IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 5, 2021, accepted November 16, 2021, date of publication November 17, 2021,
date of current version November 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129061

A Novel Intelligent Fault Diagnosis Method for Rolling Bearings Based on Compressed Sensing and Stacked Multi-Granularity Convolution Denoising Auto-Encoder

CHUANG LIANG[1,2], CHANGZHENG CHEN[1,3], YE LIU[2], AND XINYING JIA[2]
[1]School of Mechanical Engineering, Shenyang University of Technology, Shenyang 110870, China
[2]BMW Brilliance Automotive Ltd., Shenyang 110143, China
[3]Liaoning Engineering Center for Vibration and Noise Control, Shenyang 110870, China

Corresponding author: Changzheng Chen (czchen@sut.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51675350 and Grant 51705337.

ABSTRACT This paper investigates the unsupervised automatic feature extraction method with a large amount of unlabeled data for the fault diagnosis of rolling bearings in automobile production line, where the fault information is hard to identify due to the low-level features of a single category and the massive fault data is difficult to process. Different from the existing methods, which only combine the compressive sensing with single category of low-level features, or extract features from raw data, a novel intelligent fault diagnosis method for rolling bearings based on the compressive sensing and a stacked multi-granularity convolution denoise auto-encoder network is proposed, which utilizes the nonlinear projection to achieve the compressed acquisition and resolves issues with character unicity by extracting a diverse category of high-level features. Moreover, a regularization method called 'dropout' is used to prevent overfitting during the training process. The amount of measured data that contained all the information of faults is reduced and the classification accuracy is improved by extracting more robust features based on the proposed method. Finally, the effectiveness of the proposed method is validated using data sets from rolling bearings in an automotive production line and the analysis result show that it is superior to the existing methods and is able to obtain high diagnostic accuracies.

INDEX TERMS Bearing fault, intelligent diagnosis, feature extraction, compressive sensing, stacked auto-encoder.

I. INTRODUCTION

Rotating machinery plays an important role in modern automobile industry. With the upgrading of automotive production capacity, the line stop caused by the faults of rotating machinery will cause heavy economic losses and even endanger the personal safety of producers. Hence, the condition monitoring of the rotating machinery has attracted great attentions [1]. According to the statistics of historical machinery failures, almost 40% of the faults about rotating machinery come from rolling bearings, which will suffer various faults due to the harsh production conditions [2], [3]. Consequently, the reliable bearing fault diagnosis methods are meaningful and practical.

Vibration analysis is widely adopted in the bearing fault diagnosis [4]. After a literature review, signal processing and intelligent diagnosis are two main methods that have proved to be effective [5], [6]. Intelligent diagnosis methods mainly include two steps: feature extraction and fault recognition. It should be noted that feature extraction is more significant which intends to obtain representative characteristic from raw signals based on signal processing methods. For instance, spectral analysis, time-domain statistical analysis [7], [8], transform domain analysis [9], [10], entropy and adaptive decomposition [11]–[13]. Nevertheless, some insensitive or redundant information may be in these extracted features. Then some dimension reduction strategies and feature

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

selection methods are presented to obtain sensitive characteristic, which affect the computational efficiency as well as the diagnosis results, such as feature discriminant analysis [14] and principal component analysis. On the other hand, some artificial intelligence methods are adopted to identify the bearing faults. For example, support vector machine (SVM) [15], random forest [16], artificial neural network (ANN), and k-nearest neighbor [17]. The general procedure is shown as Figure 1, and the left hand side is the traditional method.

However, intelligent fault diagnosis methods still have two deficiencies. Firstly, traditional feature extraction methods not only rely heavily on diagnostic expertise and professional technology, but also need to extract features manually. Also, these methods are ordinarily studied according to one specific diagnosis issue with low generalization. Secondly, traditional artificial intelligent methods cannot distinguish the primary differences between the complex information effectively from massive raw signals. It generally uses high-dimensional signals to show the information in the complex mechanical system. The Shannon-Nyquist theorem is the common way to extract the vibration signals. In this traditional way, multiple sensors with long operation periods over high sampling could produce a large amount of data, which put forward high requirements on the transmission bandwidth, data storage, acquisition hardware, and subsequent processing [18]. Thus how to extract features effectively from massive raw data and identify the faults accurately are worth researching.

In this paper, we adopt the compressive sensing (CS) to extract the raw data, which is fundamentally different from Nyquist theory. In recent years, CS has attracted considerable attention in some areas, such as signal-pixel camera [19], radar imaging [20], and electrocardiogram [21]. CS reduces the amount of sampled data while retaining most of the useful information. To a certain extent, CS provides a new thinking in the field of data acquisition and processing because of its low requirement for the storage and computational [22]. The general procedure of the CS method introduced by some researchers includes three parts: the projection acquisition of the raw data, compressed and reconstruction. The procedure has shown in the middle of the Figure 1.

Obviously, the methods based on CS are still draw support from the conventional ways, although it reduces the demands for data storage and computational. Histon *et al.* developed the deep learning (DL) theory [23], which provides the theoretical support for the above difficulties. The fundamental of the DL is that it can map the original space data into the feature space by learning a nonlinear input function under the structure of a multilayer neural network. DL is applied in different fields successfully with the function of dealing with massive data analysis automatically since its emergence, such as the image recognition, the speech identification [24], and other applications. The development of DL has significantly reduced the dependence on expertise and the manual selection of features in intelligent diagnosis [25]. The procedure has shown in the right hand of the Figure 1. Common DL algorithms include convolution neural networks (CNN), recurrent neural network (RNN), deep belief network (DBN), and auto-encoder (AE) [26]. Among these algorithms, CNN is more popular because of its character of sparse connections and weight shares. However, this technique need a back-propagating (BP) error approach [27] to train the network using massive labeled datasets. Hence, the acquisition of data set requires extensive resources, which restricts the application and development of CNN. Under these circumstances, unsupervised learning [28] becomes a better choice, which can automatically extract features via unlabeled data. The auto-encoder (AE) has the unsupervised neural network structure. Nevertheless, too many network parameters were introduced in AE, due to its performance of the full connectivity between layers [29]–[32]. Thus both CNN and AE have a common limitation about feature extraction. To solve these problems, Shao et.al. proposed an improved convolutional deep belief network with compressed sensing [33]. He *et al.* put forward a new framework based on small labeled infrared thermal images and enhanced convolutional neural network from convolutional auto-encoder [34].

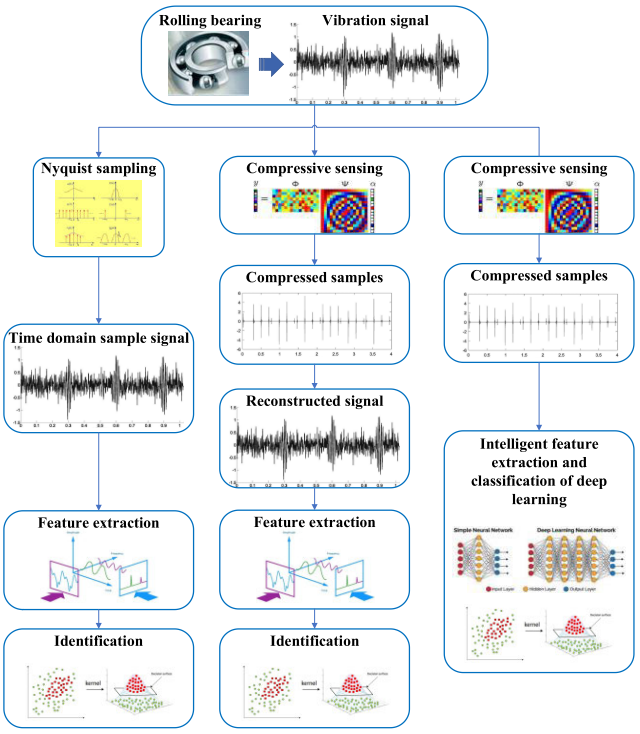

FIGURE 1. The procedure of different methods.

Summing up the above, existing DL methods have the difficulties in acquiring massive labeled data. On the other hand, feature extraction also has many limitations. Instead, unmarked data can be easily obtained. So we propose a novel fault diagnosis framework based on CS and stacked multi-granularity convolution denoising auto-encoder (SMGCDAE) method. This framework provides

a new bearing fault classification solution for bearing fault diagnosis. Firstly, CS reduces the amount of sampled data while retaining most of the useful information. Then, multi-granularity convolution denoising auto-encoders (MGCDAE) combines an ensemble learning thought called multi-granularity convolution kernel [35]–[37] and the denoising auto-encoder (DAE), which can use fewer parameters and lower computational learning cost to extract robust features from unlabeled data. We can obtain different features, because the size of the kernels varies. This approach adds the receptive fields due to the function of the convolution kernels, allowing them to acquire multifarious fault features. In addition, dropout [38] is utilized in the hidden layer of the auto-encoders to prevent the overfitting by averaging the model. In brief, the generalization performance in fault diagnosis is improved, because of the diversify of the attributes. Finally, we stack several multi-granularity convolution denoising auto-encoders (MGCDAEs) to form a SMGCDAE structure and use a pre-training method to train this network [39]. We summarize the main insights and contributions of this work as follows:

1. Employed a novel bearing fault diagnosis framework, which integrates CS with SMGCDAE method in this paper. Compared to other techniques, this framework can reduce environmental demands, transmission costs and computations. The original vibration data is linearly mapped into a lower dimension space. The small amount of compressed signal not only gets rid of the dependence on diagnostic expertise and prior knowledge, but also contains most of the information. In addition, this framework can obtain more comprehensive key features using diverse characteristics exhibited and acquire robust features based on unsupervised learning.

2. In the diagnosis case of a real data set from an automotive production line, the effects of the key parameters and the selection of the proposed method are thoroughly studied. In addition, the experiment shows the superiority of our proposed framework by comparing with traditional methods.

The remainder of this paper is organized as follows. Section 2 overviews the theory of CS and AE utilized in the proposed technique, and presents the details of the SMGCDAE with dropout. Section 3 describes the proposed CS-SMGCDAE intelligent method for rotating machinery fault diagnosis. Section 4, the performance of the raised method is verified by experiments from an automotive production line. Finally, conclusions and future work are included in section 5.

II. COMPRESSIVE SENSING

This section gives a brief introduction to CS [40], which is a special case of sparse representation. To a certain extent, CS is even an extended of sparse representation. The sample idea of CS is that so many real-world signals have sparse features in some domain, e.g., Fourier Transform (FT), we can use fewer measurements to reconstruct it under some conditions. CS has two principles: one is the sparsity of the signals; the other one is that the measurements matrix from the original signals

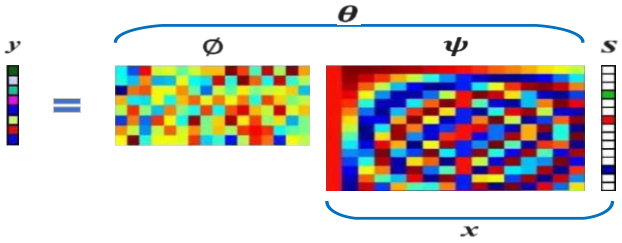

FIGURE 2. Compressive sampling framework.

could be compressed by their sparse representations. In other words, the measurements matrix in the second principle must satisfy the data minimal information loss, i.e., Restricted Isometry Property (RIP). Briefly, we describe CS as follows.

Assuming an unknown original signal X, $x_i \in R^n$, which has n data points. To allow these data points produce a set of sparse components, for a given sparse transformation matrix φ, $\varphi_i \in R^{n\times n}$, the mathematical definition of X can be expressed as Equation (1):

$$x = \sum_{i=1}^{n} \varphi_i s_i \tag{1}$$

Or more efficiently

$$x = \varphi s \tag{2}$$

where s represent the sparse elements and a $n * 1$ column vector of coefficients. When the dictionary (sparse transformation) φ is incoherent with the measurement matrix $\emptyset$, the original signal X can be reconstructed by the compressed measurements y based on the theory of the compressive sampling, and y can be written as follows:

$$y = \emptyset\varphi s = \theta s \tag{3}$$

where y is the compression measurement, represented by a $m * 1$ column vector. θ is the measurement matrix, $\theta = \emptyset\varphi$, and the matrix θ must satisfy the data minimal information loss, i.e., Restricted Isometry Property (RIP) [41].

Definition 1.1: The measurement matrix θ satisfies the Restricted Isometry Property (RIP) if there is a parameter $\delta \in (0, 1)$ as follows:

$$(1-\delta)\|s\|_2^2 \le \theta\|x\|_2^2 \le (1+\delta)\|s\|_2^2 \tag{4}$$

The size of the measurement matrix θ is $m * n$, which depends on the compressive sampling rate (α), and the length of m is significantly lower than the Nyquist rate ($m \ll n$). Figure 2 show the compressive sampling framework.

To some extents, the compressed data can cover most of the raw signal information if the measurement matrix satisfies the RIP. [42] proved that the random Gaussian matrix satisfies the RIP with good universality. Hence, the measurement matrix employs the random Gaussian matrix to obtain the compressed data.

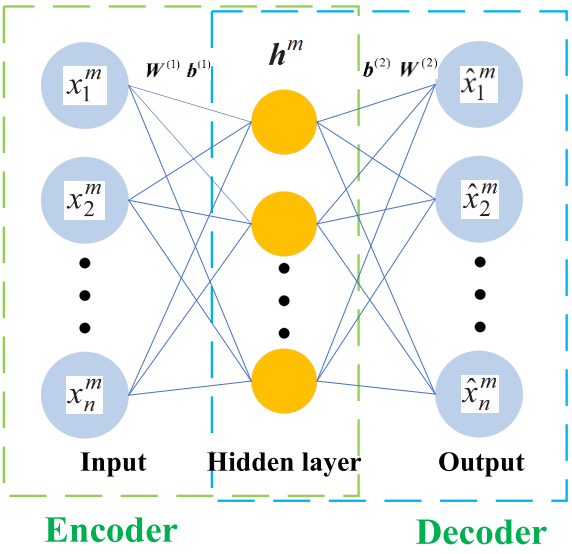

FIGURE 3. The architectural of AE.

III. STACKED MULTI-GRANULARITY CONVOLUTION DENOISING AUTO-ENCODER WITH DROPOUT

A. DEEP NEURAL NETWORK AND AUTO-ENCODERS

The deep neural network (DNN) is developed from deep learning with the deep architectures. In this network, the representative information in the approximate complex nonlinear functions and compressed measurements can be captured with small errors. In addition, DNN has the ability of amplifying the differences in the explanatory information contained from the original data and suppressing irrelevant parts that cause interference, thus can distinguish the different fault classes.

An auto-encoder is a widely used unsupervised neural network, which has three layers. The target of the output in an auto-encoder is to reconstruct the input data via the backpropagation [25]. As depicted in Figure 3, an auto-encoder consist of encoder part and decoder part like many unsupervised feature learning methods. The encoder network not only transforms the high-dimensional input data into the low-dimensional output codes, but also produces the feature vectors. The decoder network reconstructs the inputs from these feature vectors.

The encoder network can be defined as a feature extraction function f_θ. For each measured signal x^m, that can compute a feature vector h^m, as shown in Equation (5):

$$h^m = f_\theta\left(x^m\right) \tag{5}$$

where h^m is the feature representation obtained from x^m.

The decoder network can be denoted by a recovery function g_θ, which can transform the feature space h^m into the input space $\hat{x}^m$, producing a reconstruction:

$$\hat{x}^m = g_\theta\left(h^m\right) \tag{6}$$

The parameter sets of the auto-encoder are learned simultaneously on an approximation such that $\hat{x}^m$ is similar to x^m, also attempting to attain the lowest possible reconstruction error $L(x, \hat{x})$. Where the loss function $L(x, \hat{x})$ can measure the discrepancy between x and $\hat{x}$. Hence, we can obtain the following equation:

$$L\left(x, \hat{x}\right) = \frac{1}{2}\left\|x - \hat{x}\right\|^2 \tag{7}$$

In fact, affine mapping is the most common used form for auto-encoder [43], and that keep collinearity followed by nonlinearity:

$$f_\theta\left(x\right) = s_f\left(Wx + b\right) \tag{8}$$

$$g_\theta\left(x\right) = s_g\left(W^T x + d\right) \tag{9}$$

where s_f and s_g are the activation functions of the encoder and decoder, respectively, e.g. hyperbolic and sigmoid. b and d are bias vectors, and W and W^T are the weight matrices.

Although the original input data can be reconstructed by the learned feature representation perfectly, the generalization performance of the model is not good.

B. DROPOUT

Dropout is a technique used to prevent overfitting in the fully connected layers. The network will remove some hidden units in each layer randomly with a certain probability during each training iteration, thus the hidden units can change their states without the help of other hidden units. In this study, the dropout technique is applied to avoid the extraction of the same feature repeatedly and prevent complex co-adaptations on the training data.

C. SIGNAL MULTI-GRANULARITY CONVOLUTION DENOISING AUTO-ENCODER WITH DROPOUT

In real implementation, achieving sufficient feature learning is susceptible to interference because of complicated factors. For instance, instrumentation errors and inaccurate data collection could cause data deviation. Consequently, capture more information to measure the latent high-level feature representation is highly necessary. In order to learn high-level features effectively in this paper, we adopt the multi-granularity convolution kernels. Under this concept, each convolution layer contains convolution kernels of varying sizes, with each convolution kernels corresponding to a unique feature. This structure integrates high-level features to present more comprehensive information through various mappings.

In this paper, the proposed MGCDAE pipeline contains three dimensions of convolution kernels: 1×1, 3×3, and 5×5. As the number of kernel increases, the amount of computation and required runtime increases. Therefore, the 1×1 kernel was mainly applied to alleviate computational bottleneck, reduce network parameters, and decrease dimensionality. The first part of the multi-granularity is a 1×1 convolution layer. To allow local connection of each pipeline as sparse as possible, another 1×1 convolution layer is also constructed (see Figure 4).

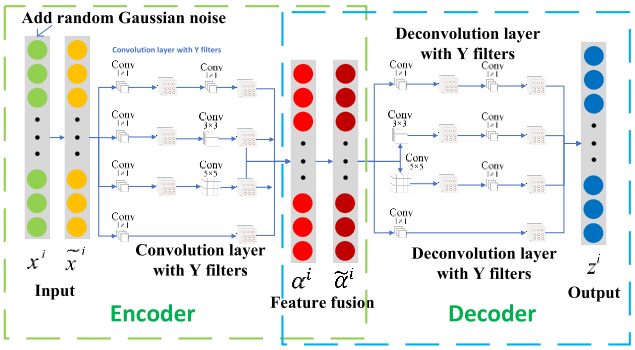

FIGURE 4. The MGCDAE architecture.

To optimize the network, convolution kernels are trained by DAE with dropout technique in the training stage of the proposed network. DAE was firstly introduced in 2008, which is an unsupervised approach used for extracting robust features. In this study, we firstly contaminate the original data to obtain the noisy data, then extract the robust features, which can ensure stability and improve generalization performance. Here, we choose random Gaussian noise to destroy the raw data.

In the encoder process as shown in Figure 4, add the Gaussian noise randomly into the original input vector x^i, thus can obtain a corrupt input vector $\tilde{x}^i$, then get into the nonlinear activation function by linear mapping. Next, the $\tilde{x}^i$ is mapped to a latent vector representation α^i by the function f.

$$\alpha^i = f\left(W_1 * \tilde{x}^i + b_1\right) \tag{10}$$

where W_1 and b_1 are the weighting matrix and encoding bias vectors, and $*$ is the convolution operation.

In brief, we use the benefits of variability of the number of convolution kernels to obtain different high-level representations. When the obtained features were extracted from a given pipeline, it could be integrated using a weighted average, because of the same dimensions. In other words, we put forward a feature fusion method by matching the dimensions of the convolution layer. This method is beneficial for improving the generalization performance of the network.

After that, to optimize the training process, the dropout technique is applied to the network, which can prevent overfitting in the fully connected layers. Technically, the ''dropout'' can be realized by omitting the neural units in the hidden layers randomly with a probability q. Then we can get a dropped representation $\tilde{\alpha}^i$ by a scalar product with a masking vector m.

$$\tilde{\alpha}^i = m \cdot \alpha^i \tag{11}$$

A unique network is trained in each iteration, since the network is updated iteratively by dropping the neurons randomly in the hidden layer. This operation improves the subsequent classification performance greatly.

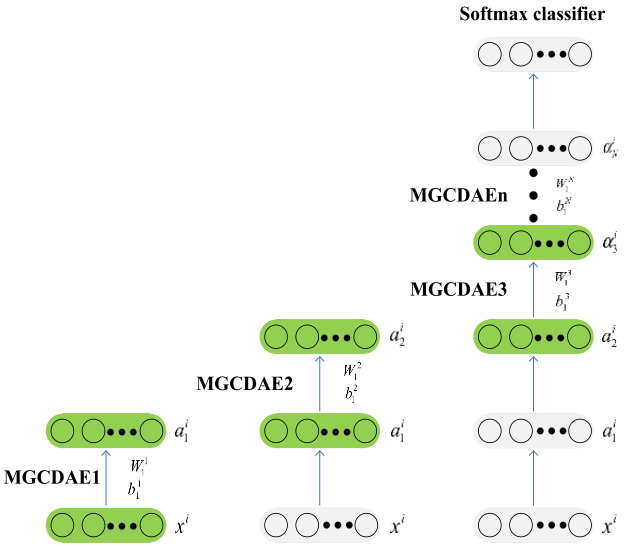

FIGURE 5. The SMGCDAE network architecture.

Instead, in the decoding process, the reconstructed input vector z^i was output by a nonlinear activation function g using the potential feature representation $\tilde{\alpha}^i$.

$$z^i = g\left(W_2 * \tilde{\alpha}^i + b_2\right) \tag{12}$$

In this expression, g, W_2, and b_2 are the decoding function, matrix, and decoding bias vector. Assuming a given training set $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\}$, the overall cost function can be defined as:

$$J(W, b) = \frac{1}{2M} \sum_{i=1}^{M} \left\| x^i - z^i \right\|^2 \tag{13}$$

D. STOCKED MULTI-GRANULARITY CONVOLUTION DENOISING AUTO-ENCODER WITH DROPOUT

Inspired by [44], [45], we find out the remarkable abstractness of the deep neural networks. Hence, multiple MGCDAEs are stacked in a deep neural network. The BP algorithm is used to train the first MGCDAE1. Then the output of the encoder α_1^i become the input for the next MGCDAE2. This process is shown in Figure 5.

Finally, a deep stacked MGCDAE (SMGCDAE) is formed by N MGCDAEs. We can calculate the latent feature representation α_N^i:

$$\alpha_N^i = f\left(W_1^N * \alpha_{N-1}^i + b_1^N\right) \tag{14}$$

where W_1^N is the weight matrix, b_1^N is the bias vector.

The aim of the SMGCDAE network is to improve the nonlinear mapping capabilities of the MGCDAE. We can obtain and fuse the high-level features by abstracting the initial feature layers. At last, the features are put into the classifier to complete the final classification.

IV. SOFTMAX CLASSIFIER

In this study, we use the softmax classifier [46], [47] to classify the fault types as follows:

$$f(x) = \frac{1}{\sum_{j=1}^{k} e_j^{\sum_{i=0}^{M} w_i x_i}} \begin{bmatrix} e_1^{\sum_{i=0}^{M} w_i x_i} \\ e_2^{\sum_{i=0}^{M} w_i x_i} \\ \cdots \\ e_k^{\sum_{i=0}^{M} w_i x_i} \end{bmatrix} \tag{15}$$

where k and w_i are the data category and the weight of the sample x_i. To minimize the reconstruction error, we also need a cost function during the training process in this deep network, thus we can obtain more similar data with the original data.

V. PROPOSED FAULT DIAGNOSIS METHOD

In consideration of the challenges caused by the restrictions of traditional approach and the difficulties in processing massive raw data in bearing fault diagnosis. This paper initially adopts a data acquisition method, which can realize fault signal acquisition by using the transform domain projection in the CS domain. Moreover, a SMGCDAE deep learning algorithm is constructed to realize intelligent diagnosis. The procedure of the proposed framework is shown in Figure 6.

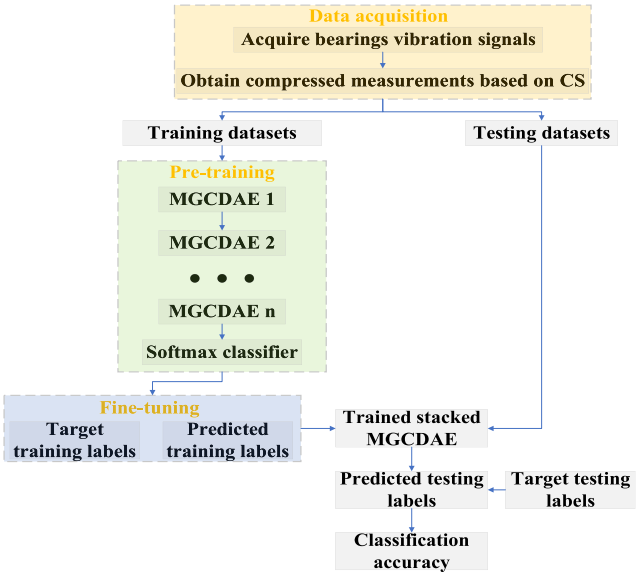

FIGURE 6. Procedure of the proposed method.

Firstly, acquire the compressed data via a specific compression ratio. Then obtain a dataset $x = \{x^i, d^i\}_{i=1}^{M}$, where M is the total number of samples, d^i is the label corresponding to x^i, x^i is the i-th compressed sample. The original dataset is divided into two parts, the training set $x_{train} = \{x^m, d^m\}_{m=1}^{C}$ and the testing set $x_{test} = \{x^k, d^k\}_{k=1}^{D}$. The former is used to train the constructed SMGCDAE network, which is a greedy training method including two main processes: one is to initialize the weight in the network through pre-training the MGCDAE. another one is to improve the performance of the network by further fine-tuning the networks with BP algorithm. The test set has responsible to validate the performance of the proposed diagnosis network.

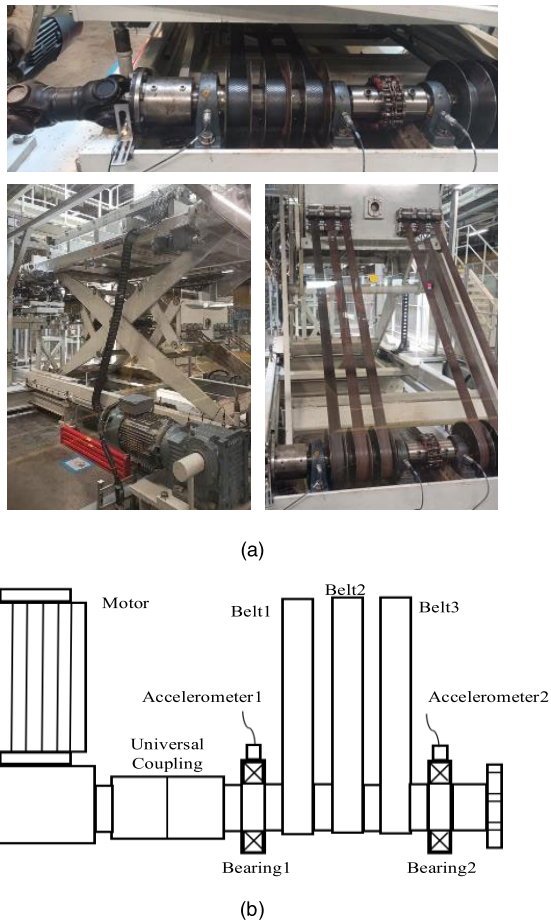

FIGURE 7. (a) Bearing test for the experiment. (b) Transmission mechanism and sensor locations.

VI. EXPERIMENTAL VERIFICATION

A. DATASET DESCRIPTION

This subsection aims to verify the superiority of the proposed method by conducting the fault diagnosis of 'scissor' lifter located in the assembly production line of an automotive company as shown in Figure 7. In the automotive production line, the 'scissor' lifter is a kind of car lifting equipment with car lifting stability and has a wide range of applications. It is mainly used for cars transportation between the height difference of production line. Hence, it will cause huge losses once this important equipment breaks down. The most easily damaged part of the equipment is the rolling bearing on the rotating spindle. The vibration signal of the bearings is extracted and analyzed.

In the experiment, seven kinds of rolling bearing conditions were existing in the test: normal (NOR), outer race fault (Stripping with a size of 40mm*3mm, ORF1), outer race fault (pitting, ORF2), inner race fault (Stripping with a size of

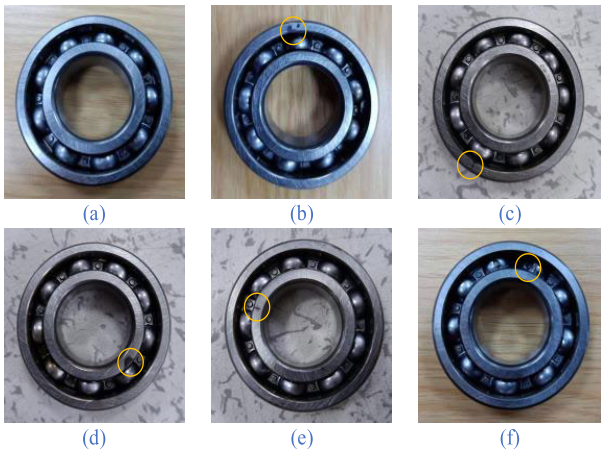

FIGURE 8. Rolling bearings with different health states.

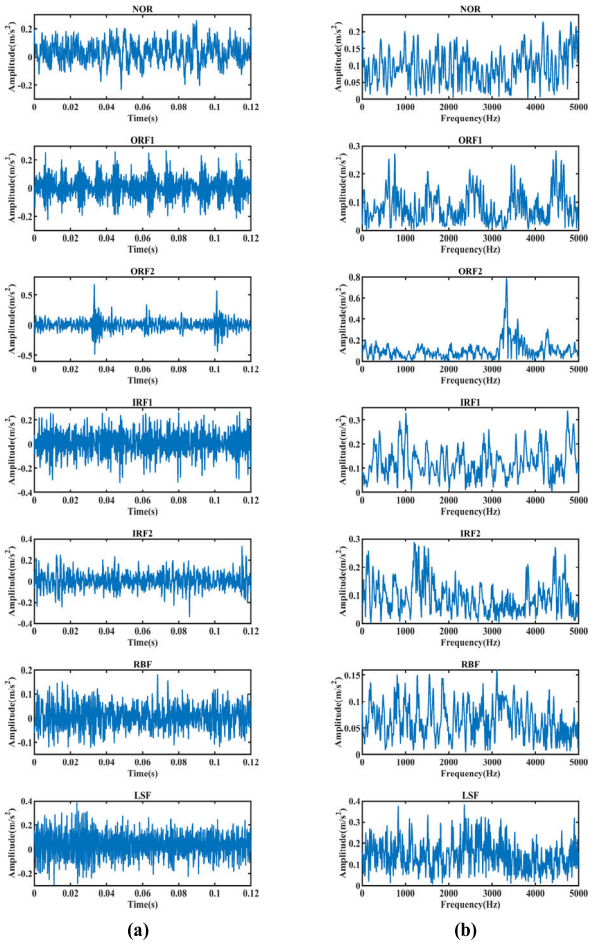

FIGURE 9. (a) The time domain waveforms; (b) the frequency domain waveforms.

40mm*3mm, IRF1), inner race fault (pitting, IRF2), rolling element fault (REF), and lubrication shortage fault (LSF), as illustrated in Figure 8. Each type has 150 samples with 4800 in length, which is marked as set A.

Then, the compressed data under different compression ratios (CRs) with measurement matrix can be obtained by compressed acquisition theory. For example, a 1440×4800 random Gaussian matrix is generated if given the CR $=$ 70%, and which matrix can be used to obtain the compressed sample A'. The raw dataset and the compressed dataset are marked as A and A' in the subsequent processing. Table 1 details the bearing datasets information.

TABLE 1. Detailed information of the bearing datasets.

Data Set (A /A')	Number of Training Samples	Number of Testing Samples	Fault Types	Fault Description	Labels
4800/1440	50	100	NOR	none	1
	50	100	ORF1	pitting	2
	50	100	ORF2	stripping with a size of 40mm*3mm	3
	50	100	IRF1	pitting	4
	50	100	IRF2	stripping with a size of 40mm*3mm	5
	50	100	REF	pitting	6
	50	100	LSF	lubrication shortage	7

As seen in Figure 9, the time domain waveforms and frequency domain waveforms of 'scissor' lifter can hardly distinguish the conditions because of the complexity test condition.

B. EFFECT OF COMPRESSION RATIO(CR)

The CR is related to the length of the original signal and the size of the measurement matrix. The sampling points required by CS decrease with the increase of CR. In a limited number of observations, it could not obtain complete information from raw data if the value of CR is too small. Hence, the CR has an upper bound based on the limited of the RIP theory. When the value of CR is less than 40%, it cannot exhibit a good compression effect. Figure 10 shows the influence of CR changes on diagnostic accuracy and computing time.

As shown in Figure 10, the computing time increases gradually with the decrease of CR within the scope of our research. However, the accuracy rate has been very high, and there is no positive or negative correlation with CR. Finally, 70% is selected as the CR by analyzing the results and influenced factors. We can conclude that a higher CR can be adopted if the computing time requirement is strict, which also could slightly reduce the accuracy. In addition, the requirements of communication and data storage are lower for the higher value of CR.

C. COMPARISON

This section mainly contains two phases of experimentation. Firstly, we investigate the differences in accuracy between

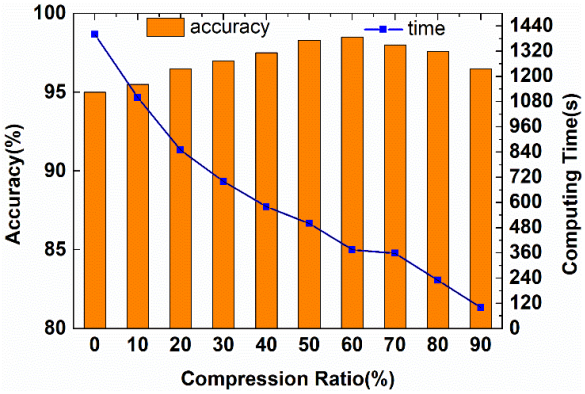

FIGURE 10. Recognition accuracy and computing time with different CRs.

TABLE 2. Classification accuracy on different datasets.

Approach	Datasets			
	dataset 1 (label 1)	dataset 2 (label 2&3)	dataset 3 (label 4&5)	dataset 4 (label 6&7)
CAE(1×1)	85%	88%	84%	81%
CAE(3×3)	87%	88.5%	85.5%	83.5%
CAE(5×5)	87%	90%	86.5%	84%
MGCAE	90%	91%	87%	85.5%
CDAE(1×1)	86%	89%	86%	82%
CDAE(3×3)	88%	90.5%	86.5%	84.5%
CDAE(5×5)	89%	92%	87%	86%
MGCDAE	91%	93%	90%	87%

our proposed method and the prototype. Then to evaluate the effectiveness of our approach, we compare the proposed method with other existing classification approaches.

The parameters of the model were conventional taken from literatures. The number of the filter is set to 96. Set stride to 2. Moreover, the activation function for neurons is typically a Leaky Relu function [48].

In this part, each method was run for 10 times. Thus we can obtain a general comparison after averaging the value of each experiment. The average classification accuracy is shown in Table 2. To facilitate analysis, we divided all the conditions into four kinds of dataset. Let normal condition be the dataset1, ORF1 and ORF2 be the dataset2, IRF1 and IRF2 be the dataset3, REF and LSF be the dataset4. We initially focus on the influence between single-grained and multi-granularity convolution kernels on classification performance. In this study, we use three convolution kernel sizes of CAE to compare with the multi-granular convolution kernel as shown in Table 2. In addition, the 20% random Gaussian noise was added into the raw data to improve the generalization performance in the process of the training, and the further comparison verifies the effective of our approach.

The results show that the accuracy of the MGCAE method on the dataset1 was 90%, the dataset2 was 91%, the dataset3 was 87%, and the dataset4 was 85.5%, respectively. The above accuracy results are meaningfully higher than CAE(5×5), CAE(3×3), and CAE(1×1). It is evident that the others across all four datasets are inferior to our approach.

In addition, the accuracy of the MGCDAE method on the dataset1 was 91%, the dataset2 was 93%, the dataset3 was 90%, and the dataset4 was 87%, respectively. Comparing with the condition of no noise, the accuracy of noise adding has been increasing by 1%, 2%, 3%, and 1.5%, respectively. Hence, we can obtain the robust features to improve the classification accuracy by adding the Gaussian noise.

Figure 11 illustrates the effect of varied noise levels on classification performance. In each dataset, the performance for different proportions of added Gaussian noise were indicated in Figure 11 (a), (b), (c), (d), respectively. Among these figures, signal MGCDAE was represented by 'MGCDAE', the stack of three MGCDAEs was represented by 'stack-3', the stack of five MGCDAEs was represented by 'stack-5', the stack of seven MGCDAEs was represented by 'stack-7'. The results show that the generalization performance of the model increased with the number of the layers. Furthermore, compared to noise-free conditions (the proportion of added Gaussian noise is 0), adding noise will improve generalization performance during the training phase. In the dataset1, when the proportion of added noise is 10%, the accuracy of the model is the highest. In other datasets, the proportion corresponding to the highest accuracy is different. Hence, the proportion of added noise has a certain impact on the prediction accuracy of the model.

To evaluate the feasibility and stability of our approach (MGCDAE-7), we compare it with different types of traditional machine learning classification methods, for instance, random forests (RF), convolution neural network (CNN), support vector machine (SVM), and deep belief network (DBN). Moreover, the 20% random Gaussian noise was added into the raw data. As shown in Table 3, the proposed SMGCDAE method has the highest average accuracy of 97%, 99%, 98%, and 93%, respectively, demonstrating superior performance, across all four datasets. These performances benefit from that the proposed method not only can extract the robust features, but also includes a sparse network structure.

D. ANALYSIS OF DROPOUT

Finally, as a supplement, we also investigate the effect of dropout on performance of the proposed method. Here, we set the step size to 0.1 and the dropout rate is changed from 0 to 0.5, the 20% random Gaussian noise was added into the raw data, and the number of the stack was 7. As shown in Figure 12, different dropout rates have different classification performances. The result show that when the dropout rate was close to 0.2, we can obtain the best classification performance. When the dropout rate was more than 0.2, the performance would decrease. This result indicated that

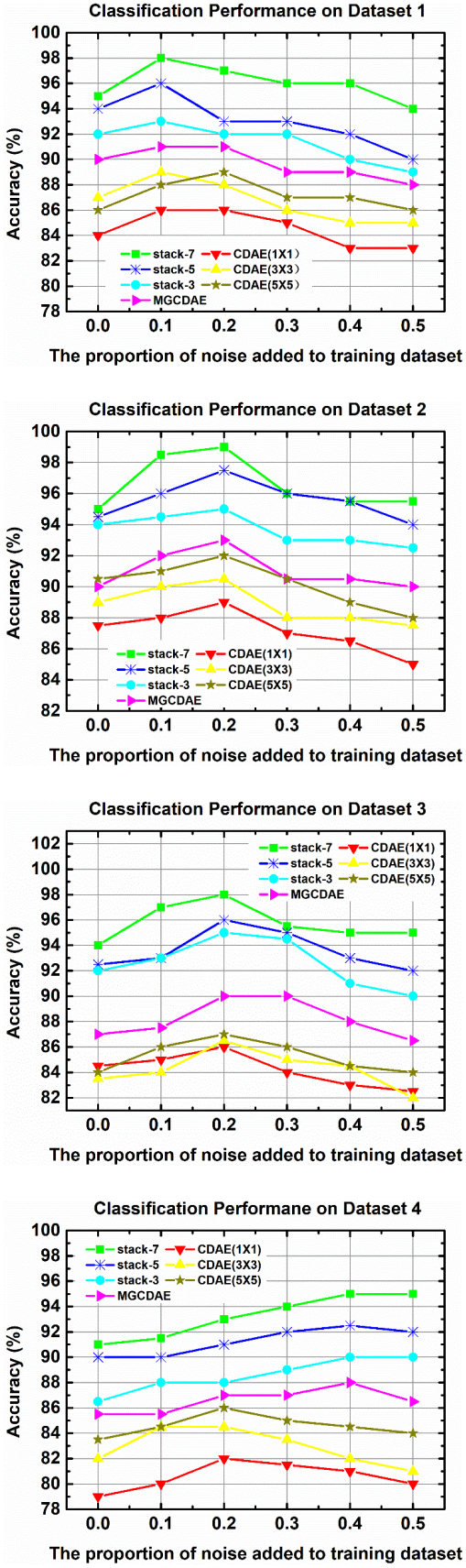

FIGURE 11. The effect of varied noise levels on classification performance.

TABLE 3. Classification accuracy for the existing traditional machine learning method.

Approach	Datasets			
	dataset 1	dataset 2	dataset 3	dataset 4
RF	92%	95%	95%	86%
CNN	94%	97%	95.5%	89%
SVM	95%	93.5%	93%	86.5%
DBN	94%	97.5%	97%	90%
Our Approach	97%	99%	98%	93%

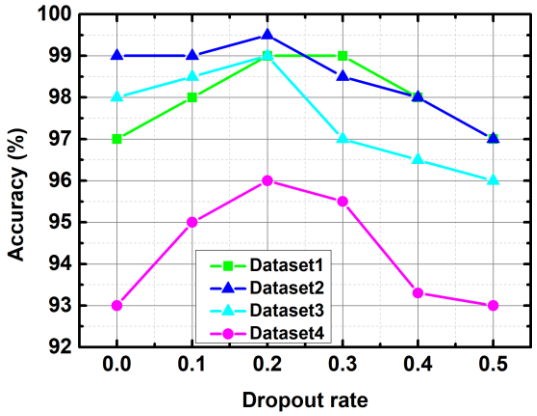

FIGURE 12. Study on the effect of dropout.

appropriate dropout is beneficial for the performance of the proposed method.

E. DISSCUSION

As reported above experiments, our approach has the better generalization performance for general classification tasks. Among these comparative experiments, we discussed the influence of some parameters on experimental results, such as the size of the convolution neural, whether to add the noise, the proportion of the added noise, the number of the MGCDAE stacked and whether to add dropout method. In a word, add certain proportion of noise and stack MGCDAE with dropout method can improve the accuracy of the model.

Although the results show that our approach achieved high quality generalization performance, there still remain some issues. For example, it is not superior to other methods in computing time. In some cases, it even exceeds the comparison method. With increase of the number of superposition on MGCDAE, the computing time could increase further, but the accuracy is unlikely to continue to grow. Furthermore, we use three size convolution kernels to optimize the model, but the further problem is that how to automatic select the type of convolution kernel for different data types. The larger the convolution kernel, the greater the time and complexity of calculation.

VII. CONCLUSION AND FUTURE WORK

This paper proposed a novel intelligent bearing fault diagnosis method based on the CS and an unsupervised feature extraction approach (SMGCDAE). The compressed data has the ability to capture the discriminative information that can be used to extract features automatically. Then, a CNN based on DAE is built to mine the useful information and finish the fault classification by softmax classifier. In addition, the SMGCDAE improved on existing approach via introducing the concept of the multi-granularity convolution kernels and used the dropout to prevent overfitting. The case studies of bearing data sets demonstrated the robustness and effectiveness of this technique. The CS-SMGCDAE intelligent diagnosis method can obtain relatively high identification accuracy with small amount of measure data. The proposed method provides a new idea for mechanical big data processing. In the future work, we will further explore a general model and apply it to other datasets.

ACKNOWLEDGMENT

The authors would like to thank BMW Brilliance Automotive Ltd. for allowing them to use their data.

REFERENCES

[1] H. Jiang, C. Li, and H. Li, "An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis," *Mech. Syst. Signal Process.*, vol. 36, no. 2, pp. 225–239, Apr. 2013.

[2] Z. Huo, Y. Zhang, P. Francq, L. Shu, and J. Huang, "Incipient fault diagnosis of roller bearing using optimized wavelet transform based multi-speed vibration signatures," *IEEE Access*, vol. 5, pp. 19442–19456, 2017.

[3] X.-X. Ren, G.-H. Yang, and X.-J. Li, "Sampled observer-based adaptive output feedback fault-tolerant control for a class of strict-feedback nonlinear systems," *J. Franklin Inst.*, vol. 356, no. 12, pp. 6041–6070, Aug. 2019.

[4] Y. Miao, M. Zhao, J. Lin, and Y. Lei, "Application of an improved maximum correlated Kurtosis deconvolution method for fault diagnosis of rolling element bearings," *Mech. Syst. Signal Process.*, vol. 92, pp. 173–195, Aug. 2017.

[5] X. Ding and Q. He, "Time–frequency manifold sparse reconstruction: A novel method for bearing fault feature extraction," *Mech. Syst. Signal Process.*, vol. 80, pp. 392–413, Dec. 2016.

[6] W. Mao, J. Wang, L. He, and Y. Tian, "Online sequential prediction of imbalance data with two-stage hybrid strategy by extreme learning machine," *Neurocomputing*, vol. 261, pp. 94–105, Oct. 2017.

[7] X. Jin, M. Zhao, T. W. S. Chow, and M. Pecht, "Motor bearing fault diagnosis using trace ratio linear discriminant analysis," *IEEE Trans. Ind. Electron.*, vol. 61, no. 5, pp. 2441–2451, May 2014.

[8] W. Li, S. Zhang, and G. He, "Semisupervised distance-preserving self-organizing map for machine-defect detection and classification," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 5, pp. 869–879, May 2013.

[9] C. Shen, D. Wang, F. Kong, and P. W. Tse, "Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier," *Measurement*, vol. 46, no. 4, pp. 1551–1564, May 2013.

[10] N. Li, R. Zhou, Q. Hu, and X. Liu, "Mechanical fault diagnosis based on redundant second generation wavelet packet transform, neighborhood rough set and support vector machine," *Mech. Syst. Signal Process.*, vol. 28, pp. 608–621, Apr. 2012.

[11] X. Zhang and J. Zhou, "Multi-fault diagnosis for rolling element bearings based on ensemble empirical mode decomposition and optimized support vector machines," *Mech. Syst. Signal Process.*, vol. 41, nos. 1–2, pp. 127–140, Dec. 2013.

[12] X.-G. Zhang and G.-H. Yang, "Optimal sensor attacks in cyber-physical systems with round-robin protocol," *Inf. Sci.*, vol. 548, pp. 85–100, Feb. 2021.

[13] G. F. Bin, J. J. Gao, X. J. Li, and B. S. Dhillon, "Early fault diagnosis of rotating machinery based on wavelet packets-empirical mode decomposition feature extraction and neural network," *Mech. Syst. Signal Process.*, vol. 27, no. 2012, pp. 696–711, Sep. 2011.

[14] M. Van and H.-J. Kang, "Wavelet kernel local Fisher discriminant analysis with particle swarm optimization algorithm for bearing defect classification," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 12, pp. 3588–3600, Dec. 2015.

[15] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, 2007.

[16] M. Cerrada, G. Zurita, D. Cabrera, R. V. Sanchez, M. Artes, and C. Li, "Fault diagnosis in spur gears based on genetic algorithm and random forest," *Mech. Syst. Signal Process.*, vols. 70–71, pp. 87–103, Mar. 2016.

[17] P. Baraldi, F. Cannarile, F. Di Maio, and E. Zio, "Hierarchical K-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions," *Eng. Appl. Artif. Intell.*, vol. 56, pp. 1–13, Nov. 2016.

[18] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[19] M. F. Duarte, M. A. Davenport, D. Takhar, and J. N. Laska, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[20] S. H. Hsieh, W. Liang, C. S. Lu, and S. C. Pei, "Distributed compressive sensing: Performance analysis with diverse signal ensembles," *IEEE Trans. Signal Process.*, vol. 68, pp. 3500–3514, 2020.

[21] Z. Zhang, T.-P. Jung, S. Makeig, and B. D. Rao, "Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ECG via block sparse Bayesian learning," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 300–309, Feb. 2013.

[22] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[23] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[24] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[25] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[26] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[28] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.

[29] X. Wang, Y. Qin, Y. Wang, S. Xiang, and H. Chen, "ReLTanh: An activation function with vanishing gradient resistance for SAE-based DNNs and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 363, pp. 88–98, Oct. 2019.

[30] J. Kim and M. Sohaib, "Reliable fault diagnosis of rotary machine bearings using a stacked sparse autoencoder-based deep neural network," *Shock Vib.*, vol. 2018, no. 12, May 2018, Art. no. 2919637.

[31] G. Liu, H. Bao, and B. Han, "A stacked autoencoder-based deep neural network for achieving gearbox fault diagnosis," *Math. Problems Eng.*, vol. 2018, Jul. 2018, Art. no. 5105709.

[32] Z. Yang, D. Gjorgjevikj, J. Long, Y. Zi, S. Zhang, and C. Li, "Sparse autoencoder-based multi-head deep neural networks for machinery fault diagnostics with detection of novelties," *Chin. J. Mech. Eng.*, vol. 34, no. 1, p. 54, Dec. 2021.

[33] H. Shao, H. Jiang, H. Zhang, W. Duan, T. Liang, and S. Wu, "Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing," *Mech. Syst. Signal Process.*, vol. 100, pp. 743–765, Feb. 2018.

[34] H. Zhiyi, S. Haidong, Z. Xiang, Y. Yu, and C. Junsheng, "An intelligent fault diagnosis method for rotor-bearing system using small labeled infrared thermal images and enhanced CNN transferred from CAE," *Adv. Eng. Informat.*, vol. 46, Oct. 2020, Art. no. 101150.

[35] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.* Berlin, Germany: Springer, 2000, pp. 1–15.
[36] Y. Yang and J. Jiang, "Hybrid sampling-based clustering ensemble with global and local constitutions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 952–965, May 2016.
[37] Y. Yang, L. J. Cao, Q. Liu, and P. Yang, "A stacked multi-granularity convolution denoising auto-encoder," *IEEE Access*, vol. 7, pp. 83888–83899, 2019.
[38] G. Hinton, N. Srivastava, A. Krizhevsky, R. R. Salakhutdinov, and I. Sutskever, "Improving neural networks by preventing co-adaptation of feature detectors," *Comput. Sci.*, vol. 3, no. 4, pp. 212–223, Jul. 2012.
[39] D. Zhao, F. Zhao, and Y. Gan, "Reference-driven compressed sensing MR image reconstruction using deep convolutional neural networks without pre-training," *Sensors*, vol. 20, no. 1, p. 308, Jan. 2020.
[40] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[41] L. Shi, G. Qu, and Q. Wang, "A method of reweighting the sensing matrix for compressed sensing," *IEEE Access*, vol. 9, pp. 21425–21432, 2021.
[42] N. Y. Yu and Y. Li, "Deterministic construction of Fourier-based compressed sensing matrices using an almost difference set," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, pp. 1–14, Dec. 2013.
[43] A. H. Elsheikh, M. F. Wheeler, and I. Hoteit, "Sparse calibration of subsurface flow models using nonlinear orthogonal matching pursuit and an iterative stochastic ensemble method," *Adv. Water Resour.*, vol. 56, pp. 14–26, Jun. 2013.
[44] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals," *IEEE Access*, vol. 6, pp. 25399–25410, 2019.
[45] S. Amini and S. Ghaemmaghami, "Towards improving robustness of deep neural networks to adversarial perturbations," *IEEE Trans. Multimedia*, vol. 27, no. 7, pp. 1889–1903, Jul. 2020.
[46] N. C. Noh and J. P. Heo, "Mutually orthogonal softmax axes for cross-domain retrieval," *IEEE Access*, vol. 8, pp. 56491–56500, 2020.
[47] N. Ahmed and M. Campbell, "Variational Bayesian learning of probabilistic discriminative models with latent softmax variables," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3143–3154, Jul. 2011.
[48] C. Zhang and P. C. Woodland, "DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5300–5304.

CHUANG LIANG was born in 1992. He is currently pursuing the Ph.D. degree jointly trained by the School of Mechanical Engineering, Shenyang University of Technology, China, and BMW Brilliance Automotive Ltd. His current research interests include signal process and intelligent fault diagnosis of rotating machine in automotive production line.

CHANGZHENG CHEN was born in 1964. He is currently a Professor and a Ph.D. Supervisor with the School of Mechanical Engineering, Shenyang University of Technology, China. His current research interests include vibration, noise, and fault diagnosis.

YE LIU, photograph and biography not available at the time of publication.

XINYING JIA, photograph and biography not available at the time of publication.