Signal Processing 130 (2017) 377–388

Contents lists available at ScienceDirect

Signal Processing

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

# Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification

Chen Lu [a,b], Zhen-Ya Wang [a,b], Wei-Li Qin [a], Jian Ma [a,b,*]

[a] School of Reliability and Systems Engineering, Beihang University, Xueyuan Road, Haidian District, Beijing China
[b] Science & Technology on Reliability & Environmental Engineering Laboratory, Beihang University, Xueyuan Road, Haidian District, Beijing China

ARTICLE INFO

*Article history:*
Received 9 March 2016
Received in revised form
22 July 2016
Accepted 25 July 2016
Available online 29 July 2016

*Keywords:*
Fault diagnosis
Stacked denoising autoencoder
Health state identification
Deep learning

ABSTRACT

Effective fault diagnosis has long been a research topic in the prognosis and health management of rotary machinery engineered systems due to the benefits such as safety guarantees, reliability improvements, and economical efficiency. This paper investigates an effective and reliable deep learning method known as stacked denoising autoencoder (SDA), which is shown to be suitable for certain health state identifications for signals containing ambient noise and working condition fluctuations. SDA has become a popular approach to achieve the promised advantages of deep architecture-based robust feature representations. In this paper, the SDA-based fault diagnosis method contains three successive steps: health states are first divided into training and testing groups for the SDA model, a deep hierarchical structure is then established with a transmitting rule of greedy training, layer by layer, where sparsity representation and data destruction are applied to obtain high-order characteristics with better robustness in the iteration learning. Validation data are finally employed to confirm the fault diagnosis results of the SDA, where existing health state identification methods are used for comparison. Rotating machinery datasets are employed to demonstrate the effectiveness of the proposed method.

## 1. Introduction

Fault diagnosis has increased in importance with the increasing demand for electro-mechanical systems, driven mostly by economic, environmental, reliability, and safety incentives [1,2]. For applications in rotary machinery engineering projects, unexpected failures of common components can result in serious loss of safety, property, and customer satisfaction. To possibly eliminate such problems, a performance assessment of system degradation by accurately classifying the current health state from monitored signals is a promising way to improve system reliability. In this case, condition-based monitoring research using sensory information from functioning systems has been proposed and widely applied, which is essential to meeting customer demands regarding up-time, health management and maintenance [3]. Continuous condition monitoring and real-time fault diagnosis play an indispensable role that not only results in detection and diagnosis of fault information in advance of damage but also enables fault prognosis to provide support for crucial decision-making regarding maintenance [4–6]. On this basis, a wide range of practical applications for condition monitoring-based failure diagnosis have been conducted in the literature, including bearings [7–9], pumps [10,11], power transmission systems [12,13], and power [14]. Previous studies have shown that the accuracy of the diagnosis results is, to a large extent, dependent on the extracted fault features usually acquired by time–frequency representations, which allow the impacts due to different types of damage to be detected, and the different visual information and calculated parameters between the impacts to be indexed by the classification methods to judge if the system is in healthy condition [15–18]. Lee analyzed the application of prognosis and health management for rotating machinery, in which the typical methods for feature extraction, fault diagnosis, performance assessment, and degradation prediction are discussed [19]. However, due to higher rotary machinery system complexity and sensory data heterogeneity, the effective diagnosis of multiple health state classifications based on sensory data with strong ambient noise and working condition fluctuations is still a problem and a major challenge for the application of the proposed methodologies in complex engineering systems due to possible information loss and external influences. In this regard, studies based on machine learning techniques and statistical inference techniques have been conducted from multiple aspects to improve the effectiveness of health state classifications, resulting in a number of classic and typical classification methods, such as support vector machines

[*] Corresponding author at: School of Reliability and Systems Engineering, Beihang University, Xueyuan Road, Haidian District, Beijing, China.
*E-mail address:* majian3129@126.com (J. Ma).

http://dx.doi.org/10.1016/j.sigpro.2016.07.028
0165-1684/© 2016 Elsevier B.V. All rights reserved.

(SVM) [20,21], random forest (RF) [22,23], filters [24], and auto-associative neural network (AANN) [25], as well as some state of the art revised techniques for the further application of fault detection and multiple classification tasks [17,26,27]. As a consequence, the most important task in such studies is to effectively learn elemental feature information from complex and heterogeneous signals as indicators and accurately identify different health states based on the learned indexes. However, a problem arises that the capacities of diagnosis algorithms with simple architectures, e.g., one hidden layer neural network, will meet limitations when faced with complex non-linear relationships in fault diagnosis issues [28].

In recent years, deep learning has emerged gradually as another isolated group of research specialized for pattern recognition, which is an effective way to imitate the human brain learning process, and shows great superiority in capturing the representative information from raw data via multiple non-linear transformations [29]. Compared to current shallow machine learning algorithms, deep learning-based methods attempt to model high-level abstractions in data using multiple processing layers with complex structures, resulting in better representations from the point of view of simplifying a learning task from input examples [30]. Based on scientific knowledge in the area of biology, cognitive humanoid autonomous methods with deep-learning-architecture have been proposed and applied over the years [31–35]. As indicated by current studies, one of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised feature learning and hierarchical feature extraction [36]. Therefore, the information important for classification with respect to diagnosis issues are learned automatically, which could reduce the amount of human labor or prior knowledge used in traditional shallow learning methods. In consideration of the similarity between health states of complex rotary machinery components and heterogeneous data in image pattern classification problems with high-dimensionality, deep learning methods may show great potential in system fault diagnosis with respect to the advantage of a dominant training mechanism and deep learning architecture [37]. In addition, deep learning is believed to be able to discover useful high-order feature representations, as well as the relevance of initial signals, which motivates the emergence of promising applications for dealing with diagnosis problems faced during classification tasks with complex and mixed system health states, both effectively and accurately [38]. Recent theoretical studies also explored the view that deep hierarchical architectures may be needed to yield a new point for complex distributions to achieve better and more robust generalization performance in challenging recognition tasks [39,40]. However, although there exists great potential, as well as a crucial need to address these challenges by utilizing the advantages of deep learning techniques, these are still rarely applied in current fault diagnosis research of electromechanical systems.

In this study, a deep learning method based on SDA that guarantees diagnosis accuracy and improves fault pattern classification robustness is proposed with respect to the complex sensory signals and ambient influence, where unsupervised self-learning and data destruction processes are used to achieve better feature representations. The SDA is established based on a series of autoencoders and learns using a deep network architecture in a layer-by-layer fashion. High-order feature representations obtained in this way are supposed to be the input for subsequent fault classifiers. The SDA based fault diagnosis method consists of three consecutive stages: first, health states are divided into separate groups for training and testing. Second, a deep hierarchical structure is established with a transmitting rule of greedy training to map the input data into useful and robust high-order representations. Third, the proposed SDA models are validated using testing datasets. The fault diagnosis accuracy of the proposed deep learning method can be used to form a knowledge base to determine if the approach is applicable for detecting and classifying the health states of complex systems with inevitable interference. Existing health state classification methods, such as SVM and RF are used for comparison. Rotating machinery datasets are employed to justify the effectiveness of the proposed methods.

This paper is organized as follows: in Section 2, a description of the SDA model is presented. Section 3 presents and discusses the results from different applied methodologies, as well as for different experimental conditions. A comparison is also made with the proposed method. Section 4 concludes the paper.

## 2. Fault diagnosis using the SDA-based health state classification

This section details the proposed SDA-based health state classification approach. Section 2.1 overviews the general architecture and the basic learning process of the SDA. Section 2.2 discusses the involved methodologies in unsupervised forward learning, such as sparse representation. Section 2.3 discusses the data destruction process with respect to robustness feature self-learning. Section 2.4 presents the overall stepwise procedure for the employed diagnosis approach.

### 2.1. Basic description

The SDA employs a multilayered architecture comprised of one input layer and multiple hidden layers, as shown in Fig. 1.

The SDA structure is similar to a stacked network of the auto-associative neural networks (AANN). As noted by the name, AANN is a kind of neural network that consists of three layers where the output and input data are requested to be as close to identical as possible, realizing a feature reconstruction as shown in Fig. 1 left. Generally the hidden layer has less neurons, of which output is usually referred to a kind of high-order feature representation for the input data in such models. Thus the idea of greedy layer-wise unsupervised learning is employed to present better feature representations based on level by level AANN models, giving rise to an initial set of parameter values for the first layer, and the output of the first hidden layer is used (a new representation for the raw input) as the input for the next layer, similarly initializing that layer. Notice that the output layer of each AANN is removed in the integration, resulting a series of cascading models named autoencoder. Each successive autoencoder in the deep structure follows the same transformation concept and passes the regularity throughout the SDA architecture as indicated in Ref. [32]. An example of SDA architecture is shown in Fig. 1 right. The SDA consists of three autoencoders, where layer 1 (input layer) and layer 2 (hidden layer of AANN 1) forms the first autoencoder, layer 2 (input layer of AANN 2) and layer 3 (hidden layer of AANN 2) forms the second autoencoder, and layer 3 (input layer of AANN 3) and layer 4 (hidden layer of AANN 3) stand for the third autoencoder. After having initialized a number of layers, the whole SDA model can be fine-tuned based on a supervised training criterion.

### 2.2. Unsupervised fault characteristic self-learning with sparse representation

An SDA model is a straightforward variation on the stacking of ordinary stacked autoencoders as described in Section 2.1. The hierarchical architecture of a deep neural network has been discussed regularly in prior work on supervised learning [41,42]; however, it has been found that the obtained training and generalization errors are extremely poor using the standard random

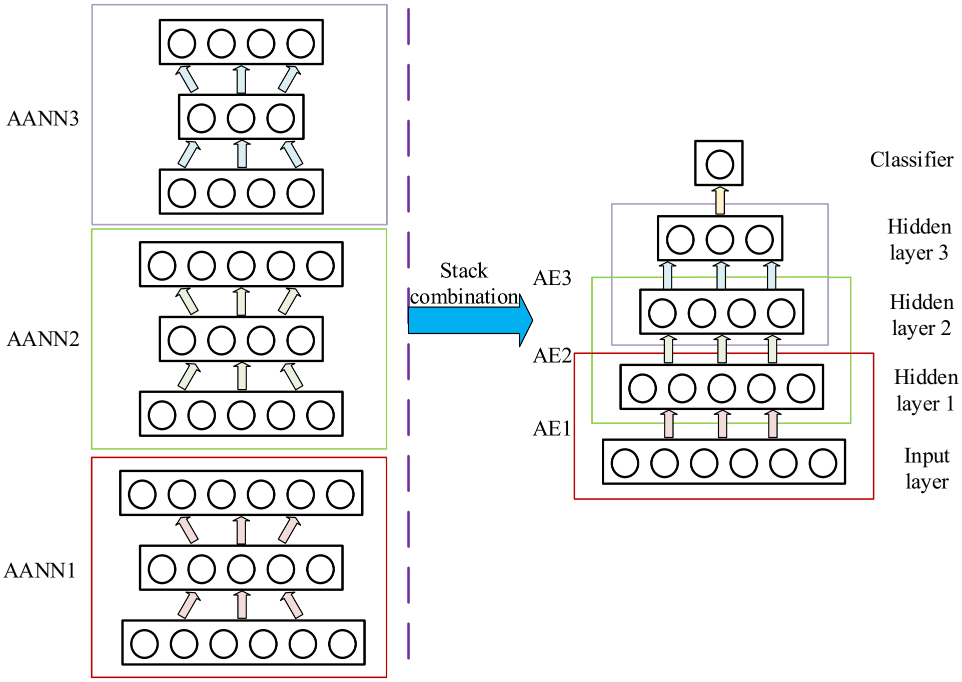

**Fig. 1.** Stacked denoising autoencoder architecture.

initialization, which means that the gradient-based training process is easily stuck in the local minimum, and it is more obvious when a deeper neural network is applied. To solve the problem, Ref. [31] discovers that much better results may be achieved when unsupervised learning algorithms are used in the pre-training of each layer, one layer after the other, starting with the input layer. The basic deep learning architecture is briefly described as follows.

Considering a stacked autoencoder with $l$ layers, the notations $\omega^{(k,l)}$ and $b^{(k,l)}$ denote the weight and bias parameters for the $kth$ autoencoder (the $kth$ hidden layer in the deep architecture), respectively. After the hierarchical structure is initialized, the encoding process for each layer in the forward order is set as below:

$$a^{(k+1)} = f\left(\omega^{(k,l)}x^{(k)} + b^{(k,l)}\right) = \frac{1}{1 + \exp\left(-b^{(k,l)} - \sum \omega^{(k,l)}x^{(k)}\right)} \tag{1}$$

where $x^{(k)}$, $a^{(k+1)}$, $\omega^{(k,1)}$, and $b^{(k,1)}$ denote the input, output, weight matrix, and bias vectors of the $kth$ autoencoder, respectively. $f(.)$ represents the sigmoid transformation with log-likelihood function from one layer to another.

For each autoencoder, the output $a^{(k)}$ is seen as a higher-order representation of the original data achieved by the feature reconstruction process in the AANN model, which is also defined as the decoding stack reverse reconstruction in deep learning. The decoding process is as follows.

$$z = g_{\theta'}(a) = s(\omega^T a + b^T),\ \theta' = \left\{\omega^T, b^T\right\} \tag{2}$$

where $z$, $a$, and $\theta$ denote the output, hidden layer output, and connection parameters of the employed AANN model, respectively. The $s(.)$ stands for the reconstruction function, aiming to enable the output $z$ to be equal to the input data.

Although the activations of the deepest layer contain the main factors of feature information, it is still viewed as a lossy compression of the input data, which means the reconstructed features should be uncorrelated in removing the redundant information. To achieve this goal, a sparsity constraint is applied to the hidden units of the each autoencoder that is used to discover the elemental feature representations in the original data, despite the number of hidden units [43]. For instance, in the case of the sigmoid function, the output of the hidden unit is 1 for activation, whereas 0 for restraint, and only the activated units can be used for feature representation. Given an input sample $x$, the average activation $\rho_j$ of the hidden unit $j$ is defined as:

$$\rho_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_j\left(x^{(i)}\right)\right] \tag{3}$$

where $a_j$ and $m$ refer to the activation of the hidden unit $x^{(i)}$ and the number of input nodes, respectively.

Assuming that $\rho$ is a constant value close to 0, a limitation based on Kullback–Leibler (KL) divergence is introduced to measure the difference between $\rho_j$ and $\rho$ in this paper, where the KL divergence increases monotonically as the diversity becomes greater. An extra penalty term is thus added to the optimization objective that penalizes $\rho_j$ deviating significantly from $\rho$ during the unsupervised learning process, as formulated in Eq. (4).

$$KL\left(\rho\|\rho_j\right) = \rho \log\frac{\rho}{\rho_j} + (1-\rho)\log\frac{1-\rho}{1-\rho_j} \tag{4}$$

where $\rho$ and $\rho_j$ stand for the sparse coefficient and average activation, respectively.

To realize the minimization of the reconstruction error, a cost function with respect to one single autoencoder is shown in Eq. (5), given the feature reconstruction function $f(.)$.

$$C(\omega, b; x, a) = \frac{1}{2}\|z - x\|^2 = \frac{1}{2}\|h_{\omega,b}(x) - x\|^2 \tag{5}$$

where $C(W, b; x, a)$ is a one-half squared-error cost function. On this basis, assuming a training set of $m$ input samples, the overall cost function of the deep architecture with $n$ layers is set to be:

$$C(\omega, b) = \left[\frac{1}{m}\sum_{i=1}^{m} C\left(\omega, b; x^{(i)}, a^{(i)}\right)\right] + \beta\sum_{j=1}^{s} KL\left(\rho\|\rho_j\right) + \frac{\lambda}{2}\sum_{l=1}^{n-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(\omega_{ji}^{(l)}\right)^2 = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h_{\omega,b}\left(x^{(i)}\right) - x^{(i)}\|^2\right)\right] + \beta\left[\rho\log\frac{\rho}{\rho_j} + (1-\rho)\log\frac{1-\rho}{1-\rho_j}\right] + \frac{\lambda}{2}\sum_{l=1}^{n-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(\omega_{ji}^{(l)}\right)^2 \tag{6}$$

where $C(\omega, b)$ is an average sum-of-squares error, in which $\lambda$ is a regularization term to help prevent overfitting by decreasing the magnitude of the weights. The weight $\beta$ determines the sparsity proportion employed in the sparse representation process. $\omega_{ji}^{(l)}$ represents the synaptic weight between the *ith* neuron in layer $l$ and *jth* neuron in layer $l+1$, and $s_l$ denotes the number of the total neurons in layer $l$. This greedy layer-wise procedure has been shown to yield significantly better local minima than random initialization of deep networks by training each autoencoder in sequence and employing the average squared-error as the evaluation indicator, thus achieving better generalization in a number of tasks.

### 2.3. Data destruction for self-learning of robustness features

A problem usually arises in the self-learning process when the ambient noise that cannot be ignored is mixed within the dynamic vibration signals, which is hard to deal with manually because of the large number of samples to be trained. In this sense, a new mechanism of stacked learning is applied in the SDA model to achieve feature representations with better robustness when the input is stochastically corrupted. However, the original input is still used as the target for the reconstruction in the deep learning process [40]. The model trained in this way can be used such that the feature self-learning process not only tries to encode the input but also captures the statistical structure by approximately maximizing the likelihood. In other words, the corruption part is supposed to be represented by capturing the statistical dependencies between the inputs. Thus, the effect of external interferences is diminished by the proposed data deconstruction process to obtain better feature representations. Taking one autoencoder as an example, the data destruction method undertaken in this study is described as follows.

Let $q^0(x)$ be the joint distribution function concerned with input samples:

$$q^0(x, \tilde{x}, a) = q^0(x) qD\left(\tilde{x}|x\right)\delta_{f_\theta(\tilde{x})}(a), \tag{7}$$

where $x$ and $\tilde{x}$ denote the initial and corrupted input data, respectively and $a$ is the deterministic function of $\tilde{x}$. Assuming that $\mu$ is defined as $\mu = f_\theta\left(\tilde{x}\right)$ and that $\upsilon$ is the destruction level, the units $\delta_\mu(a)$ are forced to be 0 when $\mu \neq \upsilon$, i.e., $\tilde{x}$ is achieved by means of a stochastic mapping of $\tilde{x}{\sim}qD\left(\tilde{x}|x\right)$. The autoencoder is thus applied for the following feature reconstruction based on the unsupervised learning process mentioned above. Note that the data destruction process is conducted in all layers of the SDA model instead of only the input layer. The destruction process of each layer during signal processing is shown in Fig. 2.

From this, the encoder and decoder processes described in Section 2.2 in the SDA based deep learning are modified, as shown in Eqs. (8) and (9), respectively:

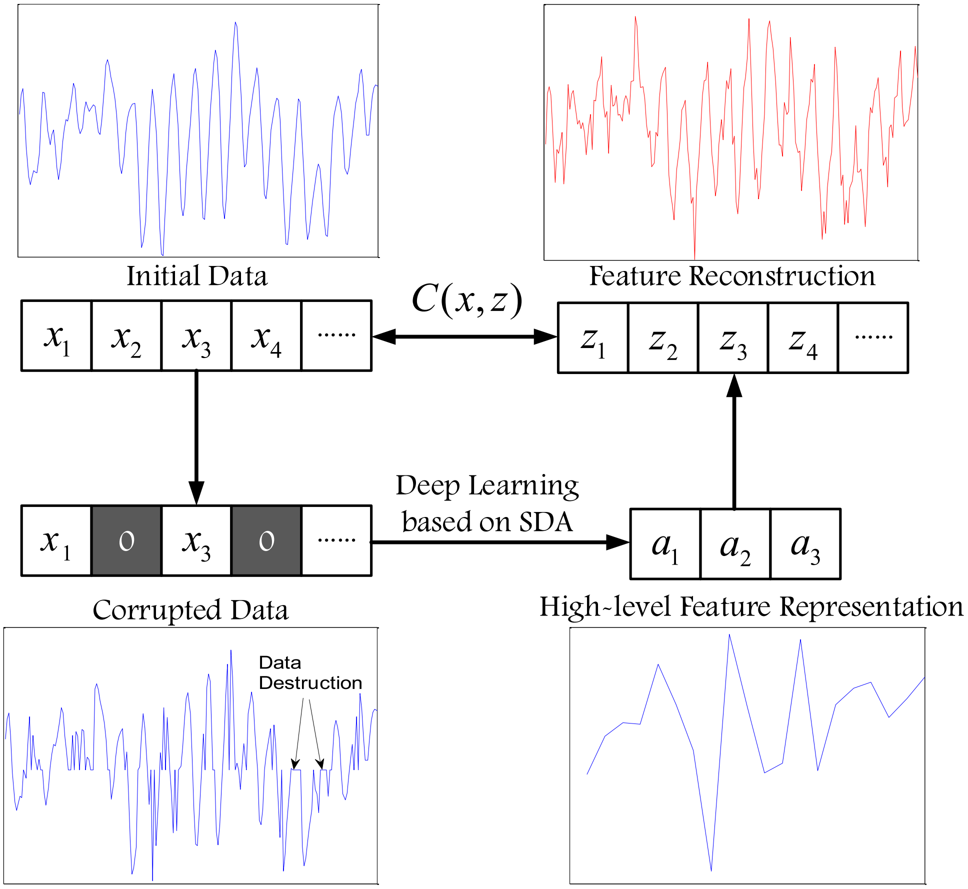
Initial Data
Feature Reconstruction
$C(x,z)$
$x_1$ $x_2$ $x_3$ $x_4$ ……
$z_1$ $z_2$ $z_3$ $z_4$ ……
Deep Learning based on SDA
$x_1$ 0 $x_3$ 0 ……
$a_1$ $a_2$ $a_3$
Corrupted Data
High-level Feature Representation
Data Destruction

**Fig. 2.** Destruction process for the input data in each layer.

$$a = f(\tilde{x}) = f(\omega\tilde{x} + b),\ \theta = \left\{\omega,\ b\right\}, \tag{8}$$

$$z = g_{\theta'}(a) = s(\omega^{T}a + b^{T}),\ \theta' = \left\{\omega^{T},\ b^{T}\right\}, \tag{9}$$

where $\tilde{x}$ denotes the corrupted input, and all remaining parameters are the same as described in Eqs. (1) and (2).

Note that the reconstructed feature $z$ is a deterministic function of $\tilde{x}$ instead of the original input $x$. However, the uncorrupted input $x$ is still the target of feature reconstruction. The cost function is modified based on Eq. (3) to consider the destruction process:

$$\begin{aligned}\theta,\ \theta' &= \underset{\theta,\theta'}{\arg\min}\frac{1}{m}\sum_{i=1}^{m}C\left(x^{(i)},\ z^{(i)}\right)\\ &= \underset{\theta,\theta'}{\arg\min}\frac{1}{m}\sum_{i=1}^{m}C\left(x^{(i)},\ s_{\theta'}\left(f_{\theta}\left(\tilde{x}^{(i)}\right)\right)\right),\end{aligned} \tag{10}$$

where $\theta$ and $\theta'$ denote the optimized weight and bias vectors in the forward order and reconstruction process that minimize the difference between the feature reconstruction representations and the corrupted input data, whereas $m$, $f_{\theta}(.\ )$, and $s_{\theta'}(.\ )$ stand for the number of layers, the mapping function, and the reconstruction function, respectively.

As described above, it is suggested that the SDA applied to a fault diagnosis corresponds to a generative model, in which layer nodes represent the random variables and arcs of the input that indicate something about the type of dependency existing between the random feature variables. Several possible generative models are also discussed in [44]. We note that the corruption process is only used during the training process, but not for propagating representations, which is discussed in the following section.

### 2.4. Back propagation learning based fine tuning

The parameters derived from multiple hidden layers are determined to be the input of a supervised classifier followed by the global back propagation optimization process.

In this study, the softmax regression algorithm is employed for multi-class classification [45]. We suppose the training samples are $(x^{(k)}, y^{(k)})$, and the label $y^{(k)}$ is treated as the training target for supervised optimization learning. Given an input $x$, the classification probability $p\left(y = j \mid x\right)$ achieved by the softmax regression algorithm for each category $j(j = 1, \ldots, k)$ can be expressed based on the following hypothesis function:

$$J_{\theta}(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1\left|x^{(i)};\ \theta\right.) \\ p(y^{(i)} = 2\left|x^{(i)};\ \theta\right.) \\ \vdots \\ p(y^{(i)} = k\left|x^{(i)};\ \theta\right.) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k}e^{\theta_j^{T}x^{(i)}}}\begin{bmatrix} e^{\theta_1^{T}x^{(i)}} \\ e^{\theta_2^{T}x^{(i)}} \\ \vdots \\ e^{\theta_k^{T}x^{(i)}} \end{bmatrix}, \tag{11}$$

where $\theta_1, \theta_2, \cdots\theta_k \in \Re^{n}$ refer to the model parameters.

To further improve the classification performance, a fine-tuning process using the back propagation algorithm is applied on the basis of traditional pattern recognition, where the parameters of the SDA model are updated to minimize the training error [32]. In this study, the fine tuning employed in the trained SDA is a global parameter adjustment in consideration of the certain optimization algorithms such as the conjugate gradient approach instead of traditional gradient method, aiming to achieve better computational efficiency and stability [31,46]. The procedure of the back propagation algorithm is listed in Table 1.

The SDA model is established based on some key parameters, including the batch size, number of layers and hidden neurons, data destruction level, sparsity proportion, and the maximum number of training epochs for the training process. The successive learning process is shown in Fig. 3.

**Table 1**
Back propagation learning procedure for the SDA model.

| | |
|---|---|
| Step 1 | Calculating the outputs of the layers in the forward direction from the first hidden layer to the output layer |
| Step 2 | Calculating the residual error of each unit in the output layer |
| Step 3 | Calculating the residual error of each unit in the hidden layers from back to front |
| Step 4 | Calculating the expected partial derivatives of the corresponding cost function |
| Step 5 | Updating the residual errors of each layer based on the partial derivatives |
| Step 6 | Updating the initial weights and bias using gradient descent. |
| Step 7 | Tuning based on the conjugate gradient approach |

To address the health state identification problems effectively with the SDA, the first step is to identify the data format and diagnosis targets. Pre-processing is then conducted as the second step based on the data requirements to ensure both the diagnosis accuracy and the computational efficiency. Pre-processed data are divided into training and testing groups consisting of a series of batches for forward and back propagation learning. The architecture parameters of the SDA are initialized with small random values, and distinct updates are calculated and integrated as diverse parts of the full deep network. The number of loops for the unsupervised learning and back propagation process is determined by the maximum number of epochs set for the model. The SDA is finally verified based on the misclassification error, which is defined as the ratio of the number of misclassified health states to the total number of diagnosis results.

In summary, the main contribution of this paper is proposing a new autonomic feature extraction based on the deep learning theory to address the diagnosis issues of rotary machinery components. Differing from hand-engineered feature extraction and noise-reduction algorithms, the unsupervised learning in the SDA model helps the salient fault characteristic mining for effective signal processing, and the ‘destruction level’ in the training process of SDA is introduced by a certain distribution of noise to receptive input nodes, which indeed improves the representation robustness. The proposed SDA model pays close attention to improve the diagnosis accuracy through feature self-learning, which has the potential to overcome the aforementioned deficiencies in the current diagnosis methods. In addition, higher diagnosis accuracy and stability are expected to be achieved for the promised advantages of destruction process and the ability to establish a richer network structure.

## 3. Health identification applications

Rolling element bearings are key components in rotary machinery components, and their health state conditions often have an important influence on the performance and reliability of the machinery they are installed in. Thus, the proposed health state identification approach is used in rolling bearing case studies with different degrees of ambient noise and working condition fluctuations as examples. The performance of fault pattern classification using the proposed SDA approach is compared with existing diagnosis algorithms, with the detailed results listed below. Rolling bearing datasets employed in the experiment are obtained from the Bearing Data Center of Case Western Reserve University for testing and verification [47]. The SDA model was established based on a deep learning toolbox as indicated in Ref. [48]. All of the experiments were conducted using the Matlab environment.

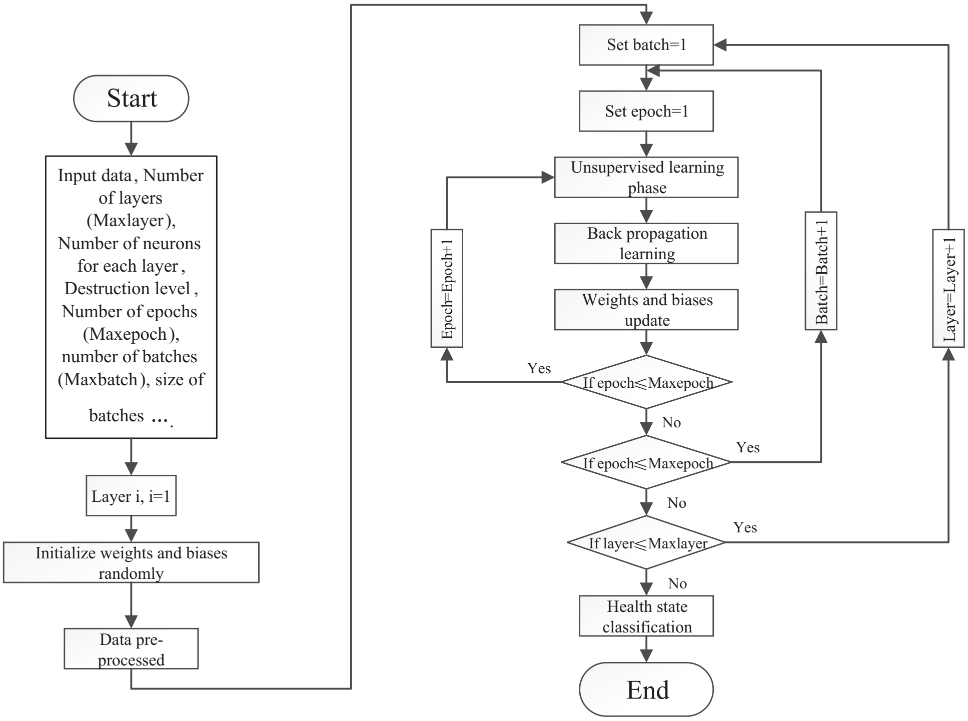
Start
Input data, Number of layers (Maxlayer), Number of neurons for each layer, Destruction level, Number of epochs (Maxepoch), number of batches (Maxbatch), size of batches ….
Layer i, i=1
Initialize weights and biases randomly
Data pre-processed
Set batch=1
Set epoch=1
Unsupervised learning phase
Back propagation learning
Weights and biases update
If epoch≤Maxepoch
Yes
Epoch=Epoch+1
No
If epoch≤Maxepoch
Yes
Batch=Batch+1
No
If layer≤Maxlayer
Yes
Layer=Layer+1
No
Health state classification
End

**Fig. 3.** Successive learning process for the SDA based fault diagnosis.

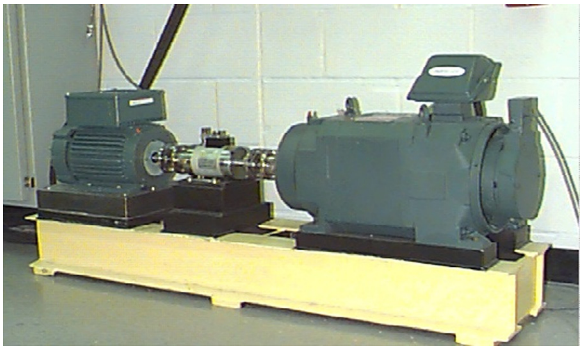

**Fig. 4.** Bearing test rig used for the experiment.

### 3.1. Case study description

The datasets are composed of multivariate vibration series generated by a bearing test-rig, as shown in Fig. 4. The test-rig consists of a 2-horsepower (hp) motor used as the prime mover to drive a shaft coupled with a bearing housing, a control circuit for speed control and to test the bearing at various speeds, a dynamometer and a torque converter / encoder for signal transformation and measurement.

In this study, the employed vibration data were collected using accelerometers attached to the housing with magnetic bases and installed at the 12 o'clock position for the DE bearings. For each test case, the dataset is assigned to one corresponding operating condition based on its operational setting speed. The complete dataset for each working condition is defined by a different unit ID, and the initial health condition for each unit is different and contaminated with measurement noise. For data acquisition, a sampling rate of 48 kHz was used for the bearing faults. Single point faults were introduced into the test bearings using electro-discharge machining with defect diameter of 7 mils and fault depth of 0.011 in. Detailed information on the structure of the bearings is listed in Table 2.

**Table 2**
DE Bearing information.

| Type | Inside diameter | Outside diameter | Thickness | Ball diameter | Pitch diameter |
|---|---|---|---|---|---|
| 6203–2RS JEM SKF | 0.9843 | 2.0472 | 0.5906 | 0.3126 | 1.537 |

**Table 3**
DE bearing working conditions.

| Operating condition | Load of Moto/HP | Rotating Speed/rpm |
|---|---|---|
| C1 | 0 | 1797 |
| C2 | 1 | 1772 |
| C3 | 2 | 1750 |
| C4 | 3 | 1730 |

### 3.2. Data preprocessing and health state definition

Four different operating conditions denoted as C1 to C4 were used, as shown in Table 3, which represent training and testing samples of different rotating speeds. The DE bearing data of the normal (ID 97, 98, 99, and 100), inner race fault (ID 105, 106, 107, and 108), outer race fault (ID 130, 131, 132, and 133), and rolling element fault (ID 118, 119, 120, and 121) conditions were acquired for classification. The time series datasets were divided into four health conditions based on the proximity to the corresponding failure state.

Two experiments were conducted to verify the effectiveness of the SDA for health state classification in the presence of working condition fluctuations. First, the training and testing datasets were both taken from the same rotating speed of 1797 rpm for a single working condition diagnosis. In the second experiment, the SDA is trained using the datasets at 1797 rpm and tested using the three other operating condition datasets. The training and testing data are listed in Table 4.

### 3.3. The SDA model establishment

The establishment of an optimal SDA model has an impact on the diagnosis effects. Considering the stack characteristics of the deep learning process, Ref. [49] sheds light on several key parameters that have an important influence on the unsupervised learning process. Some experiments on FE bearings at 1797 rpm were undertaken to analyze the effect of the changes in such parameters in the SDA model setup. In this particular section, receptive input size, number of hidden layers and units, sparsity proportion, and data destruction level were tested to determine the optimal values in the SDA model. Reconstruction errors of the autoencoders are set to be the indicators for judging the employed parameters, which could be calculated based on the cost function described in Section 2.2.

#### 3.3.1. Receptive input size

Traditionally, better representation could be achieved with larger input data. However, as is often the case, the receptive input size needs to be set based on computational constraints. To observe the changing influence of the receptive input size for each autoencoder, the experiment was first conducted based on the first autoencoder, as shown in Fig. 5(a). The reconstruction error clearly decreases when the larger receptive input size ranged from 20 to 180. However, the subsequent values are rather stable and less than 0.1 especially when the receptive input size is greater than 200, meaning that the employed input data size should cover enough information, as well as ensuring calculation efficiency for bearing diagnosis issues.

**Table 4**
Training and testing data for each cross-validation of the bearing experiments.

| | Operating condition | Normal | Inner Ring | Outer Ring | Ball |
|---|---|---|---|---|---|
| Experiment 1 | | | | | |
| Training Samples | C1 | 195,150 | 97,012 | 97,592 | 98,057 |
| Testing Samples | C1 | 48,788 | 24,253 | 24,399 | 24,514 |
| Experiment 2 | | | | | |
| Training Samples | C1 | 243,938 | 121,265 | 121,991 | 122,571 |
| Testing Samples | C2 | 483,903 | 121,991 | 122,426 | 121,410 |
| | C3 | 485,063 | 122,136 | 121,410 | 121,556 |
| | C4 | 485,643 | 122,917 | 122,571 | 121,556 |

#### 3.3.2. Number of hidden layers and nodes

The number of hidden nodes is normally of great importance to achieving high performance in the SDA model, as well as the depth of the deep architecture. Therefore, taking the first autoencoder as an example, an experiment is employed in which the receptive input size was set to 200 based on the experiment above to determine the appropriate hidden layer parameters by analyzing the influence on the reconstruction results, as illustrated in Fig. 5(b). It is noted that the feature representation results seem to be satisfactory when the number of units in the second layer is less than half that in the previous layer, which was treated as a type of dimension reduction. Taking all of the above results together, further experiments of multiple hidden layers were conducted to demonstrate the expansibility and credibility of the influence of the parameters for the deep learning structure. The best three results for each experimental condition are listed in Table 5.

In Table 5, we show that a considerable improvement in performance can be found with fewer nodes in the next layer if the number of hidden layers remains unchanged, e.g., when the first and the second layers consisted of 100 and 50 nodes, respectively, a much better result of 0.035 is obtained compared to the other experiments with two hidden layers. However, we also noted that the changing trend with a different number of hidden layers is not very distinct when the number of hidden layers was greater than 3. For example, the best results with 3, 4, 5, and 6 hidden layers are all approximately 0.029, meaning that such hidden layer numbers are capable of mining salient bearing feature representations. Thus, to ensure quicker calculation speeds and consume fewer computation resources, three hidden layers with 100, 50, and 25 hidden units are employed.

#### 3.3.3. Sparsity criterion

Sparsity representation plays an important role in unsupervised forward learning to obtain the main factors in the initial information. A suitable sparsity criterion could not only improve the forward learning ability in view of less information loss but also enhance the computing efficiency. The experimental results are displayed in Fig. 5(c). The reconstruction error fluctuates over a small scope when the sparsity criterion is greater than 0.10, whereas poor results are achieved for smaller values. This is mainly because demanding a sparsity criterion could result in difficulties in the unsupervised learning being able to capture enough information associated with the accompanying

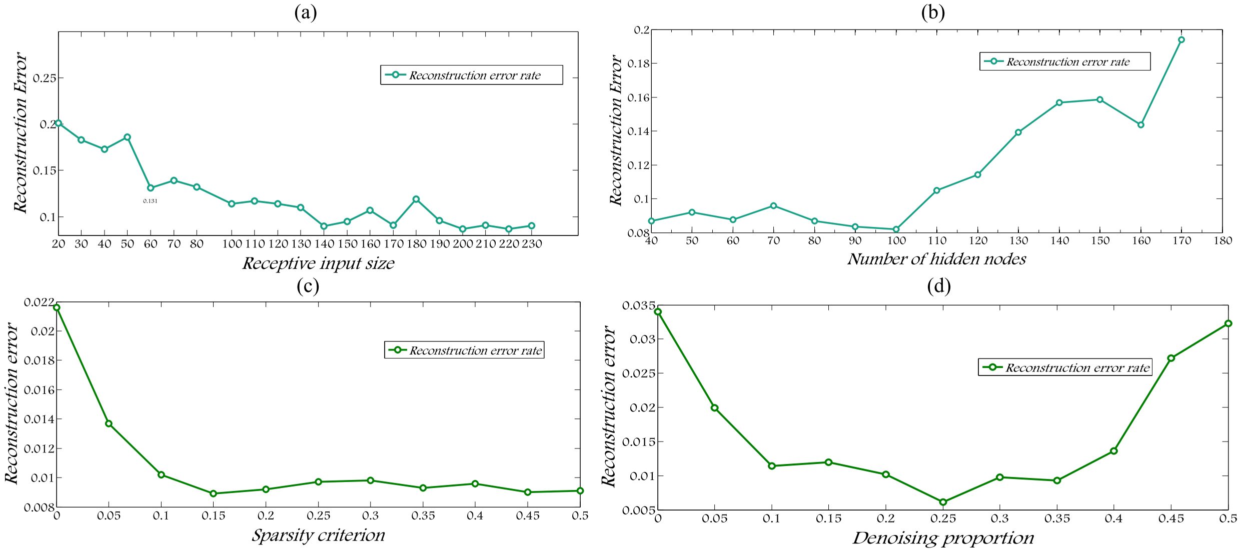

**Fig. 5.** Reconstruction error curves for different SDA model parameters: (a) receptive input data, (b) number of hidden nodes, (c) sparsity criterion, (d) data destruction level.

destruction process. In this respect, the sparsity criterion was relaxed and set to 0.15 for the bearing experiment.

### 3.3.4. Destruction level

The characteristic of destruction process is mining salient feature information by reconstructing the initial data from the destroyed part based on multilayer deep learning. However, a limitation seems to exist because an excessive destruction level may bring about the loss of useful information, resulting in difficulties during the reconstruction process. The changing trend with different destruction levels is shown in Fig. 5(d). As the destruction level varies from 0% to 20% during the initial stage, the reconstruction error clearly decreases, with a result of 0.0102. The reconstruction error shows little variation when the level of data destruction continues to increase, reaching an error of approximately 0.005–0.01. However, as the destruction proportion exceeds 35%, a clear upward trend appears and the result finally decreases to 0.0323 at a destruction level of 50%. A suggested explanation is that excessive feature information may be lost when the destroyed proportion is large, meaning that the remaining information is not robust enough to represent the initial data. In other words, the cost function used for feature reconstruction cannot converge to an expected minimum even though the noise immunity is better.

Based on the above experiments, the detailed SDA information is as follows. The input and output layers were constructed with 200 and four neurons denoting the input parameters and target fault categories, respectively. There were three hidden layers, with 100, 50, and 20 neurons in each layer for deep architecture based feature self-learning, where the transfer function was sigmoid with sparsity restrictions. The key parameters of the SDA are listed in Table 6.

### 3.4. SDA based fault state diagnosis

As stated above, the input signal was reconstructed into high-order representations via self-learning with multiple hidden layers, where destruction process was employed to improve the learning robustness. In general, the fault characteristic self-learning could also be regarded as a type of dimension reduction process, such as a principle component analysis (PCA). Ref. [29] compares the features learned by the deep learning method to those extracted by the PCA in terms of image recognition, demonstrating the effectiveness of high-level feature representation process with deep architecture. Similarly, a comparative experiment was conducted using complex non-linear bearing data as well.

In this particular section, 500 batches are randomly selected from the 1797 rpm bearing data with 15 dB noise injection. The size of each batch is 200, based on the receptive input data. The SDA, SAE, and PCA are then employed to map the input data to two-dimensional feature representations. A clustering analysis is conducted to determine the feature learning results of each method, as displayed in Fig. 6.

As shown in Fig. 6, the principal components of the different health states extracted by PCA are more mixed compared to the deep learning methods, which demonstrates that hand-engineered feature extraction methods have difficulty handing non-linear complex signals with respect to strong ambient noise. It is also noted that the high-level features of normal states and rolling element fault states are also mixed to some extent in the SAE model, mainly because the ambient noise brings about unexpected data fluctuations that result in meaningless data obfuscation of the two states. However, the SDA model performs relatively better, mainly because of the ability of the deep architecture based learning process to learn elemental feature representations to classify the similar datasets preferably, as well as the data destruction process for high robustness self-learning.

A further experiment was also conducted to investigate the learned salient fault characteristics in the SDA model, as illustrated in Fig. 7. The left and right columns indicate the initial signals and input data with 15 dB SNR in one iteration, respectively. We observe that the deep neural network with data destruction can still effectively perceive useful feature information (e.g., the marked fluctuations in Fig. 7) with strong noise, mainly due to the revised deep learning mechanism that is well suited to capture the main variations in a higher order space, such as the manifold [40].

The softmax regression algorithm was used as the top classifier of the SDA model in this study, where the sequences (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1) are labels for normal, inner race fault, outer race fault, and rolling element fault, respectively.

Similarly, the structures and parameters of the comparative models are described as follows. The SAE was an elementary stacked autoencoder model whose parameters were the same as those of the SDA except that the data destruction level was 0 to investigate the effectiveness of the data destruction process. The AE was the first autoencoder used in the deep learning architecture of the SDA, and it was trained in a nearly identical manner as the AANN. The RF was a type of ensemble learning method, where the bootstrap method is applied to improve the generalization ability. In this experiment the RF model consisted of 300 classification trees, each of which was built based on twenty percent of the input data amount. The minimum number of leaf nodes was set to 20. Normally, the diagnosis results are better with a larger number of trees. In the SVM model, the selection of the kernel function parameter and error penalty factor affected the precision significantly. A Gaussian kernel function was applied to

**Table 5**
Experimental result comparison with different hidden layer parameters.

| Receptive input size | Number of hidden layers | Number of hidden nodes | Reconstruction error |
|---|---|---|---|
| 200 | 2 | 200;100 | 0.0406 |
| 200 | 2 | 100;100 | 0.0428 |
| 200 | 2 | 100;50 | 0.035 |
| 200 | 3 | 100;100;50 | 0.0392 |
| 200 | 3 | 100;50;50 | 0.0331 |
| 200 | 3 | 100;50;25 | 0.0291 |
| 200 | 4 | 100;100;50;50 | 0.032 |
| 200 | 4 | 100;50;50;25 | 0.0301 |
| 200 | 4 | 100;50;25;15 | 0.0291 |
| 200 | 5 | 100;100;100;50;50 | 0.0319 |
| 200 | 5 | 100;50;50;25;15 | 0.0295 |
| 200 | 5 | 100;50;25;25;15 | 0.0289 |
| 200 | 6 | 100;100;50;50;25;25 | 0.0308 |
| 200 | 6 | 100;50;50;50;25;15 | 0.0300 |
| 200 | 6 | 100;50;25;25;15;10 | 0.0290 |

**Table 6**
The SDA classifier model parameters.

| | | | | | | |
|---|---|---|---|---|---|---|
| Structure parameters | Input layer neurons<br>200 | Hidden Layer 1<br>100 | Hidden layer 2<br>50 | Hidden layer 3<br>20 | Output layer<br>4 | Transfer function<br>Sigmoid |
| Learning parameters | Batch size<br>100 | Destruction level<br>0.25 | Data interval<br>2 | Sparsity criterion<br>0.15 | Epoch number<br>300 | Weight reduction factor<br>0.003 |

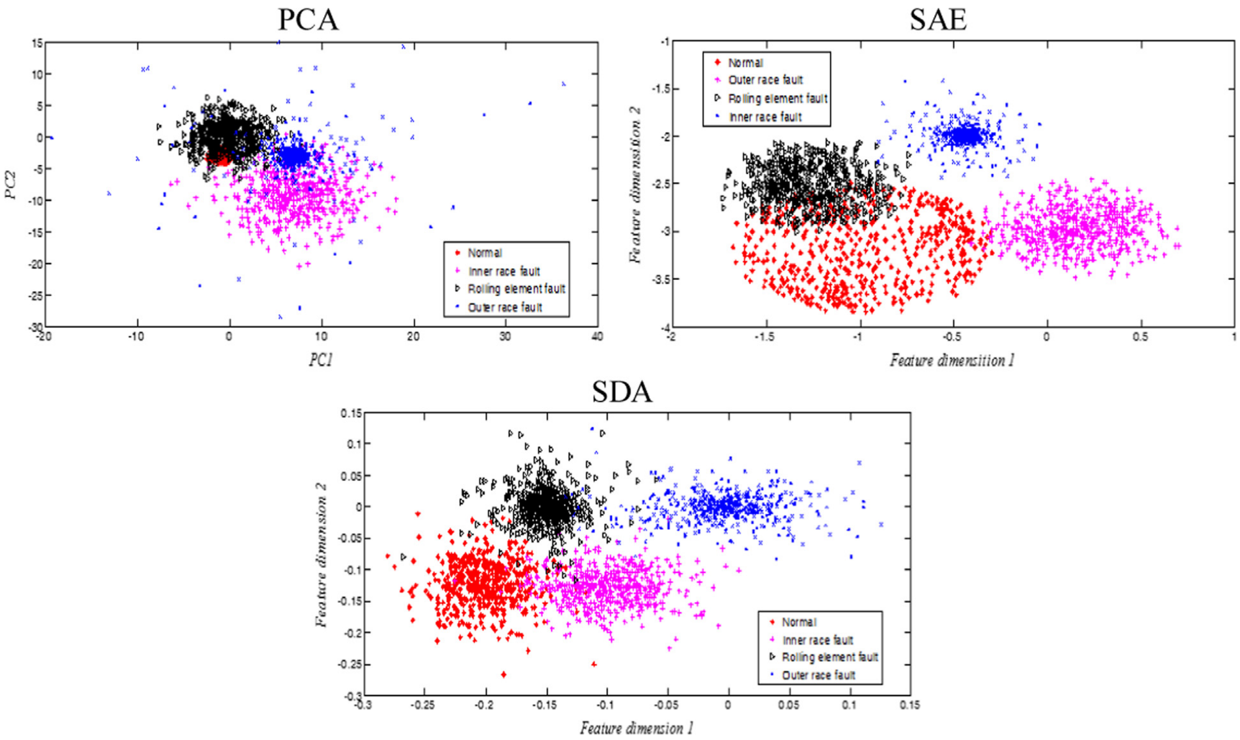

**Fig. 6.** Scatter plots of feature vectors of the comparative methods.

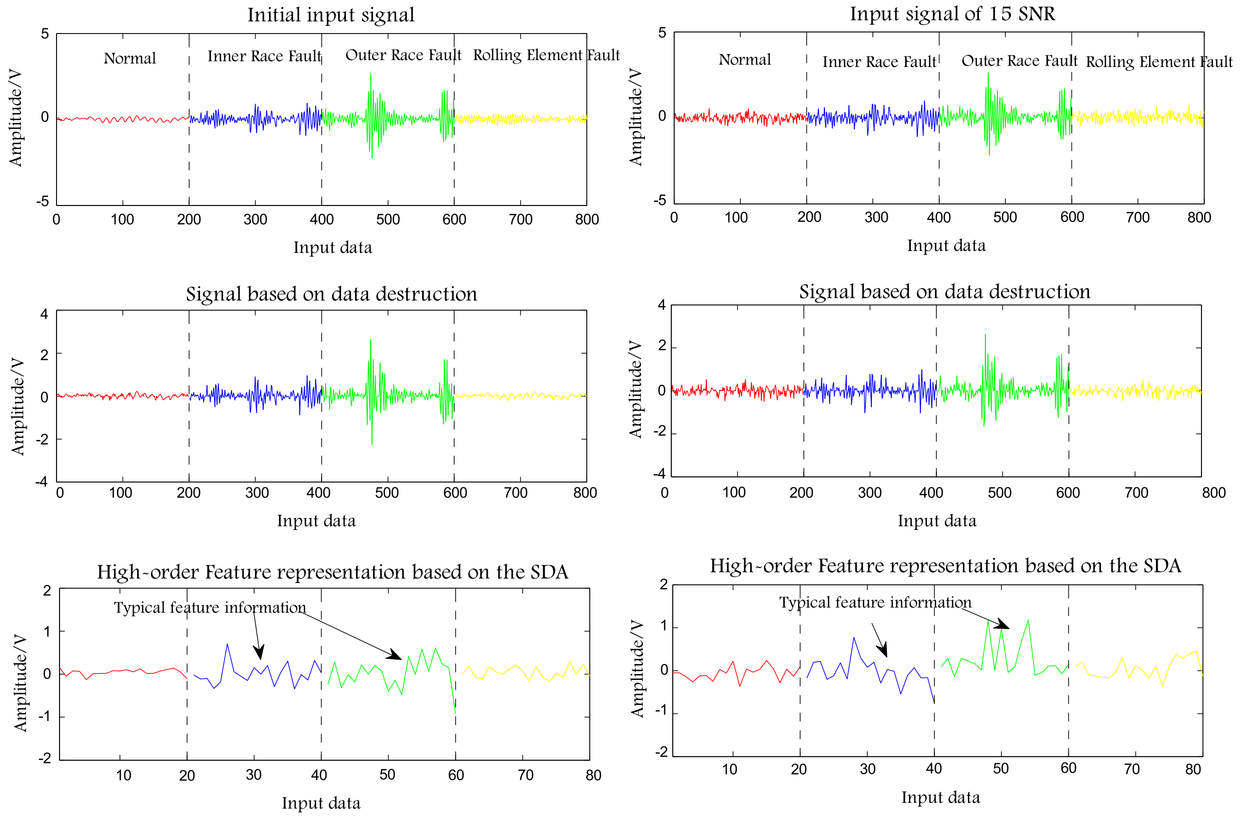

**Fig. 7.** High order feature representation based on the SDA for one iteration.

the SVM classification model, and the error penalty factor was set to 512 to achieve better results, as well as reduce overfitting. The SVM model was trained based on the one-versus-one criterion. Refs. [50,51] discuss the practical parameter selection of the comparative methods. Note that empirical mode decomposition was employed for feature extraction as data pre-processing, which extracted the energy of intrinsic mode functions for inputs of RF or SVM.

**Table 7**
Comparison classification results from experiment 1.

| Classification rate testing | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | SNR | | | | | | | | |
| | 12 dB | 13 dB | 14 dB | 15 dB | 16 dB | 17 dB | 18 dB | 19 dB | 20 dB |
| SDA | 91.67% | 93.43% | 93.51% | 94.64% | 94.18% | 95.11% | 95.93% | 98.56% | 99.83% |
| SAE | 88.61% | 91.02% | 91.66% | 91.14% | 92.21% | 94.97% | 95.63% | 96.99% | 99.57% |
| SVM | 83.58% | 84.20% | 88.12% | 89.75% | 91.75% | 93.15% | 95.25% | 96.50% | 99.75% |
| RF | 88.03% | 91.86% | 92.19% | 94.21% | 94.05% | 94.25% | 96.29% | 98.77% | 99.95% |
| AE | 79.37% | 81.16% | 82.93% | 85.13% | 87.42% | 91.25% | 93.96% | 96.87% | 99.37% |

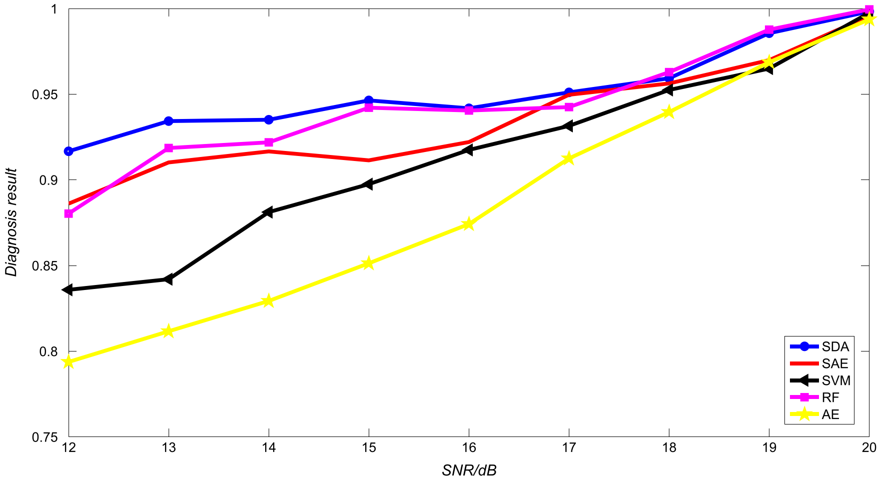

**Fig. 8.** Bearing diagnosis results of employed methods with different SNRs.

### 3.4.1. *Experiment 1*

Cross validation was used in this case, where on average, the original data were divided into five groups. Of these groups, four were set as the training data, whereas the remaining group was used for testing. The relative data sizes are shown in Table 4. This study employed ratio of correctness numbers to total categories as the classification accuracy. In this case, noise stress tests were also carried out to demonstrate the robustness and effectiveness of the employed methods with different signal to noise ratio (SNR) values, where the average values of the five tests with different training sets were considered to be the final result, as shown in Table 7.

From the diagnosis results, most of the methods exhibit high correct classification rates above 95% when the SNR is greater than 18. Taking the average classification results as an example, the diagnosis results are 99.83%, 99.57%, 99.75%, 99.95%, and 99.37%, respectively, for the SDA, SAE, SVM, RF, and AE methods at 20 dB. However, a clear downward trend subsequently appears as the noise ratio increases, as shown in Fig. 8.

It is noted that the SDA and SAE produce relative high classification accuracy and stability in most cases for different SNRs owing to the capability of learning complex non-linear mapping relationships between the input data and the health states through encoding higher order network structures with unsupervised deep learning. In addition, compared to the SAE, it seems that the data destruction process in the SDA is superior in terms of the bearing diagnosis under the working conditions with strong space noise, showing an accuracy improvement of 3.06% in the 12 dB as an example. As also indicated by the diagnosis results in this case study, the RF performs comparably accurately to the deep learning methods to some degree, mainly because the stochastic bagging processes used in the RF in this paper also possesses strong anti-noise abilities [52,53]. However, the diagnosis results of the RF in the cross validation process were not very stable compared to the SDA, leading to a relatively obvious fluctuation in the range of accuracies. For example, the RF diagnosis results became relatively worse when the SNR was 12, due to an overfitting problem when faced with strong ambient noise.

Despite the diagnosis accuracy, stable training time is also an important indicator used to judge the robustness of the employed algorithms in view of different levels of noise. In this study, the average training time for each comparative method in cross-validation is calculated and displayed in Fig. 9 for data with 12, 14, 16, 18, and 20 dB SNRs as examples.

As shown in Fig. 9, the training time for each method became shorter as the SNR increased from 12 dB to 20 dB. The training time was 357, 332, 194, 73, and 50 s for the SDA, SAE, RF, SVM, and AE, respectively, at 20 dB, whereas the calculation time increased to 379, 428, 227, 90, and 57 s at 12 dB. As aforementioned, the proposed deep learning methods could achieve better diagnosis performance; however, they appear to be more time-consuming compared to the shallow machine learning methods. This situation can be explained by their forward learning and back propagation mechanisms with a deep-learning architecture. In addition, the SDA achieved a more stable training time compared to the SAE, mainly because the data destruction can reduce the back propagation time of the learned features with better robustness.

### 3.4.2. *Experiment 2*

Fluctuations in the working conditions are inevitable during the practical application of bearings as a limitation on the classification accuracy. This experiment was conducted to investigate the applicability of the employed methods and to address such situations. In this case, all of the trained models were used to realize health state classification for all working conditions using

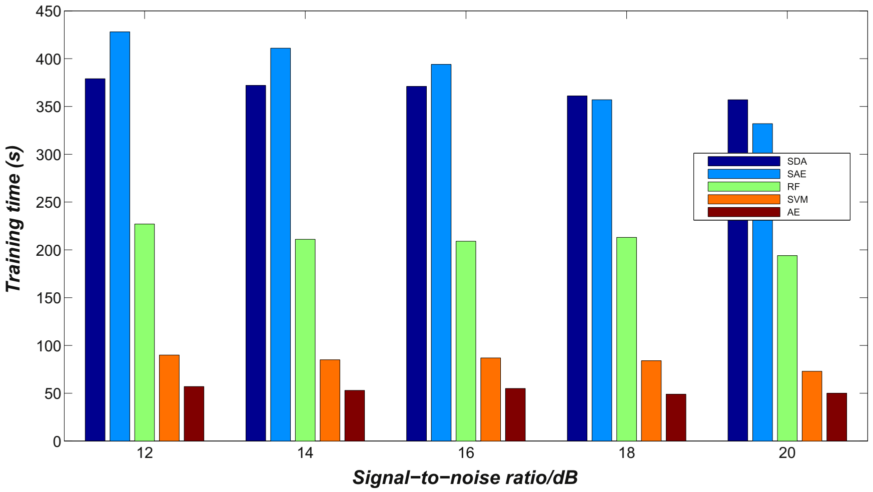

**Fig. 9.** Training time for each method with different levels of ambient noise.

**Table 8**
Comparison classification results for experiment 2.

| Classification rate testing | | | |
|---|---|---|---|
| Methods | C2 | C3 | C4 |
| SDA | 95.58% | 93.25% | 91.79% |
| SAE | 94.40% | 91.03% | 88.27% |
| SVM | 92.25% | 90.50% | 87.25% |
| RF | 95.50% | 93.75% | 90.75% |
| AE | 90.02% | 87.41% | 84.01% |

the data only from one condition, as shown in Table 4. Only the data for the DE bearing at 1797 rpm with the SNR of 17 dB were used for model training, whereas the data at the other rotating speeds were used as test samples. The comparison results are shown in Table 8.

As shown in Table 8, the diagnoses for all of the employed methods deteriorated due to the working condition fluctuations. The specific results decreased to 91.79%, 88.27%, 87.25%, 90.75%, and 84.01% for the SDA, SAE, SVM, RF, and AE methods, respectively. However, the proposed SDA based approach exhibits better performance than the other methods, obtaining a maximum difference of 1.76%, whereas the AE exhibited the most obvious decline of 7.24%. In addition, the diagnosis results of the SDA are more stable compared to the SAE and the other methods, demonstrating that the data destruction process could indeed result in a shared and robust representation with some invariance to the different working conditions, improving the stacked learning ability and having a stronger adaptability for the working condition fluctuations. Therefore, the performance still might be acceptable for a range of operating speeds due to the insensitivity of the working conditions.

## 4. Conclusions

The development of the SDA model and the comparative analysis of this method provide some valuable insights into fault diagnosis. Through the diagnosis results of the comparative experiments, we see that the proposed deep learning method is able to adaptively mine salient fault characteristics and effectively identify the health states with high diagnosis accuracy and strong robustness. Compared to traditional diagnosis algorithms, the main advantage of the proposed method is that the fault features are learned via a general-purpose learning procedure instead of hand-engineered or having prior knowledge on signal processing techniques, which is easy to apply to diagnosis issues. The diagnosis results also demonstrate that the deep architecture based learning methodology indeed provides a way to improve the health state identification accuracy, and data destruction facilitates the feature representation process to recognize the robust characteristics of the input signals containing environmental noise and working condition fluctuations. The feasibility of the SDA-based fault diagnosis approach is demonstrated using health state classification datasets from rolling bearings.

Future work will include more experimental tests to further understand the limitations of the deep learning method, particularly to larger ranges of fluctuations and the periodic variability of the employed systems. In addition, in this paper, we used a deep architecture based on a number of testing experiments. However, this is still an open problem for optimal parameter determination, especially when a deeper architecture is employed. It is very attractive to investigate the possibility of the practical applications of optimization algorithms in the model establishment process in the future. Studies on the health assessment and performance degradation prediction are also expected to evaluate if the proposed method is applicable to a greater number of research fields.

## Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant nos.51575021 and 51105019) and the Technology Foundation Program of National Defense (Grant no. Z132013B002), as well as the Innovation Foundation of BUAA for PhD Graduates.

Chen Lu and Zheng-Ya Wang contributed equally to this work and should be considered joint first authors.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.sigpro.2016.07.028.

## References

[1] E. Zio, Reliability engineering: old problems and new challenges, Reliab. Eng. Syst. Safe. 94 (2009) 125–141.

[2] H. Sun, Z. He, Y. Zi, J. Yuan, X. Wang, J. Chen, S. He, Multiwavelet transform and its applications in mechanical fault diagnosis - a review, Mech. Syst. Signal Process. 43 (2014) 1–24.

[3] S. Yin, S.X. Ding, A. Haghani, H. Hao, P. Zhang, A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process, J. Process. Contr. 22 (2012) 1567–1581.

[4] Y. Qin, B. Tang, Y. Mao, Adaptive signal decomposition based on wavelet ridge and its application, Signal Process. 120 (2016) 480–494.

[5] G. Niu, B. Yang, M. Pecht, Development of an optimized condition-based maintenance system by data fusion and reliability-centered maintenance, Reliab. Eng. Syst. Safe. 95 (2010) 786–796.

[6] N. Phuong, M. Kang, J. Kim, B. Ahn, J. Ha, B. Choi, Robust condition monitoring of rolling element bearings using de-noising and envelope analysis with signal decomposition techniques, Expert Syst. Appl. 42 (2015) 9024–9032.

[7] I. El-Thalji, E. Jantunen, A summary of fault modelling and predictive health monitoring of rolling element bearings, Mech. Syst. Signal Process. 60–61 (2015) 252–272.

[8] W. Yang, P.J. Tavner, Empirical mode decomposition, an adaptive approach for interpreting shaft vibratory signals of large rotating machinery, J. Sound Vib. 321 (2009) 1144–1170.

[9] X. Zhang, B. Wang, X. Chen, Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine, Knowl.-Based Syst. 89 (2015) 56–85.

[10] W. Du, A. Li, P. Ye, C. Liu, Fault diagnosis of plunger pump in truck crane based on relevance vector machine with particle swarm optimization algorithm, Shock Vib. 20 (2013) 781–792.

[11] Y. Ding, J. Ma, Y. Tian, Health assessment and fault classification for hydraulic pump based on LR and softmax regression, J. Vibroeng. 17 (2015) 1805–1816.

[12] F. Zakaria, D. Johari, I. Musirin, Optimized artificial neural network for the detection of incipient faults in power transformer, in: Proceedings of the 2014 IEEE 8th International Power Engineering And Optimization Conference (PEOCO), 2014, pp. 635–640.

[13] C. Lu, H. Yuan, L. Tao, H. Liu, Performance assessment of hydraulic servo system based on bi-step neural network and autoregressive model, J. Vibroeng. 15 (2013) 1546–1559.

[14] M.J. Santofimia, X. Del Toro, P. Roncero-Sanchez, F. Moya, M.A. Martinez, J. C. Lopez, A qualitative agent-based approach to power quality monitoring and diagnosis, Integr. Comput.-Aid E 17 (2010) 305–319.

[15] Y. Yu, YuDejie, J.S. Cheng, A roller bearing fault diagnosis method based on EMD energy entropy and ANN, J. Sound Vib. 294 (2006) 269–277.

[16] R. Yan, R.X. Gao, X. Chen, Wavelets for fault diagnosis of rotary machines: a review with applications, Signal Process. 96 (2014) 1–15.

[17] A. Youssef, C. Delpha, D. Diallo, An optimal fault detection threshold for early detection using Kullback–Leibler Divergence for unknown distribution data, Signal Process. 120 (2016) 266–279.

[18] J. Harmouche, C. Delpha, D. Diallo, Incipient fault detection and diagnosis based on Kullback-Leibler divergence using principal component analysis: part II, Signal Process 109 (2015) 334–344.

[19] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems-reviews, methodology and applications, Mech. Syst. Signal Process. 42 (2014) 314–334.

[20] R. Jegadeeshwaran, V. Sugumaran, Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines, Mech. Syst. Signal Process. 52–53 (2015) 436–446.

[21] X. Zhang, W. Chen, B. Wang, X. Chen, Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization, Neurocomputing 167 (2015) 260–279.

[22] W. Yan, Application of random forest to aircraft engine fault diagnosis, in: F. Sun ,H.P. Liu (Eds.),2006, pp. 468–475.

[23] B. Yang, X. Di, T. Han, Random forests classifier for machine fault diagnosis, J. Mech. Sci. Technol. 22 (2008) 1716–1725.

[24] H. Dong, Z. Wang, H. Gao, On design of quantized fault detection filters with randomly occurring nonlinearities and mixed time-delays, Signal Process. 92 (2012) 1117–1125.

[25] C. Gautam, V. Ravi, Counter propagation auto-associative neural network based data imputation, Inform. Sci. 325 (2015) 288–299.

[26] Z. Li, H. Fang, M. Huang, Diversified learning for continuous hidden Markov models with application to fault diagnosis, Expert Syst. Appl. 42 (2015) 9165–9173.

[27] Z. Wang, C. Lu, J. Ma, H. Yuan, Z. Chen, Novel method for performance degradation assessment and prediction of hydraulic servo system, Sci. Iran 22 (2015) 1604–1615.

[28] F. Jia, Y. 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. Sig. Process. 72–73 (2016) 303–315.

[29] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.

[30] B. Yoshua, L. Pascal, P. Dan, L. Hugo. Greedy Layer-Wise Training of Deep Networks, in: Editor edito. Advances in Neural Information Processing Systems 19 (NIPS'06). Pub Place; 2007.

[31] G.E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.

[32] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[33] X. Zhang, J. Wu, Denoising deep neural networks based voice activity detection, in: Proceedings of the International Conference on Acoustics Speech and Signal Processing ICASSP,2013, pp. 853–857.

[34] Y. Bengio, H. Lee, Editorial introduction to the Neural Networks special issue on Deep Learning of Representations, Neural Netw. 64 (2015) 1–3.

[35] G. Varol, A.A. Salah, Efficient large-scale action recognition in videos using extreme learning machines, Expert Syst. Appl. 42 (2015) 8274–8282.

[36] M. Lee, A. Hirose, Z. Hou, R.M. Kil, H.A. Song, S. Lee, in: M. Lee, A. Hirose, Z. Hou, R.M. Kil (Eds.), Hierarchical Representation Using NMF, Springer Berlin Heidelberg, 2013, pp. 466–473.

[37] I. Arel, D.C. Rose, T.P. Karnowski, Deep machine learning-a new frontier in artificial intelligence research, IEEE Comput. Intell. Mag. 5 (2010) 13–18.

[38] P. Tamilselvan, P. Wang, Failure diagnosis using deep belief learning based health state classification, Reliab. Eng. Syst. Safe. 115 (2013) 124–135.

[39] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, M. Iwahashi, Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification, Eurasip. J. Audio Speech (2015).

[40] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.

[41] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw./Publ. IEEE Neural Netw. Counc. 5 (1994) 157–166.

[42] D.F. Wulsin, J.R. Gupta, R. Mani, J.A. Blanco, B. Litt, Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement, J. Neural Eng. 8 (2011).

[43] Y.W. Teh, M. Welling, S. Osindero, G.E. Hinton, Energy-based models for sparse overcomplete representations, J. Mach. Learn. Res. 4 (2004) 1235–1260.

[44] T. Amaral, L.M. Silva, L.A. Alexandre, C. Kandaswamy, J.M. Santos, J.M. de Sa, Using Different Cost Functions to Train Stacked Auto-encoders, in: F. Castro, A. Gelbukh ,M.G. Mendoza (Eds.) Mexican International Conference on Artificial Intelligence-MICAI,2013, pp. 114-120.

[45] J. Yang, Y. Bai, G. Li, M. Liu, X. Liu, A novel method of diagnosing premature ventricular contraction based on sparse auto-encoder and softmax regression, Bio-Med. Mater. Eng. 261 (2015) S1549–S1558.

[46] N. Andrei, An adaptive conjugate gradient algorithm for large-scale unconstrained optimization, J. Comput. Appl. Math. 292 (2016) 83–91.

[47] W.A. Smith, R.B. Randall, Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study, Mech. Syst. Signal Process. 64–65 (2015) 100–131.

[48] Rasmusbergpalm, Deep Learning Toolbox, ⟨https://github.com/rasmusbergpalm/DeepLearnToolbox⟩, 2015 (accessed 16.07.22).

[49] A. Ciates, H. Lee, A.Y. Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning, in: Editor edito. International Conference on Artificial Intelligence and Statistics. Pub Place; 2011.

[50] C. Hsieh, R. Lu, N. Lee, W. Chiu, M. Hsu, Y.J. Li, Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks, Surgery 149 (2011) 87–93.

[51] V. Cherkassky, Y.Q. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Netw. 17 (2004) 113–126.

[52] R.E. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer, A comparison of decision tree ensemble creation techniques, IEEE Trans. Pattern Anal. 29 (2007) 173–180.

[53] P.O. Gislason, J.A. Benediktsson, J.R. Sveinsson, Random Forests for land cover classification, Pattern. Recogn. Lett. 27 (2006) 294–300.