Deep Coupling Autoencoder for Fault Diagnosis with Multimodal Sensory Data

Meng Ma, Chuang Sun, Xuefeng Chen, *Member, IEEE*

***Abstract*—Effective fault diagnosis of rotating machinery has multifarious benefits such as improved safety, enhanced reliability and reduced maintenance cost for complex engineered systems. With many kinds of installed sensors for conducting fault diagnosis, one of the key tasks is to develop data fusion strategies which can effectively handle multimodal sensory signals. Most traditional methods use hand-crafted statistical features and then combine these multimodal features simply by concatenating them into a long vector to achieve data fusion. The present study proposes a deep coupling autoencoder (DCAE) model which handles the multimodal sensory signals not residing in a commensurate space, such as vibration and acoustic data, and integrates feature extraction of multimodal data seamlessly into data fusion for fault diagnosis. Specifically, a coupling autoencoder (CAE) is constructed to capture the joint information between different multimodal sensory data, and then a DCAE model is devised for learning the joint feature at a higher level. The CAE is developed by coupling hidden representations of two single-modal autoencoders, which can capture the joint information from multimodal data. The performance of the proposed method is evaluated by two experiments, which shows that the DCAE model succeeds in efficiently utilizing multi-source sensory data to perform accurate fault diagnosis. Compared with other methods, the proposed method exhibits better performance.**

***Index Terms*—fault diagnosis, multimodal information fusion, coupling autoencoder, deep learning**

I. Introduction

INTELLIGENT fault diagnosis of mechanical components plays an important role in prognostics and health management (PHM), which is a cost-effective maintenance strategy where maintenance schedules are predicted based on the results provided from diagnostics and prognostics [1]. No matter how good a machine design is, some components deteriorate over time since they are under certain stress or load when working, usually involving randomness [2]. Unexpected breakdowns can be prohibitively costly because of the lost production, failed shipping schedule, and poor customer satisfaction [3]. In order to avoid such problems, it is essential to identify the current health state of system through effective fault diagnosis. Health state classification focuses on using sensory information acquired from different kinds of sensors to reveal the health state of system components [4]. Safety-critical mechanical systems and structures – such as aero engines, aircraft structures, rotary equipment, bearings and gears – have benefited from advanced sensor systems developed specifically for fault diagnosis， and health and usage monitoring [5]. For complex engineered systems, a single sensor can never acquire all the necessary data. In many situations, sophisticated and smart sensors installed on the complex equipment (e.g., aircraft engines, wind turbines, and power distribution transform) are able to collect sensory data related to status and performance [6]. Thus with a large amount of multi-sensory data available, data fusion techniques are widely employed because of their inherent superiority to fuse and reveal the information from multiple sensors [7], but effective diagnosis of current health state based on the sensory data is still an intricate problem and remains as a major challenge [8]. The signals collected from different kinds of sensors, such as vibration signals, acoustic signals and temperature signals, are characterized by very distinct statistical and physical properties. Thus we should consider the fact that they are known as multimodal sensory signals [9].

Different types of information carried in multimodal signals show difference in describing the machine performance. For example, the sensory signals, such as vibration, sound, acoustic emission and wear debris have been applied in monitoring the health condition of machine components. It is known that the local faults in components, such as rolling element bearings and gears, lead to impacts, which can be detected in the multimodal sensory signals. However, the acoustic emission signals are used to detect low-speed bearing faults because they are capable of capturing dynamic activity in low-energy data [10]. Although the vibration and sound signals are both effective to detect various faults, the latter can be acquired at a distance from the machine. So sound sensors have an advantage of easier installation compared with vibration sensors. Wear debris signals with a better trend can provide a robust indicator to indicate the degree of fault propagation [11]. Thus, regarding the multimodal data, one of the most essential task is to develop

Manuscript received March 27, 2017; revised July 29, 2017, October 24, 2017; accepted December 25, 2017.

This work was supported in part by the National Key Basic Program of China under Grant 2015CB054700, in part by the National Natural Science Foundation of China (51705398), in part by the China Postdoctoral Science Foundation (2017M613117), and in part by the Fundamental Research Funds for the Central Universities (XJJ2017019). (Corresponding author: Xuefeng Chen)

Meng Ma, Chuang Sun, and Xuefeng Chen are with the State Key Laboratory for Manufacturing Systems Engineering, Xi’an Jiaotong University, Xi’an 710049, China (e-mail: mamxjscu@stu.xjtu.edu.cn; ch.sun@xjtu.edu.cn; chenxf@mail.xjtu.edu.cn).

.

new algorithms that can effectively deal with multimodal sensory signals.

The development of deep learning provides a promising opportunity for big data analysis and processing, which has dramatically improved the performance in speech recognition, visual object recognition, object detection and many other domains including drug discovery and genomics [12]. Deep learning is capable of yielding useful and powerful features from the original signals that can ultimately be useful for classification or regression [13].The ability to learn complexity of deep learning grows exponentially with the increase in heterogeneity and dimensionality of acquired sensory data [14]. Deep learning represents an attractive option to process mechanical big data for multimodal information fusion because it aims at automatic discovery of important features independent of application domains.

A wide range of practical applications for health state classifications and failure diagnosis has been reported in the literature. Gebraeel *et al.* [15] developed neural-network-based models for predicting bearing failures. Yan and Gao [16] presented a new approach to monitoring machine health based on the Approximate Entropy (ApEn). Wang *et al.* [17] proposed an enhanced quality-related fault detection approach based on orthogonal signal correction (OSC) and modified-PLS. To identify the most effective features for enhancement of fault classification accuracy in rotating machines, Li *et al.* [18] proposed a novel dimension reduction algorithm, referred to as the nearest and farthest distance preserving projection (NFDPP). To characterize the subspace, a nonlinear subspace distance is defined for structural health monitoring in [19][20]. Lei *et al* [21] proposed a two-stage learning method for intelligent diagnosis of machines. A method based on discriminative deep belief network and ant colony optimization was put forward for predicting the mechanical health status in [22]. Despite the above-mentioned achievements, these methods only utilize the vibration signals collected from a single sensor to perform health state classification.

The multi-sensory data can be fused at different levels: (a) sensor-level, (b) feature-level, and (c) decision-level [23]. Each fusion strategy has its own merits and limitations. The sensor-level fusion can make full use of the information however it may not be very discriminative in nature. Much less information is preserved when signals are fused at the decision level compared with that at the feature-level. To deal with the task of multi-sensor data fusion, a lot of methods have been developed in this area. Basir et al. [24] investigated the use of Dempster-Shafer evidence theory as a tool for modeling and fusing multi-sensory pieces of evidence pertinent to engine quality. In [25], a novel method for a high-level latent and shared feature representation from neuroimaging modalities via deep learning was proposed. In [26], the group sparse representation based classifier (GSRC) was proposed for multimodal multi-feature biometric recognition. To process the huge amount of data captured by the limited lifetime biosensors, Habib et al. [27] presented a data fusion model on the coordinator level using a decision matrix and fuzzy set theory. In order to improve the fault reliability, a new multisensor data fusion technique based on sparse autoencoder and deep belief network (SAE-DBN) is improved in [28]. The effectiveness of SAE-DBN scheme is verified through bearing fault experiments on a bearing test platform. However, these fusion methods have certain limitations because the related information between different sensory signals has not been effectively explored.

In general, the fault diagnosis methods based on machine learning can be broadly categorized into supervised and unsupervised learning methods. This study focuses on the unsupervised learning through deep architecture in health state classification. The model of supervised learning defines the effect of one set of observations on another called label information. It is expected to learn a relationship between input values and desired output labels, and cannot cluster or classify data through discovering the powerful features on its own. Unsupervised learning process does not utilize labeled training data, which facilitates the learning of larger and more complex models, especially the deep architecture. The supervised learning has a limitation in learning models with deep hierarchies because the difficulty of learning task increases exponentially in the number of steps between observation data and label information.

Vibration and sound emission signals always carry dynamic information of the mechanical system and they are widely used in condition monitoring and fault diagnosis [29]. Therefore, finding the common feature representation, which combines the related information from vibration and acoustic signals, is helpful to improve the performance of information fusion. To address the challenge of multimodal data fusion, a new deep architecture called deep coupling autoencoder (DCAE) is proposed. In this study, the problems of both feature representation and multimodal data fusion for health state classification of mechanical components are considered. From the view of feature representation, unlike the traditional approaches that considered the low-level features, the proposed approach extracts high-level or abstract features to enhance its discriminative ability through exploiting a deep learning strategy. Meanwhile, from the view of a multimodal data fusion, it is noteworthy that the deep coupling autoencoder model combines the modality-specific feature extraction process and information fusion process, which is different from the conventional method that first extracts features and then fuses the information during classifier learning. It should be noted that once the process of feature extraction is finished, some common information between different modalities will be lost. Therefore, it is important to discover the common information by fully using original information in each modal data in the feature extraction process.

The rest of the paper is organized as follows. The proposed DCAE model is introduced in Section II. Section III presents in detail the application of the proposed deep architecture in health state classifications of gears and bearings. Comparisons and results with other methods are also shown in this section. Finally, some conclusions are summarized in Section IV.

II. Deep Coupling Autoencoder Model

A. Autoencoders

In the autoencoder (AE) framework, an AE is a discriminative graphical model which tries to reconstruct the input signals [30]. A feature-extracting function, denoted by f_θ, is called the encoder. It enables the straightforward and efficient computation of the feature vector $\boldsymbol{h}=f_\theta(\boldsymbol{x})$ from the input signals $\boldsymbol{x}$. For each sample $\boldsymbol{x}_i$ from the data set $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, the hidden layer is defined as follows:

$$\boldsymbol{h}_i=f_\theta(\boldsymbol{x}_i) \tag{1}$$

where $\boldsymbol{h}_i$ is a feature-vector or representation encoded from $\boldsymbol{x}_i$. The feature space is mapped back into input space by another function g_θ, called the decoder. Then the decoder process produce a reconstruction:

$$\boldsymbol{z}_i=g_\theta(\boldsymbol{h}_i) \tag{2}$$

The set of parameters θ of the encoder and decoder are trained simultaneously on the task of attempting to minimize the reconstruction error, a measure of difference between the input $\boldsymbol{x}$ and its reconstruction $\boldsymbol{z}$ over all the training datasets.

The basic training process of AE model is to find the value of parameter vector θ by minimizing the reconstruction error:

$$\theta=\arg\min\sum_{i=1}^{n}L(\boldsymbol{x}_i,\boldsymbol{z}_i) \tag{3}$$

where $\boldsymbol{x}_i$ is the training sample, n is the number of training samples and L is a function that measures the reconstruction error. Traditionally, squared error or cross-entropy is used to compute the reconstruction error:

$$L(\boldsymbol{x},\boldsymbol{z})=\|\boldsymbol{x}-\boldsymbol{z}\|^2 \tag{4}$$

$$L(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{x}^T\log(\boldsymbol{z})+(1-\boldsymbol{x})^T\log(1-\boldsymbol{z}) \tag{5}$$

Using the back propagation algorithm, this minimization is achieved by updating the parameters efficiently with stochastic gradient descent. The choice of function f_θ and g_θ depends largely on the input domain range and nature, and the most commonly used function of the encoder and decoder are affine mappings:

$$f_\theta(\boldsymbol{x})=sigm(W\boldsymbol{x}+\boldsymbol{b}) \tag{6}$$

$$g_\theta(\boldsymbol{h})=sigm(W'\boldsymbol{h}+\boldsymbol{d}) \tag{7}$$

where $sigm$ is the activation functions. In the model the parameter set is $\theta=\{W,\boldsymbol{b},W',\boldsymbol{d}\}$. The weight vectors of the encoder and decoder are W and W', the bias vectors $\boldsymbol{b}$ and $\boldsymbol{d}$.

B. Coupling Autoencoder

Since there is an inherent relation among multimodal signals acquired from different types of sensors, it is necessary to extract common features when performing representation learning from the multimodal signals. Thus a coupling autoencoder (CAE) model is constructed to deal with different multisensory data. As shown in Fig 1, the coupling autoencoder architecture is composed of two basic autoencoders. The two autoencoders are coupled by a predefined similarity measure on the hidden layers, which is used to capture the correlations from different modal signals. Each autoencoder of the CAE architecture is used to extract information from one single modality. To find the joint information in the process of feature learning, two autoencoders are coupled by a similarity measure, which denotes the relation among different modalities [31]. Although the two basic autoencoders have the same architecture, the weight parameters of two autoencoders are different from each other after the learning process. In this way, the joint information of signals from multi-sensors can be obtained using the trained model. That is to say, the multimodal data fusion has been done by the feature extraction process.

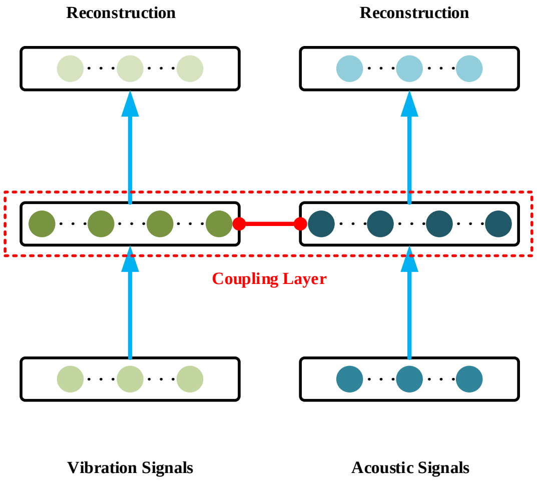


Fig. 1. Coupling autoencoder architecture.

The feature layers of the mapping function from inputs of two sensory datasets can be denoted as $f_v(x_v,\theta_v)$ and $f_a(x_a,\theta_a)$, where f_v is the vibration modality and f_a is the acoustic modality. The model parameters are denoted as θ_v and θ_a. The similarity measure between the hidden layers of two autoencoders is defined as follows:

$$S(x_v,x_a;\theta_v,\theta_a)=\left\|f_v(x_v,\theta_v)-f_a(x_a,\theta_a)\right\|^2 \tag{8}$$

where x_v and x_a are the representations of vibration signals and acoustic signals, respectively.

To learn the joint representation of the two modalities from the same object, a new loss function is constructed to train the model. The loss function of the coupling model can be defined as follows:

$$\begin{aligned} L(x_v,x_a;\theta_v,\theta_a)&=\alpha L_v(x_v;\theta_v)+\beta L_a(x_a;\theta_a)\\ &\quad+\gamma S(x_v,x_a;\theta_v,\theta_a) \end{aligned} \tag{9}$$

$$L_v(x_v;\theta_v)=\|x_v-z_v\|^2$$

$$L_a(x_a;\theta_a)=\|x_a-z_a\|^2$$

$$\alpha+\beta+\gamma=1$$

where L_v and L_a are the loss functions of the two autoencoders, which are responsible for two modalities

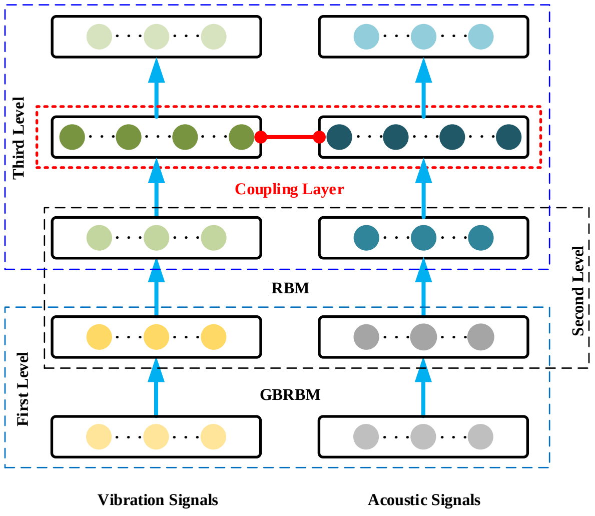


Fig. 2. Deep coupling autoencoder.

respectively, namely vibration and acoustic modality. z_v and z_a are the reconstruction data from the inputs x_v and x_a respectively. α, β and γ are the parameters that control the coupling model weights among the functions of reconstruction losses and similarity loss. The parameter γ indicates the similarity measure between two modal signals. If $\gamma = 0$, the loss function degenerates to the loss function of two separate autoencoders, thus it cannot learn the joint information between different modalities. On the contrary, if $\gamma = 1$, it means that the cost function only centers on the similar information of different modal signals while overlooking the reconstruction errors of the inputs. The vibration signals and acoustic signals collected from mechanical systems both carry useful information concerning the machine conditions. As they are closely allied together, the correlations between them can be captured through the similarity measure, which is helpful for health state assessment in a comprehensive way by information fusion strategy.

The training process can be achieved by back propagation algorithm. The gradient of the model loss function is calculated as follows:

$$\frac{\partial L}{\partial \theta_v} = \alpha \frac{\partial L_v}{\partial \theta_v} + \gamma \frac{\partial S}{\partial \theta_v} \tag{10}$$

$$\frac{\partial L}{\partial \theta_a} = \beta \frac{\partial L_a}{\partial \theta_a} + \gamma \frac{\partial S}{\partial \theta_a} \tag{11}$$

C. *Deep Coupling Autoencoder*

Since the signals acquired from different types of sensors are characterized by very distinct statistical properties that each modality may have a different kind of representation and correlational structure, it is difficult to jointly model the multimodalities with a shallow architecture. Thus a deep architecture, called deep coupling autoencoder, is proposed to address this problem. The advantage of the model lies in considering the correlation among different modalities when performing representation learning. As vibration and acoustic signals are real-valued, Gaussian Bernoulli Restricted Boltzmann Machine (GBRBM), an extension of binary RBM, is used to model the real-valued vectors.

```
Algorithm 1 RBM learning algorithm
    Input: training samples: $\boldsymbol{x}$, $\varepsilon$ .
    Output: the parameters of RBM: $\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}$.
    Update the parameters using gradient descent.
    for all training samples $\boldsymbol{x}$ do
        $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}$
        $\boldsymbol{h}^{(0)} \leftarrow \boldsymbol{sigmoid}\left(\boldsymbol{x}^{(0)}\boldsymbol{W} + \boldsymbol{c}\right) > \boldsymbol{rand}(\ )$
        $\boldsymbol{x}^{(1)} \leftarrow \boldsymbol{sigmoid}\left(\boldsymbol{h}^{(0)}\boldsymbol{W}^T + \boldsymbol{b}\right) > \boldsymbol{rand}(\ )$
        $\boldsymbol{h}^{(1)} \leftarrow \boldsymbol{sigmoid}\left(\boldsymbol{x}^{(1)}\boldsymbol{W} + \boldsymbol{c}\right) > \boldsymbol{rand}(\ )$
        $\boldsymbol{W} \leftarrow \boldsymbol{W} + \varepsilon\left(\boldsymbol{x}^{(0)}\boldsymbol{h}^{(0)} - \boldsymbol{x}^{(1)}\boldsymbol{h}^{(1)}\right)$
        $\boldsymbol{b} \leftarrow \boldsymbol{b} + \varepsilon\left(\boldsymbol{x}^{(0)} - \boldsymbol{x}^{(1)}\right)$
        $\boldsymbol{c} \leftarrow \boldsymbol{c} + \varepsilon\left(\boldsymbol{h}^{(0)} - \boldsymbol{h}^{(1)}\right)$
    end for
```

As Fig. 2 shows, the deep architecture has three stacked layers. After learning a Gaussian Bernoulli Restricted Boltzmann Machine, for the applications with continuous inputs, the activation probabilities of its hidden units are treated as the data for training the Restricted Boltzmann Machine (RBM) which is stacked one layer up. The second layer, used to learn higher-level features from input signals, consists of two basic RBMs. Then the outputs of the second layer RBM are used as the input for the third layer, the coupling layer. The training process of DACE is carried out by this efficient layer-by-layer greedy learning strategy, and this learning procedure is unsupervised and no label information is needed.

The basic RBM is an undirected graphical model with stochastic binary variables [32], and consists of two layers, namely visible layer $\boldsymbol{x}$ and hidden layer $\boldsymbol{h}$. The weight matrix between the two layers denoted by $\boldsymbol{W}$ is undirected. In addition, each unit has a bias. The probability distribution for an RBM is defined by its energy function as follows:

$$P(\boldsymbol{x},\boldsymbol{h}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{x},\boldsymbol{h})\right) \tag{12}$$

With this energy function, $E(\boldsymbol{x},\boldsymbol{h})$ and the partition function Z can be defined as follows:

$$E(\boldsymbol{x},\boldsymbol{h}) = -\sum\nolimits_i b_i x_i - \sum\nolimits_j c_j h_j - \sum\nolimits_{i,j} b_i w_{ij} c_j \tag{13}$$

$$Z(\boldsymbol{x},\boldsymbol{h}) = \sum\nolimits_{x,h} \exp\left(-E(\boldsymbol{x},\boldsymbol{h})\right) \tag{14}$$

where b and c are the biases of the visible layer and the hidden layer respectively. The sum over $\boldsymbol{x},\boldsymbol{h}$ denotes all the possible states of the model.

Under this model, if the visible units are binary-valued, the conditional probability can be written as:

$$P(x_i = 1 | h) = sigmoid\left(b_i + W_i h\right) \tag{15}$$

$$P(h_i = 1 | x) = sigmoid\left(c_i + W_i x\right) \tag{16}$$

where, *sigmoid*(⋅) is the sigmoid function.

The model is trained by maximizing the log probability of the training data, and the training process enables the model to

```
Algorithm 2 DCAE learning algorithm
        Input:  training samples: $x, v, \varepsilon$ .
        Output: the parameters of DCAE.
1:      Training the GBRBM in the first layer;
2:      Training the RBM in the second layer;
3:      Update the parameters of CAE using gradient descent
        with the loss function (9):
        $\theta_v = \theta_v - \varepsilon \frac{\partial L}{\partial \theta_v}$
        $\theta_a = \theta_a - \varepsilon \frac{\partial L}{\partial \theta_a}$
        where $\varepsilon$ is the learning rate, $\theta$ denotes the
        parameters of CAE model.
4:      Repeat steps 1 - 3 until convergence;
```

generate data like the training samples. Algorithm 1 shows the RBM learning process. ε is a learning rate and *rand*() produces random uniform numbers between 0 and 1.

The signals in practical applications are continuous, such as vibration and sound emission signals, therefore it is necessary to consider RBM with continuous visible neurons. The RBM and continuous data are connected by the visible layer. Accordingly, Gaussian neurons have been introduced in the visible layer to represent continuous inputs. The GBRBM has Gaussian distributed visible units and Bernoulli distributed hidden units [33]. The energy function of GBRBM is defined as follows:

$$E(\boldsymbol{x},\boldsymbol{h}) = \frac{1}{2\sigma^2}\sum_i x_i^2 - \frac{1}{\sigma^2}\left(\sum_i b_i x_i + \sum_j c_j h_j + \sum_{i,j} b_i w_{ij} c_j\right) \tag{17}$$

If the visible units are real-valued, the conditional probability can be written as:

$$P(x_i = 1 \mid h) = N\left(b_i + W_i h, \sigma^2\right) \tag{18}$$

$$P(h_i = 1 \mid x) = sigmoid\left(\frac{1}{\sigma^2}(c_i + W_i x)\right) \tag{19}$$

where $N(\cdot)$ is the Gaussian density.

In summary, an important character of the DCAE is to move away from separate feature representation and fusion. The training of the DCAE is carried out in a layer-wise manner from the bottom up. DCAE is capable of learning similar features from different modalities by mapping the multimodal signals into the same representation space. The joint features learned from the deep architecture can be used for health state classification. The algorithm of the proposed DCAE is shown in Algorithm 2. The gradient of the coupling autoencoder is a linear combination of two basic autoencoders, thus the updating process is converged.

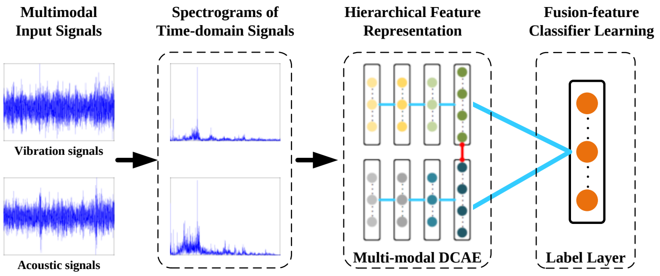
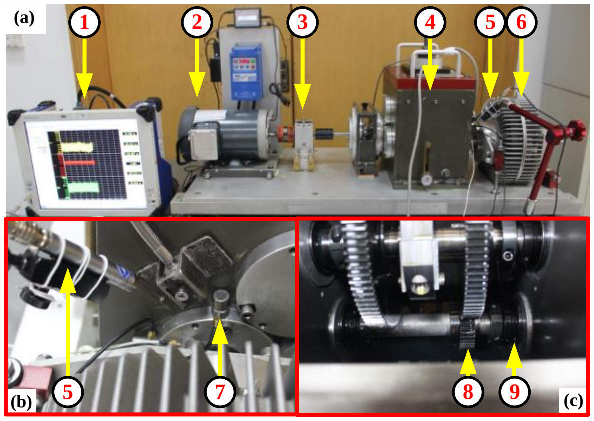
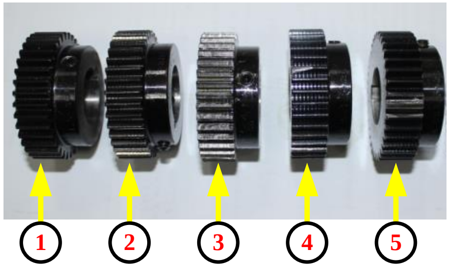

Multimodal Input Signals
Spectrograms of Time-domain Signals
Hierarchical Feature Representation
Fusion-feature Classifier Learning
Vibration signals
Acoustic signals
Multi-modal DCAE
Label Layer

Fig. 3. Schematic illustration of the proposed method



1: HBM data acquisition, 2: Driving motor, 3: Coupling,
4: Planetary gearbox, 5: Microphone, 6: Torque controller,
7: Acceleration sensor, 8: Gear #1, 9: Bearing #1.

Fig. 4. Experimental setup



1: Normal, 2: Crack in tooth root, 3: Worn tooth,
4: Chipped tooth, 5: Missing tooth.

Fig. 5. Condition patterns of the gears #1

III. Application to Health State Assessment

Fig. 3 illustrates the schematic diagram of the framework for health state classification. With vibration and acoustic signals acquired from different types of sensors, time domain signals are converted into the corresponding spectra in the frequency domain by a fast Fourier transform. Then the frequency-domain signals are used as the model's inputs. When applied to the task of health state classification, the DCAE model can be followed by or combined with a classifier. Softmax regression is used in the deep model, which has the advantage of fast computation. A final layer of variables which indicates the labels provided in the training data is added to the DCAE model. Then the classification task based on the joint features is performed by the final layer. To evaluate the effectiveness of the proposed method, the experiments for health state classification of bearings and gears are carried out.

A. Multimodal Data Acquisition

The experiments were carried out on a test-rig as shown in Fig. 4(a). To acquire the original multimodal signals, an acceleration sensor was used to collect the vibration signals and a microphone was used to collect the acoustic signals. A Swiss

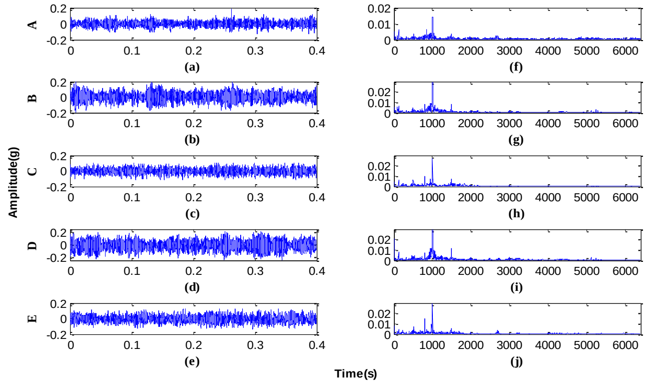


Fig. 6. The vibration signals ((a)-(e)) and their corresponding spectrum ((f)-(j))

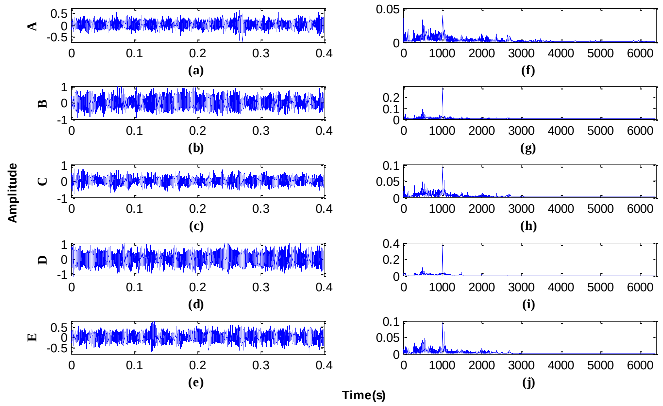
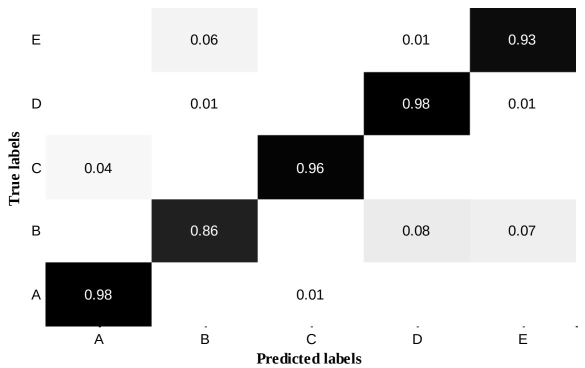


Fig. 7. The acoustic signals ((a)-(e)) and their corresponding spectrum ((f)-(j))

Kistler's 8702B100M1 accelerometer was installed on the bearing housing. Sound emission signals were measured using a B&K 4145 microphone installed close to the bearing house. The location of sensors is shown in Fig. 4(b), and the configuration of the gearbox is shown in Fig. 4(c). During the experiments, for each faulty pattern, the tests were repeated 5 times and the motor speed was 1800 rpm. In each test, the vibration and acoustic signals were collected with a sampling rate of 12.8 kHz, and the duration of sampling time was 90s.

B. Health State Classification for Gears

The proposed method was first tested for the health state assessment of gears. During the experiment, the bearing #1 was normal. In addition to the normal gear (A), there are 4 gears in faulty states, which are chipped tooth (B), crack in tooth root (C), missing tooth (D), and worn tooth (E), as shown in Fig 5. The healthy gear #1 was replaced by different defective gears and the signals were recorded for all the cases separately under the same operating conditions. To detail the collected signals by the microphone and accelerometer respectively, the time domain signals and their spectra for each pattern are shown in Fig. 6 and Fig.7. It can be clearly seen that there are some differences of vibration and acoustic signals among the five patterns. However it is difficult and impractical to manually classify the health states of gears by observing the differences, especially when a large number of samples need to be distinguished.



	A	B	C	D	E
E		0.06		0.01	0.93
D		0.01		0.98	0.01
C	0.04		0.96		
B		0.86		0.08	0.07
A	0.98		0.01		

Fig. 8. The confusion matrix of health assessment for gears

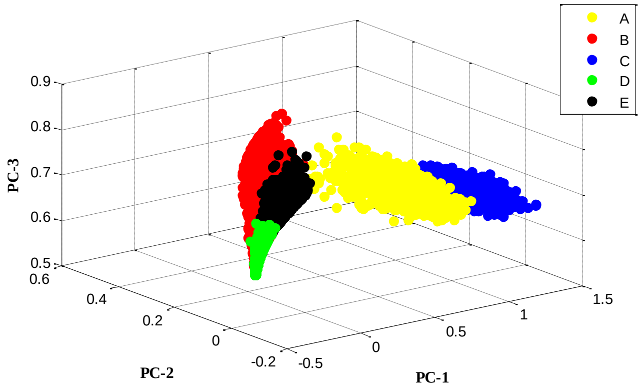


Fig. 9. Scatter plots of principal components of learned features

After the acquisition of multisensory signals, the vibration signals are preprocessed with second generation wavelet package decomposition. The frequency band signals which contain the mechanical fault characteristic frequency are chosen to be converted to spectra. Meanwhile the acoustic signals were converted into the corresponding spectra directly. To be brief, second generation wavelet package decomposition mainly includes three steps: split, prediction and update. The detailed information can be seen in [34]. The number of sub-band wavelet packet is 4 when the decomposition level of vibration signals is 2. Because of the uncertainty of the fault frequency modeling, it is difficult to determine the fault characteristic frequency accurately, especially in practical applications. Thus a sub-band that contains the relevant characteristic vibration frequencies is selected. When the motor speed is set to be 1800 rpm, the fault characteristic frequency is in the range of 0~1000 Hz. The sampling frequency is 12.8 kHz, so the frequency ranges of four sub-bands are 0~3200 Hz, 3200~6400 Hz, 6400~9600 Hz and 9600~12800 Hz, respectively. The rotation speed of the output shaft is 1.5Hz. According to the parameters of the gears, the meshing frequency is 50.7Hz. For a local gear fault, such as chipped tooth, crack in tooth and missing tooth, the defect characteristic frequency is rotational speed with an additional modulating effect on the normal frequency. For the fault of worn tooth, the defect frequency is 50.7 Hz with sidebands in the power

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2018.2793246, IEEE Transactions on Industrial Informatics

TABLE I
HEALTH STATE ASSESSMENT FOR GEARS

Category	Method	Classification accuracy
Deep learning with data fusion	Proposed DCAE with vibration and acoustic signals	94.3%
Deep learning without data fusion	DAE with vibration signals	91.3%
	DAE with acoustic signals	88.7%
Deep learning with traditional data fusion	DAE with vibration and acoustic signals.	92.6%

TABLE II
HEALTH STATE ASSESSMENT FOR BEARINGS

Category	Method	Classification accuracy
Deep learning with data fusion	Proposed DCAE with vibration and acoustic signals	96.4%
Deep learning without data fusion	DAE with vibration signals	90.1%
	DAE with acoustic signals	86.6%
Deep learning with traditional data fusion	DAE with vibration and acoustic signals.	92.7%

spectrum of the envelope signals. The sub-band signals of 0~3200 Hz are converted to spectra and used as inputs of the deep architecture. The spectra of vibration signals have a 0.5s window size with 0.1s overlap. Totally 10000 samples are constructed. There are 2000 samples for each pattern and each sample contains 5120 data points. The corresponding states of the classified data are the required label information.

When applying the DCAE for deep fusion learning, the parameters should be chosen carefully. The number of neurons is an important factor that affects the feature extraction and fusion results. To simplify the parameter selection, the number of units for all the hidden layers in the deep architecture is restricted to be the same. Grid search is used for the number of hidden units with the setting of 50, 100, 200, 400, and 500. Generally with more neurons, it would induce heavy computational burden and may bring a better performance. The validation datasets are performed to determine the best number of the hidden units. For the GBRBM layer, which is used to model the real-valued vectors for vibration and acoustic signals, the learning rate is set to be 0.1. The architecture is structured with 1024(visible)-100(hidden)-120(hidden) for the deep architecture. The number of hidden units for the coupling layer is 400. For the coupling layer, three parameters, α, β, and γ, needed to be chosen appropriately. The values are set to be 0.4, 0.4 and 0.2, respectively. To train the DCAE model, 8000 vibration and acoustic samples are randomly chosen as the inputs. With the trained DCAE model, the rest 2000 samples are exploited for the test of the assessment performance. The present DCAE model produces a classification rate of 94.3% for the gear fault diagnosis experiment. To give more information about the assessment result of the gears, the confusion matrix of health state assessment is shown in Fig. 8. It can be seen that the present method misclassifies 8% of test samples of pattern B (chipped tooth) as pattern D (missing tooth) and 7% as pattern E (worn tooth). The reason may be that the spectra with the defect characteristic frequency are similar, which can be seen in Fig.6 and Fig.7. However, it should be noted few samples in faulty states are recognized as normal ones, which can alert the operator before it develops into a catastrophic failure compared with the situation that a faulty state is identified as normal one.

To illustrate what features the DCAE extracts through hierarchical learning, the learned features are visualized by implementing principal component analysis (PCA). The first three principal components are shown in Fig. 9. It can be seen that the same health state are clustered well and different states are separated from each other except a few samples, which reveals that the proposed method can learn the discriminative and joint features of the multimodal signals well. And a few learned features of different faulty states are mixed, which corresponds to the result shown in the confusion matrix of health state assessment for gears.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2018.2793246, IEEE Transactions on Industrial Informatics

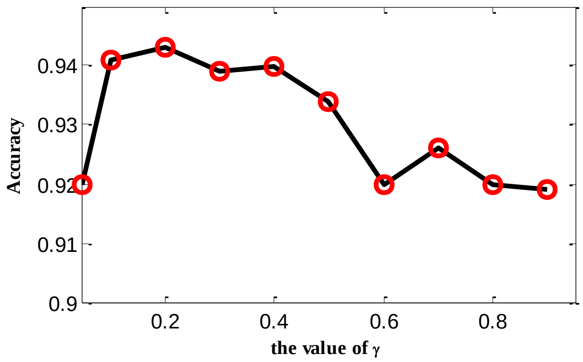


Fig. 10. The accuracy of health assessment for gears with different values of γ

Fig. 10 shows impact of parameters γ on the performance of the deep architecture. Usually α is equal to β because the reconstruction errors of vibration and acoustic modality are equally important. If γ is set too large, it means more importance is attached to the joint information between vibration and acoustic signals. On the contrary, small value of γ means the individual modality draws more attention while joint information is ignored. As shown in Fig. 10, the proposed deep architecture has a good performance in a quite large range.

C. Health State Classification for Bearings

To further validate the robustness of the model, the experiment of bearings with different health states was carried out. In this experiment, 3 faulty bearings, namely with inner race fault, outer race fault and ball fault, and 1 normal bearing were used to form 4 condition patterns. During the experiment, the healthy bearing #1 was replaced with defective bearings in turn and the gear #1 was in a normal condition. Similar with the experiment of health state assessment for gears, the vibration and acoustic signals were collected. The vibration signals were decomposed by second generation wavelet package, and the decomposition level was 2. The defect characteristic frequencies of inner race fault, outer race fault, and ball fault are 7.4Hz, 4.5Hz and 3.0Hz, respectively. The sub-band signals containing faulty characteristics were converted to the corresponding spectra. While the acquired acoustic signals were converted to frequency domain without second generation wavelet package decomposition. The total number of samples was 10000 and each pattern had 2500 samples. Among them 80% numbers were randomly chosen for training the DCAE model, and the rest for testing. The number of the hidden neurons were determined through the grid search method similar to that used in the gear health state assessment. With the trained model, 2000 testing samples were used to validate the performance, and the testing accuracy was 92.4%

D. Comparisons and Discussion

Different from the traditional methods using statistical features as the input of shallow learning model, or extracting features from each mode individually and combining them into a long vector simply, the present study considers the problems of integrating feature extraction and multimodal data fusion into a single process. The proposed deep architecture model, DCAE, is compared with deep learning methods without multimodal data fusion. In this category of comparison, the information of each individual mode is used separately. Specifically, only the acoustic signals or vibration signals are used for the health diagnosis of machine components. Under this condition, the CAE is replaced with autoencoder in the deep architecture model called Deep Autoencoder (DAE), and the rest of the architecture remains unchanged. The comparison results of health state assessment for gears and bearings are shown in TABLE I and TABLE II, respectively. Without multimodal data fusion, classification rates would drop. If only the vibration signals or acoustic signals are used for the DAE, the classification rates for gears are 91.3% and 88.7%, respectively. While for the bearing, the classification rates are 90.1% and 88.6%, respectively. It can be seen that the performance can be improved when considering the data fusion of different modalities. Since there exists inherent relations between vibration and acoustical signals, the proposed method is capable of learning the common information which is helpful for the classification of health state. In summary, the results reveal that multimodal data fusion improves the accuracy of intelligent fault diagnosis for mechanical components. On the other hand, vibration signals contribute to higher classification rate than the acoustic signals, which indicates that the vibration signals are more sensitive and effective for the health state classification of mechanical components.

Moreover, the DCAE method is compared with the traditional data fusion strategy. It concatenates the individually learned features into a long vector, then the vector is used as inputs of the softmax regression to carry out intelligent fault diagnosis. The traditional strategy has two disadvantages. First, it ignores the related information between different modalities. Second, when the individually extracted features from each model do not reside in a commensurate space, it is difficult for the classifier to determine reliable decision boundaries. TABLE I and TABLE II show that the result of traditional fusion method is better compared with the single-sensor mode, which indicates that the fusion of multimodal signals can truly improve the accuracy of health state classification. However, the proposed fusion strategy performs better than the traditional one. Thus the fusion strategy is important for analyzing the multimodal signals. This means the DCAE model is a good choice to address the problem of multimodal information fusion.

Based on the comparisons of experimental results using different methods, we can observe that the proposed method can handle the multimodal signals fusion and project the high-dimensional vectors of signals from different types of sensors into a common latent space, which enables effective and efficient classification of health state for mechanical components.

IV. Conclusion

In this paper, a novel multi-modal fusion deep architecture, called DCAE, is developed to find a joint feature between vibration and acoustic signals for the challenge of health state classification. The model employs deep learning strategy based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2018.2793246, IEEE Transactions on Industrial Informatics

on the coupling autoencoder to fuse multimodal signals acquired from different sensors, which incorporates feature learning and multimodal data fusion into a single process. Moreover, the constructed deep architecture can learn the high-level features in a self-taught way via greedy layer-wise training, which can efficiently extract correlations between vibration and acoustic signals. Experimental results of the two multimodal datasets show that the proposed method is capable of learning the joint feature from different modalities and performs accurate classification of health state in comparison with other methods.

It should be pointed out that the multimodal sensory data and appropriate fusion strategy contribute to advancing the level of health state classification. Future research is still needed in the following aspects. Firstly, different multimodal measurements, such as acoustic emission, current and thermal signals should be collected to achieve multiple sensor fusion for the health state assessment. Secondly, fusion strategy should be investigated since it is critical in the health state assessment with multimodal signals.

REFERENCES

[1] D. An, N. H. Kim, and J. H. Choi, “Practical options for selecting data-driven or physics-based prognostics algorithms with reviews,” *Reliability Engineering & System Safety,* vol. 133, pp. 223-236, Jan, 2015.

[2] J. Lee, F. J. Wu, W. Y. Zhao, M. Ghaffari, L. X. Liao, and D. Siegel, “Prognostics and health management design for rotary machinery systems-Reviews, methodology and applications,” *Mechanical Systems and Signal Processing,* vol. 42, no. 1-2, pp. 314-334, Jan, 2014.

[3] X. W. Dai, and Z. W. Gao, “From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis,” *IEEE Transactions on Industrial Informatics,* vol. 9, no. 4, pp. 2226-2238, Nov, 2013.

[4] P. Tamilselvan, and P. F. Wang, “Failure diagnosis using deep belief learning based health state classification,” *Reliability Engineering & System Safety,* vol. 115, pp. 124-135, Jul, 2013.

[5] N. M. Vichare, and M. G. Pecht, “Prognostics and health management of electronics,” *IEEE Transactions on Components and Packaging Technologies*, vol. 29, no. 1, pp. 222-229, Mar, 2006.

[6] W. Q. Meeker, and Y. L. Hong, “Reliability Meets Big Data: Opportunities and Challenges,” *Quality Engineering,* vol. 26, no. 1, pp. 102-116, Jan 2, 2014.

[7] L. D. Xu, W. He, and S. C. Li, “Internet of Things in Industries: A Survey,” *IEEE Transactions on Industrial Informatics,* vol. 10, no. 4, pp. 2233-2243, Nov, 2014.

[8] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, “Multisensor data fusion: A review of the state-of-the-art,” *Information Fusion,* vol. 14, no. 1, pp. 28-44, Jan, 2013.

[9] R. C. Farias, J. E. Cohen, and P. Comon, “Exploring Multimodal Data Fusion Through Joint Decompositions with Flexible Couplings,” *IEEE Transactions on Signal Processing,* vol. 64, no. 18, pp. 4830-4844, Sep 15, 2016.

[10] M. Kang, J. Kim, and J. M. Kim, “An FPGA-Based Multicore System for Real-Time Bearing Fault Diagnosis Using Ultrasampling Rate AE Signals,” *IEEE Transactions on Industrial Electronics,* vol. 62, no. 4, pp. 2319-2329, Apr, 2015.

[11] N. Baydar, and A. Ball, “Detection of gear failures via vibration and acoustic signals using wavelet transform,” *Mechanical Systems and Signal Processing,* vol. 17, no. 4, pp. 787-804, Jul, 2003.

[12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature,* vol. 521, no. 7553, pp. 436-444, May 28, 2015.

[13] E. J. Humphrey, J. P. Bello, and Y. Lecun, “Feature learning and deep architectures: new directions for music informatics,” *Journal of Intelligent Information Systems,* vol. 41, no. 3, pp. 461-481, Dec, 2013.

[14] I. Arel, D. C. Rose, and T. P. Karnowski, “Deep machine learning-a new frontier in artificial intelligence research [research frontier],” *IEEE Computational Intelligence Magazine,* vol. 5, no. 4, pp. 13-18, 2010.

[15] N. Gebraeel, M. Lawley, R. Liu, and V. Parmeshwaran, “Residual life, predictions from vibration-based degradation signals: A neural network approach,” *IEEE Transactions on Industrial Electronics,* vol. 51, no. 3, pp. 694-700, Jun, 2004.

[16] R. Q. Yan, and R. X. Gao, “Approximate Entropy as a diagnostic tool for machine health monitoring,” *Mechanical Systems and Signal Processing,* vol. 21, no. 2, pp. 824-839, Feb, 2007.

[17] G. Wang, and S. Yin, “Quality-Related Fault Detection Approach Based on Orthogonal Signal Correction and Modified PLS,” *IEEE Transactions on Industrial Informatics,* vol. 11, no. 2, pp. 398-405, Apr, 2015.

[18] W. H. Li, S. H. Zhang, and S. Rakheja, “Feature Denoising and Nearest-Farthest Distance Preserving Projection for Machine Fault Diagnosis,” *IEEE Transactions on Industrial Informatics,* vol. 12, no. 1, pp. 393-404, Feb, 2016.

[19] C. Sun, Z. S. Zhang, X. Luo, T. Guo, J. X. Qu, and B. Li, “Support vector machine-based Grassmann manifold distance for health monitoring of viscoelastic sandwich structure with material ageing,” *Journal of Sound and Vibration,* vol. 368, pp. 249-263, Apr 28, 2016.

[20] M. Ma, X. F. Chen, X. L. Zhang, B. Q. Ding, and S. B. Wang, “Locally Linear Embedding on Grassmann Manifold for Performance Degradation Assessment of Bearings,” *IEEE Transactions on Reliability,* vol. 66, no. 2, pp. 467-477, Jun, 2017.

[21] Y. G. Lei, F. Jia, J. Lin, S. B. Xing, and S. X. Ding, “An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data,” *IEEE Transactions on Industrial Electronics,* vol. 63, no. 5, pp. 3137-3147, May, 2016.

[22] M. Ma, C. Sun, and X. Chen, “Discriminative Deep Belief Networks with Ant Colony Optimization for Health Status Assessment of Machine,” *IEEE Transactions on Instrumentation and Measurement,* 2017.

[23] M. Zitnik, and B. Zupan, “Data Fusion by Matrix Factorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 37, no. 1, pp. 41-53, Jan, 2015.

[24] O. Basir, and X. H. Yuan, “Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory,” *Information Fusion,* vol. 8, no. 4, pp. 379-386, Oct, 2007.

[25] H. I. Suk, S. W. Lee, D. G. Shen, and A. s. D. Neuroimaging, “Latent feature representation with stacked auto-encoder for AD/MCI diagnosis,” *Brain Structure & Function,* vol. 220, no. 2, pp. 841-859, Mar, 2015.

[26] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, “Group sparse representation based classification for multi-feature multimodal biometrics,” *Information Fusion,* vol. 32, pp. 3-12, Nov, 2016.

[27] C. Habib, A. Makhoul, R. Darazi, and C. Salim, “Self-adaptive data collection and fusion for health monitoring based on body sensor networks,” *IEEE Transactions on Industrial Informatics,* vol. 12, no. 6, pp. 2342-2352, 2016.

[28] Z. Y. Chen, and W. H. Li, “Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network,” *Ieee Transactions on Instrumentation and Measurement,* vol. 66, no. 7, pp. 1693-1702, Jul, 2017.

[29] M. Ma, X. Chen, S. Wang, Y. Liu, and W. Li, "Bearing degradation assessment based on weibull distribution and deep belief network." In *Flexible Automation (ISFA), International Symposium on* (pp. 382-385). IEEE.

[30] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798-1828, Aug, 2013.

[31] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder." In *Proceedings of the 22nd ACM international conference on Multimedia,* pp. 7-16. ACM, 2014.

[32] G. E. Hinton, and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science,* vol. 313, no. 5786, pp. 504-507, Jul 28, 2006.

[33] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, Nov, 2012.

[34] M. Meng, L. Ruonan, H. Yushan, and C. Xuefeng, "Fault diagnosis of bearing running status using mutual information." In *Prognostics and System Health Management Conference (PHM-2014 Hunan), 2014,* pp. 135-139. IEEE, 2014.

Meng Ma, received the B.S. degree in mechanical engineering from Sichuan University, Chengdu, China, in 2009. He is currently working toward the Ph.D. degree in mechanical engineering in the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, China.

His research interests include prognostics and health management (PHM) and machine learning for remaining useful life prediction.

Chuang Sun, received the Ph.D. degree in 2014 in Mechanical Engineering from Xi'an Jiaotong University, China. From Mar. 2015 to Mar. 2016, he was a postdoc at Case Western Reserve University, Cleveland, USA.

He is now an assistant research fellow in school of mechanical engineering at Xian Jiaotong University, China. His research interests include sparse representation, deep learning, fault diagnosis and prognosis.

Xuefeng Chen (M'12) received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2004.He is currently a Professor of Mechanical Engineering with Xi'an Jiaotong University.

His current research interests include finite-element method, mechanical system and signal processing, diagnosis and prognosis for complicated industrial systems, smart structures, aero-engine fault diagnosis, and wind turbine system monitoring.

Dr. Chen was a recipient of the National Excellent Doctoral Dissertation of China in 2007, the Second Award of Technology Invention of China in 2009, the National Science Fund for Distinguished Young Scholars in 2012, and a Chief Scientist of the National Key Basic Research Program of China (973 Program) in 2015. He is the Chapter Chairman of the IEEE Xi'an and Chengdu Joint Section Instrumentation and Measurement Society.