Hindawi
Shock and Vibration
Volume 2020, Article ID 1971945, 10 pages
https://doi.org/10.1155/2020/1971945

*Research Article*

# Multitask Convolutional Neural Network for Rolling Element Bearing Fault Identification

**Mingxing Jia**,[1] **Yuemei Xu**,[1] **Maoyi Hong**,[1] **and Xiyu Hu**[2]

[1]*College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China*
[2]*Shenyang Special Equipment Inspection and Research Institute, Shenyang 110819, Liaoning, China*

Correspondence should be addressed to Mingxing Jia; jiamingxing@ise.neu.edu.cn

Received 15 November 2019; Revised 20 February 2020; Accepted 12 May 2020; Published 14 July 2020

Academic Editor: Paola Forte



As one of the most vital parts of rotating equipment, it is an essential work to diagnose rolling bearing failure. The traditional signal processing-based rolling bearing fault diagnosis algorithms rely on artificial feature extraction and expert knowledge. The working condition of rolling bearings is complex and changeable, so the traditional algorithm is slightly lacking adaptability. The damage degree also plays a crucial role in fault monitoring. Different damage degrees may take different remedial measures, but traditional fault-diagnosis algorithms roughly divide the damage degree into several categories, which do not correspond to the continuous value of the damage degree. To solve the abovementioned two problems, this paper proposes a fault-diagnosis algorithm based on "end-to-end" one-dimensional convolutional neural network. The one-dimensional convolution kernel and the pooling layer are directly applied to the original time domain signal. Feature extraction and classifier are merged together, and the extracted features are used to judge the damage degree at the same time. Then, the generalization ability of the model is studied under a variety of conditions. Experiments show that the algorithm can achieve more than 99% accuracy and can accurately give the damage degree of the bearing. It has good performance under different speeds, different types of motors, and different sampling frequencies, and so it has good generalization ability.

## 1. Introduction

Rotating machinery and equipment is developing in the large-scale, refined, and systematic direction. Rolling bearing is one of the most important components of rotating machinery and equipment. The probability of mechanical equipment fault caused by the failure of rolling bearings is as high as 30%, so reliable condition monitoring of rolling bearings is critical for the normal operation of mechanical equipment [1].

It is difficult to directly monitoring the state of the bearing due to its small size and the concealed mounting position. Usually, the condition of bearing is diagnosed by analyzing the vibration signal [2–5]. Recently, the machine learning technology [6], especially deep learning [7], is rapidly developing. Many models of deep learning such as Stacked Auto Encoder (SAE), Deep Belief Network (DBN), and Convolutional Neural Network (CNN) have been widely applied to the diagnosis of rolling bearings [8]. Junbo et al. [9] used the stacking self-encoder to construct a deep neural network. The fault features are extracted by stacking the intermediate hidden layers by using the characteristics of the input and output of the encoder. Lu et al. [10] used a denoising self-encoder to improve the robustness of the model. The denoising input from the encoder randomly removed some of the input signals to improve the stability of the hidden layer extraction features. Gan et al. [11] used the Deep Belief Network to propose the Hierarchical Diagnosis Network (HDN). The method first uses DBN to determine the location of the bearing damage and then uses DBN to diagnose the size of the damage.

Convolutional Neural Networks (CNNs) have had many major breakthroughs in the field of image recognition [12, 13]. In addition, they are widely used in the field of fault diagnosis [14, 15]. Janssens et al. [16] used the shallow convolution neural network model to make discrete Fourier

Transform of the original signal, transform the original time-domain vibration signal into the frequency domain, and use the extracted useful features as input to the one-dimensional convolutional neural network. Jia et al. [17] proposed a four-layer deep neural network, which performs Fast Fourier Transform on 2400 data as the network input signal, then uses the self-encoder to perform unsupervised training layer by layer, and finally carries out supervised training. Wen et al. [18] proposed a convolution neural network model, which was based on LeNet. The original vibration signal was transformed into two-dimensional gray image as input of the model, and the fault diagnosis of the bearing and centrifugal pump was realized. Yuan et al. [19] constructed a two-dimensional CNN model and extracted the time-frequency map of the rolling bearing as the input of the two-dimensional CNN. Wei et al. [20] proposed a deep one-dimensional convolutional neural network model, using dropout technology and small batch training, to achieve accurate and stable diagnosis based on original vibration signals.

The abovementioned research studies have already achieved good diagnostic results, and even some models have a diagnostic accuracy of 99%. However, they only gave a judgment on the category of the fault, without further considering the extent of the fault. Cococcioni et al. [21] proposed a method, based on classification techniques, for automatic detection and diagnosis of defects of rolling element bearings containing different severities. That method proved the high sensitivity to different types of defects and to different degrees of fault severity. This method is applied to diagnosis bearing fault containing different severities of the same fault as belonging to the same class. The degree of bearing failure is an important indicator to guide equipment maintenance. In order to accurately estimate the degree of failure during fault diagnosis, a multitask convolutional neural network model is proposed in this paper.

The proposed one-dimensional convolutional neural network model uses convolution kernels and pooling layers to directly act on the original time domain signal and trains the extraction features together with the classifier to achieve an "end-to-end" model. The model does not rely on artificial feature extraction and expert knowledge and maximizes the use of CNN for feature extraction. We improve on the basis of the model, get the fault degree information, and improve the generalization ability of the network. In the era of mechatronics big data, it is particularly important to improve the generalization ability of models. The proposed model conducts a more in-depth study of the generalization ability of the model.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction of CNN. Section 3 explains the proposed fault diagnosis method. Section 4 is concentrated on fault diagnosis experimental investigations. Finally, Section 5 gives the conclusion and future work.

## 2. A Brief Introduction of CNN

The CNN is a special structure of the feed-forward neural network, generally including a convolution layer, activation layer, pooling layer, and fully connected layer. Corresponding to the CNN model, the input signal is extracted by the convolution layer, the abstract features are obtained by the convolution and pooling operation of the input signal layer by layer, and finally, the classification or regression is performed through the fully connected layer.

The two core ideas of the CNN network are local connection and convolution kernel parameter sharing. A key role of both is to reduce the number of parameters, making the operation simple, efficient, and able to operate on very large data sets. The local connection ensures that the filter, after learning, can maximize the response to local features, and the parameter sharing reduces the risk of overfitting to some extent.

*2.1. Convolutional Layer.* The convolutional layer uses a convolution kernel to convolve the input signal or the local region of the previous layer output and generates a corresponding feature through a nonlinear activation function. It can acquire richer features with fewer parameters, compared to the fully connected layer. Due to its insensitivity to input dimensions and the ease of training of deep structures, it has become the most common method for complex high-dimensional input feature extraction. The mathematical model of convolutional layer can be described by the following equation:

$$g(i) = \sum_{x=1}^{m} \sum_{y=1}^{n} \sum_{z=1}^{p} a_{x,y,z} \times w_{x,y,z}^{i} + b^{i}, \quad i = 1, 2, \ldots, q, \quad (1)$$

where $i$ is $i_{\text{th}}$ convolution core; $g(i)$ is the characteristic graphs obtained by learning the $i_{\text{th}}$ convolution kernel; $a$ is input data; $w$ is the weighting factor; $b$ is bias of convolution kernels; and $x$, $y$, $z$ are the dimensions of input data.

*2.2. Activation Layer.* After the convolution operation, the activation function transforms the output of the convolution layer nonlinearly and maps the original linear and non-separable multidimensional features into another space. The commonly used activation functions are the Sigmoid function, ReLU function, and Tanh function. The ReLU activation function is the most widely used in CNN among them. The mathematical model of the ReLU activation function can be described by the following equation:

$$y(i) = f(g(i)) = \max\{0, g(i)\}, \quad i = 1, 2, \ldots, q, \quad (2)$$

where $g(i)$ is the output of the convolutional layer; $y(i)$ is the activation value of $g(i)$.

*2.3. Pooling Layer.* The pooling layer carries out the downsampling operation, which includes maximum pooling and average pooling. The most common pooling method for CNN is maximum pooling. The maximum pooling layer can not only reduce the training parameters but also further extract the features to improve the robustness of the features. The mathematical model of maximum pooling can be described by the following equation:

$$p^{l(i,j)} = \max_{(j-1)w<t<jw}\left\{a^{(l,t)}\right\}, \quad j = 1, 2, \ldots, q, \tag{3}$$

where $a^{l(i,t)}$ is the $t_{\text{th}}$ neuron of the $i_{\text{th}}$ feature map of the $l_{\text{th}}$ layer; $w$ is the width of the convolution kernel; and $j$ is the $j_{\text{th}}$ pooling layer.

*2.4. Fully Connected Layer.* The fully connected layer is used to classify the features obtained by the convolution kernel. The output of the last pooled layer is paved into a one-dimensional feature vector as the input of the fully connected layer. So, the input and output are fully connected, and the output layer uses the Softmax activation function. The mathematical model of the fully connected layer can be described by the following equation:

$$z^{l+1(j)} = \sum_{i=1}^{n} w_{ij}^{l} a^{l(i)} + b_j^l, \tag{4}$$

where $w_{ij}^l$ is the weight between the $i_{\text{th}}$ neuron in the $l_{\text{th}}$ layer and the $j_{\text{th}}$ neuron in the $(l+1)_{\text{th}}$ layer; $z_j^{l+1}$ is the output of the $j_{\text{th}}$ neuron in the $(l+1)_{\text{th}}$ layer; $b_j^l$ is the bias of all neurons in the $l_{\text{th}}$ layer to the $j_{\text{th}}$ neuron in the $(l+1)_{\text{th}}$ layer.

## 3. The Proposed Method

Although CNN has been applied in fault diagnosis, most of the models still need a lot of signal processing technology and rich engineering practice experience and do not make full use of the self-learning ability of convolutional neural networks. Limited by the number of feature extractions, the structure of the network is generally shallow, and the nonlinear features contained in the original signal often require a multilayer network to fully extract. Moreover, the proposed models cannot give a judgment on the degree of fault. The degree of failure is extremely important for the stable operation of the equipment. Different degrees of failure are likely to take different measures. The traditional rolling bearing fault diagnosis algorithm can only give the type of fault, but cannot predict the degree of fault. Therefore, this paper proposes a multitask convolutional neural network model. The method proposed in this paper can not only accurately judge the type of fault but also predict the degree of fault, thus giving more accurate guidance.

*3.1. The Structure of the Network.* The network structure proposed in this paper is shown in Figure 1, and it mainly consists of three parts: an input layer, feature extraction layer, and multitask layer. The input layer inputs the original vibration signal with a length of 1024 samples. The feature extraction layer contains a total of four convolution layers, and each one was followed by a ReLU activation function and a maximum pooling layer. The first convolutional layer has 16 convolution kernels, the second convolutional layer has 32 convolution kernels, the third convolutional layer has 64 convolution kernels, and the fourth convolutional layer has 128 convolution kernels. The kernel of the first convolution layer has a size of 64*1, and the kernel of the remaining convolution layers is 3*1. The purpose of the first convolution kernel is to obtain enough information from the original vibration signal to expand the receptive field. The network contains three fully connected layers. The last pooling layer is laid flat and connected to the first fully connected layer to obtain the features of the original vibration signal. The fully connected layer has 100 neurons, and it is shared by two tasks. The remaining two fully connected layers connect two tasks, respectively, and each one has 10 neurons. The multitasking layer uses the obtained features to perform classification tasks and regression tasks, respectively, to achieve fault type judgment and fault degree judgment. The purpose of joint training of two tasks is to be able to obtain more abundant features, so that the type of fault and the degree of fault can be predicted simultaneously.

*3.2. Loss Function Design.* This paper has two tasks, one is the classification task, using the cross-entropy loss function and the other is the regression task, using the minimum mean square error (MSE) loss function. The sum of these two loss functions constitutes the overall loss function of the model. The cross entropy characterizes the distance between the actual output and the expected output, which is the smaller the value of the cross entropy is, the closer the two probability distributions are. The mean square error in mathematical statistics refers to the expected value of the square of the difference between the parameter estimate and the true value of the parameter. The smaller the value of MSE, the better the accuracy of the prediction model describing the experimental data.

Once the loss function is determined, the gradient descent algorithm can be used to update the parameters. The RMSProp optimization algorithm is used in this paper. RMSProp optimizer is suitable for processing nonstationary data, which can effectively prevent premature convergence in deep learning. The basic steps are as follows in Algorithm 1:

## 4. Experimental Investigation

*4.1. Experimental Data.* The data are obtained from the Case Western Reserve University Bearing Fault Database. There are three kinds of defect locations in the data set: rolling element damage, outer ring damage, and inner ring damage. The damage diameters are 0.007 inch, 0.014 inch, and 0.021 inch. The original input signal is a one-dimensional time series signal. In order to meet the basic needs of deep learning a large number of samples, sample expansion is required. In this paper, the method of overlapping sample segmentation is used to realize sample expansion. Compared with nonoverlapping sample segmentation, the correlation between adjacent elements can be preserved as much as possible to improve the robustness of the model. The overlap rate is set to 0.3. The principle of sample expansion is shown in Figure 2. A total of 4 data sets were prepared for the experiment, as shown in Table 1. Data sets A, B, and C are data sets with rotational speeds of 1797 r/min, 1772 r/min, and 1750 r/min, respectively. We divide each data into two

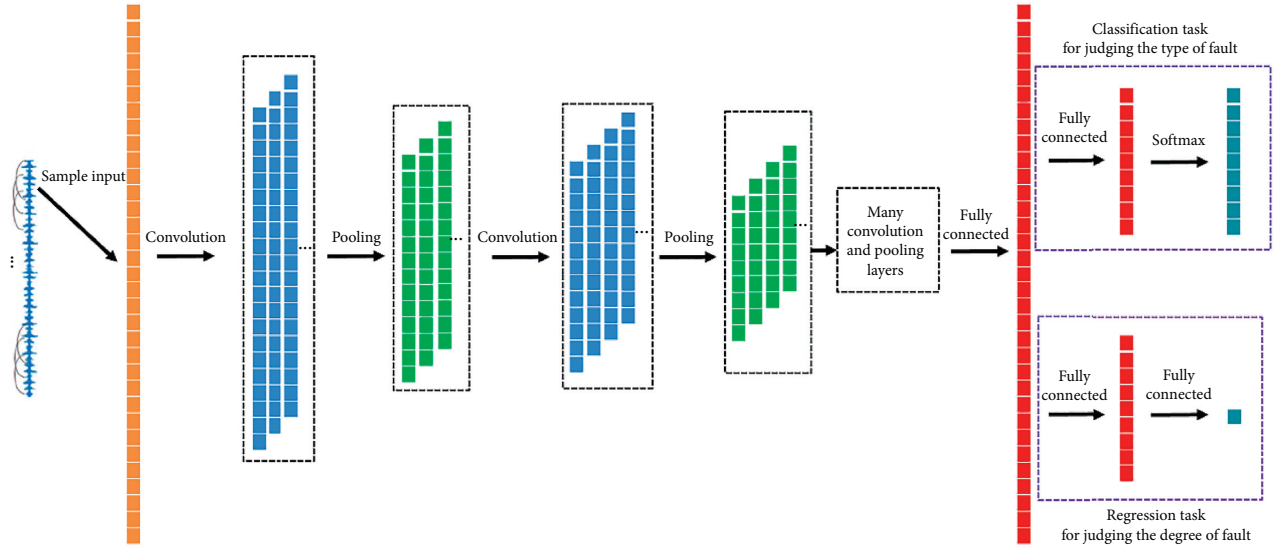


Figure 1: The multitask convolutional neural network structure.

Require: global learning rate $\varepsilon$, Decay rate $\rho$
Require: initial parameter $\theta$
Require: small constant $\delta$, usually set to $10^{-6}$
  Initialize accumulated variable $r = 0$
  while failure to reach the stop criterion do
    Small batches containing m samples $\{x^{(1)}, \ldots, x^{(m)}\}$ from the training set, corresponding target is $y^{(i)}$
    Calculating the gradient: $g \longleftarrow (1/m)\nabla_\theta \sum_i L(f(x^{(i)}; \theta), y^{(i)})$
    Cumulative square gradient: $r \longleftarrow \rho r + (1-\rho) g \odot g$
    Calculating parameter updating: $\Delta\theta = -(\varepsilon/\sqrt{\delta + r}) \odot g$
    Application update: $\theta \longleftarrow \theta + \Delta\theta$
  end while.

Algorithm 1: RMSProp optimization algorithm.

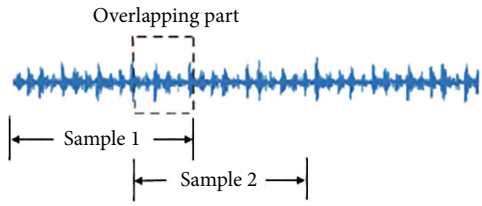


Figure 2: The principle of sample expansion.

parts, one including 1000 samples used as training data and another including 200 samples used as test data. It should be noted that since the model in this paper is multitasking, two labels are required, as shown in Table 1.

*4.2. Multitask Model Training.* Train the built model with data sets A, B, C, and D. The initial learning rate of RMSProp is 0.001. The full connection layer uses the dropout technique to suppress overfitting, the dropout rate is 0.5, the small batch mini-batch size is 64, and the cutoff training round is set to 200. In order to verify the reliability of the training and test results, the network was run 10 times, and we calculate the average of the ten runs. For each time experiment, the 1000 samples in the training data are reused as the training data set, the test data set including 100 samples which are randomly extracted from the test data by sampling replacement. Those 100 samples are randomly selected from the test set for each run for testing to prevent the network from using the same test samples each time. The experimental results are shown in Table 2.

It can be seen from the experimental results that the accuracy on each data set can almost reach 99% for the classification task. Performance on data sets A, B, and C is better than on data set D. For the regression task, the loss function value is the smallest on test set A, which can be reduced to 2.4563. The prediction result on test set A is shown in Figure 3. It can be seen that the model can relatively accurately determine the damage of the fault. Degree, at the same time, can also give the category judgment of the fault, achieving the purpose of multitasking.

In order to see if the model successfully extracted the feature, the first fully connected output is dimensioned using the t-SNE method to 2 dimensions. Figure 4 shows that useful features can be extracted from different data sets.

Table 1: Different data sets.

| Damage location | | None | Rolling element | | | Inner race | | | Outer race | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Classification label | | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 |
| Regression label | | 0 | 7 | 14 | 21 | 7 | 14 | 21 | 7 | 14 | 21 |
| Damage degree | | 0 | 0.007 | 0.014 | 0.021 | 0.007 | 0.014 | 0.021 | 0.007 | 0.014 | 0.021 |
| A | Train | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Test | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| B | Train | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Test | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| C | Train | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | Test | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| D | Train | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 |
| | Test | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |

Table 2: Training and testing results of different data sets.

| | Train set | | | Test set | | |
|---|---|---|---|---|---|---|
| | Classification accuracy (%) | Classification loss | Regression loss | Classification accuracy (%) | Classification loss | Regression loss |
| Set A | 99.21 | 0.0180 | 1.9414 | 99.11 | 0.0324 | 2.4563 |
| Set B | 99.60 | 0.0135 | 1.5936 | 98.86 | 0.0455 | 2.8531 |
| Set C | 99.56 | 0.0220 | 1.9556 | 99.12 | 0.0387 | 3.0156 |
| Set D | 98.95 | 0.0867 | 2.4862 | 98.22 | 0.1031 | 3.4741 |

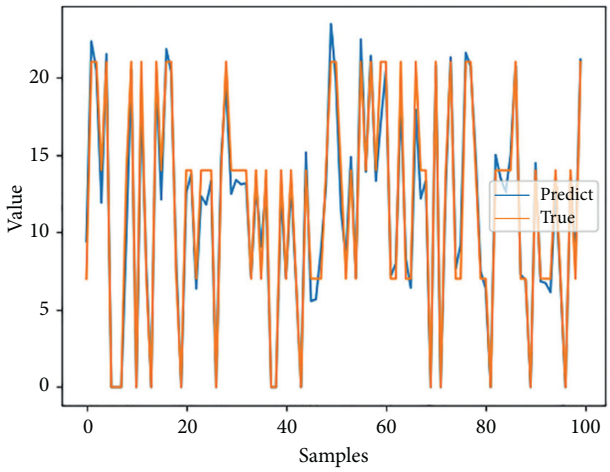


Figure 3: Prediction result of fault degree.

When detecting faults online, we should try to minimize errors. For a section of vibration signal, split the vibration signal into $n$ segments, which can overlap, and put it into the network, so we can get $n$ outputs. Voting methods are used for minimize classification errors, which means the category with the highest number of votes is selected as the result. One way to minimize the regression error is using the weighted average method. It weights the average of the results of the n-segment signals, and we get the final degree of fault of the given signal. The mathematical model can be described by the following equation:

$$y = \frac{\sum_{i=1}^{n} w_i y_i}{n}, \tag{5}$$

where $y_i$ is the output of the $i_{\text{th}}$ segment; $w_i$ is the weight of the $i_{\text{th}}$ segment.

*4.3. Multitask Weight Selection.* The model in this paper optimizes the parameters of the network by multitask simultaneous training. The loss function of the whole convolutional neural network is obtained by adding the loss function of the two tasks, but the order of magnitude of the two tasks is different, which will lead to the bias of the training of network parameters. Therefore, the selection of the weight of the loss function of the two tasks is one of the key factors affecting the diagnostic accuracy of the model. It can be seen from the training results that the loss of the classification task is smaller than the regression loss, and the weight of the regression loss function is relatively increased for training. The results of different weight ratios for data set A training are shown in Table 3 and Figure 5. It can be seen that when the weight ratio increases gradually, the classification accuracy will gradually increase and eventually stabilize, while the loss of regression tasks will gradually increase. This paper considers that it is advisable when the weight ratio is 0.8 : 1, because the classification accuracy tends to be stable and the loss function is at a turning point.

*4.4. Multitask Learning Mode.* There are many kinds of multitask learning modes in deep learning. The one used in this paper is multitask joint training and sharing the same

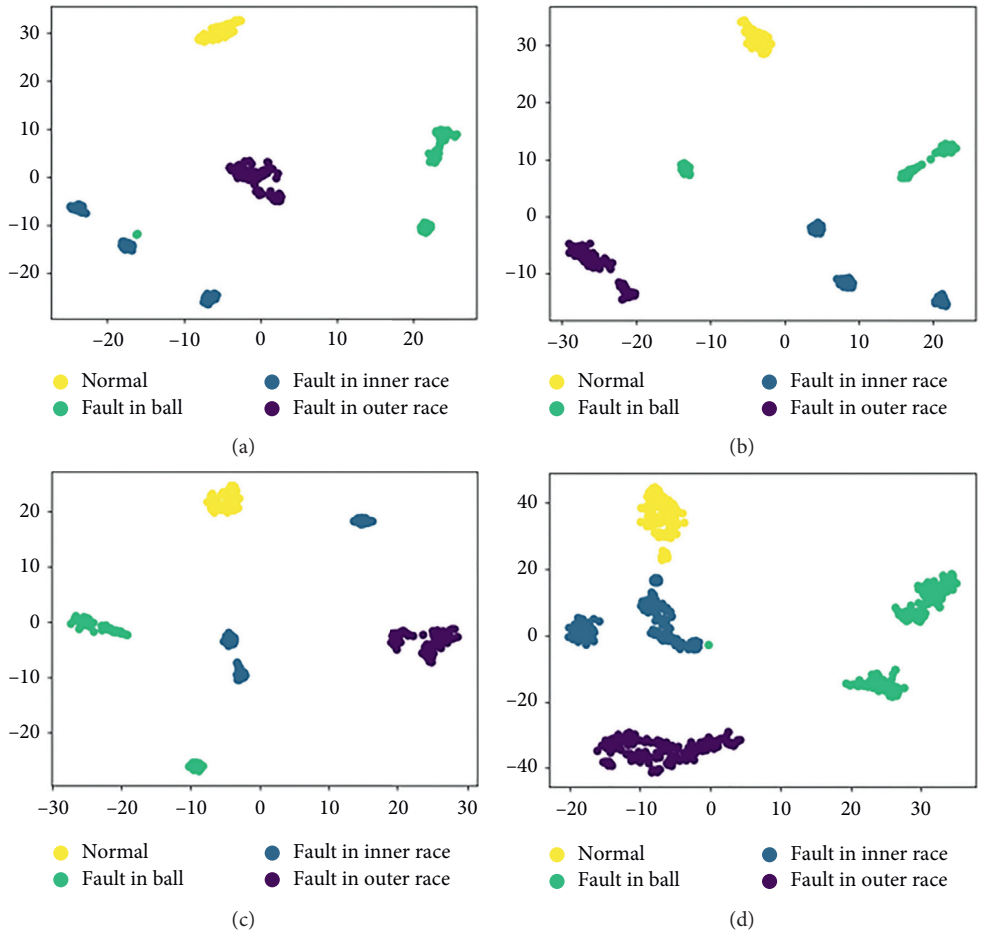


Figure 4: Fully connected layer output visualization of different data sets. (a) Set A. (b) Set B. (c) Set C. (d) Set D.

Table 3: Testing results of different weight ratios.

| Weight ratio | Classification accuracy (%) | Regression loss |
|---|---|---|
| 0.1 : 1 | 96.56 | 2.7720 |
| 0.2 : 1 | 97.40 | 1.3328 |
| 0.3 : 1 | 98.67 | 1.5603 |
| 0.4 : 1 | 98.96 | 1.6305 |
| 0.5 : 1 | 99.17 | 1.6658 |
| 0.6 : 1 | 99.18 | 1.8625 |
| 0.7 : 1 | 99.20 | 1.8820 |
| 0.8 : 1 | 99.23 | 1.9120 |
| 0.9 : 1 | 99.21 | 2.1256 |
| 1 : 1 | 99.25 | 2.3014 |

layer of hidden layers of the network, but only at the near end of the network to start bifurcation to do different tasks. Another mode is individual training, which trains one of the tasks separately, then fixes the structure and parameters of the network, and uses one layer of the hidden layer of the network as the input of another task to train another task. In order to compare the advantages and disadvantages of the two modes, the first fully connected layer is trained as the input of the second task. The experimental results of the two modes are compared as shown in Table 4.

It can be seen that individual training has better classification accuracy, but it does not perform well in the regression task, while the two tasks joint training mode is smaller in regression loss, which shows that this network structure can extract the features of fault category and fault degree more effectively. Compared with the two tasks, it is more capable of taking into account another task. It can be seen that the network proposed in this paper can extract more features. Therefore, the proposed network can extract more abundant features.

*4.5. Generalization Performance Experiment.* For today's electromechanical big data era, the optimal model must not only be able to adapt to different speeds but also apply to a variety of different conditions. This paper will study the generalization ability of the classification task model on different motors and different sampling frequencies. The fault diagnosis test platform for rotating machinery is shown in Figure 6. The fault types also include rolling element damage, outer ring damage, and inner ring damage. Vibration signals at 1450 r/min, 1400 r/min, 1350 r/min, and 1300 r/min were measured and each includes 900 samples. The CWRU data is measured at 12 kHz and 24 kHz and each includes 3600 samples. In order to study the effect of

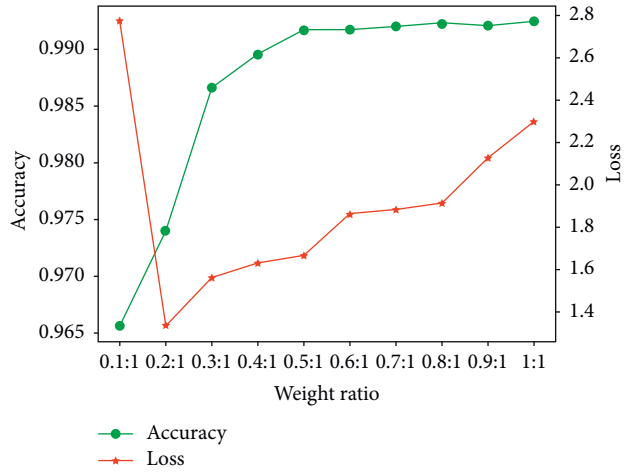


Figure 5: Testing results of different weight ratios.

Table 4: Experimental results of different modes.

| | | Joint training | Individual training |
|---|---|---|---|
| Train set | Classification accuracy | 99.21% | 99.37% |
| | Classification loss | 0.0180 | 0.0103 |
| | Regression loss | 1.9414 | 10.566 |
| Test set | Classification accuracy | 99.11% | 99.13% |
| | Classification loss | 0.0324 | 0.0189 |
| | Regression loss | 2.4563 | 13.108 |

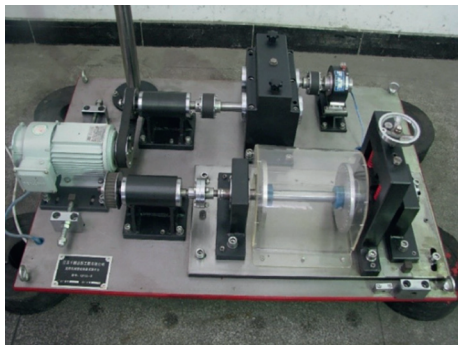

Figure 6: The fault diagnosis test platform for rotating machinery.

different sampling frequencies on the diagnostic results, we collected data of 2560 Hz and 5120 Hz on the fault diagnosis test platform and each also includes 3600 samples.

The CWRU data set and the collected data are jointly trained and tested. The samples were divided into training sets and test sets in a 5 : 1 ratio. So, the total data set includes data from different speeds, different types of motors, and different sampling frequencies, as shown in Figure 7. The CWRU data set includes three different damage degrees. However, the fault diagnosis test platform for rotating machinery only has one damage degree, which is not similar to any of the CWRU data sets. The diagnosis of the damage degree is limited because the fault diagnosis test platform only has one damage degree and it even holds half of the data. The structure of the network is modified, the task of damage degree diagnosis in multitasking is removed, leaving only fault category diagnosis, and then the network is trained. The training result is shown in Figure 8. The accuracy of the training set of the model is up to 99.07%, and the loss function value is 0.0059. The accuracy rate on the test set is up to 98.12%, and the loss function value is 0.0120.

In order to verify the model's adaptive feature learning ability for the original input signal, the t-SNE method is used to perform dimensionality reduction visual analysis on the features learned by each convolutional layer. The partial convolutional layer output visualization is shown in Figure 9. The original data is not clearly distinguished after the first convolutional layer, but as the depth of the network increases, the feature separation of the convolutional layer extraction becomes higher and higher, and the classification is getting

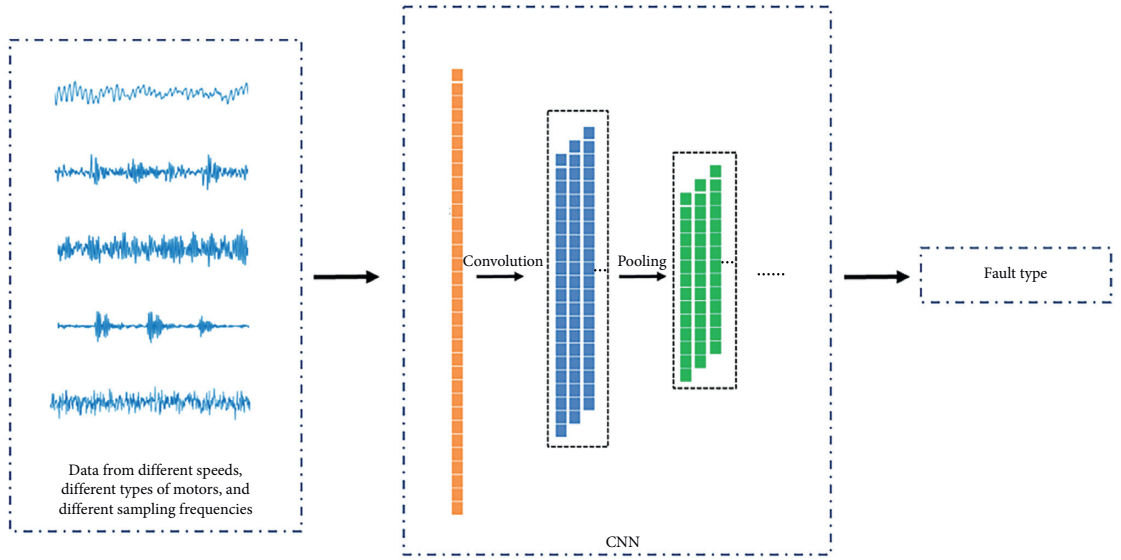


Figure 7: The network structure of generalization ability research.

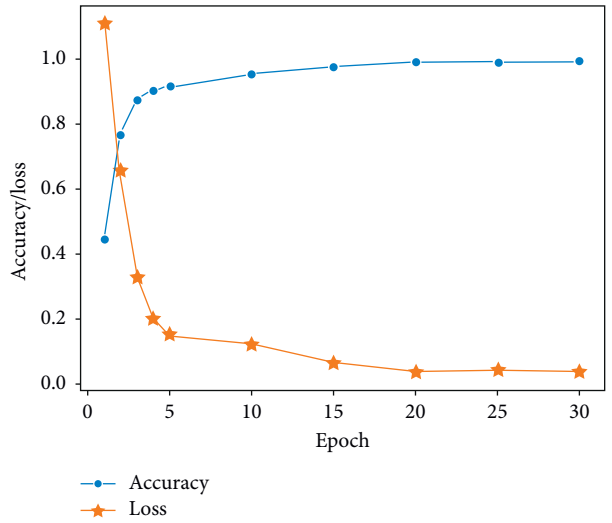


Figure 8: Training result of generalization ability research.

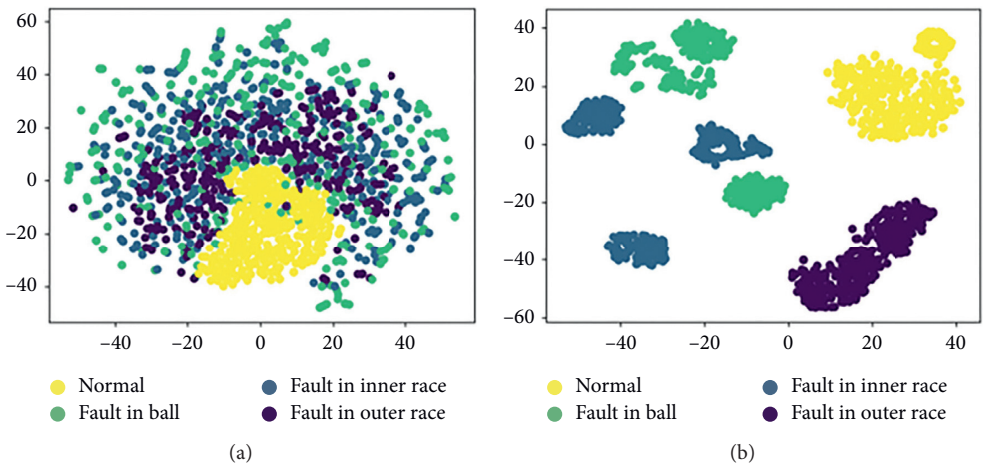


Figure 9: Continued.

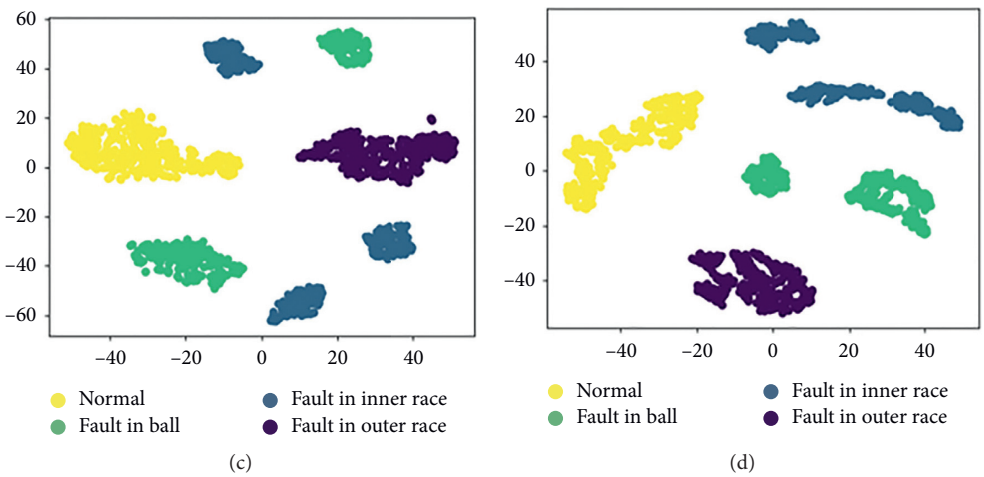


Figure 9: Partial convolutional layer output visualization. (a) Output of conv_1. (b) Output of conv_3. (c) Output of conv_4. (d) Output of fully connected layer.

better and better. Finally, the fully connected layer can be completely distinguished. It can be seen that the proposed model can learn effective features from the vibration signals of different speeds, different types of motors, and different sampling frequencies to achieve accurate fault diagnosis.

## 5. Conclusions

The multitask convolutional neural network proposed in this paper can extract effective fault features from the original signal, realize accurate diagnosis of fault types, and give the fault degree. The advantages and disadvantages of the models under different modes are compared, and it is proved that the optimal results can be achieved when using multitask joint training. The proposed method has good generalization ability and can maintain a high fault recognition rate even if the speed, the type of motor, and the sampling frequency are changed.

The research in this paper is mainly aimed at the analysis of the current state of the real time input signal, and the fault prediction is also important in rotating machinery. It is more meaningful to predict the state after the current information. The following research can be extended to the fault type and degree prediction.

## Data Availability

The CWRU data used to support the findings of this study are obtained from the Case Western Reserve University Bearing Fault Database and have been deposited in the following website: http://csegroups.case.edu/bearingdatacenter/home.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Key Research and Development Project of China (2018YFC0809005) and the Fundamental Research Funds for the Northeastern Universities (N170402009).

## Supplementary Materials

Self-collected data in generalization performance experiment. (*Supplementary Materials*)

## References

[1] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

[2] J. Zarei, H. R. Tajeddini, and H. R. Karimi, "Vibration analysis for bearing fault detection and classification using an intelligent filter," *Mechatronics*, vol. 24, no. 2, pp. 151–157, 2014.

[3] M. Feldman, "Hilbert transform in vibration analysis," *Mechanical Systems and Signal Processing*, vol. 25, no. 3, pp. 735–802, 2011.

[4] F. Hemmati, M. S. W. Orfali, and M. S. Gadala, "Roller bearing acoustic signature extraction by wavelet packet transform, applications in fault detection and size estimation," *Applied Acoustics*, vol. 104, pp. 101–118, 2016.

[5] M. Van and H. J. Kang, "Two-stage feature selection for bearing fault diagnosis based on dual-tree complex wavelet transform and empirical mode decomposition," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 230, no. 2, pp. 291–302, 2015.

[6] K. P. Murphy, "Machine learning: a probabilistic perspective," *Chance*, vol. 27, no. 2, pp. 62-63, 2012.

[7] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the Neural Information Processing Systems*, Curran Associates Inc, Lake Tahoe, CA, USA, December 2012.

[9] T. Junbo, L. Weining, A. Juneng et al., "Fault diagnosis method study in roller bearing based on wavelet transform and stacked auto-encoder," in *Proceedings of the 27th Chinese Control and Decision Conference (2015 CCDC)*, IEEE, Qingdao, China, May 2015.

[10] C. Lu, Z. Y. Wang, W. L. Qin et al., "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Processing*, vol. 130, pp. 377–388, 2017.

[11] M. Gan, C. Wang, and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 92–104, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, June 2015.

[13] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," *Computer Vision-ECCV 2018*, Springer, Berlin, Germany, 2018.

[14] 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," *Mechanical Systems and Signal Processing*, vol. 100, pp. 439–453, 2018.

[15] M. He and D. He, "A deep learning based approach for bearing fault diagnosis," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 3057–3065, 2017.

[16] O. Janssens, V. Slavkovikj, B. Vervisch et al., "Convolutional neural network based fault detection for rotating machinery," *Journal of Sound and Vibration*, vol. 377, pp. 331–345, 2016.

[17] F. Jia, Y. Lei, X. Zhou, L. Zhou, and N. Lu, "Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 303–315, 2016.

[18] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network based data-driven fault diagnosis method," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2017.

[19] J. H. Yuan, T. Han, J. Tang et al., "An approach to intelligent fault diagnosis of rolling bearing using wavelet time-frequency representation and CNN," *Machine Design and Research*, vol. 2, pp. 93–97, 2017.

[20] Z. Wei, P. Gaoliang, L. Chuanhao, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017.

[21] M. Cococcioni, S. L. B. Lazzerini, and S. L. Volpi, "Robust diagnosis of rolling element bearings based on classification techniques," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2256–2263, 2013.