Journal of Dynamics, Monitoring and Diagnostics, 2022, 1, 148-159
https://doi.org/10.37965/jdmd.2022.53

ISTP
RESEARCH ARTICLE

Long-Range Dependencies Learning Based on Nonlocal 1D-Convolutional Neural Network for Rolling Bearing Fault Diagnosis

Huan Wang,[2] Zhiliang Liu,[1,2] and Ting Ai[2]

[1]State Key Laboratory of Traction Power, Southwest Jiaotong University, Chengdu 610031, People's Republic of China

[2]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

(Received 29 October 2021; Revised 09 February 2022; Accepted 29 March 2022; Published online 12 April 2022)

Abstract: In the field of data-driven bearing fault diagnosis, convolutional neural network (CNN) has been widely researched and applied due to its superior feature extraction and classification ability. However, the convolutional operation could only process a local neighborhood at a time and thus lack the ability of capturing long-range dependencies. Therefore, building an efficient learning method for long-range dependencies is crucial to comprehend and express signal features considering that the vibration signals obtained in a real industrial environment always have strong instability, periodicity, and temporal correlation. This paper introduces nonlocal mean to the CNN and presents a 1D nonlocal block (1D-NLB) to extract long-range dependencies. The 1D-NLB computes the response at a position as a weighted average value of the features at all positions. Based on it, we propose a nonlocal 1D convolutional neural network (NL-1DCNN) aiming at rolling bearing fault diagnosis. Furthermore, the 1D-NLB could be simply plugged into most existing deep learning architecture to improve their fault diagnosis ability. Under multiple noise conditions, the 1D-NLB improves the performance of the CNN on the wheelset bearing data set of high-speed train and the Case Western Reserve University bearing data set. The experiment results show that the NL-1DCNN exhibits superior results compared with six state-of-the-art fault diagnosis methods.

Keywords: convolutional neural network; fault diagnosis; long-range dependencies learning; rolling bearing

I. INTRODUCTION

Rolling bearings are the pivot components of the rotating machinery, and the damage of them directly declines the performance of the mechanical system, and safety problems, as well as enormous economic losses, could be caused. However, the long-time process under adverse operating conditions could easily cause different kinds of damage such as crack, abrasion, and gap. Therefore, the health condition monitoring for rolling bearings is crucial to protect the machinery system from safety problems [1].

With the development of the internet of things and the demand for long-term condition monitoring, companies have obtained enormous industrial data. Since the data-driven machine learning method could extract features of the machinery system form historical data automatically, it has been widely applied in the field of rolling bearing fault diagnosis. In general, the traditional diagnosis methods [2–5] mainly include two steps: (1) feature extraction and (2) fault recognition. The feature extraction [2,6] is to obtain the features that can reflect the state of the machine through the feature extraction algorithm. Fault recognition [3,7] uses a classifier algorithm to identify and classify the obtained features. However, the manually extracted statistical features can hardly characterize the complex dynamic features of vibration signals. Moreover, most of these classifier algorithms are shallow models, which cannot learn the complex nonlinear relationship effectively. Thus, it is easy for them to make a wrong judgment.

In recent years, deep learning has attracted more and more attention in the field of fault diagnosis [8–11]. Compared with traditional methods, the deep learning method could extract features from lower level to higher level automatically based on multiple nonlinear operations, and thus it could diagnose with higher intelligence. In particular, the convolutional neural network (CNN) has achieved remarkable success in fault diagnosis tasks due to its unique feature learning mechanism [12–14]. For example, Ince et al. [15] proposed a new one-dimensional CNN (1DCNN) for the real-time fault diagnosis of motors. Peng et al. [16] used a 1D deep residual CNN to diagnose the fault status of train wheelset bearing. Chen et al. [17] combined the CNN with an extreme learning machine to improve the fault diagnosis performance of the network. These methods are based on the 1DCNN [15–21], which mainly takes signals as input and automatically extracts fault features and diagnoses fault types through 1D convolution. In addition, Xia et al. [22] proposed a multisensor-based CNN fault diagnosis method to learn spatial and temporal information from multiple sensors simultaneously to obtain better results. Wen et al. [23] used two-dimensional CNN (2DCNN) to diagnose the health status of various mechanical components. These methods are based on the 2DCNN [22–24], which recombines 1D signal into 2D image or time spectrum, and then uses 2D network architecture to get the final diagnosis results. However,

Corresponding author: Zhiliang Liu (email: zhiliang_liu@uestc.edu.cn).

compared with the 1DCNN, the network structure and operation process required by the 2DCNN are more sophisticated. Therefore, in this article, we use the 1DCNN to solve the fault diagnosis of rolling bearings.

Even though the CNN has been successfully applied in bearing fault diagnosis, it was initially introduced to solve computer vision problems such as image segmentation [25] and face recognition [26]. In order to accomplish these tasks, CNN needs to pay more attention to the relevant information of the local neighborhood. Therefore, CNN lacks sufficient attention to the relevance of long-distance information.

Nevertheless, the vibration signal of rotating machinery is significantly different from the image. It is a temporal signal with strong periodicity. In addition, because of complicated operation conditions, these signals are always with strong nonlinearity and instability. Therefore, there is a strong correlation among different time points. Among these periodicities and correlations, there may be a large quantity of valuable information hidden. For example, as shown in Fig. 1, when a bearing has a local fault, the faulty part and other components produce a periodic short-term impact and encourage the bearing system to perform high-frequency free attenuation vibration according to its resonance frequency. Therefore, if we only consider the signal within a local region, diagnosis is more likely to be interfered with by random factors [27]. Apart from this, comparing the relationship among the amplitude of impulse points in different periods and positions is considered practical to understand the information in the signal fully.

The nonlocal mean (NLM) algorithm was first introduced by Buades et al. [28] in the field of image denoising. This algorithm firstly breaks the image into patches of the same size. Then, it replaces the value at one pixel with the weighted average based on the similarity among the patch where the pixel belongs and other patches. In that way, the NLM could use the dependencies among one pixel and other pixels. Therefore, this method has a strong ability to capture long-range dependencies and has shown its extraordinary performance on image denoising. Besides, The NLM is also widely used in the denoising task of 1D time-series signals and has achieved impressive results. For instance, Van et al. [29] presented a fault diagnosis method based on NLM denoising. Kumar et al. [30] applied NLM to electrocardiogram denoising.

Recently, Wang et al. [31] combined NLM and CNN to introduce nonlocal neural networks, which have superior performance on image classification compared with other computer vision methods. We assume that it makes more sense to improve the long-range dependencies learning ability of the algorithms applied for processing time-series signals, compared with that for image data. Therefore, inspired by [31], we introduced the idea of NLM in the field of time-series signal denoising into the 1DCNN and constructed a 1D nonlocal block (1D-NLB) for capturing long-range dependencies. The 1D-NLB computes the response at a position as a weighted average of the features at all positions. It could build connections between one position and any other positions so that it could capture their dependencies. The 1D-NLB can be integrated into every 1DCNN as an efficient, simple, and general component for capturing long-term dependencies in signals. Therefore, based on 1D-NLB, we propose a nonlocal 1D-convolutional neural network (NL-1DCNN) for fault diagnosis of rolling bearings. The NL-1DCNN captures the shallow features, long-range dependencies, and high-level features of the input signal layer by layer, thereby accurately diagnosing the current health status of the bearing.

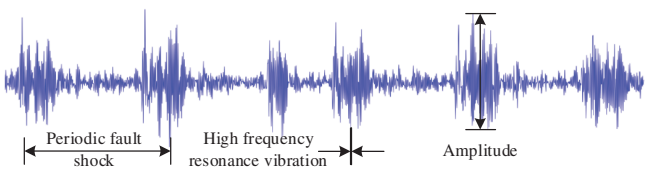

FIGURE 1. The demonstration of real faulty signal that contains both low-frequency vibration generated by the impulse of defect position and high-frequency resonance vibration.

The contributions of this paper are summarized as follows:

(1) Inspired by the NLM algorithm in the field of signal denoising, this paper proposes a nonlocal module based on the 1DCNN for capturing long-term dependencies of signals.

(2) The proposed 1D-NLB can be integrated into every 1DCNN as an efficient, simple, and universal component, thereby improving the diagnosis performance of the network.

(3) This paper proposes the 1DCNN based on 1D-NLB to diagnose the health status of rolling bearings.

(4) The NL-1DCNN has been extensively verified on the wheelset bearing data set and the Case Western Reserve University (CWRU) bearing data set [32], which has achieved better diagnostic results than six state-of-the-art fault diagnosis methods.

The rest of this paper is organized as follows. In Section II, the realization of the NLM algorithm on signal is described. In Section III, the proposed NL-1DCNN is described in detail. Section IV verifies the effectiveness and superiority of the NL-1DCNN. Section VI summarizes the whole paper.

II. REALIZATION OF NLM ON VIBRATION SIGNAL

The NLM algorithm for signal denoising is mainly based on the following procedures. First, a neighborhood block is constructed with each vibration signal point as the center, and then structural information, similar to the neighborhood block, is searched in the global range of the signal. Finally, the information is weighted and averaged to eliminate noise in the vibration signal.

Suppose the expression of the vibration signal of faulty rolling bearing is

$$y(t) = x(t) + n(t), \tag{1}$$

where $x(t)$ is the fault impulse signal, $n(t)$ is the noise generated by other factors such as resonance and $y(t)$ is the observed signal.

The mission of denoising is to eliminate $n(t)$ from the observed vibration signal $y(t)$ so that the original fault impulse signal $x(t)$ could recover. For any position t, the estimated $K(t)$ which is the weighted average of signal values within a predefined search neighborhood $N(t)$ is given by

$$K(t) = \frac{1}{Z(t)} \sum_{s \in N(t)} \omega(t,s) y(s) \tag{2}$$

where $\omega(t, s)$ is the weight associated with sth searched point and tth desired point in $N(t)$ which represents the search window centered on position t. $Z(t) = \sum_{s \in N(t)} \omega(t,s)$ is the normalized factor. The weight, as described in [33], is given by

$$\begin{aligned} \omega(t,s) &= \exp\left(-\frac{\sum_{\delta \in \Delta} (y(t+\delta) - y(s+\delta))^2}{2L_\Delta \lambda^2}\right) \\ &= \exp\left(-\frac{d^2(y(t), y(s))}{2L_\Delta \lambda^2}\right), \end{aligned} \tag{3}$$

where λ is the bandwidth parameter and Δ represents the local patch of L_Δ points surrounding the position t; the patch surrounding the position s also contains L_Δ points; d^2 represents the sum of the squares of Euclidean distances of the local patches centered on the signal points t and s. The novelty of NLM is that the weight between two local patches relies on their similarity rather than their physical distance [34]. Therefore, the denoising process of NLM is nonlocal.

III. THE PROPOSED NL-1DCNN FAULT DIAGNOSIS METHOD

In this section, the generic definition of nonlocal operation in the CNN is firstly introduced. Then we give an instance based on the definition. For the last part, the NL-1DCNN aiming at rolling bearing fault diagnosis is introduced in detail.

A. DEFINITION OF 1D NONLOCAL

Different from the implementation of NLM algorithm in the field of vibration signal denoising, the nonlocal operation in the 1DCNN takes feature signals as input, and then outputs feature signals containing global feature information. Therefore, we define a generic nonlocal operation in the 1DCNN as

$$m_i = \frac{1}{\kappa(n)} \sum_{\forall j} f(n_i, n_j) g(n_j) \tag{4}$$

where i is the index of a position on the output feature signal, and the response at that position is the value obtained after a nonlocal operation. j is the index that enumerates all possible positions. n is the input feature signal and m is the output which has the same length as n. The function f is responsible for calculating the dependency between indexes i and all indexes j of the signal. The function g computes the response of the input signal at position j. The response is normalized by a factor $\kappa(n)$.

This operation takes the relationship between position i with any position j into consideration and regards the weighted average value of the response as output. Therefore, it can make the network perceive long-range dependencies among different regions in the input feature signal at one time. By comparison, the convolutional operation could only learn the feature within a local neighborhood whose size equals the size of convolution kernels. Likewise, a recurrent neural network could only capture the dependencies among neighboring times.

The 1D nonlocal operation is very simple. The basic idea is to calculate the long-range correlation between the current position and other positions in the input signal, so that the algorithm can quickly capture the detailed local information and global information of the input signal. In addition, this operation can be easily implemented in the CNN with only a small amount of parameter increasing.

B. 1D NONLOCAL BLOCK

According to the above definition, the pivot of 1D-NLB operation is function f which calculates similarity and function g computing the response. Thus, the realization of these two functions is highly related to the performance of 1D-NLB. In this paper, for simplicity, we only consider g as a linear transformation, which means $g(n_j) = W_g n_j$, where W_g is a weight matrix to be learned. According to the implementation of nonlocal operations in [28,31], a natural choice of f is the Gaussian function. For the convenience of capturing the dependencies among different regions in the signal, we define the f as

$$f(n_i, n_j) = \exp(n_i^T n_j) \tag{5}$$

where $n^T{}_i\ n_j$ represents dot-product similarity, which is much easier to realize in various neural network platforms, and does not add any training parameters. Thus, the normalized factor is defined as

$$\kappa(n) = \sum_{\forall j} f(n_i, n_j) \tag{6}$$

Figure 2 illustrates the realization of the 1D-NLB in the 1DCNN. n is the input feature signal, $n \in R^{B \times W \times C}$, where B is batch size, W means the length of the signal, and C represents the number channels. At the very beginning, n is multiplied by n^T and get matrix v, $v \in R^{B \times W \times W}$. Then, v is fed into softmax layer to obtain the dependencies among

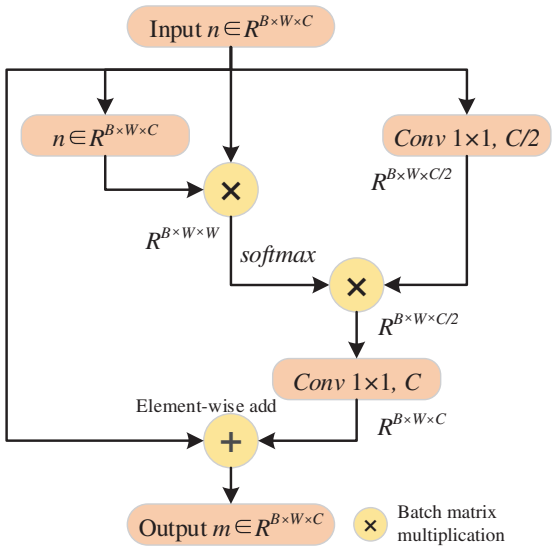

FIGURE 2. The illustration of universal architecture of the 1D-NLB. '×' denotes batch matrix multiplication and '+' represents element-wise addition. This module can well capture the long-distance dependencies of the input signal.

one position of n and other positions. The result could be expressed as

$$\hat{v}_i = \frac{\exp(v_i)}{\sum_j \exp(v_j)} \tag{7}$$

where $\hat{v} \in R^{B\times W\times W}$.

Meanwhile, n goes through a 1×1 convolutional layer to halve its channels. After that, it is multiplied by $\hat{v}$ and passes another 1×1 convolutional layer so that the number of channels could recover to C. Thus, the output m is calculated by

$$m = Conv(\hat{v}Conv(n)) \tag{8}$$

where $m \in R^{B\times W\times C}$.

At last, in order to optimize the feature signal while retaining the original information. We introduce residual connection on this basis to form a complete 1D-NLB. As a result, the output m is rewritten as

$$m = Conv(softmax(n^T n) \otimes Conv(n)) + n, \tag{9}$$

The method we proposed computes the dependencies among one local region of the input signal and the entire signal. Besides, the information could be extracted by only increasing extremely few training parameters. The 1D-NLB is very simple to be plugged into most existing 1DCNN. It could also be embedded into any layers among the network to combine the long-range dependencies with short-range information at different levels. Therefore, this allows us to build an architecture with a strong ability to learn the global information contained in the signal.

C. NONLOCAL 1D-CONVOLUTIONAL NEURAL NETWORK

The 1D-NLB can be simply embedded in the 1DCNN to improve its learning ability of long-range dependencies of input signals. Based on 1D-NLB, we propose the NL-1DCNN, which aims at rolling bearing fault diagnosis. The universal architecture of the NL-1DCNN is shown in Fig. 3.

The NL-1DCNN takes a 1D vibration signal as input. First, two shallow convolution modules are used to learn the shallow feature information in the signal. Subsequently, a 1D-NLB is used to learn the long-range dependencies features of the signal. Through the feature learning of the shallow convolution module, the input signal of 1D-NLB can encode enough semantic information, so that 1D-NLB can obtain the temporal correlation in the signal with higher effectiveness and accuracy. This is why two shallow convolution modules are used before 1D-NLB. In addition, the NL-1DCNN also uses multiple convolution modules to encode the high-level semantic features of the signal, so that different types of signals have sufficient distinction. For each convolutional module, it is consisted of a 1D convolutional layer, a batch normalization, and a ReLU activation function layer. We implement downsampling by setting a large convolution stride, which can minimize the corresponding information loss.

For the classification stage, the learned feature is sent to a global average pooling (GAP) [35] layer followed by a softmax activation. Assuming there are H different classes, the output probability Q_h for the class h is calculated by

$$Q_h = \frac{\exp(q_h)}{\sum_{h=1}^{H} \exp(q_h)}, h = 1,2,\ldots,H \tag{10}$$

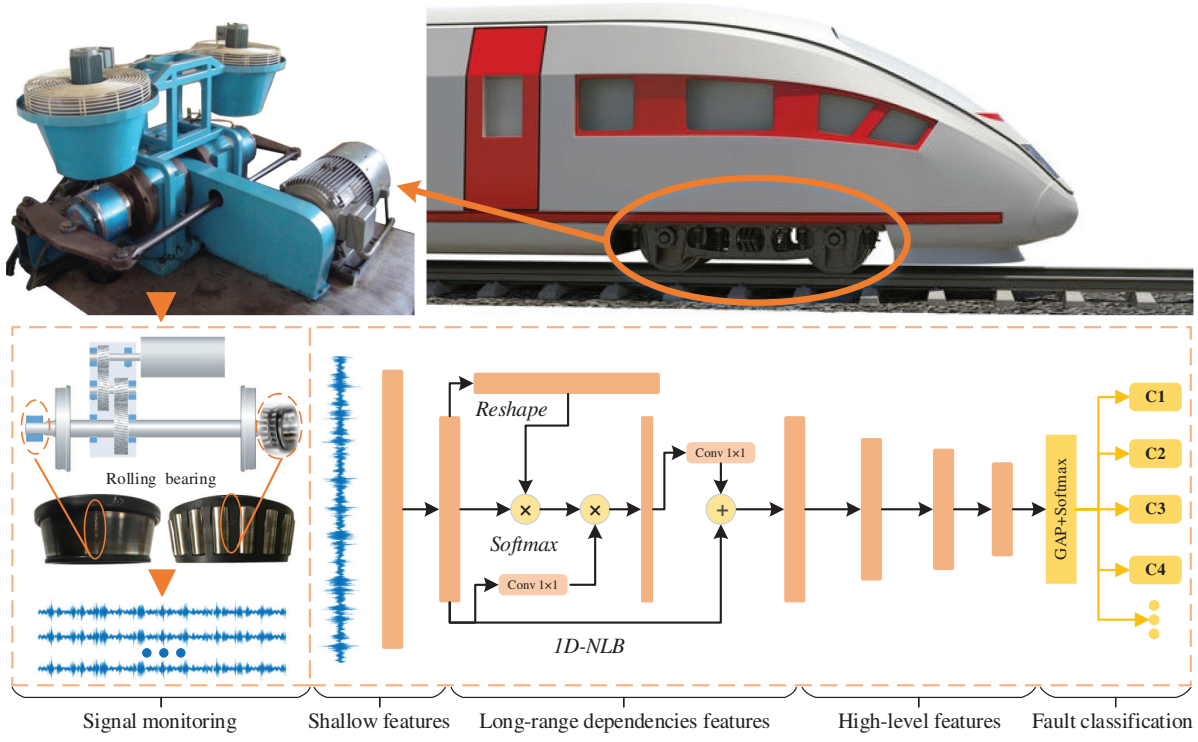

FIGURE 3. The illustration of universal architecture of the NL-1DCNN. It contains four parts: (a) signal convolutional layer for shallow features learning, (b) 1D-NLB which is used to capture long-range dependencies features, (c) multiple convolutional layers for high-level features learning, and (d) the fault classification part combined GAP with Softmax layer.

TABLE I. The detailed architecture of NL-1DCNN

Layer	Type	Kernel/ channel	Stride/ padding	Output
1	Conv+BN	24 × 1/16	4/Yes	512 × 16
2	Conv+BN	12 × 1/32	2/Yes	256 × 32
3	1D-NLB	–	–	256 × 32
4	Conv+BN	12 × 1/32	2/Yes	128 × 32
5	Conv+BN	6 × 1/64	2/Yes	64 × 64
6	Conv+BN	6 × 1/64	2/Yes	32 × 64
7	Conv+BN	3 × 1/128	2/Yes	16 × 128
8	Global Average Pooling			128
9	Softmax			12
C1	C2	C3	C4	...

where q_h is the input of the softmax layer. The diagnosis output is the fault label corresponding to the largest Q_h.

The detailed architecture of the NL-1DCNN is demonstrated in Table I. The length of the input signal of the NL-1DCNN is 2048 × 1, which can ensure that the input signal contains a complete period. Six convolutional modules are applied in the NL-1DCNN in total. Among them, the first two layers of convolution modules are used to capture the shallow information of the input signal, then 1D-NLB is used to learn long-range dependencies features, and the last four layers of convolution modules are used to learn high-level semantic features. The number of channels of the network's convolution module gradually increases from 16 to 128. Except that the stride of the first layer is set to 4, the stride of other layers is set to 2, so that the dimension of the feature signal is finally compressed to 16 × 128. Inspired by [16,19,36], we use wide convolution kernel to learn more fault-related features of the signal. In order to balance the feature extraction capability and the number of parameters of the network model, we set the size of the convolution kernel to gradually decrease, that is, the size of the convolution kernel of the network is gradually reduced from 24 × 1 to 3 × 1. The proposed network model thus uses large convolution kernels in shallow layers to obtain sufficient shallow features from the signal. The extracted features are then filtered and abstracted using small convolution kernels in the deep layers to build high-level features that can be used for device health identification. Apart from this, we use GAP layer to compress the signal into a vector, which decreases the number of trained parameters dramatically compared with using fully connected layer. The probability is outputted by the softmax function.

IV. EXPERIMENT VERIFICATION

In this section, we perform an ablation study and comparative experiments on the wheelset bearing data set and motor bearing data from the CWRU to verify the effectiveness and superiority of the proposed nonlocal operation and fault diagnosis method.

A. EXPERIMENT SETUP

Deep learning-based methods need a large quantity of samples to optimize parameters and the process of slicing the training samples with overlap proposed by [16,19] could enormously increase the number of training samples. Therefore, we adopt the same method for data augmentation. The length of each sample is 2048 while the step size of sliding segmentation is set to 128 in our experiment. 2048 is greater than the number of sampling points in one rotation cycle of the device, so each sample contains complete cycle information.

The proposed NL-1DCNN is realized in the Keras library under Python 3.5. The training and testing process are performed on a workstation with an Intel Core i7-6850K CPU and a GTX 2080 GPU. In addition, we changed the division standard deviation to division variance in z-score normalization. We find that this can make the network achieve better performance. During the training process, we adopt Adam optimizer and the learning rate is set to 0.0001. The batch size is 196 and 96 on wheelset bearing data set and motor bearing data set, respectively. In this paper, we adopt three generic performance indicators: accuracy, recall, and precision.

To better stimulate strong noise disturbance of bearings in the real circumstance, we added additional Gaussian white noise to the raw signals. The definition of SNR is shown as

$$SNR_{dB} = 10log_{10}\left(\frac{P_{signal}}{P_{noise}}\right), \tag{11}$$

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

In this paper, the NL-1DCNN is compared with six state-of-the-art deep learning-based methods. First, we compare the NL-1DCNN with dislocated time series CNN (DTS-CNN) proposed by Liu et al. [27]. The DTS-CNN uses a dislocate layer, so that the network can learn the correlation between different time series in the signal to a certain extent. In the experiments, m, n, and k of the DTS-CNN are set to 10, 512, and 30, respectively, and a dropout layer with a dropout rate of 0.2 is used in the fully connected layer to suppress overfitting. In addition, we compare the NL-1DCNN with the LSTM-based methods. The LSTM has a good learning ability of timing correlation features. In this experiment, the used LSTM has two LSTM cells, where its time steps are 64 and the input dimension is 32.

Finally, we also selected the two state-of-the-art 1DCNN-based fault diagnosis methods, namely wide first-layer kernels CNN (WDCNN) [19] and residual learning-based CNN (ResCNN) [18], which use wide convolution kernel and residual network structure, respectively; and the two state-of-the-art 2DCNN-based fault methods, namely Wen-CNN [23] and hierarchical learning rate adaptive deep CNN (ADCNN) [24], both convert 1D signals into 2D images, and then use different structures of 2D networks to learn fault features. To fairly compare the performance of different methods, we have trained and tested these methods under the same experimental conditions, and four-fold cross-validation is also applied to verify the performance of every method.

B. CASE 1: WHEELSET BEARING FAULT DIAGNOSIS

1) DATA DESCRIPTION. The wheelset bearing test rig provides the experiment data. As shown in Fig. 4, the

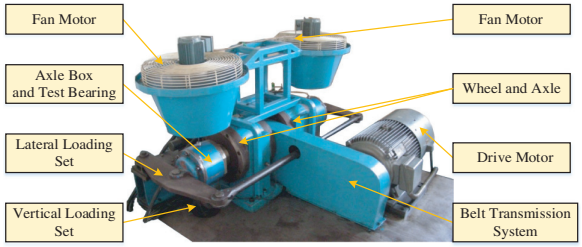

FIGURE 4. The introduction of wheelset bearing test rig.

wheelset bearing test rig is mainly composed of a drive motor, a belt transmission device, a lateral loading set, a vertical loading set, and two fan motors. The vertical and the lateral loading sets are designed to mimic two-dimensional loads in real train operation. An axle and its two supporting bearings are assembled to the test rig. Use an acceleration sensor to collect the vibration signal of rolling bearing. The acceleration sensor is fixed at 9 O'clock and 12 O'clock of the axle box, and the sampling frequency is 5120 Hz. The experimental bearings used double-row taper roller bearings. The photos and models of these faulty bearings are shown in Fig. 5. These faulty bearings are naturally produced during the operation of high-speed train.

There are various faults occurring to wheelset bearing during the real operation. Therefore, 12 different kinds of typical fault conditions combined with health conditions are set. The faults are distributed in the inner race, outer race, rolling element and cage of the wheelset bearing, and the severity of the faults is different. The information of the testing wheelset bearing is shown in Table II. For each fault type, we set five different running speeds: 60, 90, 120, 150, and 180 km/h, combined with four different vertical loads (56, 146, 236 and 272 kN) and two lateral loads (0 and 20 kN). In that case, there are 40 different working conditions for each fault type, which fully validate the robustness of the intelligent diagnosis method under various operation conditions. After data augmentation, the number of train sample and test sample is 142 420 and 45 668, respectively, for every experiment.

As shown in Fig. 6, the raw vibration signals of the 12 health conditions of the wheelset bearing data set are displayed. In addition, in order to explain the influence of noise on vibration signal, we show the vibration signal after adding different degrees of noise. As shown in Fig. 7, we added 6 dB, 0 dB, and −6 dB Gaussian white noise to the vibration signals of the two fault categories. It can be seen that when a small amount of noise is added, the noise has little effect on the vibration signal. However, when a large amount of noise

FIGURE 5. The photos of faulty bearings. These faulty bearings are naturally produced during the operation of high-speed train. The model of the experimental bearing is given in parentheses.

TABLE II. Health information for 12 experimental Wheelset bearings

Location	Fault mode	Label
None	Normal	Cl
Inner race	Pitting	C2
Rolling element	Pitting	C3
Rolling element	Flaking with a size of 3 mm × 35mm	C4
Inner race	Flaking with a size of 3 mm × 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 10 mm × 45mm	C7
Inner race	Flaking with a size of 10 mm × 45mm	C8
Outer race	Flaking with a size of 10 mm × 30mm	C9
Rolling element	Flaking with a size of 1 mm × 1 mm	C10
Cage	Cracking	C11
Outer race	Flaking with a size of 10 mm × 45mm	C12

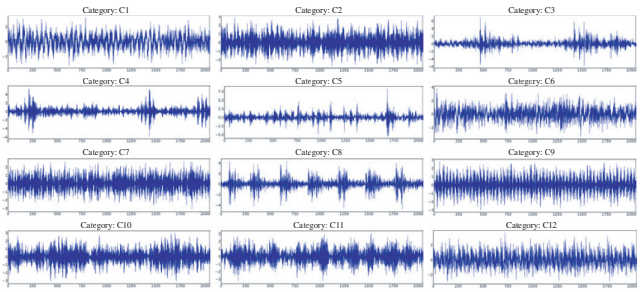

FIGURE 6. The raw data of the high-speed train wheelset bearing data set.

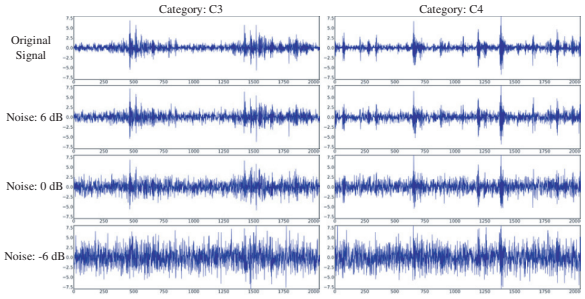

FIGURE 7. The vibration signals after adding different degrees of noise are displayed.

is added, the original waveform of the vibration signal is completely destroyed by the noise, so that it is difficult to distinguish. In actual situations, noise is inevitable. Therefore, in the following experiments, we will also discuss the influence of noise on the deep learning model and the antinoise performance of our proposed method.

2) INFLUENCE OF THE POSITION OF 1D-NLB. The proposed 1D-NLB can be embedded in any layer of the

network to capture long-range dependencies of the feature signal. However, because the length and semantic level of the feature signal in different layers are not consistent, the features learned by 1D-NLB on these layers are also different. Therefore, embedding 1D-NLB in different locations on the network brings different diagnostic performances.

In order to explore the impact on performance when embedding 1D-NLB in different layers of the network, in this experiment, we set up a total of seven different network structures, which are the 1DCNN (the same structure as the NL-1DCNN but does not include 1D-NLB), NL-1DCNN-1, NL-1DCNN-2, ..., NL-1DCNN-6, in which, the number after their name indicates the layer after which the 1D-NLB is embedded. With SNR = −6 dB, we performed experiments on these seven methods. Table III and Fig. 8 show the accuracy, recall, and precision of these methods on the wheelset bearing data set.

The experimental results show that the 1DCNN only obtains 76.80% accuracy, 74.30% recall, and 75.56% precision. After adding 1D-NLB after the first convolutional layer, the NL-1DCNN-1 achieves 81.64% accuracy, 80.13% recall, and 82.90% precision which means they are improved by 4.84%, 5.83%, and 5.75%, respectively. This is a huge improvement, which illustrates the effectiveness of the proposed 1D-NLB. The NL-1DCNN-2 has further achieved better performance, and its accuracy, recall, and precision have improved by 7.53%, 8.60%, and 8.26% over 1D-NLB, respectively. This shows that the 1D-NLB can encode enough long-distance dependencies from shallow feature signals, so that the network can achieve better performance.

In addition, we also observed that starting from NL-1DCNN-3, the diagnostic performance of the network decreased compared to NL-1DCNN-2. Furthermore, the performance of NL-1DCNN-6 is even worse than the 1DCNN. This shows that the 1D-NLB is very sensitive to its location in the network, and its performance changes with its location in the network. In summary, we can conclude that as the location of 1D-NLB in the network deepens, its performance increases first and then decreases. This phenomenon is well understood. The main role of 1D-NLB is to capture the long-range dependencies of the feature signal, and whether sufficient temporal dependencies can be captured is closely related to the input of the 1D-NLB. When the 1D-NLB is located in the shallow layer, the input feature signal has sufficient length, but the semantic level is low, so increasing the semantic level of the input signal can improve the performance of 1D-NLB. When the 1D-NLB is located in the deep layer, the length of the feature signal becomes a greater restrictive factor. In particular, the length of the feature signal outputted by the sixth convolution layer is only 16. In this case, the 1D-NLB has been unable to learn any temporal-related features from such a short feature signal. As a result, the performance of the network has declined since NL-1DCNN-3. Therefore, when designing a 1D-NLB-based fault diagnosis method, it is necessary to balance the two key factors which are the semantic level and feature signal length.

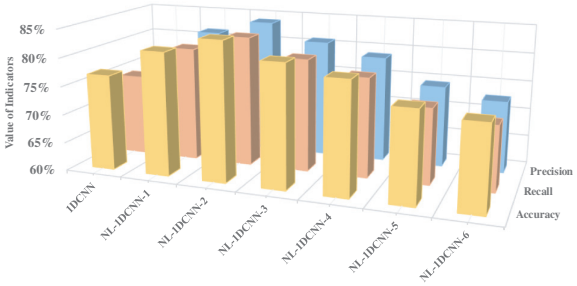

FIGURE 8. The results of the influence of the position of 1D-NLB.

In order to understand the improvement of network performance brought by 1D-NLB more clearly, we use T-SNE technology [37] to visualize the distribution of the features of NL-1DCNN-2 and the 1DCNN on a 2D space, respectively. It is worth noting that the only difference between NL-1DCNN-2 and the 1DCNN is that NL-1DCNN-2 has 1D-NLB and the 1DCNN does not. The visualization results are shown in Fig. 9. Different-colored dots represent different health conditions. According to the subfigures A1 and B1, the shallow features of these two networks are not distinguishable. Subsequently, the 1D-NLB makes the NL-1DCNN-2's features more distinguishable. Thus, the discrimination of the features of NL-1DCNN-2 is better than the 1DCNN. For example, the features in the subfigures B2 and B3 are always clustered together. The degree of dispersion of A2 is greater than that of B2, and the degree of dispersion of A3 is greater than that of B3. This shows that the features in A2 and A3 are more discriminative. Therefore, the discrimination of the features of subfigures A2 and A3 is significantly better than that of subfigures B2 and B3.

This phenomenon shows that the long-distance dependency captured by 1D-NLB is helpful for the network to distinguish and diagnose different fault categories. This not only proves the validity of 1D-NLB, but also proves that the long-distance dependence of the signal helps the network fully understand the hidden features of the signal. It is precisely because 1D-NLB learns these features that the ordinary CNN networks cannot learn, so that the network can obtain better diagnostic results.

3) INFLUENCE OF THE NUMBER OF 1D-NLBS. In order to further explore the impact of the number of 1D-NLB on diagnostic performance, we add one and two 1D-NLBs to the network on the basis of NL-1DCNN-2, which are named NL-1DCNN-2-1 and NL-1DCNN-2-2, respectively. With SNR = −6 dB, we performed experiments on these three methods. The accuracy, recall, and precision of these three methods are shown in Table IV.

TABLE III. Experimental results of the effect of 1D-NLB position on network performance (−6dB)

Indicators	1DCNN	NL-1DCNN-1	NL-1DCNN-2	NL-1DCNN-3	NL-1DCNN-4	NL-1DCNN-5	NL-1DCNN-6
Accuracy	76.80 ± 0.18	81.64 ± 0.40	**84.33** ± 0.55	81.43 ± 0.34	79.59 ± 0.70	75.88 ± 0.84	74.73 ± 0.66
Recall	74.30 ± 0.61	80.13 ± 0.60	**82.90** ± 0.72	79.81 ± 0.42	77.48 ± 0.93	73.27 ± 1.43	71.43 ± 1.08
Precision	75.56 ± 0.35	81.31 ± 0.47	**83.82** ± 0.51	80.81 ± 0.47	78.74 ± 0.77	74.33 ± 0.91	72.60 ± 0.62

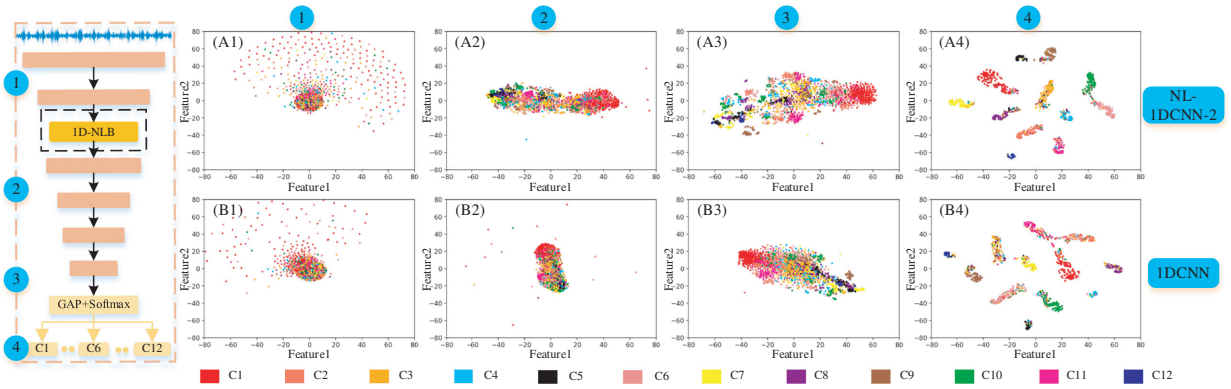

FIGURE 9. The features visualization of NL-1DCNN-2 and the 1DCNN. The figure proves that the 1D-NLB can improve the discrimination of features.

Table IV The Results of the Influence of the Number of 1D-NLBs (−6dB)

Indicators	NL-1DCNN-2	NL-1DCNN-2-1	NL-1DCNN-2-2
Accuracy	84.33 ± 0.55	**84.48** ± 0.12	83.73 ± 0.43
Recall	82.90 ± 0.72	**83.51** ± 0.53	82.20 ± 0.59
Precision	83.82 ± 0.51	**84.18** ± 0.12	83.42 ± 0.46

We find that the number of 1D-NLB has little effect on network performance. The NL-1DCNN-2, NL-1DCNN-2-1, and NL-1DCNN-2-2 achieved similar fault diagnosis performance. This shows that using only one 1D-NLB can capture adequate long-distance dependencies and greatly improve the performance of the network. Although the NL-1DCNN-2–1 is slightly better than NL-1DCNN-2, adding more modules also increases the computational burden to a certain extent. Therefore, in the subsequent experiments, the network structure of our proposed method is consistent with NL-1DCNN-2.

4) EFFECTIVENESS OF 1D-NLB IN EXISTING METHODS. In order to verify the wide applicability of 1D-NLB in the CNN-based fault diagnosis methods, this experiment continues to explore the performance of 1D-NLB in the existing CNN methods. We use the WDCNN as the baseline, and then embed 1D-NLB into different layers of the WDCNN. A total of five different network structures are designed, which are named WDCNN-1, WDCNN-2, . . . , WDCNN-5. The number after their name indicates the layer after which the 1D-NLB is embedded. With SNR = −6 dB, we performed experiments on these six methods. Table V and Fig. 10 show the accuracy, recall, and precision of these methods.

Obviously, we find that the proposed 1D-NLB can also effectively improve the fault diagnosis performance of the WDCNN. For example, the accuracy of the WDCNN-2 is improved by 4.09% compared with the WDCNN. Consistent with the phenomenon of previous experiments, as the position of 1D-NLB in the WDCNN gets deeper, the diagnostic performance of the network increases first and then decreases. This also shows that the length of the feature signal and the semantic level have a great impact on the performance of 1D-NLB. In addition, we find that the improvement of the WDCNN-2 compared with the WDCNN is smaller than that of NL-1DCNN-2 compared with 1DCNN. This is because the WDCNN used a very large downsampling rate in the first convolution layer, which caused the length of the feature signal too small, resulting in 1D-NLB being unable to achieve better performance. This also shows that in order to maximize the performance of 1D-NLB, we need to design a relatively reasonable network structure. Even though the WDCNN is not optimized for 1D-NLB, this module still considerably improves the fault diagnosis performance of the WDCNN. This strongly proves the wide applicability of 1D-NLB.

This experimental phenomenon proves that the proposed 1D-NLB can be simply embedded in other existing CNN architectures to improve their performance, even if these CNNs are not specifically optimized for 1D-NLB. Therefore, 1D-NLB has a very wide application potential, and it could be used as a general module to improve the performance of most CNN networks.

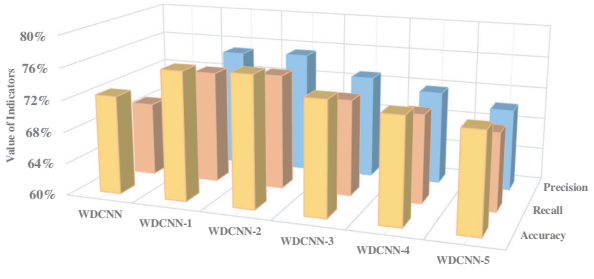

FIGURE 10. The results of influence of the number of 1D-NLBs.

TABLE V. Experimental results of the effectiveness of 1D-NLB in existing methods (−6dB)

Indicators	WDCNN	WDCNN-1	WDCNN-2	WDCNN-3	WDCNN-4	WDCNN-5
Accuracy	72.38 ± 0.78	76.20 ± 1.77	**76.47** ± 1.21	74.27 ± 1.57	73.21 ± 1.67	72.36 ± 0.83
Recall	69.30 ± 1.03	74.08 ± 1.76	**74.44** ± 1.20	72.06 ± 1.59	71.08 ± 1.75	69.71 ± 0.86
Precision	70.03 ± 1.09	74.99 ± 1.81	**75.30** ± 1.12	72.97 ± 1.46	71.71 ± 1.77	70.20 ± 0.67

5) COMPARED WITH STATE-OF-THE-ARTS METHODS. In order to verify the superiority of the proposed NL-1DCNN and explore its performance under different noise conditions, we compare the NL-1DCNN with six state-of-the-art deep learning-based fault diagnosis methods under three different noises (SNR = −6 dB, 0 dB, and 6 dB). Table VI shows the accuracy, recall, and precision of our method and comparison methods.

Obviously, the fault diagnosis performance of the NL-1DCNN under the noise of −6dB, 0dB, and 6dB is better than the other six deep learning methods. According to the experimental results, the NL-1DCNN achieves a diagnostic accuracy of 99.67% at SNR = 6 dB, which is 1.41% higher than that of Wen-CNN. This effectively proves the fault diagnosis ability of the proposed method in a weak noise environment. In addition, when SNR = −6 dB, which means the noise intensity is 3.98 times the raw signal, the NL-1DCNN can still obtain 84.33% fault diagnosis accuracy, which is 11.95% higher than Wen-CNN. This is a good proof that the NL-1DCNN has good antinoise performance even without any denoising preprocessing. In addition, we find that the LSTM with long-distance dependency learning ability has a good performance in this data set. At SNR = −6 dB, it can obtain a diagnostic accuracy of 81.06%. This also confirms that networks with long-distance dependency learning capabilities can effectively capture more essential signal features and thus obtain better fault diagnosis results when dealing with time-series signal. By contrast, the DTS-CNN only obtained 64.15% accuracy at SNR = −6 dB. This shows that the applicability of the DTS-CNN is not satisfactory, and it is difficult to adapt to the fault diagnosis task of wheelset bearing data set.

Table VI also shows the parameter quantities of our method and the comparison method. Since we only added one 1D-NLB module, the number of parameters in our model is still relatively small. Therefore, the proposed method achieves a large performance boost with little parameter increase.

C. CASE 2: MOTOR BEARING FAULT DIAGNOSIS

1) DATA DESCRIPTION. Motor bearing data set is provided by the CWRU bearing data center through fault simulation experiment. There are four health statuses of rolling bearing: normal, outer race fault, inner race fault, and ball fault. Introduced by electro-discharge machining, each fault type was set in three different fault diameters (7 mils, 14 mils, and 21 mils). Therefore, this data set contains 10 different fault types, where same fault types with different fault diameter are attached with different label. The fault information is shown in Table VIII. After data augmentation, the number of train sample and test sample is 79 924 and 26 100, respectively, for every experiment.

2) COMPARED WITH STATE-OF-THE-ART METHODS. In order to explore the applicability of the proposed method on the CWRU bearing data set, we compare the NL-1DCNN with six state-of-the-art deep learning methods under three noise situations (SNR = −6 dB, 0 dB, and 6 dB). Table VII shows the accuracy, recall, and precision of these methods.

We find that the NL-1DCNN has better fault diagnosis performance than six comparison methods under three

TABLE VI. Experimental results of the NL-1DCNN and six state-of-the-art deep learning-based methods on Wheelset bearing Data set

SNR	Indicators	NL-1DCNN	LSTM	DTS-CNN	WDCNN	Wen-CNN	ResCNN	ADCNN
6dB	Accuracy	**99.67** ± 0.13	97.65 ± 0.34	95.70 ± 0.14	97.90 ± 0.18	98.26 ± 0.21	96.34 ± 0.53	86.65 ± 0.34
	Recall	**99.67** ± 0.15	97.39 ± 0.39	95.29 ± 0.20	97.64 ± 0.19	98.08 ± 0.22	95.99 ± 0.42	84.66 ± 0.31
	Precision	**99.65** ± 0.14	97.40 ± 0.37	95.36 ± 0.15	97.62 ± 0.34	98.16 ± 0.24	95.97 ± 0.59	85.00 ± 0.44
0dB	Accuracy	**98.26** ± 0.40	94.47 ± 0.20	88.03 ± 0.50	93.13 ± 1.02	93.52 ± 0.29	89.37 ± 1.01	77.33 ± 0.79
	Recall	**98.14** ± 0.48	94.00 ± 0.23	86.93 ± 0.38	92.54 ± 1.11	92.93 ± 0.40	88.23 ± 1.41	74.81 ± 0.99
	Precision	**98.28** ± 0.43	94.18 ± 0.18	87.17 ± 0.42	92.64 ± 1.11	93.14 ± 0.16	88.40 ± 0.67	75.33 ± 0.65
−6dB	Accuracy	**84.33** ± 0.55	81.06 ± 0.80	64.15 ± 0.60	72.38 ± 0.78	71.05 ± 1.00	67.11 ± 1.59	57.03 ± 1.15
	Recall	**82.90** ± 0.72	79.43 ± 0.80	60.49 ± 0.74	69.30 ± 1.03	67.86 ± 0.91	62.26 ± 1.84	50.40 ± 1.52
	Precision	**83.82** ± 0.51	79.99 ± 0.75	60.96 ± 0.68	70.03 ± 1.09	68.72 ± 0.98	64.74 ± 0.83	52.12 ± 1.61
Parameters	–	1.10M	2.20M	10.20M	0.75M	7.20M	6.30M	0.23M

TABLE VII. Experimental results of the NL-1DCNN and six state-of-the-arts deep learning-based methods on motor bearing Data set

SNR	Indicators	NL-1DCNN	LSTM	DTS-CNN	WDCNN	Wen-CNN	ResCNN	ADCNN
6dB	Accuracy	**99.89** ± 0.09	96.24 ± 0.29	99.80 ± 0.05	99.76 ± 0.07	99.79 ± 0.05	99.82 ± 0.04	98.68 ± 0.20
	Recall	**99.90** ± 0.08	96.37 ± 0.28	99.81 ± 0.05	99.78 ± 0.07	99.80 ± 0.05	99.83 ± 0.04	98.72 ± 0.19
	Precision	**99.90** ± 0.08	96.38 ± 0.29	99.81 ± 0.05	99.78 ± 0.07	99.80 ± 0.05	99.80 ± 0.05	98.71 ± 0.19
0dB	Accuracy	**99.17** ± 0.16	88.47 ± 0.72	98.90 ± 0.09	98.37 ± 0.31	98.56 ± 0.37	98.08 ± 0.36	96.03 ± 0.47
	Recall	**99.20** ± 0.16	88.91 ± 0.75	98.94 ± 0.09	98.43 ± 0.30	98.60 ± 0.35	98.14 ± 0.35	96.15 ± 0.46
	Precision	**99.20** ± 0.15	88.95 ± 0.76	98.95 ± 0.08	98.45 ± 0.25	98.61 ± 0.36	98.12 ± 0.40	96.15 ± 0.45
−6dB	Accuracy	**91.23** ± 1.04	65.27 ± 0.81	88.69 ± 1.32	85.17 ± 0.96	86.75 ± 0.80	85.89 ± 1.22	81.45 ± 1.40
	Recall	**91.49** ± 1.00	66.50 ± 1.05	89.03 ± 1.28	85.63 ± 0.95	87.16 ± 0.78	86.30 ± 1.20	81.98 ± 1.33
	Precision	**91.34** ± 1.04	66.98 ± 1.05	89.10 ± 1.23	85.48 ± 0.92	87.01 ± 0.77	86.73 ± 1.15	81.92 ± 1.33

noise conditions. Under SNR = 6 dB, the NL-1DCNN achieved 99.89% accuracy of fault diagnosis. At SNR = 0 dB, the noise power is equal to the raw signal power, and the NL-1DCNN achieved a 99.17% accuracy for fault diagnosis. This shows the excellent fault diagnosis performance of the NL-1DCNN. Moreover, the NL-1DCNN performs better on the motor bearing data set than the wheelset bearing data set, and it can obtain 91.23% accuracy at SNR = −6 dB. In addition, we find that the LSTM obtained only 65.27% diagnostic accuracy on this data set. However, the DTS-CNN exhibited relatively good results, which achieved an accuracy of 88.69% at SNR = −6 dB. Although the performance of the DTS-CNN is still far from that of the NL-1DCNN, this proves once again the importance of long-distance dependencies for fault diagnosis tasks.

From the performance of these methods on two data sets, the performance of the DTS-CNN and the LSTM is greatly affected by the data set, and they can only exert their good performance on some data sets. The NL-1DCNN can achieve excellent performance on both data sets, which shows its good adaptability. This reflects the application potential of the NL-1DCNN in other fault diagnosis tasks of rotating machinery to a certain extent.

In order to show the performance of these methods more clearly, we use the T-SNE technology to visualize the final output distribution of the NL-1DCNN, the LSTM, the DTS-CNN, the WDCNN, the Wen-CNN, the ResCNN, and the ADCNN in a two-dimensional space. The visualization results are shown in Fig. 11, where different colors represent different health conditions of motor bearings. Obviously, the output distribution of the NL-1DCNN has the best discrimination, followed by the DTS-CNN and the Wen-CNN. This is consistent with the results of Table VIII, which shows that the proposed NL-1DCNN has better performance on the motor bearing data set.

In order to better understand the diagnostic performance of the proposed method for each health category, the confusion matrix of the proposed NL-1DCNN at SNR = 6 dB is displayed in Fig. 12. Obviously, our method can distinguish normal samples and fault samples with 100% accuracy. In addition, in the identification of fault types, the NL-1DCNN can also identify inner race fault and outer race fault with 100% accuracy, and it can accurately identify the degree of bearing failure. The NL-1DCNN only made a few misjudgments in the diagnosis of ball fault. And, these misjudgments are just judging a certain fault degree of ball fault as other fault degree. This shows that our method can accurately distinguish different fault categories, and there may be few misjudgments when determining the degree of fault.

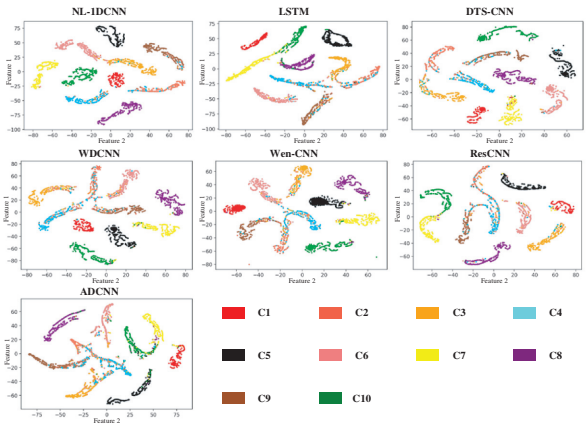

FIGURE 11. Scatter plots of the NL-1DCNN, LSTM, DTS-CNN, WDCNN, Wen-CNN, ResCNN, and ADCNN in two-dimensional space.

TABLE VIII. Description of the motor bearing Data set information

Fault location	Fault size (mil)	Load (hp)	Label
None	0	0,1,2,3	Cl
Ball fault	7	0,1,2,3	C2
Ball fault	14	0,1,2,3	C3
Ball fault	21	0,1,2,3	C4
Inner race fault	7	0,1,2,3	C5
Inner race fault	14	0,1,2,3	C6
Inner race fault	21	0,1,2,3	C7
Outer race fault	7	0,1,2,3	C8
Outer race fault	14	0,1,2,3	C9
Outer race fault	21	0,1,2,3	C10

		Predicted Label											
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Recall	Testing samples
True Label	C1	1800	0	0	0	0	0	0	0	0	0	100	1800
	C2	0	2696	0	4	0	0	0	0	0	0	99.85	2700
	C3	0	0	2700	0	0	0	0	0	0	0	100	2700
	C4	0	20	3	2677	0	0	0	0	0	0	99.15	2700
	C5	0	0	0	0	2700	0	0	0	0	0	100	2700
	C6	0	0	0	0	0	2700	0	0	0	0	100	2700
	C7	0	0	0	0	0	0	2700	0	0	0	100	2700
	C8	0	0	0	0	0	0	0	2700	0	0	100	2700
	C9	0	0	0	0	0	0	0	0	2700	0	100	2700
	C10	0	0	0	0	0	0	0	0	0	2700	100	2700
	Precision	100	99.26	99.89	99.85	100	100	100	100	100	100	—	26100

FIGURE 12. Confusion matrix of the NL-1DCNN under SNR = 6dB.

V. CONCLUSIONS

In this paper, we propose the NL-1DCNN for rolling bearing fault diagnosis. This method aims to improve the long-range dependencies learning ability of the network, so as to fully understand the hidden features of the signals. To this end, we introduced the nonlocal mean method to the CNN and built a 1D-NLB for capturing long-range dependencies. The basic idea of 1D-NLB is to calculate the long-range correlation between the current position and other positions, so that the network can quickly capture the local and global information of the input signal. We validate the effectiveness of the method on two bearing data sets. Experimental results show that the diagnostic performance of the NL-1DCNN is considerably better than the six outstanding methods. The conclusions are summarized as follows: (1) the long-distance dependence can help the network to fully understand the hidden information of the signal, and this information is also very important for fault diagnosis tasks. (2) The proposed 1D-NLB absorbs the advantages of the nonlocal mean denoising algorithm and has excellent learning ability for long-distance dependencies. It can be easily embedded in most CNN architectures to improve its fault diagnosis performance. (3) The NL-1DCNN has good fault diagnosis performance, and it has

consistent performance on two data sets, which shows its application potential in other fault diagnosis tasks.

In addition, the performance of the proposed method is still relatively low in the case of strong noise, which cannot meet the needs of practical applications. Moreover, in practical situations, it is often impossible to obtain enough fault samples, and the proposed method cannot cope with this situation well. Therefore, in future work, we will focus on improving the model's performance in strong noise environments and introduce the idea of few-shot learning to improve the performance of the diagnostic model in the case of limited labeled samples.

Acknowledgments

This work was supported by the State Key Laboratory of Traction Power, Southwest Jiaotong University (TPL2104), and the National Natural Science Foundation of China (61833002).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

[1] M. Cerrada, R. Sánchez, C. Li, F. Pacheco, D. Cabrera and J. Valente de Oliveira, et al., "A review on data-driven fault severity assessment in rolling bearings," Mech. Syst. Signal Process., vol. 99, pp. 169–196, 2018.

[2] X. Wang, Z. Yang, and X. Yan, "Novel particle Swarm optimization-based variational Mode decomposition method for the fault diagnosis of complex rotating machinery," IEEE/ASME Trans. Mechatron., vol. 23, pp. 68–79, 2018.

[3] Q. Hu, Z. He, Z. Zhang, and Y. Zi, "Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble," Mech. Syst. Signal Process., vol. 21, pp. 688–705, 2007.

[4] E. Alizadeh, N. Meskin, and K. Khorasani, "A negative selection immune system inspired methodology for fault diagnosis of wind turbines," IEEE Trans. Cybern., vol. 47, pp. 3799–3813, 2017.

[5] W. Zou, Y. Xia, and H. Li, "Fault diagnosis of Tennessee-Eastman process using orthogonal incremental extreme learning machine based on driving amount," IEEE Trans. Cybern., vol. 48, pp. 3403–3410, 2018.

[6] Y. Cheng, Z. Wang, B. Chen, W. Zhang, and G. Huang, "An improved complementary ensemble empirical mode decomposition with adaptive noise and its application to rolling element bearing fault diagnosis," ISA Trans., vol. 91, pp. 218–234, 2019.

[7] D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, "Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN," Expert Syst. Appl., vol. 40, pp. 4137–4145, 2013.

[8] H. Han, H. Wang, Z. Liu, and J. Wang, "Intelligent vibration signal denoising method based on non-local fully convolutional neural network for rolling bearings," ISA Trans., vol. 122, pp. 13–23, 2021.

[9] H. Wang, Z. Liu, Y. Ge, and D. Peng, "Self-supervised signal representation learning for machinery fault diagnosis under limited annotation data," Knowl.-Based Syst., vol. 239, pp. 107978, 2022.

[10] H. Wang, Z. Liu, D. Peng, and Z. Cheng, "Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising," ISA Trans., In Press, 2021.

[11] H. Wang, T. Men, and Y. F. Li, "Transformer for high-speed train wheel wear prediction with multiplex local-global temporal fusion," IEEE Trans. Instrum. Meas., vol. 71, pp. 1–12, 2022.

[12] 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. Inform., vol. 16, pp. 4949–4960, 2020.

[13] 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.

[14] H. Wang, Z. Liu, D. Peng, M. Yang, and Y. Qin, "Feature-level attention-guided multitask CNN for fault diagnosis and working conditions identification of rolling bearing," IEEE Trans. Neur. Net. Learn., pp. 1–13, 2021.

[15] 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, pp. 7067–7075, 2016.

[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] Z. Chen, K. Gryllias, and W. Li, "Mechanical fault diagnosis using convolutional neural networks and Extreme learning machine," Mech. Syst. Signal Process., vol. 133, p. 106272, 2019.

[18] W. Zhang, X. Li and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," ISA Trans., vol. 95, pp. 295–305, 2018.

[19] Z. Wei, P. Gaoliang, L. Chuanhao, C. Yuanhang, and Z. Zhujun, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," Sensors (Basel, Switzerland), vol. 17, p. 425, 2017.

[20] 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, 2018.

[21] Y. Han, B. Tang, and L. Deng, "An enhanced convolutional neural network with enlarged receptive fields for fault diagnosis of planetary gearboxes," Comput. Ind., vol. 107, pp. 50–58, 2019.

[22] 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. Mechatron., vol. 23, pp. 101–110, 2018.

[23] 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, pp. 5990–5998, 2018.

[24] 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, 2016.

[25] H. Wang, G. Wang, Z. Xu, W. Lei, and S. Zhang, "High- and low-level feature enhancement for medical image segmentation," in *Proc. MLMI*, Lecture Notes in Computer Science, vol. 11861. Springer, Cham, pp. 611–619, 2019.

[26] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," IEEE Trans. Pattern Anal., vol. 40, pp. 1002–1014, 2018.

[27] R. Liu, G. Meng, B. Yang, C. Sun, and X. Chen, "Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine," IEEE Trans. Ind. Inform., vol. 13, pp. 1310–1320, 2017.
[28] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *Proc. CVPR*, IEEE, pp. 60–65, 2005.
[29] M. Van, H. Kang, and K. Shin, "Rolling element bearing fault diagnosis based on non-local means de-noising and empirical mode decomposition," IET Sci. Meas. Technol., vol. 8, pp. 571–578, 2014.
[30] S. Kumar, D. Panigrahy, and P. K. Sahu, "Denoising of electrocardiogram (ECG) signal by using empirical mode decomposition (EMD) with non-local mean (NLM) technique," Biocybern. Biomed. Eng., vol. 38, pp. 297–312, 2018.
[31] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR*, IEEE, pp. 7794–7803, 2018.
[32] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," Mech. Syst. Signal Process., vol. 64–65, pp. 100–131, 2015.
[33] D. Van De Ville and M. Kocher, "SURE-based non-local means," IEEE Signal Process. Lett., vol. 16, pp. 973–976, 2009.
[34] A. Buades, B. Coll, and J. Morel, "A review of image denoising algorithms, with a new one," Multiscale Modeling Simul., vol. 4, pp. 490–530, 2005.
[35] M. Lin, Q. Chen, and S. Yan, "Network in network," [Online]. Available: https://arxiv.org/abs/1312.4400.
[36] H. Wang, Z. Liu, D. Peng, and Y. Qin, "Understanding and learning discriminant features based on multi-attention 1DCNN for wheelset bearing fault diagnosis," IEEE Trans. Ind. Inform., vol. 16, no. 9, pp. 5735–5745, 2019.
[37] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.