IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 9, 2020, accepted April 18, 2020, date of publication April 21, 2020, date of current version May 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2989371

An Adaptive Anti-Noise Neural Network for Bearing Fault Diagnosis Under Noise and Varying Load Conditions

GUOQIANG JIN, TIANYI ZHU, MUHAMMAD WAQAR AKRAM, YI JIN, (Member, IEEE), AND CHANGAN ZHU

Department of Precision Machinery and Instrumentation, University of Science and Technology of China, Hefei 230026, China

Corresponding authors: Yi Jin (jinyi08@ustc.edu.cn) and Changan Zhu (changan@ustc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51605464, and in part by the National Major Scientific Instruments and Equipments Development Project of the National Natural Science Foundation of China under Grant 61727809.

ABSTRACT Fault diagnosis in rolling bearings is an indispensable part of maintaining the normal operation of modern machinery, especially under the varying operating conditions. In this paper, an end-to-end adaptive anti-noise neural network framework (AAnNet) is proposed to solve the bearing fault diagnosis problem under heavy noise and varying load conditions, which takes the raw signal as input without requiring manual feature selection or denoising procedures. The proposed AAnNet employs the random sampling strategy and enhanced convolutional neural networks with the exponential linear unit as the activation function to increase the adaptability of the neural network. Moreover, the gated recurrent neural networks with attention mechanism improvement are further adopted to learn and classify the features processed by the convolutional neural networks part. Besides, we try to explain how the network works by visualizing the intrinsic features of the proposed framework. And we explore the effect of the attention mechanism in the proposed framework. Experiments show that the proposed framework achieves state-of-the-art results on two datasets under varying operating conditions.

INDEX TERMS Bearing fault diagnosis, convolutional neural network, deep learning, load domain adaptation, noisy conditions, recurrent neural network.

I. INTRODUCTION

Fault diagnosis in mechanical equipment has gained significant attention in the modern industry. Failures of mechanical equipment could result in economic loss and casualties [1]. Rolling element bearings are the critical target for fault diagnosis in mechanical equipment, especially in rotating machinery, which accounts for a large proportion of failures [2]. In the past few decades, fault diagnosis in rolling bearings has been widely studied. Data-driven based methods are commonly used in bearing fault diagnosis [3].

Recently, with the development of training optimization algorithms of the deeper neural networks, the difficulty of training more complex networks decreases; and various deeper neural networks begin to be widely studied [4]. Convolutional neural networks (CNNs) [5] and recurrent neural networks (RNNs) [6] are the most commonly used deep learning networks nowadays, which have been well-developed and widely applied in various tasks including bearing fault diagnosis [7]–[9]. The gated recurrent unit (GRU) [10] is one of the improved versions of RNN, which could capture time-dependent characteristics from the signals and is widely used in sequence-based tasks like language and speech-related works [11]. The vibration signal from the bearing is similar to the sentence and language with time-dependent characteristics. Thus, it is reasonable to apply the gated recurrent unit for the diagnosis of the bearing fault.

In early applications of CNNs or RNNs for bearing fault diagnosis, it was common to employ preprocessing procedure to extract features from the raw signal and then use the networks to classify bearing fault types. Abed *et al.* [12] proposed an RNN for bearing fault classification using selected discrete wavelet transforms features as the input. In [13], the author employed the discrete Fourier transform (DFT) to extract features from raw signals and then used CNN to learn features for bearing fault diagnosis. In [14], statistical features

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

were manually extracted from the vibration data and then used as input to a CNN based classifier. In [15], the proposed deep neural network (DNN) employed the frequency spectra of the raw signal as the input. The proposed method could classify not only the fault type but also the fault severity. Xie and Zhang [16] used the discrete wavelet transform (DWT) as the feature extractor for CNN and classified not only the types of bearing fault but also the severity of the fault. In [17], the proposed method took the wavelet packet energy (WPE) image from the raw signal as the input of CNN, which could perform multiclass classification for spindle bearing fault diagnosis regardless of the load fluctuation. In [18], the proposed method used stacked RNNs and long short-term memory (LSTM) networks that took frequency spectrum sequences as the input and adopted an adaptive learning rate. The proposed method could classify not only the types of bearing fault but also the severity of the fault. In [19], the proposed two-stage cross-domain fault diagnosis method based on deep generative neural networks took the frequency domain data as input. The proposed method could artificially generate fake samples on the target domain for domain adaptation missions, which performed well under varying load conditions and could classify both the types of bearing fault and the severity of the fault. In [20], the proposed deep fully convolutional neural network (DFCNN) took the spectrogram as the input and obtained classification results under seven different signals, including the fault type and corresponding severity. However, these aforementioned early studies require complicated manual feature extraction and selection processes, which increases the difficulty of bearing fault diagnosis.

With the development of deep learning and the increase in the depth of the neural network, the feature representation and learning ability of CNNs have been greatly improved. More and more networks use simple preprocessing and data augmentation methods to directly take raw signals as the input instead of complex feature selection algorithms. Qian *et al.* [21] proposed an adaptive overlapping convolutional neural network (AOCNN) that took data segmentation with overlapping as the data augmentation method and adopted an adaptive layer to overcome the shift variant and marginal problems. The proposed method performed well under a small-sized training dataset. In [22], the proposed deep convolutional neural network found that the data augmentation method of signal translation could effectively improve the performance of bearing fault diagnosis. Since RNNs can better obtain the intrinsic characteristics of time series data, and the bearing signal is also a time-dependent sequence, so the application of RNNs in bearing fault diagnosis has gained more attention in recent years. Pan *et al.* [23] implemented CNN as the feature extractor for the LSTM. The proposed end-to-end network could classify not only the fault type but also the fault severity. Zhao *et al.* [24] proposed the local feature-based gated recurrent unit (LFGRU) networks that used handcrafted local-feature extraction scheme to generate features for the bidirectional GRU network. Then, additional center-biased averaging features from the input were used as the supplementary information together with the output of GRU to compute the final result. Yu *et al.* [25] proposed a stacked LSTM network that employed the raw signal with data augmentation as the input, which could classify the fault type and fault severity. However, the aforementioned algorithms do not consider bearing fault diagnosis under noisy conditions.

Background noise is a common and unavoidable disturbance in industrial sites. Application of deep learning in bearing fault diagnosis under noisy environments has received more and more attention in recent years. Lu *et al.* [26] proposed a four-layer CNN with the time-frequency features extracted from training set as the input, which could perform the bearing fault classification under noisy conditions with the signal-to-noise ratio (SNR) between 10 dB to 50 dB. In [27], the proposed CNN based neural network employed the distance metric learning method to increase the domain adaptation ability, which took the frequency features of the raw signal as input and performed the bearing fault classification with SNR between −8 dB to 8 dB and under varying load conditions. In [28], the proposed GRU-based nonlinear predictive denoising autoencoders (GRU-NP-DAEs) employed the length loss method to enhance the robustness of models, which could utilize information from multiple sensors and achieved good accuracy with the SNR between 1 dB to 10 dB. In [29], the proposed neural network employed the residual structure to reduce the training difficulty of a deeper neural network, which performed the bearing fault classification with SNR between 0 dB to 8 dB. Qiao *et al.* [30] proposed an adaptive weighted multiscale convolutional neural network (AWMSCNN) that could adaptively extract multiscale features from raw signals, which was tested with the SNR between −3 dB to 7 dB. Li *et al.* [31] proposed a novel transfer learning method based on domain adversarial training to extract the underlying shared features across multiple source domains to diagnose the target domain. The proposed method was tested under noisy conditions with the SNR between −4 dB to 8 dB. Zhang *et al.* [32] proposed a DCNN with wide first-layer kernels (WDCNN) that took the raw signal with data augmentation as the input, which performed well with the SNR between −4 dB to 10 dB and tested under varying load conditions. In [33], our previous work proposed a structure optimized DCNN that achieved similar performance compared to the WDCNN, using less than half parameters. Zhang *et al.* [34] proposed a CNN with training interference (TICNN) with data augmentation, that achieved higher accuracy with the SNR between −4 dB to 10 dB and tested under varying load conditions. However, there is still room for improvement, especially under higher noise conditions.

After the above comprehensive review, we can conclude that although many algorithms can achieve a good classification result of bearing fault types, not all algorithms can achieve the classification of bearing fault severity. Moreover, many algorithms require sophisticated manual feature design and selection procedures, denoising procedures, and data augmentation, which increase the difficulty of bearing fault diagnosis. Furthermore, researches in complex environments, especially under noisy and varying load conditions, are still relatively few, and these methods still have room for improvement. The bearing fault classification in varying operating conditions is still a challenging task.

In this paper, we proposed an end-to-end adaptive anti-noise neural network framework (AAnNet) to address the above problems, which combines the advantages of multiple structures, which can distinguish not only the different bearing fault types but also the severity of the corresponding fault. In order to address the noise problem in the bearing fault diagnosis, the AAnNet employs the random sampling and exponential linear unit in CNN to improve the adaptability against varying levels of noise. Moreover, the proposed method combines the advantages of multiple structures, including the powerful feature extraction capability of CNN, the innate timing-dependent sequence processing capability of GRU, and the feature abstraction and fusion ability of attention mechanism. The combination improves the adaptability and generality of the neural network in complex environments.

The proposed AAnNet is an end-to-end neural network that automatically extracts features from raw signals instead of using handicraft features. It does not require expert knowledge for manual filter design and manual feature selection procedures. The main contributions of this article are listed as follows:

1) We proposed an end-to-end AAnNet that could adaptably perform the classification of bearing fault and corresponding severity without requiring manual feature selection and denoising procedures and achieved state-of-the-art results on two datasets.
2) We proposed random sampling as the data input strategy and a combination of CNN, GRU, and the attention mechanism to improve the anti-noise and domain adaptation of the network.
3) The proposed AAnNet has great performance under heavy noise conditions and has strong domain adaptation against varying load conditions.
4) We investigated the effect of using the attention mechanism and tried to explore the intrinsic features of the neural network by visualizing the kernel weight distribution and activation values of the proposed AAnNet.

The rest of this paper is organized as follows: The basic theory of the neural network is provided in Section II. The details of the proposed AAnNet is provided in Section III. In Section IV, several experiments are conducted on two datasets under noisy conditions and varying load conditions to validate the proposed framework. Then, network visualization and comprehensive analysis are performed to analyze and evaluate the applicability of the proposed method. Finally, Section V concludes the whole paper.

II. BASIC THEORY OF NEURAL NETWORK

A. CONVOLUTIONAL NEURAL NETWORK

Inspired by biological neural processes, a convolutional neural network is composed of multiple connected neurons, and each neuron is only responsible for a small partially overlap receptive field, which greatly reduces the number of parameters and the training difficulty [4], [35]. Another advantage of a convolutional neural network is that it can handle inputs with variable lengths, whereas the traditional neural network could not handle variable lengths of input. Batch normalization (BN) [36] has been widely used in the deep neural networks together with the convolutional layer. The BN layer could improve the accuracy and training speed of a deep neural network by solving the internal covariate shift problem. Dropout reduces the overfitting problem by temporarily dropping units from the neural network according to a certain probability during training [37]. This random dropping method effectively prevents units from co-adapting. Dropout could make the model more robust and reduce the impact of noise on the model, which improves the accuracy of the model under noisy conditions.

In this paper, we proposed to use a noise-robust activation to increase the adaptability under noise conditions. The activation function in the neural network gives neurons a nonlinear expression of the convolution result with the input signal. The exponential linear unit (ELU) [38] inherits the advantages of the Rectified linear unit (ReLU) [39] that solves the well-known vanishing gradients problem, whereas overcoming the "dying ReLU" problem [40]. Moreover, the ELU has non-zero activation value and gradient at the left side, which decreases the bias shift problem by pushing mean unit activations closer to zero. This operation is like batch normalization but with lower computational complexity. The ELU is defined as follows:

$$ELU(x) = \begin{cases} \alpha(exp(x) - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0, \end{cases} \tag{1}$$

where α is a positive hyper-parameter that controls saturation for negative input values. The ELU is noise-robust under deactivation state, thanks to the saturation when inputting small negative value. We will later compare the effects of ReLU and ELU on the proposed method.

B. GATED RECURRENT UNIT NEURAL NETWORK

The gated recurrent unit (GRU) is an improved version of RNN that could retain the context of the input sequence via internal states to capture long-term dependence by the structures of gates, whereas overcoming the well-known vanishing or exploding gradient problem of RNN [42]. These gates are used to remove or add information to the hidden state to decide whether to remember long-term dependence or use short-term information. As shown in Fig.1, the workflow of the GRU is described as follows:

At time step t, the reset gate $\boldsymbol{r}_t$ is computed by

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r), \tag{2}$$

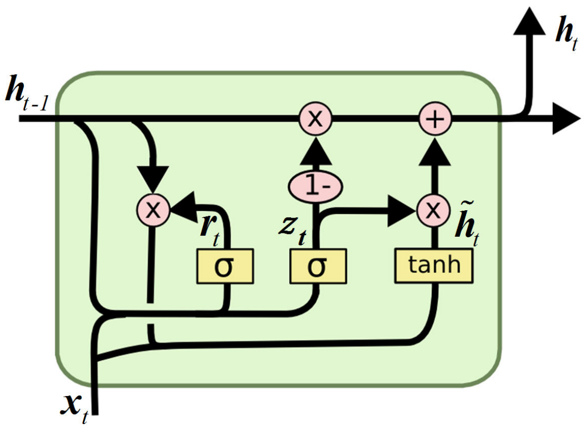


FIGURE 1. The basic structure of a GRU cell [41].

where $\boldsymbol{x}_t$ is the input, $\boldsymbol{W}_r$ and $\boldsymbol{U}_r$ are the weight matrices to be learned, $\boldsymbol{b}_r$ is bias weight, $\boldsymbol{h}_{t-1}$ is the previous activation, σ is a logistic sigmoid function. The reset gate $\boldsymbol{r}_t$ decides how much the past activation to be kept for the computation of the candidate activation $\tilde{\boldsymbol{h}}_t$ which is computed by

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_h\boldsymbol{x}_t + \boldsymbol{r}_t \odot (\boldsymbol{U}_h\boldsymbol{h}_{t-1}) + \boldsymbol{b}_h), \tag{3}$$

where $\odot$ represents an element-wise multiplication, $\boldsymbol{W}_h$ and $\boldsymbol{U}_h$ are the weight matrices. The candidate activation $\tilde{\boldsymbol{h}}_t$ is computed by the input $\boldsymbol{x}_t$ and previous step $\boldsymbol{h}_{t-1}$, which is modulated by the reset gates $\boldsymbol{r}_t$. The update gate z_t controls the proportion of the past activation and the candidate activation, expressed by

$$z_t = \sigma(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z), \tag{4}$$

where $\boldsymbol{W}_z$ and $\boldsymbol{U}_z$ are the weight matrices, $\boldsymbol{b}_z$ is bias weights, $\boldsymbol{h}_{t-1}$ is the previous activation. Finally, the new activation $\boldsymbol{h}_t$ is computed by

$$\boldsymbol{h}_t = (1 - z_{t-1}) \odot \boldsymbol{h}_{t-1} + z_t \odot \tilde{\boldsymbol{h}}_t, \tag{5}$$

where z_{t-1} is the previous update gate status. The new activation is mixed from the past activation and the current candidate activation, where the update gate controls the mix proportion.

When an important feature in the signal is detected, the update gate z_t will adaptively adjust to allow the memory content to transfer across multiple time steps to capture information at different time scales, thus easily carrying the feature over a long distance, i.e., preserving the long-term dependencies. The reset gate $\boldsymbol{r}_t$ helps the GRU to ignore the previous hidden states to capture the new short-term dependencies whenever the detected feature is not necessary anymore [43]. Thanks to the application of these control gates, a GRU could adaptively generate output based on the long-term relationship and current short-term information, thus has better performance with the time-depended sequence.

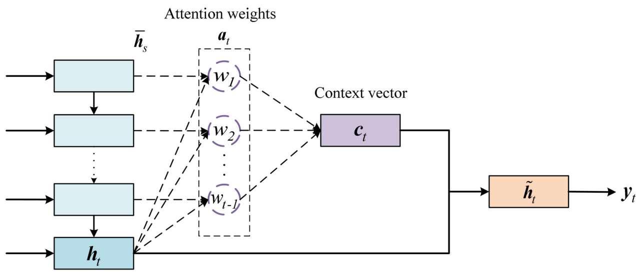


FIGURE 2. The basic structure of the attention mechanism.

C. ATTENTION MECHANISM

The attention mechanism [44] can automatically focus on relevant information and pay more attention to the inherent characteristics of the feature. Since not all features have the same contribution to the final classification result, the attention mechanism is employed to automatically select the most relevant features, thereby improving the performance of the proposed network.

As shown in Fig.2, the attention mechanism takes advantage of full information by using all time steps, which employs the weighting method for each time step to get the final result, instead of manually selecting a time step as the final output. The main calculation process of the attention mechanism is described as follows:

Given all the hidden source states $\bar{\boldsymbol{h}}_s$ and the current target state $\boldsymbol{h}_t$, the attention weights vector $\boldsymbol{a}_t$ is calculated as follows:

$$\boldsymbol{a}_t(s) = \frac{\exp(score(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s))}{\sum_{s'} \exp(score(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'}))}, \tag{6}$$

where the function *score* is designed to compare the current target hidden state $\boldsymbol{h}_t$ with each source hidden state $\bar{\boldsymbol{h}}_s$ to calculate the contribution of each time step to the final output, which is denoted as follows:

$$score(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \boldsymbol{h}_t^{\top} \boldsymbol{W}_s \bar{\boldsymbol{h}}_s, \tag{7}$$

where $\boldsymbol{W}_s$ is the trainable weight matrix. Then, by having the attention weight vector $\boldsymbol{a}_t$ and each time step $\bar{\boldsymbol{h}}_s$, the context vector $\boldsymbol{c}_t$, i.e., the weighted time steps vector is computed by:

$$\boldsymbol{c}_t = \sum_s \boldsymbol{a}_t(s)\bar{\boldsymbol{h}}_s. \tag{8}$$

Finally, we concatenate the context vector $\boldsymbol{c}_t$ and the target hidden state $\boldsymbol{h}_t$ to get the output attentional hidden state $\tilde{\boldsymbol{h}}_t$, denoted as

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t]), \tag{9}$$

where the $\boldsymbol{W}_c$ is the trainable weight matrix.

The attention mechanism could combine high-dimensional information and select the most important features to improve the classification result, which utilizes the information of different time steps from the output of an RNN. Then, the weighted time steps are processed as the final output.

The weight matrix is automatically obtained by learning without manual selection, which could reduce the difficulty of parameter selection and improve the robustness of the proposed network. Through the attention mechanism, the network could pay more attention to the discriminative features, and the bearing health features can be effectively captured [45]. Further discussion of the attention mechanism will be present in Section IV-E.3.

III. THE PROPOSED AAnNet NEURAL NETWORK

The main framework of the proposed AAnNet is shown in Fig.3. The proposed end-to-end adaptive anti-noise neural network mainly consists of four parts: data input strategy based on random sampling; feature extractor based on enhanced CNN that employs ELU as the activation function; feature classifier based on GRU; and feature post-processor based on attention mechanism and DNN. The proposed method combines the advantages of multiple structures to improves adaptability and generality in complex environments.

The input of the AAnNet is a fixed-length segment of raw data without handcrafted features or denoising procedures. The output of the AAnNet depends on the type of tasks. It can be only the types of bearing fault, which have a small number of categories, or it can be fault types with the corresponding severity, which have a large number of categories. The details of the proposed AAnNet are described in the following section.

A. ENHANCED NETWORK GENERALITY UNDER NOISY CONDITION

The proposed AAnNet uses CNN as the feature extractor to take raw signals as the input instead of manually selected features. The random sampling strategy in the proposed network is similar to the bootstrap sampling [46] in statistics, which is a simple way of getting an estimated distribution of an uncertain dataset. The bearing fault signals with noise can be regarded as an uncertain dataset. By employing the random sampling strategy, we can have a relatively easier way to estimate the bearing fault characteristics from noisy signals. On the other hand, random sampling could be regarded as the simulation of noise interference in the training process, which improves the adaptability of the neural network under noise conditions. Thus we can have a good test result from the noise-added data but only using the data that is not added with noise to train the network.

The random sampling procedure is shown in Fig.4. The sampling of each point in the data obeys the Bernoulli distribution with probability p, where p is 0.5 in this paper. The perceptual neurons in the input layer are randomly shut down to simulate the random sampling procedure to the input signal. In order to have a fixed input length, the randomly chosen data points will be set to zero, denoted as red points in Fig.4(b).

In each convolutional layer, the ELU is adopted as the activation function. Unlike the widely used ReLU, the ELU could push mean unit activations closer to zero to reduce the bias shift effect. The ELU preserves negative information and provides more information for subsequent calculations and ensures a noise-robust deactivation state that leads to higher generality under noisy conditions. Furthermore, inspired by [32], [47], larger receptive field size comes with more useful context information from noisy signals, we use wide layer kernels in the convolutional layers to enhance the visual range of the network to extract more features from the raw signal. Batch normalization is implemented to improve stability and speed up network convergence. The dropout is also implemented to make the model more robust under noisy conditions as well as to prevent the over-fitting problem.

The GRU could capture the time-dependent features from the input, unlike the traditional CNN. We use GRU to capture the latent information of the preprocessed features from CNN to generalize the classification result. The combination of CNN and GRU could better capture the latent information of the sequence and enhance the ability of the network to deal with more complex situations, especially the heavy noise conditions and varying load conditions. Compared with the network composed entirely of CNNs, the combination of CNN and GRU could take advantage of CNN's powerful feature extraction capabilities and GRU's information processing ability with time-dependent signals.

The attention mechanism and the following DNN part are employed to make full use of the output from the GRU. Usually, we need to manually specify the time step of the GRU output as the final output, which requires more parameter selection operations and increases the training difficulty. By employing the attention mechanism, the network can be automatically trained to find the appropriate time step as the final output, furthermore, focusing more on the intrinsic characteristics of the signal. The DNN part with the softmax is used to generate the final classification result.

B. IMPLEMENTATION DETAILS OF THE PROPOSED AAnNet

A typical structure of the proposed AAnNet is mainly composed of one input layer, two convolutional layers, two GRU layers, the attention mechanism layer, and the DNN layers. The proposed network uses a fixed-length input. We implement the random sampling strategy by using dropout in the input layer. After the input layer, there are two convolutional layers as the feature extractor which uses ELU as the activation function. After each convolutional layer, the batch normalization layer and dropout layer are adopted to improve stability and generality. Then, two GRU layers come after the CNN part, followed by another dropout layer. The attention mechanism layer is after the dropout layer. Finally, comes the two fully-connected layers and one softmax layer to generate the classification result. Assuming the input length is 1670, and the final output is 10 categories, the detailed network parameters are shown in Table.1.

The dropout rate is fixed to 0.5. The convolution kernel size is $128 * 1$, and the numbers of filters are 64 and 72 in the

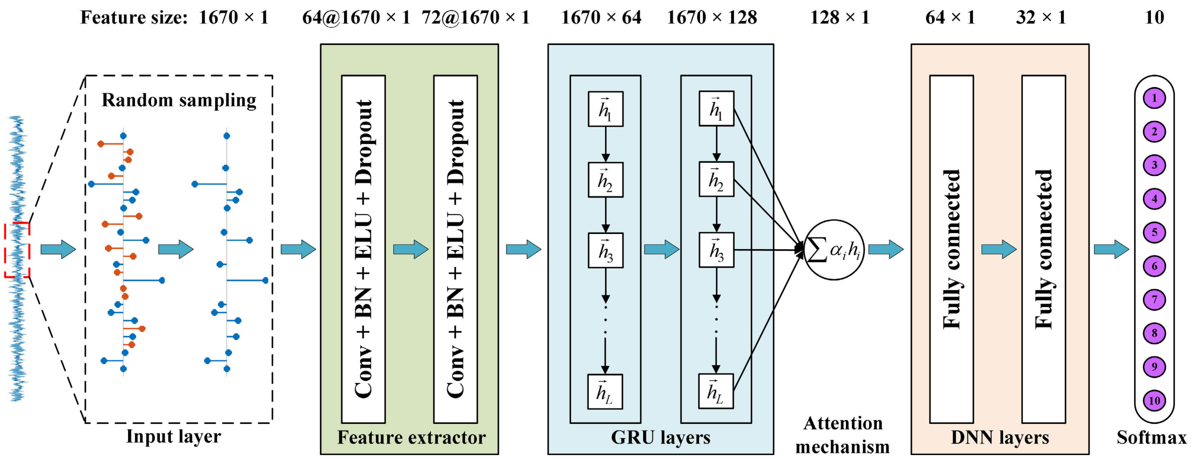


FIGURE 3. Architecture of the proposed AAnNet.

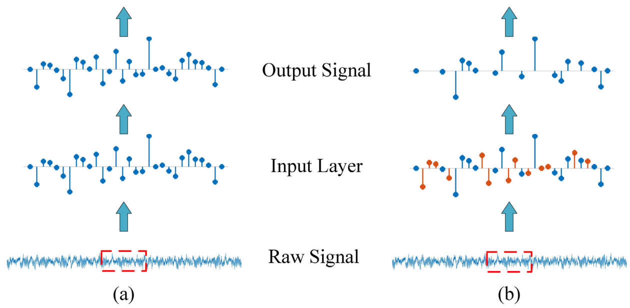


FIGURE 4. Data input methods: (a) without random sampling, (b) with random sampling.

TABLE 1. Parameters of a typical structure of the proposed AAnNet.

No.	Layer Type	Kernel Size	Output Size (Width × Depth)	Number of Kernels
1	Input Layer	-	1670*1	-
2	Dropout			
3	Convolution1	128*1	1670*64	64
4	BN + Dropout			
5	Convolution2	128*1	1670*72	72
6	BN + Dropout			
7	GRU Layer1	-	1670*64	64
8	GRU Layer2	-	1670*128	128
9	Dropout			
10	Attention Layer	-	128*1	-
11	Fully-connected1	-	64*1	64
12	Fully-connected2	-	32*1	32
13	Softmax	-	10	10

first and second convolutional layers, respectively. The stride in the convolutional layer is 1, and no padding is used in the convolutional layer. The numbers of neurons in the GRU are 64 and 128 in the first and second GRU layers, respectively. There are 128 neurons in the attentional hidden state vector. Then followed by the two fully-connected layers which have 64 and 32 neurons each. Finally comes the softmax layer with 10 neurons. We use Adam [48] as the optimizer and train each model for 10000 epochs.

To test the performance of the proposed framework to the maximum extent, each data used in training is independent. Therefore, the original signal is directly divided into the specified length without overlapping, and no data augmentation method is used. Thus, every sample of the segmented data will not have an overlapped part. Each training data will be independent. We use the minimum number of samples from the complete dataset to train the neural network. In order to verify the generality of models under noisy data, we train all the models on the original data without the addition of noise and test them on the data having different levels of noise added.

IV. VALIDATION OF PROPOSED AAnNet

In this paper, we tested our proposed algorithm on two different datasets separately. The first dataset is from the Case Western Reserve University (CWRU) Bearing Data Center [49], which has been a benchmark dataset for bearing fault diagnosis in recent years. The second dataset is from our bearing fault diagnosis experimental platform QPZZ-*II*. We conducted the bearing fault diagnosis experiment to collect four types of bearing vibration signals. We verified the performance of the proposed method on these two datasets and compared it with other methods. Each model has been trained five times with a fixed random seed. The experiments were run on an NVIDIA TITAN Xp GPU using the Keras [50] and TensorFlow frameworks.

A. PREPROCESSING SIGNAL

Since the input length of the proposed algorithm is a fixed value, so the original signal needs to be segmented into fixed-length signals. As the vibration signal is periodic, and the period is related to the motor speed, it is reasonable to segment the signal according to the period. In this paper, the input length of the signal is the number of points collected within one complete motor rotation cycle. This kind of signal segmentation method is the smallest division of the complete useful information. For example, if the sampling rate is 48000 Hz, and the motor speed is 1724 rpm, then the accelerometer will collect about 1670 points within a motor rotation cycle. In this paper, the amplitude of the input signal is normalized to $[-1, +1]$.

Typically, a feasible approach to data augmentation is to segment the raw data with overlapping, as described in [34]. In this paper, we use independent samples from the dataset, i.e., no overlapping between each sample to verify the adaptability of the proposed model.

1) ADDING NOISE TO THE SIGNAL

Additive white Gaussian noise (AWGN) [51] is widely present in real situations, which mimics the effect of many random processes that occur in nature. In order to study the impact of noise on the classification of bearing fault, we generated different levels of noisy signals by adding AWGN to the raw signals. The generated noisy signals are measured by the signal-to-noise ratio (SNR) which is defined as the ratio of the power of a meaningful signal to the power of background noise [52]. The decibel form (dB) of SNR is expressed as:

$$SNR_{dB} = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right), \tag{10}$$

where P is the average power of a periodic signal $x(t)$ in period T, which is defined as:

$$P_{\text{avg}} = \frac{1}{T}\int_0^T p(t)\mathrm{d}t = \frac{1}{T}\int_0^T |x(t)|^2\mathrm{d}t. \tag{11}$$

When the signal is in discrete form, the average power can be calculated as:

$$P_{\text{avg}} = \frac{1}{length(x)}\sum_{t\in length(x)} |x(t)|^2. \tag{12}$$

For a signal with zero mean and known variance, its power can be expressed by the variance σ_N^2, so for standard normal distributed noise, the power of the signal is 1. Thus we first calculate the power of the original signal using (12), then calculate the power of the noise signal P_{noise} to be generated using (10) with the desired SNR. Finally, we generate the additive white Gaussian noise by the following formula, then add it to the original signal to compose the signal with the desired SNR. The standard normal distribution noise is generated by *randn*(), as denoted in

$$N = \sqrt{P_{\text{noise}}} \times randn(length(x)). \tag{13}$$

In order to demonstrate the noised signal, we generated a 0 dB SNR signal using the above method based on the signal of outer race fault from the CWRU dataset. The original signal of the outer race fault, the 0 dB SNR signal after adding Gaussian noise, and the corresponding spectra are shown in Fig.5. In the frequency domain on the right part of the figure, it can be seen that the noise spectrum is superimposed on the original spectrum. The whole spectrum from low frequency to high frequency is added with a uniform noise which is called the additive white Gaussian noise. In the time domain on the left part of the figure, it can be seen that the signal with added noise is more difficult to recognize in the time domain by a human. However, an algorithm needs to work successfully under different noise conditions due to the noisy environment. Thus, to simplify the difficulty, we are going to evaluate the proposed method under noisy conditions with different SNRs. We train the neural network on the training data without noise added, and evaluate the neural network with the noise-added test data.

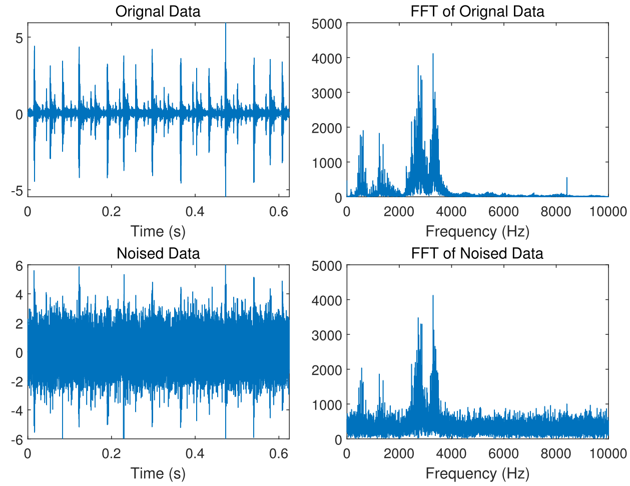


FIGURE 5. **Visualization of original and 0 dB SNR noise-added signals of outer race fault from the CWRU dataset.**

B. BASELINE ALGORITHMS

We compared the proposed algorithm with the following baseline algorithms in the experiment: 1) the TICNN with the changing random rate in dropout during training as described in the paper, which is consist of 6-layer convolutional networks; 2) the 5-layer convolutional networks named WDCNN; 3) the GRU neural networks which share the same parameters of the GRU structures in the proposed method; 4) and the support vector machine (SVM) with the Gaussian radial basis kernel. The input signal of all baseline methods is processed in the same way as the input of the proposed method, without further feature extraction procedure.

C. CASE STUDY I: CWRU DATA UNDER NOISY ENVIRONMENT

In the first case, we performed a benchmark experiment on the well-known CWRU dataset. Since there have been lots of studies on this dataset [53], it is reasonable and convenient to validate the proposed algorithm on it. The test rig of the CWRU dataset is shown in Fig.6. The vibration signal we used is collected from the drive end by the accelerometer with the sample rate of 48000 Hz. We constructed three datasets under loads of 1 hp, 2 hp, and 3 hp. As for the load at 3 hp,

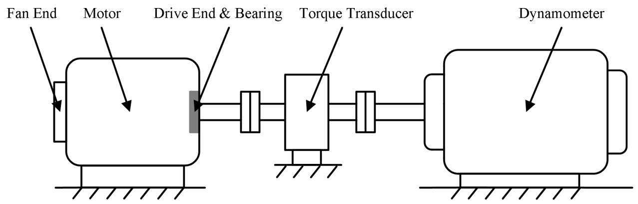


FIGURE 6. Test rig of the CWRU dataset.

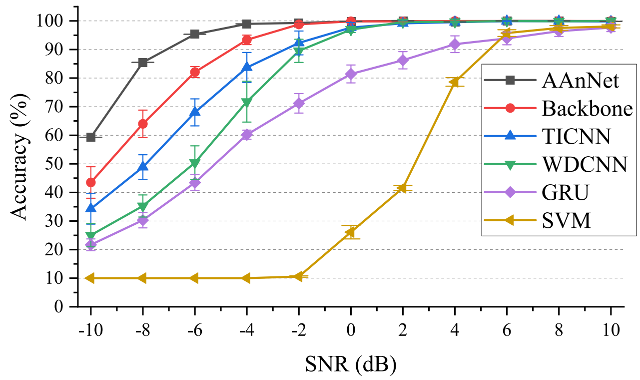


FIGURE 7. Comparison of different methods with different SNR values on the CWRU dataset.

the average motor speed is 1724 rpm is used in this paper. The bearing at the drive end is the deep groove ball bearing (6205-2RS JEM SKF).

The test bearings have four types of health conditions: (1) normal condition; (2) inner race fault; (3) ball fault; and (4) outer race fault. Each fault condition has three types of fault diameters with 7 mils, 14 mils, and 21 mils (1 mil = 0.001 inches), which are generated by electro-discharge machining. So, there are one normal condition and three fault types, each with three kinds of severity leads to 10 different classes in the dataset.

According to the preprocessing procedure aforementioned, the CWRU dataset has 10 different categories, and the length of each segment is 1670. Because we do not segment the signal with overlapping and there are fewer data in 1 hp condition, so the total number of samples is less than other loads. We randomly permutate the data and divide it into training data and test data with the ratio of 8:2. These two parts of the dataset are independent of each other. We only add noise to the test data in order to evaluate the performance of the proposed method under noisy conditions. There are 289 samples in 2 hp and 3 hp load conditions in each health condition, of which 231 are training samples, and the remaining 58 are test samples. As for the 1 hp condition, there are 181 training samples and 46 test samples in each health condition. All the samples have no overlapping parts with each other, nor other data augment methods. The 3 hp load data is used in the following experiments. Details are shown in Table 2.

1) DIAGNOSIS RESULTS AND ANALYSIS

As described before, we trained all the models on the original data without noise added, and test on the data that are added with different levels of noise. In order to study the contribution of each part in the framework to the final result, we divide the proposed AAnNet into two parts based on the network structure: (1) the backbone part and (2) the attention mechanism (AM) part. Note that these major parts have further subparts. The Backbone mainly consists of the random sampling part, the CNN part, and the GRU part. And the last time step of GRU is manually specified as output, then the Backbone is followed by a fully-connected layer that uses the Softmax as the activation function. The AM part includes not only the attention mechanism layer but also the following DNN part because there are additional fully-connected layers after the attention layer to help to generate the final output. Thus fully-connected layers are slightly different in Backbone and the proposed whole network framework (Backbone + AM). We will compare the effects of each part in the Ablation study section.

The comparison of the proposed method with different batch sizes under different SNRs is shown in Table 3. As seen from the table, when the SNR of signals is below −6 dB, the classification accuracy of AAnNet with batch size 128 is the highest. When the SNR is between −4 dB and 0 dB, the accuracy with batch size 64 is the highest. When the SNR is above 2 dB, the accuracy with all batch sizes is almost the same. For each batch size, the proposed AAnNet outperforms the Backbone, providing that the attention mechanism could improve the classification accuracy. As for the Backbone, when the batch size is 256, the classification accuracy is lowest compared with the batch size of 128 and 64, while the classification accuracy with the batch size of 128 is almost the same as with the batch size of 64. One explanation is that a larger batch size tends to converge to sharp minimizers in the training process, which leads to lower accuracy [54]. However, employing a relatively larger batch size would improve the training speed because the number of iteration in each epoch decreases, which can make full use of graphics card with large memory. We choose the batch size of 128 for the subsequent experiments as the trade-off between the accuracy and training time.

The comparison of different methods with different SNR values trained on the CWRU dataset is shown in Fig.7. As shown in the figure, the proposed AAnNet and the backbone part of the AAnNet achieve the highest accuracy among all other baseline methods. The accuracy of Backbone is up to 16.85 percentage higher than the baseline model TICNN, and the accuracy of AAnNet is up to 38.37 percent higher than baseline model TICNN with very low variance, which indicates that the proposed method significantly outperforms other baseline methods under high noise conditions.

The accuracy of SVM is 98.07% when the SNR is 10 but drops quickly when the noise increases. When the SNR is below −2 dB, the classification results of SVM are no longer valid. The accuracy of GRU without random sampling and

TABLE 2. Description of the CWRU dataset.

Health condition		Normal	Inner race fault			Ball fault			Outer race fault		
Fault size (in.)		0	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021
Class labels		1	2	3	4	5	6	7	8	9	10
No. of 1 hp load	Train	181	181	181	181	181	181	181	181	181	181
	Test	46	46	46	46	46	46	46	46	46	46
No. of 2 hp load	Train	231	231	231	231	231	231	231	231	231	231
	Test	58	58	58	58	58	58	58	58	58	58
No. of 3 hp load	Train	231	231	231	231	231	231	231	231	231	231
	Test	58	58	58	58	58	58	58	58	58	58

TABLE 3. Classification accuracy of the proposed AAnNet and the backbone part of the AAnNet with different batch sizes and SNR values.

Batch size	Models	SNR(dB)										
		-10	-8	-6	-4	-2	0	2	4	6	8	10
256	AAnNet	55.69 ± 0.00	72.93 ± 0.00	92.24 ± 0.00	98.28 ± 0.00	99.48 ± 0.00	99.83 ± 0.00	100.00 ± 0.00	99.83 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.48 ± 0.00
	Backbone	31.59 ± 1.70	46.38 ± 1.64	71.48 ± 1.56	88.14 ± 2.44	96.90 ± 0.90	99.48 ± 0.30	99.69 ± 0.22	99.86 ± 0.22	99.93 ± 0.15	99.96 ± 0.08	99.83 ± 0.32
128	AAnNet	$\mathbf{59.31 \pm 0.00}$	$\mathbf{85.52 \pm 0.00}$	$\mathbf{95.24 \pm 0.09}$	98.97 ± 0.00	99.41 ± 0.09	99.83 ± 0.00	$\mathbf{100.00 \pm 0.00}$	99.66 ± 0.00	$\mathbf{100.00 \pm 0.00}$	$\mathbf{100.00 \pm 0.00}$	99.83 ± 0.00
	Backbone	43.48 ± 5.10	64.00 ± 6.41	82.10 ± 2.62	93.38 ± 1.88	98.79 ± 0.34	99.86 ± 0.08	99.83 ± 0.00	99.96 ± 0.08	99.96 ± 0.08	99.96 ± 0.08	$\mathbf{99.96 \pm 0.08}$
64	AAnNet	51.55 ± 0.00	67.07 ± 0.00	93.55 ± 0.09	$\mathbf{99.31 \pm 0.00}$	$\mathbf{99.48 \pm 0.00}$	$\mathbf{100.00 \pm 0.00}$	100.00 ± 0.00	99.83 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.83 ± 0.00
	Backbone	42.59 ± 3.04	66.29 ± 5.95	82.59 ± 2.79	92.59 ± 3.22	99.05 ± 0.65	99.87 ± 0.12	99.83 ± 0.00	$\mathbf{100.00 \pm 0.00}$	100.00 ± 0.00	99.96 ± 0.08	99.91 ± 0.15

CNN part is the lowest among the neural network-based methods. In contrast, the performance of the Backbone with the random sampling and enhanced CNN feature extractor is better than the original GRU with the accuracy up to 38.66 percent higher. The random sampling and enhanced CNN feature extractor could automatically select the latent features from the raw signal without manual selection and improve the anti-noise performance. Then the processed features by the CNN part could be used more efficiently by the GRU with higher accuracy. Moreover, by employing the attention mechanism and the subsequent DNN layers, the classification accuracy of AAnNet is signification improved, and the accuracy variance is very small, the result remains highly consistent. This competent result proves that the proposed framework is effective in bearing fault diagnosis under high noise conditions.

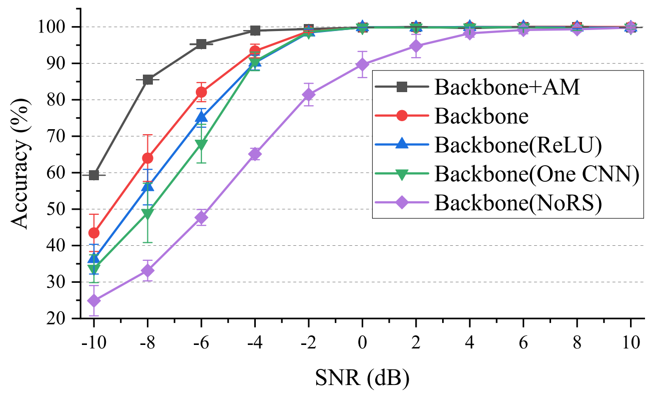


FIGURE 8. Comparison of different structures with different SNR values on the CWRU dataset.

2) ABLATION STUDY

In order to validate the enhancement schemes used in the proposed methods, we conducted experiments on several models with different structures to verify and select the optimal model. The comparison results are shown in Fig.8, where Backbone stands for the backbone part of the AAnNet as described before, Backbone(NoRS) stands for the backbone part without random sampling in the input layer, Backbone(ReLU) stands for the backbone part that uses ReLU as the activation function in the CNN layer, Backbone(One CNN) stands for the backbone part that only uses one CNN layer as the feature extraction part, Backbone+AM stands for the network with all parts, i.e., the random sampling part, the backbone part, and the attention mechanism part.

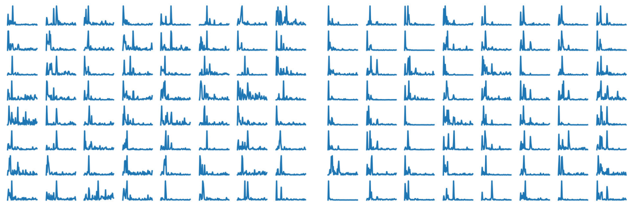
As shown in Fig.8, when there is no random sampling in the input layer, the classification accuracy drops quickly when the noise rises. The accuracy of the model decreases up to 35.38 percent compared to the backbone part. This proves that random sampling can effectively improve the noise-robust performance of the network. When there is only one CNN layer before the GRU part, the accuracy decreases up to 15.03 percentage points compared to the backbone part. It is worth noting that Backbone(One CNN) has nearly the same performance as the TICNN when the SNR is below -6 dB and outperforms the TICNN when the SNR is between -4 dB and 2 dB. When using the commonly used ReLU activation function in the feature extractor, i.e., the CNN part, the accuracy decreases up to 7.96 percentage points and has lower variance compared to the ELU version, i.e., the Backbone. This proves that ELU has better performance than the ReLU, especially under high noise conditions. When using the proposed AAnNet, the result achieves the highest accuracy, which is up to 21.52 percentage points compared to the backbone part and the lowest variance compared to other structures, providing that this framework is suitable for the bearing fault classification task, especially under high noise conditions.

3) NETWORKS VISUALIZATIONS

To better understand the intrinsic characteristics of the neural network, we demonstrate the kernel weight distribution of the proposed neural network. The feature extraction layers based on CNN play an important role in the proposed model, which can automatically extract features from the signal. Each convolution kernel in CNN is equivalent to a filter, and the weight of the convolution kernel is equivalent to the parameter of the filter. Therefore, figuring out the kernel weight distribution of a trained network is very important for understanding how the neural network works. Because the kernel size in the convolutional layer is 128×1, so we can visualize each kernel by plotting the kernel weight. To make it more clear to show how each kernel works, we demonstrate the kernel weight of the CNN part in frequency form, using the power spectrum. Note that the kernel weights are the internal parameters of the neural network, not the features extracted from the input.

As shown in Fig.9, each subgraph represents the kernel weight spectrum of a convolution kernel in the convolutional layer. Since the frequency value in each subgraph does not have actual meaning, we only show the curve in the graph and hide the coordinate axis to illustrate the frequency distribution differences among convolution kernels more clearly. Specifically, there are 64 convolution kernels in the first convolutional layer, and all of them are included in Fig.9(a), whereas only part of the kernels from the second convolutional layer are included in Fig.9(b).

It can be seen that the convolution kernel weight distribution corresponding to different convolutional layers is different. In the first layer, the filters pay more attention to multiple frequency bands, and each spectrum looks like the FFT spectrum of bearing fault signal, which composes of unique signal frequency and uniformly distributed noise spectrum, as shown in Fig.5. Therefore, the first layer can be seen as a transform from the time domain to the feature domain of the signal, somehow like the way FFT works. In the second layer, the spectrum is more sparse with less characteristic frequency. Each kernel is only interested in some particular frequency and can be regarded as a specific filter for the signal. Besides, the noise spectrum in the second layer is suppressed compared to the first layer, indicating that the effect of noise has been reduced in the second layer after the first layer's feature extraction. Therefore, the second convolutional layer can be regarded as a feature selector in the feature domain. By visualizing the kernel weight distribution of the convolutional layers, we can see that the CNN part has good effects on feature selection and noise suppression. Besides, the parameters of convolutional layers are automatically learned by the network, which does not require expert knowledge for complex filtering and feature selection design.



(a) Conv Layer 1 (b) Conv Layer 2

FIGURE 9. Visualization of kernel weights of 1st and 2nd convolutional layers by power spectrum. Each subgraph in (a) and (b) represents the power spectrum of the kernel weight in the respective convolutional layer.

To better understand how the neural networks process the raw signal, we visualize the process of bearing fault signals been coding and processing by the neural network via t-distributed stochastic neighbor embedding (t-SNE) [55]. The t-SNE is a common method for visualizing high-dimensional data. In this study, we first use the principal component analysis (PCA) to reduce the high-dimensional features into 100 dimensions to speed up the subsequent calculation. Then the t-SNE algorithm is adopted to map the 100 dimensions features to 2 dimensions to demonstrate the relationship between each feature. We use the bearing signals of -4 dB SNR as the input of the proposed network, then use t-SNE to visualize the output of the neural network layer by layer, as shown in Fig.10. There are 10 different classes of data named C1 to C10 in figure with different colors, respectively.

As shown in Fig.10(a), the raw signal is almost inseparable in the feature space. In contrast, the output features of each layer become more and more separable as the hierarchical goes deeper. This means that the proposed network has an excellent ability to extract useful features from the raw signal that could distinguish different types of bearing faults under noisy conditions. It is obvious that after processed by the GRU layers, the features are well separated, whereas just similar features are aggregated after the CNN layers. This proves that GRU works as the decoder or the classifier to do the feature classification task, whereas the CNN works as the encoder to extract and encode the features from the raw signals. Furthermore, by applying the attention mechanism and DNN after the GRU part, the output features become more separable and stable, as shown in Fig.10(e,h).

It is worth noting that similar classes are first aggregated and then separated, as shown in Fig.10(b,c,d,e). This indicates the network firstly puts inputs with similar characteristics together and then separates them step by step. Similar to human operation, it is a classification method from large to small, from whole to part, which could relatively reduce the difficulty of classification. By visualizing the workflow of the neural network layer by layer, it would be helpful to understand how the neural network works.

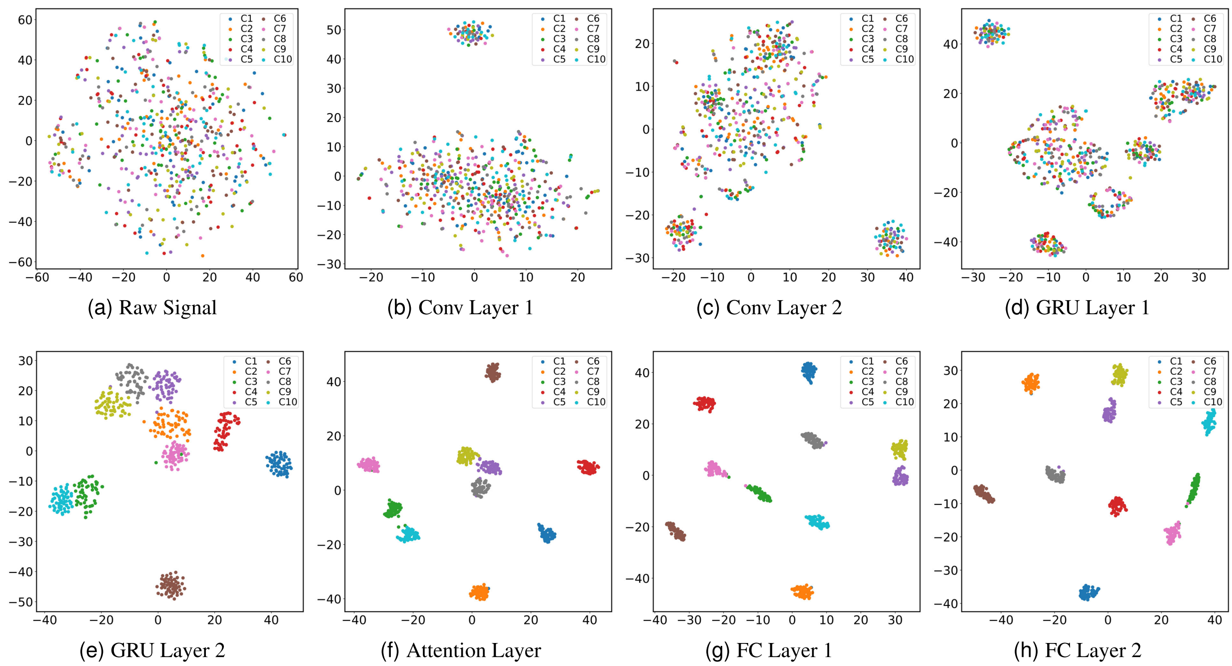


FIGURE 10. Feature visualization of neural network layer by layer via t-SNE at −4 dB SNR.

D. CASE STUDY II: CWRU DATA UNDER VARYING LOAD CONDITIONS

In order to verify the domain adaptive ability of the proposed method under varying load conditions, we trained each model with one load condition and tested with other load conditions. There are 3 load conditions in the CWRU dataset, which are 1 hp, 2 hp, and 3 hp. We did not add noise to the data nor using the manually selected features in this study. In the original CWRU dataset, there are fewer data in 1 hp condition than in other conditions, and we do not segment the signal with overlapping nor using other data augment methods. Thus the final data under 1 hp condition is less than in other conditions. The training data and test data are randomly selected with the ratio of 8:2. A detailed description of the data is shown in Table.2.

The comparison result is shown in Fig.11. We conducted 6 different experiments to cover all situations with domain adaptation under three loads conditions. The $1 \rightarrow 2$ stands for these models are trained on 1 hp load condition and tested on 2 hp load condition. The AVG stands for the average accuracy of each model. As shown in the figure, the SVM, whose average accuracy is 36.61%, has the worst performance among these methods, just like it in the noise case study. The baseline model TICNN, whose average accuracy is 52.87%, unlike it in the noise case study, performs worse than another baseline model WDCNN.

The proposed Backbone has the average accuracy of 61.68% that outperforms all other methods in every load domain adaptation scenarios providing the proposed method has good domain adaptive ability under varying load conditions. The Backbone with attention mechanism performs worse than Backbone, with an average accuracy

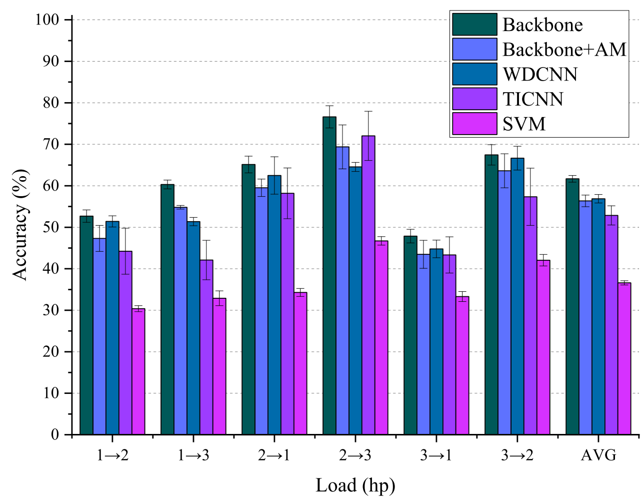


FIGURE 11. Comparison of different models with different loads on the CWRU dataset.

of 56.26%, even worse than the baseline model WDCNN with an average accuracy of 56.88%. It is interesting that Backbone+AM has the best performance against noise interference but performs worse in load domain adaptation. In the noise case study, because the proposed random sampling and the enhanced CNN could already have a strong enough anti-noise ability, then the intrinsic characteristics of the signal can be relatively well extracted from the raw signal, so the attention mechanism can better combine high-dimensional information and get a good classification accuracy. But in the load domain adaptation mission, it may be that after the attention mechanism has been trained under one load

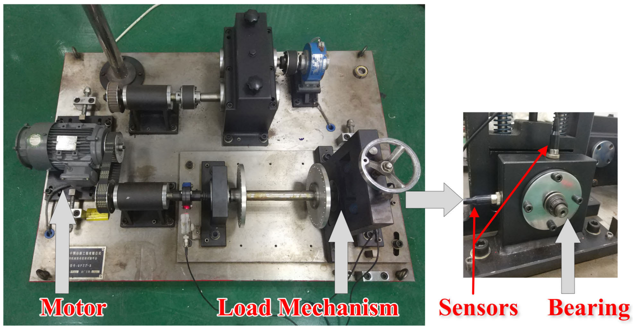


FIGURE 12. Test rig of the experiment.

TABLE 4. Specification of the test bearing.

Model	Number of rollers	Pitch diameter(mm)	Roller diameter(mm)	Contact angle(deg)
N205EM	12	38	8	0

condition, its internal attention weights are fixed, which may not be suitable for another load condition, thus degenerates the domain adaptation ability under varying load conditions. Although employing attention mechanism does not bring improvement, the proposed method still has the best domain adaptive ability than other baseline methods.

E. CASE STUDY III: EXPERIMENT DATA UNDER NOISY ENVIRONMENT

In order to verify the universality and generality of the proposed method, we conducted another bearing fault diagnosis experiment on the second dataset. The second dataset is from our bearing fault diagnosis experimental platform QPZZ-*II*. As shown in Fig.12, the test rig is mainly consisting of a motor, some couplings, and a load mechanism with the testing bearing. The vibration signal is collected by the accelerometer with the sample rate of 20000 Hz under loading conditions. The average motor speed is 1487 rpm. We use the signals collected from the accelerometer installed at 12 o'clock (directly in the load zone) in the following experiments. The test bearing installed inside the load mechanism at the end is the cylindrical roller bearing (N205EM HRB). The specification of the bearing is shown in Table 4. The test bearings have four types of health conditions: (1) normal condition; (2) ball fault; (3) inner race fault; and (4) outer race fault. The fault conditions of the test bearings are generated by electro-discharge machining. There are one normal condition and three types of fault conditions, totally four different types of data.

By using the preprocessing procedure aforementioned, we added different levels of AWGN to the signals collected from the experimental platform to verify the performance of the proposed algorithm, which is similar to the method of generating different levels of noise in the case study I. The outer race fault signal, the 0 dB SNR Gaussian-noise-added signal, and the corresponding spectra are shown in Fig.13. It can be seen that, similar to the CWRU data, the noise spectrum is superimposed on the original spectrum with uniform distribution. And the noised data is more difficult to recognize in the time domain by a human compared with the original data. The length of each segment is 807, according to the motor speed and the sampling rate of the accelerometer. Each segment has no overlapping with others. The data is randomly permuted and divided into training data and test data with the ratio of 8:2. There are 973 samples in each category, of which 778 are training samples, and the remaining 195 are test samples. Details are shown in Table 5. Same as the CWRU experiment, the neural network is trained with the training data without noise added and evaluated with the noise-added test data.

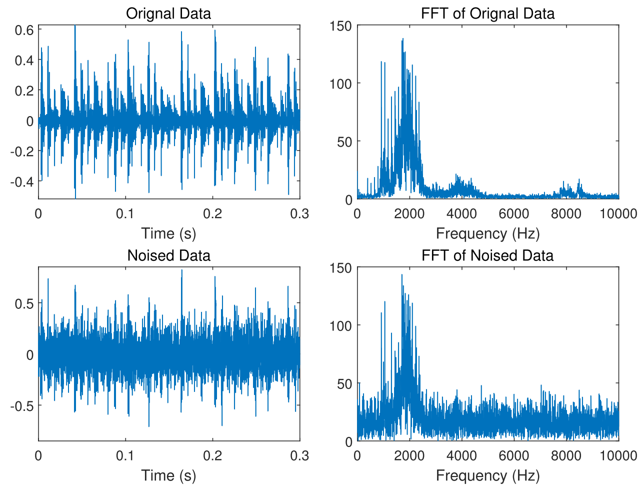


FIGURE 13. Visualization of original and 0 dB SNR noise-added signals of outer race fault from the experiment dataset.

TABLE 5. Description of the experiment bearing dataset.

Health condition		Normal	Inner race fault	Ball fault	Outer race fault
Class labels		1	2	3	4
Number of samples	Train	778	778	778	778
	Test	195	195	195	195

1) DIAGNOSIS RESULTS AND ANALYSIS

The classification result of the proposed method under different SNRs is shown in Fig.14, where Backbone+AM stands for the AAnNet with attention mechanism, Backbone stands for the backbone part of the AAnNet that without attention mechanism.

From the figure, we can see that the network-based methods maintain high generality under different levels of noise compared with the SVM which is no longer valid when the SNR value is below −4 dB. The accuracy of GRU is similar to the TICNN. The proposed method significantly outperforms other methods under high noise conditions with low variance. The accuracy of Backbone is up to 19.72 percent higher than the baseline TICNN. Notably, the Backbone outperformed the Backbone with the attention mechanism that achieves the

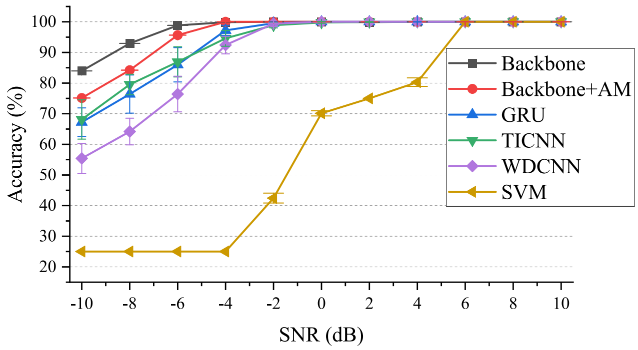


FIGURE 14. Comparison of different methods with different SNR values on the experiment dataset.

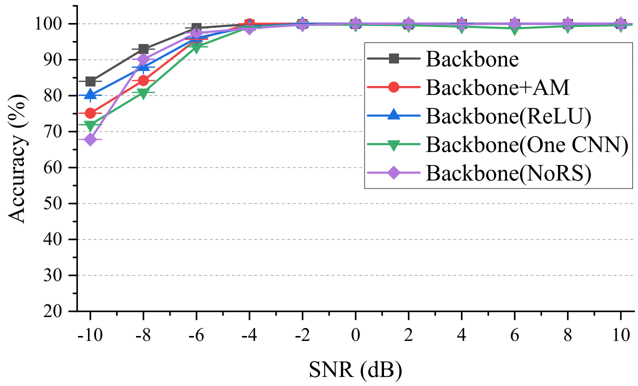


FIGURE 15. Comparison of different structures with different SNR values on the experiment dataset.

highest score in the CWRU dataset. We will further discuss the attention mechanism in Section IV-E.3.

2) ABLATION STUDY

We also conducted experiments with different structures to verify the proposed schemes on the experiment dataset. The comparison results are shown in Fig.15, where Backbone stands for the backbone part of the AAnNet as described before, Backbone+AM stands for the network with all parts, i.e., the random sampling part, the backbone part, and the attention mechanism part. Backbone(ReLU) stands for the backbone part that uses ReLU as the activation function in the CNN layer, Backbone(One CNN) stands for the backbone part that only uses one CNN layer as the feature extraction part, Backbone(NoRS) stands for the backbone part without random sampling in the input layer.

As shown in Fig.15, when there is no random sampling in the input layer, the classification accuracy begins to decrease quickly when the level of noise rises; and has the lowest accuracy when SNR is −10 dB. This proves that random sampling can effectively improve the noise-robust performance of the network. When there is only one CNN layer before the GRU part, the accuracy decreases up to 12.05 percentage points compared to the backbone part. The performance of this model is the lowest when SNR above −8 dB. When using the commonly used ReLU activation function in the feature extractor, the accuracy decreases up to 3.85 percentage points compared to the ELU version, providing that ELU is better

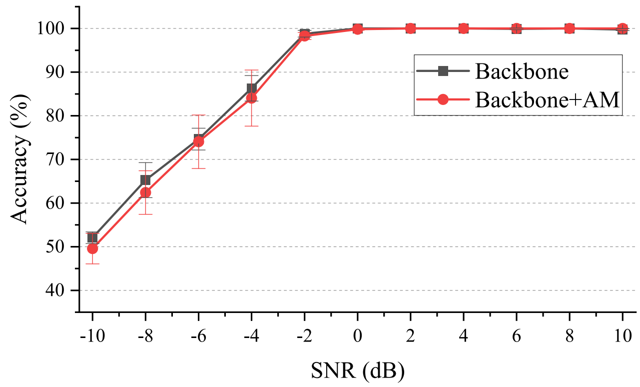


FIGURE 16. Comparison of the proposed method with or without attention mechanism with different SNR values on the 4 class CWRU dataset.

than ReLU in the bearing fault classification. When using the Backbone with the attention mechanism, the result accuracy decreases up to 8.85 percentage points compared to the Backbone.

3) DISCUSSION ON THE ATTENTION MECHANISM

Although the proposed model with the attention mechanism has achieved excellent results in the CWRU experiment under noise conditions, it does not perform well on the experimental dataset. One possibility is that the attention mechanism works less efficiently when the number of classes is small. There are 10 classes in the CWRU dataset, while there are only 4 classes in the experiment dataset.

To verify the conjecture, we conduct another experiment that only including different fault types of data from the CWRU dataset, totally 4 classes. This verification dataset includes the (1) normal condition; (2) inner race fault; (3) ball fault; and (4) outer race fault. To be specific, each fault type only contains data with fault diameters of 7 mils. There are 289 samples in each category, of which 231 are training data, and the remaining 58 are test samples. The classification results of attention mechanism based method and the method without attention mechanism are shown in Fig.16.

As shown in Fig.16, when there are only 4 classes from the CWRU dataset, the accuracy of Backbone is up to 2.84 percent higher compared with the Backbone+AM, which can prove the conjecture described above. The attention mechanism works less efficiently when there are fewer classes. The classification difficulty decreases when the number of classes is small, which would cause overfitting that makes the attention mechanism less efficient. It would be harder to calculate the appropriate weights in the attention weight because there are less data when there are fewer classes. Therefore, the appropriate network structure should be chosen according to the actual task and the complexity of the task, in order to make full use of the neural network. In the bearing fault diagnosis, it may be less efficient using the attention mechanism combined with GRU in the load domain adaptation tasks and the noise conditions when the number of the categories is small.

V. CONCLUSION

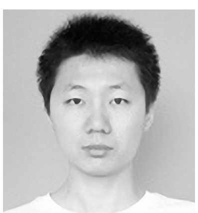
In this paper, we proposed an end-to-end neural network framework that combined the advantages of different structures of neural networks to classify the bearing fault signals under heavy noise and varying load conditions. The experiments on the CWRU dataset and our experiment dataset proved that the proposed neural network framework achieved state-of-the-art results under heavy noise or varying load conditions. Furthermore, we have found through experiments that the attention mechanism combined with GRU is more suitable when the number of categories is large, whereas less improvement is observed when the number of categories is small or in the load domain adaption task. Additionally, we explored the intrinsic characteristics of the neural network by visualizing the kernel weight distribution and activation values of the proposed method, which helps to understand how neural networks work.

In future work, we will further study the characteristics of the attention mechanism found in this paper. Since domain adaptation and transfer learning are promising ways to diagnose bearing fault, we will focus on theses methods to handle more difficult tasks like noise and varying load conditions occurring at the same time as well as the imbalanced training data problem.

REFERENCES

[1] N. Tandon and A. Choudhury, "A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings," *Tribology Int.*, vol. 32, no. 8, pp. 469–480, Aug. 1999.
[2] A. Rai and S. H. Upadhyay, "A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings," *Tribology Int.*, vol. 96, pp. 289–306, Apr. 2016.
[3] M. Cerrada, R.-V. Sánchez, C. Li, F. Pacheco, D. Cabrera, J. V. de Oliveira, and R. E. Vásquez, "A review on data-driven fault severity assessment in rolling bearings," *Mech. Syst. Signal Process.*, vol. 99, pp. 169–196, Jan. 2018.
[4] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[6] 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.
[7] D.-T. Hoang and H.-J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, Mar. 2019.
[8] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Machine learning and deep learning algorithms for bearing fault diagnostics - a comprehensive review," 2019, *arXiv:1901.08247*. [Online]. Available: http://arxiv.org/abs/1901.08247
[9] N. F. Waziralilah, A. Abu, M. Lim, L. K. Quen, and A. Elfakharany, "A review on convolutional neural network in bearing fault diagnosis," in *Proc. MATEC Web Conf.*, Les Ulis, France: EDP Sciences, vol. 255, 2019, Art. no. 06002.
[10] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078
[11] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[12] W. Abed, S. Sharma, R. Sutton, and A. Motwani, "A robust bearing fault detection and diagnosis technique for brushless DC motors under non-stationary operating conditions," *J. Control, Autom. Electr. Syst.*, vol. 26, no. 3, pp. 241–254, Jun. 2015.
[13] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vibrat.*, vol. 377, pp. 331–345, Sep. 2016.
[14] M. Bhadane and K. I. Ramachandran, "Bearing fault identification and classification with convolutional neural network," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Apr. 2017, pp. 1–5.
[15] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 303–315, May 2016.
[16] Y. Xie and T. Zhang, "Feature extraction based on DWT and CNN for rotating machinery fault diagnosis," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 3861–3866.
[17] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 8, pp. 1926–1935, Aug. 2017.
[18] H. Jiang, X. Li, H. Shao, and K. Zhao, "Intelligent fault diagnosis of rolling bearings using an improved deep recurrent neural network," *Meas. Sci. Technol.*, vol. 29, no. 6, Jun. 2018, Art. no. 065107.
[19] X. Li, W. Zhang, and Q. Ding, "Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5525–5534, Jul. 2019.
[20] W. Zhang, F. Zhang, W. Chen, Y. Jiang, and D. Song, "Fault state recognition of rolling bearing based fully convolutional network," *Comput. Sci. Eng.*, vol. 21, no. 5, pp. 55–63, Sep. 2019.
[21] W. Qian, S. Li, J. Wang, Z. An, and X. Jiang, "An intelligent fault diagnosis framework for raw vibration signals: Adaptive overlapping convolutional neural network," *Meas. Sci. Technol.*, vol. 29, no. 9, Sep. 2018, Art. no. 095009.
[22] X. Li, W. Zhang, Q. Ding, and J.-Q. Sun, "Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation," *J. Intell. Manuf.*, vol. 31, no. 2, pp. 433–452, Feb. 2020.
[23] H. Pan, X. He, S. Tang, and F. Meng, "An improved bearing fault diagnosis method using one-dimensional CNN and LSTM," *Strojniski Vestn.-J. Mech. Eng.*, vol. 64, pp. 443–452, May 2018.
[24] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.
[25] L. Yu, J. Qu, F. Gao, and Y. Tian, "A novel hierarchical algorithm for bearing fault diagnosis based on stacked LSTM," *Shock Vibrat.*, vol. 2019, pp. 1–10, Jan. 2019.
[26] C. Lu, Z. Wang, and B. Zhou, "Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification," *Adv. Eng. Informat.*, vol. 32, pp. 139–151, Apr. 2017.
[27] X. Li, W. Zhang, and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77–95, Oct. 2018.
[28] H. Liu, J. Zhou, Y. Zheng, W. Jiang, and Y. Zhang, "Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders," *ISA Trans.*, vol. 77, pp. 167–178, Jun. 2018.
[29] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Trans.*, vol. 95, pp. 295–305, Dec. 2019.
[30] H. Qiao, T. Wang, P. Wang, L. Zhang, and M. Xu, "An adaptive weighted multiscale convolutional neural network for rotating machinery fault diagnosis under variable operating conditions," *IEEE Access*, vol. 7, pp. 118954–118964, 2019.
[31] X. Li, W. Zhang, Q. Ding, and X. Li, "Diagnosing rotating machines with weakly supervised data using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1688–1697, Mar. 2020.
[32] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017.
[33] G. Jin, D. Li, Y. Wei, W. Hou, H. Chen, Y. Jin, and C. Zhu, "Bearing fault diagnosis using structure optimized deep convolutional neural network under noisy environment," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 630, Oct. 2019, Art. no. 012018.
[34] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.

[35] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, nos. 5–6, pp. 555–559, Jun. 2003.
[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[38] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289
[39] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[40] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013, p. 3.
[41] C. Olah. *Understanding LSTM Networks*. Accessed: Oct. 23, 2018. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
[42] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2013, pp. 1310–1318.
[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 2067–2075.
[44] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*. [Online]. Available: http://arxiv.org/abs/1508.04025
[45] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal Process.*, vol. 161, pp. 136–154, Aug. 2019.
[46] B. Efron, "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 569–593.
[47] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[49] X. Lou and K. A. Loparo, "Bearing fault diagnosis based on wavelet transform and fuzzy inference," *Mech. Syst. Signal Process.*, vol. 18, no. 5, pp. 1077–1095, Sep. 2004.
[50] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io
[51] P. A. Frost, "Estimation in continuous-time nonlinear systems," Ph.D. dissertation, Dept. Elec. Engrg., Stanford Univ., Stanford, CA, USA, 1968.
[52] D. H. Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, p. 2088, 2006.
[53] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.
[54] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," 2016, *arXiv:1609.04836*. [Online]. Available: http://arxiv.org/abs/1609.04836
[55] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.



GUOQIANG JIN was born in Lanzhou, Gansu, China, in 1992. He received the B.S. degree in mechanical engineering from the University of Science and Technology of China, Hefei, China, in 2015, where he is currently pursuing the Ph.D. degree in precision machinery and instrumentation.

His research interests include intelligent fault diagnosis, signal processing, intelligent manufacturing, and deep learning.

TIANYI ZHU was born in Hefei, Anhui, China, in 1993. She received the B.S. degree in mechanical engineering from the Hefei University of Technology, Hefei, China, in 2015, and the M.S. degree in mechanical engineering from the University of Science and Technology of China, Hefei, in 2018.

Her current research interests include intelligent fault diagnosis, intelligent manufacturing, signal processing, data mining, and deep learning.

MUHAMMAD WAQAR AKRAM received the M.S. degree from the University of Agriculture Faisalabad, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in precision machinery and instrumentation with the University of Science and Technology of China, Hefei, China.

His research interests include intelligent manufacturing, solar energy, farm machinery, and deep learning.

YI JIN (Member, IEEE) was born in Hefei, Anhui, China, in 1984. He received the B.S. degree from Jiangnan University, Wuxi, China, in 2008, and the Ph.D. degree in mechanical engineering from the University of Science and Technology of China, Hefei, China, in 2013.

He is currently an Associate Professor with the University of Science and Technology of China. His current research interests include intelligent manufacturing, signal processing, pattern recognition, and deep learning.

CHANGAN ZHU was born in Wuhu, Anhui, China, in 1957. He received the B.S. degree from the Hefei University of Technology, Hefei, China, in 1982, the M.S. degree from Xidian University, Xi'an, China, in 1985, and the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 1989.

He is currently a Professor with the School of Engineering Science, University of Science and Technology of China. His research interests include intelligent manufacturing, signal processing, control theory, and deep learning.