $\mathrm{MDAE}_i(\mathbf{x}_j)$ represents the predicted label of MDAE_i $(i = 1, 2, \ldots \tilde{K})$ for sample $\mathbf{x}_j$. Obviously, if the predicted label for sample $\mathbf{x}_j$ is determined as T finally, it must satisfy the following constraints

$$\mathrm{Pre\_L}\_\mathbf{x}_j = T(T = 1, 2, \ldots C) \tag{10}$$

with

$$Score\_T\_\mathbf{x}_j = \max_{c\in\{1,2,\ldots C\}}\{Score\_c\_\mathbf{x}_j\} \tag{11}$$

where $\mathrm{Pre\_L}\_\mathbf{x}_j$ represents the final predicted label of sample $\mathbf{x}_j$.

Step 4: The results of the EDAEs are always expected to be superior to the best performance of all the individual DAEs, even that some of the individual DAEs unlikely show extremely better results than others. Thus, the combined results can be calculated as

$$\mathrm{Com\_}Acc = \max_{i\in\{1,2,\ldots,\tilde{K}\}}\left\{\frac{\sum_{j=1}^{N} Num(Pre\_L\_\mathbf{x}_j, \mathrm{L}_j)}{N}, Acc_i\right\} \tag{12}$$

with

$$Num(Pre\_L\_\mathbf{x}_j, \mathrm{L}_j) = \begin{cases} 1, if\ Pre\_L\_\mathbf{x}_j = \mathrm{L}_j \\ 0, if\ Pre\_L\_\mathbf{x}_j \neq \mathrm{L}_j \end{cases} \tag{13}$$

Step 5: In order to maintain the stability of the combined results, repeated trials are carried out. Finally, the average value of the combined results are expressed as

$$\mathrm{Final\_}Acc = \frac{\sum_{i=1}^{Q} \mathrm{Com\_}Acc_i}{Q} \tag{14}$$

where Q is the number of repeated trials, and $\mathrm{Com\_}Acc_i$ is the combined result for ith trial.

3.3. General procedure of the proposed method

In this paper, a novel method called ensemble DAEs is proposed for the intelligent fault diagnosis of bearings. The framework of the proposed method is shown in Fig. 5 and the general procedures are summarized as follows.

Step 1: Collect the vibration data of rolling bearings with acquisition device.
Step 2: Without any signal preprocessing or feature extraction, the raw vibration data is divided into training and testing samples.
Step 3: A series of auto-encoders with different properties are designed with different activation functions, and then each auto-encoder is used to construct its corresponding deep auto-encoder, respectively.
Step 4: Construct the ensemble DAEs using various deep auto-encoders, and design the combination strategy.
Step 5: The ensemble DAEs are used for feature learning and fault diagnosis based on the training samples.
Step 6: Validate the performance of the trained ensemble DAE using the testing samples.

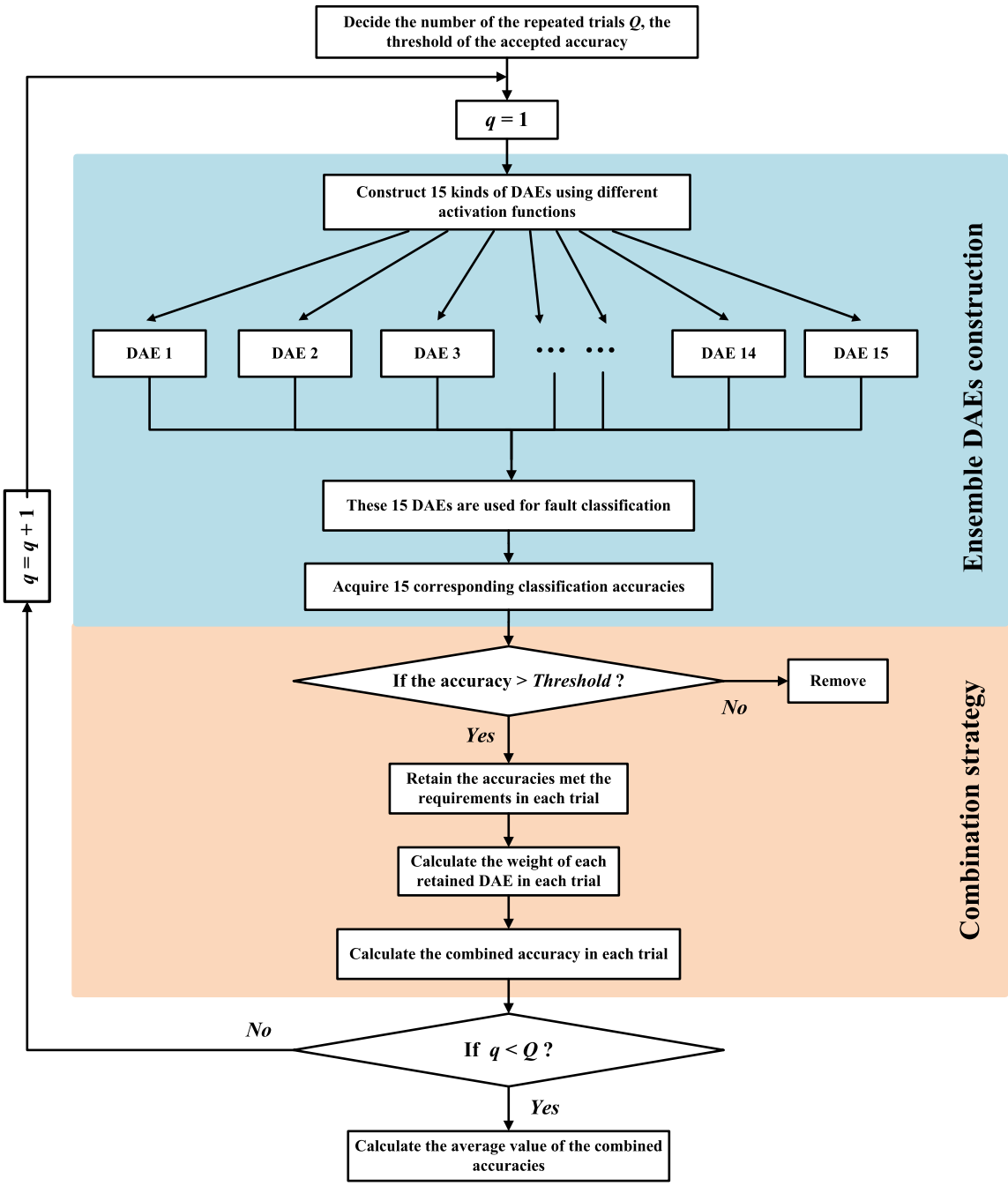

Fig. 4. Combination strategy of the ensemble DAEs (EDAEs).

4. Experimental verification

4.1. Rolling bearing experimental data

In this study, the experimental vibration data of rolling bearings is from Case Western Reserve University Lab [34]. As shown in Fig. 6, the experimental setup mainly contains an induction motor, an accelerometer, testing bearings and a loading motor. Each bearing (6205-2RS JEM SKF) was tested under four different loads (0, 1, 2 and 3 hp), and single point faults were introduced to the bearings with fault diameters of 0.007, 0.014, 0.021 and 0.028 inches (1 in. = 25.4 mm). An accelerometer is placed near the drive end to collect the vibration signals at a sampling frequency of 12 kHz.

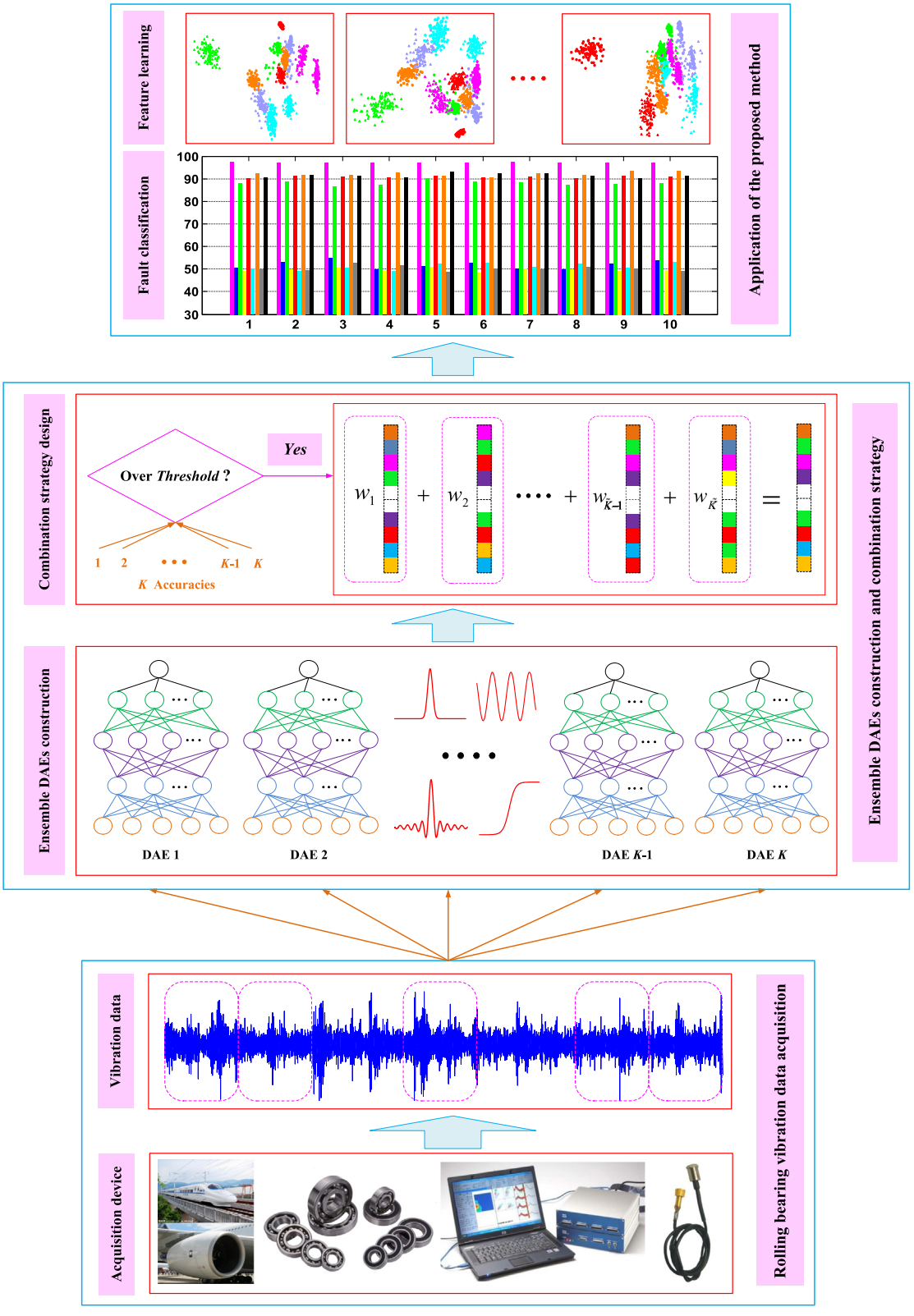

Fig. 5. The flowchart of the proposed method.

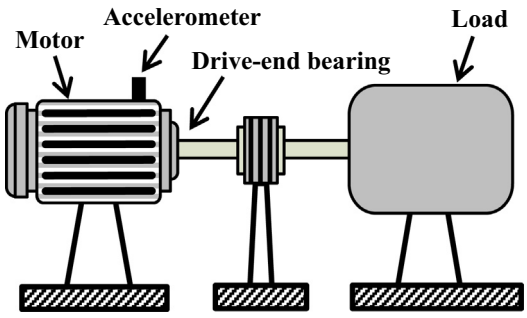

Fig. 6. The rolling bearing experimental setup.

The collected vibration data under 1797 rpm (0 hp) is used for fault classification. Twelve kinds of rolling bearing working conditions are selected, including different fault types, different fault severities and different fault orientations. Each condition contains 300 samples, and each sample is a collected vibration signal segment consists of 400 sampling data points. In order to avoid the particularity and contingency, the random 200 samples of each condition are used for training and the remaining 100 for testing. More details about the twelve bearing conditions are listed in Table 2. The raw time-domain waveforms of the twelve bearing working conditions are shown in Fig. 7 (first 80,000 sampling points).

4.2. Diagnosis results and analysis

In order to demonstrate the superiority of the proposed method, four other traditional methods are also used to analyze the same data set for comparison, including BP neural network (BPNN), SVM, random forest (RF) and Boosting. The following points need to be emphasized:

√ The proposed method focuses on bearing fault diagnosis without any signal preprocessing or feature extraction, which is totally different from the traditional methods.
√ The input of the proposed method is always the raw vibration data (400-dimension).
√ BPNN, SVM, RF and Boosting have two types of inputs. One is the raw vibration data, and the other is 24 features (11 time-domain features and 13 frequency-domain features). More details about the 24 feature parameters can be seen in Ref. [35].

In order to show the effectiveness and robustness of the proposed method, ten trials are carried out for diagnosing each bearing data set. The average training accuracies and testing accuracies are listed in Table 3, and Fig. 8 shows the detailed diagnosis results of the different methods on the testing samples in each trial. It can be seen from Table 3 that the proposed method shows the highest training accuracy (99.15%, 23,796/24,000) and the smallest standard deviation (0.0562) compared with other methods. The average testing accuracy of the proposed method is 97.18% (11,661/12,000), and it is much higher than BPNN, SVM, RF and Boosting using the raw vibration data (400-dimension), which are 51.96% (6235/12,000), 49.71% (5965/12,000), 51.17% (6140/12,000) and 50.31% (6037/12,000), respectively. After manual feature extraction (24-dimension), though the average testing accuracies of BPNN, SVM, RF and Boosting increase to 88.22% (10,586/12,000), 90.81% (10,897/12,000), 92.07% (11,048/12,000) and 91.43% (10,972/12,000), respectively, their performance still cannot be compared with the proposed method. The standard deviation of the proposed method on the testing samples is

Table 2
Description of the twelve bearing working conditions.

Bearing operating condition	Fault diameter (inches)	Outer race fault orientation	The number of training /testing samples	Condition label
Normal	0	–	200/100	1
Ball	0.007	–	200/100	2
Ball	0.014	–	200/100	3
Ball	0.021	–	200/100	4
Inner race	0.007	–	200/100	5
Inner race	0.021	–	200/100	6
Inner race	0.028	–	200/100	7
Outer race	0.007	Vertical @3:00	200/100	8
Outer race	0.007	Center @6:00	200/100	9
Outer race	0.014	Center @6:00	200/100	10
Outer race	0.021	Vertical @3:00	200/100	11
Outer race	0.021	Center @6:00	200/100	12

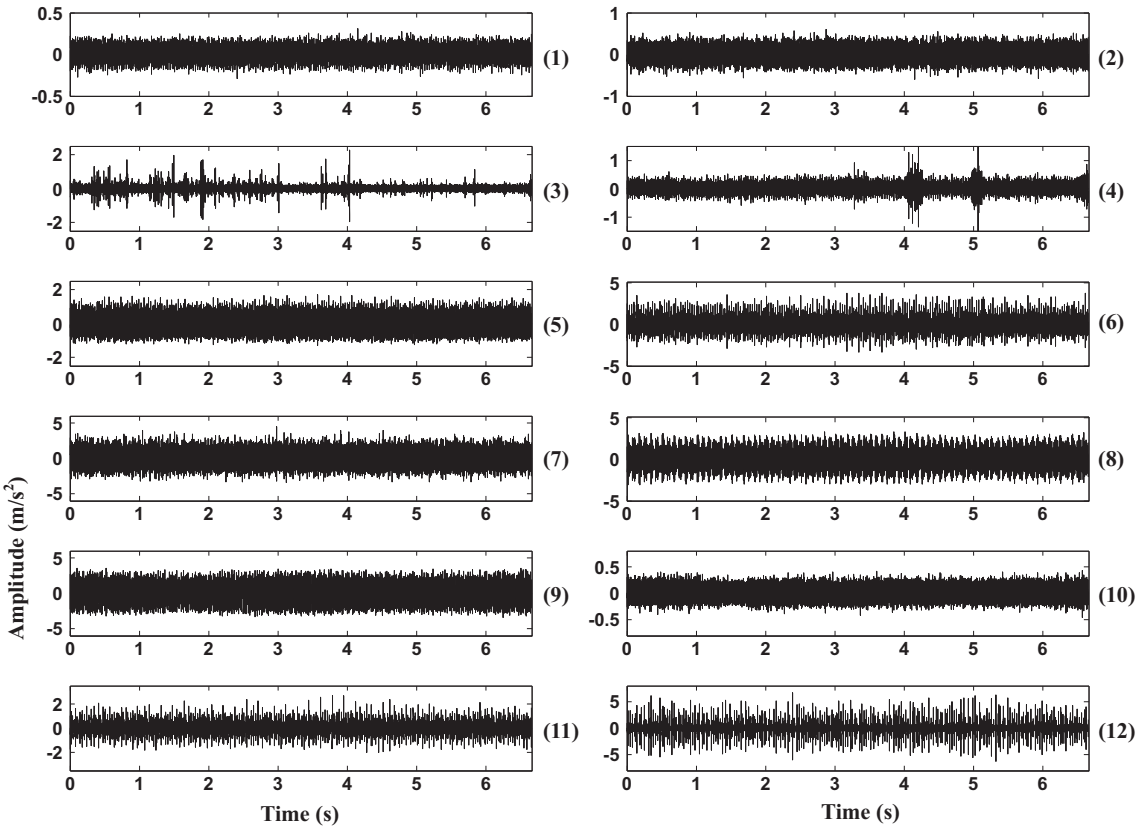

Fig. 7. Vibration signals of the twelve bearing operating conditions.: (1) Normal; (2) Ball fault (0.007); (3) Ball fault (0.014); (4) Ball fault (0.021); (5) Inner race fault (0.007); (6) Inner race fault (0.021); (7) Inner race fault (0.028). (8) Outer race fault (0.007@3:00); (9) Outer race fault (0.007@6:00); (10) Outer race fault (0.014@6:00); (11) Outer race fault (0.021@3:00); (12) Outer race fault (0.021@6:00).

Table 3
Diagnosis results of different methods.

Methods	The dimension of input	Average accuracy (%) ± standard deviation	
		Training samples	Testing samples
Method 1 (Proposed)	**400**	**99.15% ± 0.0562**	**97.18% ± 0.1142**
Method 2	400	53.88% ± 1.7454	51.96% ± 1.8063
Method 3	24	90.80% ± 0.8960	88.22% ± 1.0895
Method 4	400	51.54% ± 0.7859	49.71% ± 0.8317
Method 5	24	93.10% ± 0.3471	90.81% ± 0.3810
Method 6	400	58.29% ± 1.4082	51.17% ± 1.5159
Method 7	24	94.86% ± 0.8999	92.07% ± 0.9220
Method 8	400	52.06% ± 1.3153	50.31% ± 1.2856
Method 9	24	92.31% ± 0.8569	91.43% ± 0.8313

Remarks: **Method 1–The proposed method**; Method 2–BPNN with raw data; Method 3–BPNN with 24 features; Method 4–SVM with raw data; Method 5–SVM with 24 features; Method 6–RF with raw data; Method 7–RF with 24 features; Method 8–Boosting with raw data; Method 9–Boosting with 24 features.

0.1142, and it is much lower than other eight methods, which are 1.8063, 1.0895, 0.8317, 0.3810, 1.5159, 0.9220, 1.2856 and 0.8313, respectively. Thus, the proposed method can stably distinguish the different fault types, different fault severities and different fault orientations of rolling bearings.

From Fig. 8, the testing accuracy of the proposed method for each trial is 97.33%, 97.17%, 97.08%, 97.00%, 97.08%, 97.25%, 97.33%, 97.08%, 97.17% and 97.25%, respectively. Fig. 9 is the multi-class confusion matrix of the proposed method for the first trial. The multi-class confusion matrix thoroughly records the diagnosis classification results of the different bearing conditions, including both classification information and misclassification information. The ordinate axis of the multi-class confusion matrix represents the actual label of each bearing condition, and the horizontal axis represents the predicted

label. Therefore, the element on the main diagonal of the multi-class confusion matrix represents the diagnosis classification accuracy of each condition. It can be seen from Fig. 9 that the lowest accuracy happens in condition 7.

From the comparison results, it can be concluded that the proposed method shows higher accuracy than BPNN, SVM, RF and Boosting. The main reason is that the proposed method can effectively learn the essential features from the input data, while the performance of the traditional intelligent methods such as BPNN, SVM, RF and Boosting depends heavily on manual feature extraction. Their diagnosis results will be further improved after selecting the most sensitive features from the original feature set or designing some new features with excellent properties, however, it is a time-consuming, blind and subjective task. Compared with manual feature extraction and tedious feature selection, unsupervised feature learning from the measured vibration signals is more attractive and powerful.

In this study, the main parameters of the proposed method are available in Table 4. The structure design of deep learning model is still a great challenge, and there is not a theoretical method to solve this problem [11]. In this paper, the architecture of each individual DAE is 400-200-100-80-12, which is decided by experimentation and a simple idea similar to Ref. [3].

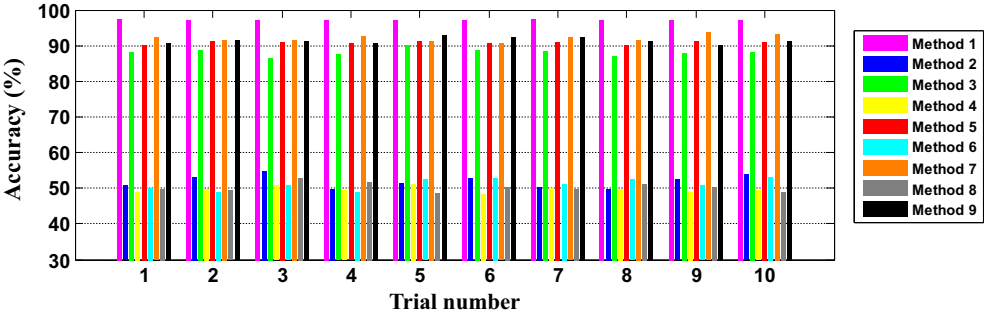

Fig. 8. Detailed testing results of different methods for ten trials.

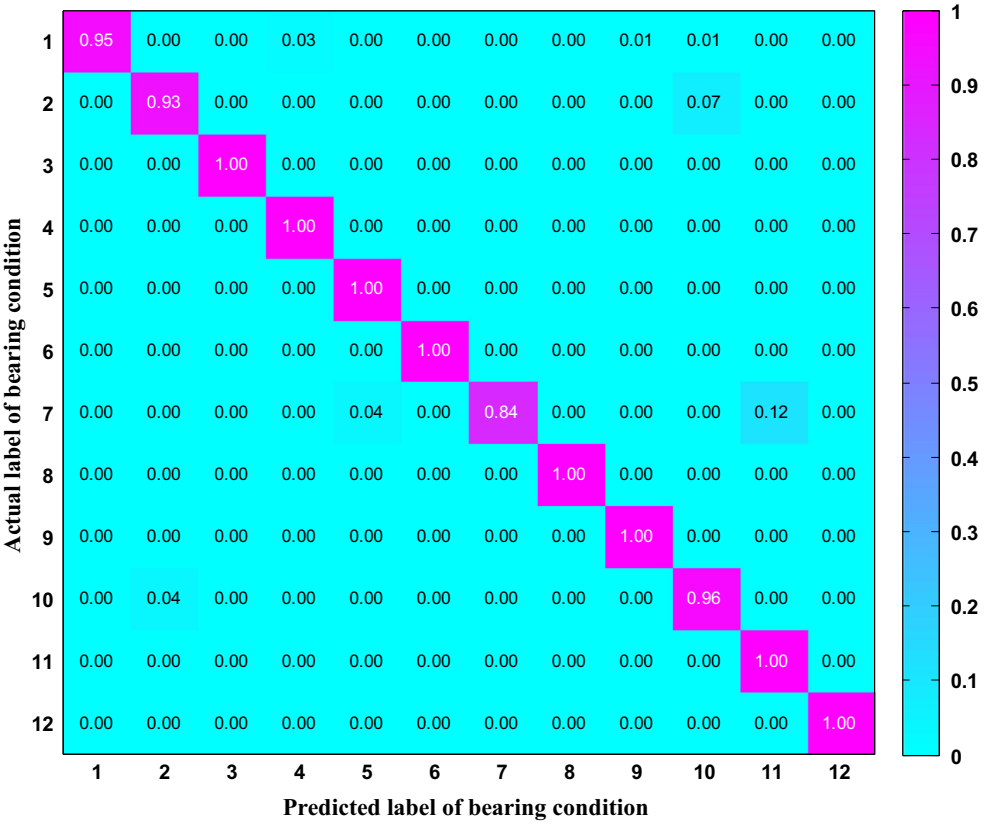

Fig. 9. Multi-class confusion matrix of the proposed method for the first trial.

Table 4
Parameters used in rolling bearing fault diagnosis.

Description	Value
The units of the input layer (The dimension of each sample)	400
The number of the different DAEs	15
The number of hidden layers in each DAE	3
The units of the first hidden layer	200
The units of the second hidden layer	100
The units of the third hidden layer	80
Learning rate of each DAE	0.05
Iteration times of each DAE	60
Sparsity parameter of each DAE	0.2
Sparse penalty factor of each DAE	6
The threshold of the accepted accuracy	0.91
The number of the repeated trials	10

Sparse penalty factor β and sparse parameter ρ in Eq. (6) are determined with a cross-validation technique. Fig. 10 shows the relationship between the testing accuracy and parameter set (β, ρ) for the second trial (97.17%, 1166/1200), where the candidate set of β is selected as [1,2,3,4,5,6,7,8,9] and ρ is [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9].

The threshold of the accepted accuracy is important for the designed combination strategy and standard majority voting strategy. Fig. 11 shows the influence of the accepted threshold on the testing accuracy. From Fig. 11, it can be seen that smaller threshold values will lead more individual DAEs to be considered, while those DAEs with low performance may affect the combined results. Larger threshold values will lead less individual DAEs to be considered, and less complementary information will be provided at the same time. Therefore, it is significant to select a reasonable threshold of the accepted accuracy in practical applications.

The main parameters of the other methods are described as follows.

- Method 2 (BPNN with raw data): The architecture is 400-801-12, which is decided by the guiding principles and experiences. The learning rate is 0.1 and the iteration number is 800.
- Method 3 (BPNN with 24 features): The architecture is 24-49-12, the learning rate is 0.1 and the iteration number is 500.
- Method 4 (SVM with raw data): RBF kernel is applied. The penalty factor and the radius of the kernel function are set to 30 and 0.15, respectively. Each of them is determined through a 5-fold cross validation.
- Method 5 (SVM with 24 features): RBF kernel is applied. The penalty factor and the radius of the kernel function are set to 40 and 0.25, respectively.
- Method 6 (RF with raw data): The number of trees is 500, and the number of predictors (*mtry*) randomly sampled at each split node is 20.
- Method 7 (RF with 24 features): The number of trees is 500, and the number of predictors (*mtry*) randomly sampled at each split node is 4.
- Method 8 (Boosting with raw data): The number of weak learners is 200, which is decided through cross-validation on the training samples, which means that the number of ensemble learning cycles is equal to 200. The most popular decision stump is used as a component in boosting, and the threshold error of every weak learner is 0.5.

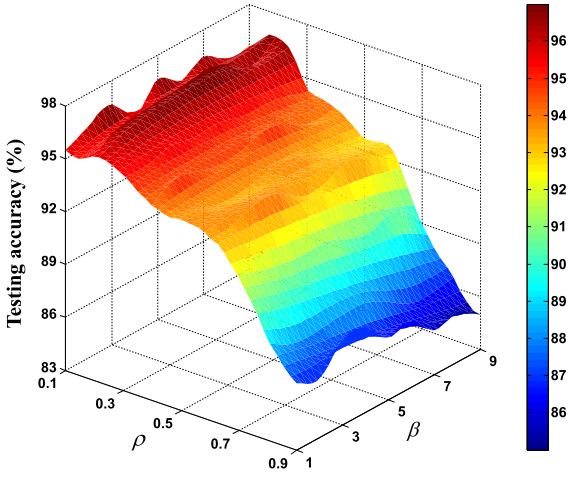

Fig. 10. The relationship between testing accuracy and parameter set (ρ, β) for the second trial.

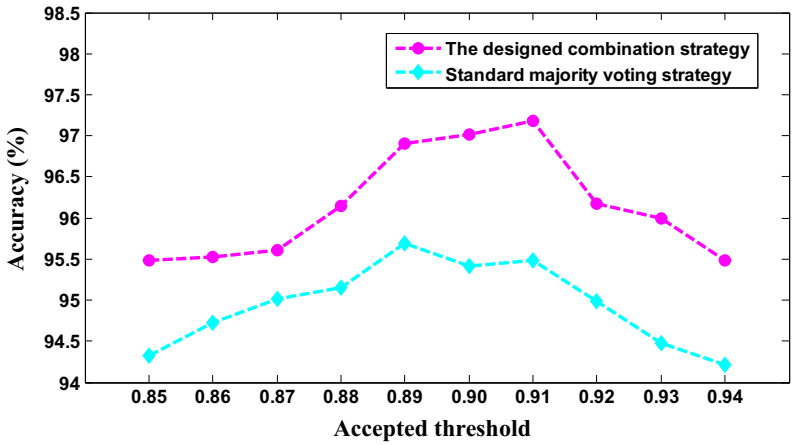

Fig. 11. The influence of the accepted threshold on the testing accuracy.

- Method 9 (Boosting with 24 features): The number of weak learners is set as 200, and the threshold error of every weak learner is 0.5

4.3. Comparison with the individual deep learning methods

It is necessary to compare the performance of the proposed method with various individual deep learning methods (individual DAE, DBN and CNN). Ten trials are carried out for each method based on the raw data, and the detailed weight assignments of the 15 DAEs in each trial are listed Table 5 (Accepted threshold = 0.91). The average testing accuracies of different methods are listed in Table 6, and Fig. 12 shows the detailed results of the 15 DAEs and EDAEs (the proposed method) in each trial. Besides, the standard deviations and average computing time are available in Table 6 (Core i5, 16-GB memory). It can be found that the testing accuracies of the 15 individual DAEs range from 82.20% to 93.98% and the average value is 90.36% (162,655/180,000). The testing accuracy of the proposed method is improved to 97.18%, which is 3.1% higher than the highest accuracy of the 15 DAEs and 6.72% higher than the average value. Compared with standard DBN and CNN, the proposed method also shows better diagnosis performance. The standard deviation of the proposed method is 0.1142, which is much smaller than other methods. However, the computing time of the proposed method is more than others due to the increase of individual models. As the modern hardware technology and training algorithm develop rapidly, we can accomplish ensemble deep learning methods much more efficiently soon [3]. Therefore, the proposed method is more accurate and stable than individual DAEs, DBN and CNN. The superiority of the proposed method arises from two main aspects: (1) The multiple non-linear transformations can automatically learn the deep features from the input data. Considering that the learned deep features (third-layer features) are high-dimensional data (80-dimension), linear discriminant analysis (LDA) is used for two-dimensional visualizations of the deep features learned by the 15 DAEs for the first trial, as shown

Table 5
Weight assignment of the 15 DAEs in each trial (Accepted threshold = 0.91).

Different DAEs	Weight assignment in each trial (Accurate to four decimal places)									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
DAE 1	0	0	0.1668	0	0	0.1382	0.1677	0	0	0
DAE 2	0.1236	0.1636	0	0	0	0	0	0	0.1222	0.1636
DAE 3	0	0	0	0.1253	0	0	0.1644	0	0.1251	0
DAE 4	0.1227	**0.1714**	0.1638	0.1228	0.1674	0	0	0	0	0
DAE 5	0	0	0.1673	0.1257	0	0	0.1635	0.1231	0	0
DAE 6	**0.1310**	0	0	**0.1269**	0	**0.1455**	0	0.1226	0	0.1668
DAE 7	0.1272	0	0.1664	0	0.1677	0	0	0	0.1280	0.1615
DAE 8	0	0	0	0.1240	0	0.1435	0	**0.1282**	**0.1289**	0
DAE 9	0	0.1627	0	0	0.1656	0.1442	0	0	0.1218	**0.1706**
DAE 10	0.1231	0.1663	0	0.1262	0	0.1419	0	0.1271	0	0
DAE 11	0.1231	0	0	0.1240	0.1644	0	**0.1694**	0.1233	0.1262	0
DAE 12	0	0.1672	0	0	0.1659	0	0	0.1244	0	0.1698
DAE 13	0.1227	0	0.1667	0	0	0.1429	0.1685	0.1258	0.1224	0
DAE 14	0	0	0	0	0	0	0	0	0	0
DAE 15	0.1267	0.1687	0.1689	0.1251	**0.1690**	0.1438	0.1665	0.1254	0.1254	0.1677

Bold values represent the largest weight assignment in each trial.

Table 6
Diagnosis results of different deep learning methods (Accepted threshold = 0.91).

Deep learning methods	Testing diagnosis result (%)	Average time (s)
Standard DAE 1 (Identity)	89.57 ± 2.4132 (10,748/12,000)	22.44
Standard DAE 2 (ArcTan)	82.20 ± 3.7686 (9864/12,000)	38.90
Standard DAE 3 (TanH)	90.00 ± 2.4957 (10,800/12,000)	53.38
Standard DAE 4 (Sinusoid)	88.98 ± 3.3393 (10,678/12,000)	33.31
Standard DAE 5 (Softsign)	89.38 ± 3.4641 (10,726/12,000)	50.16
Standard DAE 6 (ReLU)	91.75 ± 3.4785 (11,010/12,000)	29.83
Standard DAE 7 (Leaky ReLU)	91.58 ± 2.3626 (10,990/12,000)	30.07
Standard DAE 8 (PReLU)	91.05 ± 4.0765 (10,926/12,000)	31.11
Standard DAE 9 (ELU)	91.15 ± 2.6971 (10,938/12,000)	30.39
Standard DAE 10 (SoftPlus)	91.53 ± 2.1912 (10,984/12,000)	49.28
Standard DAE 11 (Bent identity)	91.28 ± 2.3175 (10,954/12,000)	26.17
Standard DAE 12 (SoftExponential)	90.95 ± 2.7194 (10,914/12,000)	53.66
Standard DAE 13 (Sinc)	90.72 ± 3.1181 (10,886/12,000)	33.77
Standard DAE 14 (Sigmoid)	91.33 ± 2.5252 (10,960/12,000)	53.83
Standard DAE 15 (Gaussian)	93.98 ± 0.6288 (11,278/12,000)	52.87
Average of the 15 DAEs	90.36 (162,655/180,000)	39.28
Standard DBN	86.98 ± 2.9810 (10,438/12,000)	59.67
Standard CNN	91.97 ± 1.0206 (11,036/12,000)	78.84
The proposed method (Ensemble DAEs)	97.18 ± 0.1142 (11,661/12,000)	589.17

(Note: the format of the diagnosis result is average accuracy ± standard deviation.)

in Fig. 13, in which LDA1 and LDA2 represent the first two principle components, respectively. The annotations 1-12 corresponding to the bearing condition labels listed in Table 2. (2) The proposed method can take full advantage of the complementary information provided by different individual DAEs.

The main parameters of standard DBN and CNN are described as follows: (1) standard DBN: the structure of DBN is 400-250-100-100-12, which is determined by experimentation. The learning rate, momentum and iteration number are 0.1, 0.9 and 200, respectively. (2) Standard CNN: the structure of CNN consists of an input layer, two convolutional layers, two pooling layers and an output layer. The size of the input layer is $20 * 20$, the first convolutional layer and the second convolutional layer has 6 kernels and 12 kernels, respectively. The scales of two pooling layers are both set to 2, the learning rate and iteration number are 0.1 and 200, respectively.

In this section, we also research the influence of sample dimension on the performance of the proposed method. Fig. 14 shows the evolution curve of testing accuracy as the dimension of each sample increases from 150 to 600 by 50. Under different cases, the numbers of the training samples and testing samples are always set to 150 and 50, respectively. From

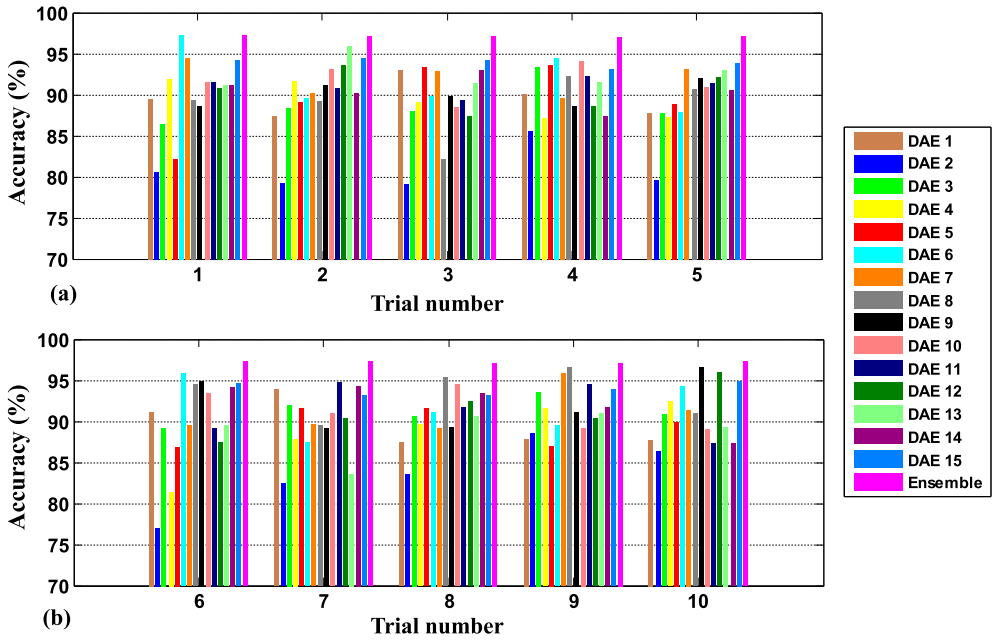

Fig. 12. Detailed diagnosis results of different individual DAEs and ensemble DAEs for ten trials. (a) From the first trial to the fifth trial; (b) from the sixth trial to the tenth trial.

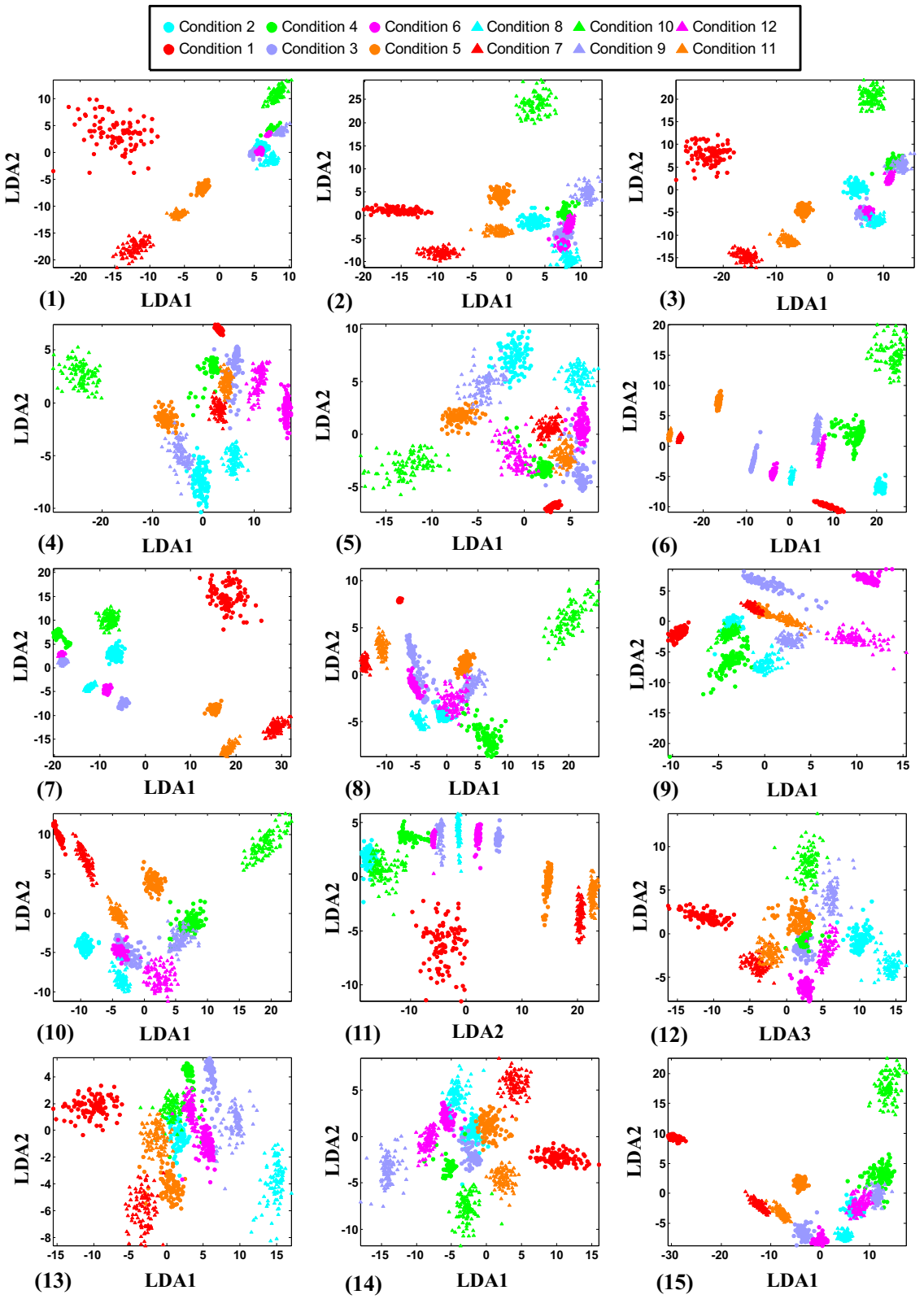

Fig. 13. Two-dimensional visualizations of the deep features learned by the 15 kinds of DAEs using LDA for the first trial. (1) DAE 1; (2) DAE 2; (3) DAE 3; (4) DAE 4; (5) DAE 5; (6) DAE 6; (7) DAE 7; (8) DAE 8; (9) DAE 9; (10) DAE 10; (11) DAE 11; (12) DAE 12; (13) DAE 13; (14) DAE 14; (15) DAE 15.

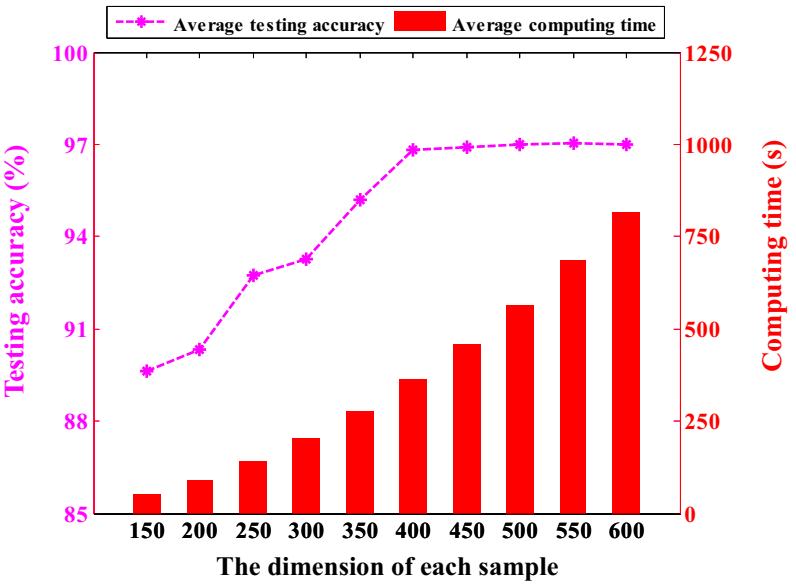

Fig. 14. The relationship between the diagnosis performance and the dimension of each sample.

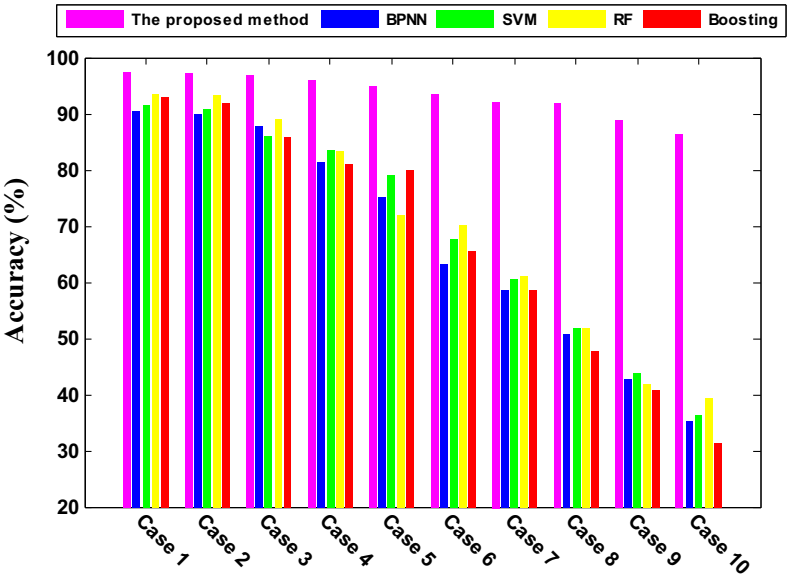

Fig. 15. Detailed diagnosis results of different methods on ten unbalanced cases.

Fig. 15, it can be clearly found that the testing accuracies greatly increase as the dimensions increase from 150 to 400. When the dimension are larger than 400, the testing accuracies keep basically stable. Unlike the testing accuracies, the computing time always increases greatly from 150 to 600. Generally, higher dimension of each sample will achieve higher testing accuracy and more computing time for deep learning models. Therefore, reasonable dimension of each sample is very important.

4.4. Influence of the unbalanced training dataset

In Section 4.2, we only consider bearing fault classification based on the balanced training dataset. In practical engineering, the percentage of normal sample is usually much larger than each kind of fault sample. Thus, how to effectively solve the fault classification problem with the unbalanced dataset has been a great challenge in mechanical fault diagnosis field. In this section, we investigate the diagnosis performance of the proposed method in dealing with the unbalanced training dataset.

As listed in Table 7, ten kinds of unbalanced cases are designed for comparing the performance of 5 different methods (The proposed method, BPNN with 24 features, SVM with features, RF with 24 features, Boosting with 24 features). Under the ten cases, the number of the testing samples in each condition is always set to 100, while the ratios of the normal sample and each kind of fault sample in training samples are 200:180, 200:170, 200:160, 200:150, 200:140, 200:130, 200:120, 200:110 and 200:100, respectively. It should be noted that Case 1 is the same as Section 4.2, i.e., each bearing condition contains 200 training samples and 100 testing samples. Ten trials are performed for analyzing each unbalanced dataset based on the five methods, and the best result of each method among the ten trials is reported as its corresponding diagnosis result. Fig. 16 shows the detailed results of the 5 methods in each trial. In Case 1, the diagnosis accuracy of the proposed method is 97.33%, compared with other four methods, which are 90.33%, 91.42%, 93.58% and 92.83%, respectively. In Case 6, the diagnosis accuracy of the proposed method is 93.50%, and it is much higher than others, which are 93.50%, 63.25%, 67.67%, 70.08% and 65.42%, respectively. It can be clearly found that despite highly unbalanced training dataset will result in low diagnosis performance of different methods, compared with other methods, the proposed method shows better generalization performance.

In order to quantitatively research the classification performance of different methods based on the unbalanced dataset, precision rate, recall rate and F-measure are calculated, which are expressed as

$$\mathrm{P} = \mathrm{precision} = \frac{TP}{TP + FP} \times 100 \tag{15}$$

$$\mathrm{R} = \mathrm{recall} = \frac{TP}{TP + FN} \times 100 \tag{16}$$

$$\mathrm{F-\ measure} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} = \frac{2TP}{2TP + FP + FN} \tag{17}$$

where TP represents the number of true positive instances, FP represents the number of false positive instances, and FN represents the number of false negative instances. It can be seen from Eq. (17) that F-measure index contains the information of precision rate and recall rate, whose value ranges from 0 (worst) to 1 (best) [36].

Table 7
Description of the unbalanced training dataset.

Unbalanced cases	Size of normal sample		Size of each kind of fault sample	
	Training sample	Testing sample	Training sample	Testing sample
Case 1	200	100	200	100
Case 2	200	100	180	100
Case 3	200	100	170	100
Case 4	200	100	160	100
Case 5	200	100	150	100
Case 6	200	100	140	100
Case 7	200	100	130	100
Case 8	200	100	120	100
Case 9	200	100	110	100
Case 10	200	100	100	100

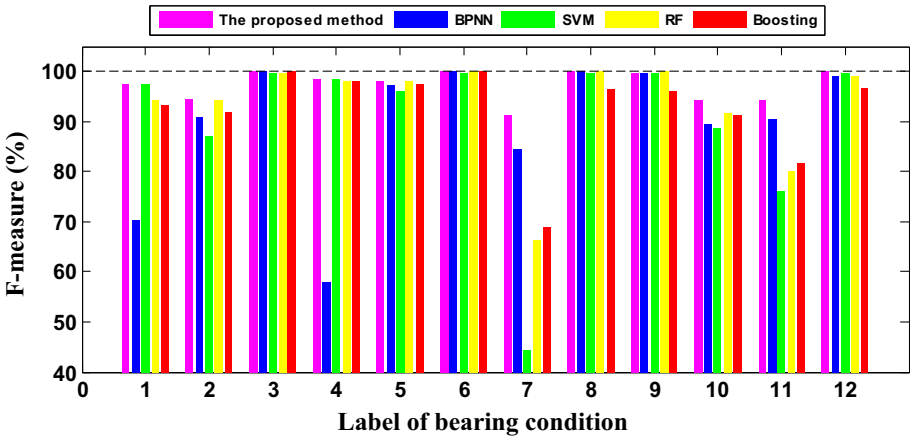

Fig. 16. F-measures of the bearing data using different methods in Case 1.

Table 8 lists the precision and recall rates of the five methods in Case 1, and Fig. 16 shows the corresponding F-measure values (Percentage form). Table 9 lists the precision and recall rates of the five methods in Case 2, and Fig. 17 shows their corresponding F-measure values. In Case 1, the F-measure values of the 12 bearing conditions using the proposed method are 97.44%, 94.42%, 100%, 98.52%, 98.04%, 100%, 91.30%, 100%, 99.50%, 94.12%, 94.34% and 100%, respectively, which are

Table 8
Precision rate and recall rate using different methods in Case 1.

Bearing condition	The proposed method		BPNN		SVM		RF		Boosting	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Condition 1	100	95	59.86	85	100	95	100	89	100	87
Condition 2	95.88	93	100	83	100	77	97.85	91	100	85
Condition 3	100	100	100	100	99.01	100	100	99	100	100
Condition 4	97.09	100	89.58	43	98.99	98	98	98	98.98	97
Condition 5	96.15	100	94.34	100	92.59	100	100	96	97.03	98
Condition 6	100	100	100	100	100	99	100	100	100	100
Condition 7	100	84	100	73	93.55	29	91.23	52	94.74	54
Condition 8	100	100	100	100	99.01	100	100	100	100	93
Condition 9	99.01	100	99.01	100	100	99	100	100	92.59	100
Condition 10	92.31	96	80.65	100	79.37	100	85.96	98	84.03	100
Condition 11	89.29	100	82.64	100	61.35	100	66.67	100	68.97	100
Condition 12	100	100	98.04	100	99.01	100	98.04	100	93.46	100

(Note: P and R represent the precision rate and recall rate, respectively.)

Table 9
Precision rate and recall rate using different methods for Case 6.

Bearing condition	The proposed method		BPNN		SVM		RF		Boosting	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Condition 1	100	95	100	58	100	60	100	63	100	55.56
Condition 2	89.33	67	50.96	53	54	54	58.62	51	65.26	62
Condition 3	100	100	26.19	11	72	90	67.91	91	68.52	74
Condition 4	98.04	100	95.24	100	100	95	87.72	50.51	98.68	75
Condition 5	98.04	100	92.31	12	90.91	20	53.27	57	60.22	56
Condition 6	100	100	49.47	94.95	78.05	96	54.81	57	73.79	76
Condition 7	95.29	81	27.19	31	54.72	29	80	36	39.84	49
Condition 8	100	89	85	34	36.40	87	63.85	83	46.55	27
Condition 9	90.91	100	59.75	95	58.64	95	58.33	91	52.98	80
Condition 10	73.02	92	78.89	71	81.44	79	82.05	64	82.86	58
Condition 11	85.22	98	56.18	100	38.89	7	74.44	99	51.41	73
Condition 12	99.01	100	94.34	100	94.34	100	94.29	99	80	100

(Note: P and R represent the precision rate and recall rate, respectively.)

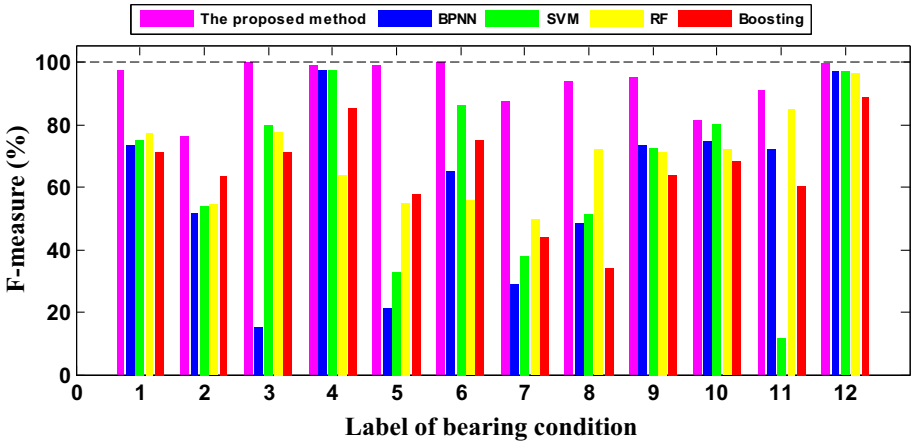

Fig. 17. F-measures of the bearing data using different methods in Case 6.

slightly higher than other four methods. In Case 6, the F-measure values of the 12 bearing conditions using the proposed method are 97.44%, 76.57%, 100%, 99.01%, 99.01%, 100%, 87.57%, 94.18%, 95.24%, 81.42%, 91.16% and 99.50%, respectively, which are much higher than other methods. The similar phenomena are more evident in Cases 7–10, where the unbalances between the normal sample and each kind of fault sample are more serious. The comparison results confirm the effectiveness of the proposed method in dealing with unbalanced dataset, which has outstanding advantages in precision rate, recall rate and F-measure over BPNN, SVM, RF and Boosting.

5. Conclusions

In this paper, a novel method called ensemble deep auto-encoders (EDAEs) is proposed for intelligent fault diagnosis of rolling bearings. The proposed method can be divided into three main steps: Firstly, different activation functions are employed as the hidden functions to design a series of auto-encoders with different characteristics. Secondly, the EDAEs are built with various auto-encoders for unsupervised feature learning from the measured vibration signals. Finally, the learned deep features are successively fed into *Softmax* classifiers for accurate and stable fault classification based on a combination strategy.

The proposed method is applied to diagnose the experimental bearings vibration data. The results confirm that the proposed method can get rid of the dependence on manual feature extraction and overcome the limitations of the individual deep learning models, which is more effective and robust than the existing intelligent diagnosis methods. It is very interesting to combine deep learning and ensemble learning for new applications. The authors would continue to research this topic in the future.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 51475368), Shanghai Engineering Research Center of Civil Aircraft Health Monitoring Foundation of China (No. GCZX-2015-02), and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (No. CX201710).

References

[1] H.K. Jiang, C.L. Li, H.X. Li, An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis, Mech. Syst. Signal Process. 36 (2013) 225–239.
[2] Y.H. Miao, M. Zhao, J. Lin, Y.G. Lei, Application of an improved maximum correlated kurtosis deconvolution method for fault diagnosis of rolling element bearings, Mech. Syst. Signal Process. 92 (2017) 173–195.
[3] F. Jia, Y.G. Lei, J. Lin, X. Zhou, N. Lu, Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, Mech. Syst. Signal Process. 72–73 (2016) 303–315.
[4] H.D. Shao, H.K. Jiang, H.W. Zhao, F.A. Wang, A novel deep autoencoder feature learning method for rotating machinery fault diagnosis, Mech. Syst. Signal Process. 98 (2017) 187–204.
[5] Z.X. Wei, Y.X. Wang, S.L. He, J.D. Bao, A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection, Knowl.-Based Syst. 116 (2017) 1–12.
[6] X.X. Ding, Q.B. He, N.W. Luo, A fusion feature and its improvement based on locality preserving projections for rolling element bearing fault classification, J. Sound Vib. 335 (2015) 367–383.
[7] W. Li, Z.C. Zhu, F. Jiang, G.B. Zhou, G.A. Chen, Fault diagnosis of rotating machinery with a novel statistical feature extraction and evaluation method, Mech. Syst. Signal Process. 50–51 (2015) 414–426.
[8] J. Singh, A.K. Darpe, S.P. Singh, Bearing damage assessment using Jensen-Rényi Divergence based on EEMD, Mech. Syst. Signal Process. 87 (2017) 307–339.
[9] Y.G. Lei, Z.J. He, Y.Y. Zi, EEMD method and WNN for fault diagnosis of locomotive roller bearings, Exp. Syst. Appl. 38 (2011) 7334–7341.
[10] X.L. Zhang, W. Chen, B.J. Wang, X.F. Chen, Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization, Neurocomputing 167 (2015) 260–279.
[11] H.D. Shao, H.K. Jiang, F.A. Wang, H.W. Zhao, An enhancement deep feature fusion method for rotating machinery fault diagnosis, Knowl.-Based Syst. 119 (2017) 200–220.
[12] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[13] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Networks 61 (2015) 85–117.
[14] H.D. Shao, H.K. Jiang, X. Zhang, M.G. Niu, Rolling bearing fault diagnosis using an optimization deep belief network, Meas. Sci. Technol. 26 (2015) 115002.
[15] M. Gan, C. Wang, C.A. Zhu, Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings, Mech. Syst. Signal Process. 72–73 (2016) 92–104.
[16] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, Convolutional neural network based fault detection for rotating machinery, J. Sound Vib. 377 (2016) 331–345.
[17] W.T. Mao, L. He, Y.J. Yan, J.W. Wang, Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine, Mech. Syst. Signal Process. 83 (2017) 450–473.
[18] X.L. Zhang, B.J. Wang, X.F. Chen, Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine, Knowl.-Based Syst. 89 (2015) 56–85.
[19] J.D. Zheng, H.Y. Pan, J.S. Cheng, Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines, Mech. Syst. Signal Process. 85 (2017) 746–759.
[20] M. Cerrada, G. Zurita, D. Cabrera, R.V. Sánchez, M. Artés, C. Li, Fault diagnosis in spur gears based on genetic algorithm and random forest, Mech. Syst. Signal Process. 70–71 (2016) 87–103.
[21] N. Morizet, N. Godin, J. Tang, E. Maillet, M. Fregonese, B. Normand, Classification of acoustic emission signals using wavelets and Random Forests, Mech. Syst. Signal Process. 70–71 (2016) 1026–1037.
[22] Y. LeCun, Y. Bengio, G.E. Hinton, Review: deep learning, Nature 521 (2015) 436–444.
[23] H. Schulza, K. Chob, T. Raikob, S. Behnkea, Two-layer contractive encodings for learning stable nonlinear features, Neural Network 64 (2015) 4–11.

[24] M. Nielsen, Neural Networks and Deep Learning, Determination Press, 2015.
[25] X.S. Ding, J.D. Cao, A. Alsaedi, F.E. Alsaadi, T. Hayat, Robust fixed-time synchronization for uncertain complex-valued neural networks with discontinuous activation functions, Neural Networks 90 (2017) 42–55.
[26] S.S. Liew, M. Khalil-Hani, R. Bakhteri, Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems, Neurocomputing 216 (2016) 718–734.
[27] X.M. Jiang, S. Mahadevan, Y. Yuan, Fuzzy stochastic neural network model for structural system identification, Mech. Syst. Signal Process. 82 (2017) 394–411.
[28] N. Vinod, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, Int. Conf. Mach. Learn. (2010) 807–814.
[29] X.J. Jin, C.Y. Xu, J.S. Feng, Y.C. Wei, J.J. Xiong, S.C. Yan, Deep learning with S-shaped Rectified linear activation units, Comput. Sci. 3 (2015) 1–8.
[30] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, Delving deep into rectifiers: surpassing human-Level performance on ImageNet classification, IEEE Int. Conf. Comput. Vis. IEEE (2016) 1026–1034.
[31] G.E. Hinton, S. Osindero, Y.W. The, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[32] J. Tang, W. Yu, T.Y. Chai, Z. Liu, X.J. Zhou, Selective ensemble modeling load parameters of ball mill based on multi-scale frequency spectral features and sphere criterion, Mech. Syst. Signal Process. 66–67 (2016) 485–504.
[33] L. Wang, C. Wu, Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map, Knowl.-Based Syst. 121 (2017) 99–110.
[34] http://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website.
[35] 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 171 (2016) 837–853.
[36] Y.G. Lei, F. Jia, J. Lin, S.B. Xing, S.X. Ding, An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data, IEEE Trans. Industr. Electron. 63 (2016) 3137–3147.