IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 13, 2020, accepted November 24, 2020, date of publication December 1, 2020,
date of current version December 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041735

Enhanced Lightweight Multiscale Convolutional Neural Network for Rolling Bearing Fault Diagnosis

YAOWEI SHI, AIDONG DENG, MINQIANG DENG, JING ZHU, YANG LIU, AND QIANG CHENG

National Engineering Research Center of Turbo-Generator Vibration, School of Energy and Environment, Southeast University, Nanjing 210096, China

Corresponding author: Aidong Deng (dnh@seu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51875100.

ABSTRACT The vibration signals collected from rolling bearings in industrial systems are highly complex and contain intense environmental noise, which challenges the performance of traditional fault diagnosis methods. Moreover, the applicability of the model in engineering practice, especially in the Industrial Internet of Things context, puts forward higher requirements for its storage and computational costs. Considering these challenges, this article proposes an enhanced lightweight multiscale convolutional neural network (CNN) for rolling bearing fault diagnosis. Our contributions mainly fall into three aspects. Firstly, the proposed model is modular and easy to expand, which combines the idea of multiscale learning with attention mechanism and residual learning, enabling the network to extract more abundant and discriminative fault features directly from the raw vibration signal. Consequently, the proposed model can perform better. Secondly, the interpretability of the multiscale learning mechanism is explored by visualizing the extraction process of multiscale features. Finally, for the first time, we introduce the depthwise separable convolution into multiscale CNN to reduce the storage and computational costs of the model, which realizes the lightweight of the model and improves its applicability in the Industrial Internet of Things context. The experimental results on the rolling bearing dataset demonstrate that, compared with the state-of-the-art multiscale CNN models, the proposed model has better discriminative fault feature extraction ability and anti-noise ability, and is more suitable for practical industrial systems.

INDEX TERMS Rolling bearing fault diagnosis, multiscale convolutional neural network, attention mechanism, residual learning, lightweight.

I. INTRODUCTION

Rolling bearings are one of the most common components in rotating machines, and their health conditions are strictly related to the safe and stable operation of mechanical equipment [1], [2]. However, under the complex conditions of high speed and high workload for a long time, rolling bearings are prone to occur wear, spalling, or other faults, which quickly lead to performance degradation of the equipment and even cause significant economic losses and severe casualties [3]. Therefore, it is of tremendous realistic significance to estimate the health conditions of rolling bearings and reduce the risk of unplanned shutdowns.

In the past years, a great deal of research on fault diagnosis of rolling bearings based on machine learning has been prompted and achieved good results to some extent. However, there are still some apparent drawbacks, such as being difficult to be optimized as a whole and relying heavily on specific domain knowledge and expert experience. Deep learning, as an end-to-end method has achieved a series of breakthroughs in the field of fault diagnosis, providing a powerful solution to the above drawbacks [4]. Different deep learning methods such as the recurrent neural network (RNN) [5], [6], convolutional neural network (CNN) [7], [8], deep belief network (DBN) [9], [10], and autoencoder [11], [12] have been widely used in the fault diagnosis of rolling bearings and achieved high diagnostic accuracy. In particular, CNN has achieved remarkable success in fault diagnosis due to its

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

powerful ability of automatic feature extraction and classification. Some researchers [13]–[15] convert one-dimensional (1D) signals into two-dimensional (2D) images, and then feed them into 2DCNN to achieve diagnostic results. Recently, compared with 2DCNN, 1DCNN is considered more suitable for fault diagnosis due to its attractive characteristics, such as uncomplicated and extracting fault features directly from the collected vibration signals without worrying about the loss of useful information caused by the 1D-to-2D conversion process. Hao *et al.* [8] combined 1DCNN and long short-term memory (LSTM) network to extract spatial and temporal features for more effective bearing fault diagnosis. Reduce the computational complexity by using 1DCNN in front of the LSTM layer. Peng *et al.* [16] proposed a 1D residual block by introducing the idea of residual learning into the traditional 1DCNN for fault diagnosis of wheelset bearings, effectively solved the problem of performance degradation for the deeper CNN. Zhang *et al.* [17] proposed a novel 1DCNN model with wide first-layer kernels for the fault diagnosis of rolling bearing. In the first convolutional layer of the model, wide kernels were used to extract features and suppress high-frequency noise.

The attention mechanism has been proposed as a kind of contribution screening of information and successfully applied to a variety of research tasks such as document classification [18], handwriting synthesis [19], and image caption generation [20], and so on. It can assist the neural network to focus more on the task-related information and ignore the information that contributes less to the task [21]–[23]. In this way, the learning ability and interpretability of neural networks can be improved using the attention mechanism. In the fault diagnosis field, attention-based neural networks are getting more and more attention. Wang *et al.* [24] adopted the multi-head attention mechanism to optimize the CNN structure and developed a convolutional network model for intelligent bearing fault diagnosis. Li *et al.* [23] constructed a deep learning model for rolling bearing fault diagnosis, which combines the convolutional network and LSTM. The attention mechanism was introduced to assist the model in locating the important features and visualizing the learned diagnosis knowledge.

Recently, with the rapid development and maturity of the Industrial Internet of Things technology, the concept of lightweight has arisen from the need for models with lower storage and computational costs in actual applications [25]. Liu *et al.* [26] proposed a lightweight MT-1DCNN for exploring the possibility of using auxiliary tasks to improve the performance of the fault diagnosis task. To reduce the proposed model's complexity, the number and size of convolution kernels are reasonably reduced. Yao *et al.* [25] proposed a SIRCNN for bearing fault diagnosis. The depthwise separable convolution and inverted residual structure were adopted to ensure the accuracy of the model in noisy environments while achieving lightweight.

In the real industrial systems, rolling bearings usually work under complex operation conditions, such as variable loads and speeds. Hence, the vibration signals measured from the bearing are nonlinear and nonstationary with strong coupling and contain intense environmental noise. On the other hand, the impact segment containing pulse components in vibration signals reflect the fault behavior of rolling bearings. The frequency of the pulse components caused by faults of different positions and severity levels varies greatly, which result in the features that sensitive to different faults are distributed on different time scales of vibration signals [27], [28]. Therefore, vibration signals usually exhibit multiscale characteristics and contain intricate patterns on multiple time scales [29]–[31]. However, the traditional CNN architecture cannot capture multiscale features of vibration signals due to its single scale characteristics. Thus, it is difficult to apply it to the diagnosis task of rolling bearings. To overcome the limitation, researchers have proposed some methods to introduce multiscale learning into CNN. Jiang *et al.* [32] proposed an MSCNN architecture. It performs multiple downsampling and smoothing operations on raw vibration signals in parallel to obtain multiscale signals. Then convolution and pooling operations were used to extract features of different scales. The results demonstrate that the MSCNN has better fault feature extraction ability and robustness than single scale CNN, and as more scales are incorporated, it can achieve better and more reliable performance. However, due to the defects of its multiscale signals acquisition method, the number of scales is constrained by the length of the input sample. In [33], an MK-ResCNN architecture was proposed, which provides a solution to the above problem. In MK-ResCNN, convolutional kernels with different sizes were utilized to extract multiscale features from vibration signals in parallel, and identity mapping and residual mapping were introduced to overcome the degradation problem caused by deep networks. In [27], a multiscale feature extraction method similar to the MK-ResCNN was adopted. Besides, an adaptive weight vector was introduced to emphasize the scale feature sensitive to faults.

Although aforementioned CNN-based multiscale learning methods make up for the defects of feature extraction ability of traditional signal scale CNN methods and achieved good diagnosis performance, there are still the following challenges. 1) Lacking reinforcement mechanism for discriminative fault features. As mentioned above, the impact segment in vibration signals carries the fault information of rolling bearings and has certain uniqueness for different faults; thus, it can be understood as the ''ID'' of the faults. However, due to the weight sharing strategy of convolutional kernels, more attention cannot be paid to these fault-related signal segments. 2) Poor interpretability. Although the multiscale CNN has been demonstrated to have higher diagnostic accuracy and stronger anti-noise ability than traditional single scale CNN, its learning mechanism is still a ''black box'' that will significantly restrict its further development in the field of fault diagnosis. 3) Limited by storage and computational costs. Some studies have shown that more scales and the wide convolutional kernel can usually improve the diagnosis

performance of the model [23]–[32], but also increase storage and computational costs of the model due to more parameters and more complex structure. Furthermore, the advancement of the Industrial Internet of Things puts forward the requirement of ''small, light, and fast'' for deep learning models to improve their applicability [34], [35]. Therefore, existing multiscale models have to make a trade-off between diagnosis performance and the cost of storage and computation. As a result, they usually use fewer scales (four scales were used in [27], [32], [36], and three scales were used in [33]), or stack multiple convolutional layers with narrow kernels instead of a layer with wide kernel [33].

To overcome the abovementioned challenges, we conducted a more in-depth study and exploration on the application of multiscale CNN in the field of fault diagnosis in terms of overall network structure, discriminative feature extraction ability, interpretability, and applicability. This article proposes an enhanced lightweight multiscale feature extraction module (ELMFEM) and develop a stacked ELMFEMs (S-ELMFEMs) model for rolling bearing fault diagnosis. The main contributions of this article are summarized as follows.

1) This paper proposed a stackable ELMFEM for rolling bearing fault diagnosis. Through multiscale and the overall easy-to-stack structural design, it can extract abundant and complementary features in raw vibration signals from multiple time scales while overcoming the defects of the complex structure of existing multiscale CNNs. Stacking multiple ELMFEMs will cause the network depth to increase rapidly. To avoid gradient explosion/vanishing caused by the above situation, residual connection technique is introduced.

2) The multiscale CNN is still a ''black box''. As far as we know, there is no research to show its work mechanism intuitively. However, interpretability is very important to control this technology fully. Therefore, we introduce the attention mechanism into the ELMFEM. On the one hand, by visualizing the attention weight, the multiscale CNN's working mechanism is displayed more intuitively to explore its interpretability. On the other hand, the attention mechanism makes up for the shortcomings of convolution kernel weight sharing and improves the ability of discriminative fault feature extraction.

3) More scale branches will inevitably bring more parameters, which will increase the model's computational and storage costs. To meet the requirements of ''small, light, and fast'' proposed by the Industrial Internet of Things, the depthwise separable convolution (DSC) technique was introduced into the ELMFEM. In this way, the lightweight of the model is realized. Besides, the existence of residual connection and the local use of DSC effectively solve the representational bottleneck problem of the features and lower accuracy caused by the characteristics of DSC.

The rest of this article is organized as follows. In Section 2, the proposed ELMFEM and S-ELMFEMs fault diagnosis model are described in detail. Data description and experimental verification setup are illustrated in Section 3. In Section 4, the effectiveness and superiority of the proposed method are verified. Finally, Section 5 concludes the whole article.

II. PROPOSED METHOD

As mentioned above, the fault features of vibration signals measured from the rolling bearing under different operation conditions and intense noise have the characteristics of time scale diversity and unobvious, which challenges the fault features extraction ability of single scale CNN. Therefore, some researchers combined multiscale learning with CNN to achieve better diagnosis performance. However, for multiscale CNN, insufficient attention has been paid to how to effectively improve the extraction ability of discriminative fault features, the interpretability of learning mechanism, and the applicability in the Industrial Internet of Things context. Given these, this article proposes a stackable ELMFEM to incorporate attention mechanism, residual learning, and lightweight convolution to the multiscale CNN. In this way, the proposed diagnosis model (S-ELMFEMs) constructed by the ELMFEM has the characteristics of high diagnostic accuracy, strong anti-noise ability, and lightweight, and its learning mechanism is also more interpretable. The details of the proposed model are elaborated in the following subsections.

A. 1D DEPTHWISE SEPARABLE CONVOLUTION

Depthwise separable convolution (DSC) is a typical lightweight convolution, which factorizes standard convolution operations into depthwise convolution and pointwise convolution [38]. In short, depthwise convolution applies filters, i.e., convolutional kernels, to each channel of the input layer by layer, and then pointwise convolution aggregate the output of depthwise convolution with 1×1 convolutional kernel. Experimental results in Xception [39] and MobileNet [40] have proved that DSC can be used to CNN on a large scale to reduce parameters and computational cost. In this article, the 1D depthwise sparable convolution (1DDSC) is used. The schematic diagram of standard 1D convolution and 1DDSC are shown in Fig. 1. Ignoring the influence of bias, the parameters and computation amount of standard 1D convolution and 1DDSC are shown in Table 1.

where L_{F} is the length of the input, L_{K} is the size of the convolutional kernel, M is the number of input channels, N is the number of convolutional kernels. The ratio of the computation amount of the 1DDSC and the standard 1D convolution is:

$$\frac{L_{\mathrm{K}} \times M \times L_{\mathrm{F}} + M \times N \times L_F}{L_K \times M \times N \times L_F} = \frac{1}{N} + \frac{1}{L_K} = \frac{N + L_K}{N \times L_K} \quad (1)$$

As can be seen from (1), the 1DDSC significantly reduces the computational cost, and when N is fixed, the wider the convolutional kernel size, i.e., the larger L_{K}, the higher the reduction in calculation. It should be pointed out that the 1DDSC is only used in the multiscale feature extraction stage in this article.

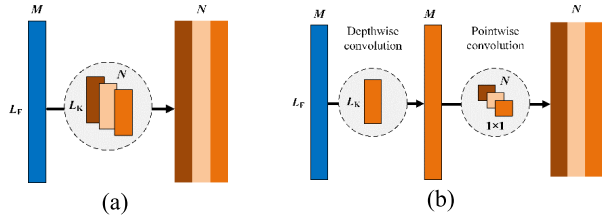


FIGURE 1. The schematic diagram of (a) standard 1D convolution and (b) 1DDSC.

TABLE 1. Parameters and calculation amount of standard 1D convolution and 1DDSC.

	Parameters	Computation amount
Standard 1D convolution	$L_k \times M \times N$	$L_k \times M \times N \times L_F$
1DDSC	$L_k \times M + 1 \times M \times N$	$L_k \times M \times L_F + M \times N \times L_F$

B. DISCRIMINATIVE FEATURE REINFORCEMENT MECHANISM (DFRM)

Since the impact segments in vibration signals reflect the fault behavior of the rolling bearing, more attention should be paid to it to highlight discriminative fault information, thereby improving the pertinence and reliability of feature extraction and the interpretability of multiscale learning mechanisms.

The basic structure of the DFRM is shown in Fig. 2. It is actually the attention mechanism. The shape of the input feature $\boldsymbol{O}'$ is $(T \times F)$, where T denotes the length of $\boldsymbol{O}'$ and F denotes the number of channels of $\boldsymbol{O}'$. The feature $\boldsymbol{Y}$ is obtained through an s $\times 1$ 1DDSC layer with channel number F/r, where r is the dimension reduction factor to simplify the model and accelerate training. Whereafter, the feature information of all the activation maps across channels in $\boldsymbol{Y}$ is projected and compressed on the temporal signal through a 1×1 1D convolutional layer with 1 channel and the Sigmoid activation function, and then the optimization vector $\boldsymbol{V}$ is obtained. Specifically, assume that the input $\boldsymbol{O}'$ represented as $[o^1, o^2, \ldots, o^T]$, where $o^i \in \mathbb{R}^{1\times F}$ and $i = 1, 2, \ldots, T$. After the above transformation, i.e., (2), $\boldsymbol{V} = [v^1, v^2, \ldots, v^T]$ is obtained, where v^i corresponds to the ith element in the optimization vector $\boldsymbol{V}$ and the value range is [0, 1]. The value of v^i indicates the importance of the ith time-series point.

$$\boldsymbol{V} = \sigma\left(W_2 f\left(W_1 \boldsymbol{V} + b_1\right) + b_2\right) \tag{2}$$

The obtained optimization vector $\boldsymbol{V}$ can be dynamically followed and adjusted according to the input signal's change. The values of v^i in the corresponding positions of the discriminative fault features are large, and the values of v^i in the corresponding positions of the irrelevant features are small. Finally, through (3), the discriminative fault features in $\boldsymbol{Y}$ are reinforced, the irrelevant features are weakened, and then the optimized feature $\boldsymbol{Y}'$ is obtained.

$$\boldsymbol{Y}' = [y'^1, y'^2, \ldots, y'^T] = \boldsymbol{Y} \cdot \boldsymbol{V} \tag{3}$$

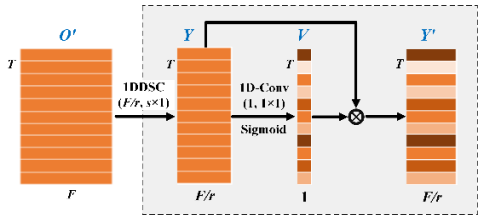


FIGURE 2. The basic structure of the DFRM.

where '.' denotes element-wise multiplication operation. Therefore, the optimized feature $\boldsymbol{Y}'$ has more discriminative fault information and interpretable learning mechanism.

C. ENHANCED LIGHTWEIGHT MULTISCALE FEATURE EXTRACTION MODULE (ELMFEM)

The structure of the ELMFEM is graphically illustrated in Fig. 3. The goal of the EMFEM is to extract complementary multiscale fault features from vibration signals in a parallel manner, so multiple convolutional layers with different kernel sizes are used to form multiple scale branches (referred to as multiscale 1DDSC in Fig. 3). The original feature set $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, .., \boldsymbol{Y}_S\}$ is obtained though S parallel 1DDSC layers with different kernel sizes, where $\boldsymbol{Y}_j \in \mathbb{R}^{\mathrm{T}\times(F/r)}(j = 1, 2, \ldots, S)$ and S denotes the number of scales. In order to extract detailed information and keep the feature $\boldsymbol{Y}_j$ in the same shape during the above operation, the stride of all 1DDSC layers is set to 1. The padding strategy is adopted to make $\boldsymbol{Y}_j$ (j $= 1, 2, \ldots, S$) have the same length. Then the original feature set $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, .., \boldsymbol{Y}_S\}$ is reconstructed by DFRM to obtain the optimized feature set $\{\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, \ldots, \boldsymbol{Y}'_S\}$, and $\boldsymbol{Y}'_j$ is concatenated along the channel dimension to form a multiscale feature map $\boldsymbol{Y}_\mathbf{m}$, which is denoted as

$$\boldsymbol{Y}_m = \left[\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, \ldots, \boldsymbol{Y}'_S\right] \tag{4}$$

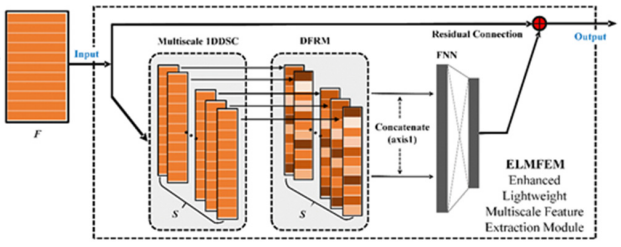


FIGURE 3. Framework of the proposed ELMFEM.

It contains all the discriminative fault information of different time scales. Furthermore, $\boldsymbol{Y}_\mathbf{m}$ is fed into the feedforward neural network (FNN) containing two fully connected layers (number of neurons: 256 and F) to integrate hidden layer features, thereby further extracting useful features and removing redundant features. It is worth mentioning that the number of neurons in the second fully connected layer is set to F so that the output of the FNN has the same shape as the input of the ELMFEM. Finally, the residual connection is introduced

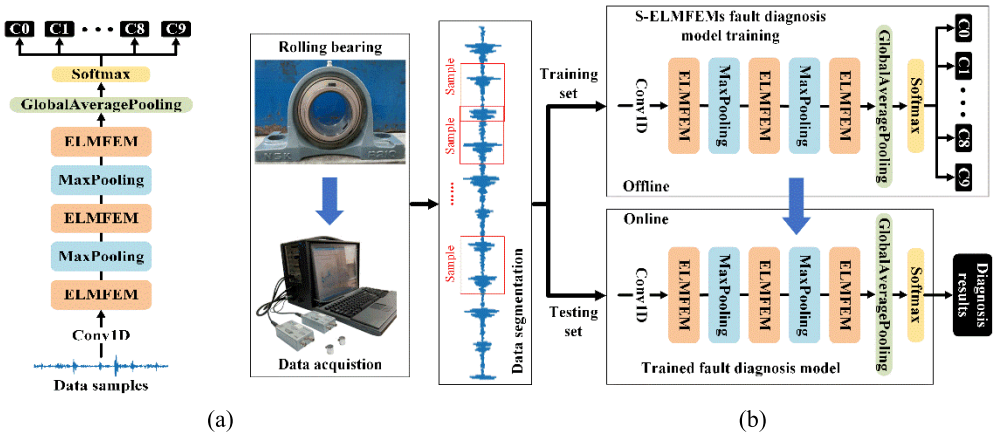


FIGURE 4. (a) Architecture of the S-ELMFEMs and (b) flowchart of the S-ELMFEMs based fault diagnosis method.

to add the output of FNN and the input of EMFEM to get the final output of ELMFEM. There are two reasons why the residual connection is used here. First, retain the original information to prevent the loss of useful information; and second, to overcome the performance degradation problem of deep networks.

D. STACKED ELMFEMS (S-ELMFEMS) BASED FOR FAULT DIAGNOSIS

As mentioned before, fault features of rolling bearings in industrial systems are highly coupled with intense environment noise, and the feature distribution under different operating conditions is quite different. Although complementary fault features can be extracted from multiple time scales, sufficient depth of the network is still indispensable. According to Lin *et al.* [41], the deeper networks can extract more abstract and higher-level fault features, which have strong condition expression ability and robustness, and can significantly improve model performance. Therefore, this article constructs a deep multiscale network (S-ELMFEMs) by stacking multiple ELMFEMs for rolling bearing fault diagnosis.

The architecture of S-ELMFEMs is shown in Fig. 4(a), where three ELMFEMs are used. In the experiment, 1D vibration signals are used as inputs of the model directly. First, the various type features of raw input signals are extracted through a 1D convolutional layer (Conv1D) to expand its channel dimension. In addition to being regarded as a feature extractor, the Conv1D can also implement residual connection operation in the first ELMFEM of S-ELMFEMs. The number of kernels in Conv1D can be adjusted according to the actual situation, which is set to 64 in this article. Whereafter, the output of the Conv1D is fed into the stacked ELMFEMs. It is worth mentioning that we use the max-pooling layer with a pooling length of 2 between every two ELMFEMs to capture local invariant features and accelerate training. Finally, the global average pooling layer is used to replace the full connection layer to prevent overfitting. A softmax layer is used to output a conditional probability for each category, which is defined as

$$P_c = \frac{\exp(\theta_c x)}{\sum_{c=1}^{C} \exp(\theta_c x)}, \quad c = 1, 2, \ldots, C \tag{5}$$

where θ_c is the model parameter and $\sum_{c=1}^{C} P_c = 1$.

Some structural parameters of the S-ELMFEMs in this article are summarized as follows: the number of kernels in Conv1D is 64, the dimension reduction factor r is set to 4, and the number of neurons in the two fully connected layers of FNN is 256 and 64, respectively.

The flowchart of the S-ELMFEMs based fault diagnosis method is shown in Fig. 4(b). It should be pointed out that both offline training and online testing are performed in a single specific task in this article, which satisfied the general assumption that the training data and testing data have the same feature distribution. If the offline training and online testing of the proposed model are used for different tasks, it is necessary to adopt techniques such as fine-tuning in the online testing part to realize online learning. The general procedures are summarized as follows.

1) *Step 1*: Collect vibration signals of rolling bearings in different health conditions through a data acquisition system. Then, segment vibration signals into multiple small segments of the same length as samples for training.
2) *Step 2*: We build an end-to-end fault diagnosis model based on S-ELMFEMs. The raw vibration signals are directly used as input. After the Conv1D, the extracted multi-dimensional feature maps are fed into stacked ELMFEMs and classifiers, and the corresponding diagnosis results are finally output. Offline training is completed until the maximal epoch is met.
3) *Step 3*: Collect test samples from vibration signals of the test bearing in the same way as training samples and input them into the well-trained model for online diagnosis. The model finally outputs the current health condition of the test bearing.

III. DATA DESCRIPTION AND EXPERIMENT SETUP

A. DATA DESCRIPTION

The fault dataset was acquired from the bearing data center of Case Western Reserve University (CWRU). The test stand in CWRU is shown in Fig. 5. The dataset includes ten health conditions: normal condition (N), inner race fault (IF) (fault diameters: 7 mil, 14 mil, 21 mil), ball fault (BF) (fault diameters: 7 mil, 14 mil, 21 mil) and 6 o'clock outer race fault (OF) (fault diameters: 7 mil, 14 mil, 21 mil), where the same health condition under different loads and speeds is treated as one category. The detailed description of the dataset and condition labels is summarized in Table 2. The vibration signals were collected from the drive end with a sampling frequency of 12 kHz.

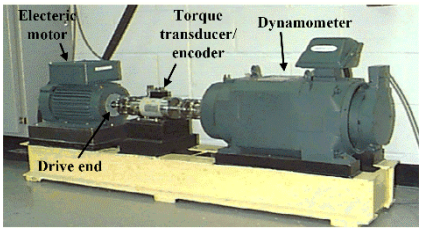


FIGURE 5. The test stand in CWRU.

TABLE 2. Description of the dataset and condition labels.

Label	Fault Description	Motor Speed (rpm)	Motor Load (HP)
C0	N	1797/1772/1750/1730	0/1/2/3
C1	IF_07	1797/1772/1750/1730	0/1/2/3
C2	OF_07	1797/1772/1750/1730	0/1/2/3
C3	BF_07	1797/1772/1750/1730	0/1/2/3
C4	IF_14	1797/1772/1750/1730	0/1/2/3
C5	OF_14	1797/1772/1750/1730	0/1/2/3
C6	BF_14	1797/1772/1750/1730	0/1/2/3
C7	IF_21	1797/1772/1750/1730	0/1/2/3
C8	OF_21	1797/1772/1750/1730	0/1/2/3
C9	BF_21	1797/1772/1750/1730	0/1/2/3

To prevent ''test leaky'', the raw vibration signal is first divided into five equal-length parts, one part is used as a test data subset to generate test samples, and the remaining part is used as a training data subset to generate training samples. Then the sliding window segmentation approach with overlap is used on the two subsets to obtain data samples. The window size (sample length) and sliding step size are set to 1200 and 600, respectively, ensuring the difference between the two adjacent samples. In this way, a total of 7920 data samples are obtained, including 6360 training samples and 1560 test samples. In the experiment, to ensure the reliability of the experimental results, the construction of the above data samples and each experiment are repeated five times, with each of the five equal-length parts used once as the test data subset. The average diagnosis results of the test set over five times are recorded.

B. EXPERIMENTAL SETUP

In this article, two evaluation indicators, testing accuracy and $F1$ score, are adopted to indicate the diagnosis performance, which are defined as:

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{7}$$

where TP, TN, FP, and FN represent the true positive samples, true negative samples, false positive samples, and false negative samples, respectively.

To simulate the noisy operating environment in real industrial systems, Gaussian white noise is added to the raw vibration signals with different signal-to-noise ratios (SNRs). The definition of SNR is shown as (8).

$$SNR_{\text{dB}} = 10\log_{10}(\frac{P_{\text{signal}}}{P_{\text{noise}}}) \tag{8}$$

where P_{signal} and P_{noise} denote the power of the raw signal and the noise, respectively.

The proposed model is implemented in the Keras (Tensorflow backend) and trained on a PC with a GTX 1060 GPU, 8GB of RAM, and an Intel Core i7-8750H CPU. In the experiment, the categorical cross-entropy loss function and Adam optimization algorithm with an initial learning rate of 0.0001 are adopted, and the batch size is set to 150. Besides, a learning rate decay operation is also adopted during training, that is, when loss of more than 3 epochs do not decrease, the learning rate is multiplied by 0.7.

IV. RESULTS AND DISCUSSION

In this section, the effectiveness and superiority of the proposed S-ELMFEMs are validated and discussed through five experiments on the above dataset. First, we explore the influence of the number of scales and ELMFEMs on diagnosis performance and determine the model structure mainly used in subsequent experiments. We then verify the effectiveness of the DFRM, residual learning, and 1DDSC in improving model performance and applicability through comparative experiments. Finally, the proposed model is compared with four existing multiscale CNN models to illustrate its superiority.

A. INFLUENCE OF THE NUMBER OF SCALES AND ELMFEMS ON DIAGNOSIS PERFORMANCE

The S-ELMFEMs can flexibly adjust the number of scales and network depth (i.e., the number of the ELMFEMs). Both parameters have a considerable influence on diagnosis performance. Hence, it is of great significance to study the influence of scale and depth on diagnosis performance.

In this study, ten different kernel sizes are set to be 1×1, 4×1, 9×1, 16×1, 25×1, 36×1, 49×1, 64×1, 81×1, and 100×1, and the number of all kernels is 16. Additionally, different numbers of ELMFEMs range from one to five are considered. The experiment is carried out under the noise

of −10 dB. The average results of accuracy over five folds are shown in Fig. 6. In the legend, 1s refers to using only 1×1 kernel, 2s refers to using 1×1 and 4×1 kernels, and so on, and 10s uses all ten kernels. Overall, the testing accuracy of the proposed model increases as the number of scales and ELMFEMs increases. When the model consists of four ELMFEMs, and each ELMFEM contains ten scales (denoted as the model (4, 10)), the testing accuracy is the highest, reaching 93.42%. This implies that the deeper network with more scales can learn more comprehensive and higher-level features, which enable the model to achieve better diagnosis performance. On the other hand, when the model contains more scales and ELMFEMs, the testing accuracy tends to be saturated. For example, the testing accuracy of the model (3, 8) is improved by nearly 11.42% compared to model (1, 5); however, model (5, 10) is only improved by about 0.53% compared to model (3, 8). The above phenomena can be explained from the following two aspects: 1) Limited by the training sample size. When the training sample size is fixed, and model parameters increase to a certain value, these parameters are difficult to continue to be optimized. 2) Limited by the length of training samples. According to the structure of the S-ELMFEMs, more max-pooling layers are used as the network depth increases, which greatly reduces the length of the extracted feature maps. As a result, the convolutional layers with the wide kernel (large scale branches) in the ELMFEMs at the end of the model can hardly extract useful features.

Therefore, for the challenging diagnosis tasks in practical engineering, the above problems can be solved using the following approaches to further improve the diagnosis performance: increasing the number and length of training samples, removing some max-pooling layers, and properly adjusting the kernel size. From the comprehensive consideration of the characteristics of the samples used in this article, the diagnosis performance, and training speed of the model, the model (8, 3) is mainly used in subsequent experiments. In the absence of special instructions, the S-ELMFEMs mentioned in this article refers to the model (8, 3).

To intuitively understand the influence of the number of scales and network depth on diagnosis performance, the t-SNE technique [42] is adopted to visualize the learned feature maps in a three-dimensional space for easier comparison.

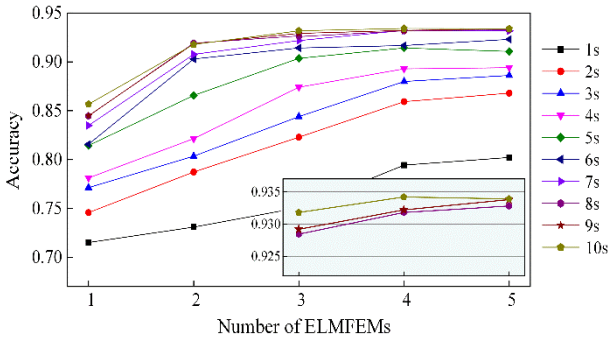


FIGURE 6. Influence of the number of scales and ELMFEMs on performance using the S-ELMFEMs.

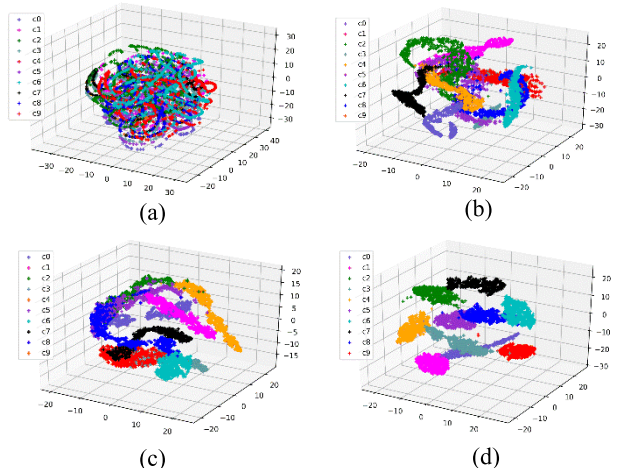
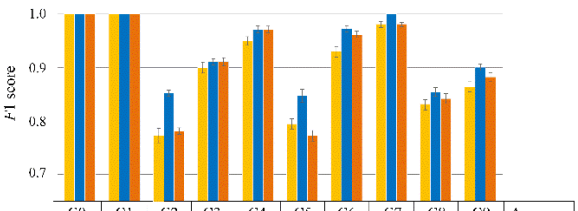


FIGURE 7. Feature visualization of the proposed model for different health conditions: (a) input data, (b) output of the model composed of single ELMFEM with single scale (64 × 1), (c) the model (8, 1), and (d) the model (8, 3).

As shown in Fig. 7, where different colors distinguish features of different health conditions. From Fig. 7(a), we can observe that different health conditions are very chaotic and overlap each other, which means that the input data's feature information is hardly separable. In Fig. 7(b), features in different health conditions are gradually separated, and features in the same condition are gradually clustered, representing that the distinguishing features are initially extracted. As the number of scales (c) and ELMFEMs (d) increase, more distinguishing features are learned, and features in the same condition present better clustering effects. In Fig. 7(d), the health conditions are clearly clustered into ten clusters, and only a few samples are misclassified. The visualization results further prove that the S-ELMFEMs can learn more discriminative and robust fault features from raw vibration signals, and has excellent fault classification ability.

B. EFFECTIVENESS OF THE DFRM

This section verifies the effectiveness of the DFRM under SNR = −10 dB. In this experiment, two model structures, i.e., the S-ELMFEMs and the model-I, are set up. It should be noted that the model-I only lacks the DFRM compared to the S-ELMFEMs. Both of them adopt the same training strategy and are trained and tested on the same dataset. The average results of the $F1$ score and accuracy over five times are shown in Fig. 8, where the error bar represents the standard deviation



	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	Accuracy
Model-I	1.00	1.00	0.77	0.90	0.95	0.79	0.93	0.98	0.83	0.86	0.9011
S-ELMFEMs	1.00	1.00	0.85	0.91	0.97	0.85	0.97	1.00	0.85	0.90	0.9286
Model-II	1.00	1.00	0.78	0.91	0.97	0.77	0.96	0.98	0.84	0.88	0.9095

FIGURE 8. Performance of the proposed model and comparison model.

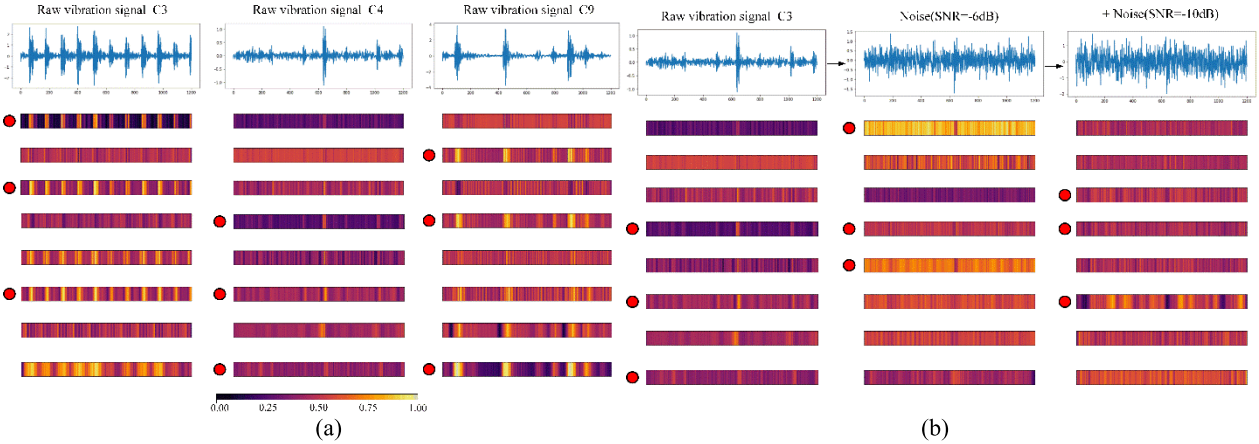


FIGURE 9. The visualization of optimization vectors with (a) raw vibration signals of rolling bearings in different health conditions as input and (b) vibration signals of different SNRs as input.

and shows the stability of diagnosis performance. Obviously, the performance of the S-ELMFEMS always outperforms the model-I. Specifically, compared to the model-I, the accuracy is improved by 2.75%, and the $F1$ score of different health conditions are improved by 0-0.08. In addition, it can also be found that a smaller standard deviation for each condition can be noticed for S-ELMFEMs. These results prove that DFRM can effectively improve the discriminative fault feature extraction ability, and enable the model to obtain better and more stable diagnosis performance.

To further understand the learning mechanism of the DFRM and multiscale CNN, the optimization vectors of the well-trained model are visualized for different inputs, as shown in Fig. 9. To obtain a better correspondence with input signals, the visualized optimization vectors selected here come from eight scales of the first ELMFEM. It can be seen in Fig. 9(a) that almost all optimization vectors can adaptively locate the impact segments and reinforce the fault-related features reflected by them according to the intrinsic characteristics of the input data itself so that the model can pay more attention to them. In Fig. 9(b), when noise is added, the impact segments are overwhelmed. However, there are still some optimization vectors that can locate them. This fault feature reinforcing mechanism enables the model to efficiently extract more discriminative fault information, thus improving the diagnosis performance. The experiment results further prove the effectiveness of the DFRM.

Furthermore, the extraction ability of different scales to discriminative fault features is defined through the location accuracy and reinforcement degree (judged by the color contrast) of the corresponding optimization vector to impact segments. According to the above definition, the three scales with the highest extraction ability are marked with red dots in Fig. 9. It can be found that the marked scales are not fixed. This indicates that different scales have different discriminative features extraction ability for different input signals (such as different health conditions and noise levels). This proves that the superiority of multiscale CNN, which can extract

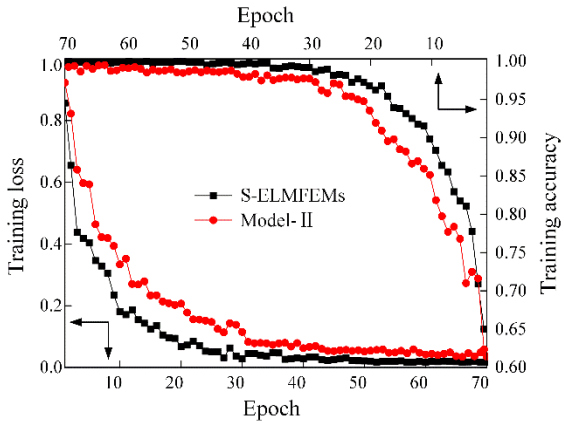


FIGURE 10. Training accuracy and loss of the proposed model and comparison model with respect to epoch.

abundant and complementary fault features from different time scales to make up for the defects of a single scale, thus has stronger feature learning and anti-noise ability than traditional single scale CNN. What is more, these results make the learning mechanism of multiscale CNN more interpretable. It can be understood as presetting multiple feature extractors with different scales to match an unknown vibration signal, resulting in a greatly increased probability of extracting discriminative fault features. Compared with the single scale CNN, it has higher robustness and more reliable diagnosis results.

C. EFFECTIVENESS OF THE RESIDUAL LEARNING

This section verifies the effectiveness of residual learning under SNR $= -10$ dB. In this experiment, two model structures, i.e., the S-ELMFEMs and the model-II, are set up. The model-II only lacks the residual connection compared to the S-ELMFEMs. Both of them adopt the same training strategy and are trained and tested on the same dataset. The results are summarized in Fig. 8. It can be easily observed

TABLE 3. Parameters and size of the proposed model with different number of ELMFEMs. (The number of scales per ELMFEM is set to 8).

Number of ELMFEMs	Standard 1D convolution			1DDSC		
	Parameters	Size(MB)	Accuracy	Parameters	Size(MB)	Accuracy
1	263,602	3.97	0.8577	75,954	1.72	0.8567
2	526,170	7.82	0.9158	150,874	3.35	0.9179
3	788,738	11.61	0.9302	225,794	4.96	0.9318
4	1,051,306	15.56	0.9381	300,714	6.61	0.9342
5	1,313,874	19.43	0.9341	375,634	8.22	0.9339

TABLE 4. Parameters and size of the proposed model with different number of scales. (The number of ELMFEMs is set to 3).

Number of scales	Standard 1D convolution			1DDSC		
	Parameters	Size(MB)	Accuracy	Parameters	Size(MB)	Accuracy
2	92,600	1.50	0.8241	68,624	1.41	0.8229
4	195,686	3.07	0.8729	89,654	2.14	0.8740
6	411,092	6.80	0.9183	118,460	3.01	0.9143
8	788,738	11.61	0.9304	158,498	4.96	0.9286

that the performance of S-ELMFEMs always outperforms the model-II. Specifically, compared with the model-II, the accuracy is improved by 1.91%, and the $F1$ score of different health conditions are improved by 0-0.08. It indicated that the performance degradation is effectively alleviated in the S-ELMFEMs. Fig. 10 shows the training accuracy and loss of the first 70 epochs. It can be found that the training accuracy of both models is almost 100% and higher than the testing accuracy, but the testing accuracy of the S-ELMFEMs is higher than model-II, which means that the S-ELMFEMs has better generalization. Moreover, compared with the model-II, the training loss curve of the S-ELMFEMs is steeper, and the final loss is smaller, which means that the S-ELMFEMs converges quicker and better during the training process. The above results further prove that residual learning can effectively alleviate the performance degradation in the deeper network and reduce the training difficulty.

D. EFFECTIVENESS OF THE 1DDSC

In this section, the effectiveness of 1DDSC is verified by comparing it with standard 1D convolution on the proposed model. The parameters and size of the model with different number of ELMFEMs and scales are list in Table 3 and Table 4, respectively, and their change trends are shown in Figure 11. Besides, the diagnostic accuracy of models is also given in Table 3 and Table 4 for comparison. It can be seen that as the number of ELMFEM increases, the parameters and size of the model increase linearly. Compared with standard 1D convolution, the 1DDSC reduces model parameters by about 70% and model size by more than 55% under the same number of ELMFEM. Moreover, with the introduction of more scales and wider convolutional kernels, the model's parameters and size increase exponentially, and the advantages of 1DDSC become more prominent. The parameters and size of the proposed model are dramatically reduced by using 1DDSC instead of standard 1D convolution, so the storage and computational costs of the model are reduced, and its applicability is improved. It can also be found

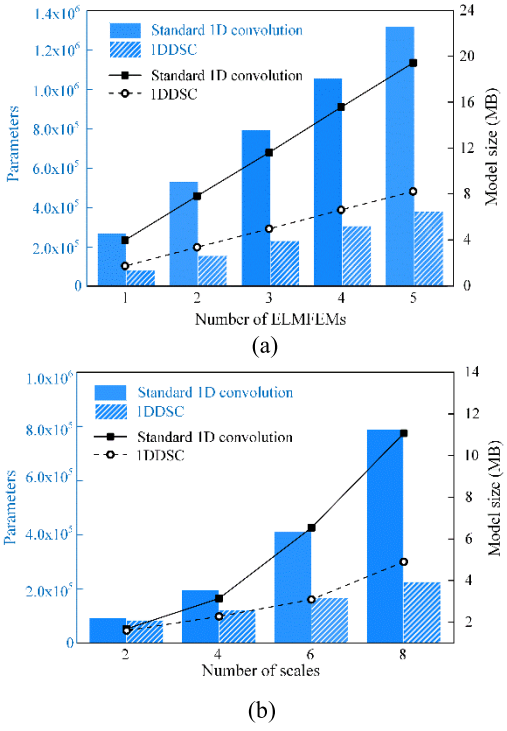


FIGURE 11. Parameters and size of the proposed model with different number of (a) ELMFEMs and (b) scales.

that the diagnostic accuracy of the proposed model under these convolution operations is almost the same, and the use of 1DDSC can maintain the excellent diagnosis performance of the model. The experiment results prove the effectiveness of the 1DDSC, which also explains lightweight in the name of the ELMFEM.

E. COMPARISON WITH EXISTING MULTISCALE CNN MODELS

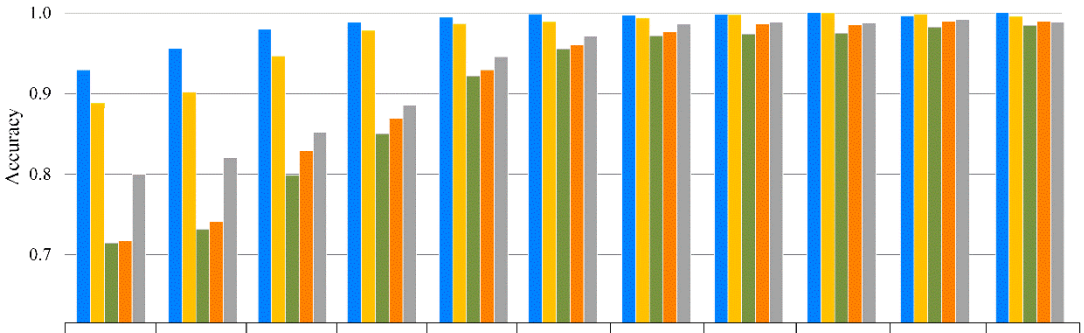
To verify the superiority of the proposed model, four existing multiscale CNN models, including MK-ResCNN [33],

AWMSCNN [27], MS-DCNN [36], and MSCNN[32] are implemented as comparisons in this study. The MK-ResCNN uses three scale branches composed of three different kernel sizes (3×1, 5×1, and 7×1). Each scale branch extracts features through three CNN blocks, and each CNN block is composed of two convolution layers and a residual connection. The AWMSCNN consists of a denoising layer, four-scale feature learning block (2×1, 4×1, 6×1, and 8×1), multiscale feature weighting layer, and a feature fusion layer. The MS-DCNN consists of three pairs of convolutional and pooling layers, four-scale feature extraction block (1×1, 3×1, 5×1, and 4×1 pooling), and two fully connected layers. In the MSCNN, multiple coarse-grained layers are adopted to represent the raw vibration signal. Then two pairs of convolutional and pooling layers and a global average pooling layer are used to extract features.

The experiments are carried out in different noise environments with the SNRs ranging from -10 dB to 10 dB, and the training strategies of the five models are the same. The experimental results are summarized in Fig. 12.

Obviously, the S-ELMFEMs is superior to the other four CNN models, and its diagnostic accuracy exceeds 95.5% at almost full noise level (from -8 dB to 10 dB). In general, the accuracy of the five models increases with the increase of SNR, and after the SNR exceeds 0 dB, the change of accuracy tends to be stable. It is because the diagnostic accuracy mainly depends on the ability of the model itself to extract fault features when the noise is small. When the SNR changes from 0 to 10 dB, only S-ELMFEMs always maintain an accuracy of over 99.9%, which indicates that the proposed model has excellent and stable discriminative fault feature extraction ability. What is more, when the SNR decreases to -10 dB, the S-ELMFEMs can still achieve accuracy of 92.86%, which is 3.59%, 21.55%, 21.37%, and 12.79% higher than MK-ResCNN, AWMSC, MS-DCNN, and MSCNN respectively. Therefore, the S-ELMFEMs has strong anti-noise ability even without any additional denoising process. Furthermore, SNR $= -10$ dB indicates that the noise power is much larger than the raw signal power, which means that the proposed model can effectively detect extremely weak faults.

Although the S-ELMFEMs has only a small improvement over the MK-ResCNN (the maximum improvement is 5.28% when SNR $= -8$ dB), the storage and computational costs of the S-ELMFEMs are much less than the MK-ResCNN. The proposed model has two orders of magnitude less floating-point computation amount than MK-ResCNN. As mentioned in Section 1, the model's storage and computational costs are two crucial factors for developing a real-time diagnosis system in the Indusial Internal of Things context. As shown in Table 5, the size, training time (per epoch), and testing time (per sample) of the five models are recorded. The size of S-ELMFEMs is less than one-fifth of the MK-ResCNN. Thus, it has higher applicability and broader application. Also, both the training time and testing time of the MK-ResCNN are almost twice that of the S-ELMFEMs, which means that S-ELMFEMs has higher operating efficiency, lower diagnosis system development cost, and faster releases of updates when much more training data are available. Especially for real-time diagnosis, the S-ELMFEMs can help operators make timely decisions due to faster testing speed. According to Table 5, the AWMSC, MS-DCNN, and MSCNN have a smaller size and consume less training and testing time than S-ELMFEMs. It is not a surprise because the S-ELMFEMs has more scales and deeper network structure than the other four models. Even so, it still has only 4.96 MB model size, 13.55 s training time, and 2.76 ms testing time. Hence, considering the diagnosis



	-10 dB	-8 dB	-6 dB	-4 dB	-2 dB	0 dB	2 dB	4 dB	6 dB	8 dB	10 dB
S-ELMFEMs	0.9286	0.9563	0.9736	0.9803	0.9886	0.9998	0.9992	0.9996	1.0000	0.9992	1.0000
MK-ResCNN	0.8927	0.9035	0.9389	0.9712	0.9795	0.9896	0.9968	0.9992	1.0000	0.9996	0.9979
AWMSC	0.7131	0.7379	0.8006	0.8528	0.9217	0.9589	0.9758	0.9798	0.9789	0.9875	0.9898
MS-DCNN	0.7149	0.7499	0.8326	0.8767	0.9311	0.9637	0.9802	0.9897	0.9906	0.9937	0.9948
MSCNN	0.8007	0.8265	0.8624	0.8923	0.9551	0.9727	0.9891	0.9904	0.9927	0.9962	0.9936

FIGURE 12. Performance of S-ELMFEMs and four comparison models in different noise environment.

TABLE 5. Size and cost time of different models.

	Floating-point computations	Model size (MB)	Training time (s)	Testing time (ms)
S-ELMFEMs	7.4×10^7	4.96	13.55	2.76
MK-ResCNN	1.2×10^9	25.18	27.42	5.25
AWMSC	6.5×10^5	0.32	1.48	0.43
MS-DCNN	1.1×10^6	0.92	1.78	0.67
MSCNN	3.9×10^7	2.15	2.91	0.85

performance and the cost of storage and computation, the proposed model has more enormous advantages, especially for practical industrial applications.

V. CONCLUSION

Targeting the fault diagnosis of the rolling bearing under complex operation conditions and intense noise, this article proposes an ELMFEM and develops an S-ELMFEMs fault diagnosis model. The S-ELMFEMs is easy to expand, and the techniques involved, such as discriminative feature reinforcement mechanism (DFRM), residual learning, and 1D depthwise separable convolution (1DDSC), etc., form complementary advantages, improve the ability of fault feature extraction and the applicability in the Industrial Internet of Things context. Besides, the interpretability of multiscale CNN is also preliminarily explored in the article. It provides a solution to overcome the three challenges about multiscale CNN mentioned in Section 1.

The proposed S-ELMFEMs fault diagnosis model is evaluated on the rolling bearing dataset. The experimental results illustrate the effectiveness of DFRM, residual learning, and 1DDSC. They can significantly improve the multiscale CNN performance while dramatically reducing the parameters and size of the model. Additionally, the S-ELMFEMs shows significant advantages over existing multiscale CNN models. Our work shows that the proposed model not only has excellent discriminative fault features extraction ability and strong anti-noise ability but also has low storage and computational costs. It indicates that the proposed model is more promising in practical engineering, especially in the Industrial Internet of Things context. Moreover, the proposed multiscale CNN diagnosis framework can also be applied to other industrial systems without any processing.

For further work, first, we are going to introduce the transfer learning to improve the model's domain adaptation ability. Second, we will further explore the learning mechanism of multiscale CNN to improve the pertinence of scale parameter setting according to specific diagnosis tasks, thereby further improving the diagnosis performance of the model and simplifying the model structure.

REFERENCES

[1] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 485–520, 2011.

[2] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, nos. 1–2, pp. 108–126, Feb. 2013.

[3] I. El-Thalji and E. Jantunen, "A summary of fault modelling and predictive health monitoring of rolling element bearings," *Mech. Syst. Signal Process.*, vols. 60–61, pp. 252–272, Aug. 2015.

[4] Z. Liu, Z. Jia, C.-M. Vong, S. Bu, J. Han, and X. Tang, "Capturing high-discriminative fault features for electronics-rich analog system via deep learning," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1213–1226, Jun. 2017.

[5] 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.

[6] R.-B. Sun, Z.-B. Yang, L.-D. Yang, B.-J. Qiao, X.-F. Chen, and K. Gryllias, "Planetary gearbox spectral modeling based on the hybrid method of dynamics and LSTM," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106611.

[7] S. Hao, F.-X. Ge, Y. Li, and J. Jiang, "Multisensor bearing fault diagnosis based on one-dimensional convolutional long short-term memory networks," *Measurement*, vol. 159, Jul. 2020, Art. no. 107802.

[8] Y. Chang, J. Chen, C. Qu, and T. Pan, "Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels," *Renew. Energy*, vol. 153, pp. 205–213, Jun. 2020.

[9] M. Gan, C. Wang, and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 92–104, May 2016.

[10] H. Shao, H. Jiang, H. Zhang, and T. Liang, "Electric locomotive bearing fault diagnosis using a novel convolutional deep belief network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 3, pp. 2727–2736, Mar. 2018.

[11] Z. Meng, X. Zhan, J. Li, and Z. Pan, "An enhancement denoising autoencoder for rolling bearing fault diagnosis," *Measurement*, vol. 130, pp. 448–454, Dec. 2018.

[12] 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.

[13] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.

[14] 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 Vib.*, vol. 377, pp. 331–345, Sep. 2016.

[15] M. Xia, T. Li, L. Xu, L. Liu, and C. W. de Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, Feb. 2018.

[16] D. Peng, Z. Liu, H. Wang, Y. Qin, and L. Jia, "A novel deeper one-dimensional CNN with residual learning for fault diagnosis of wheelset bearings in high-speed trains," *IEEE Access*, vol. 7, pp. 10278–10293, 2019.

[17] 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, Feb. 2017.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 2727–2736.

[19] K. B. Charbonneau and O. Shouno, "Neural trajectory analysis of recurrent neural network in handwriting synthesis," 2018, *arXiv:1804.04890*. [Online]. Available: http://arxiv.org/abs/1804.04890

[20] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

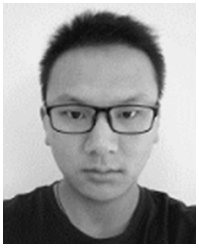
[21] Y. Qin, S. Xiang, Y. Chai, and H. Chen, "Macroscopic–microscopic attention in LSTM networks based on fusion features for gear remaining life prediction," *IEEE Trans. Ind. Electron.*, vol. 67, no. 12, pp. 10865–10875, Dec. 2020.

[22] Y. Qin, D. Chen, S. Xiang, and C. Zhu, "Gated dual attention unit neural networks for remaining useful life prediction of rolling bearings," *IEEE Trans. Ind. Informat.*, early access, Jun. 2, 2020, doi: 10.1109/TII.2020.2999442.

[23] 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.
[24] H. Wang, J. Xu, R. Yan, C. Sun, and X. Chen, "Intelligent bearing fault diagnosis using multi-head attention-based CNN," *Procedia Manuf.*, vol. 49, pp. 112–118, Jan. 2020.
[25] D. Yao, H. Liu, J. Yang, and X. Li, "A lightweight neural network with strong robustness for bearing fault diagnosis," *Measurement*, vol. 159, Jul. 2020, Art. no. 107756.
[26] Z. Liu, H. Wang, J. Liu, Y. Qin, and D. Peng, "Multitask learning based on lightweight 1DCNN for fault diagnosis of wheelset bearings," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.
[27] 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.
[28] D. Peng, H. Wang, Z. Liu, W. Zhang, M. J. Zuo, and J. Chen, "Multibranch and multiscale CNN for fault diagnosis of wheelset bearings under strong noise and variable load condition," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4949–4960, Jul. 2020.
[29] L. Zhang, G. Xiong, H. Liu, H. Zou, and W. Guo, "Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 6077–6085, Aug. 2010.
[30] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mech. Syst. Signal Process.*, vol. 85, pp. 746–759, Feb. 2017.
[31] H. Liu and M. Han, "A fault diagnosis method based on local mean decomposition and multi-scale entropy for roller bearings," *Mechanism Mach. Theory*, vol. 75, pp. 67–78, May 2014.
[32] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019.
[33] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3797–3806, Jun. 2020.
[34] R. Fernandez Molanes, K. Amarasinghe, J. Rodriguez-Andina, and M. Manic, "Deep learning and reconfigurable platforms in the Internet of Things: Challenges and opportunities in algorithms and hardware," *IEEE Ind. Electron. Mag.*, vol. 12, no. 2, pp. 36–49, Jun. 2018.
[35] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24411–24432, 2018.
[36] Z. Zilong and Q. Wei, "Intelligent fault diagnosis of rolling bearing using one-dimensional multi-scale deep convolutional neural network based health state classification," in *Proc. IEEE 15th Int. Conf. Netw., Sens. Control (ICNSC)*, Mar. 2018, pp. 1–6.
[37] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5353–5360.
[38] X. Li, J. Li, C. Zhao, Y. Qu, and D. He, "Gear pitting fault diagnosis with mixed operating conditions based on adaptive 1D separable convolution with residual connection," *Mech. Syst. Signal Process.*, vol. 142, Aug. 2020, Art. no. 106740.
[39] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[41] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 818–833.
[42] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

YAOWEI SHI received the M.S. degree from Nanjing Tech University, Nanjing, China. He is currently pursuing the Ph.D. degree in engineering with Southeast University, Nanjing. His current research interests include signal processing, deep learning, and fault diagnosis.

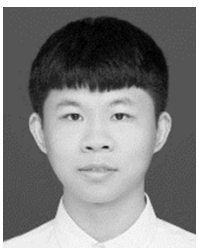
AIDONG DENG received the Ph.D. degree in engineering from Southeast University, Nanjing, China. He is currently a Professor and a Doctoral Advisor with Southeast University. His current research interests include signal processing, machine learning, and mechanism analysis.



MINQIANG DENG received the M.S. degree from Southeast University, Nanjing, China, where he is currently pursuing the Ph.D. degree in engineering. His current research interests include signal processing, machine learning, and fault diagnosis.

JING ZHU received the M.S. degree from Southeast University, Nanjing, China, where she is currently pursuing the Ph.D. degree in engineering. Her current research interests include signal processing, machine learning, and vibration control.

YANG LIU received the B.S. degree from the Qingdao University of Science and Technology, Qingdao, China. He is currently pursuing the M.S. degree in engineering with Southeast University, Nanjing, China. His current research interests include deep learning and artificial intelligence.



QIANG CHENG received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, China. He is currently pursuing the M.S. degree in engineering with Southeast University, Nanjing. His current research interests include machine learning and data mining.