IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received November 27, 2018, accepted December 10, 2018, date of publication December 20, 2018,
date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2888842

A Novel Deeper One-Dimensional CNN With Residual Learning for Fault Diagnosis of Wheelset Bearings in High-Speed Trains

DANDAN PENG[1], ZHILIANG LIU[1,2], HUAN WANG[1], YONG QIN[2], (Member, IEEE), AND LIMIN JIA[2]

[1]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

Corresponding authors: Zhiliang Liu (zhiliang_liu@uestc.edu.cn) and Yong Qin (yqin@bjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1200401, in part by the National Natural Science Foundation of China under Grant 51505066 and Grant 61833002, and in part by the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, under Grant RCS2018K002.

ABSTRACT The health condition of a wheelset bearing, the key component of a railway bogie, has a considerable impact on the safety of a train. Traditional bearing fault diagnosis techniques generally extract signals manually and then diagnose the bearing health conditions through the classifier. However, high-speed trains (HSTs) are usually faced with variable loads, variable speeds, and strong environmental noise, which pose a huge challenge to the application of the traditional bearing fault diagnosis methods in wheelset bearing fault diagnosis. Therefore, this paper proposes a 1D residual block, and based on the block, a novel deeper 1D convolutional neural network (Der-1DCNN) is proposed. The framework includes the idea of residual learning and can effectively learn high-level and abstract features while effectively alleviating the problem of training difficulty and the performance degradation of a deeper network. Additionally, for the first time, we fully use the wide convolution kernel and dropout technology to improve the model's ability to learn low-frequency signal features related to the fault components and to enhance the network's generalization performance. By constructing a deep residual learning network, Der-1DCNN can adaptively learn the deep fault features of the original vibration signal. This method not only achieves very high diagnostic accuracy for the fault diagnosis task of wheelset bearings in HSTs under strong noise environment, but also its performance is quite superior when the train's working load changes without any domain adaptation algorithm processing. The proposed Der-1DCNN is evaluated on the dataset of the multi-operating conditions of the wheelset bearings of HSTs. Experiments show that this method shows a better diagnostic performance compared with the state-of-the-art deep learning methods of bearing fault diagnosis, which proves the method's effectiveness and superiority.

INDEX TERMS High-speed trains, wheelset bearings fault diagnosis, deep learning, one-dimensional residual block, wide convolutional kernel.

I. INTRODUCTION

Recently, high-speed trains (HSTs) have experienced a rapid development throughout the world. However, maintaining the safety and reliability of HSTs is challenging with the increasing speed. The health condition of a wheelset bearing, the core component of a railway bogie, has a considerable impact on the safety of trains. Once the bearing fails, it may endanger the normal machine operation and cause significant economic losses. Sometimes it even results in a serious safety accident. Consequently, condition monitoring and the fault diagnoses of wheelset bearings are valuable and meaningful for maintaining the normal operation and safety of HSTs. However, due to the complexity and variability of the working environments of HSTs, the vibration signals collected from wheelset bearings are susceptible to noise and other components. Therefore, the fault diagnoses of these types of bearings are more challenging when comparing with the bearings in common industrial equipment.

Traditional intelligent fault diagnosis methods mainly include three steps: data acquisition, feature extraction and fault identification. It is worth noting that feature extraction and fault identification are two important steps for bearing fault diagnosis, directly affecting the accuracy of fault classification. Generally, time domain features (kurtosis [1], entropy [2] and so forth) and time-frequency domain features (wavelet packet [3], Hilbert spectrum [4] and so forth) are manually extracted. Then, these statistical parameters are fed into machine learning algorithms, such as a support vector machine (SVM) [4], [5], k-Nearest neighbor (kNN) [1] and artificial neural network (ANN) [2]. For the fault diagnosis of HSTs, Cao *et al.* [6] applied the empirical wavelet transform (EWT) for wheelset bearing fault diagnosis and obtained good performance in the detection of outer race faults, roller faults and the compound fault of outer race and roller. Wang *et al.* [7] proposed ensemble empirical mode decomposition (EEMD) and then extracted the kurtosis features of each component to diagnose the wheelset bearing fault. Qin *et al.* [8] introduced fuzzy entropy and EEMD to analyze the fault features of HSTs. In [9], the feature selection of HST bogie fault signals was performed with wavelet entropy, and then, an SVM was used as the model of fault recognition. Liu *et al.* [10] employed the SVM framework as the fault classifier of the braking systems of HSTs and obtained relatively high accuracy.

However, the intelligent fault diagnosis methods mentioned above still have some drawbacks: 1) the diagnosis performances rely heavily on the design of feature extraction methods, which often require experts who have strong domain knowledge and rich practical experiences. For every specific fault diagnosis task, feature extraction methods must be redesigned, so it is time consuming and labor intensive. 2) Extracting features using domain knowledge cannot guarantee that the statistical parameters can fully represent the complex dynamic characteristics. This is mainly due to the following three aspects. First, compared with the bearing vibration signals of general equipment, those collected from wheelset bearings are nonlinear and non-stationary with stronger noise. Second, the working conditions are changing during the equipment's operation, so the fault features are variable and complex. Third, the feature distribution of data samples under various working conditions are generally different, typically different load conditions. 3) These machine learning classifiers, for example, an SVM, kNN and ANN, employ shallow networks, so it is difficult to learn adequate features. In addition, the nonlinear relationship of fault signals may not be effectively learned, thus causing one to make misjudgments.

Based on the above discussions, deep learning techniques with their powerful automatic feature learning ability are expected to provide an effective solution for the intelligent fault diagnosis of wheelset bearings in HSTs. The core of deep learning is feature learning [11], whose aim is to obtain hierarchical feature information through a hierarchical network, thus solving an important problem that required features to be manually extracted in the past. A convolutional neural network (CNN) [12] is a powerful deep learning method that has been successfully applied in various fields, such as computer vision [13], speech recognition [14], natural language processing [15] and one-dimensional (1D) signal processing [16]. Generally, the CNN is a hierarchical model that uses raw data as input and extracts high-level features layer by layer from the original data through convolution operations, pooling operations, nonlinear activation function mapping, and so on. Compared with the traditional fully connected neural network, a CNN can learn more robust features and has better generalization performance. Meanwhile, it can also save on training costs through weight sharing and pooling operations. Therefore, this paper's aim is to develop an end-to-end wheelset bearing fault diagnosis method based on CNNs, and to the best of our knowledge, CNN technology is first used in the fault diagnosis of wheelset bearings in HSTs in this paper.

First, as mentioned before, the fault features of wheelset bearings are highly coupled with noise, and the feature distribution under various working loads are quite different, which significantly increases the difficulty of fault feature extraction from simple shallow CNN models. To effectively extract fault-related features from vibration signals with complex interference, the network should learn higher-level and more abstract signal features so as to filter fault-related features from complex signal features. CNNs naturally integrate low/mid/high-level features and classifiers in an end-to-end multilayer fashion, and the "levels" of features can be enriched by the number of stacked layers [17], so a deeper network can learn richer and higher-level signal features. However, the deeper CNN has its own defects in addition to its powerful feature learning ability. First, backpropagation calculates the gradient through the chain rule, which easily leads to an exponential decrease/increase of the gradient as the layer increases. Therefore, the deeper CNN often encounters the vanishing or exploding gradient problem, and its training becomes more difficult [18]. Second, network degradation is another major problem, which leads to an increase of the training error of training samples [17]. These two problems greatly limit the deeper CNN's development in the field of fault diagnosis.

Therefore, this paper proposes the 1D residual learning block, which is a further improvement and application of Resnet's [17] take on 1D vibration signals. Based on this block, a novel deeper 1D CNN (Der-1DCNN) is constructed. The framework can extract higher-level signal features from raw vibration signals via a deep network architecture and has better fault discrimination abilities. Moreover, a Der-1DCNN can not only effectively alleviate the problem of training difficulty for 1D deeper CNNs but also dynamically adapt to datasets of various sizes by adjusting the number of 1D residual blocks.

Additionally, when the wheelset bearings of HSTs fail, low-frequency impact components and modulation components are introduced. These fault features are easily overwhelmed by strong noises and inherent high-frequency vibration signals with dominant amplitudes. Therefore, how to improve the CNN architecture so that it can effectively learn low-frequency fault-related features is another challenge for the fault diagnosis of wheelset bearings. In the convolutional layer of a CNN, the maximum range of input signals that each convolution operation can perceive is closely related to the size of the convolution kernel. Generally, a wide convolution kernel can better learn the global low-frequency trend features of vibration signals without local interference features misleading it.

Therefore, we first propose fully using wide convolution kernels of various sizes in the convolutional layer of a 1D CNN, instead of the narrow convolution kernels (1×3) used in traditional CNNs, to enhance the learning of low-frequency features and global features. In addition, as the network depth and the size of the convolution kernel increase, the network parameters inevitably increase, so overfitting is inevitable. To prevent overfitting, we introduce the dropout technique [19], which is commonly used in the fully connected layer, into the convolutional layer. Its key idea is to randomly drop out the convolutional kernel during training, which prevents parameters from co-adapting too much and makes the model learn more robust features. The introduction of dropout is equivalent to training multiple ''thinned'' networks and then using the idea of model integration to effectively prevent the problem of overfitting and to improve the network's generalization performance.

The key contributions of this paper are summarized as follows. 1) First, we propose a 1D residual block by introducing the idea of residual learning into the traditional 1D CNN. This block can effectively solve the problem of training difficulty and performance degradation for the deeper CNN. Additionally, the introduction of a wide convolutional kernel and dropout further enhances the feature learning ability of the deeper CNN in a noisy environment. 2) This paper constructs a novel Der-1DCNN framework based on a 1D residual block and develops an end-to-end bearing intelligent fault diagnosis method. This system regards the raw vibration signal as input, which can automatically learn the high-level features and classify various health conditions simultaneously. It does not require any additional signal processing or expert knowledge, thus effectively improving the applicability of intelligent fault diagnosis systems. 3) The proposed Der-1DCNN is evaluated through experiments on the wheelset bearing test rig of HSTs with a comprehensive performance evaluation. Compared with state-of-the-art deep learning methods applied to bearing fault diagnosis, a Der-1DCNN can significantly improve the diagnostic performance of bearing fault diagnosis in HSTs in a strong noise environment and has strong domain adaptation ability under changing load conditions.

The rest of this paper is structured as follows. Section 2 describes related works. In Section 3, the proposed Der-1DCNN is detailed. In Section 4, the wheelset bearing datasets are employed to demonstrate the effectiveness and superiority of the proposed method. Additionally, the validity of a Der-1DCNN is investigated and discussed in Section 5. Finally, conclusions are drawn and future works are discussed in Section 6.

II. RELATED WORKS

The related works described in this section contain intelligent fault diagnosis methods for HSTs and CNNs for the fault diagnosis of rotating machinery.

A. INTELLIGENT FAULT DIAGNOSIS METHODS FOR HIGH SPEED TRAINS

In the field of the intelligent fault diagnosis of HSTs, Zhao *et al.* [20] introduced empirical mode decomposition (EMD) and fuzzy entropy to extract the signal features of HSTs, and a back propagation (BP) neural network was applied as the model for fault diagnosis. However, this method leads to the wrong state recognition in some working conditions. Xie *et al.* [21] adopted fast Fourier transform (FFT) to extract the signal features of bogie HSTs, and k-DBNs based on kNNs and deep belief networks (DBNs) are proposed to classify faults, but recognition accuracy is about 54% in a real environment. Pang *et al.* [22] used the denoising autoencoder (DAE) and BP neural network to recognize the bogie faults of HSTs after signal preprocessing using the discrete Fourier transform (DFT), but it has poor performance in an actual running environment. Guo *et al.* [23] proposed a novel DBN for the fault analysis of an HST under a single failure condition, whose input is the FFT signal. However, the FFT technique can be used to process only stationary signals, whereas the time-domain signals of bogies are non-stationary. Yin and Zhao [24] developed a DBN model to provide for the real-time monitoring and diagnosis of vehicle on-board equipment. We can clearly see that up to now, few studies have been done on the wheelset bearing fault diagnosis using deep learning techniques. Even though some scholars have studied the fault diagnosis of key components of the HSTs of bogies, several problems still exist. For example, the diagnosis accuracy is low, and the method's performance is poor under complex conditions. Therefore, it is valuable to study wheelset bearing fault diagnosis under complex working conditions.

B. CNN FOR FAULT DIAGNOSIS OF ROTATING MACHINERY

Recently, CNNs have been widely used in rotating machinery fault diagnosis due to excellent feature learning ability. In 2015, Chen *et al.* [25] applied a CNN for the fault diagnosis of gearboxes. However, manual feature extraction was still needed. Janssens *et al.* [26] proposed a three-layer CNN model for bearing fault recognition, in which the input of a CNN is a DFT signal. Guo *et al.* [27] proposed a hierarchical adaptive deep CNN for bearing fault diagnosis. Ince *et al.* [28] established a 1D-CNN model for motor fault detection from raw time series data, which successfully avoids the time-consuming feature extraction process. Considering multi-sensor data information, Xia *et al.* [29] designed the 1D-CNN model to diagnose faults of rotating machinery, and it proved that the diagnosis result of

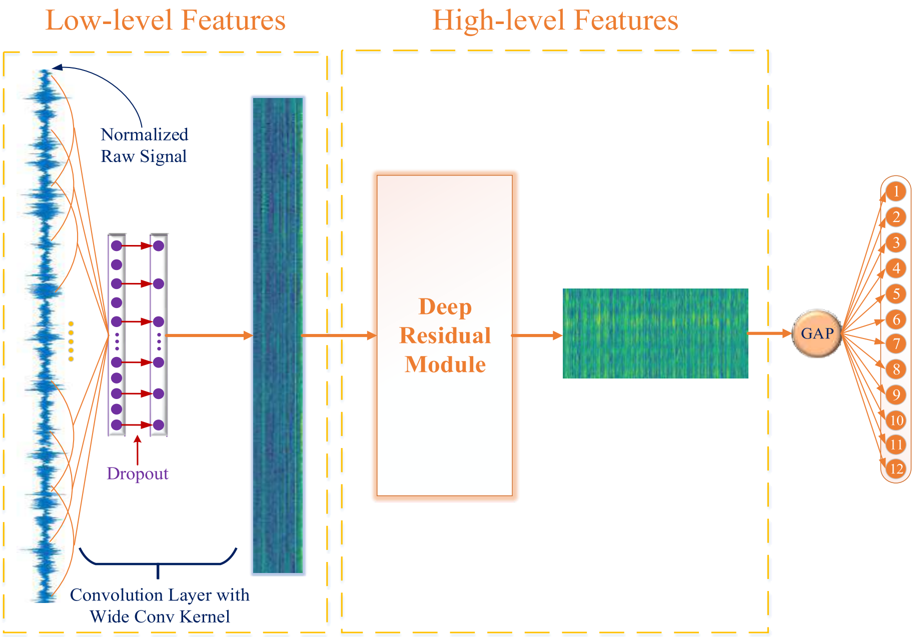


FIGURE 1. Architecture of the proposed Der-1DCNN framework.

multi-sensor data is better than that of the single sensor. Zhang *et al.* [30] proposed a WDCNN model from the raw signals for the bearing fault diagnosis. In 2018, he also proposed the TICNN model with training interference, which can work on raw noisy data and under different working loads for bearing fault diagnosis [31]. Recently, Jiang *et al.* [32] proposed multiscale CNNs for fault diagnosis of wind turbine gearbox by combining multi-scale learning with deep learning techniques, and it achieved high accuracy. Through converting signals into two-dimensional (2D) images, Wen *et al.* [33] designed a CNN based on LeNet-5 and applied it in a motor bearing dataset. It can be found that the above networks all adopt shallow-layer CNNs (<10). The vibration signals of wheelset bearings are complex, and the fault signals are highly coupled with other irrelevant signals. Therefore, it is very promising to design a much deeper network to learn higher-layer and more abstract fault-related features.

III. PROPOSED Der-1DCNN-BASED FAULT DIAGNOSIS METHOD

In this section, the proposed Der-1DCNN-based fault diagnosis method is detailed, and its overall architecture is presented in Figure 1. To extract effective fault-related features, a Der-1DCNN constructs a much deeper 1D CNN to learn features of different levels from the raw signal. Specifically, a Der-1DCNN uses two wide convolutional layers and a deep residual module to learn low-level and high-level features, respectively. Second, this paper introduces a wide convolutional kernel and dropout into the proposed Der-1DCNN framework to improve the model's anti-noise ability. Finally, the extracted high-level features are put into a global average pooling [34] layer and a fully connected layer with softmax to obtain the classification result, thereby completing the fault diagnoses of wheelset bearings.

A. HIGH-LEVEL FEATURES LEARNING WITH STACKED 1D RESIDUAL BLOCKS

1) RESIDUAL LEARNING

The nonlinear layer composed of multiple stacked convolutional layers and activation layers can fit very complex nonlinear functions, which is one of the reasons that CNNs have powerful performance. Therefore, we attempt to construct a much deeper 1D CNN to diagnose the wheelset bearing fault of HSTs under complex working conditions. Nevertheless, the problems of training difficulty and performance degradation of the deeper CNN indicate that a deeper network cannot exert powerful learning ability. To address this problem, residual learning [17] is proposed. Its key idea is to convert complex function $L(x)$ fitted by multiple nonlinear layers into residual function ($R(x) = L(x) - x$), where x is the input to these layers. In addition, an input-to-output identity mapping is constructed through shortcut connections. Through the shortcut connections, the network enables the flow of information across layers without the attenuation that would stem from multiple stacked non-linear transformations, thereby improving the network's training speed. Obviously, it is easier to learn the residual function than to learn a new complex function. This can effectively exert the powerful learning ability of the deeper CNN and solve the problem of performance degradation. Based on residual learning, we will introduce the 1D residual block proposed in this paper and the Der-1DCNN architecture in the following section.

2) 1D RESIDUAL BLOCK

As shown in Figure 2, the 1D residual block has two branches. One is to fit the residual function via two 1D weight layers, and the other is to complete the identity mapping of the input signal through shortcut connections. The corresponding elements of two branches are added together, and then pass through the ReLU [35] nonlinear activation function to form the entire 1D residual block.

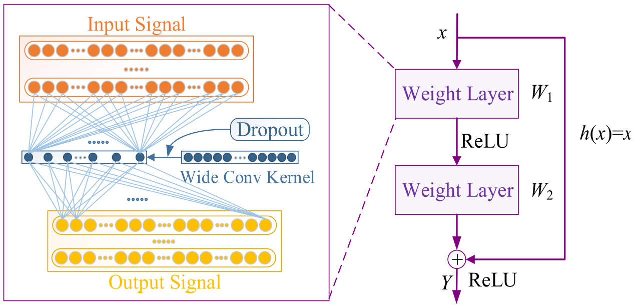


FIGURE 2. 1D residual block.

The input signal of the first weight layer (W_1) is output feature signal $x = \{x_1, x_2, \ldots, x_l\}$ of the previous layer, and the signal length is l. In this weighting layer, we slide the convolutional kernel with the size of $1\times k$ and a stride of s on the input feature signal, and we obtain output feature signal Y_1 accordingly. Output y_i of the ith node in feature signal Y_1 is defined as (1).

$$y_i = w^{\mathrm{T}} x_{i:i+k-1} + b_1, \tag{1}$$

where w is the convolutional kernel vector; b_1 denotes the bias term of the first weight layer; and $x_{i:i+k-1}$ is a k-length sub-signal of input signal x starting from the i-node.

Then, nonlinear output y'_i is obtained through ReLU nonlinear activation function f shown as (2).

$$y'_i = f(w^{\mathrm{T}} x_{i:i+k-1} + b_1), \tag{2}$$

where ReLU nonlinear activation function f is defined as (3).

$$f(z) = \begin{cases} z & z > 0 \\ 0 & z \leq 0 \end{cases} \tag{3}$$

Therefore, final output feature vector Y'_1 is defined as (4).

$$Y'_1 = \{y'_1, y'_2, \cdots, y'_m\}, \tag{4}$$

where m denotes the length of Y'_1. When $s = 1$, $m = l$, and when $s \geq 2$, $m = l/s$. After the first weight layer and the nonlinear activation function, we obtain output feature signal Y'_1, which is regarded as the input of the second weight layer (W_2). Repeat the above (1) operation, and obtain Y_2. Next, Y_2 is added to identity mapping $h(x)$ and then passes through the ReLU nonlinear activation function to obtain final output Y of the residual block shown in Eq. (5).

$$Y = f(Y_2 + h(x)) = f(W_2 Y'_1 + b_2 + h(x)) \tag{5}$$

Therefore, we assume that function $R(x)$ represents the residual function learned by the stacked weight layers. The following (6) and (7) are the definition of the 1D residual block.

$$R(x) = W_2 f(W_1 x + b_1) + b_2 \tag{6}$$

$$L(x) = f(h(x) + R(x)) \tag{7}$$

Obviously, data stream x can flow in the network without attenuation, and it is easier to learn residual function $R(x)$ than to learn the new complex function $L(x)$. Therefore, the 1D residual block can effectively exert the powerful learning ability of the deeper CNN and solve the problem of performance degradation.

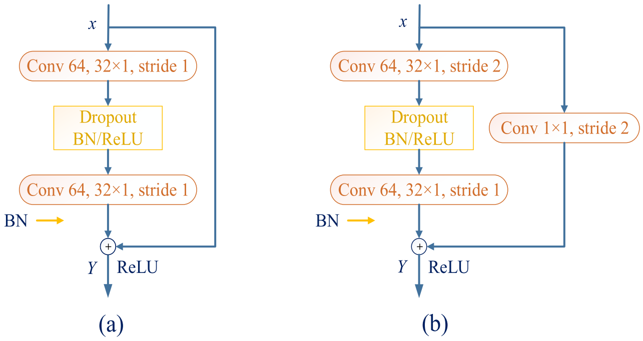


FIGURE 3. Architecture of 1D residual block. (a) Identity-block. (b) Down-block.

Figure 3 is the architecture of the 1D residual block. Specially, 1D residual block "Identity-block" includes two convolution operations, where the wide convolutional kernel has the size of $1\times k$ and a stride of 1. Batch normalization (BN) [36] and dropout [19] are used to improve the performance of the 1D residual block, as shown in Figure 3(a). However, the input and output data streams of "Identity-block" must have the same dimensions; otherwise, the addition operation of two data streams cannot be done. If this is not the case, 1D residual block "Down-block" shown in Figure 3(b) is adopted, where the linear projection of the input data stream is performed via a convolution operation with a 1×1 convolutional kernel and a stride of 2 on the shortcut connections to match the dimensions.

3) THE DEEP RESIDUAL MODULE OF DER-1DCNN

In this section, we introduce the deep residual module of a Der-1DCNN, which is composed of stacked 1D residual blocks for high-level fault feature extraction. To effectively learn fault features from the raw signals measured under variable loads, variable speeds and a strong noise environment, the key idea of this module is to construct a much deeper 1D network to learn and abstract the signal features layer by layer. Finally, the high-level feature information is extracted. Therefore, we build a deep residual module via stacked 1D residual blocks. The module has the following three advantages: 1) powerful learning ability: it can effectively exert the powerful feature learning ability of the much deeper network.

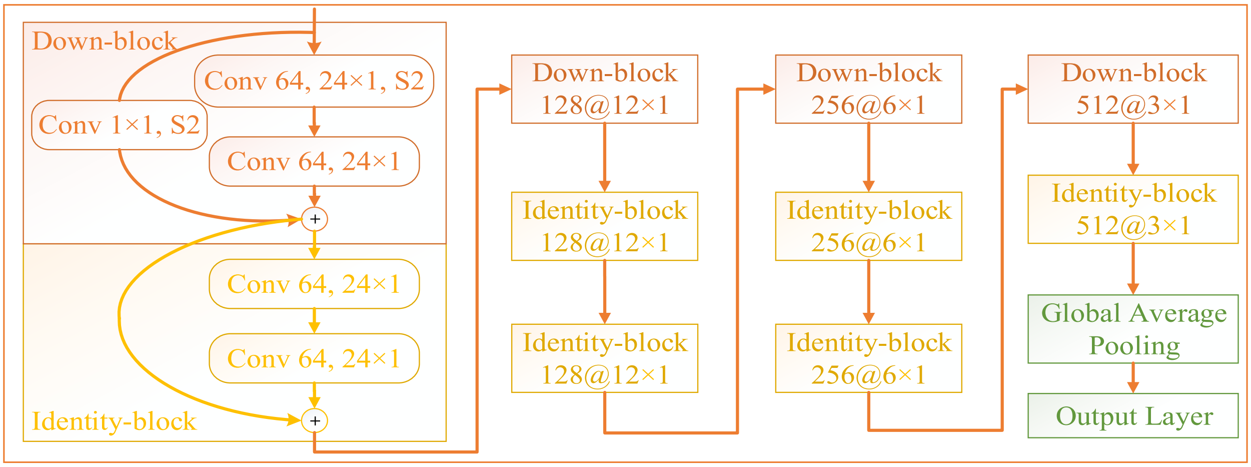


FIGURE 4. Deep residual module with stacked 1D residual blocks.

2) Convenience: by simply stacking the 1D residual blocks, we can obtain a deeper network with a powerful feature learning ability. 3) Adaptability: the deep residual module is universal and flexible. It may have different depths and different convolutional kernel sizes. For example, we can construct 30-layer, 50-layer or even much deeper 1D CNNs based on different signal lengths and dataset sizes.

Figure 4 shows an architecture of a 24-layer Der-1DCNN with 10 stacked 1D residual blocks, where two shallow convolutional layers are not shown. It can be seen that building the Der-1DCNN framework follows the following simple design rules: 1) the network depth is adjusted by the number of 1D residual blocks. 2) For the same output feature signal length, the convolutional layers have the same number of convolutional kernels. 3) If the input feature signal length of the convolutional layers is halved, the number of convolutional kernels is doubled, while the size of the wide convolutional kernels is halved. 4) Down-sampling is performed in ''Down-block'' directly via convolutional layers with a stride of 2. As shown in Figure 4, S2 represents a stride of 2. 5) The residual learning network ends with a global average pooling layer and a fully connected layer with softmax. It is worth noting that the parameter settings of the network are not fixed and can be adjusted according to actual working conditions. For example, to better extract the features of complex signals, the number of network layers or the size of wide convolutional kernel can be increased, but the overall design principles of the Der-1DCNN framework need to follow the above rules.

B. CONVOLUTIONAL LAYER WITH WIDE CONVOLUTIONAL KERNEL

The convolutional layer is the core of a CNN, consisting of multiple convolutional kernels. It performs the convolution operation on the input signal to produce an output to the next layer. Suppose that w is a convolutional kernel vector with size $1\times k$, and x is the discrete vibration signal. Let $X \cdot W$ be the result of 1D discrete convolution. The ith element of the result is given by (8).

$$(X \cdot W)[i] = \sum_{n=1}^{k} w_{k-n+1} \cdot x_{i+n-1} \tag{8}$$

It can be seen that, for different convolutional kernels, the convolution operation is capable of extracting different insightful information from the raw signal. If the 1×3 narrow convolutional kernel used in the traditional 1D-CNN is adopted, each output feature value can obtain only the feature relationship among the adjacent three values of the input signal, which will greatly limit the network's ability to learn low-frequency signal features. However, the introduction of a wide convolutional kernel allows one convolution operation to obtain the feature relationship in a longer sequence. Through the learning of multi-layer convolutions, the network can extract better low-frequency fault-related features and suppress the noise interference. However, a wide convolutional kernel increases network parameters, which is not beneficial for making the network deeper. Hence, we adopt the 1×48 wide convolutional kernel in the first convolutional layer, and for the subsequent convolutional layer, the width of the convolutional kernel is gradually reduced.

Another advantage of a wide convolutional kernel is that the network has fewer parameters and computations when the receptive field is same. In a 1D-CNN, the receptive field refers to the length of the region where the signal point of the feature signal in each layer is mapped on the original signal. The calculation formula is in (9),

$$RF_i = (RF_{i+1} - 1) \times s_i + k_i, \tag{9}$$

where RF_i is the receptive field of the ith convolutional layer; RF_{i+1} is the receptive field of the $(i+1)$th layer; s is the stride of the convolution operation; and k is the size of the convolutional kernel. For a 2D-CNN, three stacked 3×3 convolutional layers can use only 27 weighting parameters to obtain the same receptive field as a 7×7 convolutional layer. However, for a 1D-CNN, five stacked 1×3 convolutional layers use 15 weighting parameters to obtain the same receptive field as a 1×11 convolutional layer. Therefore, it is unwise for the 1D signals processing to use a narrow convolutional kernel. In this paper, we adopt a wide convolutional kernel to learn signal feature, and we adjust the size of the convolutional kernel according to the specific task.

C. 1D RESIDUAL BLOCK WITH DROPOUT

In this paper, the dropout is introduced into the first layer convolution operation of the proposed 1D residual block. It means that the convolutional kernel is randomly dropped out via probability p during training. To ensure the validity of the wide convolutional kernel of each layer, dropout rate p decreases as the size of a wide convolutional kernel decreases. Applying dropout to a neural network amounts to sampling a "thinned" network from it. A neural net with R units can be seen as a collection of 2^R possible thinned neural networks [19]. Thus, the dropout applied to a wide convolutional kernel can train many "thinned" networks, and through the integration of these "thinned" models, the network can learn more robust fault features. Therefore, we adopt wide convolutional kernels to further make dropout more efficient. Additionally, the introduction of dropout is equivalent to adding random noise to the input feature signal, which will improve the network's anti-noise ability. The dropout is expressed in (10) and (11).

$$y_i = r \cdot w^T x_{i:i+k-1} + b_1, \tag{10}$$

$$y_i' = f(y_i), \tag{11}$$

where r follows Bernoulli distribution as shown in Eq.(12), which is used to decide whether the convolutional kernel is dropped out. Therefore, residual function $R(x)$ is expressed by Eq. (13).

$$r \sim Bernoulli(p) \tag{12}$$

$$R(x) = W_2 f(r \cdot W_1 x + b_1) + b_2 \tag{13}$$

D. CLASSIFICATION

1) GLOBAL AVERAGE POOLING

In the Der-1DCNN model, global average pooling is adopted to replace the traditional fully connected layers in a CNN. It can effectively avoid the overfitting problem that easily occurs in the fully connected layers, and thus improve the network's generalization ability. It takes the average of each feature signal, and the resulting vector is fed directly into the softmax layer. Compared with the fully connected layer, global average pooling is more native to the convolution structure by enforcing correspondences between feature signals and categories. Moreover, there is no parameter to optimize in the global average pooling; thus, overfitting is avoided in this layer.

2) SOFTMAX AND LOSS FUNCTION

Obviously, the wheelset bearing fault diagnosis is a multi-classification task, so the softmax activation function is applied. It maps the output of multiple neurons to the range of (0, 1) and sums up to 1, so it is generally used as the classifier to estimate the probability distribution belonging to different classes. We assume that K is the total number of different health conditions (12 in this paper). The softmax function is expressed in (14).

$$Q_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_j}}, \tag{14}$$

where z_j is the jth input feature of the softmax activation function, and $Q_j(z)$ is the estimated probability distribution of observation z belonging to the jth class.

Then, cross-entropy loss function [37] is adopted to evaluate the error of the estimated softmax output probability distribution and the target class probability distribution. Suppose that $P(z)$ is the target distribution and that $Q(z)$ is the estimated distribution; the cross-entropy between $P(z)$ and $Q(z)$ is expressed as (15).

$$Loss = E(P(z), Q(z)) = -\sum_{j=1}^{K} P_j(z) log(Q_j(z)) \tag{15}$$

Finally, the Adam [38] optimization algorithm is used to reduce the value of the cross-entropy loss function during training, so the estimated distribution and the target distribution draw closer and closer, thereby gradually improving the model's prediction accuracy.

E. THE FAULT DIAGNOSIS METHOD BASED ON Der-1DCNN

In this section, an end-to-end wheelset bearing fault diagnosis method of HSTs based on the proposed Der-1DCNN framework is presented. The flow chart of the fault diagnosis method is shown in Figure 5, and its general operation processes are summarized as follows.

a) First, the vibration signals of bearings under various working conditions and different faults are collected through multiple acceleration sensors installed on the axle box of an HST. Then, each vibration signal is segmented into small segments by the data expansion method, thereby obtaining training samples and testing samples of the model.

b) According to the size of the sample, select the appropriate network depth to avoid the problem of wasting computing resources.

c) The raw vibration signals of training samples are regarded as the input of the Der-1DCNN model. In addition, the Adam [38] algorithm is employed to optimize all network parameters to complete high-level feature extraction and fault classification, so the end-to-end wheelset bearing fault diagnosis model based on a Der-1DCNN is obtained.

d) Input the testing samples to the well-trained fault diagnosis method to automatically extract high-level features and

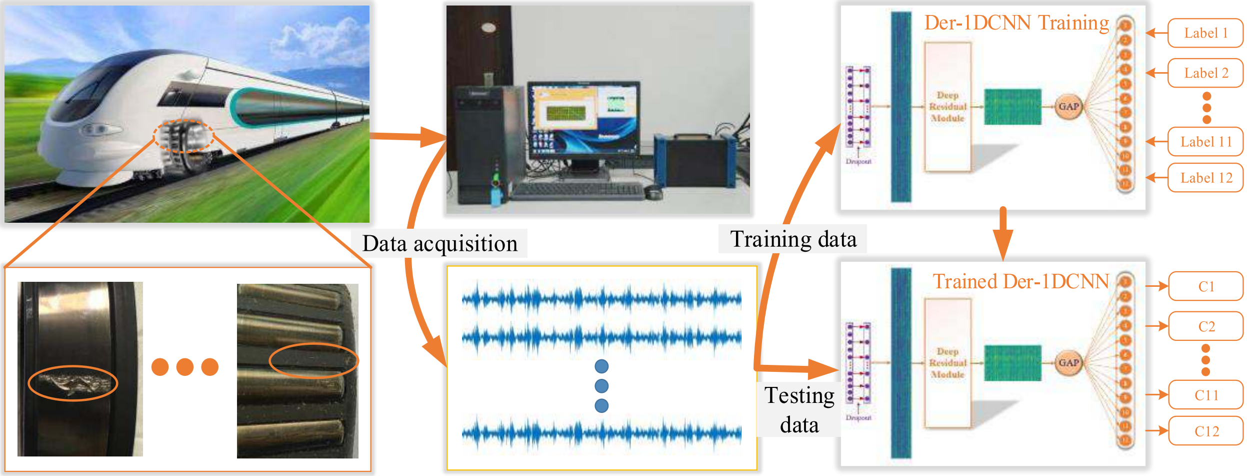


FIGURE 5. The flow chart of the wheelset bearing fault diagnosis system.

directly diagnose the health conditions of wheelset bearings in HSTs.

IV. EXPERIMENTAL VERIFICATION OF PROPOSED Der-1DCNN

In the real operation of HSTs, the working conditions of wheelset bearings vary greatly. First, the generation of random noise is inevitable, and the fault-related signal is easily overwhelmed by high-intension environment noise. Therefore, it is very important and challenging to have the ability to perform high-precision fault diagnosis in a noise environment. Second, the working load may constantly change, and the signal features change accordingly. However, it is unrealistic to collect and label enough training samples. If the fault diagnosis method has the ability to classify samples of other loads by learning training samples under the existing load, it can greatly improve the efficiency and applicability of the diagnosis method. In this section, we verify the merits of the proposed model from these two aspects of the HST test rig.

A. DATA DESCRIPTION

The experimental data come from the wheelset bearing test rigs of HSTs located in the Qingdao Sifang Institute. The vibration signals are collected via different accelerometers mounted on the axle boxes of HSTs. As shown in Figure 6, the wheelset bearing test rig is mainly composed of a drive motor, a belt transmission system, a vertical loading set, a lateral loading set, two fan motors and a control system. The vertical and lateral loading sets are designed to mimic two-dimensional loads in real train operation. The fan motor can generate wind that is opposite of the train's running direction. An axle and its two supporting bearings are assembled to the test rig. Vibration signals sampled at 5120 Hz are collected via two accelerometers that ensure that horizontal and vertical movements are measured in the wheelset bearing.

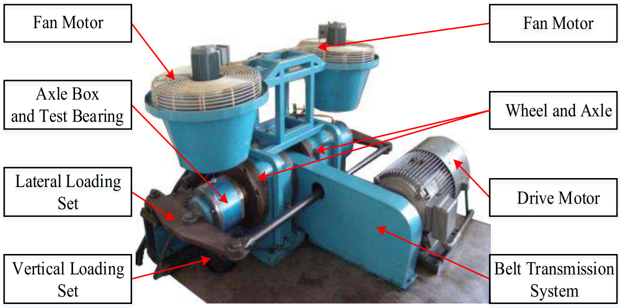


FIGURE 6. The wheelset bearing test rig.

For the sake of verifying the merits of the proposed Der-1DCNN model, 12 different health conditions of wheelset bearings are set in consideration of multiple and mixed fault patterns. The fault information with respect to the test bearings is listed in Table 1, where the labels are C1, C2, C3, ... and C12, respectively. To simulate the complex and changing working conditions of HSTs during their operation as much as possible, under each health condition, five running speeds are designed: 60, 90, 120, 150 and 180 km/h, and four different vertical loads are conducted: 56, 146, 236 and 272 kN, and also two lateral loads: 0 and 20 kN are set. Therefore, each health condition includes 40 kinds of working conditions. After the raw signal is extended through the data augmentation technique, there is a total of 329,752 samples. The samples are randomly divided into training samples and testing samples. A total of 284,260 samples are training samples, and 45,492 samples are testing samples.

B. EXPERIMENTAL SETUP

1) DATA AUGMENTATION

Fault diagnosis based on deep learning often requires a large number of data samples, so the data augmentation technique is very meaningful for the training of deeper CNNs. We next

TABLE 1. Description of twelve health conditions.

Location	Fault Description	Class Label
None	Normal	C1
Inner race	Pitting	C2
Rolling element	Pitting	C3
Rolling element	Flaking with a size of 3mm×35mm	C4
Inner race	Flaking with a size of 3mm×45mm	C5
Rolling element	Cracking	C6
Outer race and rolling element	Mixed fault with outer race flaking and rolling element pitting, and the flaking size is 10mm×45mm	C7
Inner race	Flaking with a size of 10mm×45mm	C8
Outer race	Flaking with a size of 10mm×30mm	C9
Rolling element	Flaking with a size of 1mm×1mm	C10
Cage	Cracking	C11
Outer race	Flaking with a size of 10mm×45mm	C12

introduce a sliding segmentation approach to expand the original data. Suppose that we have a collected vibration sequence $x[n]$ with n sample points. Then, we define another three parameters: $L_{overlap}$ = the length of sample overlap for two neighbor segments; L_{seg} = the length of each segment; and N = the number of segments. Their relationships are defined in (16). Once the aforementioned four parameters $(n, L_{overlap}, L_{seg}, N)$ are specified, the whole signal $x[n]$ can be split into N segmentation signal $\{x_1[n], x_2[n], ..., x_N[n]\}$ shown in Figure 7. In this paper, $L_{overlap}$ and L_{seg} are typically defined as 1920 and 2048, respectively, which can guarantee that each segmentation signal contains at least one period of vibration. For example, a vibration signal with 51,200 sample points can obtain 385 training samples.

$$n = (N - 1) \times (L_{seg} - L_{overlap}) + L_{seg} \tag{16}$$

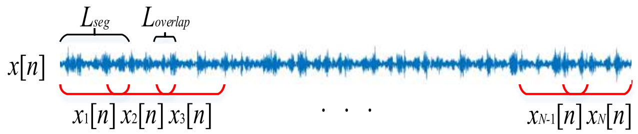


FIGURE 7. Data augmentation.

2) BASELINE SYSTEM

To validate the effectiveness and superiority of the proposed Der-1DCNN, the following four state-of-the-art methods are regarded as the benchmark methods.

a) Wen-CNN [33]: Through converting the raw vibration signals into 2D images, this method designed a new CNN based on LeNet-5 to diagnose the fault of motor bearings.

b) MSCNN [32]: The method introduced a new multi-scale CNN structure for multiscale feature extraction and classification. In this paper, MSCNN applies three multi-scale branches.

c) WDCNN [30]: This method proposed a CNN for the bearing fault diagnosis, and it applied wide convolutional kernels in the first convolutional layer for extracting features and suppressing high-frequency noise.

d) ADCNN [27]: The method converted the 1D signal into two dimensions and then adopted an adaptive deep CNN to diagnose bearing faults and determine their severity.

In each experiment, we use the same training samples and testing samples, and 80 epochs are trained in the same training strategy for these four benchmark methods and the Der-1DCNN. Because the data length of each sample is 2048 and the number of health conditions for the wheelset bearings is 12, the input and output dimensions of these benchmark methods have to be modified accordingly.

3) IMPLEMENTATION DETAILS

We implement the proposed Der-1DCNN using the Keras library and Python 3.5. Network training and testing are performed on a workstation with an Ubuntu 16.04 operating system, an Intel Core i7-6850K central processing unit, 32GB random access memory and a GTX 1080Ti graphics processing unit. To accelerate the convergence speed of the network, each original signal x_o is normalized using the z-score standardization method, which can be expressed in (17).

$$x = \frac{x_o - \mu}{\sigma}, \tag{17}$$

where μ is the mean of the sample data and σ is the standard deviation of the sample data. Finally, during the training, we adopt the cross-entropy loss function and Adam optimization algorithm with a learn rate of 0.0001 and batch size 96.

4) PERFORMANCE METRICS

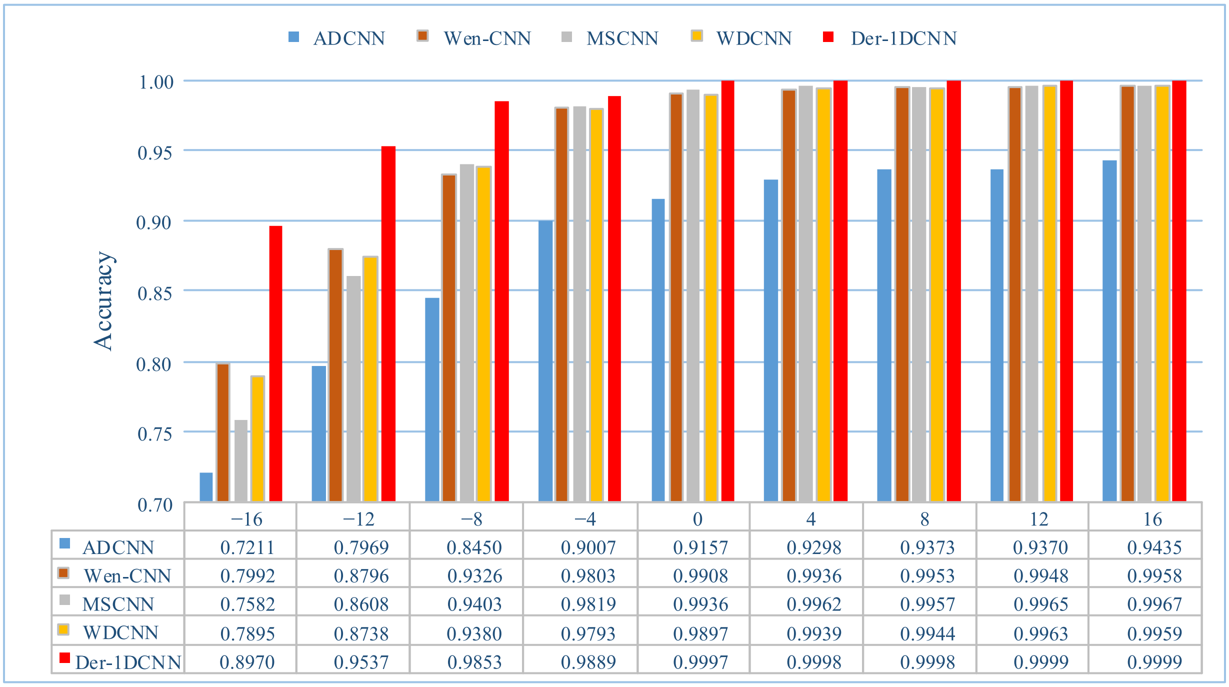
In this paper, we adopt the evaluation indicator: accuracy. It is a commonly comprehensive metric that measures the performance of classification methods. This metric is defined in Eq. (18).

$$accuracy = \frac{TP + TN}{V + T}, \tag{18}$$

where TP (true positive) is correctly classified as positive samples, FP (false positive) is misclassified as positive samples, TN (true negative) is correctly classified as negative samples, FN (false negative) is misclassified as negative samples, V is the number of positive samples shown in (19) and T is the number of negative samples shown in (20) .

$$V = TP + FN \tag{19}$$

$$T = FP + TN \tag{20}$$



	−16	−12	−8	−4	0	4	8	12	16
ADCNN	0.7211	0.7969	0.8450	0.9007	0.9157	0.9298	0.9373	0.9370	0.9435
Wen-CNN	0.7992	0.8796	0.9326	0.9803	0.9908	0.9936	0.9953	0.9948	0.9958
MSCNN	0.7582	0.8608	0.9403	0.9819	0.9936	0.9962	0.9957	0.9965	0.9967
WDCNN	0.7895	0.8738	0.9380	0.9793	0.9897	0.9939	0.9944	0.9963	0.9959
Der-1DCNN	0.8970	0.9537	0.9853	0.9889	0.9997	0.9998	0.9998	0.9999	0.9999

FIGURE 8. Performance of Der-1DCNN and four benchmark methods in noise environment (SNR = −16 dB).

C. CASE STUDY I: ROBUSTNESS AGAINST NOISE

The experimental data are collected under different working conditions and different health conditions of bearings on the HST test rig, so the original vibration signal already contains certain noise. To better simulate the strong noise interference in the real operation of trains, we add white Gaussian noise to the original signals with different signal noise ratios (SNRs). The definition of SNR is shown as (21).

$$\mathrm{SNR} = 10log_{10}\left(\frac{P_{signal}}{P_{noise}}\right), \tag{21}$$

where P_{signal} and P_{noise} are the power of signal and noise respectively, and we assume that P_{signal} is 0 dBW.

In this experiment, we will verify the effectiveness of the proposed Der-1DCNN in different noise environments with the SNRs ranging from −16 dB to 20 dB. The experimental results are shown in Figure 8. Obviously, the Der-1DCNN is superior to the other four benchmark methods and achieves the best diagnostic performance in any noise environment. Moreover, the model has more than 95% diagnostic performance under all noise levels, except for 89.7% accuracy at −16 dB. When the SNR is large, the Wen-CNN, MSCNN and WDCNN achieve similar diagnostic performance levels, and the accuracy does not increase as the SNR increases. This indicates that the diagnosis error is mainly due to the similarity of the fault feature itself when the noise is small. That is to say, under different working conditions, some different fault signals may have similar features, leading to misjudgment. However, the accuracy of the Der-1DCNN is close to 100%, which means that the Der-1DCNN has better fault feature learning ability and recognition ability, and it can extract the most essential differences among various fault features. On the other hand, although the performance of all methods decreased with the increase of noise, the Der-1DCNN still exhibits excellent anti-noise ability in a strong noise environment. Specially, the Der-1DCNN achieves nearly 90% diagnostic performance at SNR = −16 dB, which is a nearly 10% improvement over the Wen-CNN, with the best diagnostic performance in four benchmark methods. Therefore, the Der-1DCNN has stronger anti-noise ability and fault feature learning ability than the traditional CNN in a strong noise environment. Therefore, the Der-1DCNN model is robust to noise without any additional denoising preprocessing, and it is more suitable for wheelset bearing fault diagnosis in the real operation of HSTs.

To clearly present the anti-noise ability of the Der-1DCNN in a strong noise environment, we use the t-SNE [39] technique to visualize the feature distributions that these five methods ultimately learned in a 2D space. The result is shown in Figure 9, where different colors represent the different health conditions of wheelset bearings. It can be seen that the high-level fault features that the Der-1DCNN learned have the best discrimination, which indicates that the Der-1DCNN can learn more distinguishable fault features from complex vibration signals.

D. CASE STUDY II: DOMAIN ADAPTATION OF VARIOUS WORKING LOADS AND SPEEDS

In this experiment, we first verify the domain adaptation ability of the Der-1DCNN under different working loads.

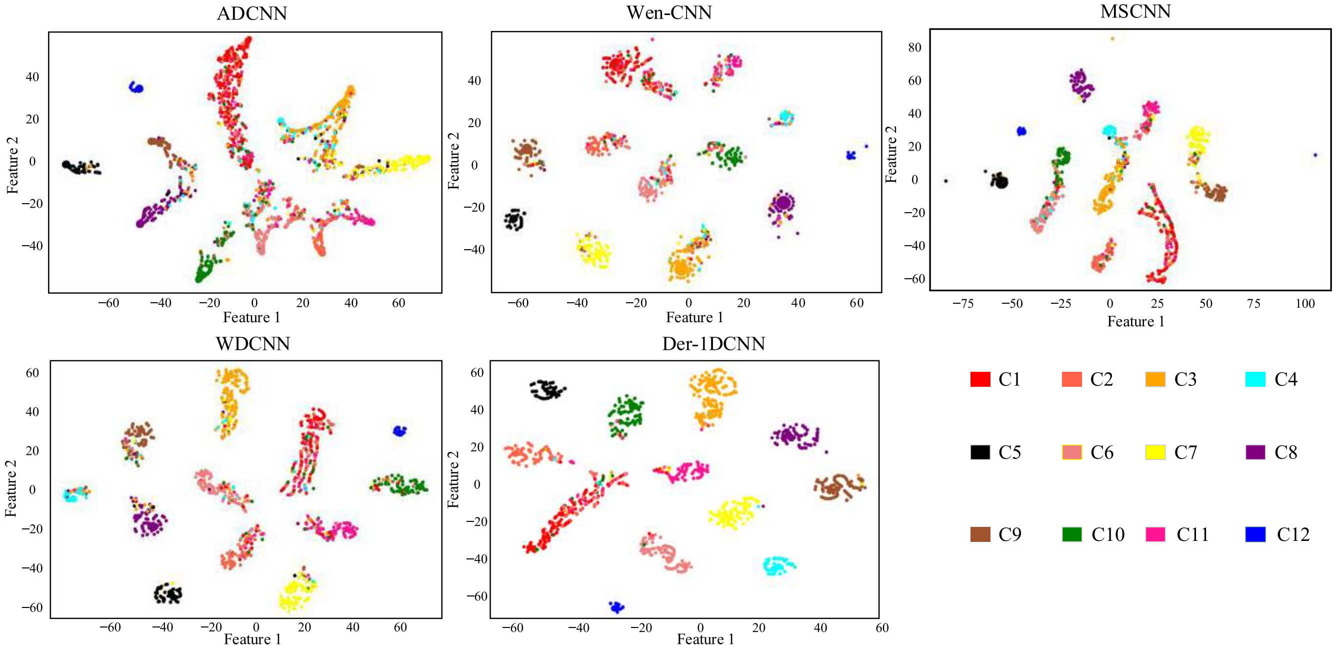
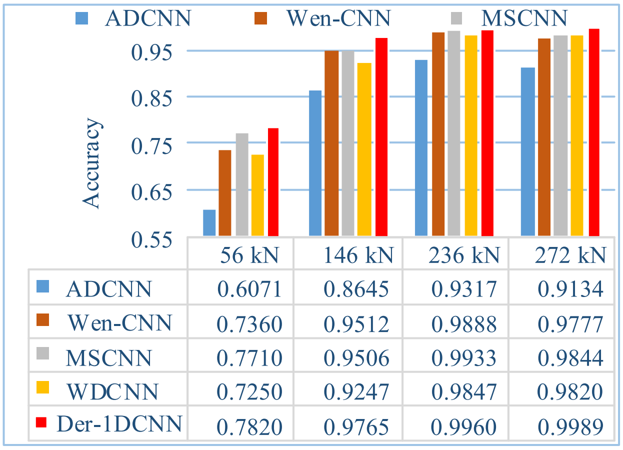


FIGURE 9. Visualization of these five methods in noise environment (SNR = −16 dB).



	56 kN	146 kN	236 kN	272 kN
ADCNN	0.6071	0.8645	0.9317	0.9134
Wen-CNN	0.7360	0.9512	0.9888	0.9777
MSCNN	0.7710	0.9506	0.9933	0.9844
WDCNN	0.7250	0.9247	0.9847	0.9820
Der-1DCNN	0.7820	0.9765	0.9960	0.9989

FIGURE 10. Performance of Der-1DCNN and four benchmark methods under different working loads.

First, we select the data samples under four vertical loads as a dataset, including 56, 146, 236 and 272 kN. Then, we take one kind of load data as a testing sample, and three other load data as training samples, so we obtain four sets of experimental data. For example, data with the load of 56 kN is employed as testing samples, and the data of other load conditions (146, 236 and 272 kN) are regarded as training samples.

The experimental results are shown in Figure 10. Clearly, the Der-1DCNN shows the strongest domain adaptation ability under four load conditions, indicating that the Der-1DCNN has quite superior diagnostic performance under the changing working loads without any domain adaptation algorithm processing. From the trend of performance for each method under different load conditions, it can be found that the smaller the load, the worse the diagnosis performance. The reason for this is that the smaller the load, the weaker the corresponding fault feature, and the strong fault features learned under huge load conditions cannot be effectively adapted to the identification of weak features. Therefore, the superior performance of the Der-1DCNN under small load conditions can better reflect the generalization of the high-level features that the Der-1DCNN extracted. On the other hand, as the load increases, the noise that the mechanical system itself generates also increases, so the accuracy of these four benchmark methods decreases when the load is 272 kN. In contrast, the accuracy of the Der-1DCNN does not decline but rather is closer to 100%. This implies that the Der-1DCNN model is more promising for domain adaptation diagnosis under changing load conditions and in changing noise environments.

In addition, we try to apply the proposed method and comparison methods to the task of wheelset bearing speed domain adaptation. Similarly, we use the data at a certain speed as a testing set, and the data at other speeds as a training set. Table 2 shows the fault diagnosis results of

TABLE 2. The speed domain adaptation results of Der-1DCNN and four benchmark methods.

Method Accuracy Speed	ADCNN	Wen-CNN	MSCNN	WDCNN	Der-1DCNN
60 km/h	0.4847	0.4222	0.5352	0.4847	***0.6605***
90 km/h	0.4537	0.5427	***0.7726***	0.6429	0.7406
120 km/h	0.4718	0.6302	***0.8242***	0.6413	0.7634

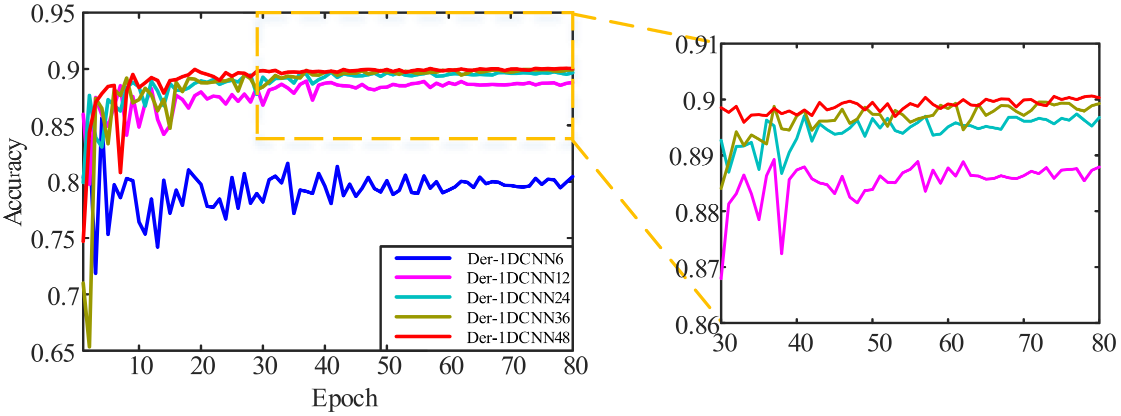


FIGURE 11. Performance of the Der-1DCNN model under different network depths.

each method for the testing dataset at 60, 90 and 120 km/h. As can be seen, the diagnostic accuracy improves with the increasing speed. This is because in a certain number of sampling points, the faster the speed, the more fault impact components it contains, so the higher the diagnosis accuracy. Additionally, it is difficult for the existing CNN methods to get satisfactory results about the speed domain adaptation task. Relatively speaking, the MSCNN method shows good domain adaptability, especially at the speed of 120 km/h. At different speeds, the number of sampled points per signal period is different, and the fault characteristic frequency is also different. Therefore, compared with the load domain adaptation task, the spatial distribution of sample features varies greatly at different speeds, which makes it difficult for CNNs to learn the feature correlation between the same fault samples at different speeds and to classify them into one category. Because of this, CNNs perform poorly while the MSCNN achieves better diagnostic accuracy due to its multi-scale operation that is the sample resampling to reduce the differences of spatial distribution of sample features among different speeds. Therefore, we can introduce the idea of multi-scale learning into our network to further enhance the model's performance.

V. DISCUSSIONS

In this section, the validity of the proposed Der-1DCNN model is discussed and investigated through three experiments. First, we explore the effects of network depth on network performance to prove that a deeper network enables higher-level learning and more abstract features. Then, we evaluate the importance of residual learning in network performance through the comparison experiments of the Der-1DCNN and Der-1DCNN without residual learning. Finally, we compare the Der-1DCNN with the Der-1DCNN without a wide convolutional kernel and/or dropout to illustrate the advantages of a wide convolutional kernel and dropout.

A. DISCUSSIONS ON EFFECTS OF NETWORK DEPTH

The proposed Der-1DCNN model can adjust the network depth simply by increasing or decreasing the number of 1D residual blocks. The network depth has a considerable influence on the feature abstraction level. Moreover, the low-level features obtained from the original signal may be affected by the changing speeds or loads and environment noise. Therefore, the abstraction levels of features can significantly affect the classification results. In this section, we study the effect of network depth on the model's performance. Five depths of the Der-1DCNN model, namely 6, 12, 24, 36 and 48 layers, are defined, and they are called the Der-1DCNN6, Der-1DCNN12, Der-1DCNN24, Der-1DCNN36 and Der-1DCNN48, respectively. The experiment is carried out under the noise of −16 dB, and the network depth is determined by the number of 1D residual blocks proposed in this paper.

The testing results of each epoch for these five networks are shown in Figure 11. Obviously, the diagnostic performance of the Der-1DCNN increases as the network depth increases. Among them, the accuracy of the Der-1DCNN24 has a nearly 10% improvement compared with one of the Der-1DCNN6. This is because the Der-1DCNN model can learn and extract more abstract and robust fault features at a higher level, thus enabling the network to accurately distinguish fault features and other useless features. Additionally, it can be seen that the Der-1DCNN36 and Der-1DCNN48 have similar performance (about 90%) in terms of classification accuracy, and about 0.5% improvement compared with that of the Der-1DCNN24. However, the Der-1DCNN36 and Der-1DCNN48 have more parameters and consume more computing resources. Hence, for the dataset of this paper, we adopt the 24-layer Der-1DCNN model to diagnose the wheelset bearing fault of HSTs in this paper.

The above phenomenon indicates that if the number of training samples is specific, the increase of diagnostic performance will become slower and slower as the network depth increases. Because the diagnostic performance always increases, the high-level feature extraction ability of a deeper

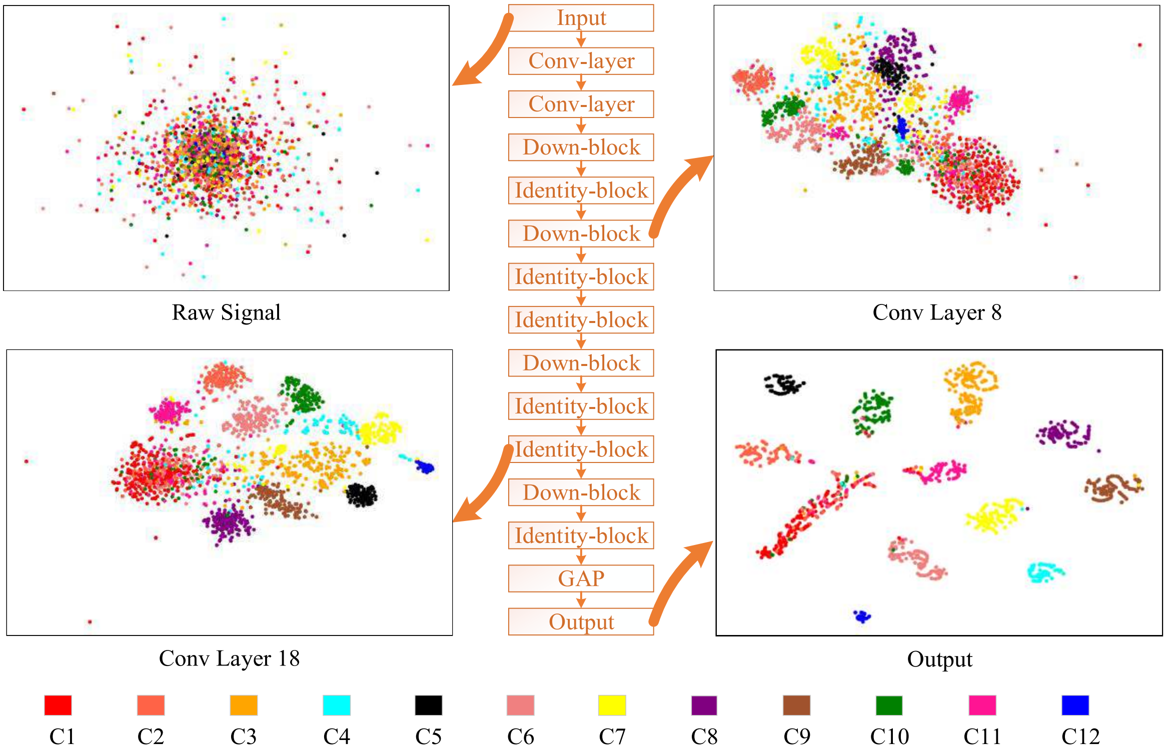


FIGURE 12. Visualization of different layer of the proposed Der-1DCNN.

network also increases, and thus, the phenomenon is not the result of the feature extraction ability. For the dataset of this paper, when network parameters reach a certain value as the network depth increases, these parameters cannot be fully optimized to better local optimal values, and thus, the phenomenon stems from the limitation of sample size. In summary, for practical engineering applications, to deal with more challenging and complex diagnostic tasks, especially when a larger amount of data is available, the diagnostic performance can be further improved by building the Der-1DCNN model.

To more clearly show the effect of depth on diagnostic performance, we apply the t-SNE [39] technique to provide a 2D representation of the output features at different levels for the Der-1DCNN24 model adopted in this paper. The results are shown in Figure 12, where different colors represent different health conditions. As can be seen, the distribution of the various health conditions of the raw signal is very turbulent, but as the network layer increases, the feature distribution of 12 types of health conditions is gradually separated. To be specific, as the network layer increases, the network can extract more abstract and higher-level fault features, making it easier to distinguish between different health conditions.

B. EVALUATION OF RESIDUAL LEARNING

Residual learning completes the identity mapping of input by using shortcut connections so as to make the stacked nonlinear layers learn the simple residual function, and then improve the training speed and accuracy of the deeper CNN. Multiple stacked 1D residual blocks are introduced to construct the much deeper 1D CNN to suppress the problems of training difficulty and performance degradation. To evaluate the effect of the residual learning of the Der-1DCNN on model performance, under noise of −16 dB, we test the 48-layer Der-1DCNN (Der-1DCNN48) and the 48-layer Der-1DCNN without residual learning (CNN48), and we record the training accuracy, testing accuracy and training loss of these two models. It is worth noting that the CNN48 lacks only the function of residual learning, and to be specific, it does not have shortcut connections to complete the identity mapping of the input. For other parameters, such as convolutional kernel size and number, dropout rate, etc., the CNN48 and Der-1DCNN48 are the same.

The training results are shown in Figure 13. From Figure 13 (a), we can clearly see that the testing accuracy of the Der-1DCNN48 has increased by an average of 3.5% compared with the CNN48. More importantly, the testing accuracy of any epoch of the Der-1DCNN48 is higher than that of the CNN48. This implies that the network degradation problem is well solved in the Der-1DCNN48 model. Comparing Figure 13(b) and Figure 13(a), obviously, the training accuracy of these two models is higher than the testing accuracy, indicating that both networks have the overfitting problem in a strong noise environment (−16 dB). However, the training

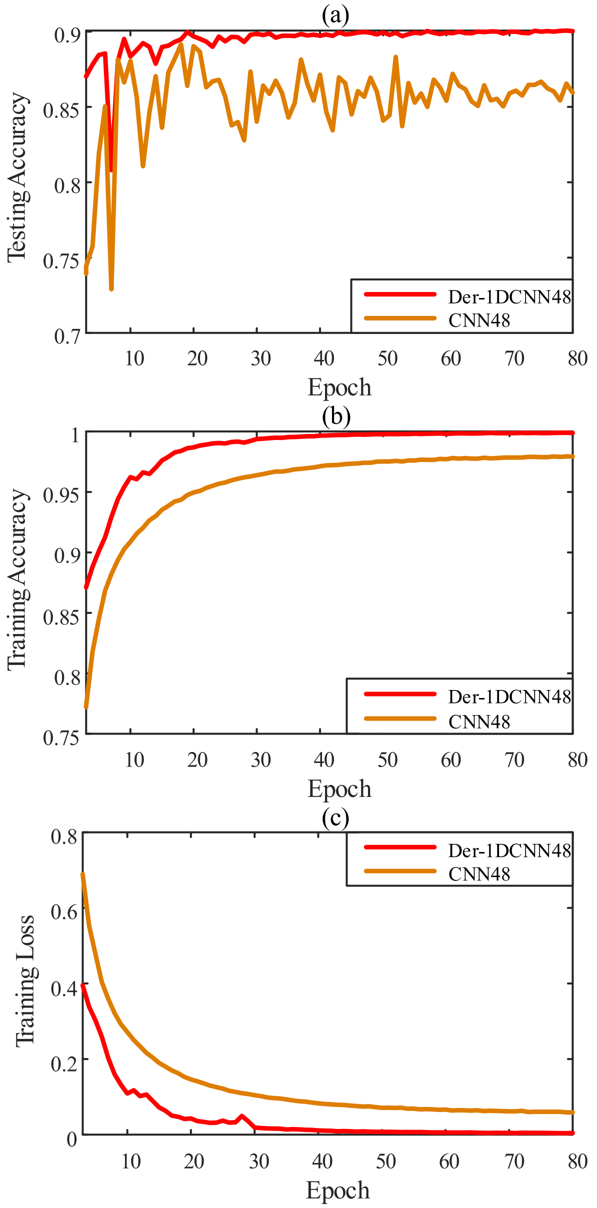


FIGURE 13. Performance of Der-1DCNN48 (with residual learning) and CNN48 (without residual learning): (a) Testing accuracy; (b) Training accuracy; (c) Training loss.

accuracy and testing accuracy of the Der-1DCNN48 are significantly higher than those of the CNN48. This means that the Der-1DCNN48 not only extracts high-level fault features more effectively but also has better generalization. Figure 13(c) illustrates the change in the training loss of these two models. It is worth noticing that compared with the CNN48, the training loss of the Der-1DCNN48 decreases more rapidly, and the final loss is closer to zero, so the Der-1DCNN48 converges quickly and well during the training process. This further proves that residual learning can effectively solve the problem of training difficulty and performance degradation for a deeper network and enable the network to find better local optimal solutions.

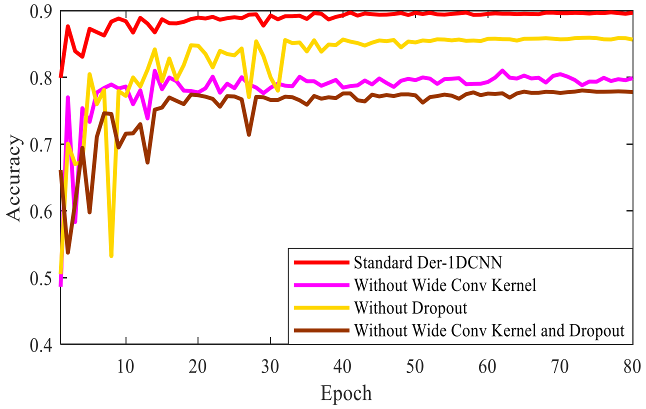


FIGURE 14. Performance of standard Der-1DCNN and one without wide convolutional (Conv) kernel, without dropout, and without wide convolutional kernel and dropout.

TABLE 3. Testing time of a sample for these five methods.

Method	ADCNN	Wen-CNN	MSCNN	WDCNN	Der-1DCNN
Testing time / ms	0.116	0.144	1.000	0.336	1.000

C. DISCUSSIONS ON EFFECTS OF NETWORK DEPTH

To evaluate the influence of a wide convolutional kernel and dropout on the performance of the Der-1DCNN model, we do the comparison experiments of the standard Der-1DCNN, a Der-1DCNN without a wide convolutional kernel, a Der-1DCNN without dropout and a Der-1DCNN without a wide convolutional kernel and dropout under noise of −16 dB. The diagnosis results are recorded during the entire training process and are shown in Figure 14. It can be seen that in the case of the absence of a wide convolutional kernel, the accuracy is far lower than that of the standard Der-1DCNN, and the final diagnostic result is about 10% lower than that of the standard Der-1DCNN. Therefore, in a strong noise environment, the introduction of a wide convolutional kernel can make the network effectively learn the low-frequency fault-related features. In addition, the final diagnostic result of the standard Der-1DCNN is about 4% higher than that of a Der-1DCNN without dropout, which indicates that the application of dropout to the convolutional layer can also effectively improve the model's performance in a complex noise environment. Finally, in the case of the absence of dropout and a wide convolutional kernel, the model exhibits the worst performance, demonstrating that the combination of a wide convolutional kernel and dropout can further improve the model's performance.

D. DISCUSSIONS ON THE MODEL'S STABILITY AND COMPLEXITY

The model's stability and complexity are very important evaluation indicators in practical engineering applications. Therefore, in this section, we further evaluate the model's stability by repeating many experiments, and we also evaluate

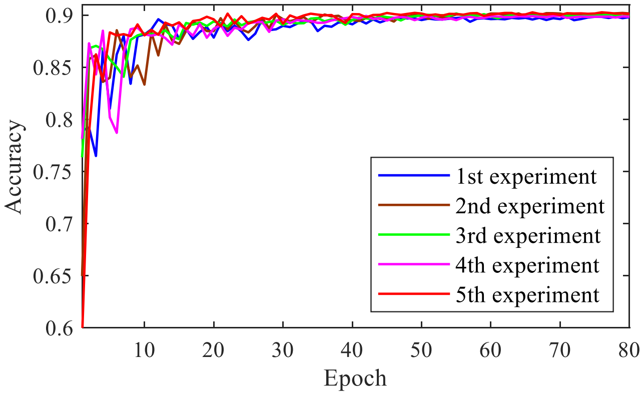


FIGURE 15. The stability evaluation of the Der-1DCNN model under SNR = −16 dB.

the model's complexity by comparing the testing time of the Der-1DCNNC with other CNN methods.

First, we perform five repeated experiments for the Der-1DCNN method under the strong noise environment (SNR= −16 dB). The diagnosis results are recorded during the entire training process and are shown in Figure 15. As can be seen, the Der-1DCNN model can obtain about 90% accuracy in these five repeated experiments, and the final testing results of these five experiments have a variance of ±0.0015. This shows that the proposed Der-1DCNN model has very good testing stability.

Then, we separately record the average time to diagnose a testing sample when comparison methods and the Der-1DCNN model perform parallel operations on the workbench. As shown in Table 3, the Der-1DCNN model consumes more time than other CNN methods in terms of testing time, which is easy to understand. Because the Der-1DCNN model has a deeper network structure and more parameters, it will inevitably consume more time when processing testing data. However, it can be seen that the average diagnostic time of the proposed model for a testing sample is only 1ms, which is fully acceptable in practical engineering applications. Therefore, the proposed Der-1DCNN method is also suitable for real-time monitoring and fault diagnosis.

VI. CONCLUSIONS AND FUTURE WORKS

Targeting the original vibration signals collected from the complex working conditions of variable loads, variable speeds and strong noise, this paper proposes a novel Der-1DCNN framework to learn high-level fault features, which are applied to the intelligent fault diagnosis of wheelset bearings in HSTs. The key contribution of this paper is to propose 1D residual blocks and to develop a novel Der-1DCNN model. The model can effectively exert the powerful learning abilities of deeper networks and automatically learn the most essential high-level fault features from raw signals. Meanwhile, we introduce a wide convolutional kernel and dropout into the proposed model to further enhance the anti-noise ability.

In this paper, the Der-1DCNN model is evaluated on the wheelset bearing test rigs of HSTs. The experimental results show that the Der-1DCNN has the most excellent performance and most powerful feature learning ability than four comparison methods applied to bearing fault diagnosis in a strong noise environment. Moreover, the Der-1DCNN model has superior domain adaptation ability without any pre-processing. This indicates that the proposed model is promising in the intelligent fault diagnosis of HSTs. Additionally, it provides a novel fault diagnoses method for the field of fault diagnosis, and it can be applied to the fault diagnosis tasks of other mechanical or industrial systems without any processing.

In our future work, first, we consider introducing transfer learning and multi-scale learning to further improve the network's domain adaptation ability. Second, because the number of normal samples is far greater than the fault samples in the real operation of HSTs, we will further study the intelligent fault diagnoses of HSTs in the case of unbalanced samples.

REFERENCES

[1] J. Tian, C. Morillo, M. H. Azarian, and M. Pecht, "Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1793–1803, Mar. 2016.

[2] Y. Yang, D. Yu, and J. Cheng, "A roller hearing fault diagnosis method baesd on EMD energy entropy and ANN," *J. Sound Vib.*, vol. 294, nos. 1–2, pp. 269–277, Jun. 2006.

[3] N. G. Nikolaou and I. A. Antoniadis, "Rolling element bearing fault diagnosis using wavelet packets," *NDT&E Int.*, vol. 35, no. 3, pp. 197–205, Apr. 2009.

[4] Y. Yang, D. Yu, and J. Cheng, "A fault diagnosis approach for roller bearing based on IMF envelope spectrum and SVM," *Measurement*, vol. 40, nos. 9–10, pp. 943–950, Nov./Dec. 2007.

[5] J. Yang, Y. Zhang, and Y. Zhu, "Intelligent fault diagnosis of rolling element bearing based on SVMs and fractal dimension," *Mech. Syst. Signal Process.*, vol. 21, no. 5, pp. 2012–2024, Jul. 2007.

[6] H. Cao, F. Fan, K. Zhou, and Z. He, "Wheel-bearing fault diagnosis of trains using empirical wavelet transform," *Measurement*, vol. 82, pp. 439–449, Mar. 2016.

[7] X. Wang, C. Peng, and Z. Zhang, "Application of EEMD-based resonance demodulation technology in train bearing fault diagnosis," *Modern Electron. Technol.*, vol. 38, no. 21, pp. 24–27, Nov. 2015.

[8] N. Qin, W. Jin, J. Hung, and Z. Li, "Ensemble empirical mode decomposition and fuzzy entropy in fault feature analysis for high-speed train bogie," *Control Theory Appl.*, vol. 31, no. 9, pp. 1245–1251, Sep. 2014.

[9] N. Qin, W. D. Jin, J. Huang, P. Jiang, and Z. M. Li, "High speed train bogie fault signal analysis based on wavelet entropy feature," *Adv. Mater. Res.*, vols. 753–755, no. 12, pp. 2286–2289, Aug. 2013.

[10] J. Liu, Y.-F. Li, and E. Zio, "A SVM framework for fault detection of the braking system in a high speed train," *Mech. Syst. Signal Process.*, vol. 87, pp. 401–409, Mar. 2017.

[11] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[12] 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.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2012, pp. 1097–1105.

[14] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[15] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2014, pp. 1746–1751.

[16] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2015, pp. 2377–2385.

[19] 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, Jan. 2014.

[20] J. Zhao, Y. Yang, T. Li, and W. Jin, "Application of empirical mode decomposition and fuzzy entropy to high-speed rail fault diagnosis," in *Foundations of Intelligent Systems* (Advances in Intelligent Systems and Computing), Nov. 2014, pp. 93–103.

[21] J. Xie, T. Li, Y. Yang, and W. Jin, "Learning features from high speed train vibration signals with deep belief networks," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2014, pp. 2205–2210.

[22] R. Pang, Z. Yu, W. Xiong, and H. Li, "Faults recognition of high-speed train bogie based on deep learning," *J. Railway Sci. Eng.*, vol. 12, no. 6, pp. 1283–1288, Dec. 2015.

[23] C. Guo, Y. Yang, H. Pan, T. Li, and W. Jin, "Fault analysis of high speed train with DBN hierarchical ensemble," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2016, pp. 2552–2559.

[24] J. Yin and W. Zhao, "Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach," *Eng. Appl. Artif. Intell.*, vol. 56, pp. 250–259, Nov. 2016.

[25] Z. Chen, C. Li, and R.-V. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock Vib.*, vol. 2015, no. 2, pp. 1–10, Oct. 2015.

[26] O. Janssens *et al.*, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016.

[27] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, pp. 490–502, Nov. 2016.

[28] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016.

[29] 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.

[30] 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, pp. 425–446, Feb. 2017.

[31] 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.

[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.*, to be published.

[33] 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.

[34] M. Lin, Q. Chen, and S. Yan, "Network in network," *Comput. Sci.*, to be published.

[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Jun. 2010, pp. 807–814.

[36] I. Goodfellow, Y. Bengio, and A. Courville, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Computer Science*. Nov. 2016, pp. 173–177.

[37] I. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2016.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf Learn. Represent.*, May 2015.

[39] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

DANDAN PENG is currently pursuing the B.S. degree with the University of Electronic Science and Technology of China, Chengdu, China. Her research interests include deep learning, signal processing, and their application in the bearing fault diagnosis.

ZHILIANG LIU was born in Rizhao, Shandong, China, in 1984. He received the Ph.D. degree from the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013. From 2009 to 2011, he was a Visiting Scholar with the University of Alberta. From 2013 to 2015, he was an Assistant Professor with the School of Mechanical and Electrical Engineering, UESTC, where he has been an Associate Professor, since 2015. He published more than 40 papers, including over 20 journal papers. His research interests include fault diagnosis and prognostics of rotating machinery by using advanced signal processing and data mining methods. He currently holds seven research grants from the National Natural Science Foundation of China, Open Grants of National Key Laboratory, and China Postdoctoral Science Foundation.

HUAN WANG is currently pursuing the B.S. degree with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include medical image processing, image recognition, deep learning, and machine learning.

YONG QIN received the B.Sc. and M.Sc. degrees in transportation automation and control engineering from Shanghai Railway University, China, in 1993 and 1996, respectively, and the Ph.D. degree in information engineering and control from the China Academy of Railway Sciences, in 1999. He is currently a Professor and the Vice Dean of the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He has authored or co-authored more than 100 publication papers (SCI/EI), 1 ESI highly cited paper, and 5 books, and holds 23 patents granted, including 2 U.S. patents. His research area mainly focused on prognostics and health management for railway transportation system, transportation network safety and reliability, and rail operation planning and optimization. He is a member of the IEEE ITS and RS and a Senior Member of IET. He received the 11 Science and Technology Progress Award of the Ministry.

LIMIN JIA received the B.Sc. degree from Shanghai Railway University, China, in 1984, and the M.Sc. and Ph.D. degrees in information engineering and control from the China Academy of Railway Sciences, in 1987 and 1991, respectively. He is currently a Professor and the Dean of the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He has authored or co-authored about 70 publication papers (SCI/EI) and 20 books and holds about 20 patents. His research area mainly focused on prognostics and health management for railway transportation system, rail operation planning and optimization, and transportation network safety and reliability.