A wide kernel CNN-LSTM-based transfer learning method with domain adaptability for rolling bearing fault diagnosis with a small dataset

Yingmou Zhu¹, Hongming chen², Wei Meng², Qing Xiong³ and Yongjian Li²

Abstract
It is difficult to obtain sufficient data for some machines, in addition, different working conditions result in different distributions of training data and test data, which lead to the failure of traditional deep learning methods in engineering applications. To solve these problems, we propose a novel deep learning framework called 1D-WCLT for rolling bearing fault diagnosis that combines wide kernel deep convolutional neural network and long short-term memory (WDCNN-LSTM). In this approach, a wide convolution kernel is utilized for local convolution, because the local convolution receptive field is increased, the fault feature contained in the low-middle frequency component is extracted. It is worth mentioning that the network complexity has not increased. Then, the transfer learning is applied to solve related but different domain problems, the useful knowledge learned by WDCNN-LSTM model under sufficient data conditions is transferred to diagnosis tasks with small dataset. After that, some state-of-the-art methods are applied to compare with the proposed method. At the end, experimental results showed that the proposed approach is an excellent algorithm for fault feature extraction of machinery and has much better identification accuracy and applicability than the other existing techniques.

Keywords
Rolling bearing, fault diagnosis, small dataset, CNN-DLSTM, transfer learning

Date received: 14 June 2022; accepted: 4 October 2022

Handling Editor: Chenhui Liang

Introduction
Rolling bearings are widely used in modern mechanical equipment. Because of the complex and harsh operating environments, it is easy to cause unexpected failures.¹ Therefore, to ensure the safe operation of mechanical equipment, it is of great significance to implement effective state monitoring and fault diagnosis.²

As an important branch of machine learning methods, deep learning is characterized by adaptive extraction of data features based on deep neural network model for target tasks.³⁴ Compared with traditional machine learning methods, deep learning has advantages in model construction, feature extraction and generalization performance,⁵ and has been widely applied in many

¹State Key Laboratory of Heavy Duty AC Drive Electric Locomotive Systems Integration, Zhuzhou, China
²School of Railway Tracks and Transportation, Wuyi University, Jiangmen, China
³School of Intelligent Manufacturing and Automobile, Chengdu Vocational & Technical College of Industry, Chengdu, China

Corresponding author:
Yongjian Li, School of Railway Tracks and Transportation, Wuyi University, 22#, Dongcheng Village, Pengjiang District, Jiangmen 529020, China.
Email: guinong9339@163.com
fields, such as speech recognition, image processing, and signal processing.

Generally, convolutional neural network (CNN) is mostly applied to two-dimensional image processing due to its strong nonlinear characterization ability and its ability to extract local spatial features of input data. Guo et al. proposed a hierarchical images of rolling bearings as input for fault diagnosis achieved good performance. Lu et al. proposed a CNN for small sample damage diagnosis, which was investigated using a deep learning technique under small dataset. Su et al. introduced data reconstruction hierarchy in bearing using deep learning technique under small environment.

In addition, long and short-term memory (LSTM) network, as a variant of recurrent neural network (RNN), is more suitable for processing long-term dependent information than CNN due to its recurrent unit structure, and can extract sequential features of input data. Liu et al. proposed a diagnostic model that combined the advantages of LSTM network and statistical process analysis to predict aeroengine bearing faults with multi-stage performance degradation. Gao et al. used a recurrent unit combining multiple residual blocks and one LSTM block to simultaneously extract sequential features and spatial features from raw signals, solving the problem of weak feature extraction in bearing fault diagnosis. Wu et al. designed a batch normalized LSTM (BNLSTM) network model that could learn the mapping relationship between two data sets to generate auxiliary samples to solve the problem of degraded model performance caused by insufficient labeled data for bearing fault diagnosis.

The superior performance of the method mentioned above is mainly based on sufficient failure samples, independent and uniformly distributed dataset. However, it is difficult to obtain abundant fault data, also a little amount of data can be marked. To solve these problems, some methods, such as information fusion, transfer learning, and so on, were developed. Recently, researchers focused on fault detection of rolling bearing using deep learning technique under small dataset. Su et al. introduced data reconstruction hierarchy in recurrent meta-learning model that can realize fault recognition with sample data. Yang et al. used structural similarity GAN and improved MobileNetv3 CNN for small sample damage diagnosis, which was achieved good performance. Lu et al. proposed a novel approach based on relation network and transfer learning to address the problem of fault identification, which improved the classification accuracy under small dataset.

In this paper, to improve the fault diagnosis accuracy with insufficient effective data, we propose a hybrid neural network combining wide-kernel CNN and LSTM as a diagnostic model. The model is divided into three parts: one-dimensional CNN part, LSTM part and full connection layer part. First, the one-dimensional CNN part will extract the local spatial features of the input data and reduce its dimension. Then, the LSTM part will further extract the sequential features of the encoding features after dimensionality reduction to avoid the slow LSTM training problem caused by data dimension explosion. Finally, the encoding features output by the LSTM part are classified into the fully connected layer part. Moreover, in the convolutional layer of the one-dimensional CNN part, a wide convolution kernel is used for local convolution, which enables the diagnostic model to obtain a large local convolution receptive field with fewer parameters, so as to better extract the low and medium frequency features of input data. In addition, we use the transfer learning strategy based on the proposed hybrid neural network model to improve the domain adaptability of the diagnostic model. Here, useful knowledge related to the target task obtained through sufficient data training is transferred to the target task to obtain high-precision diagnostic performance in case of insufficient data, and there is no need to rebuild the diagnostic model due to changes in operating conditions. Compared with the previous data-driven deep learning methods, the transfer learning strategy adopted in this paper has more advantages in solving the changeable working conditions in engineering practice across domains.

The rest of this paper is arranged as follows. Section “Theoretical background” introduces the relevant background knowledge theory. Section “The proposed method” describes the proposed method in detail. Section “Experimental validation” designs experiments to verify the proposed 1D-WCLT method and compares its diagnostic performance with other state-of-the-art methods under different working conditions with insufficient data. Section “Conclusion” summarizes the conclusion.

**Theoretical background**

**Convolutional neural network**

A CNN is a deep learning model inspired by the structures of biological vision systems. The main structure includes three parts: a convolutional layer, a pooling layer and a fully connected layer (FC layer). Among them, the main function of the convolutional layer is to extract local features, the main function of the pooling layer is to reduce feature dimensions, and the main function of the fully connected layer is to classify. As shown in Figure 1, in the process of CNN forward propagation, the convolutional layer first extracts low-level generalization features from the data and then reduces the dimension of generalization features through the pooling layer. After multiple convolution-pooling feature extraction, the output abstract features of the data
are fed into the fully connected layer for classification and output results. In addition, activation functions, batch normalization layers (BN layers), and dropout algorithms are often adopted for CNNs to further optimize the model training process.

Since all the experiments in this paper take a one-dimensional vibration signal as input, the basic unit of the one-dimensional CNN will be introduced in this section.

Convolutional layer. The main function of the convolutional layer is to extract the local features of the data. When data are input, the convolutional layer performs the convolution operation on the data in the form of weight sharing. In the convolution operation process, the convolution kernel performs ergodic convolution on the input data in a fixed stride to obtain the results (Figure 2). The calculation process of one-dimensional convolution is shown in equation (1).

\[
y_{l+1}^i(j) = f \left\{ \sum_{k=1}^{W} K_l^i(j) \times X_l^i(j) + b_l^i \right\} 
\]

where \( y_{l+1}^i(j) \) represents the value \( j \) of feature map \( i \) of layer \( l + 1 \); \( K_l^i(j) \) represents the weight value \( j \) of convolution kernel \( i \) at layer \( l \); \( X_l^i(j) \) represents the \( j \)th input value of the \( n \)th convolved region at layer \( l \); \( b_l^i \) represents the bias value of the \( i \)th convolution kernel at layer \( l \); and \( W \) represents the width of the convolution kernel.

**1D-Maxpooling layer.** The main function of the pooling layer is to reduce the dimension of the feature map output by the convolutional layer through a downsampling operation and extract more discriminant features in the feature maps. Similar to the convolution process, the downsampling operation of the pooling layer traverses the input data with a fixed stride through the pooling window to obtain the corresponding feature maps. In CNNs, commonly used pooling layers include the max pooling layer and average pooling layer. The one-dimensional max pooling layer is adopted in this paper, and the downsampling operation process is shown in Figure 3. The downsampling operation is performed on the input vector of \( 1 \times 5 \) according to the stride size 1 of the \( 1 \times 3 \) pooling window, and the maximum value of all neurons in the corresponding area of each pooling window is taken as the output. The specific calculation process of the downsampling operation is shown in equation (2).

\[
p_{l+1}^t = \max_{(j-l)W+1 \leq t \leq jW} \{ a_l^t(t) \} 
\]

where \( a_l^t(t) \) represents the activation value of neuron \( t \) in layer \( l \) of channel \( i \) and \( W \) represents the width of the pooling window.
Fully connected layer. The function of the fully connected layer (FC layer) is to fully connect and classify the high-level abstract features output by the previous hidden layer. For the whole CNN, the fully connected layer plays the role of "classifier." As shown in Figure 4, when the fully connected layer is fully connected and classified, the multidimensional feature vector output by the last hidden layer is tiled into a one-dimensional feature vector, and then the flattened one-dimensional feature vector is used as the input for full connection. Finally, the softmax activation function is used to transform the input neurons into probability distributions with and equal to 1. To facilitate the establishment of a multiclassification objective function. The specific expression of the fully connected layer is shown in equation (3).

\[ y^n = \sigma(w^n p^n + b^n) \]  

where \( y^n \) is the output of the \( n \) th fully connected layer, \( w^n \) is the weight matrix of the fully connected layer, \( p^n \) is the feature vector output of the last hidden layer, and \( b^n \) is the bias vector of the fully connected layer.

The softmax activation function is used in the fully connected layer to achieve classification. The expression of the softmax activation function is shown in (4).

\[ S_i = \frac{e^i}{\sum e^j} \]  

where \( S_i \) represents the output probability value of neuron \( i \) activated by the softmax function and \( e^i \) represents the output value of the \( i \) th neuron.

Long short-term memory network

LSTM is a special type of RNN that can effectively solve the problems of gradient disappearance and gradient explosion encountered by traditional RNNs when dealing with long-term dependent information. The reason why these two problems occur in traditional RNNs is that during the process of error backpropagation (BP), the gradient is constantly attenuated or enhanced, resulting in gradient disappearance or gradient explosion. However, in LSTM, to avoid these problems, as shown in Figure 5(a), a single LSTM memory unit is equipped with a forgetting gate, an input gate and an output gate, which can filter, update and output information, respectively. Each gate structure is composed of multiple hidden neurons, and the errors can be propagated with constant values among these hidden neurons. Therefore, the gradient disappearance or explosion of the network is avoided. In addition, as shown in Figure 5(b), LSTM can self-transfer and utilize its own state and output vector in the time dimension, so the sequence features of the input are retained in LSTM.

The working principle of the LSTM memory unit is shown in Figure 5(a). a time \( t \) is taken as an example. \( C_t \) denotes storage that accumulates the historical information states, and this storage can retain input long-term dependent information. In the forgetting gate, the
input data at time $t$ can be calculated by equation (5): information $f_t$ should be forgotten in the storage cell $C_{t-1}$ at time $t-1$. In the input gate, the input data at time $t$ can be calculated by equation (6): the information obtained through the joint decision of $i_t$ and $s_t$ should be updated and stored in the storage cell $C_t$ at time $t-1$, which becomes the storage cell $C_t$ at time $t$ after filtering and updating information. In the output gate, the input data at time $t$ can be calculated by equation (7): the storage cell $C_t$ outputs information $h_t$ according to $o_t$.

$$f_t = \sigma(W_f x_t + V_f h_{t-1} + b_f)$$  \hspace{1cm} (5)

$$i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$$

$$s_t = \tanh(W_s x_t + V_s h_{t-1} + b_s)$$ \hspace{1cm} (6)

$$C_t = f_t \odot C_{t-1} + i_t \odot s_t$$

$$o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$  \hspace{1cm} (7)

where $i_t$, $f_t$ and $o_t$ are the outputs of the input gate, the forget gate and the output gate, respectively; $W \in \mathbb{R}^{d \times k}$ and $V \in \mathbb{R}^{d \times d}$ are the shared weight matrices, and $b \in \mathbb{R}^d$ is the shared bias vector, which is updated iteratively; $k$ signifies the dimensions of the hidden vectors; $\sigma$ represents the activation function; and $\odot$ is the elementwise product operation.

**Transfer learning**

Transfer learning has been widely used in speech recognition, image classification and medical diagnosis due to its ability to solve problems in related but different domains. Based on the transfer learning strategy, the useful knowledge related to the target task learned in the source domain is transferred to the target task in the target domain so that it can complete the fault diagnosis task based on a small dataset under multiple working conditions. The application of transfer learning in rolling bearing fault diagnosis is taken as an example. Although it is difficult to obtain sufficient effective data in engineering practice, similar valid data can be obtained in large quantities in the laboratory. Therefore, with a large number of valid label data obtained in the laboratory as the auxiliary, the useful knowledge learned through experimental data can be transferred to the target task by transfer learning.

To solve the problem of failure diagnosis methods caused by the lack of valid data and different data distributions in engineering practice. Moreover, transfer learning is characterized by using similar auxiliary tasks to solve the target task. When encountering a new task, there is no need to retrain the weight parameters of the model to extract features but only input the weight parameters of the model trained in the source task through transfer learning and fine-tune them.

**The proposed method**

To solve the problem of bearing fault diagnosis methods caused by insufficient data and different data distribution, this paper proposes a transfer learning method based on WCNN-LSTM model, which takes one-dimensional vibration signals from small datasets as input and adopts fine-tuned transfer learning strategy. It is noted that we uses a wide convolution kernel for convolution, which can not only obtain a large receptive field with fewer parameters so that the model can extract the low-frequency features of signals but also inhibit the overfitting of the model. The proposed method is named 1D-WCLT, and its related theory is as follows.

**Data augmentation**

One of the characteristics of neural networks is that they need to be trained by massive data to improve the generalization ability of the model. To obtain sufficient data samples for the experiment, the utilized data augmentation method is overlap sampling. As shown in Figure 6, a fixed-length window with a fixed step length is used to slide and intercept samples, and one sample is obtained every time the window is slid. Therefore, if the signal with length $L$ is intercepted by sliding step $S$, ($L-X)/S$ samples with length $X$ can be obtained. It is worth mentioning that to ensure that each sample contains complete bearing fault features, the number of hash data points (the length of the sample) contained in the sample should not be less than the number of data points recorded by the acceleration sensor for one revolution of the bearing. In addition, the sample length used in the experiment in this paper is 1024.

**Wide kernel convolution operation**

Generally, when the convolutional layer extracts features through a convolution operation, the larger the receptive field is, the richer the extracted regional information will be. In bearing fault diagnosis, it is difficult to capture the low-frequency features of vibration signals, so it is expected to obtain a large receptive field.

**Figure 6.** The process of overlap sampling.
during feature extraction by convolution to increase the characterization information of extracted features and better extract the low-frequency features of vibration signals.\textsuperscript{32} The definition of a receptive field is that a pixel point on the feature map corresponds to the region on the input map, as shown in Figure 7. The receptive field size of the $1 \times 3 \times 1$ pixel point on the second feature map corresponds to the input map is $1 \times 3 \times 5$.

In classical two-dimensional convolutional neural networks such as VGG\textsuperscript{33} and ResNet,\textsuperscript{34} multiple small-size convolutional kernels are used to replace large-size convolutional kernels for feature extraction. For example, two $3 \times 3$ convolutional kernels are used to replace a $5 \times 5$ convolutional kernel to obtain a receptive field of the same size. Similarly, three $3 \times 3$ convolution kernels can be used to replace a $7 \times 7$ convolution kernel to obtain the same size of the receptive field. For these neural networks with two-dimensional images as input, the advantage of this method is not only to deepen the network depth and improve the network characterization ability but also to reduce the weight parameters of the network to prevent the network from overfitting (weight parameters $2 \times 3 \times 3 = 18 < 5 \times 5 = 25$, weight parameters $3 \times 3 = 27 < 7 \times 7 = 49$).

However, in the convolutional layer of the model proposed in this paper, one-dimensional convolution is carried out with one-dimensional data as input, as shown in Figure 7. In the two-layer convolution, a total of two $1 \times 3$ convolution kernels are used to obtain a $1 \times 5$ receptive field, and a total of six weight parameters are used. Different from two-dimensional convolution, a $1 \times 5$ receptive field can also be obtained with a $1 \times 5$ convolution kernel in the first layer of one-dimensional convolution, but only five weight parameters are used. By analogy, in one-dimensional convolution, a $1 \times 7$ receptive field convolved with three $1 \times 3$ convolution kernels requires 9 weight parameters, while convolved with a $1 \times 7$ convolution kernel requires only seven weight parameters. The above phenomenon shows that in a one-dimensional convolutional layer, within a certain range, the larger the convolution kernel size, the larger the receptive field, and the stronger the feature extraction capability of the model.

**WDCNN-LSTM model**

In order to obtain the superior performance of the diagnostic model, we propose a hybrid neural network combining wide-kernel CNN and LSTM as a diagnostic model, as shown in Figure 8, the one-dimensional CNN part (light blue part) and the LSTM part (light green

---

Figure 7. Multilayer one-dimensional convolution process.

Figure 8. Structure of WDCNN-LSTM model.
part) will extract the better spatial and sequential features form input data, then the fully connected layer part (light orange part) will classified these features and output what types of failures are they. Moreover, in the convolutional layer of the one-dimensional CNN part, a wide convolution kernel is used for local convolution, which enables the convolutional layer to better extract the low and medium frequency features of input data.

In the training process of the model, the forward propagation process for the input data involves extracting features from the low level to the high level. For example, in the image classification task, the low-level convolutional layer of a CNN can only extract features such as the curves and edges of images, while the high-level convolutional layer can extract the corresponding abstract features of images. Similarly, in the pretraining process of the WDCNN-LSTM model, low-level fault generalization features are extracted at the lower layers, while abstract features of fault type are extracted at the higher layers. Because the low-level generalization features of different types of faults are basically the same under different operating conditions, the weight parameters of lower layers learned to extract generalization features in source conditions can be transferred to target conditions by using transfer learning.

The structure and parameters of the WDCNN-LSTM model are shown in Figure 8, in which the parameters of the A(B) structure are expressed as follows: the “A” part outside the brackets represents the parameter layer, and the “B” part inside the brackets represents the setting parameters of the parameter layer. Taking Conv1d_1 (32, 7 x 1, 2) and FC_1 (units = 32) in the model as an example, Conv1d_1 represents the first one-dimensional convolutional layer, and (32, 7 x 1, 2) represents the number of convolution kernels. The sizes of the convolution kernels and convolution strides of the convolutional layer are 32, 7 x 1 and 2, respectively; FC_1 represents the first fully connected layer, and (Units = 32) indicates that there are 32 neurons, and the expressions of the other parameters are the same. During the pre-training of the WDCNN-LSTM model, the input data are first processed by multilayer convolution and pooling, and the local spatial features of the input data are extracted layer by layer from the low level to the high level. Then, the encoding spatial features are fed into the LSTM layer as input, and the sequential features inside the encoding features are further learned and extracted in the LSTM layer as output. Finally, the output features of the LSTM layer are fed into the fully connected layers as input for classification. After the expected results are output by forward propagation, the model weight parameters are optimized and adjusted continuously by backward propagation (BP) of the objective function and optimization algorithm to obtain the best model parameters for extracting data features.

Transfer learning strategy

The transfer learning strategy adopted in this paper is fine tuning, the fine-tuning process is to transfer the appropriate weight parameters of the model pretrained in Source Task (T_s) to Target Task (T_t) and then fine-tune them using a small amount of training data of the target domain in T_t. Fine-tuning strategies vary according to different situations. (1) When the data distributions of the source domain and target domain are different and the data quantity of the target domain is insufficient, the fine-tuning strategy is usually used to freeze only the convolutional part and fine-tune the part of the other higher layer and fully connected layers; (2) When the source domain data are similar to the target domain data distribution and the target domain data are sufficient, the usual fine-tuning strategy is to fine-tune the whole network of the model. (3) When the source domain data are similar to the target domain data distribution and the target domain data are insufficient, the usual fine-tuning strategy is to fine-tune only the last output classification layer.

Since the experiment in this paper is based on a small dataset for bearing fault diagnosis, the Fine-tuning strategy used is as follows: First, the first 8 layers of the transfer model are frozen, that is, the model weight parameters of extracting low-level generalization features obtained by these layers in T_s are retained. Then, the weight parameters of the fully connected layers of the model are initialized and fine-tuned with a small amount of target domain training data to obtain the weight of fully connected layer parameters consistent with T_t. Notably, since the fully connected layers partially act as a "classifier," the essence of fine-tuning is to replace the classifier and retrain the new classifier.

Implementation flow chart of the proposed method

The main process of the proposed method is divided into four parts: data preprocessing, model pre-training, model fine-tuning and testing. As shown in Figure 9, during data preprocessing, overlapping sampling of bearing vibration signals is first carried out to obtain data samples for experiments. In the T_s, a sufficient number of samples are collected for model pretraining. In the T_t, a small number of data samples are collected to fine-tune and test the transfer model. Then the model is pre-training to obtain appropriate weight parameters for transfer learning. Next, according to the transfer learning strategy, the appropriate model parameters obtained by pre-training in T_s are transferred to T_t, and a small amount of target domain training data in T_t is used to fine-tune the model. Finally, using the test dataset of target domain to test the fine-tuned model and obtain the fault diagnosis results.
Experimental validation

This section is mainly divided into two parts. First, the setting process of the validation experiment is introduced, including the bearing dataset adopted, experimental implementation details and contrast methods. Then, the experimental results are analyzed, including selecting the best wide convolution kernel parameters, analyzing the fault diagnosis performance when the data come from the same device and the fault diagnosis performance when the data come from different devices.

Data description

The vibration acceleration signal of the bearing contains the operation information of the bearing, and the health status of the bearing can be determined by feature extraction. Therefore, the vibration signal is used as the experimental data of bearing fault diagnosis in this paper. In the fault diagnosis experiment in this paper, two different kinds of bearing datasets are used to verify the validity of the proposed method: one is the bearing dataset of Case Western Reserve University (CWRU), and the other is the dataset from the high-speed train axle box bearing test platform built independently.

CWRU dataset. The CWRU bearing dataset is a kind of public bearing dataset\textsuperscript{35–37} and is also one of the authoritative datasets for verifying bearing fault diagnosis methods. The data acquisition test rig for this dataset is shown in Figure 10, which consists of a 2 HP motor, a torque sensor, a power meter and an electronic controller. An acceleration sensor is installed on the motor’s driving end and fan end housing to collect vibration signals. The sampling frequencies of the bearing fault at the driving end are 12 and 48 kHz, and the sampling frequency of the bearing fault at the fan end is only 12 kHz. In the experiment, an SKF6205 deep groove ball bearing was taken as the research object,
and the fault damage was obtained by electric discharge machining (EDM). The main fault locations included the inner ring, outer ring and ball of the bearing, and the damage at each fault location had three diameter sizes of 0.007, 0.014, and 0.021 in, respectively. Therefore, a total of nine fault damage types can be processed in the experiment. It is worth noting that there are 3 O’clock, 6 O’clock, and 12 O’clock damage positions in the outer ring fault, but the 6 O’clock fault data are most commonly used in fault diagnosis.

In this paper, the data of the driving end with a sampling frequency of 12 kHz are selected as the experimental validation data, as shown in Table 1, which includes the datasets of four working conditions A, B, C, and D. Motors in different working conditions have different loads and speeds, so the vibration signals collected in different working conditions also have different feature distributions. In addition, as shown in Table 2, the datasets of each working condition contain bearing data of 10 different health conditions (bearing data of 9 damage types plus bearing data of normal conditions).

Table 1. Information on the four working conditions of the CWRU bearing dataset.

| Dataset | Motor load (hp) | Motor speed (rpm) |
|---------|----------------|------------------|
| A       | 0              | 1797             |
| B       | 1              | 1772             |
| C       | 2              | 1750             |
| D       | 3              | 1730             |

Table 2. Fault information of the CWRU bearing dataset.

| Name  | Label | Fault location                  | Fault diameter (in) |
|-------|-------|---------------------------------|---------------------|
| Normal| 0     | Normal                          | -                   |
| IR007 | 1     | Inner race (Centered at 6:00)   | 0.007               |
| IR014 | 2     | Inner race                      | 0.014               |
| IR021 | 3     | Inner race                      | 0.021               |
| OR007 | 4     | Outer race (Centered at 6:00)   | 0.007               |
| OR014 | 5     | Outer race (Centered at 6:00)   | 0.014               |
| OR021 | 6     | Outer race (Centered at 6:00)   | 0.021               |
| BO007 | 7     | Ball                            | 0.007               |
| BO014 | 8     | Ball                            | 0.014               |
| BO021 | 9     | Ball                            | 0.021               |

Train axle box bearing dataset. The data acquisition test rig of the train axle box bearing dataset is shown in Figure 11, which is mainly composed of a transmission system, electrical control system, data acquisition device and computer. During the test, F-807811.02. A TAROL double-row tapered bearing assembled by a CRH380 train was taken as the experimental object. The motor fixed on the base provided a driving force to drive the spindle to rotate to simulate the bearing speed corresponding to any running speed of the high-speed train.

To simulate the real train bearing fault morphology, similar to the previous section, different types of fault damage were obtained by manual processing. Similar to the common fault damage positions of train bearings in practice, the damage positions processed in the experiment included the inner ring, outer ring and ball, and the damage types of each position included early failure (0.2 mm) and severe damage (0.8 mm). Therefore, experimental data of six damage types could be obtained. In addition, the fault damage positions of machining in the test also include cage and some compound fault damage, which includes the inner and outer ring (IO) and outer ring and ball (OB), respectively, and their damage depth is 0.6 mm.

The train axle box bearing data selected in this paper were collected under the condition that the spindle speed was 590 r/min, corresponding to the actual train speed of 100 km/h, and the sampling frequency of the collected data was 5120 Hz. As shown in Table 3, the adopted dataset E contains bearing data of 10 different health conditions to verify the diagnostic classification performance of the proposed method.

Figure 11. Train axle box bearing data acquisition test rig.

Experimental setup

Dataset setup. To verify the performance of the 1D-WCLT method for bearing fault diagnosis based on small datasets, two types of datasets are used for validation in this paper. The sources of these two types of datasets are different. One of them is from the CWRU bearing dataset, which contains datasets A, B, C, and D of four different working conditions. The other is derived from the dataset of train axle
box bearings, which only contains the dataset E of one working condition. Since the transfer learning strategy is adopted in the 1D-WCLT method, it is necessary to consider whether the source domain data and the target domain data come from the same device. When source domain data and target domain data come from the same device, the feature distribution difference is small. When source domain data and target domain data come from different devices, the feature distribution difference is large. For example, datasets A, B, C, and D come from the same device, while dataset E comes from another device. Therefore, if the 1D-WCLT method can still maintain good performance in fault diagnosis based on small datasets in the above two cases, it indicates that the method can not only accurately complete fault diagnosis when data are insufficient but also has strong domain adaptability.

**Implementation details.** The detailed parameter settings utilized in the experiments are as follows: during the pretraining and fine-tuning of the model, the mini-batch size of training is set as 16, the loss function is cross-entropy, the optimizer is Adam, the initial adaptive learning rate used in the training process is 0.0001, and the number of training iterations is 50 epochs. To avoid accidental experimental results, the average of five repeated experimental test results was taken as the final test result for each experiment.

**Contrast method.** To compare performance differences between the 1D-WCLT method and other state-of-the-art fault diagnosis methods, five deep learning methods are adopted in this paper for comparison, including common bearing fault diagnosis methods 1D-DCNN and 1D-WDCNN based on the DCNN model and WDCNN model, respectively, with a one-dimensional vibration signal as input; a method for bearing fault diagnosis based on the DCNN-LSTM model with a 2D continuous wavelet transform (CWT) time-frequency diagram as input; and the bearing fault diagnosis methods 2D-DCNN-TL and 2D-DCNN-LSTM-TL based on the DCNN model and DCNN-LSTM model for transfer learning using the 2D CWT time-frequency diagram as input. The main structure and parameters of the diagnostic model mentioned are shown in Table 4, where Conv5 represents the convolutional layer with a convolution kernel size of 5 (other Conv + numbers have similar meanings), Pool represents the maximum pooling layer, and FC64 represents the fully connected layer with a neuron number of 64 (other FC + numbers have similar meanings).

**Experimental results and analysis**

**Parameter selection of the wide convolution kernel.** In order to verify the necessity of WCNN-LSTM model using wide convolution kernel, and try to find out the number of wide convolution layers and the size of wide convolution kernel that are most suitable for WCNN-LSTM model, this section first sets four WCNN-LSTM models with different number of wide-kernel convolution layers to verify the optimal number of wide-kernel convolution layers. Then, six different convolution kernel sizes were set according to the optimal number of convolution layers obtained, so as to obtain the optimal convolution kernel size.

**Selection of the number of wide kernel convolutional layers.** In this section, four WCNN-LSTM models with different number of wide-kernel convolution layers are set to verify the optimal number of wide-kernel convolution layers. The number of wide kernel convolution layers included by the four models is 1, 2, 3 and 4 respectively, and the convolution kernel size corresponding to the wide kernel convolution layer in each model is 32, 32-16, 32-16-8 and 32-16-8-4 respectively. Where, the number of values separated by “-” corresponds to the number of the wide convolution kernel layers, and the values represents the size of the wide convolution kernel of the corresponding layer. It can be noted that the size of the convolution kernel of the first layer is larger than that of the second. That is because the first layer of convolution extracts relatively low-level generalization features, and the local convolution region needs to be added to extract more global information to better

### Table 3. Details of the train axle box dataset.

| Dataset | Name | Label     | Fault location | Fault diameter (mm) |
|---------|------|-----------|----------------|---------------------|
| E       | N    | Normal    | -              | -                   |
|         | IR02 | Inner race| 0.2            |                     |
|         | IR08 | Inner race| 0.8            |                     |
|         | OR02 | Outer race| 0.2            |                     |
|         | OR08 | Outer race| 0.8            |                     |
|         | B02  | Ball      | 0.2            |                     |
|         | B08  | Ball      | 0.8            |                     |
|         | C    | Cage      | 0.6            |                     |
|         | IO   | Inner race & Outer race | 0.6 | |
|         | OB   | Outer race & Ball | 0.6 | |

### Table 4. The main structure and parameters of the diagnostic model.

| Mode      | Structure                          |
|-----------|------------------------------------|
| DCNN      | Conv5-Pool-FC128-FC64-FC10         |
| WDCNN     | Conv64-Pool-Conv3-Pool-Conv3-Pool-FC64-FC10 |
| DCNN-LSTM | Conv3-Pool-Conv3-Pool-LSTM64-FC64-FC10 |
capture low-frequency features, while the features extracted by the second layer are relatively abstract, so the local convolution region needs to be reduced to obtain more detailed feature information.

In the above setting experiment, the dataset of working condition A and working condition E of the two types of bearing data will be used for validation respectively in this section. The dataset of each working condition will be divided into training set and test set. The training set contains 2500 samples (250 for each fault type) for model training. The test set contains 1000 samples (100 for each fault type) for model testing.

The A working condition data in CWRU dataset and the E working condition data in train axle box dataset were used to repeat each of the above experiments for 5 times to obtain their average test accuracy and average test loss. As shown in Figures 12 to 15, it can be concluded from the experimental results of the two datasets that when the wide kernel convolution layer of the WCNN-LSTM model is 2, the obtained bearing fault diagnosis accuracy is the maximum and the loss value is the minimum. When the number of model layers is 1, the model is underfitting, which leads to the failure of extracting data features comprehensively and the performance is insufficient. When the number of model layers is more than 2, the model overfitting, which leads to the failure to extract effective features accurately and the performance deteriorates. Therefore, the WCNN-LSTM model with a wide convolution kernel layer of 2 was selected as the diagnostic model in this paper.

Selection of size of wide convolution kernel. In this section, different convolution kernel sizes are set for the two convolution layers in the WCNN-LSTM model for

![Figure 12. Average test accuracy of four WCNN-LSTM models based on CWRU dataset.](image1)

![Figure 13. Average test loss value of four WCNN-LSTM models based on CWRU dataset.](image2)

![Figure 14. Average test accuracy of four WCNN-LSTM models based on train axle box bearing dataset.](image3)

![Figure 15. Average test loss value of four WCNN-LSTM models based on train axle box bearing datasets.](image4)
verification. Six different convolution kernel size groups are set, which are 5-3, 8-4, 16-8, 32-16, 64-32, and 128-64.

The experimental Settings are consistent with the above, and the results are as follows. Figures 16 and 17 show the average test accuracy and average test loss of each group of experiments based on the A condition data in the CWRU bearing dataset, which are repeated for five times. It can be seen that the larger the convolution kernel width is, the larger the average test accuracy and the smaller the average test loss. When the convolution kernel width reaches 64-32, the model’s average test accuracy is the largest, reaching 100%, and the average test loss value is the smallest, 0.02743. Of course, the convolution kernel width of the model cannot be increased blindly. When it reaches 128-64, the average test accuracy decreases, and the average test loss value increases.

The results of experimental validation using dataset E in train axle box bearing data are shown in Figures 18 and 19. Similar to the validation results using dataset A, the average test accuracy of the WDCNN-LSTM model increases with increasing convolution kernel width, and the average test loss value decreases with increasing convolution kernel width. When the convolution kernel width increases to 64-32, the average test accuracy reaches a maximum of 99.48%, and the average test loss value reaches a minimum of 0.05638. When the convolution kernel width increases to 128-64, the average test accuracy decreases, and the average test loss value increases. Thus, it can be inferred from the above results that when the convolution kernel is too small, it is difficult for the convolution process to capture the low- and medium-frequency features of vibration signals; when the convolution kernel is too large, the time-domain resolution is low, resulting in
Table 5. Six transfer conditions when source domain and target domain data come from the same device.

| Source domain | Target domain |
|---------------|---------------|
| B             | C             | D             |
| B             | -             | B → C         | B → D         |
| C             | C → B         | -             | C → D         |
| D             | D → B         | D → C         | -             |

In all the validation experiments in this section, 2500 samples of source domain data (250 samples for each fault type) are taken as training sets for pretraining for the method involving transfer learning. Target domain data are divided into a training set and a test set, which are used for the fine-tuning and testing of the transfer model, including 200/300/500/800/1000/1500/2000/2500 (each fault type 20/30/50/80/100/150/200/250 of the sample). Six different numbers of training sets are used to fine-tune the transfer model. A test set with 1000 data (100 samples for each fault type) was used to test the transfer model. However, for the method that does not involve transfer learning, as the source domain data are not required for pretraining, the samples only need to be divided into training sets and test sets. The number of samples in the training set is 200/300/500/800/1000/1500/2000/2500 (20/30/50/80/100/150/200/250 for each fault type), and the number of samples in the test set is 1000 (100 for each fault type).

Domain adaptability comparison. Experiments were carried out for validation, and the experimental results obtained are shown in Figures 20 and 21, which are the average test accuracy and average test loss values of each comparison method under different conditions, respectively. Among them, since the 1D-DCNN method, 1D-WDCNN method and 2D-DCNN-LSTM do not carry out transfer learning, the test results of these three methods remain unchanged in the transfer condition with the same target domain (because no transfer learning is carried out, the same target domain means the same working condition. For example, the test results of these three methods in the C → B condition are consistent with those in the D → B condition. As seen from Figure 20 no matter under any transfer condition, the red curve representing the average test accuracy of the 1D-WCLT method is at the top compared with other methods. Moreover, it can be seen that the accuracy obtained by other methods in different working conditions based on the same number of data sets is too different. For example, the accuracy obtained by 2D-CNN-TL method in working conditions B → D and D → B based on the training set with sample size of 200 is nearly 50% different. Similarly, it can be seen from Figure 21 that the loss value of 1D-WCLT method is the smallest under any working condition, indicating that it can achieve better performance under various working conditions.

In summary, the 1D-WCLT method has strong domain adaptability in bearing fault diagnosis because it can achieve high diagnostic accuracy under a variety of different working conditions.

Stability comparison. As shown in Figure 20, compared with other methods, the fluctuation of the average test accuracy curve of the 1D-WCLT method is the smallest in each condition. In addition, it can also be seen from Figure 20 that the 1D-WDCNN method and 2D-CNN-TL method are greatly affected by the amount of data in the training set and show obvious large fluctuations. Similarly, in Figure 21, the fluctuation of the average test loss curve of the 1D-WCLT method in each working condition is the smallest, which also indicates that
Figure 20. The average test accuracy of each comparison method based on different data amounts under six different conditions: (a) B → C, (b) B → D, (c) C → B, (d) C → D, (e) D → B, and (f) D → C.

Figure 21. The average test loss value of each comparison method based on different data amounts under six different conditions: (a) B → C, (b) B → D, (c) C → B, (d) C → D, (e) D → B, and (f) D → C.
the diagnostic performance of the 1D-WCLT method in rolling bearing fault diagnosis is very little affected by the sample size of the training set and has strong stability.

Performance comparison of fault diagnosis based on small datasets. The experimental results of rolling bearing fault diagnosis by all comparison methods based on small datasets (the sample size of training set is 200 or 300) are shown in Figure 20. The average test accuracy of the 1D-WCLT method is significantly higher than that of the other methods under all transfer conditions. Similarly, in Figure 21, compared with other methods, the 1D-WCLT method has the smallest average test loss value for fault diagnosis under any transfer condition.

Furthermore, in order to further compare the performance of fault diagnosis based on small data sets, 30 samples of each fault type are taken in this subsection to form a training set with a total amount of 300 for verification. And the T-distributed Stochastic Neighbor Embedding (T-SNE) algorithm is used to make clustering visualization of their classification results. It can be seen from Figure 20 that the average test accuracy values of each method under B→D working condition have the smallest difference, so it is more convincing to choose the classification results under B→D working condition for visualization. As shown in Figure 22, the clustering performance of the classification results of the 1D-WCLT method is better than that of the other methods. The samples of the 10 fault types in the cluster diagram are not only clearly divided into 10 categories, but the centers of each category are also far away from each other, and there is no overlapping disorder. This indicates that among all the methods, the 1D-WCLT method has the best classification performance for fault diagnosis based on small datasets.

Therefore, it can be concluded from the above phenomenon that in the case that the source domain and target domain data come from the same device, compared with other state-of-the-art rolling bearing fault diagnosis methods, the 1D-WCLT method can not only better complete the fault diagnosis task based on small datasets, but also has strong domain adaptability.

Fault diagnosis performance analysis for data from the different devices. Similar to the previous section to verify the performance of the 1D-WCLT method in rolling bearing fault diagnosis when source domain data and target domain data come from different devices, this section divides the experimental validation into two parts. First, to verify the bearing fault diagnosis performance of each comparison method under different transfer conditions, four datasets A, B, C, and D derived from CWRU bearing data were paired with dataset E derived from train axle box bearing data for transfer learning, as shown in Table 6. There are a total of four transfer condition datasets. Then, to verify the bearing fault diagnosis performance of each comparison method based on small datasets, eight different numbers of training sets were set in each transfer condition for training the diagnosis models of each method.

Figure 22. Visual clustering diagram of test results based on small datasets for each comparison method (B→D): (a) 1D-DCNN, (b) 1D-WDCNN, (c) 2D-DCNN-TL, (d) 2D-DCNN-LSTM, (e) 2D-DCNN-LSTM-TL, and (f) 1D-WCLT.
Moreover, in all validation experiments in this section, 2500 samples of source domain data are taken as training sets for pretraining in the method involving transfer learning. Target domain data are divided into a training set and a test set, including 200/300/500/800/1000/1500/2000/2500. Eight different numbers of training sets are used to fine-tune the transfer model, and a test set with 1000 data points is used to test the transfer model. In addition, for the methods that do not involve transfer learning, the training set of sample numbers for 200/300/500/800/1000/1500/2000/2500 and the test set of the sample is 1000.

**Domain adaptability comparison.** When the source domain data and target domain data come from different devices, the results obtained by the experiment are shown in Figures 23 and 24, which are the average test accuracy and average test loss values obtained by each comparison method under different transfer conditions based on different data sizes. As seen from Figure 23, compared with other methods, the average test accuracy curves of the 1D-WCLT method are all at the top in the four transfer conditions, indicating that its diagnostic performance is better. Moreover, when the number of training sets is more than 300, the accuracy can be more than 90% under the four working conditions, and even 100% when the amount of training set data is 2500. Similar to the average test accuracy result, As shown in Figure 24, the average test loss value of 1D-WCLT method is minimum in every condition, and even drops to close to 0 when the training set data amount is 2500. Therefore, it can be concluded that, compared with other state-of-the-art methods, the 1D-WCLT method can achieve higher diagnostic accuracy in rolling bearing fault diagnosis and has strong domain adaptability.

**Stability comparison.** As shown in Figure 23, compared with other methods, the 1D-WCLT method can achieve

| Source domain | A | B | C | D |
|---------------|---|---|---|---|
| E             | E→A | E→B | E→C | E→D |

Figure 23. The average test accuracy of each comparison method based on different data amounts under four different conditions: (a) E→A, (b) E→B, (c) E→C, and (d) E→D.
high accuracy and minimum fluctuation in fault diagnosis based on training sets of different data amounts under any transfer conditions. For example, under E→C condition, the accuracy value of 2D-CNN-TL method fluctuates irregularly with the change of the number of training sets. In E→B, E→C, and E→D conditions, it can be seen that the accuracy value of 1D-DCNN method differs greatly when the number of training sets is 200 and 2500. In the same way, it can also be seen in Figure 24 that the fluctuation of the average test loss curve of the 1D-WCLT method is minimal no matter under any transfer condition. Therefore, it also indicates that in rolling bearing fault diagnosis, the diagnostic performance of 1D-WCLT method is little affected by the sample size of training set, and has very stable diagnostic performance.

**Performance comparison of fault diagnosis based on small datasets.** The experimental results of rolling bearing fault diagnosis by each comparison method based on small datasets (the sample size of training set is 200 or 300) are shown in Figure 23. In the four transfer conditions given, the average test accuracy of 1D-WCLT based on small datasets is greater than that of other methods based on small datasets. Moreover, as shown in Figure 24, the average test loss value of 1D-WCLT method based on small datasets is significantly smaller than that of other methods based on small datasets under the four working conditions.

Moreover, under E→D condition, except 2D-CNN-TL method, the average test accuracy values of the other methods for bearing fault diagnosis based on small datasets have small differences. Therefore, the classification results under E→D condition are selected for clustering visualization. As shown in Figure 25, based on the dataset (the sample size of the training set is 300) used to execute fault diagnosis, the clustering performance of the classification results of the 1D-WCLT method is better than that of the other methods. The samples of the 10 fault types in the cluster diagram are not only clearly divided into 10 categories, but the centers of each category are also far away from each other. Therefore, it also indicates that among all the methods, the 1D-WCLT method has the best performance in fault diagnosis based on small datasets.

It can be concluded from the above phenomenon that, compared with other state-of-the-art rolling bearing fault diagnosis methods, when source domain data and target domain data come from different devices, the 1D-WCLT method can not only accurately
complete the fault diagnosis task based on small data-
sets, but also has strong domain adaptability.

**Computational complexity comparison of the methods.** The computational complexity of fault diagnosis methods is generally measured by the number of trainable parameters. The larger the number of trainable parameters is, the larger the computation amount is required and the longer the running time is.

The number of pre-training parameters and fine-tuning parameters of each method is shown in Table 7, where the total trainable parameters of the transfer learning methods includes the number of pre-training parameters and fine-tuning parameters, while the total trainable parameters of the non-transfer learning methods only includes the number of pre-training parameters. It can be seen that the total number of trainable parameters of the diagnosis method using two-dimensional image as input is less than that using one-dimensional vibration signal as input. In addition, the transfer learning methods will increase the total number of trainable parameters due to plus fine-tuning parameters. However, compared with the non-transfer learning method, the transfer learning method has strong domain adaptability, so it does not need to repeat the pre-training in the new working conditions, and only needs pre-training one time to complete the fault diagnosis in a variety of working conditions by fine tuning. Therefore, for multi-condition fault diagnosis, the 1D-WCLT method does not need to be re-pretrained, and the number of calculated parameters to be fine-tuned is 2410, while the 1D-DCNN, 1D-WDCNN and 2D-DCNN-LSTM methods need to be re-pretrained under new conditions, and the number of trainable parameters are 4,171,338, 1,957,834, and 99,466, respectively. Although the 2D-DCNN-TL and 2D-DCNN-LSTM-TL methods do not require re-pretraining, their fine-tuning parameters are 59,134 and 4810, respectively, which are much larger than the 2410 of the 1D-WCLT method.

In summary, for multi-working condition bearing fault diagnosis, transfer learning method does not need to repeat pre-training in each working condition, so the number of trainable parameters for fault diagnosis is small, the memory occupied is small, and the

![Figure 25. Visual clustering diagram of test results based on small datasets for each comparison method (E→D): (a) 1D-DCNN, (b) 1D-WDCNN, (c) 2D-DCNN-TL, (d) 2D-DCNN-LSTM, (e) 2D-DCNN-LSTM-TL, and (f) 1D-WCLT.](image_url)
computational complexity is also small. Moreover, among the six methods, the 1D-WCLT method has the least number of fine-tuning parameters, so it has the least amount of calculation for fault diagnosis under multiple working conditions.

Conclusion

To solve the problem of insufficient valid data and different data distributions under multiple working conditions, a fault diagnosis method called 1D-WCLT based on the WDCNN-LSTM model is developed. The proposed method adopts a wide convolution kernel for local convolution, which can effectively increase the types of extracted sequential and spatial features and achieve higher classification accuracy for the fault diagnosis task with small valid data in another working condition. The validation of the CWRU dataset and train axle box bearing dataset show that, compared with other state-of-the-art deep learning methods, whether the training set and test set come from the same device or from different devices, the proposed method can accurately complete the fault diagnosis task based on a small dataset under various working conditions. It is noted that the 1D-WCLT will be a new alternative tool for intelligent fault detection of rotary machines.

Acknowledgements

We would like to thank CWRU for offering bearing data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported in part by the Opening Project of The State Key Laboratory of Heavy Duty AC Drive Electric Locomotive Systems Integration (No. 2020ZJKF05), and in part by the Natural Science Foundation of Sichuan Province (No. 2022 NSFSC0400).

ORCID iD

Yongjian Li https://orcid.org/0000-0001-5570-9242

References

1. Shen S, Lu H, Sadoughi M, et al. A physics-informed deep learning approach for bearing fault detection. Eng Appl Artif Intell 2021; 103: 104295.
2. Milo MW, Harris B, Bjerke B, et al. Anomaly detection in rolling element bearings via hierarchical transition matrices. Mech Syst Signal Process 2014; 48: 120–137.
3. Jia F, Lei Y, Lin J, et al. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech Syst Signal Process 2016; 72–73: 303–315.
4. Jia F, Lei Y, Guo L, et al. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. Neurocomputing 2018; 272: 619–628.
5. Lei Y, Yang B, Jiang X, et al. Applications of machine learning to machine fault diagnosis: a review and road-map. Mech Syst Signal Process 2020; 138: 106587.
6. Tu YH, Du J, Sun L, et al. An iterative mask estimation approach to deep learning based multi-channel speech recognition. Speech Commun 2019; 106: 31–43.
7. Maier A, Syben C, Lasser T, et al. A gentle introduction to deep learning in medical image processing. Z Med Phys 2019; 29: 86–101.
8. Ahmed HOA and Nandi AK. Three-stage hybrid fault diagnosis for rolling bearings with compressively sampled data and subspace learning techniques. IEEE Trans Ind Electron 2019; 66: 5516–5524.
9. He M and He D. Simultaneous bearing fault diagnosis and severity detection using a LAMSTAR network-based approach. IET Sci Meas Technol 2018; 12: 893–901.
10. Kiranyaz S, Avci O, Abdeljaber O, et al. 1D convolutional neural networks and applications: a survey. Mech Syst Signal Process 2021; 151: 107398.
11. Choudhary A, Mian T and Fatima S. Convolutional neural network based bearing fault diagnosis of rotating machine using thermal images. Measurement 2021; 176: 109196.
12. Guo X, Chen L and Shen C. Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. Measurement 2016; 93: 490–502.
13. Zhang W, Li C, Peng G, et al. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mech Syst Signal Process 2018; 100: 439–453.
14. Lei J, Liu C and Jiang D. Fault diagnosis of wind turbine based on long short-term memory networks. Renew Energy 2019; 133: 422–432.
15. Liu J, Pan C, Lei F, et al. Fault prediction of bearings based on LSTM and statistical process analysis. Reliab Eng Syst Saf 2021; 214: 107646.
16. Gao D, Zhu Y, Ren Z, et al. A novel weak fault diagnosis method for rolling bearings based on LSTM considering quasi-periodicity. Knowl Based Syst 2021; 231: 107413.
17. Wu Z, Jiang H, Lu T, et al. A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data. Knowl Based Syst 2020; 196: 105814.
18. Guo L, Lei Y, Xing S, et al. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. IEEE Trans Ind Electron 2019; 66: 7316–7325.
19. Hoang DT and Kang HJ. A motor current signal-based bearing fault diagnosis using deep learning and information fusion. IEEE Trans Instrum Meas 2020; 69: 3325–3333.
20. Li C, Zhang S, Qin Y, et al. A systematic review of deep transfer learning for machinery fault diagnosis. Neurocomputing 2020; 407: 121–135.
21. Su H, Xiang L, Hu A, et al. A novel method based on meta-learning for bearing fault diagnosis with small sample learning under different working conditions. *Mech Syst Signal Process* 2022; 169: 108765.

22. Yang X, Liu B, Xiang L, et al. A novel intelligent fault diagnosis method of rolling bearings with small samples. *Measurement* 2022; 203: 111899.

23. Lu N, Hu H, Yin T, et al. Transfer relation network for fault diagnosis of rotating machinery with small data. *IEEE Trans Cybern* 2021; 52: 11927–11941.

24. Zhang W, Peng G, Li C, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors* 2017; 17: 425.

25. Olshausen BA and Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 1996; 381: 607–609.

26. Xia M, Li T, Xu L, et al. Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks. *IEEE/ASME Trans Mechatron* 2018; 23: 101–110.

27. Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* 2019; 31: 1235–1270.

28. Muzaffar S and Alfshari A. Short-term load forecasts using LSTM networks. *Energy Prog* 2019; 158: 2922–2927.

29. Graves A and Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005; 18: 602–610.

30. Pan SJ and Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010; 22: 1345–1359.

31. Jin T, Yan C, Chen C, et al. Light neural network with fewer parameters based on CNN for fault diagnosis of rotating machinery. *Measurement* 2021; 181: 109639.

32. Han Y, Tang B and Deng L. An enhanced convolutional neural network with enlarged receptive fields for fault diagnosis of planetary gearboxes. *Comput Ind* 2019; 107: 50–58.

33. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* preprint arXiv 2014; 1409.1556.

34. Shehab LH, Fahmy OM, Gasser SM, et al. An efficient brain tumor image segmentation based on deep residual networks (ResNets). *J King Saud Univ Eng Sci* 2021; 33: 404–412.

35. Smith WA and Randall RB. Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. *Mech Syst Signal Process* 2015; 64–65: 100–131.

36. Li Y, Wang X, Si S, et al. Entropy based fault classification using the Case Western Reserve University data: a benchmark study. *IEEE Trans Reliab* 2020; 69: 754–767.

37. Neupane D and Seok J. Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: a review. *IEEE Access* 2020; 8: 93155–93178.

38. Song X, Cong Y, Song Y, et al. A bearing fault diagnosis model based on CNN with wide convolution kernels. *J Ambient Intell Humaniz Comput* 2022; 13: 4041–4056.

39. Peng C, Cheng W, Song Z, et al. A noise-robust modulation signal classification method based on continuous wavelet transform. In: 2020 IEEE 5th Information technology and mechatronics engineering conference (ITOEC), Chongqing, China, 12–14 June 2020, pp.745–750. New York: IEEE.

40. Jadhav P, Rajguru G, Datta D, et al. Automatic sleep stage classification using time–frequency images of CWT and transfer learning using convolution neural network. *Biocybern Biomed Eng* 2020; 40: 494–504.