ORIGINAL ARTICLE

Open Access

Sparse Autoencoder-based Multi-head Deep Neural Networks for Machinery Fault Diagnostics with Detection of Novelties

Zhe Yang[1,2], Dejan Gjorgjevikj[3], Jianyu Long[1], Yanyang Zi[2], Shaohui Zhang[1] and Chuan Li[1*]

Abstract

Supervised fault diagnosis typically assumes that all the types of machinery failures are known. However, in practice unknown types of defect, i.e., novelties, may occur, whose detection is a challenging task. In this paper, a novel fault diagnostic method is developed for both diagnostics and detection of novelties. To this end, a sparse autoencoder-based multi-head Deep Neural Network (DNN) is presented to jointly learn a shared encoding representation for both unsupervised reconstruction and supervised classification of the monitoring data. The detection of novelties is based on the reconstruction error. Moreover, the computational burden is reduced by directly training the multi-head DNN with rectified linear unit activation function, instead of performing the pre-training and fine-tuning phases required for classical DNNs. The addressed method is applied to a benchmark bearing case study and to experimental data acquired from a delta 3D printer. The results show that its performance is satisfactory both in detection of novelties and fault diagnosis, outperforming other state-of-the-art methods. This research proposes a novel fault diagnostics method which can not only diagnose the known type of defect, but also detect unknown types of defects.

Keywords: Deep learning, Fault diagnostics, Novelty detection, Multi-head deep neural network, Sparse autoencoder

1 Introduction

With the objective of increasing the availability and reducing operation and maintenance cost of mechanical systems, Prognostics and Health Management (PHM) approaches has been getting more and more attention [1–3]. Fault diagnostics is one of the fundamental tasks of PHM, which aims at detecting and diagnosing machinery failure using model-based or data-driven approaches [4]. In the era of Industry 4.0, since mechanical systems are getting more and more complex, it is very difficult and expensive to develop physics-based degradation models required for model-based approaches. Whereas the increased availability of data collected from multiple monitoring sensors and

the grown ability of processing data by artificial intelligence algorithms have brought the great potential for the development of advanced data-driven approaches [5]. Data-driven approaches are typically based on the development of an empirical classification model trained on monitoring data. Chine et al. [6] proposed a fault diagnostics approach for the photovoltaic system based on Artificial Neural Networks (ANNs). Malik et al. [7] used Empirical Mode Decomposition (EMD) for feature extraction, and an ANN was trained using the extracted features for gearbox fault diagnostics. He et al. [8] extracted statistical features from monitoring signals and Support Vector Machines (SVMs) were developed for fault diagnostics for the 3D printer. Li et al. [9] utilized wavelet packet decomposition and SVM for the diagnostics of machinery faults of high-voltage circuit breaker failures. Liu et al. [10] proposed a bearing diagnostics approach which combined EMD and auto-regressive model to extract features from

*Correspondence: chuanli@dgut.edu.cn
[1] School of Mechanical Engineering, Dongguan University of Technology, Dongguan 523808, China
Full list of author information is available at the end of the article

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 2 of 12

vibration signals and used random forests to set an effective classification model. Hu et al. [11] developed a wind turbine bearing fault diagnosis based on multi-masking EMD and fuzzy c-means clustering. Other approaches, such as k-nearest neighbor [12], naïve Bayes [13], linear discriminant analysis [14], fuzzy petri nets [15], extreme learning machines [16], have also been successfully developed for fault diagnostics.

Although these intelligent fault diagnostics approaches have shown great progress, there are still some limitations. On the one hand, they rely on the identification of handcrafted features requiring expert knowledge or computationally demanding feature selection methods. On the other hand, the shallow learning architecture of these approaches leads to poor performance for complex classification problems. To address these difficulties, recently Deep Learning (DL) methods have been extensively applied for machinery fault diagnostics [17–20]. DL methods are expected to automatically provide high-level representation by using neural networks with multiple layers of non-linear transformations, without requiring human-designed and labor-intensive analyses of the data [21, 22]. Jia et al. [23] fed the frequency spectra of vibration signals into stacked AutoEncoders (AEs)-based Deep Neural Networks (DNNs) for rotating machinery diagnostics. Chen et al. [24] proposed Sparse AutoEncoder (SAE) and deep belief network for fault diagnosis of bearings. Lu et al. [25] employed denoising AEs for fault diagnostics of rotating machinery components. Shao et al. [26] presented a deep belief network by stacking multiple restricted Boltzmann machines for fault diagnostics of induction motors. Wang et al. [27] employed Convolutional Neural Networks (CNNs) for fault diagnostics of motors, where the input is the time-frequency map of the vibration signal. Jiang et al. [28] presented a method incorporating multiscale learning into the traditional CNN architecture for defect identification of wind turbine gearbox. Chang et al. [29] proposed a concurrent CNN composed of parallel convolution layers with multi-scale kernels for fault diagnosis of wind turbine bearings. Yuan et al. [30] investigated the development of different Recurrent Neural Networks (RNNs) including vanilla RNN, Long Short-Term Memory (LSTM) and gated recurrent unit for fault diagnostics and prognostics of aero engine. Chen et al. [31] utilized multi-scale CNNs to extract features that are then fed to a LSTM for bearing fault diagnostics. Echo State Networks (ESNs), another type of RNN characterized by high training efficiency, were successfully developed for fault diagnostics in Refs. [32] and [33], respectively. Zhang et al. [34] proposed a novel approach named deep hybrid state network integrating sparse AE and double-structure ESN for fault diagnosis of 3-D printers.

Even though these studies have outperformed other state-of-the-art fault diagnostics methods, the occurrence of novel conditions, which is common in real applications since the available monitoring data typically cannot cover all the possible types of defects and operating conditions, is seldom considered.

In this context, the objective of the present work is to develop a fault diagnostics method with the following characteristics: (i) it does not require the application of feature selection and extraction techniques; (ii) it can detect novel conditions not been recorded in the available dataset; and (iii) it can accurately diagnose known normal and faulty conditions.

We consider SAE-based method as a possible solution for this objective. An AE is a neural network with a symmetrical architecture, composed of an "encoder" and a "decoder" network. The "encoder" network transforms the large-dimensional input data into a small set of features and the "decoder" network reconstructs the data from the extracted features [35]. Since many natural signals show the sparsity property, which states that signals are with useful patterns occurred sparsely. The sparse modeling of signals has proven to be effective to extract the inherent low-dimensional features [36]. SAE, a variant of AE, employs sparsity penalty to encourage the extraction of discriminative features, which prevents the AE from simply copying the inputs and makes the features more representative for classification [37, 38]. Since SAE tends to perform a poor reconstruction with data different from those used for its training, the reconstruction error of the input data is expected to be an indicator of novelty detection [39]. A SAE can be transformed into a SAE-based DNN for diagnostics [40] by: (i) pre-training multiple 1-hidden-layer SAEs and stacking them to build a multi-layer stacked SAE, and (ii) taking only the multi-layer encoder network of the stacked SAE and adding a classification layer on it to build a DNN, then fine-tuning the DNN using input-output data. The "pre-training" and "fine-tuning" are used mainly due to the gradient vanishing/exploding problems when directly training the DNN caused by commonly adopted tanh or sigmoid nonlinear activation functions [35].

Various studies have demonstrated the success of DNNs for machinery fault diagnostics [23, 41–46]. However, they didn't consider the detection of previously unseen conditions, and most of them construct the DNN through the way of pre-training and fine-tuning which requires many computational efforts. In Ref. [39], Principi et al. proposed a novelty detection approach based on the reconstruction error of stacked AE. But the stacked AE is trained using normal data only, instead of data including normal and known multiple faulty conditions, which is common in fault diagnostics problems.

To further explore the capability of AE and AE-based DNNs for both novelty detection and fault diagnostics, we propose a SAE-based multi-head DNN for addressing the two problems. The multi-head DNN uses an encoder network to jointly learn a shared encoding representation, based on which a decoder network and a classification module are employed for unsupervised reconstruction and supervised classification, respectively. The novelty detection is realized by comparing the reconstruction error with a pre-defined threshold. In addition, the Rectified Linear Unit (ReLU) activation function [35], which is widely used for the training of CNNs to relieve the gradient vanishing/exploding problems, is adopted for the multi-head DNN to make it possible to directly train the DNN instead of following the conventional way of pre-training and fine-tuning.

The proposed method is validated by using two case studies about machinery fault diagnostics. The first case is from a benchmark considering bearings with different types of defects operating under different loads. The second one is a real experiment in which the monitoring data of a delta 3D printer with different types of defects are collected. The performance of the proposed method is compared to that of other commonly used novelty detection and fault diagnostics methods.

The remaining of this paper is organized as follows. Section 2 presents the problem statement. The proposed fault diagnostics method is illustrated in Section 3. Section 4 shows the applications of the proposed method to a benchmark bearing diagnostics case study and to experimental data of a delta 3D printer. Finally, conclusions are drawn in Section 5.

2 Problem Statement

The objective of this work is to develop a machinery fault diagnostics method being able to identify the unknown faulty conditions, and to diagnose the known normal and faulty conditions among C different classes. We assume to have available the measurements of S signals collected during the operation of the machinery under all the already known C conditions. For ease of notation, we assume all the signals are collected at the same sampling rate. Let $x_s^k(\tau)$, $\tau = 1, \ldots, T^k$, be the τ th data point collected in the s-th signal, $s \in \{1, \ldots, S\}$, under the kth condition, $k \in \{1, \ldots, C\}$, where T^k is the time at which the last data point under the k th condition is collected.

Each signal x_s^k is segmented into pieces containing M data points using a non-overlapping window, and $\lfloor T/M \rfloor$ pieces are obtained. Then, all the S signals belonging to the same piece are gathered together as a sample $x_{i_k} = \left\{ x_1^{i_k}, \ldots, x_s^{i_k}, \ldots, x_S^{i_k} \right\}$, $i_k = 1, \ldots, N^k$, where $x_s^{i_k}$ is the i_k th piece of the sth signal collected under the k th condition.

Finally, all the available $N = \sum_{k=1}^{C} N^k$ samples are lumped together to form a dataset of N input-output pairs (x_i, y_i), $i = 1, \ldots, N$, where the output $y_i \in \{1, \ldots, C\}$ is the corresponding label of the sample class.

The proposed method receives the test sample $x^{TEST} = \{x_1^{TEST}, \ldots, x_s^{TEST}, \ldots, x_S^{TEST}\}$ collected from test equipment as the input, and is required to identify whether it operates in any of already known conditions, if no, a novel condition is detected, otherwise the class $y^{TEST} \in \{1, \ldots, C\}$ is diagnosed.

3 The Proposed Method

3.1 Sparse Autoencoder

An autoencoder is an unsupervised neural network with a symmetrical structure [35], as shown in Figure 1.

The input D-dimensional sample x is transformed into its hidden representation $a = \left[a_1, a_2, \ldots, a_{D_1} \right]$ from the input layer to hidden layer, known as the "encoder":

$$a = \sigma(W_1 x + b_1), \tag{1}$$

where σ, W_1, b_1 are the activation function, the weight matrix and the bias vector of the encoder, respectively.

Then, the decoder, i.e., the network from hidden layer to the output layer, reconstructs the input x to $\hat{x}$ based on the feature vector a:

$$\hat{x} = \sigma(W_2 a + b_2), \tag{2}$$

where W_2, b_2 are the decoder weight matrix and the bias vector, respectively. As a variant of the autoencoder, a sparse autoencoder encourages the extraction of discriminative features by adding a sparse restriction for the network training [37]. Given the available input samples x_i, $i = 1, \ldots, N$, the training objective of SAE is to minimize the following cost function:
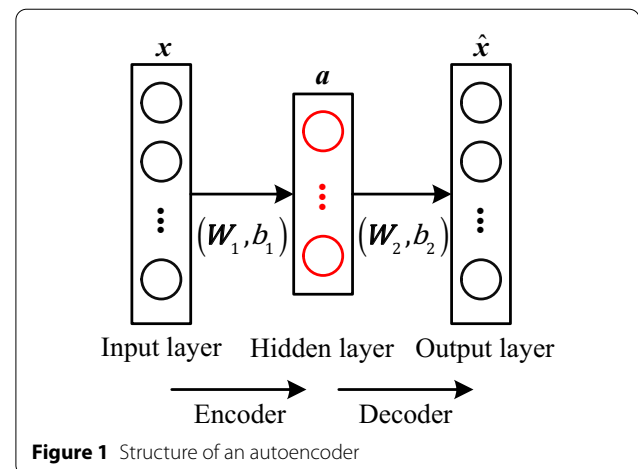

Figure 1 Structure of an autoencoder

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 4 of 12

$$F_{cost} = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 + \beta R_{sparse}, \quad (3)$$

where the first term is the reconstruction error, the second term is the $L2$ regularization term where $\boldsymbol{W}$ is the SAE weight matrix, R_{sparse} is the sparsity regularization, and λ and β are coefficients control the importance of the corresponding terms.

It has been found that constraining hidden neurons to be inactive most of the time makes them respond to different patterns lying in the data, i.e., the extracted features $\boldsymbol{a}$ are discriminative [37]. Let $\widehat{p}_j$ be the average activation of the jth hidden neuron of the SAE hidden layer, considering all the input samples $\boldsymbol{x}_i, i = 1, \cdots, N$:

$$\widehat{p}_j = \frac{1}{N}\sum_{i=1}^{N} a_{i,j}, \quad (4)$$

where $a_{i,j}$ is the jth element of the ith hidden representation $\boldsymbol{a}_i, j = 1, \cdots, D_1$. The sparsity regularization in Eq. (3), R_{sparse}, is calculated using the Kullback-Leibler (KL) divergence function to measure whether $\widehat{p}_j$ is close to a desired small sparsity proportion p:

$$\begin{aligned} R_{sparse} &= \sum_{j=1}^{D_1} \mathrm{KL}\left(p\|\widehat{p}_j\right) \\ &= \sum_{j=1}^{D_1}\left[p\log\frac{p}{\widehat{p}_j} + (1-p)\log\frac{1-p}{1-\widehat{p}_j}\right]. \end{aligned} \quad (5)$$

The KL function is zero when all $\widehat{p}_j$ are equal to p and increases when they diverge.

3.2 Multi-head Deep Neural Network

We propose a SAE-based multi-head DNN for both novelty detection and diagnostics of known conditions. The multi-head DNN consists of three modules: an encoder, a decoder and a classification module, as shown in Figure 2. The encoder extracts high-level representations from input data using multiple layers of non-linear transformations. The extracted representations are shared by the decoder and the classification module as their input. The decoder aims at reconstructing the input data whereas the classification module predicts the label of the input data. More specifically, the construction of the multi-head deep neural network includes:

(1) Encoder

The encoder receives the sample $\boldsymbol{x}$ as the input and extracts high-level representation of $\boldsymbol{x}$ using L hidden layers, $L\epsilon\mathbb{N}_+$. Let $\boldsymbol{a}^{(1)}, ..., \boldsymbol{a}^{(L)}$ be hidden representations extracted from the corresponding hidden layer, their dimensions $D_1, ..., D_L$, are set as following: (i) D_1 is typically set larger than that of input layer, D, to obtain a sparse-overcomplete representation at the first hidden layer, which has been shown able to extract independent
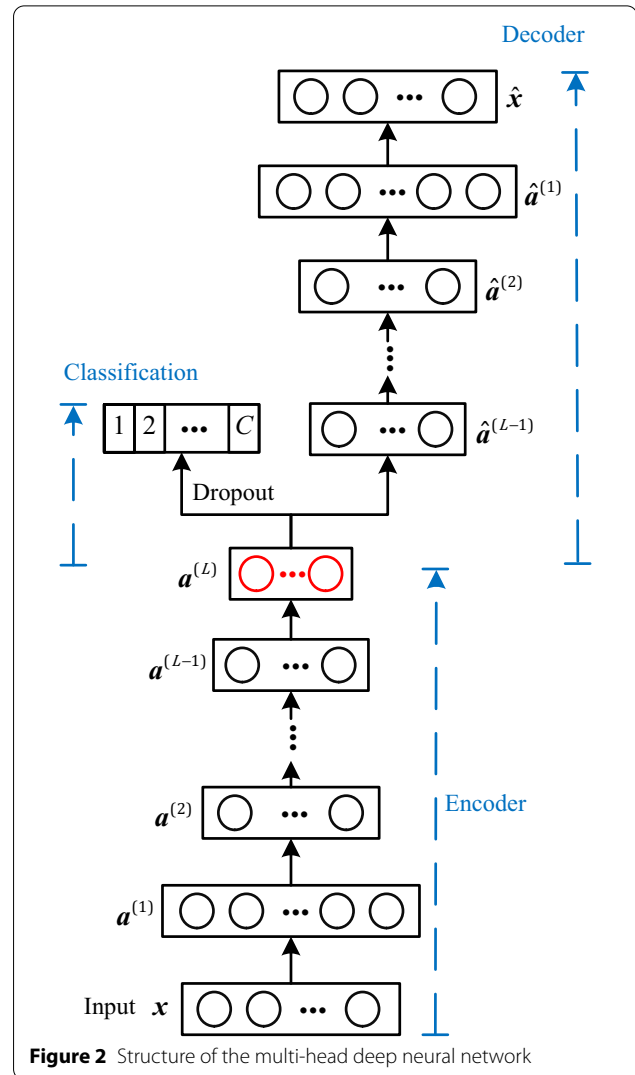

Figure 2 Structure of the multi-head deep neural network

basis functions for input data [47]; (ii) the dimension D_l, $l = 2, 3, \ldots, L$ of the remaining hidden layers should be smaller than D_{l-1} to obtained compressed representation. The ReLU activation function is employed in the input layer and all the hidden layers.

The sparsity regularization defined in Eq. (5) is applied to all the hidden layers to encourage the extraction of discriminative features, and the $L2$ regularization is employed to constraint the weights of the encoder.

(2) Decoder

The structures of the decoder and the encoder, i.e., the number and dimensions of hidden layers, are exactly symmetrical. The decoder aims at recovering the input $\boldsymbol{x}$ using $\boldsymbol{a}^{(L)}$. We use ReLU activation function in all the hidden layers and the sigmoid activation function in the output layer for reconstruction. The $L2$ regularization is used to constraint the weights of the decoder.

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 5 of 12

(3) Classification module

The classification module employs a softmax layer with C neurons, representing different conditions, to solve the C-class classification problem. Given the representation $\boldsymbol{a}^{(L)}$, the softmax layer gives a vector $[\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_C]$ as the output. The kth unit of the output, $\widehat{y}_k$, $k \in \{1, 2, \ldots, C\}$, is typically regarded as a number proportional to the conditional probability that the machinery is in the kth condition given sample $\boldsymbol{x}$:

$$\widehat{y}_k = \frac{\exp(z_k)}{\sum_{j=1}^{C} \exp(z_j)}, \tag{6}$$

where $0 \leq \widehat{y}_k \leq 1$, $\sum_{k=1}^{C} \widehat{y}_k = 1$ and z_k is the kth output unit before applying the softmax activation function:

$$z_k = \boldsymbol{w}_k \cdot \boldsymbol{a}^{(L)} + b_k, \tag{7}$$

where $\boldsymbol{w}_k$ and b_k are weights and bias of the kth neuron of the softmax layer.

To prevent over-fitting, in the classification module, we employ the dropout regularization on the hidden layer $\boldsymbol{a}^{(L)}$. Dropout randomly sets to zero a proportion p_{drop} of the hidden neurons during forward and backpropagation [35]. Therefore, the following equation is used for computing z_k considering the dropout, instead of using Eq. (7):

$$z_k = \boldsymbol{w}_k \cdot \left(\boldsymbol{a}^{(L)} \circ \boldsymbol{r} \right) + b_k, \tag{8}$$

where $\circ$ is the element-wise multiplication operator and $\boldsymbol{r} \in \mathbb{R}^{D_L}$ is a 'masking' vector of Bernoulli random variables with probability p_{drop} of being 0. Gradients are backpropagated only through the unmasked neurons.

To be associated with the neurons of the softmax layer, in the training set $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, N$, each label y_i is transformed into a one-hot C-dimensional vector $(y_{i,1}, y_{i,2}, \ldots, y_{i,C})_{i=1,\ldots,N}$, where

$$y_{i,k} = \begin{cases} 1, & y_i = k, \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, \ldots, C. \tag{9}$$

The training objective of the multi-head DNN over the training set $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, N$, is to minimize the following cost function:

$$\begin{aligned} F_{cost} = & \frac{\eta_1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|^2 + \frac{\lambda}{2} \|\boldsymbol{W}\|^2 \\ & + \beta \sum_{j=1}^{L} R_{sparse}^{(j)} - \frac{\eta_2}{N} \sum_{i=1}^{N} \sum_{k=1}^{C} y_{i,k} \log \widehat{y}_{i,k}, \end{aligned} \tag{10}$$

where the first term is the reconstruction error, the second term is the $L2$ regularization term where $\boldsymbol{W}$ is the weight matrix of the whole multi-head DNN, $R_{sparse}^{(j)}$ is the sparsity regularization applied on the jth hidden layer

of the encoder, the last term is the cross-entropy loss measuring the performance of the classification module, λ, β, η_1 and η_2 are coefficients controlling the importance of the corresponding terms.

3.3 Overview of the Proposed Method

The flow chart of the proposed fault diagnostics method using the multi-head DNN is shown in Figure 3. The original monitoring signals are segmented into data samples (Section 2). Then a multi-head DNN is built and trained using data samples collected from already known C classes associated with normal and faulty conditions. Given a test data sample, the multi-head DNN first identifies whether it belongs to the already known C classes. If yes, an unknown condition is reported, otherwise the multi-head DNN diagnoses the label of the test sample among C already known classes.
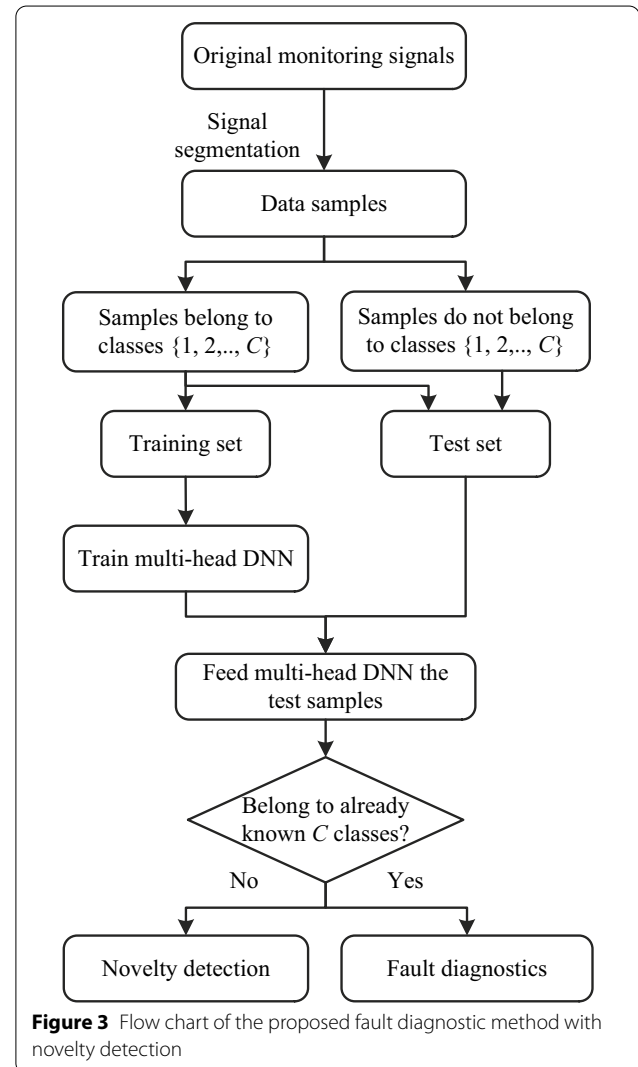

Figure 3 Flow chart of the proposed fault diagnostic method with novelty detection

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 6 of 12

With respect to the novelty detection, a test sample x^{TEST} is detected as unknown when the magnitude of its reconstruction error is larger than a certain threshold δ:

$$e^{TEST} = \|x^{TEST} - \widehat{x}^{TEST}\|^2 > \delta. \tag{11}$$

In this work, we set the threshold as:

$$\delta = Q3 + 1.5(Q3 - Q1), \tag{12}$$

where $Q1$ and $Q3$ are the 25th and 75th percentiles of $\left\{\|x_i - \widehat{x}_i\|^2, i = 1, \ldots, N\right\}$, i.e., reconstruction errors over the training set. Notice that δ is the upper whisker in the box-plot method [48].

4 Experimental Evaluations

The proposed method has been verified with respect to data collected from a bearing fault benchmark, before its application to the experimental data acquired on a delta 3D printer. All computations have been performed using an Intel Core i5-5200 CPU at 2.2 GHz processor with 4 GB RAM in Python 3.6 environment.

4.1 Evaluation of the Proposed Method Using Benchmark Data

We consider the benchmark bearing diagnostics dataset provided by the Case Western Reserve University, which contains vibration data collected from an experimental rig with defective bearings operating under four different loads [49]. During the experiment, besides the normal condition, three different kinds of fault, i.e., inner race fault, outer race fault (at 6 o'clock) and ball fault, were introduced the drive-end bearing of the motor with fault diameters of 0.18 mm, 0.36 mm and 0.54 mm, respectively. Table 1 shows the detailed description of the dataset. Vibration signals were collected at sampling frequency 12 kHz, by an accelerometer attached through the magnetic base at the drive-end.

The vibration signal of each condition is segmented into samples using a non-overlapping fixed-length time window containing 1024 data points. We implement Fast Fourier Transformation (FFT) on each sample to get the 1024 Fourier amplitude coefficients. Since the coefficients are symmetric, the first 512 coefficients are used for each sample. The last column of Table 1 listed the number of samples obtained for each bearing condition.

We assume normal, inner race faults and balls faults are already known conditions (classes 1–7), and the outer race faults (classes 8–10) are unknown conditions. The training set is composed of 70% of the data randomly selected from classes 1–7, respectively. The remaining data of classes 1–7 and all the data of classes 8–10 are used as the test set. During the training of the multi-head DNN, 5% of the training data is randomly selected as the validation set to prevent the overfitting of the model. The developed multi-head DNN is formed by an encoder with an input layer of $D = 512$ neurons and $L = 3$ hidden layers of $D_1 = 600, D_2 = 100, D_3 = 10$ neurons, a symmetric decoder, and a classification module with a softmax layer of 7 neurons associated with the known 1–7 classes in the training set. The hyperparameters of the multi-head DNN are set as follows: $p = 0.05$, $\lambda = 1 \times 10^{-7}$, $\beta = 1, \eta_1 = 80, \eta_2 = 1$ and $p_{drop} = 0.3$.

(1) Novelty detection

With respect to novelty detection, the objective is to isolate samples of unknown classes from those of known classes in the test set. The box plot of reconstruction errors on different datasets is shown in Figure 4. The threshold for novelty detection, δ, is set to 5.25×10^{-4}
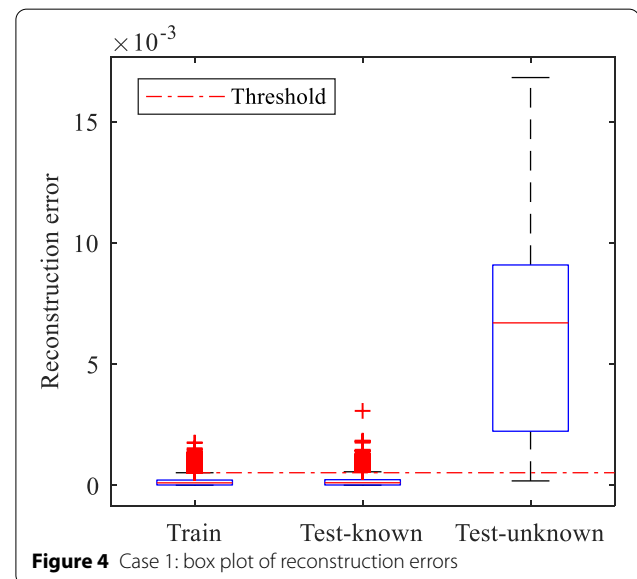
Table 1 Bearing conditions with assigned labels

Bearing condition	Fault diameter (mm)	Label /Class	Load (hp)	Number of samples
Normal	–	1	0,1,2,3	1657
Inner race fault	0.18	2	0,1,2,3	476
Inner race fault	0.36	3	0,1,2,3	472
Inner race fault	0.54	4	0,1,2,3	474
Balls fault	0.18	5	0,1,2,3	473
Balls fault	0.36	6	0,1,2,3	475
Balls fault	0.54	7	0,1,2,3	475
Outer race fault	0.18	8	0,1,2,3	475
Outer race fault	0.36	9	0,1,2,3	474
Outer race fault	0.54	10	0,1,2,3	476

Figure 4 Case 1: box plot of reconstruction errors

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 7 of 12

in this case study based on the reconstruction errors on the training set, as described in Eq. (12). Notice that the majority of reconstruction errors of known class test samples are below the threshold whereas those of unknown class test samples are above the threshold, i.e., the known and unknown class test samples are well separated.

In order to evaluate the novelty detection performance of the proposed method quantitatively, we denote test samples of known (classes 1–7) and unknown (classes 8–10) classes as positive and negative, respectively. And we use True Positive (*TP*) to represent the number of correctly classified positive samples, False Negative (*FN*) to represent the number of positive samples misclassified as negative, True Negative (*TN*) to represent correctly classified negative samples and False Positive (*FP*) to represent the number of negative samples misclassified as positive. The novelty detection performance of the developed multi-head DNN is evaluated considering the True Positive Rate (*TPR*), True Negative Rate (*TNR*) and the F_1 score:

$$True\ Positive\ Rate(TPR) = \frac{TP}{TP + FN}, \tag{13}$$

$$True\ Negative\ Rate\ (TNR) = \frac{TN}{TN + FP}, \tag{14}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}. \tag{15}$$

The *TPR* and *TNR* are the proportions of correctly classified positive and negative samples, respectively. The F_1 score is widely used as a performance metric for binary classification models, whose value ranges in $[0, 1]$ where the larger value indicates better performance.

The result of novelty detection of the proposed method is shown in Table 2. The proposed method has been compared with two popular one-class learning methods, one-class SVM and Isolation forest. One-class SVM aims at constructing a smooth boundary around the majority of probability mass of data [50, 51]. Since one-class SVM is a shallow model which prefer low-dimensional input data, the $D_3 = 10$ dimensional feature vectors $\boldsymbol{a}_i^{(3)}$,

$i = 1, \ldots, N$, extracted by the encoder from the training set are used as its input. The Radial Basis Function (RBF) kernel is used for one-class SVM, and the parameter v, i.e., the assumed proportion of negative samples in the training set, is set to 0.01, which has been optimized with the objective of maximizing the F_1 score by trial-and-error considering as possible options $\{0, 0.01, \ldots, 0.2\}$, respectively. The isolation forest employs decision trees for novelty detection, each tree is constructed by randomly splitting features and the anomalous data will produce significant shorter paths in trees. The isolation forest is also fed by $D_3 = 10$ dimensional feature vectors $\boldsymbol{a}_i^{(3)}$, $i = 1, \ldots, N$, extracted by the encoder. Its parameter N_{tree}, the number of trees, is set to 40, which has been optimized with the objective of maximizing the F_1 score by trial-and-error considering as possible options $\{10, 20, \ldots, 100\}$, respectively.

As shown in Table 2, the proposed method provides a satisfied *TNR*=98.53%, which means that nearly all the samples of the unknown classes are correctly detected, and the *TPR* is 88.57% which means that most of the samples of known classes can be identified. The *TPR* of one-class SVM is large whereas its *TNR* is only 4.14%, indicating that most of the samples of unknown classes cannot be detected. The isolation forest gets the smallest *TPR* indicating only about 61.32% samples of known classes can be identified. Moreover, the proposed method gets a larger F_1 score which also indicates that its performance is better than that of one-class SVM and isolation forest.

(2) Fault classification

The fault classification performance of the developed multi-head DNN is evaluated considering the "accuracy" of classification on the known class samples in the test set, which computes the proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}. \tag{16}$$

Table 3 shows the obtained results. The fault classification performance of the proposed method is satisfactory, characterized by a 100% accuracy. The proposed method has been compared with two state-of-the-art

Table 2 Comparison of performance of novelty detection

	The proposed method	One-class SVM fed by multi-head DNN features	Isolation forest fed by multi-head DNN features
TPR(%)	88.57	97.69	61.32
TNR(%)	98.53	4.14	100
F_1 score	0.93	0.65	0.76

Table 3 Comparison of performance of classification

Performance	Accuracy (%)
The proposed method	100
SVM fed by multi-head DNN features	100
ANN fed by multi-head DNN features	99.85
kNN fed by multi-head DNN features	100

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 8 of 12

fault diagnostics methods based on the use of a SVM, a 1-hidden-layer ANN and a k-nearest neighbor (kNN) model. Since SVMs, ANNs and kNNs are shallow models which prefer low-dimensional input data, the $D_3 = 10$ dimensional feature vectors $\boldsymbol{a}_i^{(3)}$, $i = 1, \ldots, N$, extracted by the encoder are used as their input. The RBF kernel is employed for the SVM and the regularization parameter of SVM is set to 1, which is optimized by considering as possible options $\{0.1, 0.2, \ldots, 5\}$. The number of hidden neurons of the ANN has been optimized by trial-and-error considering as possible options $\{6, 7, \ldots, 50\}$. An ANN model with layers of (10, 10, 7) neurons and ReLU activation functions have been selected. The number of reference neighbors of kNN is set to 3, which has been optimized by trial-and-error considering as possible options $\{1, 2, \ldots, 5\}$. The accuracies of the SVM, ANN and kNN are 100%, 99.85% and 100%, respectively, which are comparable with the proposed methods. These results confirm that the proposed method is accurate on classification problem and the features extracted by the multi-head DNN is discriminative enough to help shallow models to achieve good performance.

(3) Selection of hyperparameters and activation function

The proposed method has six parameters. The λ, β, η_1, η_2 are used to balance the values of terms in the cost function (Eq. (10)) and are set by calculating the magnitude ratio of the terms in the training phase. The p_{drop} is used to prevent overfitting of the classification module and is typically suggested to be around 0.5. We selected p_{drop} to be 0.3 by experience since it does not influence the classification accuracy a lot.

The sparsity proportion p is the most sensitive hyperparameter for SAE-based DNNs. During model training, a validation set formed by 5% of the training data is randomly selected, and p is selected based on the performance over the validation set considering the possible options $\{0.05, 0.1, 0.2, 0.3, 0.4\}$. Table 4 reports the performance of the proposed method on validation set regarding different values of p. And $p = 0.05$ with the smallest reconstruction error and largest classification accuracy is selected.

In addition, we have investigated the use of a more cutting-edge activation function, Leaky ReLU, in the proposed method instead of ReLU. Leaky ReLU is a variant of ReLU, which maps negative inputs to small negative values instead of zeros. However, with Leaky ReLU, the proposed method gets F_1 score 0.74, *TNR* 66.74%, *TPR* 80.40% and classification accuracy 99.93%. Compared with the results obtained with ReLU (Tables 2 and 3), the performance of Leaky ReLU is comparable regarding the classification accuracy, but much worse with respect to novelty detection. A possible reason is that the input of the multi-head DNN is the FFT coefficients which are positive. The Leaky ReLU keeps negative values during the computation, which weakens the data reconstruction and leads to poorer novelty detection.

4.2 Experiment Evaluation for 3D Printer Diagnostics

In this Section, the proposed method is applied to diagnose the fault of a delta 3D printer (SLD-BL600-6) [52]. The extruder nozzle of the delta 3D printer was controlled to perform a predefined circular movement with a radius 75mm. A multi-channel attitude sensor was mounted on the moving platform to monitor its 3-axial angular acceleration, vibration acceleration and magnetic field intensity (Figure 5).

The wear of joint bearings and synchronous belts were considered as faulty conditions. The faults of joint bearings were introduced by loosening the screw of each joint bearing by half-turn, i.e., 0.35 mm. And the faults of synchronous belts were injected by relaxing the length of two teeth, i.e., 3 mm, for each belt. In each fault condition, we consider exclusively the fault of one joint bearing or one synchronous belt. As listed in Table 5, 15 faulty conditions are simulated in total, including faults of 12 joint bearings and 3 synchronous belts. Printing tests were performed under these faulty conditions and we found that the printing quality of the
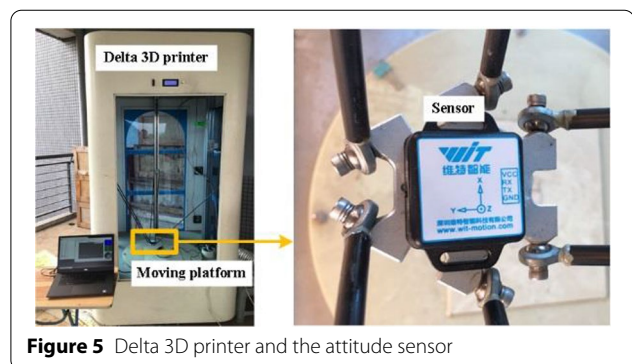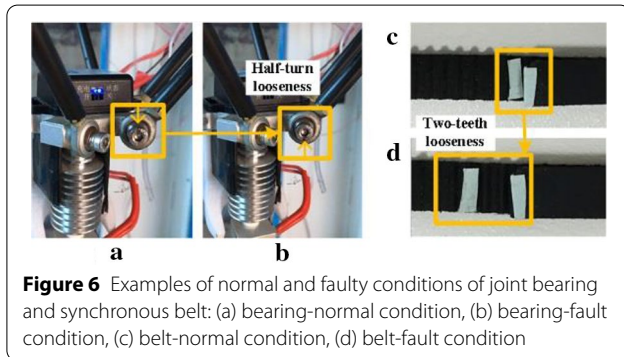
Table 4 Performance of the proposed method on validation set regarding values of p

p	Reconstruction error	Accuracy (%)
0.05	1.92×10^{-4}	100
0.1	2.20×10^{-4}	100
0.2	2.38×10^{-4}	99.82
0.3	2.62×10^{-4}	100
0.4	2.56×10^{-4}	99.92

Figure 5 Delta 3D printer and the attitude sensor

Yang *et al. Chin. J. Mech. Eng.* (2021) 34:54

Page 9 of 12

Table 5 3D printer conditions with assigned labels

Condition	Label/Class	Number of examples
Normal	1	759
Joint bearing A–L	2–13	759 for each class
Belt M–O	14–16	759 for each class

Figure 6 Examples of normal and faulty conditions of joint bearing and synchronous belt: (a) bearing-normal condition, (b) bearing-fault condition, (c) belt-normal condition, (d) belt-fault condition

Figure 7 Case 2: box plot of reconstruction errors

Table 6 Comparison of performance of novelty detection

	The proposed method	One-class SVM fed by multi-head DNN features	Isolation forest fed by multi-head DNN features
TPR(%)	98.41	98.68	97.55
TNR(%)	94.69	32.67	58.04
F₁ score	0.97	0.79	0.81

3D printer was affected seriously. Figure 6 shows examples of the normal mode and faulty mode of the joint bearing and synchronous belt, respectively.

The 9-channel monitoring data were collected under normal and all the faulty conditions at the sampling frequency of 100 Hz. For each condition, an experiment was performed to collect monitoring data for 20 circular movements, each channel of which contains 32400 data points. The data were then divided into 253 samples using a non-overlapping fixed-length time window, each channel of which contains 128 data points. The experiment was repeated for 3 times, therefore, 759 samples were collected for each condition. For each sample, the FFT was implemented on each of its channels to get the 128 Fourier amplitude coefficients. Considering the symmetry of the coefficients, the first 64 coefficients are used for each channel. Then, coefficients of all the channels are concatenated to form a 576-dimensional vector, which is used as the representation of the sample.

We assume normal and joint bearing faults are already known conditions (classes 1–13), and the synchronous belt faults (classes 14–16) are unknown conditions. The training set is composed of 70% of the data randomly selected from classes 1–13, respectively. The remaining data of classes 1–13 and all the data of classes 14–16 are used as the test set. During the training of the multi-head DNN, 5% of the training data is randomly selected as the validation set to prevent overfitting. The developed multi-head DNN is formed by an encoder with an input layer of $D = 576$ neurons and $L = 3$ hidden layers of $D_1 = 600$, $D_2 = 100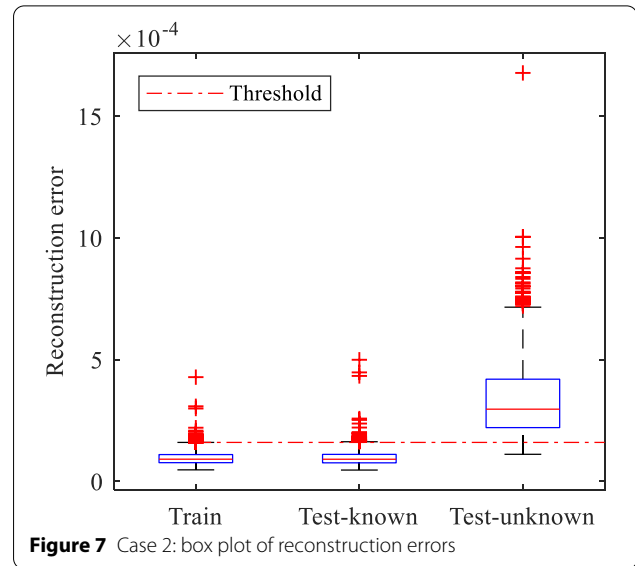$, $D_3 = 30$ neurons, a symmetric decoder, and a classification module with a softmax layer of 13 neurons associated with the known 1–13 classes in the training set. The hyperparameters of the multi-head DNN are set as follows: $p = 0.05$, $\lambda = 1 \times 10^{-7}$, $\beta = 1$, $\eta_1 = 80$, $\eta_2 = 1$ and $p_{drop} = 0.3$.

(1) Novelty detection

The box plot of reconstruction errors on different datasets is shown in Figure 7. The threshold for novelty detection, δ, is set to 1.60×10^{-4} based on the reconstruction errors on the training set, as described in Eq. (12). The known and unknown class test samples are well separated by the threshold, since the majority of reconstruction errors of known class test samples are smaller than the threshold whereas those of unknown class test samples are larger than it.

Table 6 shows the result of novelty detection of the proposed method and those of the one-class SVM and isolation forest. The $D_3 = 30$ dimensional feature vectors $\boldsymbol{a}_i^{(3)}$, $i = 1, \ldots, N$, extracted by the encoder from the training set are used as the input of one-class SVM and isolation forest. A one-class SVM model with the RBF kernel and $\nu = 0.01$ has been selected. The parameter ν has been optimized to maximize the F_1 score by trial-and-error considering as possible options $\{0, 0.01, \ldots, 0.2\}$,

respectively. The parameter of isolation forest N_{tree} is chosen as 50, which has been optimized with the objective of maximizing the F_1 score by trial-and-error considering as possible options $\{10, 20, \ldots, 100\}$, respectively.

Both the *TNR* and *TPR* of the proposed method are larger than 90%, which means that most of the samples of the unknown classes and known classes are correctly detected. The *TPR* of one-class SVM is 98.68% whereas the *TNR* is 32.67%, indicating that more than half of the samples of unknown classes cannot be detected. With similar problem, only 58.04% of the samples of unknown classes can be detected by the isolation forest. Moreover, the F_1 score of the proposed method is larger than that of the one-class SVM, indicating that its performance is better.

(2) Fault classification

Table 7 shows the obtained results. The fault classification performance of the proposed method is 97.56%. The proposed method has been compared with fault diagnostic methods based on the use of a SVM, a 1-hidden-layer ANN and a kNN model. Similar to Section 4.1, the $D_3 = 30$ dimensional feature vectors $\boldsymbol{a}_i^{(3)}$, $i = 1, \ldots, N$, extracted by the encoder are used as the input of SVM, ANN and kNN. The RBF kernel is employed for the SVM and the regularization parameter of SVM is set to 1, which is optimized by considering as possible options $\{0.1, 0.2, \ldots, 5\}$. The number of hidden neurons of the ANN has been optimized by trial-and-error considering as possible options $\{6, 7, \ldots, 80\}$. An ANN model with layers of (30, 30, 13) neurons and ReLU activation functions have been selected. The number of reference neighbors of kNN is set to 3, which has been optimized by trial-and-error considering as possible options $\{1, 2, \ldots, 5\}$. The accuracy of SVM is 97.56%, the same as that of the proposed method. The accuracies of ANN and kNN are 96.98% and 96.92%, respectively, which are slightly worse. These results confirm that the classification performance of the proposed method is no less than the other commonly used diagnostic methods.

5 Conclusions

This paper contributes to addressing the problem of fault diagnostics with novelty detection capability based on the use of SAE-based multi-head DNNs. The proposed method allows jointly performing two tasks: i) data reconstruction for novelty detection and ii) classification for diagnostics, using a single model, where features shared by these two tasks are automatically extracted from high-dimensional data. Furthermore, the use of ReLU activation function allows the reduction of computational burden by direct training of the DNN, instead of requiring conventional training procedures including pre-training and fine-tuning.

The fault diagnostics method has been verified on two case studies. The results obtained in the case studies show that: i) the proposed method can be applied for novelty detection of the machinery where multiple conditions are already known, and performs significantly better than one-class SVM and isolation forest; ii) the diagnostics accuracy of the proposed method is satisfactory, no less than the SVM and ANN-based fault diagnostic methods.

The setting of hyperparameters of the multi-head DNN is based on trial-and-error and experience. In future studies, we will focus on designing a systematic strategy for the hyperparameter setting to further facilitate the application of the proposed method. Furthermore, the proposed method detects novelties based on the reconstruction error, which is effective but indirect. A more direct way is to extract very representative features of available data, which should be compact within a clear boundary, but still be separable between those of different known types of defects. And the boundary could be employed for the detection of novelties. Therefore, the design of advanced cost function for representative feature extraction will be considered in our future work.

Authors' contributions
ZY wrote the manuscript and did the analysis. DG and YZ revised the manuscript. JL and SZ contributed to the experiments. CL was in charge of the whole work. All authors read and approved the final manuscript.

Authors' Information
Zhe Yang received the B.E. degree in Measurement & Control & Instrument and M.Sc. degree in Mechanical Engineering from *Xi'an Jiaotong University, Xi'an, China*, in 2012 and 2015, respectively. He received the Ph.D. degree in Mechanical Engineering from *Politecnico di Milano, Milan, Italy*, in 2020. He is currently a Postdoctoral Fellow with *Xi'an Jiaotong University, Xi'an, China*, and *Dongguan University of Technology, Dongguan, China*. His research interests include the development of methods and techniques for prognostics and health management of industrial components.

Dejan Gjorgjevikj received the B.Sc. degree in Electrical and Computer Engineering, Ms.El. Eng. and Ph.D. degrees in Computer Engineering from *Ss. Cyril and Methodius University, Skopje, Republic of Macedonia*, in 1992, 1997 and 2004, respectively. He is currently a Full Professor of Computer Science and Engineering with *Ss. Cyril and Methodius University, Skopje, Republic of Macedonia*. His research interests include pattern recognition, image processing, data mining and software engineering.

Table 7 Comparison of performance of classification

Performance	Accuracy (%)
The proposed method	97.56
SVM fed by multi-head DNN features	97.56
ANN fed by multi-head DNN features	96.98
kNN fed by multi-head DNN features	96.92

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 11 of 12

Jianyu Long received the B.E. and Ph.D. degrees in metallurgical engineering from *Chongqing University, Chongqing, China*, in 2012 and 2017, respectively.He is currently a Teacher of Industrial Engineering with the *School of Mechanical Engineering, Dongguan University of Technology, Dongguan, China*. From 2015 to 2016, he was a Visiting Scholar with the Center for Applied Optimization, *Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA*. His research interests include production scheduling, and prognostics and system health management.

Yanyang Zi received the B.Eng. degree in manufacturing processes and equipment and the Ph.D. degree in mechanical engineering from *Xi'an Jiaotong University, Xi'an, China*, in 1993 and 2001, respectively. He is currently a Full Professor of mechanical engineering with *Xi'an Jiaotong University*, *Xi'an, China*. His research interests include condition monitoring, fault diagnosis, prognostics, and health management of rotating machines. Dr. Zi is a Senior Member of the Chinese Mechanical Engineering Society. He is an Editorial Board Member of the International Journal of Comprehensive Engineering.

Shaohui Zhang received the B.E. degree in vehicle engineering from *Huaqiao University, Quanzhou, China*, in 2009, and the Ph.D. degree in vehicle engineering from the *South China University of Technology, Guangzhou, China*, in 2014. He is currently a Professor of Mechanical Engineering with the *School of Mechanical Engineering, Dongguan University of Technology, Dongguan, China*. His research interests include machine learning, fault diagnosis, and remaining useful life prediction.

Chuan Li received the Ph.D. degree in mechatronics engineering from *Chongqing University, Chongqing, China*, in 2007. He was a Postdoctoral Fellow with the *University of Ottawa, Ottawa, ON, Canada*, a Research Professor with *Korea University, Seoul, South Korea*, and a Senior Research Associate with the *City University of Hong Kong, Hong Kong, China*. He is currently a Professor of Mechanical Engineering with the *School of Mechanical Engineering, Dongguan University of Technology, Dongguan, China*. His research interests include prognostics and health management, and intelligent systems.

Author Details
[1]School of Mechanical Engineering, Dongguan University of Technology, Dongguan 523808, China. [2]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China. [3]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia.

References

[1] E Zio. Some challenges and opportunities in reliability engineering. *IEEE Trans. Reliab.*, 2016, 65(4): 1769-1782, https://doi.org/10.1109/TR.2016.2591504.

[2] D Wang, K-L Tsui, Q Miao. Prognostics and health management: A review of vibration based bearing and gear health indicators. *IEEE Access*, 2018. https://doi.org/10.1109/ACCESS.2017.2774261.

[3] J Zhong, J Long, S Zhang, et al. Flexible kurtogram for extracting repetitive transients for prognostics and health management of rotating components. *IEEE Access*, 2019, 7: 55631-55639, https://doi.org/10.1109/ACCESS.2019.2912716.

[4] P Baraldi, F Cadini, F Mangili, et al. Model-based and data-driven prognostics under different available information. *Probabilistic Eng. Mech.*, 2013, 32: 66-79, https://doi.org/10.1016/j.probengmech.2013.01.003.

[5] H Wang, J Xu, R Yan, et al. A new intelligent bearing fault diagnosis method using SDP representation and SE-CNN. *IEEE Trans. Instrum. Meas.*, 2020, 6(5): 2377-2389, https://doi.org/10.1109/TIM.2019.2956332.

[6] W Chine, A Mellit, V Lughi, et al. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. *Renew. Energy*, 2016, 90: 501-512, https://doi.org/10.1016/j.renene.2016.01.036.

[7] H Malik, Y Pandya, A Parashar, et al. Feature extraction using EMD and classifier through artificial neural networks for gearbox fault diagnosis. *Applications of Artificial Intelligence Techniques in Engineering*, 2019, 697: 309-317, https://doi.org/10.1007/978-981-13-1822-1_28.

[8] K He, Z Yang, Y Bai, et al. Intelligent fault diagnosis of delta 3D printers using attitude sensors based on support vector machines. *Sensors (Switzerland)*, 2018. https://doi.org/10.3390/s18041298.

[9] X Li, S Wu, X Li, et al. Particle swarm optimization-support vector machine model for machinery fault diagnoses in high-voltage circuit breakers. *Chinese J. Mech. Eng.*, 2020, 33: 6, https://doi.org/10.1186/s10033-019-0428-5.

[10] G Liu, H Li, W Liu. Bearing fault detection in varying operational conditions based on empirical mode decomposition and random forest. *2018 Prognostics and System Health Management Conference (PHM-Chongqing 2018)*, 2018: 851-854, https://doi.org/10.1109/PHM-Chongqing.2018.00152.

[11] Y Hu, S Thang, A Jiang, et al. A new method of wind turbine bearing fault diagnosis based on multi-masking empirical mode decomposition and fuzzy C-means clustering. *Chinese J. Mech. Eng.*, 2019, 32: 46, https://doi.org/10.1186/s10033-019-0356-4.

[12] P Baraldi, F Cannarile, F Di Maio, et al. Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions. *Eng. Appl. Artif. Intell.*, 2016, 56: 1-13, https://doi.org/10.1016/j.engappai.2016.08.011.

[13] J Yu, M Bai, G Wang, et al. Fault diagnosis of planetary gearbox with incomplete information using assignment reduction and flexible naive Bayesian classifier. *J. Mech. Sci. Technol.*, 2018, 32(1): 37-47, https://doi.org/10.1007/s12206-017-1205-y.

[14] A Yang, Y Wang, Y Zi, et al. An enhanced trace ratio linear discriminant analysis for fault diagnosis: An illustrated example using HDD data. *IEEE Trans. Instrum. Meas.*, 2019, 68(12): 4629-4639, https://doi.org/10.1109/TIM.2019.2900885.

[15] M Tan, J Li, X Chen, et al. Power grid fault diagnosis method using intuitionistic fuzzy Petri nets based on time series matching. *Complexity*, 2019, https://doi.org/10.1155/2019/7890652.

[16] K Li, L Su, J Wu, et al. A rolling bearing fault diagnosis method based on variational mode decomposition and an improved kernel extreme learning machine. *Appl. Sci.*, 2017, https://doi.org/10.3390/app7101004.

[17] S Khan, T Yairi. A review on the application of deep learning in system health management. *Mech. Syst. Signal Process.*, 2018, 107: 241-265, https://doi.org/10.1016/j.ymssp.2017.11.024.

[18] R Zhao, R Yan, Z Chen, et al. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.*, 2019, 115: 213-237, https://doi.org/10.1016/j.ymssp.2018.05.050.

[19] C Li, S Zhang, Y Qin, et al. A systematic review of deep transfer learning for machinery fault diagnosis. *Neurocomputing*, 2020, 407: 121-135, https://doi.org/10.1016/j.neucom.2020.04.045.

[20] R Yan, F Shen, C Sun, et al. Knowledge transfer for rotary machine fault diagnosis. *IEEE Sens. J.*, 2020, 20(15): 8374-8393, https://doi.org/10.1109/JSEN.2019.2949057.

[21] Y LeCun, Y Bengio, G Hinton. Deep learning. *Nature*, 2015, 521(7553): 436-444, https://doi.org/10.1038/nature14539.

[22] J Dai, J Tang, S Huang, et al. Signal-based intelligent hydraulic fault diagnosis methods: Review and prospects. *Chinese J. Mech. Eng.*, 2019, https://doi.org/10.1186/s10033-019-0388-9.

[23] F Jia, Y Lei, J Lin, et al. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.*, 2016, 72–73: 303-315, https://doi.org/10.1016/j.ymssp.2015.10.025.

[24] Z Chen, W Li. Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network. *IEEE Trans. Instrum. Meas.*, 2017, 66(7): 1693-1702, https://doi.org/10.1109/TIM.2017.2669947.

Yang *et al. Chin. J. Mech. Eng.*   (2021) 34:54

Page 12 of 12

[25] C Lu, Z-Y Wang, W-L Qin, et al. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. *Signal Processing*, 2017, 130: 377-388, https://doi.org/10.1016/j.sigpro.2016.07.028.

[26] S-Y Shao, W-J Sun, R-Q Yan, et al. A deep learning approach for fault diagnosis of induction motors in manufacturing. *Chinese J. Mech. Eng.*, 2017, 30(6): 1347-1356, https://doi.org/10.1007/s10033-017-0189-y.

[27] L-H Wang, X-P Zhao, J-X Wu, et al. Motor fault diagnosis based on short-time fourier transform and convolutional neural network. *Chinese J. Mech. Eng.*, 2017, 30(6): 1357-1368, https://doi.org/10.1007/s10033-017-0190-5.

[28] G Jiang, H He, J Yan, et al. Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. *IEEE Trans. Ind. Electron.*, 2019, 66(4): 3196-3207, https://doi.org/10.1109/TIE.2018.2844805.

[29] Y Chang, J Chen, C Qu, et al. Intelligent fault diagnosis of Wind Turbines via a deep learning network using parallel convolution layers with multi-scale kernels. *Renew. Energy*, 2020, 153: 205-213, https://doi.org/10.1016/j.renene.2020.02.004.

[30] M Yuan, Y Wu, L Lin. Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network. *2016 IEEE/CSAA International Conference on Aircraft Utility Systems (AUS)*, 2016: 135-140.

[31] X Chen, B Zhang, D Gao. Bearing fault diagnosis base on multi-scale CNN and LSTM model. *J. Intell. Manuf.* (2021), 32: 971-987. https://doi.org/10.1007/s10845-020-01600-2.

[32] J Long, S Zhang, C Li. Evolving deep echo state networks for intelligent fault diagnosis. *IEEE Trans. Ind. Informatics*, 2020, 16(7): 4928-4937, https://doi.org/10.1109/TII.2019.2938884.

[33] J Long, Z Sun, C Li, et al. A novel sparse echo autoencoder network for data-driven fault diagnosis of delta 3-D printers. *IEEE Trans. Instrum. Meas.*, 2020, 69(3): 683-692, https://doi.org/10.1109/TIM.2019.2905752.

[34] S Zhang, Z Sun, C Li, et al. Deep hybrid state network with feature reinforcement for intelligent fault diagnosis of delta 3-D printers. *IEEE Trans. Ind. Informatics*, 2020, 16(2): 779-789, https://doi.org/10.1109/TII.2019.2920661.

[35] Y Bengio, A Courville, P Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, 35(8): 1798-1828, https://doi.org/10.1109/TPAMI.2013.50.

[36] B Gao, W L Woo, Y He, et al. Unsupervised sparse pattern diagnostic of defects with inductive thermography imaging system. *IEEE Trans. Ind. Informatics*, 2016, 12(1): 371-383, https://doi.org/10.1109/TII.2015.2492925.

[37] A Ng. *Sparse autoencoder*. CS294A Lect. Notes, 2011, 72: 1–19.

[38] J Long, J Mou, L Zhang, et al. Attitude data-based deep hybrid learning architecture for intelligent fault diagnosis of multi-joint industrial robots. *J. Manuf. Syst.*, 2020, https://doi.org/10.1016/j.jmsy.2020.08.010.

[39] E Principi, D Rossetti, S Squartini, et al. Unsupervised electric motor fault detection by using deep autoencoders. *IEEE-CAA J. Autom. Sin.*, 2019, 6(2): 441-451, https://doi.org/10.1109/JAS.2019.1911393.

[40] Y Qi, C Shen, D Wang, et al. Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery. *IEEE Access*, 2017, 5: 15066-15079, https://doi.org/10.1109/ACCESS.2017.2728010.

[41] Y Qu, M He, J Deutsch, et al. Detection of pitting in gears using a deep sparse autoencoder. *Appl. Sci. (Switzerland)*, 2017, 7(5): 515. https://doi.org/10.3390/app7050515.

[42] M Sohaib, J-M Kim. Reliable fault diagnosis of rotary machine bearings using a stacked sparse autoencoder-based deep neural network. *Shock Vib.*, 2018, https://doi.org/10.1155/2018/2919637.

[43] H O A Ahmed, M L D Wong, A K Nandi. Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete features. *Mech. Syst. Signal Process.*, 2018, 99: 459-477, https://doi.org/10.1016/j.ymssp.2017.06.027.

[44] H Zhu, J Cheng, C Zhang, et al. Stacked pruning sparse denoising autoencoder based intelligent fault diagnosis of rolling bearings. *Appl. Soft Comput.*, 2020, 88, 106060, https://doi.org/10.1016/j.asoc.2019.106060.

[45] L Xu, M Cao, B Song, et al. Open-circuit fault diagnosis of power rectifier using sparse autoencoder based deep neural network. *Neurocomputing*, 2018, 311: 1-10, https://doi.org/10.1016/j.neucom.2018.05.040.

[46] Y Zheng, T Wang, B Xin, et al. A sparse autoencoder and softmax regression based diagnosis method for the attachment on the blades of marine current turbine. *Sensors*, 2019, 19(4): 826, https://doi.org/10.3390/s19040826.

[47] B A Olshausen, D J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.*, 1997, 37(23): 3311-3325.

[48] D F Williamson, R A Parker, J S Kendrick. The box plot: a simple visual method to interpret data. *Ann. Intern. Med.*, 1989, 110(11): 916-921.

[49] X Lou, K A Loparo. Bearing fault diagnosis based on wavelet transform and fuzzy inference. *Mech. Syst. Signal Process.*, 2004, 18(5): 1077-1095, https://doi.org/10.1016/S0888-3270(03)00077-3.

[50] K Yan, Z Ji, W Shen. Online fault detection methods for chillers combining extended kalman filter and recursive one-class SVM. *Neurocomputing*, 2017, 228: 205-212.

[51] S Amraee, A Vafaei, K Jamshidi, et al. Abnormal event detection in crowded scenes using one-class SVM. *Signal, Image Video Process.*, 2018, 12(6): 1115-1123.

[52] S Zhang, Z Sun, J Long, et al. Dynamic condition monitoring for 3D printers by using error fusion of multiple sparse auto-encoders. *Comput. Ind.*, 2019, 105: 164-176, https://doi.org/10.1016/j.compind.2018.12.004.