Knowledge-Based Systems 119 (2017) 200–220

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An enhancement deep feature fusion method for rotating machinery fault diagnosis

Haidong Shao, Hongkai Jiang*, Fuan Wang, Huiwei Zhao

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

ARTICLE INFO

Article history:
Received 10 August 2016
Revised 9 December 2016
Accepted 11 December 2016
Available online 13 December 2016

Keywords:
Deep feature fusion
Feature enhancement
Fault diagnosis
Rotating machinery
Locality preserving projection

ABSTRACT

It is meaningful to automatically learn the valuable features from the raw vibration data and provide accurate fault diagnosis results. In this paper, an enhancement deep feature fusion method is developed for rotating machinery fault diagnosis. Firstly, a new deep auto-encoder is constructed with denoising auto-encoder (DAE) and contractive auto-encoder (CAE) for the enhancement of feature learning ability. Secondly, locality preserving projection (LPP) is adopted to fuse the deep features to further improve the quality of the learned features. Finally, the fusion deep features are fed into *softmax* to train the intelligent diagnosis model. The developed method is applied to the fault diagnosis of rotor and bearing. The results confirm that the proposed method is more effective and robust compared with the existing methods.

1. Introduction

Rotating machinery has a wide range of applications in modern industry. After a long-term running under the complex and variable operating conditions, the key parts of rotating machinery will inevitably get varieties of faults, which may affect the performance of the whole machine and result in serious security accidents [1]. Therefore, it is meaningful to accurately and automatically diagnose the different faults which may happen in rotating machinery.

Intelligent fault diagnosis of rotating machinery is a typical pattern recognition problem [2]. Traditional pattern recognition methods are shallow learning models such as artificial neural network (ANN) and support vector machine (SVM). The diagnosis performance of the shallow learning models depends heavily on the quality of the extractWed features from the collected vibration signals. Currently, the most popular method for feature extraction is manually construction [3]. Lei et al. [4] calculated ten statistical parameters to reflect bearing conditions, and then input the selected features into ANN for fault classification. Liu et al. [5] extracted multi-scale entropy and adopted a back propagation (BP) neural network to detect rolling bearing faults. Four singular values were designed by Muruganatham et al. [6] for monitoring the bearing conditions, and then BP neural network was employed for distinguishing the faults. Yu et al. [7] constructed empirical mode decomposition energy entropy as input feature and used ANN as the fault classifier. Zhang et al. [8] extracted nineteen statistical parameters from the measured vibration signals, and carried out SVM to recognize the roller bearing operation conditions. Hussein et al. [9] introduced coefficients of linear time invariant autoregressive model as input features and applied nearest neighbor classifier for fault identification. Hierarchical fuzzy entropy was presented by Li et al. [10] for representing the rolling bearing conditions and then improved SVM was adopted for fault diagnosis. Liu et al. [11] proposed the most relevance atoms as features and used SVM for bearing fault diagnosis. Volterra series was extracted by Xia et al. [12] to characterize working conditions of rotor system and ANN was applied as the fault classifier. Lu et al. [13] calculated non-dimensional symptom parameters for reflecting rotor working conditions, and then the selected sensitive features were fed into K-means classifier for fault diagnosis. Chen et al. [14] developed order tracking for feature extraction and ANN for rotor fault classification. Keskes et al. [15] presented stationary wavelet packet transform for feature extraction and SVM for rotor bar fault diagnosis. Mustafa et al. [16] adopted spectral analysis for feature extraction and SVM for rotor fault detection. Although manual feature extraction has been successfully used in traditional intelligent methods, it has two obvious shortcomings: (1) The features are extracted by diagnostic experts from the collected signals, largely depending on the advanced and complex signal processing techniques. (2) Lots of time is needed to select the most sensitive features according to the specific diagnosis issue, however, they are probably unreliable for the new issues. In other words, manual feature extraction is time-consuming and labor-intensive [3], which encourages re-

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

http://dx.doi.org/10.1016/j.knosys.2016.12.012
0950-7051/

searchers to explore new methods to automatically learn the representative features from the raw data.

Deep learning is a new machine learning method, which has the great potential to overcome the inherent shortcomings of traditional intelligent methods. The most attractive advantage of deep learning methods is that they can adaptively learn useful features from the raw data through multiple nonlinear transformations [17]. At present, deep learning methods can be mainly divided into three categories: deep auto-encoder, deep belief network (DBN) and convolutional neural network (CNN). In the last three years, different deep learning models have been gradually applied to machinery fault diagnosis. Shao et al. proposed optimized DBN for rolling bearing fault diagnosis [18]. Tamilselvan et al. applied DBN for aircraft engine fault diagnosis [19]. Tran et al. combined DBN and Teager–Kaiser energy operator for reciprocating compressor valves fault diagnosis [20]. Janssens et al. adopted CNN for rotating machinery fault detection [21]. Chen et al. used CNN for gearbox fault identification and classification [22]. Sun et al. designed sparse auto-encoder-based deep neural network approach for induction motor faults classification [23]. Compared with standard DBN and CNN, deep auto-encoder is a purely unsupervised feature learning model, which is stacked with several auto-encoders (AEs). Denoising auto-encoder (DAE) and contractive auto-encoder (CAE) are two extensions of the basic auto-encoder, which have been successfully used in various applications [24-27]. DAE can reconstruct the clean data and learn useful information from the corrupted inputs. By adding the contractive penalty term, CAE can learn more robust feature representations. In other words, DAE tends to encourage the robustness of reconstruction, and CAE explicitly encourages the robustness of representation. However, current deep auto-encoders are mostly constructed with some simple and same base models, which is incapable of taking full advantages of different base models simultaneously. Therefore, the development of new deep auto-encoders for feature learning and enhancement has become an urgent task.

Despite deep learning models can automatically capture the effective features from raw vibration data to a great extent, the learned deep features are usually high-dimensional and contain redundant information [28], which may decrease the diagnosis performance and lead to more training time. In order to further improve the quality of the learned deep features and diagnosis efficiency, locality preserving projection (LPP), a popular feature fusion method, is adopted to fuse the deep features to extract the most representative information and reduce the dimension in this paper. Several studies have confirmed the remarkable superiorities of LPP for feature fusion compared with principle component analysis (PCA) [29–31].

In this paper, an enhancement deep feature fusion method is developed for rotating machinery fault diagnosis. The proposed method is applied to the fault diagnosis of rotor and electrical locomotive bearing. The results show that the proposed method can not only automatically learn the effective features from the raw vibration data, but also achieve superior diagnosis accuracy to the existing methods. The main contributions of this paper can be summarized as follows.

(1) In order to get rid of the dependence on signal processing techniques and manual feature extraction, we propose deep learning to automatically learn the useful features from the raw vibration data.
(2) In order to take full advantages of different auto-encoders for the enhancement of feature learning ability, a new deep auto-encoder is constructed with denoising auto-encoder and contractive auto-encoder.
(3) In order to improve the diagnosis efficiency, locality preserving projection is adopted to fuse the deep features to extract the most representative information.

The rest of the paper is organized as follows. In Section 2, the theory of basic auto-encoder is briefly introduced. The proposed method is described in Section 3. In Section 4, the experimental diagnosis results for rotor are analyzed and discussed. The engineering application of the proposed method is given in Section 5. Finally, general conclusions are presented in Section 6.

2. Basic auto-encoder

The basic auto-encoder is unsupervised neural network, consists of an encoder and a decoder. The auto-encoder aims to learn a compressed representation of input data. Given a training sample $\mathbf{x}=[x_1, x_2, \ldots, x_N]^T \in \Re^N$, the encoder maps the input vector $\mathbf{x}$ to a hidden representation $\mathbf{h}=[h_1, h_2, \ldots, h_M]^T \in \Re^M$ through the sigmoid function [32]

$$\mathbf{h} = f_\theta(\mathbf{x}) = s_f(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{1}$$

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

where $s_f(t)$ is the activation function of encoder, and $\theta=\{\mathbf{W}, \mathbf{b}\}$ is the parameters of encoder. $\mathbf{W} \in \Re^{M\times N}$ is the weight matrix, and $\mathbf{b} \in \Re^M$ is a bias vector.

Then, the hidden vector $\mathbf{h}$ is transformed back into a reconstruction vector $\mathbf{z}=[z_1, z_2, \ldots, z_N]^T \in \Re^N$ by the decoder as follows

$$\mathbf{z} = g_{\theta'}(\mathbf{h}) = s_g\left(\mathbf{W}'\mathbf{h} + \mathbf{b}'\right) \tag{3}$$

$$s_g(t) = 1/(1+e^{-t}) \tag{4}$$

where $s_g(t)$ is the activation function of decoder, and $\theta'=\{\mathbf{W'}, \mathbf{b'}\}$ is the parameters of decoder. Weight matrix $\mathbf{W'}=\mathbf{W}^T$, and $\mathbf{b'} \in \Re^N$ is a bias vector.

The parameter set $\boldsymbol{\theta}=\{\theta, \theta'\}=\{\mathbf{W}, \mathbf{b}, \mathbf{W'}, \mathbf{b'}\}$ of the basic auto-encoder is optimized to minimize the average reconstruction error [32]

$$J_{\mathrm{AE}}(\theta) = L(\mathbf{x}, \mathbf{z}) = -\sum_{i=1}^{N}[x_i \log(z_i) + (1-x_i)\log(1-z_i)] \tag{5}$$

where $L(\mathbf{x}, \mathbf{z})$ is the loss function which measures the difference between $\mathbf{x}$ and $\mathbf{z}$.

3. The proposed method

In this paper, we develop an enhancement deep feature fusion method for rotating machinery fault diagnosis. This method includes three parts: new deep auto-encoder construction for the enhancement of feature learning, deep feature fusion using LPP and the general procedure of the proposed method.

3.1. New deep auto-encoder construction for the enhancement of feature learning

DAE can reduce the influence of background noises and learn the robust reconstruction of real vibration data. CAE can strengthen the feature learning ability of hidden layer units, and learn the robust representation. In this paper, we combine the advantages of DAE and CAE to further improve the feature learning ability. Specifically, DAE is used to learn the low-layer features from the raw vibration data, and CAEs are used to learn the deep features based on the learned low-layer features. Finally, the learned deep features are robust to small perturbation to the inputs.

DAE is trained to reconstruct the original input sample $\mathbf{x}$ from its corrupted version $\tilde{\mathbf{x}}$. The most popular corruption choice is additive Gaussian noise. Similar to the basic auto-encoder, the main objective for DAE is to minimize the reconstruction error defined as [25]

$$J_{\mathrm{DAE}}(\theta) = L(\mathbf{x}, \tilde{\mathbf{z}}) = -\sum_{i=1}^{N} [x_i \log(\tilde{z}_i) + (1 - x_i)\log(1 - \tilde{z}_i)] \tag{6}$$

with

$$\tilde{\mathbf{z}} = g_{\theta'}(\tilde{\mathbf{h}}) = s_g(\mathbf{W}'\tilde{\mathbf{h}} + \mathbf{b}') \tag{7}$$

$$\tilde{\mathbf{h}} = f_\theta(\tilde{\mathbf{x}}) = s_f(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b}) \tag{8}$$

$$\tilde{\mathbf{x}} = \mathbf{x} + N(0, \sigma^2 \mathrm{I}) \tag{9}$$

where $\mathbf{x}$ is the original input sample, $\tilde{\mathbf{x}}$ is the corrupted sample, and σ is the noise level parameter, which represents the degree of the corruption.

CAE is trained to learn the robust intrinsic feature representations by adding the contractive penalty term. The loss function of CAE can be expressed as [26]

$$J_{\mathrm{CAE}}(\theta) = L(\mathbf{x}, \mathbf{z}) + \lambda \|J_{\mathbf{h}}(\mathbf{x})\|^2 \tag{10}$$

with

$$\|J_{\mathbf{h}}(\mathbf{x})\|^2 = \left\| \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right\|^2 = \sum_{j=1}^{M}\sum_{i=1}^{N} \left(\frac{\partial h_j}{\partial x_i} \right)^2 \tag{11}$$

$$J_{\mathbf{h}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} \cdots \frac{\partial h_1}{\partial x_N} \\ \vdots \ddots \vdots \\ \frac{\partial h_M}{\partial x_1} \cdots \frac{\partial h_M}{\partial x_N} \end{bmatrix} \tag{12}$$

where $J_{\mathbf{h}}(\mathbf{x})$ is the Jacobian matrix of the hidden units at the training sample, $\|J_{\mathbf{h}}(\mathbf{x})\|^2$ is the contractive penalty term, and λ is the regularization coefficient parameter. By minimizing the$\|J_{\mathbf{h}}(\mathbf{x})\|^2$, the learned hidden representations can be more invariant to small changes in the input.

Fig. 1 shows the layer-by-layer feature learning process of the constructed new deep auto-encoder with three hidden layers ((a)→(b)→(c)→(d)), which is similar to the basic deep auto-encoder [33]. Fig. 2 is the flowchart of the feature learning and enhancement based on the new deep auto-encoder. The detailed steps are summarized as follows. (1) Obtain the collected vibration data of rotating machinery equipment under different operating conditions, and determine the number of DAE and CAE. (2) DAE is pre-trained to automatically learn the first hidden layer features from the raw vibration data, which are used as the inputs of the first CAE. (3) The first CAE is pre-trained to learn the second hidden layer features. (4) Continue the feature learning steps until the last CAE is trained. In this way, all the hidden layers of the new deep auto-encoder have been pre-trained. (5) BP algorithm is adopted to fine-tune the parameters of the new deep auto-encoder to further improve the feature learning ability. (6) Output the learned deep features (highest hidden layer features).

By now, the constructed new deep auto-encoder has learned the effective deep features from the raw vibration data, and established complex non-linear relationships between the deep features and various operation conditions. The next step is deep feature fusion.

3.2. Deep feature fusion using LPP

The deep features learned by the new deep auto-encoder are usually high dimension and contain superfluous information. LPP is a novel feature fusion method, which can effectively recover the significant aspects of the intrinsic manifold structure. Therefore, LPP is used for removing the redundant information of the high-dimensional deep features to find the valuable low-dimensional representations in this paper. After the feature fusion using LPP, the diagnosis efficiency can be improved.

Let $F=[f_1, f_2, \ldots, f_n] \in \Re^D$ denote the learned deep feature matrix, where n and D represent the sample size and original feature dimension, respectively. The main idea of LPP is to find a transformation matrix A to project the high-dimensional input data set F into a low-dimensional space $Y=[y_1, y_2, \ldots, y_n] \in \Re^d$, where $d(d < D)$ is the fused feature dimension.

$$y_i = A^T f_i, \quad i = 1, 2, \ldots, n \tag{13}$$

where $A \in \Re^{D\times d}$ is the transformation matrix. A popular choice for determining the transformation is to minimize the objective function as following [29]

$$\min \sum_{i,j=1}^{n} (y_i - y_j)^2 S_{ij} \tag{14}$$

with

$$S_{ij} = \begin{cases} \frac{\exp(\|f_i - f_j\|^2)}{t}, & f_j \in N_k(f_i) \text{ or } f_i \in N_k(f_j) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where $\mathbf{S}=\{S_{ij}\} \in \Re^{n\times n}$ is the weight matrix calculated through the nearest-neighbor graph, and t is a positive parameter. $f_j \in N_k(f_i)$ denotes that f_j is among the k nearest neighbors of f_i.

Following some algebraic steps, the minimum optimization problem in Eq. (14) can be transformed into the following equation

$$\begin{aligned} \frac{1}{2}\sum_{i,j=1}^{n} (y_i - y_j)^2 S_{ij} &= \frac{1}{2}\sum_{i,j=1}^{n} (A^T f_i - A^T f_j)^2 S_{ij} \\ &= A^T F(\Phi - S)F^T A = A^T F L F^T A \end{aligned} \tag{16}$$

where Φ is a n-order diagonal matrix, i.e., $\Phi_{ii} = \sum_{j=1}^{n} S_{ij}$, and $L=\Phi - S$ is the Laplacian matrix.

Then, the transformation matrix can be acquired by solving the generalized eigenvalue problem

$$FLF^T A = \beta FDF^T A \tag{17}$$

where β is the eigenvalue and A is the corresponding eigenvector. The first d eigenvectors $a_1, a_2, \ldots, a_d$ are associated with their d smallest non-zero eigenvalues $\beta_1, \beta_2, \ldots, \beta_d$. The final low-dimensional embedding can be expressed by

$$Y = A^T F, \; A^T = (a_1, a_2, \ldots, a_d) \tag{18}$$

The fused low-dimensional vector Y preserves the local important information of the original deep features and will be used as the final input features of the fault classifier. *Softmax* classifier is a good choice for fault classification based on the deep features. In a word, in the proposed method, new deep auto-encoder is used as the feature extractor, LLP serves as the feature refiner, and *Softmax* as the fault classifier.

3.3. The procedure of the proposed method

In this paper, an enhancement deep feature fusion method is developed for rotating machinery fault diagnosis. The flowchart of the proposed method is shown in Fig. 3 and the general procedures are summarized as follows.

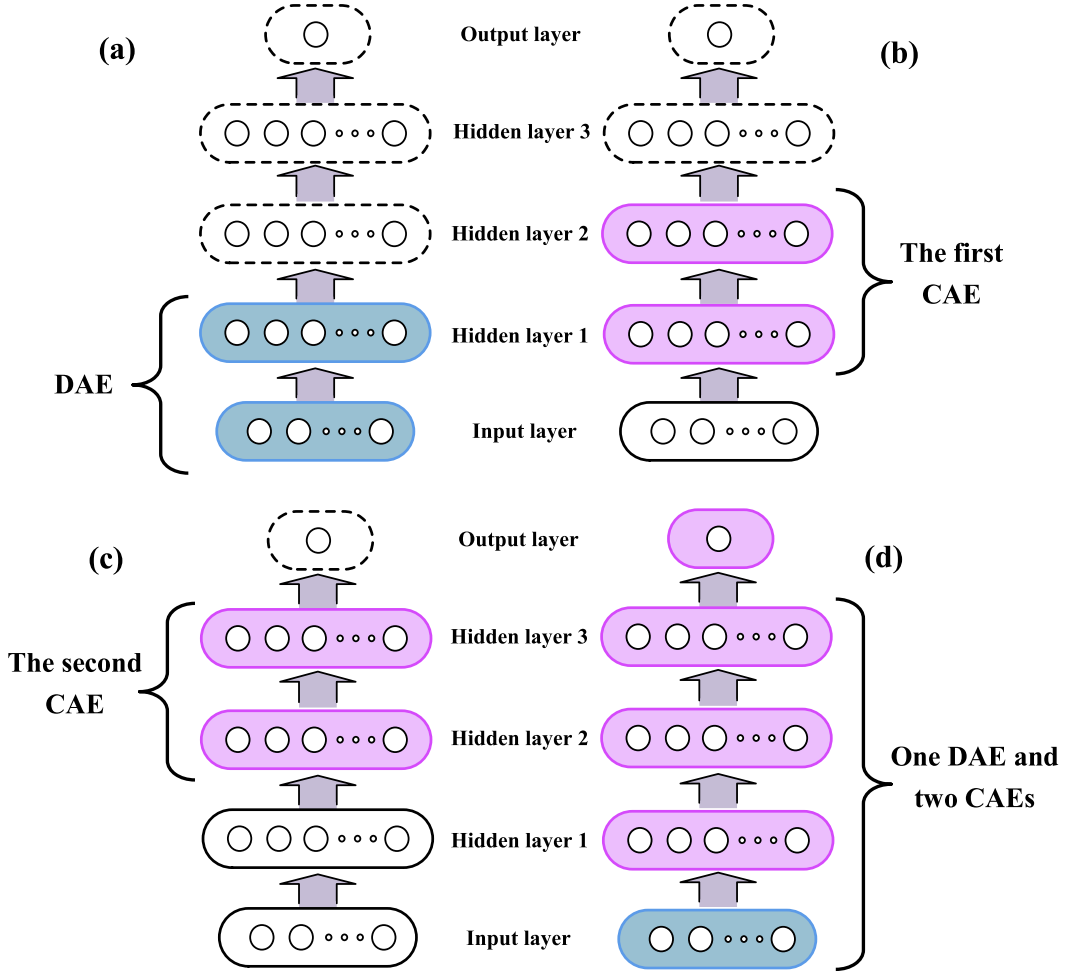

Fig. 1. Layer-by-layer feature learning process of the new deep auto-encoder with three hidden layers.

- Step 1: The vibration data of rotating machinery equipment is measured by sensors and collected by data acquisition system.
- Step 2: Without any signal preprocessing and manual feature extraction, the raw vibration data is divided into training samples and testing samples.
- Step 3: New deep auto-encoder is constructed with DAE and CAE to enhance the feature learning ability. Specifically, DAE is used to learn low-layer features from the training samples, and CAEs are used to learn deep features based on the low-layer features.
- Step 4: LPP is adopted for fusing the deep features to further remove the redundant information and improve diagnosis efficiency.
- Step 5: The fused deep features are fed into *Softmax* classifier for fault diagnosis.
- Step 6: Testing samples are used to verify the feasibility of the proposed method.

4. Experimental verification

4.1. Rotor fault test rig

Rotor is one of the key parts in rotating machinery. In this case study, rotor vibration data is used to verify the effectiveness of the proposed method. The custom-built rotor fault test rig is shown in Fig. 4, which mainly consists of motor, rotor, mass disk and eddy current sensor. The eddy current sensor is used for displacement measurement, and the measured vibration data is stored by the NI data acquisition system. The sampling frequency is 10 kHz, and the motor speed is about 1800 rpm.

Table 1
Description of the rotor operation conditions.

Rotor operation conditions	Size of training/testing samples	Label
Normal condition	140/60	1
Rub fault	140/60	2
Unbalance fault (2 *2.5 g)	140/60	3
Unbalance fault (2 *2.5 g + 1 *2.9 g)	140/60	4
Unbalance fault (2 *2.5 g + 2 *2.9 g)	140/60	5
Unbalance fault (2 *2.5 g + 3 *2.9 g)	140/60	6
Compound faults (rub and unbalance faults)	140/60	7

In this case study, seven operating conditions are considered, including normal condition, rub fault condition, four types of unbalance fault conditions and compound faults condition. The unbalance fault conditions are simulated by screwing different mass blocks (2.5 g or 2.9 g) into the threaded holes near the edge of the mass disk.

Details about the rotor operation conditions are available in Table 1. Each condition contains 200 samples, and each sample is a measured vibration signal consists of 1000 sampling data points. Therefore, the percentage of fault rotor in the training samples or testing samples is 85.7% (6/7), and in this case study, we just consider the fault diagnosis with balanced training dataset. In order to avoid the particularity and contingency of the diagnosis results, the

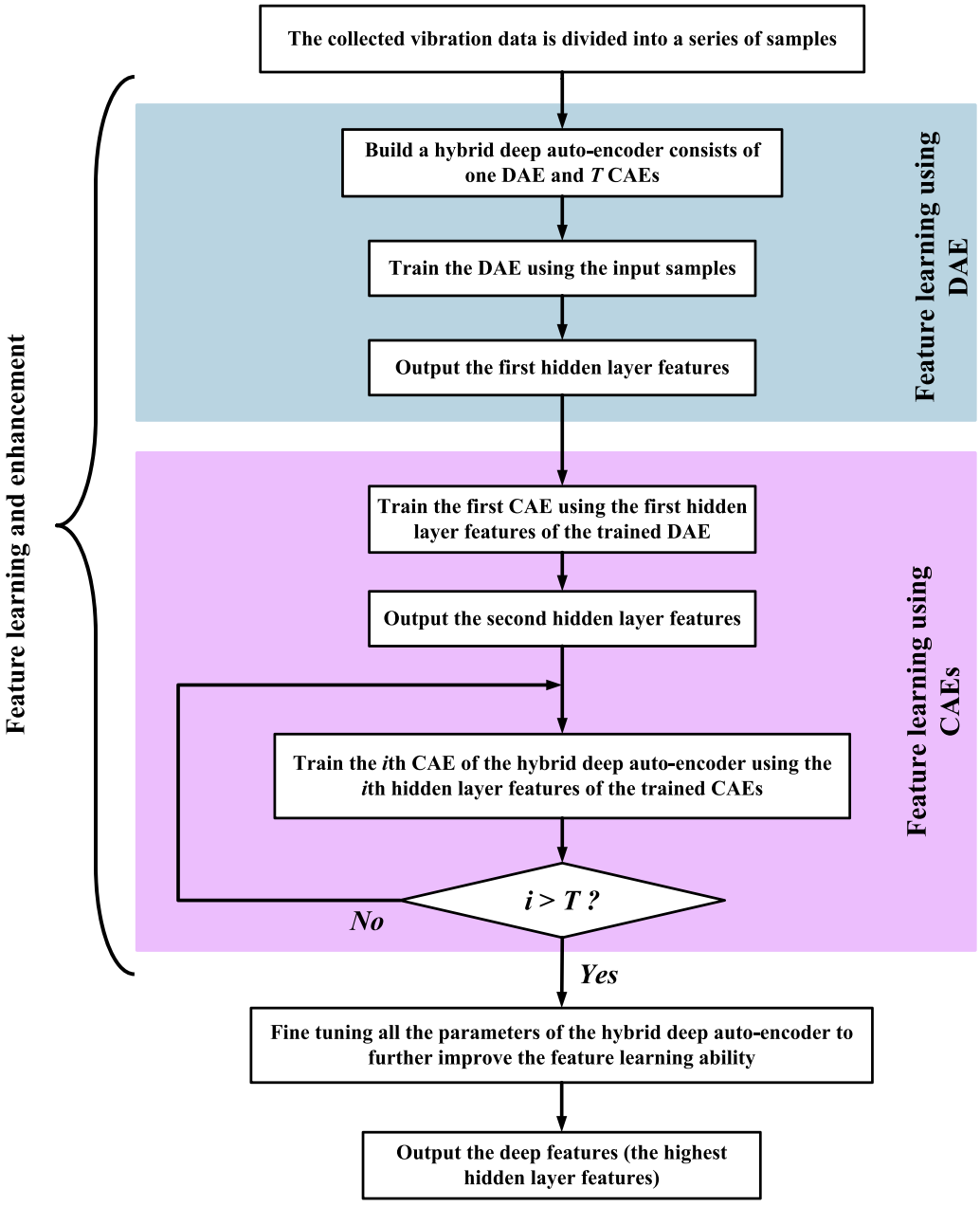

Fig. 2. Feature learning and enhancement based on the new deep auto-encoder.

random selection strategy of training samples introduced in many literatures [3,4,8,23] is also adopted in this paper. Besides, ten trials are carried out for diagnosing each dataset to show the stability of the proposed method. Specifically, random 140 (70%) samples of each condition sever as training samples, and the remaining 60 (30%) samples as testing samples. The collected vibration signals (first 20,000 data points) of the seven rotor operation conditions are shown in Fig. 5. It is very hard to distinguish the seven operation conditions only using the collected vibration signals.

For comparison, four other methods are also used for rotor fault diagnosis, including the standard deep DAE, standard deep CAE, BP neural network and SVM. In this case study, three experiments are taken into account, listed in Table 2. The specific goals of the three experiments are described as follows.

- Experiment 1: Without any signal preprocessing or manual feature extraction, the normalized vibration data is directly used as the input features for fault diagnosis.

Table 2
The descriptions of three designed experiments.

Experiments	Input features	Dimension
Experiment 1	Raw vibration data (without any manual feature extraction)	1000
Experiment 2	19 features extracted from eight frequency-band signals	152 (19*8)
Experiment 3	19 features extracted from the most sensitive frequency-band signal	19

- Experiment 2: Adopt three-layer wavelet packet (*db*5) to preprocess the vibration data, and manually extract 19 feature parameters the same as [8] from each frequency-band signal. Then the extracted features are fed into different methods for fault classification.
- Experiment 3: Adopt three-layer wavelet packet (*db*5) to preprocess the vibration data, and select the most sensitive

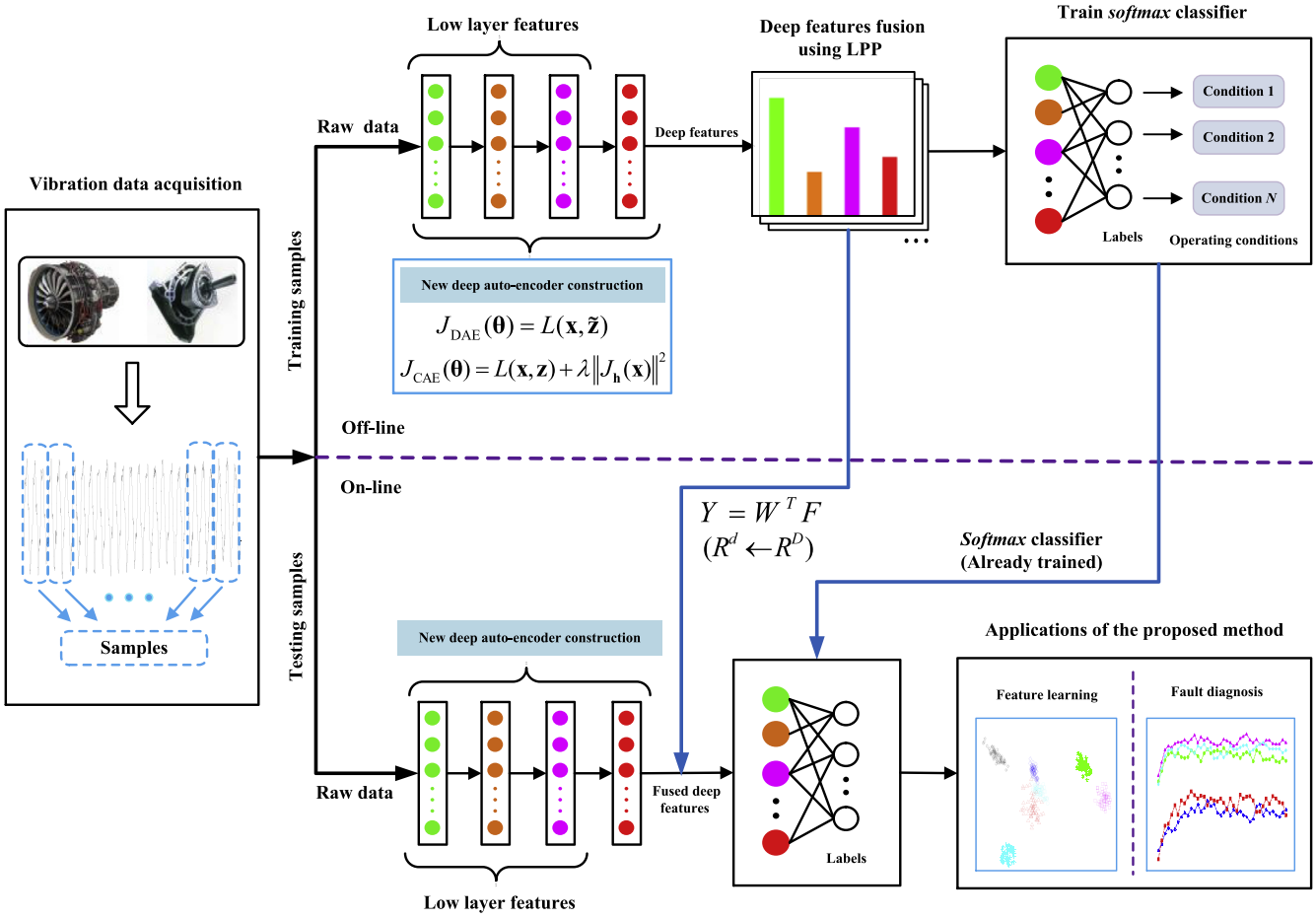

Fig. 3. The framework of the proposed method.

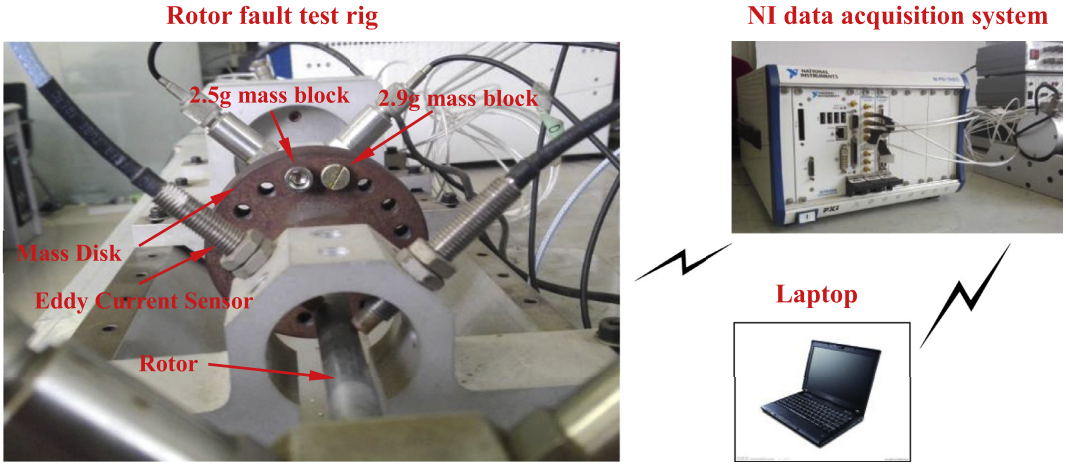

Fig. 4. Rotor fault test rig.

frequency-band signal based on frequency-band energy decomposition. Then, manually extract 19 feature parameters from the most sensitive frequency-band signal, which are used as the input features.

Two key points need to be pointed out. (1) This case study focuses on Experiment 1, i.e., rotor fault diagnosis without any signal preprocessing or manual feature extraction. (2) In each Experiment, we perform two kinds of diagnosis patterns, i.e., fault diagnosis before feature fusion and after feature fusion.

4.2. Diagnosis results and analysis

In this case study, we have to solve the seven-class classification problem, including not only fault categories but also fault severities. Ten trials are carried out for rotor fault diagnosis in each experiment. Tables 3–5 are the average testing accuracies in Experiments 1–3, respectively, and Figs. 6–8 are the detailed diagnosis results of all methods in each trial. From Table 3 we can find that the average testing accuracy of the proposed method is 90.29% before feature fusion, which is slightly higher than standard deep DAE and

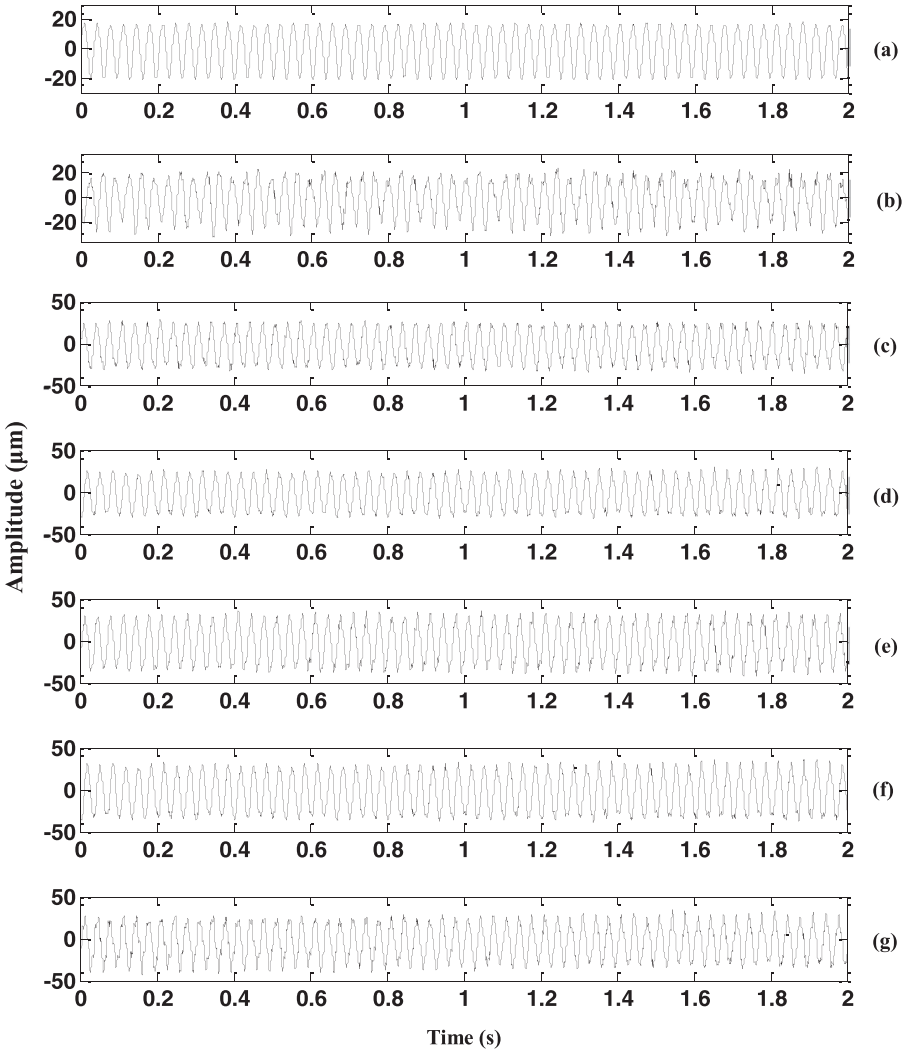

Fig. 5. Vibration signals of the seven rotor operating conditions: (a) normal condition; (b) rub fault condition; (c) unbalance fault condition (2 *2.5 g); (d) unbalance fault condition (2 *2.5 g + 1*2.9 g); (e) unbalance fault condition (2 *2.5 g + 2*2.9 g); (f) unbalance fault condition (2 *2.5 g + 3*2.9 g) and (g) compound faults condition (rub and unbalance faults).

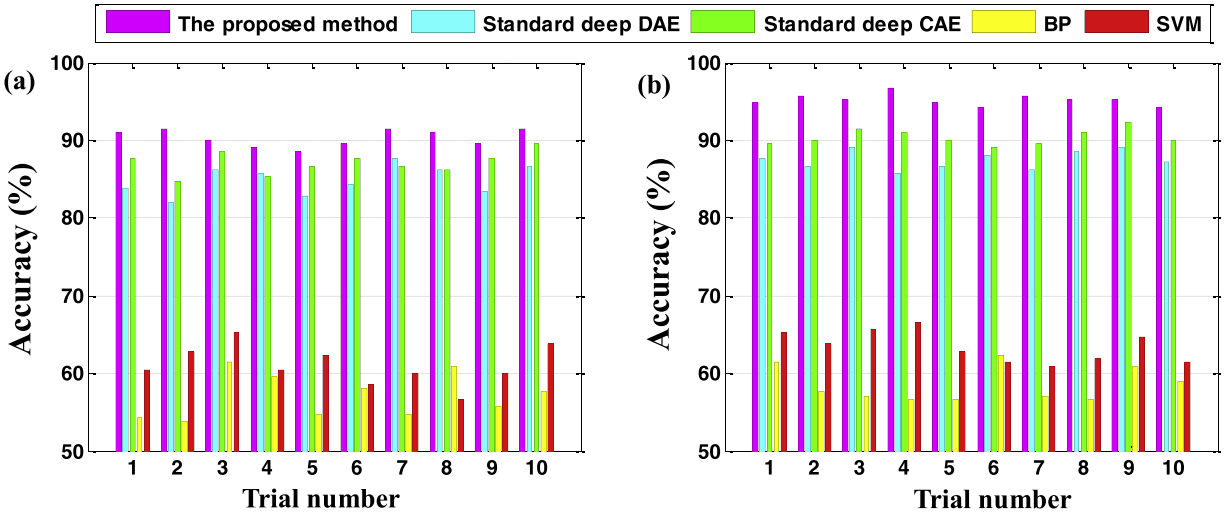

Fig. 6. Detailed diagnosis results of the 10 trials in Experiment 1. (a) Before feature fusion and (b) after feature fusion.

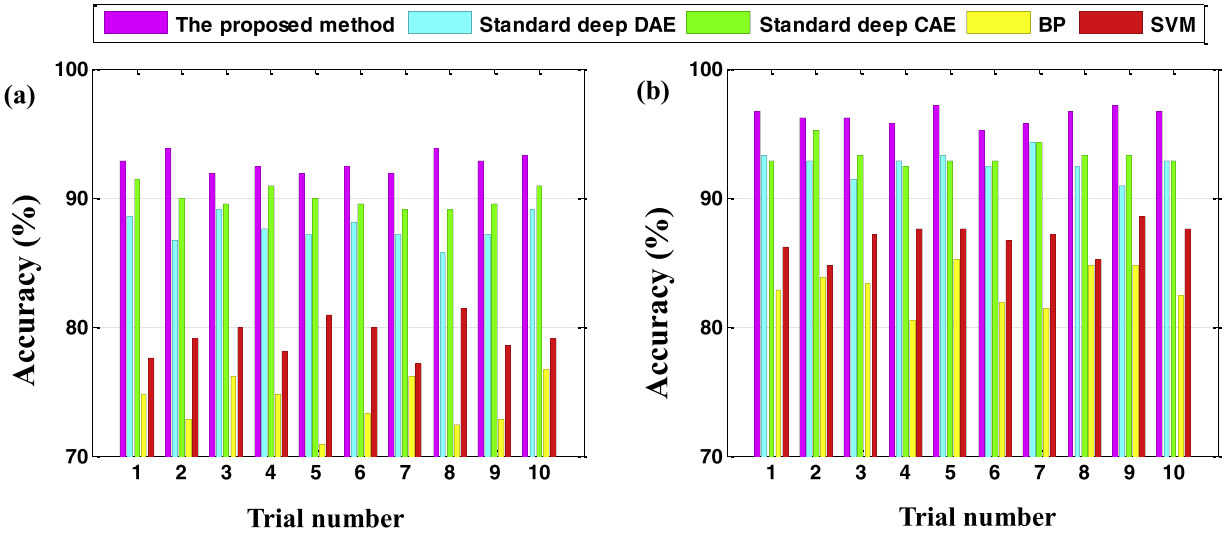

Fig. 7. Detailed diagnosis results of the 10 trials in Experiment 2. (a) Before feature fusion and (b) after feature fusion.

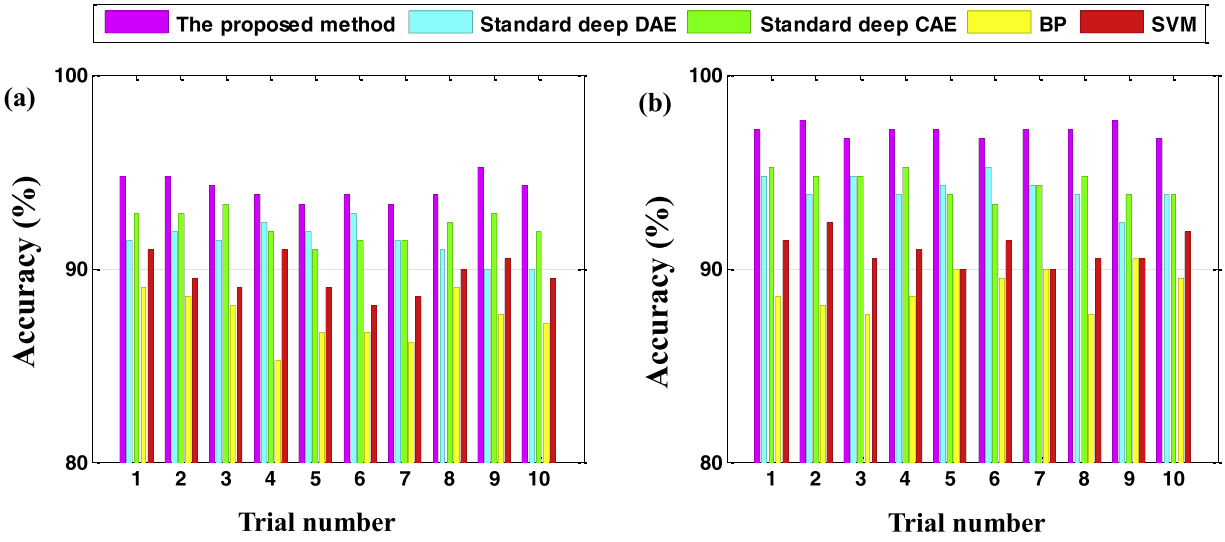

Fig. 8. Detailed diagnosis results of the 10 trials in Experiment 3. (a) Before feature fusion and (b) after feature fusion.

Table 3
Average testing accuracy comparison of the five methods in Experiment 1.

Methods	Average testing accuracy (%)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	90.29 (3792/4200)	95.19 (3998/4200)
Standard deep DAE	84.86 (3564/4200)	87.48 (3674/4200)
Standard deep CAE	87.05 (3656/4200)	90.38 (3796/4200)
BP	57.10 (2398/4200)	58.57 (2460/4200)
SVM	61.05 (2564/4200)	63.48 (2666/4200)

Table 4
Average testing accuracy comparison of the five methods in Experiment 2.

Methods	Average testing accuracy (%)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	92.71 (3894/4200)	96.33 (4046/4200)
Standard deep DAE	87.62 (3680/4200)	92.67 (3892/4200)
Standard deep CAE	89.52 (3760/4200)	93.33 (3920/4200)
BP	74.10 (3112/4200)	83.10 (3490/4200)
SVM	79.19 (3326/4200)	86.86 (3648/4200)

CAE, and much higher than BP and SVM. After feature fusion, the average testing accuracy of the proposed method is 95.19%, compared with other four methods, which are 87.48%, 90.38%, 58.57% and 63.48%, respectively. The similar phenomena are also shown in Tables 4 and 5. From Figs. 6–8 we can find that the accuracies of the proposed method are always higher than other methods in each trial.

Figs. 9 and 10 are the multi-class confusion matrixes of the proposed method for the second trial in Experiment 1 (Accurate to two decimal places). The former is the diagnosis result before feature fusion, and the latter is after feature fusion. The multi-class confusion matrix recodes the classification results of all the conditions in detail, which contains classification accuracy and misclassification error. The ordinate axis of the confusion matrix refers to actual label of classification, and the horizontal axis predict label of classification. The color bar in right illustrates the correspondence between colors and numbers from 0 and 1. From Figs. 9 and 10, we can find that mostly there is confusion between conditions

Table 5
Average testing accuracy comparison of the five methods in Experiment 3.

Methods	Average testing accuracy (%)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	94.14 (3954/4200)	97.10 (4078/4200)
Standard deep DAE	91.43 (3840/4200)	94.10 (3952/4200)
Standard deep CAE	92.19 (3872/4200)	94.38 (3964/4200)
BP	87.43 (3672/4200)	89.00 (3738/4200)
SVM	89.62 (3764/4200)	90.95 (3820/4200)

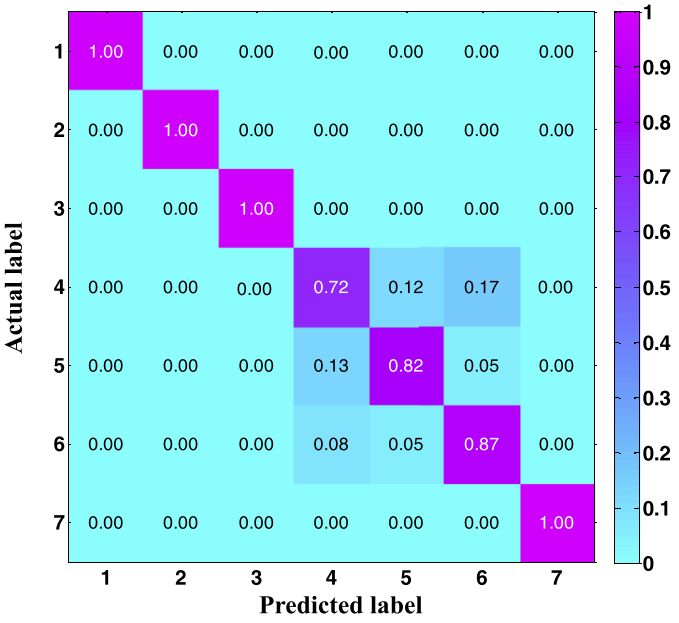

Fig. 9. Multi-class confusion matrix of the proposed method for the second trial before feature fusion in Experiment 1. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

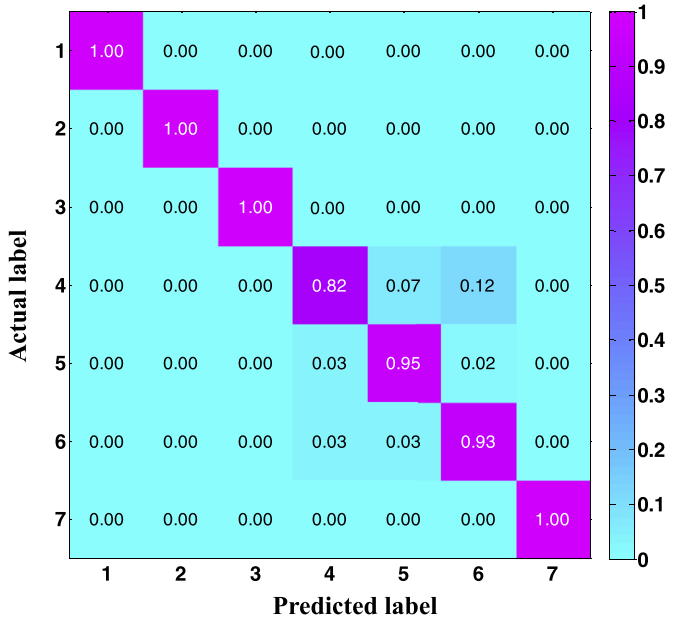

Fig. 10. Multi-class confusion matrix of the proposed method for the second trial after feature fusion in Experiment 1. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

Table 6
Average computing time comparison of the five methods in Experiment 1.

Methods	Average computing time (s)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	113.27	91.33
Standard deep DAE	108.45	87.51
Standard deep CAE	116.08	94.14
BP	30.70	14.82
SVM	17.51	5.93

4–6. The reason is that conditions 4–6 are all unbalance faults and their signal characteristics are very similar.

We can conclude that: (1) The diagnosis performance of the shallow learning models such as SVM and BP depends heavily on the manual feature extraction and feature selection. (2) Deep learning methods such as various types of deep auto-encoders show much higher diagnosis accuracy than shallow learning models. The reason is that deep learning methods can adaptively learn the valuable information from the input data through layer-by-layer feature transformation [17]. (3) Compared with the standard deep DAE and CAE, the proposed method has obvious superiority for recognizing the different fault categories and severities of rotor. The main reason is that the combined advantages of DAE and CAE further enhance the feature learning ability of the proposed method. (4) LPP feature fusion method can effectively remove the redundant information to further improve the quality of the deep features and diagnosis efficiency [30].

Table 6 lists the average computing time of all methods in Experiment 1. It can be found from Table 6 that the average computing time of the proposed method is 113.27 s before feature fusion, compared with other methods, which are 108.45 s, 116.08 s, 30.70 s and 17.51 s, respectively. After feature fusion, the average computing time of the proposed method is 91.33 s. The other four methods are 87.51 s, 94.14 s, 14.82 s and 5.93 s, respectively.

By comparison, we can conclude that: (1) The proposed method suffers from an obvious shortcoming: deep architecture needs more training time than shallow learning models due to the increase of hidden layers [34]. (2) LPP can reduce the dimension of the deep features and improve the computational efficiency. Though the computational complexity is the main shortcoming of deep learning methods, most trials can be finished in less than 2 min (Core i3, 4 GB memory). With the rapid development of training algorithm and hardware technology, we can build deep learning models more efficiently in future [3].

Figs. 11 and 12 show the standard deviations of different methods in Experiments 1–3. From Fig. 11 we can observe that the standard deviation of the proposed method is 1.081 in Experiment 1 before feature fusion, and it is smaller than other methods, which are 1.875, 1.451, 2.835 and 2.548, respectively. In Experiments 2 and 3, the standard deviations of the proposed method are still smaller than other methods. The similar phenomena also happen in Fig. 12. Therefore, the proposed method can stably distinguish the different rotor operating conditions.

In Experiment 1, the main parameters of different methods are listed in Table 7. The structure selection of deep learning model is a great challenge, and there is not a mature method in theory to select the optimal structure of deep learning models at present [35]. In this paper, when determining the deep architecture, we follow a simple idea similar to [25]. We investigate how the proposed deep model behaves as we increase the capacity both in depth (number of hidden layers) and in breadth (number of units per hidden layer). Fig. 13 shows the evolution of diagnosis accuracy as we increase the number of hidden layers (from 1 to 5) and

Table 7
Parameter description of the five methods in Experiment 1.

Methods	Parameter description
The proposed method (Deep learning)	The architecture is 1000–300–300–300–300 (one DAE and three CAEs). Learning rate is 0.1, momentum is 0.9, pre-training iteration number is 100, noise level $\sigma=0.5$, and regularization coefficient $\lambda=0.7$. In the LPP algorithm, $k=12$ and $d=28$.
Standard deep DAE (Deep learning)	The architecture is 1000–300–300–300–300 (four DAEs). Learning rate is 0.1, momentum is 0.9, pre-training iteration number is 100, and noise level $\sigma=0.5$. In the LPP algorithm, $k=12$ and $d=21$.
Standard deep CAE (Deep learning)	The architecture is 1000–300–300–300–300 (four CAEs). Learning rate is 0.1, momentum is 0.9, pre-training iteration number is 100, and regularization coefficient $\lambda=0.7$. In the LPP algorithm, $k=12$ and $d=33$.
BP (Shallow learning)	The architecture is 1000–1500–7. Learning rate is 0.05, momentum is 0.95, and iteration number is 800. In the LPP algorithm, $k=12$ and $d=16$.
SVM (Shallow learning)	RBF kernel is applied. Penalty factor is 20, and kernel radius is 0.7017. In the LPP algorithm, $k=12$ and $d=11$.

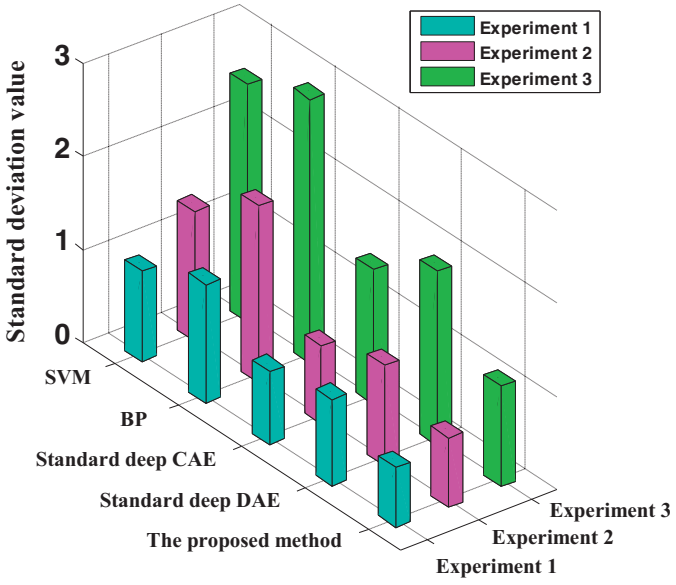

Fig. 11. The standard deviations of ten trials before feature fusion.

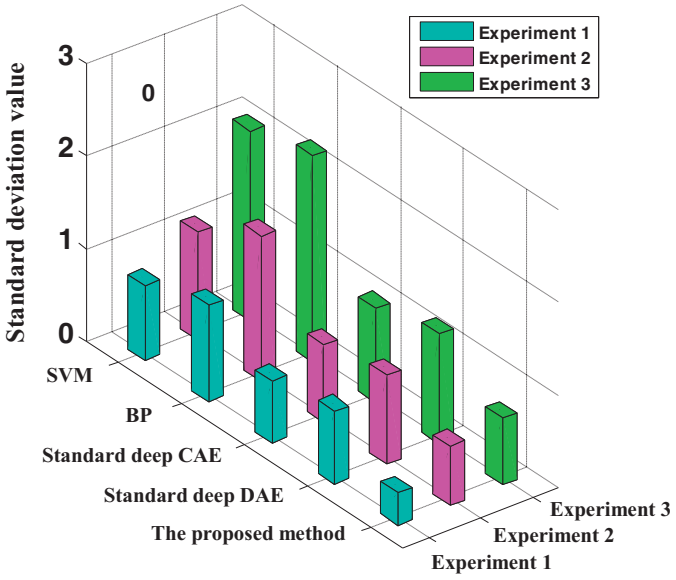

Fig. 12. The standard deviations of ten trials after feature fusion.

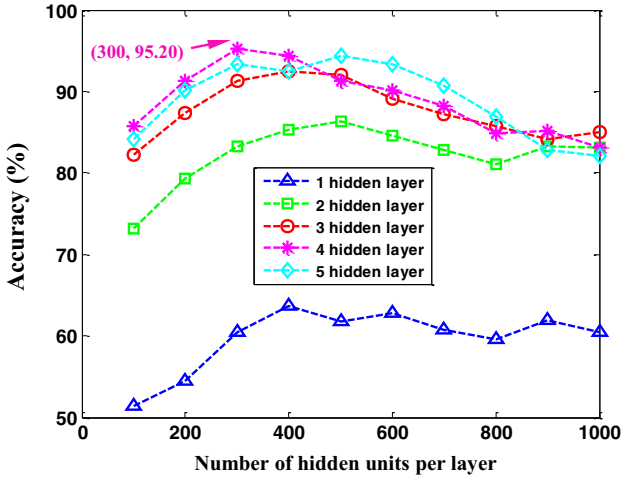

Fig. 13. The relationship between accuracy and the proposed deep architecture in Experiment 1.

the number of units per hidden layer (from 100 to 1000). It can be found that appropriate hidden layers and units contribute to feature learning and fault diagnosis. However, excessive hidden layers and units may not be a good choice. Thus, in this case study, the architecture of the proposed method is 1000–300–300–300–300. In other words, it has 4 hidden layers, and each hidden layer contains 300 units. The unit number of the input layer is decided by the dimension of the samples, and the unit number of the output layer is determined by the number of rotor operating conditions.

In the proposed new deep auto-encoder, there are other two important parameters which need to be determined. That is, noise level parameter σ in DAE and regularization coefficient λ in CAE. In this paper, we adopt cross-validation to select the optimal parameters. The candidate sets of σ and λ are both set as [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. Fig. 14 shows the evolution of diagnosis accuracy as we increase parameter values of σ and λ. It can be found that the accuracy is mainly influenced by the noise level parameter σ, and a reasonable noise level is very important. The main reason is that low noise level may not fully motivate the robustness of reconstruction [25], and high noise level will probably leads to serious corruption, which is not beneficial to the enhancement of feature learning. When determining the parameters of SVM and BP, a lot of skills are applied. RBF kernel, the most popular kernel function is applied, and the parameters of SVM are determined through a 10-fold cross validation. The structure of BP

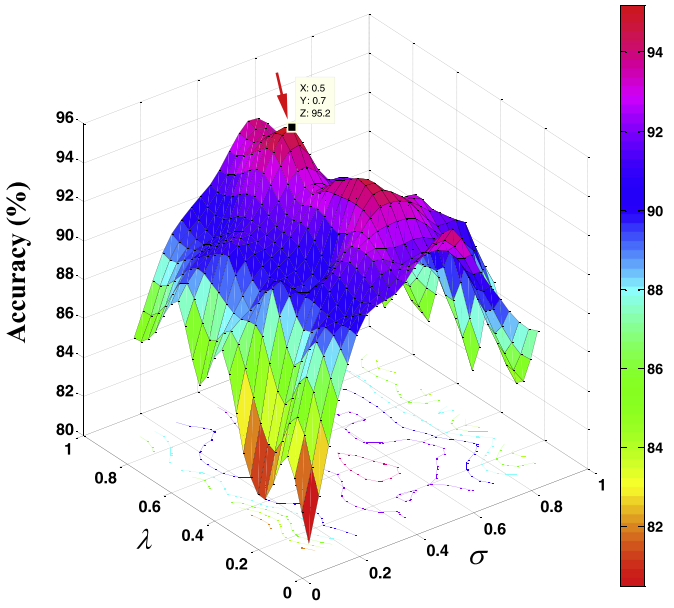

Fig. 14. The relationship between accuracy and parameters σ and λ in Experiment 1.

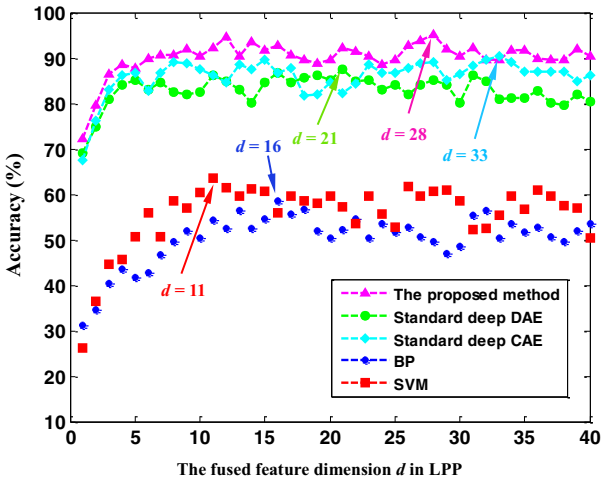

Fig. 15. The relationship between accuracy and fused feature dimension in Experiment 1.

is decided by the guiding principles, and learning rate, momentum and iteration number are set by experiences.

In the LPP algorithm, the fused feature dimension d and the number of the nearest neighbors k should be predetermined. In most cases, k is set to 12 by experience. Fig. 15 shows the evolution of diagnosis accuracy as we increase the fused feature dimension d (from 1 to 40). It can be seen that the best choice for d of the proposed method is 28. The fused feature dimensions of other four methods are set to 21, 33, 16 and 11, respectively.

In Experiment 2, the main parameters of different methods are described as follows. (1) The proposed method: The architecture is 152–100–100–100 (one DAE and two CAEs), which is determined by experimentation. In the LPP algorithm, $k=12$ and $d=11$. The other parameters are the same as Experiment 1. (2) Standard deep DAE: The architecture is 152–100–100–100 (three DAEs). In the LPP algorithm, $k=12$ and $d=16$. The other parameters are the same as Experiment 1. (3) Standard deep CAE: The architecture is 152–100–100–100 (three CAEs). In the LPP algorithm, $k=12$ and $d=18$. The other parameters are the same as Experiment 1. (4) BP: The architecture is 152–200–7, which is decided by the guiding principles and experiences. In the LPP algorithm, $k=12$ and $d=7$. The other parameters are the same as Experiment 1. (5) SVM: RBF kernel is applied. Two main parameters, the penalty factor and the radius of the kernel function, are set to 20 and 0.081, respectively. Each of them is determined through a 10-fold cross validation. In the LPP algorithm, $k=12$ and $d=6$.

In Experiment 3, the main parameters of different methods are described as follows. (1) The proposed method: The architecture is 19–30–20–10 (one DAE and two CAEs). In the LPP algorithm, $k=12$ and $d=5$. The other parameters are the same as Experiment 1. (2) Standard deep DAE: The architecture is 19–30–20–10 (three DAEs). In the LPP algorithm, $k=12$ and $d=4$. The other parameters are the same as Experiment 1. (3) Standard deep CAE: The architecture is 19–30–20–10 (three CAEs). In the LPP algorithm, $k=12$ and $d=4$. The other parameters are the same as Experiment 1. (4) BP: The architecture is 19-39-7. In the LPP algorithm, $k=12$ and $d=3$. The other parameters are the same as Experiment 1. (5) SVM: RBF kernel is applied. Two main parameters, the penalty factor and the radius of the kernel function, are set to 30 and 0.025, respectively. Each of them is determined through a 10-fold cross validation. In the LPP algorithm, $k=12$ and $d=3$.

4.3. Layer-by-layer feature learning process

In order to demonstrate the powerful ability of the proposed method in automatically learning the valuable features from the raw vibration data, in this case study, the layer-by-layer feature learning process of the proposed method is investigated.

Taking the second trial in Experiment 1 for example, considering that the learned features in each hidden layer are all high-dimensional data (300 dimensions), thus the first two projected vectors are used for visualizing. For comparison, principle component analysis (PCA) feature fusion method is also implemented in this section. As shown in Figs. 16–19, LPP1 and LPP2 represent the first two projected vectors obtained by LPP, PCA1 and PCA2 denote the first two principle components given by PCA, and the annotations 1–7 corresponding to the condition labels listed in Table 1.

By comparing Figs. 16–19, we can conclude that: (1) Higher-layer features can represent the input data in a more precise and identifiable way than lower-layer features do. The main reason is that hidden layers allow the deep architectures to be more powerful in modeling the complex nonlinear relationship hidden in the input data [3]. (2) Conditions 4–6 cannot be completely distinguished using the proposed method. The main reason is that conditions 4–6 belong to the same fault category with different fault severities, thus the fault features of the three conditions are too similar to be completely distinguished. (3) Compared with PCA method, LPP is more effective to fuse deep features for fault classification. The main reason is that PCA aims to discover the global information hidden in the data set, while LPP aims to preserve the local structure. In the field of mechanical fault diagnosis, the local structure existing in the data set is usually more valuable and important [30].

5. Engineering application

5.1. Electrical locomotive bearing experimental setup

Bearing is the most widely used component in the rotating machinery. In this case study, the proposed method is used to diagnose the electrical locomotive bearing faults. Fig. 20 is the electrical locomotive bearing test rig, and four kinds of faulty bearings are shown in Fig. 21. The vibration acceleration signal is collected at a sample frequency of 12.8 kHz. More parameters of the electrical locomotive bearing are listed in Table 8.

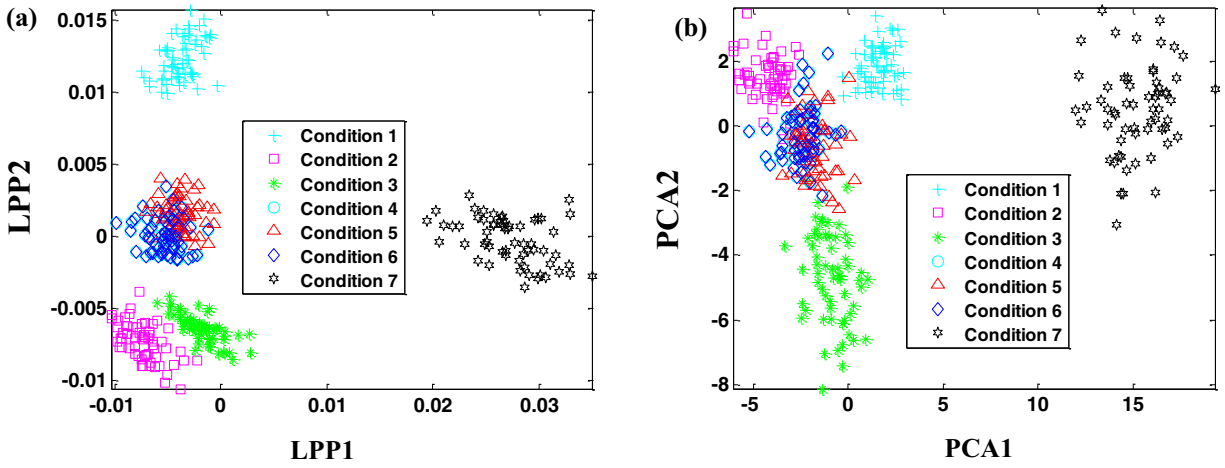

Fig. 16. 2D projection of the learned features in the first hidden layer. (a) LPP and (b) PCA.

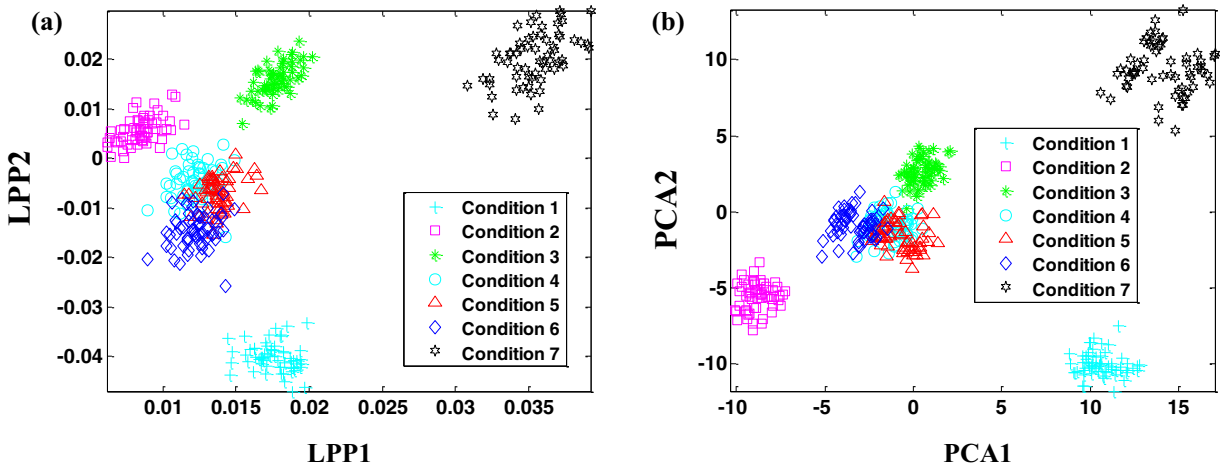

Fig. 17. 2D projection of the learned features in the second hidden layer. (a) LPP and (b) PCA.

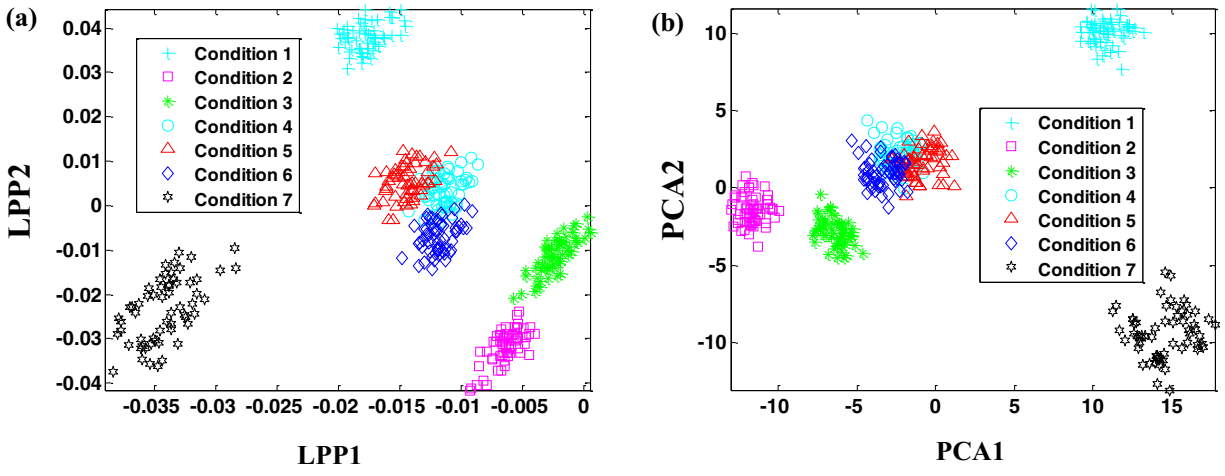

Fig. 18. 2D projection of the learned features in the third hidden layer. (a) LPP and (b) PCA.

In this case study, nine bearing operating conditions are created, listed in Table 9. Each condition consists of 400 samples, the random 300 samples of each condition are used for training and the rest 100 samples for testing. Each sample is a collected vibration signal containing 1600 data points. Fig. 22 shows the raw vibration signals of the nine bearing conditions (first 12,800 data points).

5.2. Diagnosis results and analysis

In this case study, another three experiments are taken into account. Unlike the Section 4, the specific goals of the three experiments are described as follows.

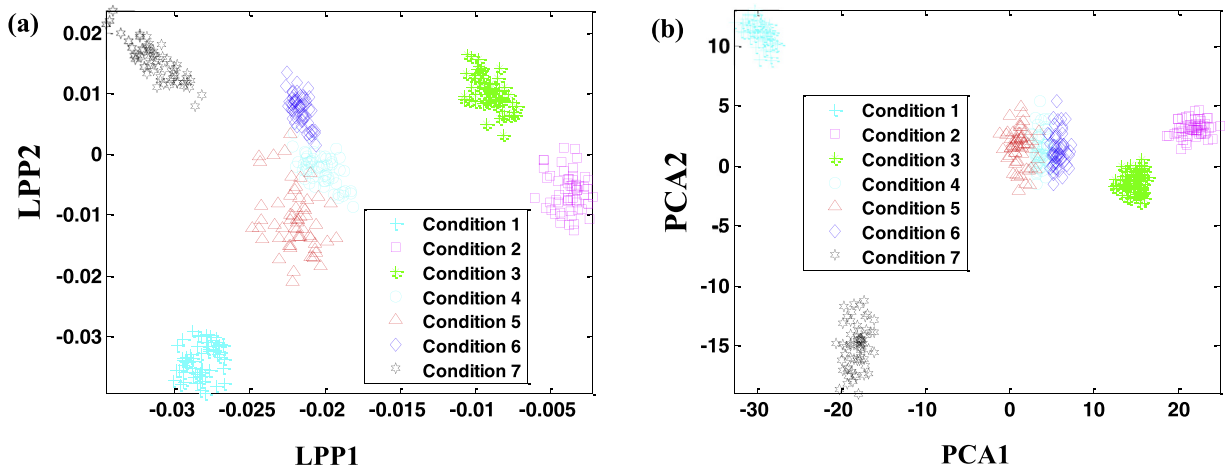

Fig. 19. 2D projection of the learned features in the fourth hidden layer. (a) LPP and (b) PCA.

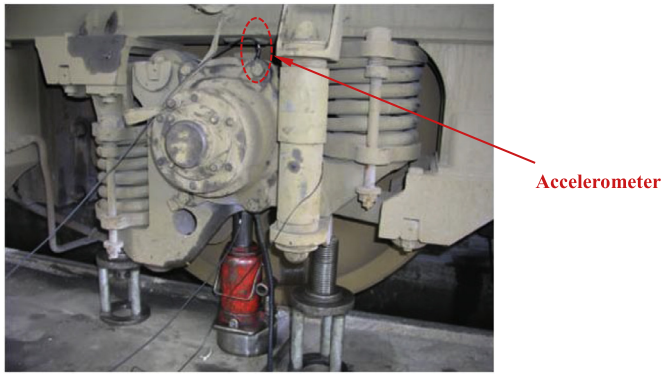

Fig. 20. Electrical locomotive bearing experimental setup.

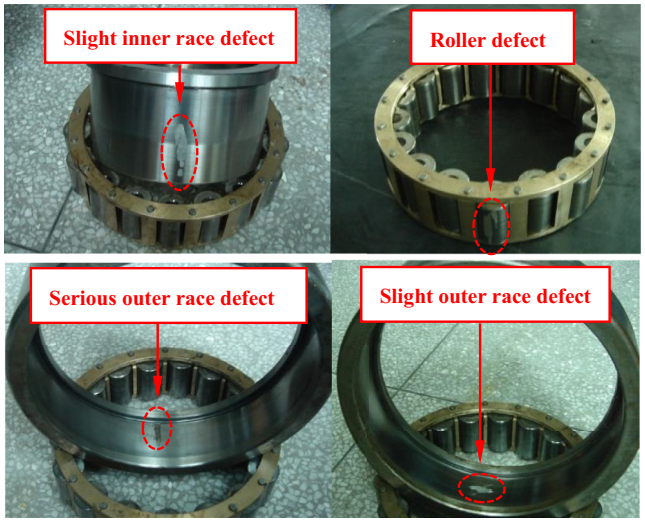

Fig. 21. Faults in the electrical locomotive bearings.

- Experiment 4: Similar to Experiment 1, without any signal preprocessing or manual feature extraction, the proposed method is compared with the standard deep DAE, standard deep CAE, BP neural network and SVM for electrical locomotive bearings fault diagnosis.

Table 8
Parameters of the electrical locomotive bearing.

Parameter	Value
Bearing specs	552732QT
Inner race diameter	160 mm
Outer race diameter	290 mm
Roller diameter	34 mm
Roller number	17
Contact angle	0°

Table 9
Description of the electrical locomotive bearing operation conditions.

Electrical locomotive bearing operation conditions	Motor speed (rpm)	Size of training/testing samples	Label
Normal condition	490	300/100	1
Slight outer race defect	490	300/100	2
Serious outer race defect	481	300/100	3
Inner race defect	498	300/100	4
Roller defect	531	300/100	5
Compound faults (outer and inner races)	525	300/100	6
Compound faults (outer race and roller)	521	300/100	7
Compound faults (inner race and roller)	640	300/100	8
Compound faults (outer and inner races and roller)	549	300/100	9

- Experiment 5: The diagnosis performance of the proposed method in dealing with the unbalanced training dataset is investigated.
- Experiment 6: The diagnosis performance of the proposed method and other deep learning methods (standard DBN and CNN) are compared based on different data sets.

It is worth noting that this case study focuses on Experiment 4, i.e., bearing fault diagnosis without any signal preprocessing or manual feature extraction, which includes single fault and compound faults.

Ten trials are employed for classifying bearing operation conditions in Experiment 4. The average testing accuracies and standard deviations are calculated in Table 10, and Table 11 shows the average computing time of all methods (Core i3, 4 GB memory). Fig. 23 is the detailed classification result in each trial before and after feature fusion, respectively. It can be found that the average testing accuracy of the proposed method is higher than other methods,

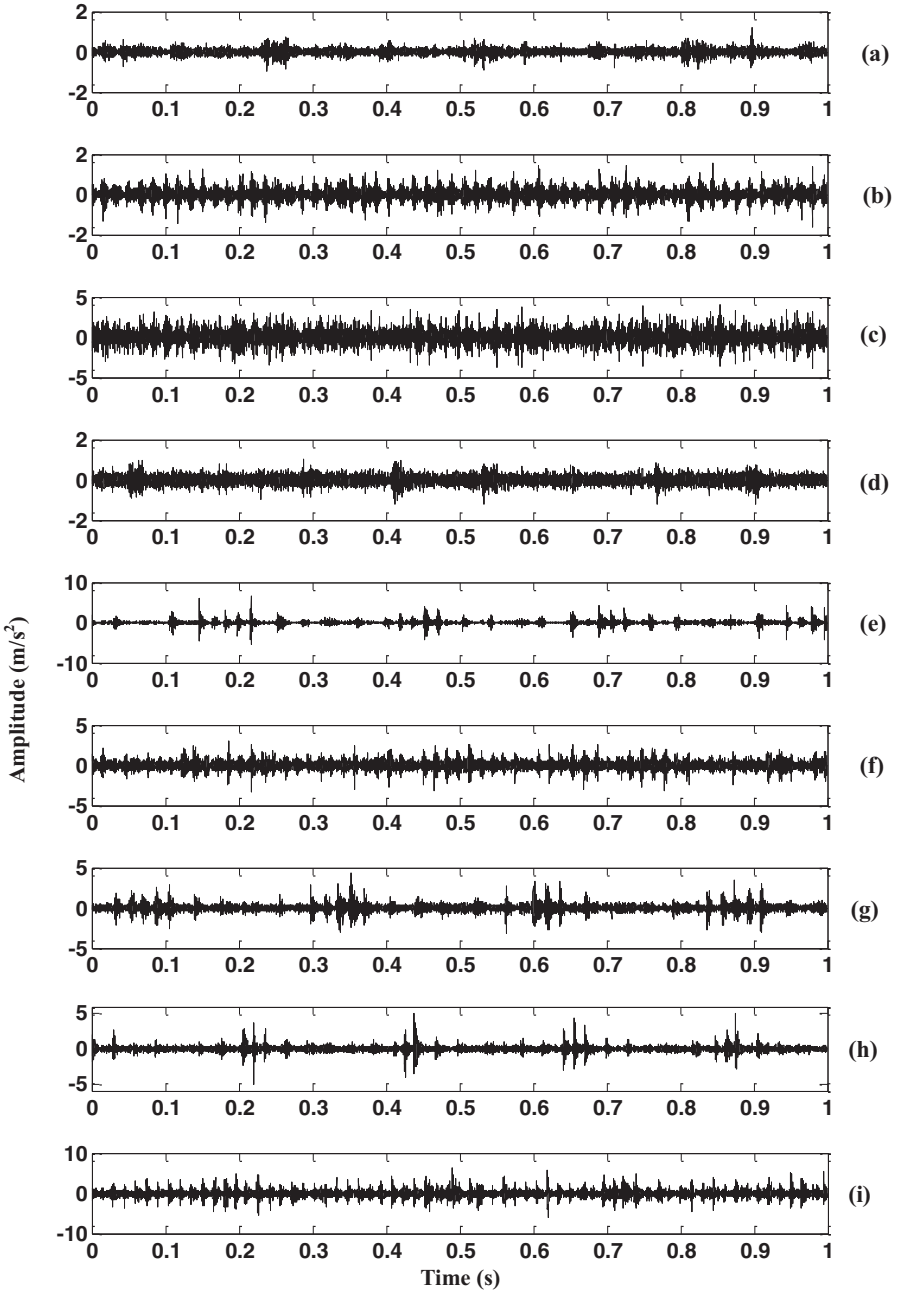

Fig. 22. Vibration signals of the nine electrical locomotive bearing conditions. (a) Normal condition, (b) slight outer race defect, (c) serious outer race defect, (d) inner race defect, (e) roller defect, (f) compound faults (outer and inner races), (g) compound faults (outer race and roller), (h) compound faults (inner race and roller) and (i) compound faults (outer and inner races and roller).

Table 10
Diagnosis results of electrical locomotive bearing in Experiment 4.

Methods	Average testing accuracy (%)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	87.90 ± 1.10 (7911/9000)	91.90 ± 0.74 (8271/9000)
Standard deep DAE	83.10 ± 2.03 (7479/9000)	84.60 ± 1.58 (7614/9000)
Standard deep CAE	83.30 ± 1.83 (7497/9000)	85.10 ± 1.73 (7659/9000)
BP	47.70 ± 6.18 (4293/9000)	49.70 ± 5.58 (4473/9000)
SVM	56.90 ± 4.23 (5121/9000)	57.60 ± 3.60 (5184/9000)

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

Table 11
Average computing time comparison for electrical locomotive bearing fault diagnosis in Experiment 4.

Methods	Average computing time (s)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	255.66	162.43
Standard deep DAE	241.62	151.76
Standard deep CAE	266.23	183.50
BP	87.29	34.95
SVM	38.17	13.84

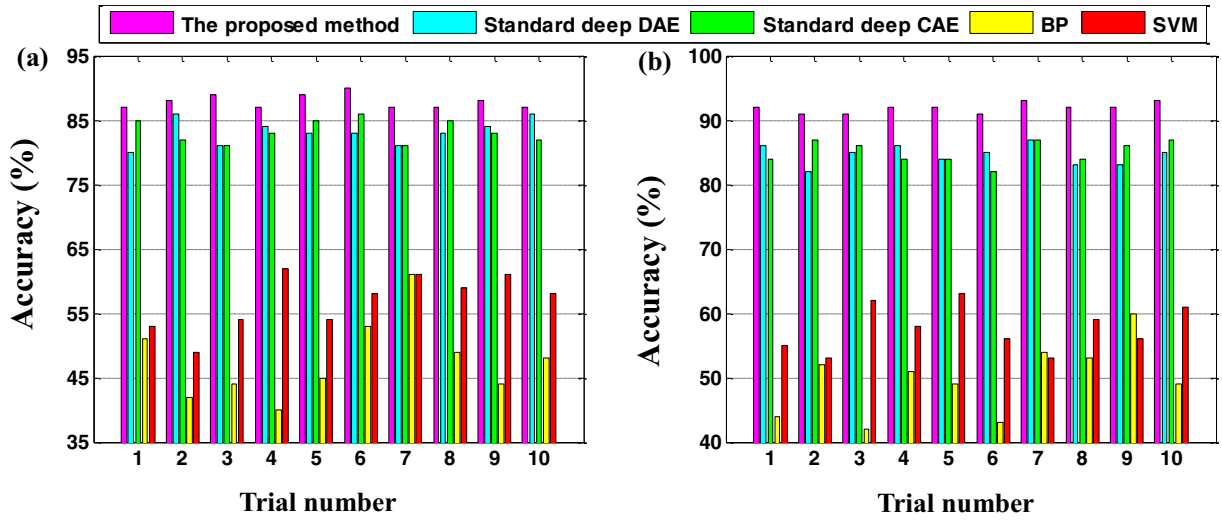

Fig. 23. Detailed diagnosis results of the 10 trials. (a) Before feature fusion and (b) after feature fusion.

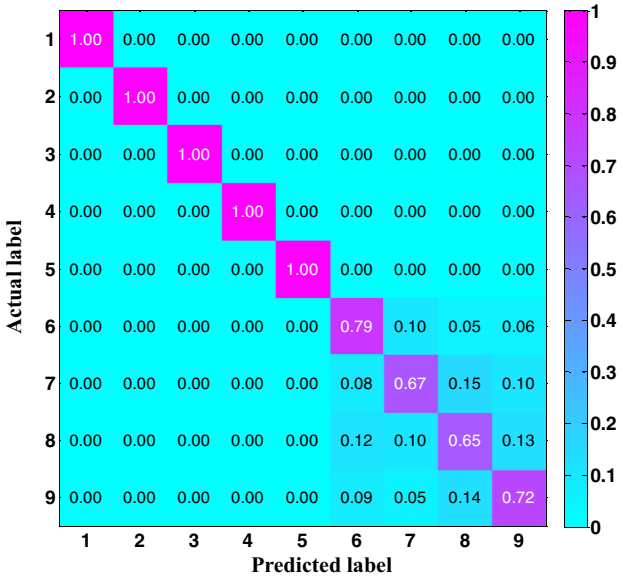

Fig. 24. Multi-class confusion matrix of the proposed method for the seventh trial before feature fusion in Experiment 4.

Table 12
Parameter description of the five methods for bearing fault diagnosis in Experiment 4.

Methods	Parameter description
The proposed method (Deep learning)	The architecture is 1600–400–400–400 (one DAE and two CAEs). Learning rate is 0.1, momentum is 0.9, pre-training iteration number is 100, noise level $\sigma = 0.4$, and regularization coefficient $\lambda = 0.5$. In the LPP algorithm, $k = 12$ and $d = 18$.
Standard deep DAE (Deep learning)	The architecture is 1600–400–400–400 (three DAEs). Learning rate is 0.1, momentum is 0.9, pre-training iteration number is 100, and noise level $\sigma = 0.4$. In the LPP algorithm, $k = 12$ and $d = 25$.
Standard deep CAE (Deep learning)	The architecture is 1600–400–400–400 (three CAEs). Learning rate is 0.1, momentum is 0.9, pre-training iteration number is 100, and regularization coefficient $\lambda = 0.5$. In the LPP algorithm, $k = 12$ and $d = 34$.
BP (Shallow learning)	The architecture is 1600–1250–9. Learning rate is 0.05, momentum is 0.95, and iteration number is 1000. In the LPP algorithm, $k = 12$ and $d = 7$.
SVM (Shallow learning)	RBF kernel is applied. Penalty factor is 30, and kernel radius is 0.092. In the LPP algorithm, $k = 12$ and $d = 9$.

and the standard deviation is lower compared with other methods. Figs. 24 and 25 are the multi-class confusion matrixes of the proposed method for the seventh trial in Experiment 4 (Accurate to two decimal places). From Figs. 24 and 25, we can find that mostly there is confusion between conditions 6–9. The reason is that conditions 6–9 are all compound faults and their signal characteristics are very similar. In addition, the feasibility of LPP is also demonstrated by the testing results. Therefore, the proposed method can accurately and stably diagnose not only single fault but also compound faults of electrical locomotive bearings.

In Experiment 4, the related parameters of all the methods are described in Table 12. The basic idea of parameter selection is similar to Experiment 1 in Section 4. Fig. 26 shows the evolution of accuracy as we increase the number of hidden layers (from 1 to 5) and the number of units per hidden layer (from 100 to 1000). It can be found that the optimal architecture of the proposed method is selected as 1600–400–400–400. Fig. 27 shows the evolution of accuracy as we increase parameter values of σ and λ (candidate set [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]). Fig. 28 shows the evolution of accuracy as we increase the fused feature dimension d (from 1 to 40).

By comparing and analyzing Sections 4 and 5, we can find that the optimal structures and parameters of deep learning models depend on the specific diagnosis task [3]. The changeable number of hidden layers and units provide the designer a lot of freedom.

5.3. Layer-by-layer feature learning process

Taking the seventh trial in Experiment 4 for example, Figs. 29 and 30 show the 2D-LPP and 3D-LPP projections of the features in the input layer, first hidden layer, second hidden layer and third hidden layer learned by the proposed deep method. The annotations 1–9 corresponding to the bearing condition are listed in Table 9. The results further demonstrate the great ability of the proposed

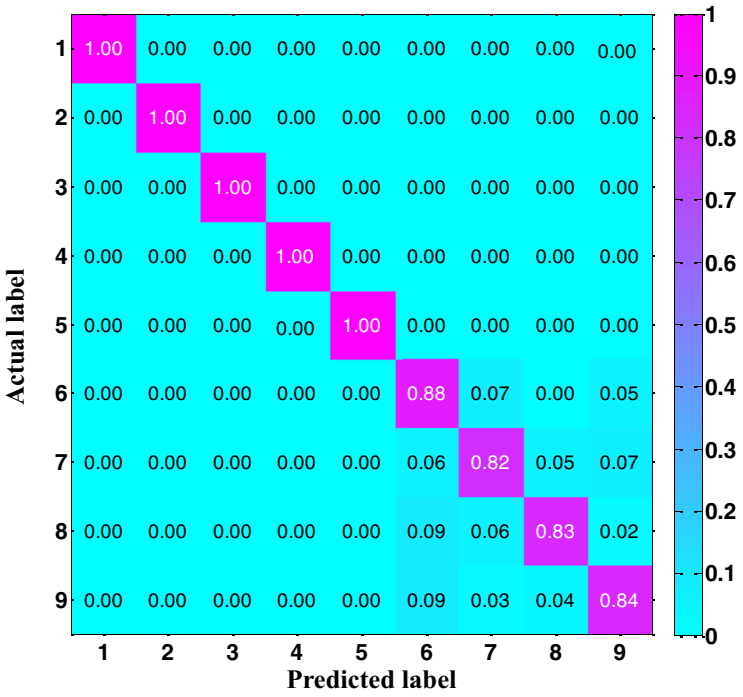

Fig. 25. Multi-class confusion matrix of the proposed method for the seventh trial after feature fusion in Experiment 4.

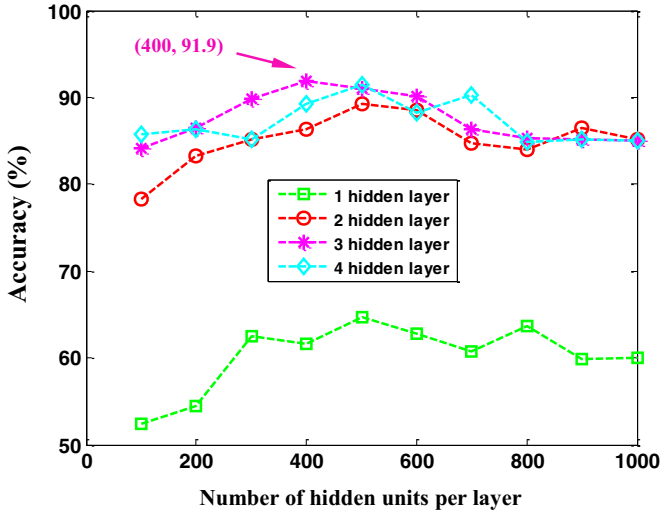

Fig. 26. The relationship between accuracy and the proposed deep architecture.

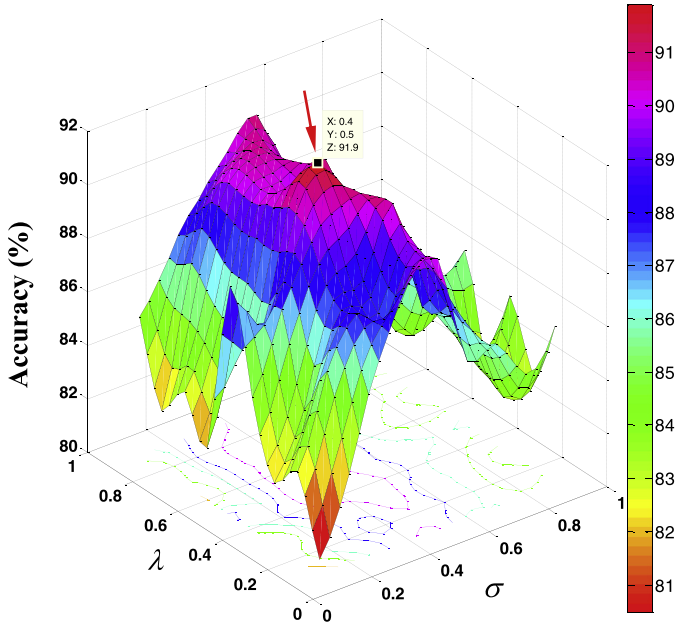

Fig. 27. The relationship between accuracy and parameters σ and λ.

method in automatically learning the useful features from the raw vibration data.

5.4. Influence of the unbalanced training dataset

In Section 4, we just consider the fault diagnosis with balanced training dataset. However, in most cases, the percentage of normal samples is much larger than fault samples in mechanical fault diagnosis area. Therefore, it is very practical and meaningful to investigate the diagnosis performance of the proposed method in dealing with the unbalanced training dataset. As shown in Table 13, ten unbalanced cases are employed in Experiment 5 to compare the performance of different diagnosis methods. It should be pointed out that Case 1 is the same as Experiment 4, i.e., each bearing condition (normal or fault) consists of 300 training samples and 100 testing samples. From Case 1 to Case 10, the size of normal sample in training set is always set to 300, while, the size of each kind of fault sample in training set decreases from 300 to 30. In order words, the normal percentage in the training set gradually increases from Case 1 to Case 10.

The parameters set in each cases of Experiment 5 are the same as Experiment 4, and ten trials are considered for classifying bearing operation conditions. Fig. 31 shows the average diagnosis accuracy as we increase the normal percentage (from 11.11% to 55.56%) before feature fusion in Experiment 5. It can be found that: (1) Highly unbalanced training dataset trends to results low diagnosis performance of different methods. The main reason is that most classifiers are sensitive to training sample distribution, i.e., they

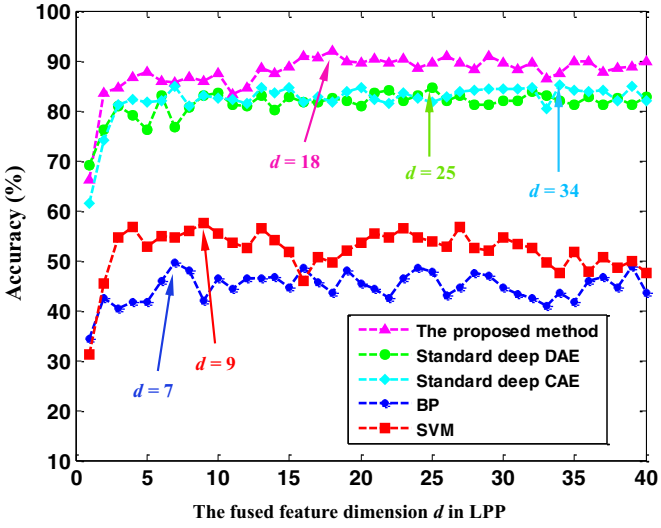

Fig. 28. The relationship between accuracy and fused feature dimension.

show higher accuracy rate in larger samples than fewer samples. (2) Compared with other methods, the proposed method shows better stability and generalization performance in dealing with the unbalanced training dataset. Therefore, effective techniques are still needed to further improve the performance of the proposed method in dealing with the unbalanced training dataset.

Table 13
Description of the unbalanced training set in Experiment 5.

Unbalanced cases	Size of normal sample		Size of each kind of fault sample		Normal percentage in the training set (%)
	Training set	Testing set	Training set	Testing set	
Case 1	300	100	300	100	11.11 (300/2700)
Case 2	300	100	270	100	12.20 (300/2460)
Case 3	300	100	240	100	13.51 (300/2220)
Case 4	300	100	210	100	15.15 (300/1980)
Case 5	300	100	180	100	17.24 (300/1740)
Case 6	300	100	150	100	20.00 (300/1500)
Case 7	300	100	120	100	23.81 (300/1260)
Case 8	300	100	90	100	29.41 (300/1020)
Case 9	300	100	60	100	38.46 (300/780)
Case 10	300	100	30	100	55.56 (300/540)

5.5. Diagnosis performance comparison of different deep learning models

In order to further verify the advantage of the proposed method, the diagnosis performance of the proposed model and other standard deep learning models is compared in this section. DBN is a generative graphical model, which is composed by multiple restricted Boltzmann machines (RBMs). CNN is a type of multilayer feed-forward neural network, which is constructed with convolutional layers and subsampling layers. Compared with standard DBN, CNN is more suitable for time series signal modeling and analysis through the convolution operations. In addition, CNN

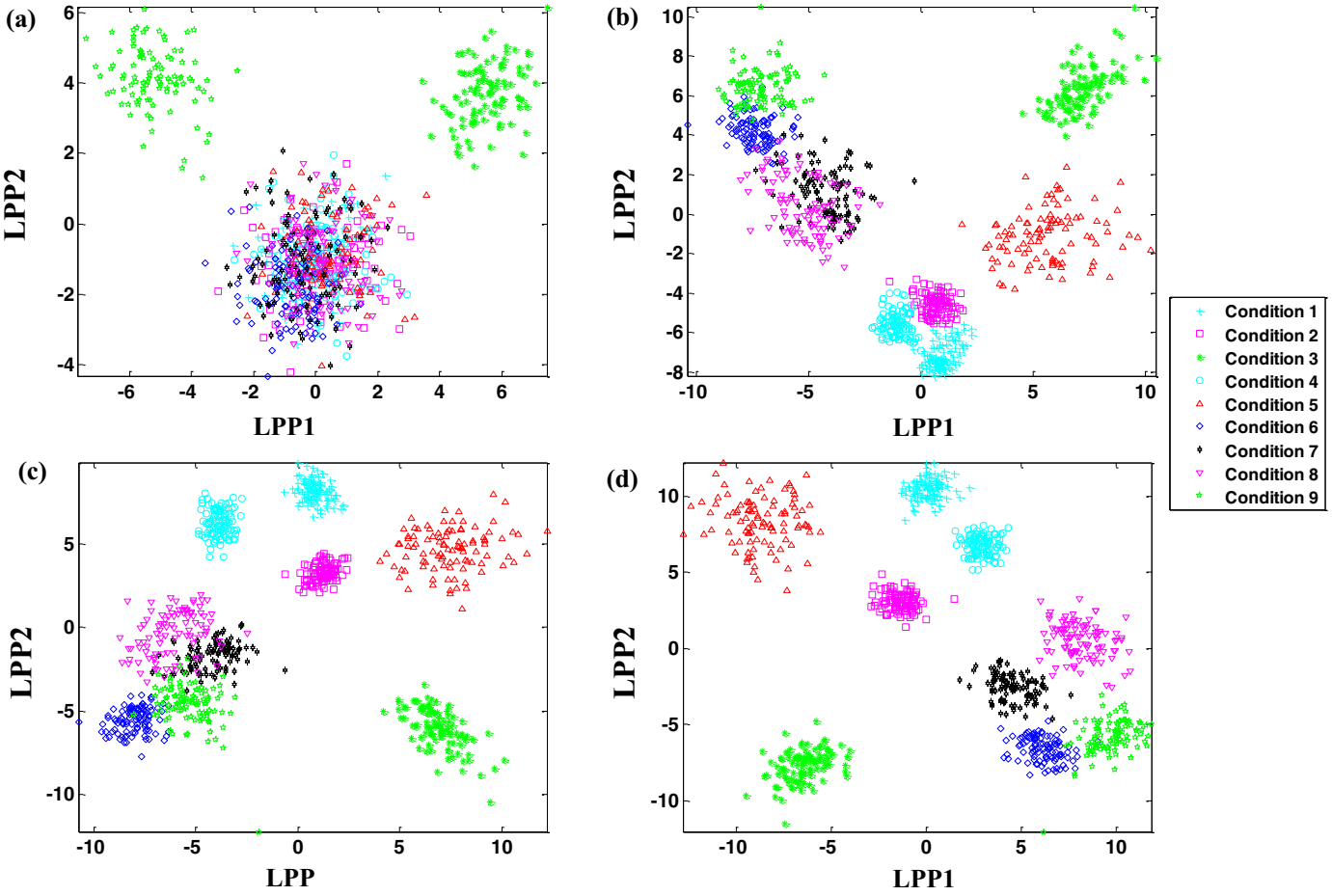

Fig. 29. 2D-LPP projection of the learned deep features. (a) Raw data features, (b) features in the first hidden layer, (c) features in the second hidden layer and (d) features in the third hidden layer.

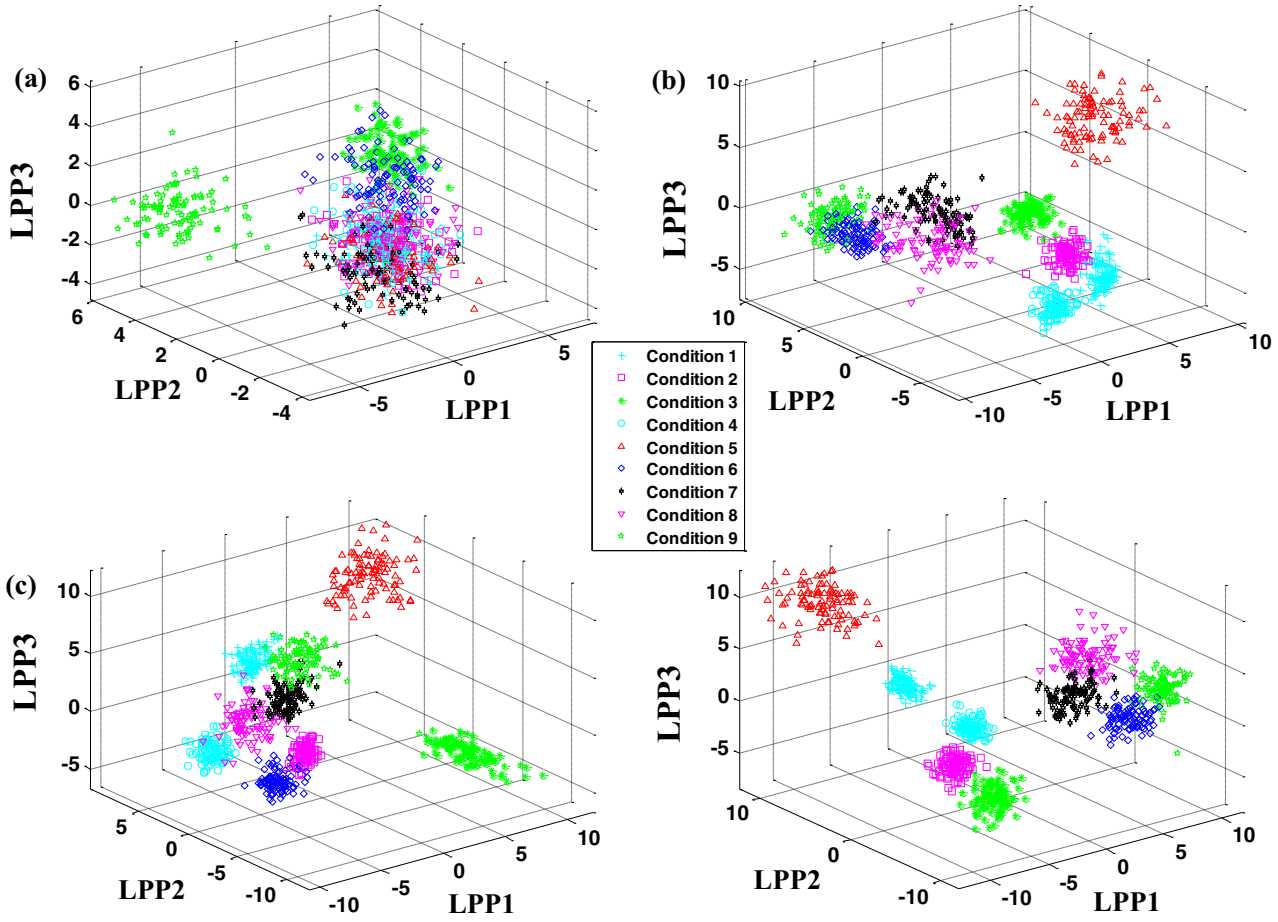

Fig. 30. 3D-LPP projection of the learned deep features. (a) Raw data features, (b) features in the first hidden layer, (c) features in the second hidden layer and (d) features in the third hidden layer.

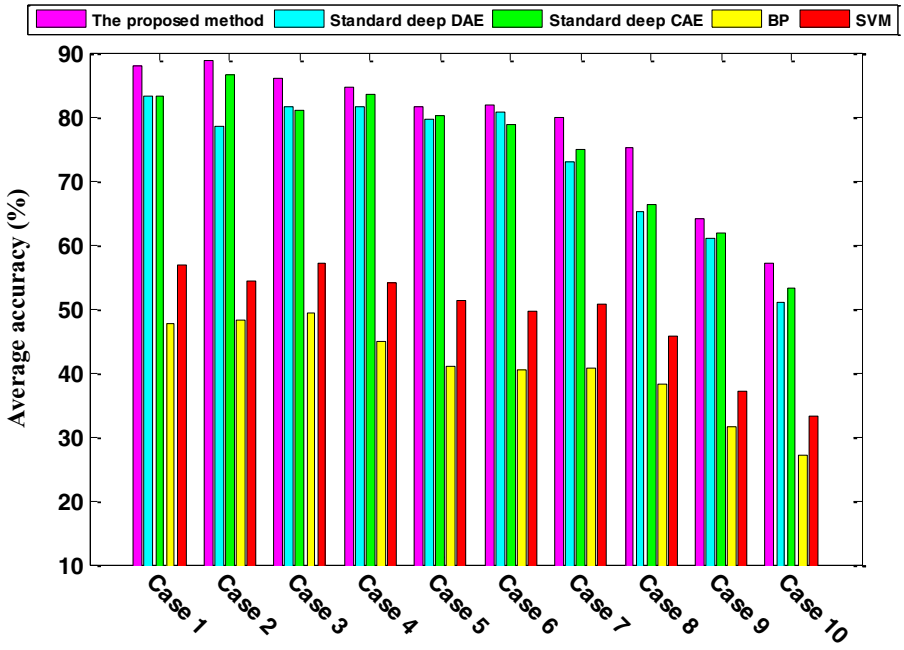

Fig. 31. Influence of the unbalanced training dataset in Experiment 5.

Table 14
Diagnosis results comparison of different deep learning models for electrical locomotive bearing fault diagnosis.

Deep learning methods	Average testing accuracy (%)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	87.90 (7911/9000)	91.90 (8271/9000)
Standard deep DAE	83.10 (7479/9000)	84.60 (7614/9000)
Standard deep CAE	83.30 (7497/9000)	85.10 (7659/9000)
Standard DBN	80.50 (7245/9000)	82.30 (7407/9000)
Standard CNN	86.80 (7812/9000)	91.50 (8235/9000)

Table 15
Diagnosis results comparison of different deep learning models for rotor fault diagnosis.

Deep learning methods	Average testing accuracy (%)	
	Before feature fusion using LPP	After feature fusion using LPP
The proposed method	90.29 (3792/4200)	95.19 (3998/4200)
Standard deep DAE	84.86 (3564/4200)	87.48 (3674/4200)
Standard deep CAE	87.05 (3656/4200)	90.38 (3796/4200)
Standard DBN	84.50 (3549/4200)	85.69 (3599/4200)
Standard CNN	86.90 (3650/4200)	92.88 (3901/4200)

has some excellent properties such as weight sharing and shift-invariant for complex data feature learning [21,36].

In Experiment 6, we preliminarily investigated the diagnosis performance of different deep learning models. Ten trials are performed for classifying bearing operation conditions. Table 14 shows the average diagnosis accuracy of different deep learning methods in Experiment 6. The average testing accuracy of the proposed method is 87.9% before feature fusion, which is slightly higher than other deep learning methods. After feature fusion, the average testing accuracy of the proposed method is 91.9%, compared with other four methods, which are 84.6%, 85.1%, 82.3% and 91.5%, respectively. The main parameters of different methods in Experiment 6 are described as follows. (1) The proposed method: all the parameters are the same as Experiment 4. (2) Standard deep DAE: all the parameters are the same as Experiment 4. (3) Standard deep CAE: all the parameters are the same as Experiment 4. (4) Standard DBN: the architecture of DBN is 1600-400-400-400. The pre-training of each RBM is completed using 200 iterations. The learning rate and momentum are selected as 0.1 and 0.9, respectively. (5) Standard CNN: the architecture of the CNN contains input layer, convolutional layer C1, pooling layer P2, convolutional layer C3, pooling layer P4 and output layer. The size of the input feature map is 40*40, C1 layer contains 6 kernels, C3 contains 12 kernels, and the scales of P2 layer and P4 layer are both set to 2. The learning rate is 0.1 and iteration number is 200.

In addition, different deep learning methods are also used to analyze rotor fault data set used in Section 4. Table 15 shows the average diagnosis accuracy of different deep learning methods for rotor operation conditions classification. The average testing accuracy of the proposed method is 90.29% before feature fusion, which is higher than other deep learning methods. After feature fusion, the average testing accuracy of the proposed method is 95.19%, compared with other methods, which are 87.48%, 90.38%, 85.69% and 92.88%, respectively. The main parameters of different methods in this experiment are described as follows. (1) The proposed method: all the parameters are the same as Experiment 1 in Section 4. (2) Standard deep DAE: all the parameters are the same as Experiment 1 in Section 4. (3) Standard deep CAE: all the parameters are the same as Experiment 1 in Section 4. (4) Standard DBN: the architecture of DBN is 1000–300–300–300–300. The iteration number of each RBM is 200. The learning rate and momentum are selected as 0.1 and 0.9, respectively. (5) Standard CNN: The size of the input feature map is 30*30 (first 900 data points in each raw sample). The other parameters are the same as Experiment 6.

It can be seen from Tables 14 and 15 that: (1) Compared with shallow learning models such as SVM and BP, different deep learning models all show unique advantages for rotating machinery fault diagnosis and feature learning. The main reason is that different deep learning models can adaptively learn the most useful information from the input data through layer-by-layer feature transformation. (2) It seems that the proposed method shows slightly better performance than standard CNN. The reason is that the proposed method is an improved form of the standard deep auto-encoder, which belongs to unsupervised method and shows more powerful feature learning ability compared with standard CNN. However, they show different performance and statistical differences on different data sets.

In order to further confirm the effectiveness and robustness of the proposed deep learning model, another two bearing data sets are applied to research the differences between the proposed model and standard CNN. One is taken from Case Western Reserve University (CWRU) Bearing Data Center [37], and the other is downloaded from Prognostics Center Excellence, NASA [38]. The two data sets are the benchmark data sets, which have widely applied and cited in a large amount of papers. More details about these two data sets can be found in [37–40].

Similar to Section 4, in this case study, three kinds of data sets are created based on CWRU bearing data and NASA bearing data, respectively, i.e., the raw data, the 152-dimensional (152-D) feature data and the 19-dimensional (19-D) feature data. Sixteen bearing operation conditions are created from each CWRU data set, which are the same as [39]. Each condition consists of 150 samples, the random 100 samples of each condition for training and the rest 50 samples for testing. Each raw sample is a collected vibration signal containing 784 data points. Three bearing operation conditions are created from each NASA data set, which are the same as [40]. Each condition consists of 100 samples, the random 50 samples of each condition for training and the rest 50 samples for testing. Each raw sample is a collected vibration signal containing 1225 data points. The length of each sample is determined by experiments.

It must be pointed out the goal of this section (Section 5.5) is to totally compare the differences between different deep learning models (the proposed deep auto-encoder and standard CNN), thus, we only consider the analysis results before feature fusion.

Ten trials are performed for analyzing each bearing data set. Table 16 shows the average diagnosis accuracy of the two deep learning models using CWRU data, and Table 17 is the classification result for NASA data. From Tables 16 and 17, we can find that the proposed model is more effective than standard CNN, especially

Table 16
Performance comparison for CWRU data (16-class classification problem).

Deep learning models	Average testing accuracy (%)		
	Raw data	152-D feature data	19-D feature data
The proposed model	91.51 (7321/8000)	94.11 (7529/8000)	95.23 (7618/8000)
Standard CNN	88.30 (6664/8000)	90.03 (7202/8000)	90.80 (7264/8000)

(*Note:* 152-D means152-dimensional).

Table 17
Performance comparison for NASA data (3-class classification problem).

Deep learning models	Average testing accuracy (%)		
	Raw data	152-D feature data	19-D feature data
The proposed model	97.27 (1459/1500)	98.53 (1478/1500)	100.0 (1500/1500)
Standard CNN	94.13 (1412/1500)	96.20 (1443/1500)	96.67 (1450/1500)

Table 18
Description of the twelve similar data sets.

Data sources	Data sets	Descriptions
Rotor data	Data set 1 (DS-1)	Raw data (1000-dimensional)
	Data set 2 (DS-2)	Feature data (152-dimensional)
	Data set 3 (DS-2)	Feature data (19-dimensional)
Electrical locomotive bearing data	Data set 4 (DS-4)	Raw data (1600-dimensional)
	Data set 5 (DS-5)	Feature data (152-dimensional)
	Data set 6 (DS-6)	Feature data (19-dimensional)
CWRU bearing data	Data set 7 (DS-7)	Raw data (784-dimensional)
	Data set 8 (DS-8)	Feature data (152-dimensional)
	Data set 9 (DS-9)	Feature data (19-dimensional)
NASA bearing data	Data set 10 (DS-10)	Raw data (1225-dimensional)
	Data set 11 (DS-11)	Feature data (152-dimensional)
	Data set 12 (DS-12)	Feature data (19-dimensional)

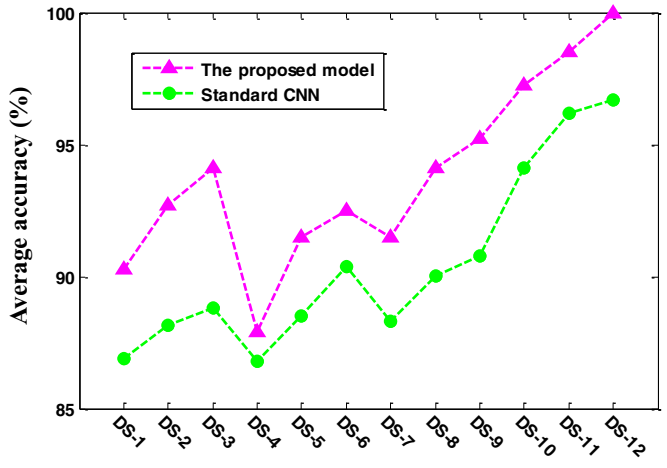

Fig. 32. The accuracies of the two deep learning models based on the twelve similar data sets.

for complex multi-class (16-class) fault classification problem using the raw data. The enhanced feature learning ability makes the proposed deep learning model more easily and accurately to distinguish the different conditions.

In the case of the raw data from CWRU, the main parameters of the two models are described as follows. (1) The proposed method: the architecture is 784–200–200–200. Other parameters are the as Experiment 4. (2) Standard CNN: The size of the input feature map is 28*28, C1 layer contains 6 kernels, C3 contains 12 kernels, and the scales of P2 layer and P4 layer are both set to 2. The learning rate is 0.1 and iteration number is 200. In the case of the raw data from NASA, the parameters are described as follows. (1) The proposed method: the architecture is 1225–300–300–300. Other parameters are the as Experiment 4. (2) Standard CNN: The size of the input feature map is 35*35, C1 layer contains 6 kernels, C3 contains 12 kernels, and the scales of P2 layer and P4 layer are both set to 2. The learning rate is 0.1 and iteration number is 200. The idea of parameter determination when dealing with other data sets is similar. It should be pointed out that the input of the standard CNN is a square matrix. Thus, the first 144-dimensional data is selected from the 152-dimensional feature data, and the first 16-dimensional data is selected from the 19-dimensional feature data.

Conducting a statistical test on differences between accuracies is absolutely crucial to evaluate the performance of different classifiers [41,42]. In this case study, paired *t*-test technique is adopted to further compare the statistical performance of the proposed method and standard CNN. Twelve similar data sets from rotating machinery fault diagnosis area are listed in Table 18, including four kinds of raw data sets, four kinds of 152- dimensional data sets and four kinds of 19- dimensional data sets. It is obviously that the data sets from different data sources are statistically independent. To some extent, the three kinds of data sets from the same data source can be considered independent each other. Table 19 and Fig. 32 are the classification results of the two models using different data sets. We can clearly find that the proposed deep learning model always show higher accuracies on different data sets compared with standard CNN.

Let Acc_i^1 and Acc_i^2 be the performance scores (average accuracies) of the proposed model and standard CNN on the ith out of N ($N=12$) data sets, respectively, shown in Table 19. The calculation process using paired *t*-test statistics can be expressed as

$$Dcc_i = Acc_i^1 - Acc_i^2 \tag{19}$$

$$S_{\overline{Dcc}} = \frac{S_{Dcc}}{\sqrt{N}} = \sqrt{\frac{N\sum_{i=1}^{N}(Dcc_i)^2 - \left(\sum_{i=1}^{N} Dcc_i\right)^2}{N(N-1)}} \tag{20}$$

$$t = \frac{\sum_{i=1}^{N} Dcc_i}{NS_{\overline{Dcc}}} \tag{21}$$

where N is the number of data sets, Dcc_i is the difference of Acc_i^1 and Acc_i^2, and S_{Dcc} is the standard deviation of Dcc_i. The t statistics can be calculated based on the Eqs. (19)–(21), and the value is 2.8673, which is larger than $t_\alpha(11)=2.201$ ($\alpha=0.05$).

Therefore, there is a statistically significant difference between the proposed deep learning models and standard CNN. There is no doubt that more data sets and experiments are needed to acquire the detailed statistical characteristics of the two models.

6. Conclusions

In this paper, an enhancement deep feature fusion method is developed for rotating machinery fault diagnosis. Firstly, the new deep auto-encoder is constructed with DAE and CAE for the enhancement of feature learning ability. Secondly, locality preserving projection is adopted to fuse the learned deep features to further improve the diagnosis efficiency. Finally, the fusion deep features are fed into *softmax* to train the intelligent fault diagnosis model.

The proposed method is applied to the fault diagnosis of rotor and electrical locomotive bearing. The results confirm that the proposed method is much more effective and robust for feature learning and fault diagnosis than shallow learning methods. Compared with standard CNN, the proposed method shows slightly better performance. It is very interesting to investigate the differences between different deep learning models. The authors would continue to investigate this topic in the future.

Table 19
The average accuracies of the two deep learning models based on the twelve similar data sets.

Deep learning models	The average accuracy (%)											
	DS-1	DS-2	DS-3	DS-4	DS-5	DS-6	DS-7	DS-8	DS-9	DS-10	DS-11	DS-12
Proposed model	90.29	92.71	94.14	87.90	91.51	92.51	91.51	94.11	95.23	97.27	98.53	100.0
Standard CNN	86.90	88.16	88.82	86.80	88.52	90.40	88.30	90.03	90.80	94.13	96.20	96.67

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 51475368) and Shanghai Engineering Research Center of Civil Aircraft Health Monitoring Foundation of China (No. GCZX-2015-02).

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] X. Jin, M. Zhao, T.W.S. Chow, Motor bearing fault diagnosis using trace ratio linear discriminant analysis, IEEE Trans. Ind. Electron. 61 (2014) 2441–2451.
[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] Y.G. Lei, Z.J. He, Y.Y. Zi, EEMD method and WNN for fault diagnosis of locomotive roller bearings, Expert Syst. Appl. 38 (2011) 7334–7341.
[5] H.H. Liu, M.H. Han, A fault diagnosis method based on local mean decomposition and multi-scale entropy for roller bearings, Mech. Mach. Theory 75 (2014) 67–78.
[6] B. Muruganatham, M.A. Sanjith, B. Krishnakumar, Roller element bearing fault diagnosis using singular spectrum analysis, Mech. Syst. Signal Process. 35 (2013) 150–166.
[7] Y. Yu, D.J. Yu, J.S. Cheng, A roller bearing fault diagnosis method based on EMD energy entropy and ANN, J. Sound Vib. 294 (2006) 269–277.
[8] 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.
[9] A. Hussein, T. Irina, A fault diagnosis methodology for rolling element bearings based on advanced signal pretreatment and autoregressive modelling, J. Sound Vib. 369 (2016) 246–265.
[10] Y.B. Li, M.Q. Xu, H.Y. Zhao, W.H. Huang, Hierarchical fuzzy entropy and improved support vector machine based binary tree approach for rolling bearing fault diagnosis, Mech. Mach. Theory 98 (2016) 114–132.
[11] R.N. Liu, B.Y. Yang, X.L. Zhang, S.B. Wang, X.F. Chen, Time-frequency atoms–driven support vector machine method for bearings incipient fault diagnosis, Mech. Syst. Signal Process. 75 (2016) 345–370.
[12] X. Xia, J.Z. Zhou, J. Xiao, A novel identification method of Volterra series in rotor-bearing system for fault diagnosis, Mech. Syst. Signal Process. 66-67 (2016) 557–567.
[13] N. Lu, Z.H. Xiao, O.P Malik, Feature extraction using adaptive multiwavelets and synthetic detection index for rotor fault diagnosis of rotating machinery, Mech. Syst. Signal Process. 52-53 (2015) 393–415.
[14] C.S. Chen, J.S. Chen, Rotor fault diagnosis system based on sGA-based individual neural networks, Expert Syst. Appl. 38 (2011) 10822–10830.
[15] H. Keskes, A. Braham, Z. Lachiri, Broken rotor bar diagnosis in induction machines through stationary wavelet packet transform and multiclass wavelet SVM, Electr. Power Syst. Res. 97 (2013) 151–157.
[16] M.O. Mustafa, D. Varagnolo, G. Nikolakopoulos, Detecting broken rotor bars in induction motors with model-based support vector classifiers, Control Eng. Pract. 52 (2016) 15–23.
[17] Y. LeCun, Y. Bengio, G.E. Hinton, Review: deep learning, Nature 521 (2015) 436–444.
[18] 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.
[19] P. Tamilselvan, P.F. Wang, Failure diagnosis using deep belief learning based health state classification, Reliab. Eng. Syst. Saf. 115 (2013) 124–135.
[20] V.T. Tran, F. AlThobiani, A. Ball, An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks, Expert Syst. Appl. 41 (2014) 4113–4122.
[21] 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.
[22] Z.Q. Chen, C. Li, R.V. Sanchez, Gearbox fault identification and classification with convolutional neural networks, Shock Vib. 2 (2015) 1–10.
[23] W.J. Sun, S.Y Shao, R. Zhao, R.Q. Yan, X.W. Zhang, X.F. Chen, A sparse auto-encoder-based deep neural network approach for induction motor faults classification, Measurement 89 (2016) 171–178.
[24] J.W. Leng, P.Y. Jiang, A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm, Knowl.-Based Syst. 100 (2016) 188–199.
[25] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.A. Manzagol, stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[26] H. Schulza, K. Chob, T. Raikob, S. Behnkea, Two-layer contractive encodings for learning stable nonlinear features, Neural Netw. 64 (2015) 4–11.
[27] P. Xiong, H.R. Wang, M. Liu, S.P. Zhou, Z.G. Hou, X.L. Liu, ECG signal enhancement based on improved denoising auto-encoder, Eng. Appl. Artif. Intell. 52 (2016) 194–202.
[28] J. Zabalza, J.C Ren, J.B Zheng, H.M Zhao, C.M Qing, Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging, Neurocomputing 185 (2016) 1–10.
[29] 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.
[30] J.B. Yu, Bearing performance degradation assessment using locality preserving projections and Gaussian mixture models, Mech. Syst. Signal Process. 25 (2011) 2573–2588.
[31] Q. Fan, D.Q. Gao, Z. Wang, Multiple empirical kernel learning with locality preserving constraint, Knowl.-Based Syst. 105 (2016) 107–118.
[32] Y. Bengio, A. Courville, Representation Learning, A review and new perspectives, IEEE Trans. Softw. Eng. 35 (2013) 1798–1828.
[33] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11 (2010) 625–660.
[34] E.D.L Rosa, W. Yu, Randomized algorithms for nonlinear system identification with deep learning modification, Inf. Sci. 364–365 (2016) 197–212.
[35] S. Kim, Y. Choi, M. Lee, Deep learning with support vector data description, Neurocomputing 165 (2015) 111–117.
[36] J.H. Sun, Z.W. Xiao, Y.X. Xie, Automatic multi-fault recognition in TFDS based on convolutional neural network, Neurocomputing 222 (2017) 127–136.
[37] http://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website, 2003.
[38] http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/, 2007.
[39] L.A. Wulandhari, A. Wibowo, M.I. Desa, Condition diagnosis of multiple bearings using adaptive operator probabilities in genetic algorithms and back propagation neural networks, Neural Comput. Appl. 26 (2015) 57–65.
[40] H. Qiu, J. Lee, J. Lin, G. Yu, Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics, J. Sound Vib. 289 (2006) 1066–1090.
[41] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[42] S. Garcia, F. Herrera, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, J. Mach. Learn. Res. 9 (2008) 2677–2694.