# Construction of a batch-normalized autoencoder network and its application in mechanical intelligent fault diagnosis

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **30** (2019) 015106 (14pp) https://doi.org/10.1088/1361-6501/aaf319

# Construction of a batch-normalized autoencoder network and its application in mechanical intelligent fault diagnosis

**Jinrui Wang[1], Shunming Li[1], Baokun Han[2], Zenghui An[1], Yu Xin[1], Weiwei Qian[1] and Qijun Wu[3]**

[1] College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, People's Republic of China
[2] College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao, People's Republic of China
[3] China Ship Development and Design Center, Wuhan, People's Republic of China

E-mail: wangjr33@163.com

Received 11 September 2018, revised 11 November 2018
Accepted for publication 22 November 2018
Published 14 December 2018

**Abstract**

Among various fault diagnosis methods, deep learning has shown state-of-the-art performance in processing mechanical big data. This paper investigates a reliable deep learning method known as autoencoder, which is most suitable for automatic feature extraction of fault signals. However, traditional autoencoders have two deficiencies: (1) the multi-layer structure of autoencoder has an internal covariate shift problem, which will cause great difficulty for the network training. (2) The application of autoencoder in the case of rotating speed fluctuation is not mature. To overcome the aforementioned deficiencies, batch normalization strategy is employed in every layer of the autoencoder network to obtain a steady distribution of activation values during training. It can regularize the network without parameter adjustment, and deal with the speed fluctuation problem perfectly. So, a new network named batch-normalized autoencoder is first proposed for intelligent fault diagnosis. The raw vibration signals are directly fed into the network and the extracted features are employed to train a softmax classifier for health state identification. A bearing and a gearbox data set are finally used to confirm the effectiveness of the proposed method. The results manifest that the proposed method can extract salient features from the raw signals and handle the fault diagnosis problem under the speed fluctuation problem.

Keywords: fault diagnosis, deep learning, autoencoder, batch normalization, speed fluctuation

(Some figures may appear in colour only in the online journal)

## 1. Introduction

In modern industries, internet of things, smart manufacturing and data-driven systems have been revolutionizing manufacturing through enabling computers to collect the massive amount of data from monitored machines and turn the mechanical big data into operational information [1]. At the same time, machines have also been more precise than ever before, so mechanical fault diagnosis has sufficiently embraced the big data revolution in a condition monitoring system. In contrast with top-down modeling proposed by physics-based fault diagnosis models, data-driven systems provide a bottom-up model for detecting the occurrence of machinery faults [2, 3]. As is well known, the physics-based methods are unable to be updated online with measured data and also cannot deal well with large-scale data. On the other hand, with fast-developing computer systems and sensors, data-driven fault diagnosis systems have drawn increasing public attention [4, 5]. Therefore, our work focuses on the data-driven model in this paper.

 

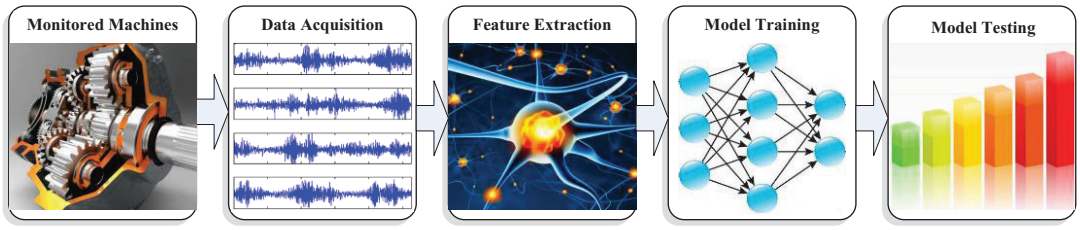


**Figure 1.** Framework of traditional data-driven fault diagnosis system.

The basic framework of the data-driven system is shown in figure 1, which usually consists of four consecutive stages: data acquisition, feature extraction, model training and model testing [6]. Conventional data-driven methods are usually to design a right set of features, and then put the features into some shallow machine learning models such as Naive Bayes [7], support vector machines (SVMs) [8] and logistic regression [9]. But these works always focus on manual feature extraction, such as statistical features, frequency and time-frequency features, that need human labor and cannot achieve updating online [10]. So, it is hard for these methods to extract intrinsic features behind the raw time-domain data. And the selection of these features always needs prior knowledge and feature engineering, which is also a difficult problem. As the hottest subfield of machine learning, deep learning has been regarded as an effective and powerful solution for an intelligent fault diagnosis system to deal with big data and learn high-level abstractions through multiple networks. Compared to traditional methods, deep learning-based methods do not need human labor and expert knowledge for feature extraction, and all the hyperparameters in model training and pattern classification modules are able to be trained jointly. Therefore, deep learning-based methods can be employed to address machinery fault diagnosis in a very general way.

As one of the widely used deep learning techniques, autoencoder has been investigated as a common component of deep neural networks (DNNs) by Bengio *et al* [11]. Meanwhile, autoencoder models have attracted extensive attention in fault diagnosis. Jia *et al* [12] proposed stacked autoencoder (SAE)-based DNNs for roller bearing and planetary gearbox fault diagnosis with input as frequency spectra after Fourier transform. Guo *et al* [13] employed multi-domain statistical features of the raw vibration signals as the input of SAEs, which can be viewed as a kind of feature fusion. Liu *et al* [14] fed the normalized spectrograms created by short-time Fourier transform into SAEs for roller bearing fault diagnosis. In the work presented in [15], the nonlinear soft threshold approach and digital wavelet frame were used to process the measured signals and then fed into SAEs for rotating machinery diagnosis. In the above works, it can be seen that these works still have a preprocess step on the raw signals. The high input dimension of raw signals may lead to heavy computation cost and an overfitting problem in the training process. On the other hand, the raw time-domain signals have a shift-variant property that will lead to low diagnosis accuracy.

But the raw time-domain signals possess the richest information of machinery health condition, and the feature extraction process will inevitably lose some meaningful information of the vibration signals. To solve this problem, some researchers have focused on directly extracting features from raw signals. Sun *et al* [16] adopted a regularization technique called 'dropout' to mask portions of output neurons randomly in the sparse autoencoder, so as to reduce overfitting in the training process. In the work of [17], rectified linear unit (ReLU) activation function and dropout were adopted in the SAE-based model. The experimental result demonstrated their effective performance in decreasing overfitting and gradient vanishing. Jia *et al* [18] constructed a local connection network based on normalized sparse autoencoder. In this experiment, the shift-invariant property was eliminated by locally extracting various features, and the L1 norm was employed to find sparse features. Although these methods can achieve a high classification performance, the parameters of these techniques still require manual selection, i.e. dropout rate and regular parameter. And the selection of these parameters is crucially important for the training process, so that plenty of time is spent on the parameter adjustment. In addition, internal covariate shift is also a tricky problem during the training process [19]. This means that the data distribution of each hidden layer changes with the parameters of the previous layer during training, which requires all layers to continuously adapt to the new distribution. The more weights and biases are altered in the network, the more severe internal covariate shift becomes. As more and more inputs are transferred to the saturated nonlinear region, the learning speed will decrease. Therefore, a lower learning rate and careful parameter initialization are essential, leading to it becoming notoriously hard for the training with saturating nonlinearities. Since the autoencoder network has three layers, it is inevitable that it will suffer from the internal covariate shift problem. Furthermore, the intelligent fault diagnosis problem under rotating speed fluctuation condition has not been resolved. The rotating speed fluctuation condition generally exists in practical engineering applications. As a result, the mapping between signal signs and fault modes becomes extremely complex, which brings great challenges to the health monitoring of mechanical equipment.

As one of the most emotional innovations in the deep learning optimization techniques, batch normalization has aroused great concern over recent years [20]. Actually, it is an adaptive reparametrization algorithm without parameter adjustment. Through several recent studies [19–21], batch normalization has shown its effectiveness in decreasing the internal covariate shift problem. Therefore, we can reduce the iteration number towards convergence and accelerate the

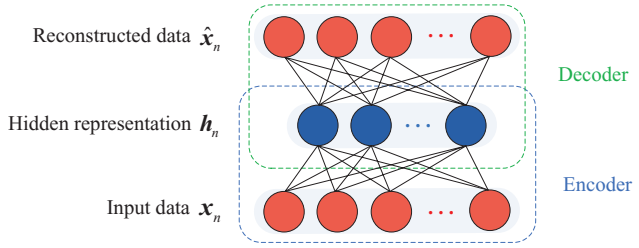


**Figure 2.** Structure of an autoencoder.

training process. Inspired by the prior researches, a new fault diagnosis framework with batch normalization is proposed. Meanwhile, considering the scale and shift abilities of batch normalization on normalizing input values, we present the first empirical study to apply it to intelligent fault diagnosis under rotating speed fluctuation condition.

The rest of the paper is organized as follows. Section 2 briefly introduces the algorithms of autoencoder and batch normalization. Section 3 is dedicated to detailing the content of the proposed batch-normalized autoencoder (BNAE) network. In section 4, the diagnosis cases of bearing and gear data sets are adopted to validate the effectiveness of the proposed method. Furthermore, the superiority of the proposed method is exhibited by comparing it with some other traditional methods. Finally, some conclusions are drawn in section 5.

## 2. Theoretical background

### 2.1. Autoencoder

Autoencoder is a kind of unsupervised learning structure that has three layers: input layer, hidden layer and output layer, as shown in figure 2. The process of an autoencoder training consists of two parts: encoder and decoder. Encoder is used for mapping the input data into hidden representation, and decoder is referred to reconstruct input data from the hidden representation. Given the unlabeled input data set $\{\boldsymbol{x}_n\}_{n=1}^{N}$ where $\boldsymbol{x}_n \in R^{m\times 1}$, $\boldsymbol{h}_n$ represents the hidden encoder vectors calculated from $\boldsymbol{x}_n$ and $\widehat{\boldsymbol{x}}_n$ is the decoder vectors of the output layer. Hence, the encoding process is as follows:

$$\boldsymbol{h}_n = f(\boldsymbol{W}_1\boldsymbol{x}_n + b_1), \tag{1}$$

where $f$ is the encoding function, $\boldsymbol{W}_1$ is the weight matrix of the encoder and $b_1$ is the bias vector.

The decoder process is defined as follows:

$$\hat{\boldsymbol{x}}_n = g(\boldsymbol{W}_2\boldsymbol{h}_n + b_2), \tag{2}$$

where $g$ is the decoding function, $\boldsymbol{W}_2$ is the weight matrix of the decoder and $b_2$ is the bias vector.

The parameter sets of the autoencoder are optimized to minimize the reconstruction error:

$$\phi(\Theta) = \arg\min_{\theta,\theta'} \frac{1}{n}\sum_{i=1}^{n} L\left(\boldsymbol{x}^i, \hat{\boldsymbol{x}}^i\right), \tag{3}$$

where $L$ represents a loss function $L(\boldsymbol{x},\hat{\boldsymbol{x}}) = \|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2$.

### 2.2. Batch normalization

Batch normalization can reparametrize almost any deep network in an elegant way, and it is able to be employed in any input or hidden layer of a network. The primary target of batch normalization is to gain a steady distribution of activation values during the learning process, and it applies normalization technique on every training mini-batch. For a layer with $n$-dimensional input $\boldsymbol{x} = (x_1 \ldots x_n)$, to improve the training and reduce the internal covariate shift, two necessary simplifications are taken by batch normalization.

First, each scalar feature is normalized independently by making it have zero mean and unit variance.

$$\hat{x}_i = \frac{x_i - E[x_i]}{\sqrt{\mathrm{Var}[x_i]}}, \tag{4}$$

where $E[x_i]$ is the mean of each unit and $\sqrt{\mathrm{Var}[x_i]}$ denotes the standarddeviation.

However, the simple normalization of each input in a layer can still change the representation of the layer. So, two parameters $\gamma_i$ and $\beta_i$ are employed for each activation $x_i$, which aims to scale and shift the normalized value:

$$f_i = \gamma_i\hat{x}_i + \beta_i. \tag{5}$$

The $\gamma_i$ and $\beta_i$ are learned along with the raw model parameters and restore the representation power of the network. Note that the raw activations can be recovered by setting $\gamma_i = \sqrt{\mathrm{Var}[x_i]}$ and $\beta_i = E[x_i]$. In this case, the steady distribution of activation values can be guaranteed during each training.

Second, consider a particular activation $x_i$ which owns $m$ values in a mini-batch $\varphi = \{x_{1\ldots m}\}$, $\hat{x}_{1\ldots m}$ that denotes the normalized values and the corresponding linear transformations are $y_{1\ldots m}$. Therefore, the batch normalization transform $\mathrm{BN}_{\gamma,\beta} : x_{1\ldots m} \rightarrow y_{1\ldots m}$ can be displayed as follows:

$$E[x_\varphi] = \frac{1}{m}\sum_{j=1}^{m} x_j, \tag{6}$$

$$\mathrm{Var}[x_\varphi] = \frac{1}{m}\sum_{j=1}^{m}(x_j - E[x_\varphi])^2, \tag{7}$$

$$\hat{x}_j = \frac{x_j - E[x_\varphi]}{\sqrt{\mathrm{Var}[x_\varphi] + \varepsilon}}, \tag{8}$$

$$\hat{y}_j = \gamma\hat{x}_j + \beta, \tag{9}$$

where $\varepsilon$ is a constant, which is imposed to avoid meeting with the undefined gradient of $\sqrt{s}$ at $s = 0$.

Furthermore, the gradient of loss $\ell$ should be backpropagated by BN transformation during training:

$$\frac{\partial\ell}{\partial\hat{x}_j} = \frac{\partial\ell}{\partial y_j}\gamma, \tag{10}$$

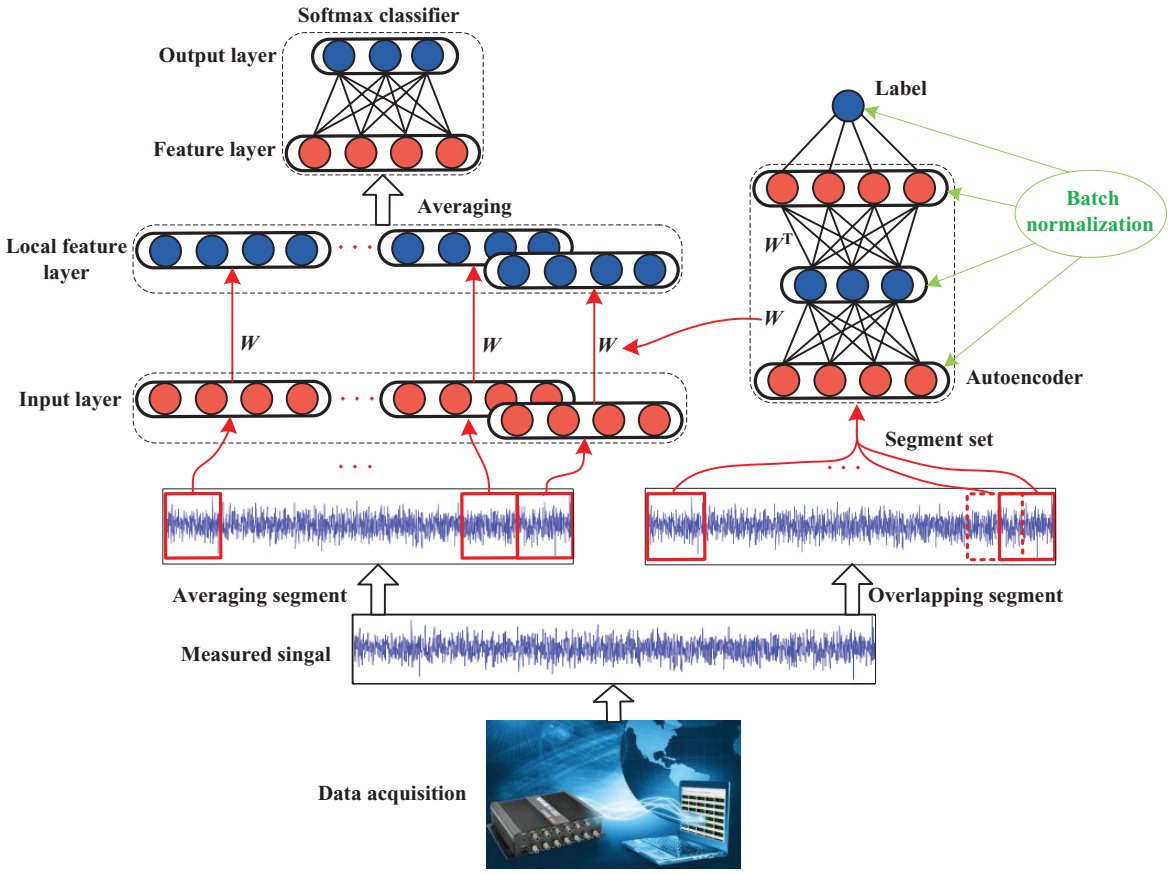


**Figure 3.** Flowchart of the proposed method.

$$\frac{\partial \ell}{\partial \mathrm{Var}[x_\varphi]} = \sum_{j=1}^{m} \frac{\partial \ell}{\partial \hat{x}_j}(x_j - E[x_\varphi]) \cdot \left[-\frac{1}{2}(\mathrm{Var}[x_\varphi] + \varepsilon)^{-\frac{3}{2}}\right], \tag{11}$$

$$\frac{\partial \ell}{\partial E[x_\varphi]} = \sum_{j=1}^{m} \frac{\partial \ell}{\partial \hat{x}_j} \frac{-1}{\sqrt{\mathrm{Var}[x_\varphi] + \varepsilon}}, \tag{12}$$

$$\frac{\partial \ell}{\partial x_j} = \frac{\partial \ell}{\partial \hat{x}_j} \frac{-1}{\sqrt{\mathrm{Var}[x_\varphi] + \varepsilon}} + \frac{\partial \ell}{\partial \mathrm{Var}[x_\varphi]} \cdot \frac{2(x_j - E[x_\varphi])}{m} + \frac{\partial \ell}{\partial E[x_\varphi]} \cdot \frac{1}{m}, \tag{13}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_j} \cdot \hat{x}_j, \tag{14}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_j}. \tag{15}$$

Therefore, BN transformation introduces normalized activations into the network, and ensures the layers can continue learning on input distributions that reduce the influence of internal covariate shift, so as to accelerate the training.

## 3. Proposed framework

This section details the proposed BNAE-based intelligent fault diagnosis method. In the method, BNAE is proposed to overcome the two deficiencies of autoencoder in the training process.

### 3.1. BNAE

We apply a batch normalization layer immediately before the activation, in this case, the nonlinear units are able to generate activation with a stable distribution and avoid overfitting. This is so that the gradients from early in the training process will originate from extremely shallow paths. Generally, the mapping function of autoencoder is an affine transformation, as show in equation (1). We employ the batch normalization transform to normalize $\boldsymbol{W}_1\boldsymbol{x}_n + b_1$, which is immediately before the activations. Note that the bias $b_1$ is not used, since its effect will be canceled in the following mean subtraction process [22]. So, equation (1) is replaced with

$$\boldsymbol{h} = f_r(BN\boldsymbol{W}_1\boldsymbol{x}_n). \tag{16}$$

Here, the BN transform is used independently to each dimension of $\boldsymbol{W}_1\boldsymbol{x}_n$, with a separate pair of learned parameters $\gamma_n$ and $\beta_n$ per dimension. In addition, the activation function $f_r(\cdot)$we applied is ReLU [23], as shown in equation (17). It can be seen as a linear function, and the gradient will not decrease with the increasing of input $s$. So, the network will not suffer from gradient vanishing, which is more suitable for representing and learning the complex vibration signals.

$$f_r(s) = \max(0, s). \tag{17}$$

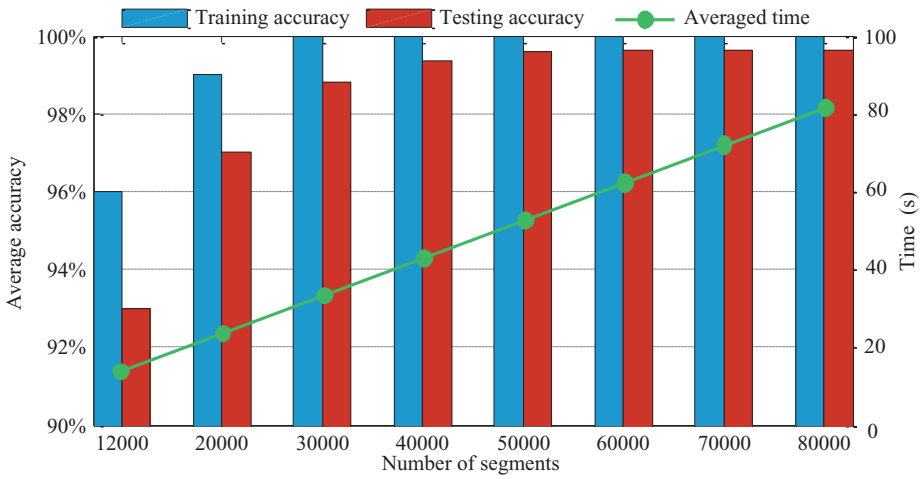


**Figure 4.** Diagnosis results using different segment numbers.

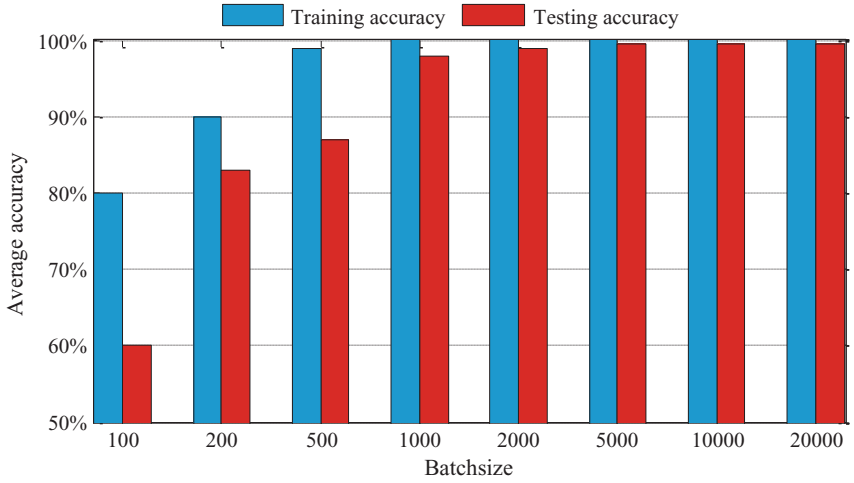


**Figure 5.** Diagnosis results using different batch sizes.

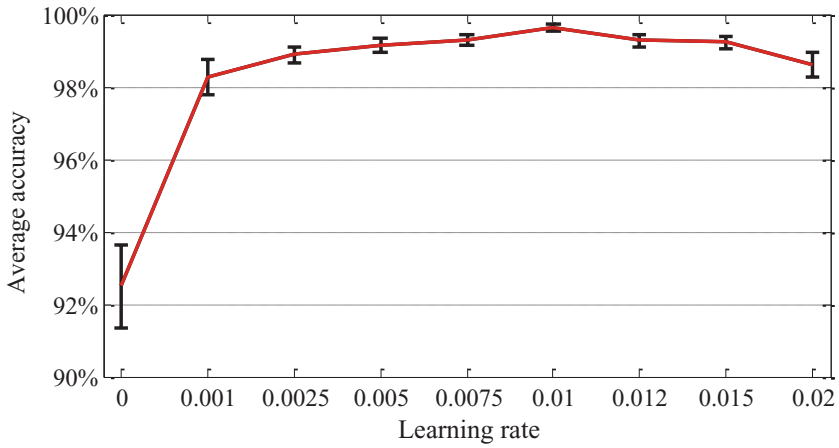


**Figure 6.** Diagnosis results using different learning rates.

### *3.2. Two-stage learning framework*

The framework and illustration of the proposed BNAE-based method are displayed in figure 3. In the first learning stage, the BNAE network is applied to train the weight matrix from raw vibration signals, and then the weight matrix is adopted to extract local feature from each sample. Afterwards, the learned feature is calculated by averaging. In the second learning stage, softmax regression is used for classification by the learned features.

First, the vibration signals under different health conditions are composed of the training set $\left\{\boldsymbol{X}^i, l^i\right\}_{i=1}^{M}$, where $M$ is the number of samples, $\boldsymbol{X}^j \in \Re^{N\times 1}$ is the $i$th sample containing $N$

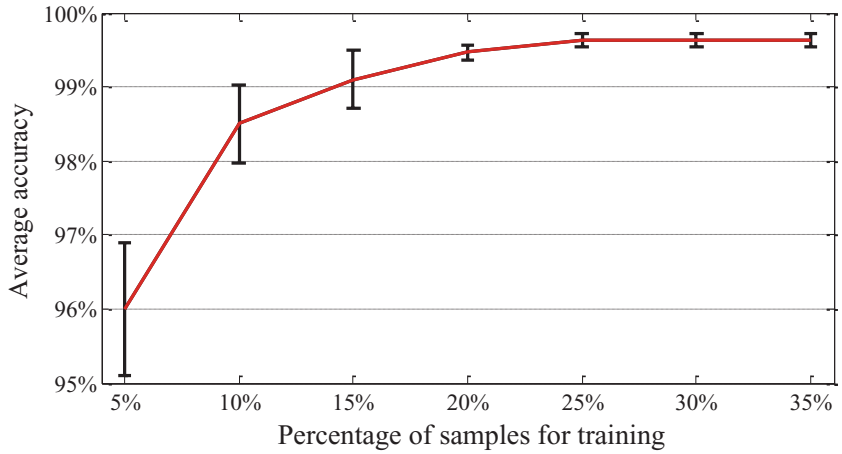


**Figure 7.** Diagnosis results using different learning rates.

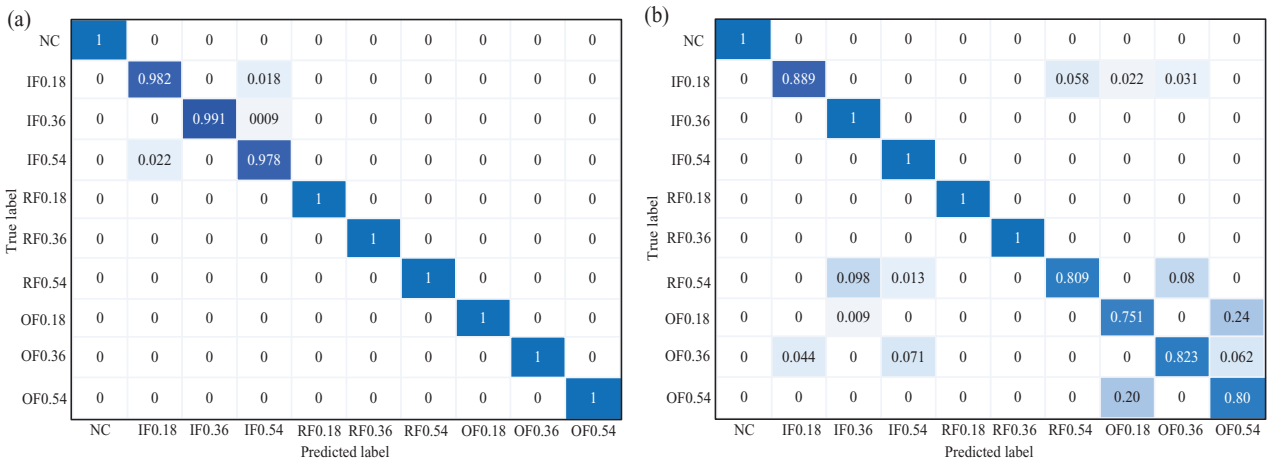


**Figure 8.** Classification results of the bearing data set: (a) confusion matrix of BNAE, (b) confusion matrix of AE.

data points and $l^i$ is the health label of $\boldsymbol{X}^i$. The training set is used to train the autoencoder through following steps:

(1) Use the overlapped method to sample $\boldsymbol{S}$ segments from the training samples. The goal of overlapping is to eliminate the effect of shift-variant property in raw vibration data. These segments constitute the unsupervised training set $\left\{\boldsymbol{S}^j\right\}_{j=1}^{M}$, where $\boldsymbol{S}^j \in \Re^{N_{in}\times 1}$ is the $j$th segment consisting of $N_{\text{in}}$ data points. The segment set $\left\{\boldsymbol{S}^j\right\}_{j=1}^{M}$ is combined into a matrix form $\boldsymbol{T} \in \Re^{N_{in}\times M}$. It is worth mentioning that there is no need of whitening due to the employment of batch normalization, which can save plenty of time in calculation.

(2) The matrix $T$ is adopted to train the BNAE network in order to obtain the weight matrix $W$. ReLU is employed as the activation function in autoencoder, and the bias is not used.

(3) Utilize BP algorithm to update weights with labeled training set $\left\{\boldsymbol{X}^i, l^i\right\}_{i=1}^{M}$ through minimizing the error between extracted features and health labels.

(4) Each training sample is averagely divided into $K$ segments and compose a segment set $\left\{\boldsymbol{x}^i\right\}_{i=1}^{K}$, where $K$ is equal to $N/N_{\text{in}}$. For each segment, we can get the local feature $\boldsymbol{f}_l^i \in \Re^{N_{out}\times 1}$ by a non-linear function [24] as follows:

$$\boldsymbol{f}_l^i = \log(1 + (\boldsymbol{W}\cdot\boldsymbol{x}^i)^2). \tag{18}$$

(5) An average is used to calculate the learned feature of each sample as follows:

$$\boldsymbol{f}^i = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{f}_l^i. \tag{19}$$

Then, we combine the learned feature set with the health label set $\left\{\boldsymbol{f}^i, l^i\right\}_{i=1}^{N}$ to train softmax regression [25]. For each

**Table 1.** Classification comparison of the bearing data set.

| Method | Training samples | No. of classes | Testing accuracy |
|---|---|---|---|
| [27] | 40% | 4 | 95.80% |
| [28] | 75% | 10 | 88.90% |
| [29] | N/A | 12 | 96.67% |
| [30] | 50% | 7 | 99.07% |
| Raw AE | 25% | 10 | 93.37% |
| Proposed | 25% | 10 | **99.63%** |

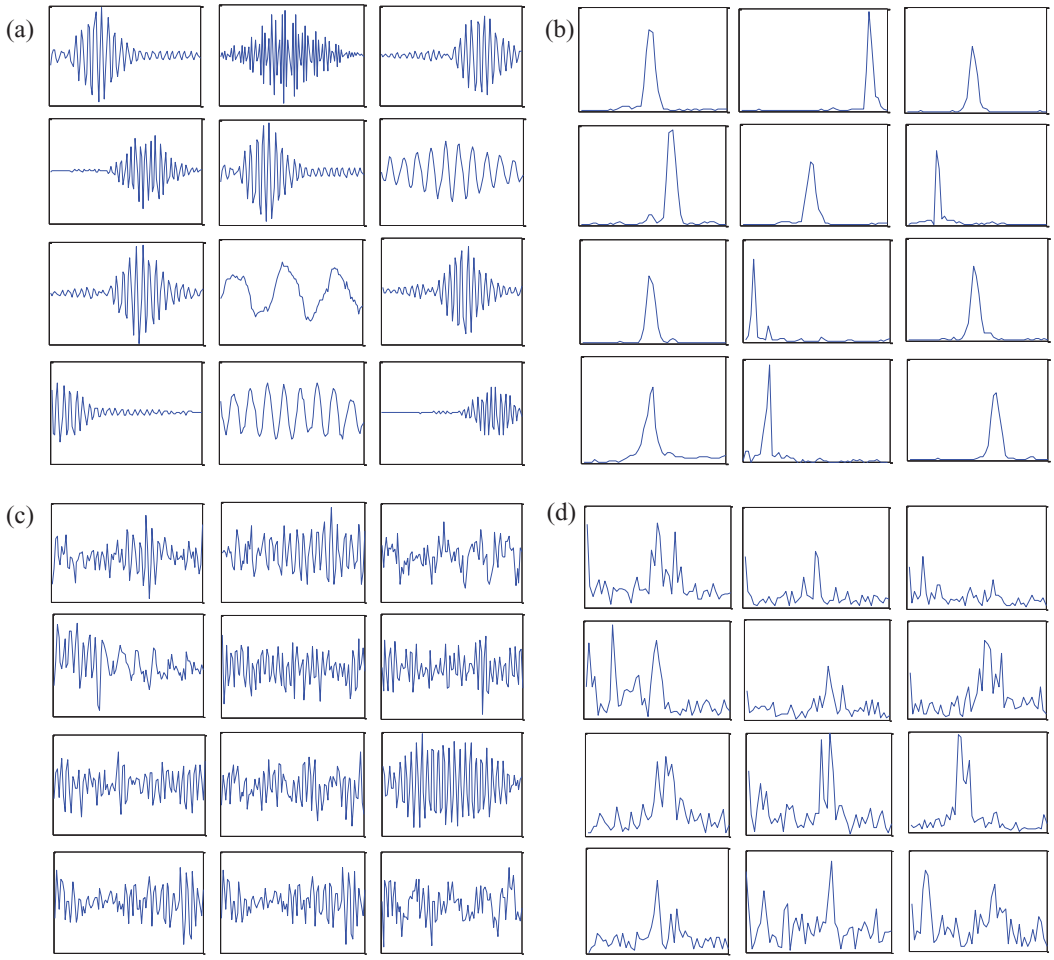


**Figure 9.** Weight vectors for the motor bearing data set: (a) vectors of BNAE in the time domain, (b) vectors of BNAE in the frequency domain, (c) vectors of raw AE in the time domain, (d) vectors of raw AE in the frequency domain.

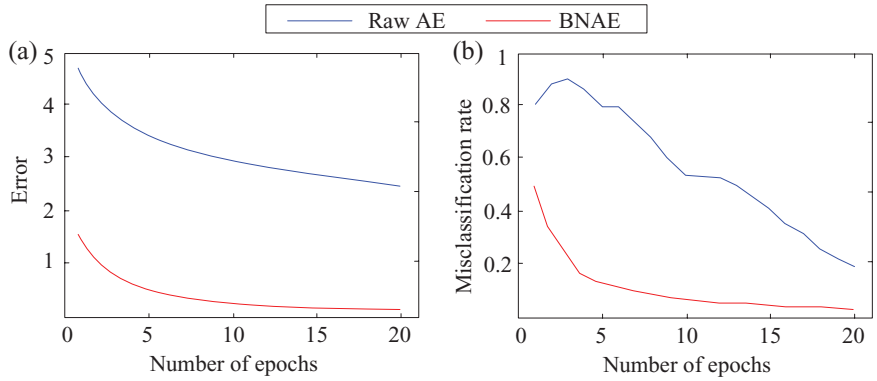


**Figure 10.** (a) Training errors for the bearing data set, (b) misclassification rates for the bearing data set.

$\boldsymbol{f}^i$, the softmax classifier computes the probability $p$ for each label as follows:

$$h_\theta\left(\boldsymbol{x}^i\right) = \begin{bmatrix} p\left(l^i = 1|\boldsymbol{f}^i;\boldsymbol{\theta}\right) \\ p\left(l^i = 2|\boldsymbol{f}^i;\boldsymbol{\theta}\right) \\ \vdots \\ p\left(l^i = K|\boldsymbol{f}^i;\boldsymbol{\theta}\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k}\exp\left(\boldsymbol{\theta}_j^T\boldsymbol{f}^i\right)} \begin{bmatrix} \exp\left(\boldsymbol{\theta}_1^T\boldsymbol{f}^i\right) \\ \exp\left(\boldsymbol{\theta}_2^T\boldsymbol{f}^i\right) \\ \vdots \\ \exp\left(\boldsymbol{\theta}_k^T\boldsymbol{f}^i\right) \end{bmatrix}, \tag{20}$$

where $\theta_1, \theta_2, \cdots \theta_k$ are the parameters of the model. $\sum_{j=1}^{k}\exp\left(\boldsymbol{\theta}_j^T\boldsymbol{f}^i\right)$ is used to normalize the distribution to make sure the sum of $p\left(l^i = j|\boldsymbol{f}^i\right)$ equals 1.

After the two learning stages, the test samples are used to verify the effectiveness of the proposed method. First, the test samples are also divided into segments alternately. Next, the local features are extracted from the segments by the trained BNAE network. Then, averaging is applied to calculate the

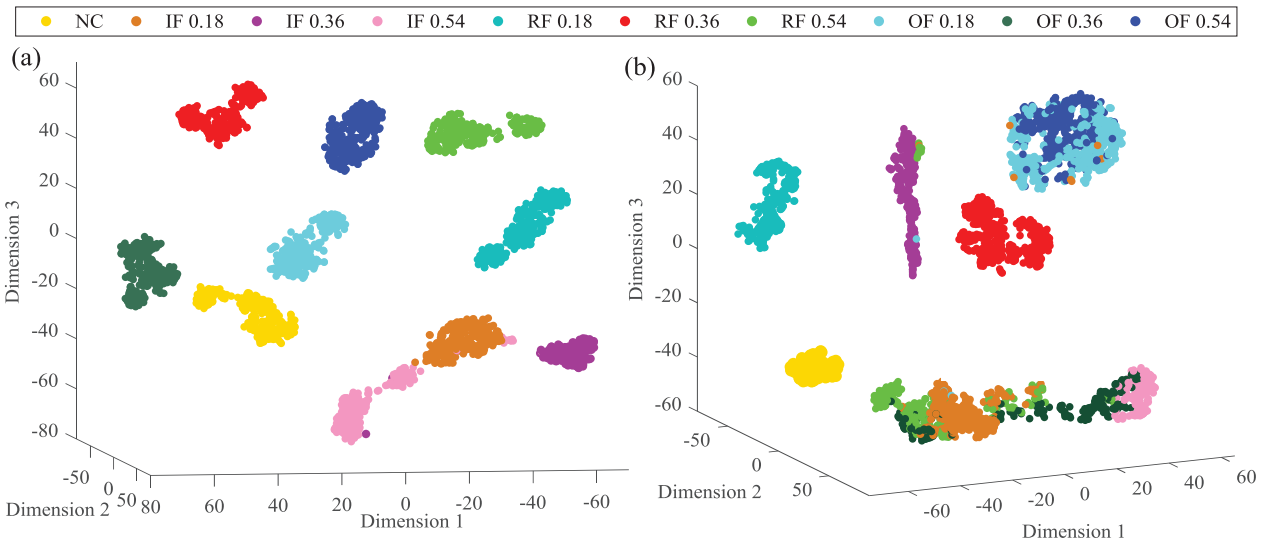


**Figure 11.** Feature visualization map by *t*-SNE for the motor bearing data set: (a) BNAE, (b) raw AE.

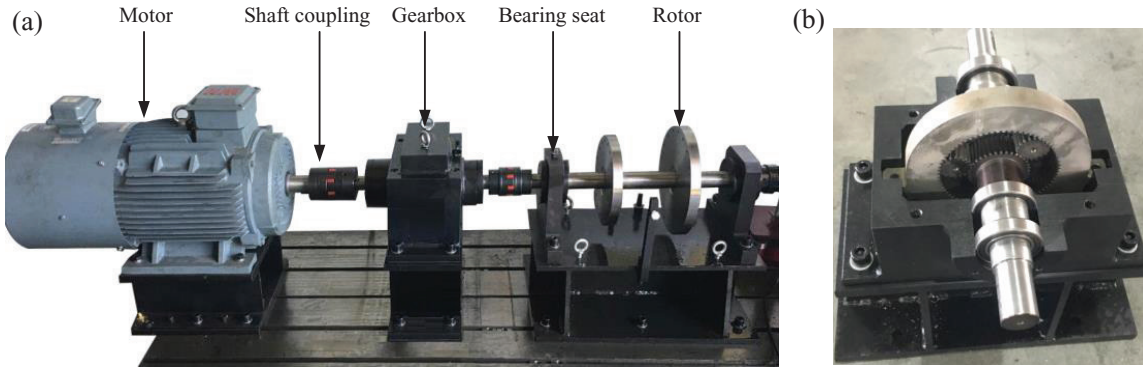


**Figure 12.** (a) Bench of fault planetary gearbox, (b) planetary gearbox.

learned features of each test sample. Finally, the learned features are input to the trained softmax regression model for classification.

## 4. Fault diagnosis using the proposed method

### *4.1. Case 1: fault diagnosis of a motor bearing*

*4.1.1. Data description.* The bearing data set provided by Case Western Reserve University [26] is analyzed in this section. The vibration signals were collected using an accelerometer from the drive end of a motor under four different operating conditions: normal condition (NC); inner race fault (IF); outer race fault (OF) and roller fault (RF). There are three different severity levels (0.18, 0.36 and 0.53 mm) for IF, OF and RF cases. All the samples were collected under four different loads (0, 1, 2 and 3 hp) and the sampling frequency was 12 kHz. Therefore, the data set includes ten health states under four loads, and we treat the same health state under different loads as one class. 100 samples are obtained from each health state under one load, and each sample contains 1200 data points. So, the total sample number of the data set is 4000.

*4.1.2. Parameter selection of the proposed method.* The parameters in the proposed method contain $N_{\text{in}}$, $N_{\text{out}}$, learning rate $\alpha$ and the segment number. The selection of $N_{\text{in}}$ and $N_{\text{out}}$ we refer to the achievement in [18], which are all set as 100. Due to the strong generalization ability of batch normalization, the iteration number is fixed as 20. In addition, 15 trials are carried out for the experiments in order to reduce the effect of randomness.

First, we investigate the selection of the segments of autoencoder. 25% of samples are randomly selected to train the network, and the rest are used for testing. The diagnosis accuracies are shown in figure 4. The result reveals that the accuracies are increasing with the increasing of the segment number, but the time spent almost increases linearly. Consequently, to take the tradeoff between the classification accuracy and time spent, we use 50 000 as the segment number, since the accuracies are not obviously increasing after segment number 50 000. Figure 5, shows the diagnosis accuracies using various batch sizes. It can be seen that when the batch size increases, the testing accuracy is higher. Since the increasing of the accuracies is not obvious after the batch size equals 5000, we choose 5000 as the batch size. Then, we investigate the selection of the learning rate $\alpha$. The diagnosis

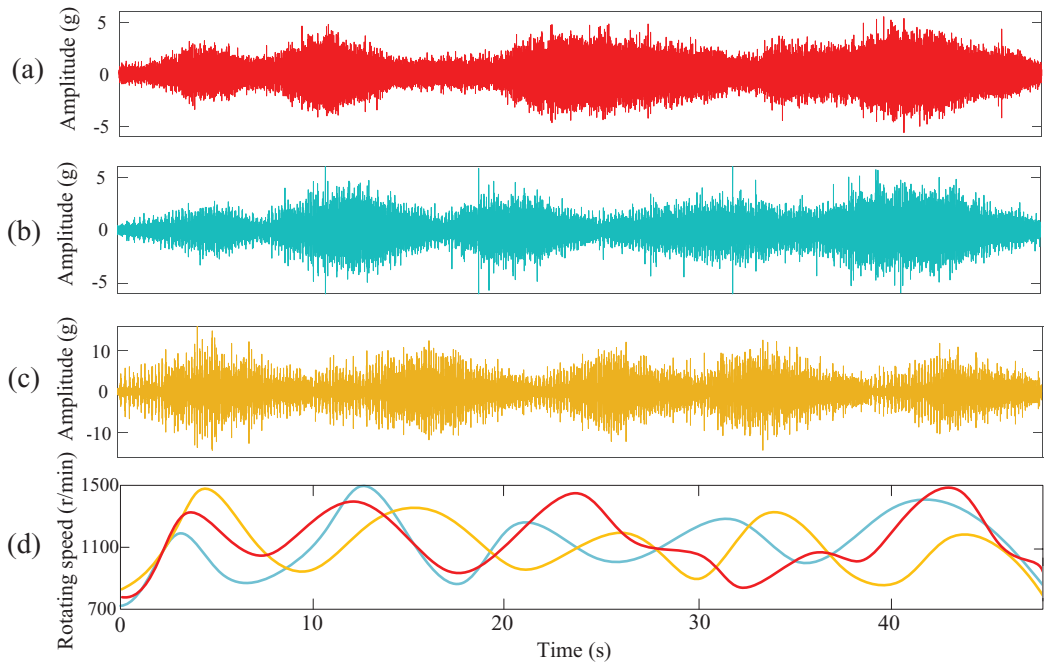


**Figure 13.** Speed fluctuation condition of three gearbox health conditions: (a) NC (b) WW (c) WWPW (d) rotating speeds.

results using different values of $\alpha$ are displayed in figure 6. It is clear that when $\alpha$ is 0.01, the diagnosis accuracy is the highest. The reason is that the bigger $\alpha$ is, the more unstable the system will be. Otherwise, if it is too small, the training epoch time will take longer. In general, the error can converge asymptotically by a suitable learning rate. The accuracy and standard deviation are stable after $\alpha = 0.01$, so 0.01 is chosen as the learning rate in this experiment. Finally, the diagnosis results using the proposed method trained by different percentages of samples are shown in figure 7. It is certain that the testing accuracy increases and its standard deviation decreases with the rise of percentage of samples. When the percentage increases to 25%, the testing accuracy is 99.63% with a small standard deviation of 0.10%. Thereafter, the accuracies are no longer increasing, so we use 25% of samples for training.

*4.1.3. Diagnosis results.* To detail the classification results of each health condition, the confusion matrices of the testing accuracies for a bearing data set by the proposed BNAE model and raw AE model are displayed in figure 8. In figure 8(a), it can be seen that some samples of IF0.18 and IF0.36 are misclassified as IF0.54, and some samples of IF0.54 are misclassified as IF0.18 by the proposed BNAE model. In figure 8(b), compared with BNAE, raw AE classifies RF0.54, OF0.18, OF0.36 and OF0.54 with the accuracies of only 80.9%, 75.1%, 82.3% and 80%, respectively. Therefore, this indicates the superiority of BNAE in the classification of a bearing data set. The average testing accuracy of 15 trials is 99.63%, as shown in table 1, which indicates that the BNAE network can directly obtain discriminative features from the raw signals to a high accuracy. For comparison, several diagnosis methods are also presented in table 1. Li *et al* [27] proposed a method that combined 19 time-domain and frequency-domain features with a self-organizing map for the bearing diagnosis under four loads, and achieved 95.8% testing accuracy. In [28], wavelet leaders multifractal features and an SVM model were used to represent ten health conditions of bearings under four loads, and finally obtained 88.9% classification accuracy. Lin *et al* [29] proposed a bearing diagnosis method using multifractal detrended fluctuation and achieved 96.67% accuracy. Xu *et al* [30] applied multiple domain features and ensemble fuzzy ARTMAP neural networks to distinguish the health conditions and 99.07% accuracy was achieved. Furthermore, the raw autoencoder without batch normalization (raw AE) is also adopted for comparison. The testing accuracy is 93.37%, which exhibits the effectiveness of batch normalization in feature extraction. Compared with the methods above, the proposed BNAE method can not only automatically distinguish the ten health conditions of bearings from the raw vibration signals, but achieve a higher accuracy with a lower percentage of training samples.

**Table 2.** Classification comparison of the gearbox data set.

| Method | Training accuracy (%) | Testing accuracy (%) |
|---|---|---|
| Raw AE | 85.65 | 70.12 |
| Proposed | 100 | 99.23 |

To illustrate the strong feature extraction ability of the BNAE network, 12 weight vectors of $\boldsymbol{W}$ trained by BNAE are displayed in figure 9(a), which is used to show the patterns of the network weights learning [31]. The corresponding frequency spectra of these vectors are shown in figure 9(b). It can be seen that these weight vectors are time-localized and have narrow spectral bandwidths, which are able to serve as good bandpass bases for mechanical signals. For comparison, we randomly plot 12 weight vectors of $\boldsymbol{W}$ trained by raw AE in figures 9(c) and (d). It is easy to find that these vectors also have some time-frequency properties, but the properties are

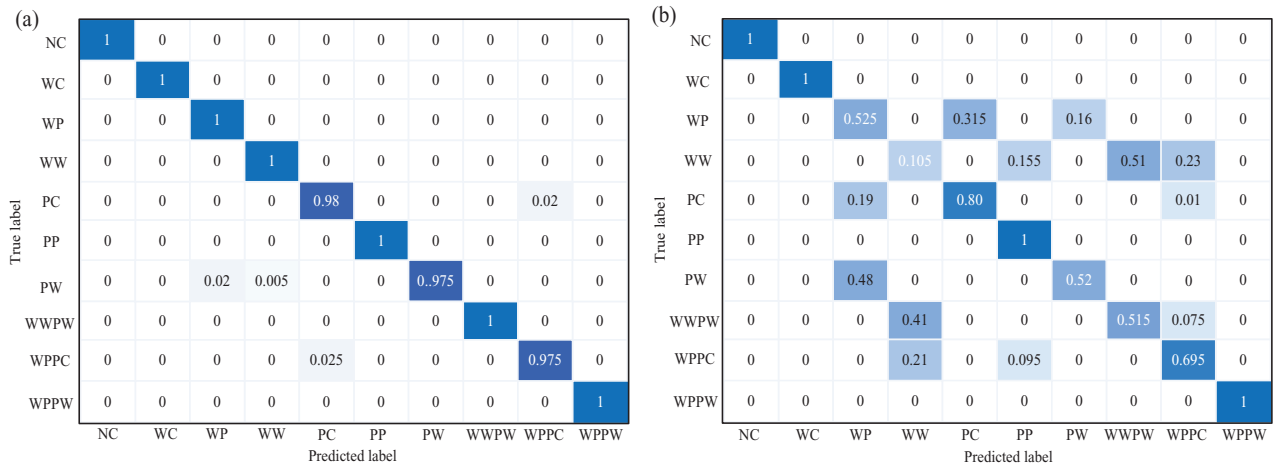


**Figure 14.** Classification results of the gearbox data set: (a) confusion matrix of BNAE, (b) confusion matrix of AE.

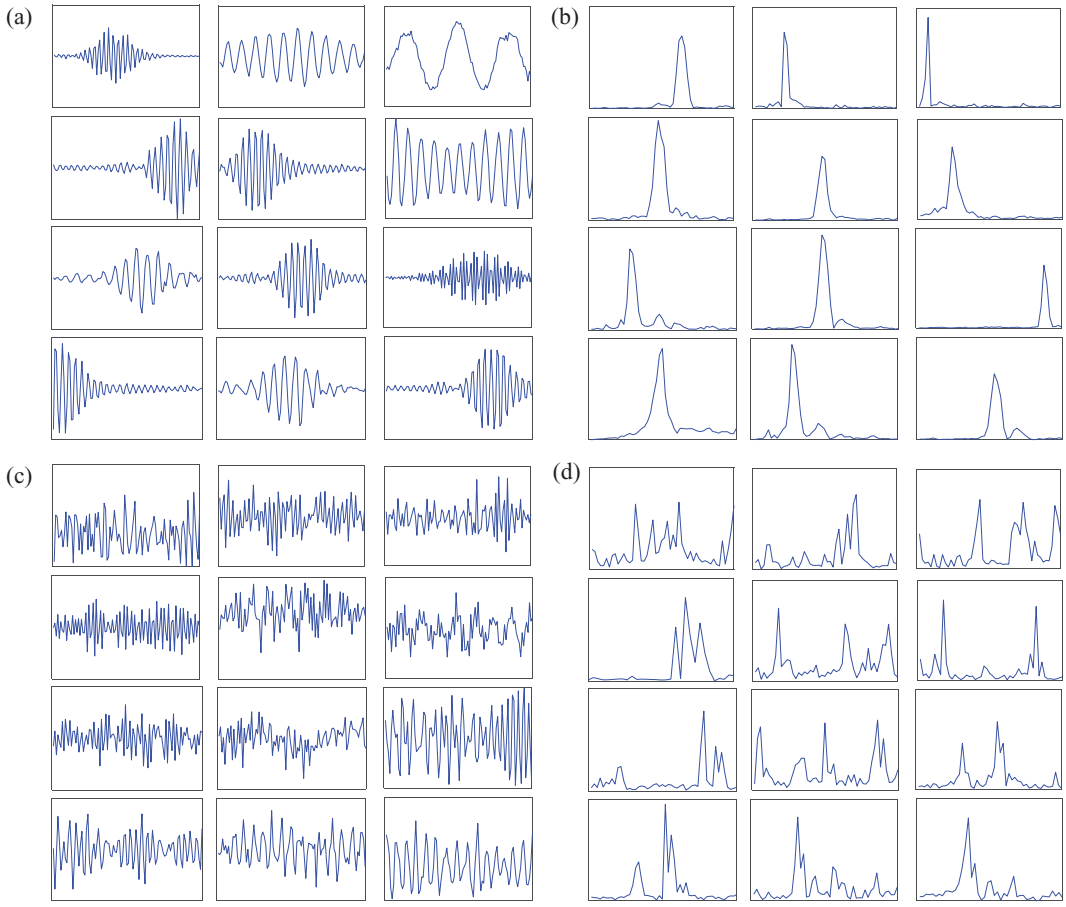


**Figure 15.** Weight vectors for the gearbox data set: (a) vectors of BNAE in the time domain, (b) vectors of BNAE in the frequency domain, (c) vectors of raw AE in the time domain, (d) vectors of raw AE in the frequency domain.

too similar. Furthermore, figure 10(a) indicates the training error of 20 epochs by the two methods, and figure 10(b) is the corresponding misclassification rate. It can be seen that the training error of the proposed model converges to almost 0 after 20 epochs, and the misclassification rate also gets close to 0. In contrast, the performance of the raw AE model is not as good as the proposed method, which cannot update to 0 and obtain a lower misclassification rate with the same epochs. So, it is no wonder that the raw AE is not able to achieve as high a diagnosis accuracy as the proposed method. To further investigate the learned features in the BNAE network, *t*-SNE [32] is employed to achieve feature visualization, as shown in figure 11(a). It is easy to see that different fault types are separated perfectly and the same fault-type samples are almost clustered together. And several samples of IF 0.18, IF 0.36 and IF 0.54 are mixed, which is also the same as the result in figure 7. Figure 10(b) displays the features extracted by the raw AE model. It can be found that the result has a

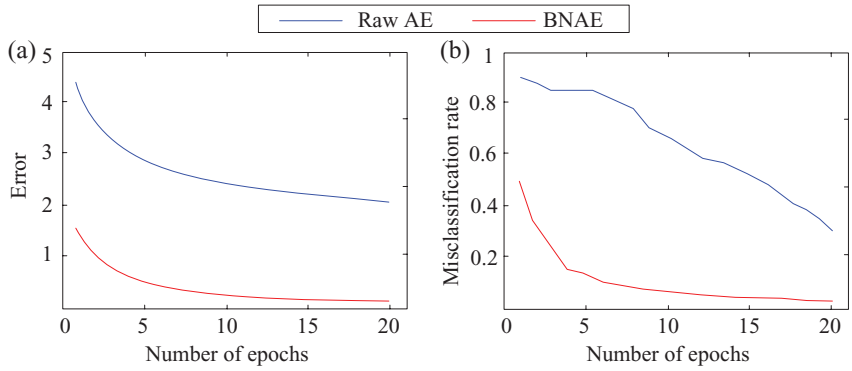


**Figure 16.** (a) Training errors for the gearbox data set, (b) misclassification rates for the gearbox data set.

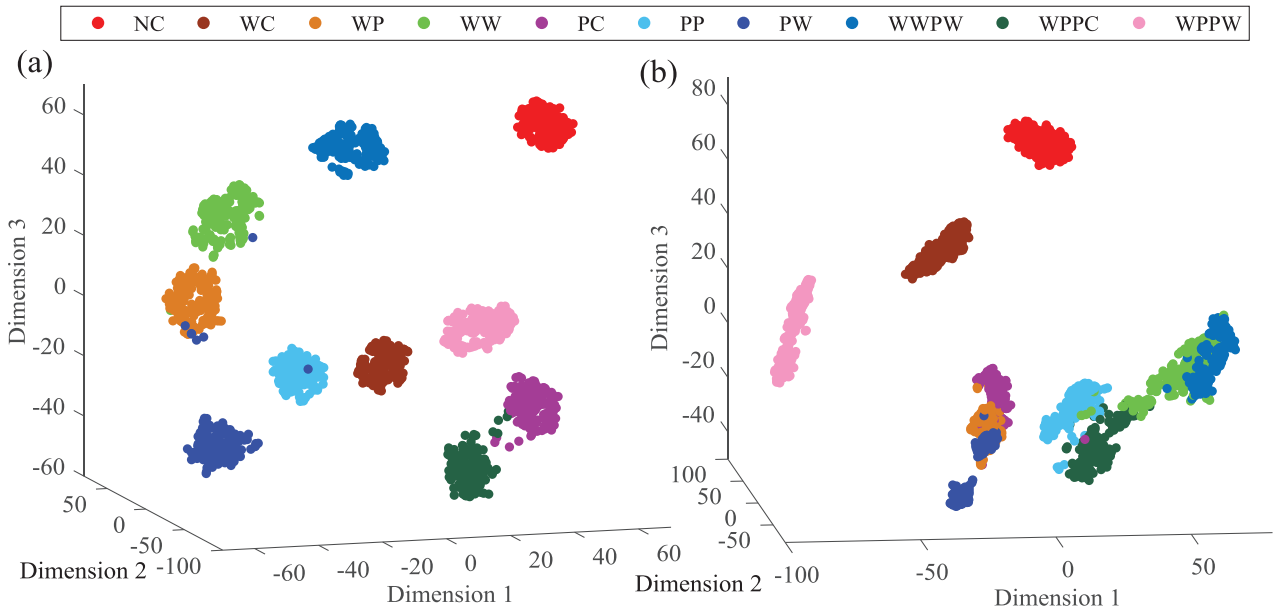


**Figure 17.** Feature visualization map by *t*-SNE for the gearbox data set: (a) BNAE model, (b) raw AE model.

serious aliasing phenomenon, which exhibits the poor feature extraction ability of the raw AE model. It also demonstrates the effectiveness of batch normalization in the employment of AE.

### *4.2. Case 2: fault diagnosis of a planetary gearbox under speed fluctuation*

Considering the scale and shift abilities of batch normalization on normalizing input values, we present the first empirical study and apply it to intelligent fault diagnosis under rotating speed fluctuation condition. As is known, the amplitude and frequency of fault signals under different rotating speeds will generate great changes, which will bring great changes for fault diagnosis. However, the batch normalization technique possesses scale and shift abilities, which can be employed to solve this problem perfectly. In this section, a planetary gearbox experimental data set with speed fluctuation condition is employed to validate the effectiveness of the proposed BNAE network. The vibration signals were collected on a specially designed bench, which consisted of a planetary gearbox, motor, two bearing seats and two shaft couplings, as shown in figure 12. An accelerometer sensor was mounted on the flat surface of the gearbox. There are ten health conditions: NC three health conditions of sun wheel (crack, pit and worn tooth), which are named WC, WP and WW, three health conditions of pinion (crack, pit, and worn tooth), which are named PC, PP and PW, and three coupled faults (wheel worn and pinion worn, wheel pit and pinion crack, wheel pit and pinion worn), which are named WWPW, WPPC and WPPW. As shown in figure 13, we use three examples (NC, WW, WWPW) to illustrate the speed fluctuation condition: the rotating speed fluctuated irregularly between 700 and 1500 r $min^{-1}$, and the speed fluctuation varies with different health conditions. 500 data samples are collected from each health condition, so a total of 5000 samples are obtained from the designed bench and each sample contains 1200 data points. The sensor is a piezoelectric accelerometer (DH131E) that is mounted on the flat surface of the gearbox and the sampling frequency is 12.8 kHz.

The structure of the designed network and the parameter set used for the gearbox data set are all the same as those used in the bearing data set. We use 60% of the gearbox vibration samples for training and the rest for testing, with the average

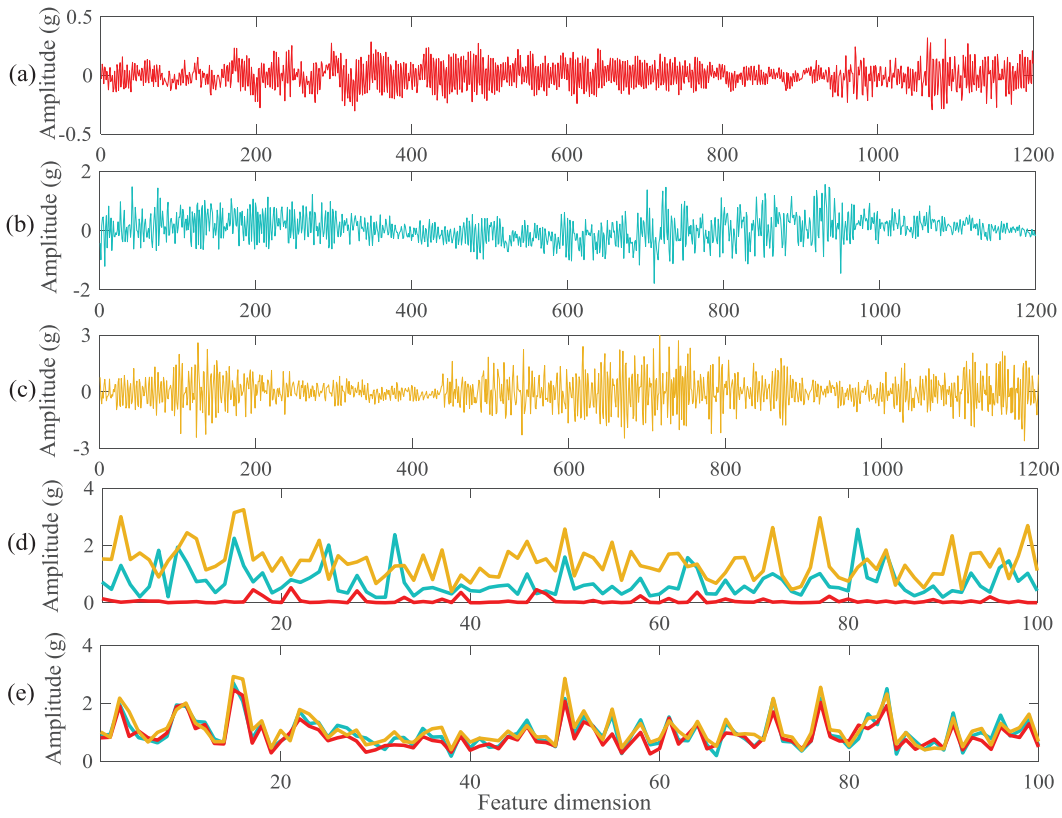


**Figure 18.** Feature vectors of three WW fault samples under different rotating speeds: (a) 700 r min$^{-1}$, (b) 1100 r min$^{-1}$, (c) 1500 r min$^{-1}$, (d) learned features of the raw AE method, (e) learned features of the proposed method.

diagnosis results of 15 trials by BNAE and raw AE networks displayed in table 2. The average training and testing accuracies of 15 trials by the AE network are only 85.65% and 70.12%, respectively. In contrast, the accuracies of the BNAE network are 100% and 99.23%, respectively, which indicates that the proposed BNAE network can indeed deal with the fault diagnosis problem under rotating speed fluctuation condition.

To detail the classification results of each health condition, we also plot the confusion matrices of the testing accuracies for the gearbox data set. Figure 14(a) displays the confusion matrix of the proposed BNAE method, and it can be seen that only several samples of three health conditions are misclassified. In contrast, the confusion matrix of the raw AE method displayed in figure 14(b) is quite terrible. It can be seen that raw AE misclassifies 47.5% of the samples of WP as PC and PW, 89.5% of the samples of WW as other health conditions, 48% of the samples of WP as WP, and also only 51.5% of the samples of PW are correctly classified. Here, we also randomly select 12 weight vectors of the two trained models, as shown in figure 15. It can be seen that the weight vectors trained by the BNAE network are time-localized and have narrow spectral bandwidth, but the raw AE network still performs badly. As shown in figure 16, the training error and misclassification rate by the proposed method all get close to 0 with 20 epochs. Unfortunately, the results of the raw AE network are extremely poor, as before. In addition, the feature visualization result by the BNAE method is displayed in figure 17(a). It can be clearly observed that all the fault types are distinguished perfectly and the same health condition samples are very well clustered. In figure 17(b), the features of most health conditions are overlapped seriously and do not cluster well, which matches with the confusion matrix displayed in figure 14(b). From the foregoing, we can come to the conclusion that the proposed BNAE model is quite able to handle the fault diagnosis problem under speed fluctuation with the help of batch normalization.

Furthermore, in order to illustrate the ability of batch normalization for handling the fault diagnosis problem under speed fluctuation, we randomly select three WW fault samples under three different rotating speeds (700, 1100 and 1500 r min$^{-1}$), as shown in figures 18(a)–(c). It can be seen that the three samples have different amplitudes and frequencies due to the effect of different rotating speeds. Then, we input these samples into the two networks and obtain 100 dimensional feature vectors, as shown in figures 18(c) and (d). Figure 18(c) displays the learned features by the raw AE method. It is easy to find that the three feature vectors still have different amplitudes and are chaotic with each other, which is hopeless for fault classification. That is the reason the raw AE performs so terribly in gearbox fault diagnosis. By contrast, due to the scale and shift abilities of batch normalization on normalizing input values, we find that the learned feature vectors obtained by the BNAE network exhibit almost the same trend as shown in figure 18(d). To be more specific, the batch normalization technique can shrink the amplitude of high-speed samples and enlarge the amplitude of low-speed samples, and can also shift the same feature points in different

locations to cluster together. Therefore, we can understand why the BNAE method can deal well with the fault diagnosis problem under rotating speed fluctuation.

## 5. Conclusions

A BNAE network is presented for bearing and gearbox fault diagnosis. Within this framework, batch normalization is employed in each layer of autoencoder to improve the training process of the network. The model can directly extract meaningful features from raw vibration signals and reduce the internal covariate shift problem in the network. In particular, it is able to deal with the fault diagnosis problem under rotating speed fluctuation. Experimental studies manifest that the proposed BNAE network outperforms the raw autoencoder network and some other traditional methods. Furthermore, three same health condition samples under different rotating speeds are employed to illustrate the scale and shift abilities of batch normalization on normalizing input values.

In this study, the training process still needs fine-tuning by labeled data to obtain the ideal result. Thus, it is desirable to develop an unsupervised learning method to improve it. Meanwhile, it is our first empirical study to achieve intelligent fault diagnosis under rotating speed fluctuation condition. We consider that this problem addressed is interesting and of importance in engineering, and we will keep on investigating this topic in future study.

## Acknowledgments

The research was supported by the National Natural Science Foundation of China (51675262) and the Project of National Key Research and Development Plan of China 'New energy-saving environmental protection agricultural engine development' (2016YFD0700800), the Advanced Research Field Fund Project of China (6140210020102), the Fundamental Research Funds for the Central Universities(NP2018304) and the Major national science and technology projects (2017-IV-0008-0045).

## ORCID iDs

Jinrui Wang https://orcid.org/0000-0001-8690-0672
Shunming Li https://orcid.org/0000-0002-1271-6036
Zenghui An https://orcid.org/0000-0001-8482-5234
Yu Xin https://orcid.org/0000-0003-4358-8396
Weiwei Qian https://orcid.org/0000-0002-1015-7920

## References

[1] Zhao R *et al* 2019 Deep learning and its applications to machine health monitoring *Mech. Syst. Signal Process.* **115** 213–37
[2] Jiang X *et al* 2019 A coarse-to-fine decomposing strategy of VMD for extraction of weak repetitive transients in fault diagnosis of rotating machines *Mech. Syst. Signal Process.* **116** 668–92
[3] Qian W *et al* 2018 An intelligent fault diagnosis framework for raw vibration signals: adaptive overlapping convolutional neural network *Meas. Sci. Technol.* **29** 095009
[4] Wang J *et al* 2018 Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis of machines *Neurocomputing* accepted (https://doi.org/10.1016/j.neucom.2018.10.049)
[5] Qi Y *et al* 2017 Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery *IEEE Access* **99** 1
[6] Jiang H *et al* 2018 Intelligent fault diagnosis of rolling bearing using improved deep recurrent neural network *Meas. Sci. Technol.* **29** 065107
[7] Muralidharan V and Sugumaran V 2012 A comparative study of Naive Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis *Appl. Soft Comput.* **12** 2023–9
[8] Widodo A and Yang B-S 2007 Support vector machine in machine condition monitoring and fault diagnosis *Mech. Syst. Signal Process.* **21** 2560–74
[9] Yan J and Lee J 2005 Degradation assessment and fault modes classification using logistic regression *J. Manuf. Sci. Eng.* **127** 912–4
[10] Feng Z, Liang M and Chu F 2013 Recent advances in time–frequency analysis methods for machinery fault diagnosis: a review with application examples *Mech. Syst. Signal Process.* **38** 165–205
[11] Bengio Y and LeCun Y 2007 Scaling learning algorithms towards AI *Large-Scale Kernel Machines* (Cambridge, MA: MIT Press) pp 321–59
[12] Jia F, Lei Y, Lin J, Zhou X and Lu N 2016 Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data *Mech. Syst. Signal Process.* **72** 303–15
[13] Guo L, Gao H, Huang H, He X and Li S 2016 Multifeatures fusion and nonlinear dimension reduction for intelligent bearing condition monitoring *Shock Vib.* **2016** 1–10
[14] Liu H, Li L and Ma J 2016 Rolling bearing fault diagnosis based on STFT-deep learning and sound signals *Shock Vib.* **2016** 12
[15] Junbo T, Weining L, Juneng A and Xueqian W 2015 Fault diagnosis method study in roller bearing based on wavelet transform and stacked auto-encoder *Control and Decision Conf.* (IEEE) pp 4608–13
[16] Sun W, Shao S, Zhao R, Yan R, Zhang X and Chen X 2016 A sparse auto-encoder-based deep neural network approach for induction motor faults classification *Measurement* **89** 171–8
[17] Zhu H *et al* 2016 Fault diagnosis of hydraulic pump based on stacked autoencoders *IEEE Int. Conf. on Electronic Measurement & Instruments* (IEEE) pp 58–62
[18] Jia F, Lei Y, Guo L, Lin J and Xing S 2017 A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines *Neurocomputing* **272** 619–28
[19] Simon M, Rodner E and Denzler J 2016 ImageNet pre-trained models with batch normalization preprint (arXiv:1612.01452)
[20] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Int. Conf. on Machine Learning* pp 448–56
[21] Littwin E and Wolf L 2016 The loss surface of residual networks: ensembles and the role of batch normalization preprint (arXiv:1611.02525)
[22] Memisevic R and Krueger D 2015 Zero-bias autoencoders and the benefits of co-adapting features *Proc. of Int. Conf. on Learning Representations* pp 1–11
[23] Nair V and Hinton G 2010 Rectified linear units improve restricted Boltzmann machines *Int. Conf. on Machine Learning, Omnipress* pp 807–14
[24] Ngiam J *et al* 2011 Sparse filtering *Int. Conf. on Neural Information Processing Systems, Curran Associates Inc.* pp 1125–33

[25] Jiang M *et al* 2016 Text classification based on deep belief network and softmax regression *Neural Comput. Appl.* pp 1–10
[26] Lou X and Loparo K A 2004 Bearing fault diagnosis based on wavelet transform and fuzzy inference *Mech. Syst. Signal Process.* **18** 1077–95
[27] Li W, Zhang S and He G 2013 Semisupervised distance-preserving selforganizing map for machine-defect detection and classification *IEEE Trans. Instrum. Meas.* **62** 869–79
[28] Du W, Tao J, Li Y and Liu C 2014 Wavelet leaders multifractal features-based fault diagnosis of rotating mechanism *Mech. Syst. Signal Process.* **43** 57–75
[29] Lin J and Chen Q 2013 Fault diagnosis of rolling bearings based on multifractal detrended fluctuation analysis and Mahalanobis distance criterion *Mech. Syst. Signal Process.* **38** 515–33
[30] Xu Z *et al* 2016 A selective fuzzy ARTMAP ensemble and its application to the fault diagnosis of rolling element bearing *Neurocomputing* **182** 25–35
[31] Jia F *et al* 2018 Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization *Mech. Syst. Signal Process.* **110** 349–67
[32] Maaten L and Hinton G 2008 Visualizing data using *t*-SNE *J. Mach. Learn. Res.* **9** 2579–605