Research Article

Fault Diagnosis of Gearbox in Multiple Conditions Based on Fine-Grained Classification CNN Algorithm

Pengcheng Jiang,¹ Hua Cong,¹ Jing Wang ² and Dongsheng Zhang²

¹Vehicle Engineering Department, Academy of Army Armored Force, Beijing, China

²Xi'an Jiaotong University, Xi'an, China

Correspondence should be addressed to Jing Wang; 1509332851@qq.com

Received 20 September 2019; Revised 29 December 2019; Accepted 15 February 2020; Published 20 June 2020

Academic Editor: Adam Glowacz

Copyright © 2020 Pengcheng Jiang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The use of the convolutional neural network for fault diagnosis has been a common method of research in recent years. Since this method can automatically extract fault features, it has played a good role in some research studies. However, this method has a clear drawback that the signals will be significantly affected by working conditions and sample size, and it is difficult to improve diagnostic accuracy by directly learning faults, regardless of working conditions. It is therefore a research orientation worthy of a diagnosis of high precision defect in various working conditions. In this article, using a fine-grained classification algorithm, the operating conditions of the object system are considered an approximate classification. A specific failure in different working conditions is considered a beautiful classification. Samples of different faults in different working conditions are learned uniformly and the common characteristics are extracted from the convolutional network so that different faults of different working conditions can simultaneously be identified on the basis of the entire sample. Experimental results show that the method effectively uses the set of samples of the working conditions of the variables to obtain the dual recognition of defects and specific working conditions and the accuracy of the recognition is significantly higher than the method of learning regardless of working conditions.

1. Introduction

Traditional methods of fault diagnosis, whether in time domain or frequency domain analysis, are highly dependent on physical experience. In recent years, because convolution neural network (CNN) was proposed [1], deep learning algorithms have developed rapidly; their powerful end-to-end learning ability enables feature extraction work that needs experience to be completed independently by CNN, which becomes a new direction of fault diagnosis research. For rotating machine, especially for gearbox, using CNN to learn the vibration signals was the main method of fault diagnosis [2–5]. However, most of these studies are only based on specific conditions. Obviously, many systems have very strong characteristic changes under different working conditions. Although some research studies try to use CNN's ability to directly model the multi-working conditions of the

object, the settings of this working condition change were very limited [6]. In fact, in some practical fault diagnosis problems, the influence of working conditions on signal characteristics is greater than that of fault types [7]. Therefore, extracting fault features directly without considering working conditions can seriously reduce the accuracy of classification [8, 9]. In order to solve this problem, some studies use the method of separated conditions modeling [7, 9], which divides the entire problem into independent problems that are not related to each other and reduces the utilization efficiency of samples. The latter method addresses another problem, which means that the sample size must be quite large. But the diagnosis of fault is often difficult to sample; this means that a separate model must be formed on a smaller sample, so accuracy is limited. Transfer learning is used to solve this problem, including adjusting the source model to adapt to new conditions

[10, 11] and using the source model to accelerate learning [12]; the premise is that there is a source model. These methods have achieved excellent results, but they are all based on an important premise, that the sample size is large enough to support diagnostic modeling in a single working condition. However, in many problems, this is a difficult requirement. A neglected message is that, although under different conditions, they have the same characteristics as the same target system. Because the influence of working conditions and defects on the system response is interactive and the influence of working conditions is greater than that of defects, the diagnosis of defects in different working conditions can be considered a fine-grained classification problem, which is a different structure from the CNN traditional. This structure is very effective in dealing with hierarchical problems [12–15] and has been applied in the diagnosis of defect [16], but only applied as a means of improving information. However, this structure was not used to solve the problem of defect diagnosis of limited samples under the disturbance state. Because the influence of working conditions is more important than health states, the model is designed as shown in Figure 1.

In the traditional method (Figure 1(a)), when the sample size is sufficient, the modeling effect has been achieved by using sub-working condition modeling. According to the traditional method (Figure 1(b)), the coupling of the working conditions to the fault is ignored, which has a significant impact. In method in Figure 1(b), although it seems that more samples are used to support the training of a single network, the modeling accuracy will often be seriously reduced instead. This means that although it is necessary to find a more efficient way to use the sample, the effect of distinguishing working conditions on fault classification must also be guaranteed at the same time.

The fine-grained method (Figure 1(c)) uses the same convolutional layers to extract common features of faults. Through the two-level fully connected network and the two-stage loss function, recognize the working condition features with stronger significance as coarse classification, and the fault features are subdivided into specific faults under different working conditions as a fine-grained identification. In order to increase the recognition accuracy, the function of increasing class spacing is designed and added to the loss. All the above operations eliminate the impact of changes in operating conditions on fault diagnosis. Our work has the following advantages:

- (i) High utilization efficiency of the data: all samples unified are used to train a single feature extractor, the sample utilization ratio is doubled compared with the traditional division of labor modeling, the convolution layer can obtain more adequate training to achieve better feature extraction effect, and the working condition label originally used for manual division is also used in learning
- (ii) Under the constraint of fine-grained model, the fault under different working conditions is automatically distinguished, and the influence of working conditions on fault diagnosis is solved while all samples are trained uniformly

Experiments show that the gearbox is a system with significant differences in working conditions and rare effective fault samples and its fault diagnosis problem is solved.

2. Research Object

Gearbox is an object with distinct characteristics. Because different working conditions mean completely different mesh states, working conditions are a kind of information that cannot be ignored. Gearbox is also a kind of object with a high cost of collecting the sample, due to the low failure rate; hundreds of hours of work can often encounter a single failure as an effective sample. This means that it is not possible to form the model through a large sample size. An appropriate method must take into account the economic cost. In this paper, the planetary gearbox for a particular type of vehicle is utilized as an object. The gearbox transmission principle is shown in Figure 2. There are three planetary gears K1, K2, and K3 that need to be analyzed. The test bench and sensor arrangement are shown in Figure 3.

The health states corresponding to the three planetary wheels include (a) normal, (b) planetary wheel tooth fault (hereinafter referred to as fault (1)), (c) planetary wheel tooth fault (hereinafter referred to as fault (2)), (d) planetary wheel tooth fault (herein after referred to as fault (3)), and (e) solar wheel tooth fault (herein after referred to as fault (4)). There are 4 kinds of corresponding working conditions: 1 to 4 gears. The signal characteristics of the same fault under different working conditions are not exactly the same. For example, normal and fault 3 and fault 4 conditions include 1/2/3/4/5 and reverse gear. The possible working conditions of fault 1 and fault 2 include gear 2 and gear 3. However, the conventional convolutional network can only label the 5 situations that need to be identified. The relationship between different working conditions and the change of fault signal characteristics by working conditions cannot be connected. The variation of working conditions becomes a disruptive factor in fault diagnosis, which reduces the accuracy of the diagnosis. Fine-grained classification algorithm can distinguish the influence of working conditions and health state from principles. Its performance depends on the design of the loss function.

In [12, 13], the contrast loss function is utilized. In [14, 15], a triple loss function is used to learn a feature that maximizes the distance between classes while minimizing the distance between classes and improve the precision of the fine-grained classification algorithm. However, when the amount of data is relatively large, the computation amount of these two algorithms increases exponentially when constructing binary or triple groups. Experiments show that the quality of binary or triple sets has a major influence on the final classification accuracy. Therefore, problems such as slow model convergence, large computation, increased training complexity, and increased uncertainty of results will occur when using the above methods. In order to solve the above problems, a new loss function has been proposed in the literature [17], which has improved the classification accuracy of fine-grained classification problems of the following two aspects: (1) a structure of cascading classification

FIGURE 1: The comparison between the traditional CNN fault diagnosis method and the fine-grained fault diagnosis algorithm. (a, b) The framework of the traditional CNN methods. “fc” denotes the fully connected layer. “C1”~“C4” denote 1~4 gear position, respectively. (c) The fine-grained fault diagnosis algorithm with two-level hierarchy label structure.

FIGURE 2: Planetary gearbox transmission schematic.

was designed to better describe the hierarchical relationship between fine-grained and coarse categories; (2) a wide margin loss method is proposed, which aims to minimize the inner distance and maximize the distance between classes, and the distance between fine-grained classes belonging to the same class of parents should be smaller than the distance between classes with fine-grained belonging to different classes of parents. The algorithm in [18] is applied to the gearbox fault diagnosis to explore its effect.

3. Methodology

Before introducing our method, we give some definitions in order to explain our method in a mathematical way. Given a

signal data set $\Gamma = \{S_i, \mathbb{Y}_i\}_{i=1}^n$, where S_i is the i th input signal and n is the total number of the training sets, each label of the input signal S_i has a hierarchical label structure $\mathbb{Y}_i = \{y_i^j\}_{j=1}^l$, where $y_i^j \in \{1, 2, \dots, Y^{(j)}\}$ is the class label of the j th level, l is the number of label levels of the hierarchical label set, and $Y^{(j)}$ is the number of classes of the j th level. We suppose that the first level denotes the fine-grained label. So y_i^1 is the fine-grained label corresponding to signal S_i , and $Y^{(1)}$ is the total number of fine-grained classes. For each input signal S_i , we define the preliminary layer of the CNN model as its feature vector, denoted as s_i . For the problem of fault signal diagnosis to be solved in this paper, the hierarchical label structure consists of two levels ($l = 2$). The first level is the fine-grained class level, which identifies the specific sources of failure ($Y^{(1)}$ = the total number of health states). The second level is the coarse class level ($Y^{(2)} = 6$) which denotes different working conditions. After explaining the basic definitions above, we introduce the two main parts of our method, respectively.

3.1. Cascaded Softmax Loss. In the problem of fine-grained classifications with two hierarchical labels (working condition coarse class classification and health state fine-grained class classification), we divided the last classification layer of the CNN model into two fully connected layers and used a cascaded softmax loss for training. Figure 4 shows the whole framework of our method. The number of neurons in fc6 layer (health state fine-grained class classification layer) and fc7 layer (working condition coarse class classification layer)

FIGURE 3: Sensor layout.

FIGURE 4: Structure of the proposed fine-grained fault signal classification model.

is $Y^{(1)}$ and $Y^{(2)}$, respectively. For each input signal S_i , the output of fc5 layer is the feature vector s_i of the signal S_i . The output of the fc6 layer and fc7 layer is the probability scores $p(y_i^1/y_i^1)$ of the health state fine-grained class $y_i^1 \in \{1, 2, \dots, Y^{(1)}\}$ and the probability scores $p(y_i^2/s_i)$ of the working condition coarse class $y_i^2 \in \{1, 2, \dots, Y^{(2)}\}$.

The purpose of adding the skipping connection between fc5 layer and fc7 layer is that it could provide not only the features of the health state fine-grained class but also the probability scores $p(y_i^1/s_i)$ (the output of fc6 layer) of it. Intuitively, using the above-mentioned two different types of information to conduct the working condition coarse level classification will be better than the one using only the health state fine-grained level classification results, since the former not only explores the semantic information of the fault signals (i.e., the learned features), but also learns the hierarchical label structure of the fault signals. Besides, in the iterative learning procedure, the error flows of fc7 layer backpropagate to fc6 layer, fc5 layer, and the first few layers of the CNN model; this can help to improve the health state fine-grained class classification accuracy.

To train the network shown in Figure 4, the cascaded softmax loss of fc6 layer and fc7 layer is as follows:

$$\text{csm}(\mathcal{W}, S_i, \mathbb{Y}) = \sum_{j=1}^l \text{softmax}(\mathcal{W}, S_i, y_i^j), \quad (1)$$

where $\mathcal{W}$ denotes the parameters of the whole network. For fine-grained classifications problem with two hierarchical labels (working condition coarse classification and health state fine-grained classification) ($l = 2$), $\text{softmax}(\mathcal{W}, S_i, y_i^1)$ and $\text{softmax}(\mathcal{W}, S_i, y_i^2)$ are used to train fc6 layer and fc7 layer, respectively. In fact, the cascaded softmax loss can be seen as a multi-task learning problem. One task is the health state fine-grained class classification; the other is the working condition coarse class classification. In the joint training procedure, these two tasks could improve each other by sharing the feature representation.

The whole loss function of training the CNN network in Figure 4 can be defined as

$$L = \frac{1}{n} \sum_{i=1}^n \text{csm}(\mathcal{W}, S_i, \mathbb{Y}_i) + \lambda \mathcal{M}(\mathcal{W}, \mathcal{S}, \mathcal{Y}), \quad (2)$$

where $\text{csm}(\mathcal{W}, S_i, \mathbb{Y}_i)$ is the cascaded softmax loss defined in formula (1), and $\mathcal{M}(\mathcal{W}, \mathcal{S}, \mathcal{Y})$ represents the GLM (generalized large-margin) loss, used to train the feature layer (fc5 layer) of the network. The input of the GLM loss is the training feature set $\mathcal{S} = \{s_1, \dots, s_n\}$ and hierarchy structure label set $\mathcal{Y} = \{\mathbb{Y}_1, \dots, \mathbb{Y}_n\}$. λ is the hyperparameter which is used to balance the cascaded softmax loss and the GLM loss.

3.2. Generalized Large-Margin Loss. For each health state fine-grained class y , we define two groups $\text{SF}(y)$ and $\text{SF}(y)$ which make up the remaining health state fine-grained classes. These two groups consist of the health state fine-grained classes which share and do not share the same parent working condition coarse class with class y , respectively. The purpose of the GLM loss includes two aspects: (1) the distance between health state fine-grained class y and the nearest health state fine-grained class in $\text{SF}(y)$ larger than the intraclass distance of health state fine-grained class y by a pre-defined margin; (2) the distance between health state fine-grained class y and its nearest health state fine-grained class in $\text{SF}(y)$ is larger than the distance between health state fine-grained class y and its farthest health state fine-grained class in $\text{SF}(y)$ by a predefined margin. In the following, we will first define the intraclass distance and interclass distance and then use these definitions to describe the GLM loss. The principle is shown in Figure 5.

The features belonging to health state fine-grained class y in the training set are defined as

$$\mathcal{F}_y = \{s_i \mid i \in \tau_y\}, \quad (3)$$

FIGURE 5: The hierarchical and the interclass structure of states.

where τ_y denotes the index set belonging to the health state fine-grained class y in the training data set. The mean vector of $\mathcal{F}_y$ is defined as

$$m_y = \frac{1}{n_y} \sum_{i \in \tau_y} s_i, \quad (4)$$

where $n_y = |\tau_y|$. The intraclass distance of $\mathcal{F}_y$ is

$$D^{(W)}(\mathcal{F}_y) = \frac{1}{n_y} \sum_{i \in \tau_y} \|s_i - m_y\|. \quad (5)$$

We define $\mathcal{F}_m$ and $\mathcal{F}_n$ as two feature sets:

$$\begin{aligned} \mathcal{F}_m &= \{s_i \mid i \in \tau_m\}, \\ \mathcal{F}_n &= \{s_i \mid i \in \tau_n\}. \end{aligned} \quad (6)$$

Then the interclass distance between $\mathcal{F}_m$ and $\mathcal{F}_n$ can be defined as

$$\begin{aligned} D^{(B)}(\mathcal{F}_m, \mathcal{F}_n) &= \frac{1}{2k} \sum_{i,j=1}^n O_{ij}^{(m,n)} \|s_i - s_j\|^2 \\ &= \frac{1}{k} \text{tr}(Q\Phi^{(m,n)}Q^T), \end{aligned} \quad (7)$$

$$O_{ij}^{(m,n)} = \begin{cases} 1, & \text{if } (s_i, s_j) \in \zeta_k(\mathcal{F}_m, \mathcal{F}_n) \text{ or } (\mathcal{F}_m, \mathcal{F}_n), \\ 0, & \text{else,} \end{cases} \quad (8)$$

where $O_{ij}^{(m,n)}$ is the (i, j) th item in affine matrix $O^{(m,n)}$ between $\mathcal{F}_m$ and $\mathcal{F}_n$, $O^{(m,n)} = (O_{ij}^{(m,n)})_{p \times p}$, $\zeta_k(\mathcal{F}_m, \mathcal{F}_n)$ is the set consisting of k nearest pairs of samples in the set of sample pairs $\{(s_i, s_j) \mid s_i \in \mathcal{F}_m, s_j \in \mathcal{F}_n\}$, $Q = [s_1, \dots, s_p]$, $\Phi^{(m,n)}$ is the Laplacian Matrix of $O^{(m,n)}$, i.e., $\Phi^{(m,n)} = Z^{(m,n)} - O^{(m,n)}$, $Z^{(m,n)} = \text{diag}(z_{11}^{(m,n)}, \dots, z_{pp}^{(m,n)})$, and $\text{tr}(\cdot)$ is the matrix trace.

After we describe the definitions mentioned above, the two constraints in the GLM loss can be expressed as

$$\mathcal{M}_y^{(1)} = \left[D^{(W)}(\mathcal{F}_y) - D^{(B)}(\mathcal{F}_y, \mathcal{F}_{\text{SF}(y)}^{(\min)}) + \alpha_1 \right]_+, \quad (9)$$

$$\mathcal{M}_y^{(2)} = \left[D^{(B)}(\mathcal{F}_y, \mathcal{F}_{\text{SF}(y)}^{(\max)}) - D^{(B)}(\mathcal{F}_y, \mathcal{F}_{\text{SF}(y)}^{(\min)}) + \alpha_2 \right]_+, \quad (10)$$

where α_1 and α_2 are two predefined margin values, $[a]_+ = \max\{a, 0\}$, and

$$\mathcal{F}_{\text{SF}(y)}^{(\min)} = \arg \min_{\mathcal{F}_i, i \in \text{SF}(y)} D^{(B)}(\mathcal{F}_y, \mathcal{F}_i), \quad (11)$$

$$\mathcal{F}_{\text{SF}(y)}^{(\max)} = \arg \max_{\mathcal{F}_i, i \in \text{SF}(y)} D^{(B)}(\mathcal{F}_y, \mathcal{F}_i), \quad (12)$$

$$\mathcal{F}_{\text{SF}(y)}^{(\min)} = \arg \min_{\mathcal{F}_i, i \in \text{SF}(y)} D^{(B)}(\mathcal{F}_y, \mathcal{F}_i), \quad (13)$$

In formulas (11)–(13), $\text{SF}(y)$ consists of health state fine-grained classes that share the same parent (i.e., working condition coarse class) with health state fine-grained class y . And $\mathcal{F}_{\text{SF}(y)}^{(\min)}$ is the feature vector set of fault signal samples which is the closest to health state fine-grained class y in the health state fine-grained classes of $\text{SF}(y)$, and $\mathcal{F}_{\text{SF}(y)}^{(\max)}$ is the feature vector set of fault signal samples which is the farthest to health state fine-grained class y in the health state fine-grained classes of $\text{SF}(y)$. Besides, $\text{SF}(y)$ consists of the health state fine-grained classes that do not share the same parent working condition coarse class with health state fine-grained class y and $\mathcal{F}_{\text{SF}(y)}^{(\min)}$ is the feature vector set of the fault signal samples which is the closest to health state fine-grained class y in the health state fine-grained class of $\text{SF}(y)$ (Figure 5).

Using the definitions mentioned above, the GLM loss which contains two-level label structure can be described as

$$\mathcal{M} = \frac{1}{Y^{(1)}} \sum_{y=1}^{Y^{(1)}} (\mathcal{M}_y^{(1)} + \mathcal{M}_y^{(2)}). \quad (14)$$

3.3. Optimization. The CNN model formation algorithm we use here is the standard backpropagation algorithm (BP) based on the mini-batch. The full loss function is defined in formula (2). By minimizing this objective function, the CNN model can learn fault characteristics that can better distinguish different health conditions in different working conditions, thus improving the classification accuracy of faults. Finally, therefore, we need to calculate the gradients of the entire loss function with respect to the activations of all CNN layers, which are called the error streams of the

corresponding layers. As it is very simple to calculate the gradient of softmax loss, we only provide the gradient bypass process for the loss of GLM with respect to the following. In this paper, we use a two-level hierarchical label structure to describe the fault diagnosis problem. Therefore, the derivatives of the GLM loss regarding s_i can be computed as follows:

$$\frac{\partial \mathcal{M}}{\partial s_i} = \frac{1}{Y^{(1)}} \sum_{c=1}^{C^{(1)}} \left(\frac{\partial \mathcal{M}_y^{(1)}}{\partial s_i} + \frac{\partial \mathcal{M}_y^{(2)}}{\partial s_i} \right), \quad (15)$$

$$\frac{\partial \mathcal{M}_y^{(1)}}{\partial s_i} = \Delta_1(y) \left[I(i \in \tau_y) \frac{2}{n_y} \left(1 - \frac{1}{n_y} \right) (s_i - m_y) \right] - \frac{2}{k} \left(Q\Phi(y, y_{SF}^{(\min)}) \right)_{(:,j)}, \quad (16)$$

$$\frac{\partial \mathcal{M}_y^{(2)}}{\partial s_i} = \Delta_2(y) \frac{2}{k} Q \left[\Phi(y, y_{SF}^{(\max)}) - \Phi(y, y_{SF}^{(\min)}) \right], \quad (17)$$

where $I(\cdot)$ is an indicator function, which equals one if the condition is true, and zero otherwise. The j th column of the matrix is represented by a subscript $(:, j)$, and

$$\Delta_1(y) = \begin{cases} 1, & D^{(B)}(F_y, F_{SF(y)}^{(\min)}) - D^{(W)}(F_y) < \alpha_1, \\ 0, & \text{otherwise}, \end{cases} \quad (18)$$

$$\Delta_1(y) = \begin{cases} 1, & D(\mathcal{F}_y, \mathcal{F}_{SF(y)}^{(\max)}, \mathcal{F}_{SF(y)}^{(\min)}) < \alpha_2, \\ 0, & \text{otherwise}, \end{cases} \quad (19)$$

$$D(\mathcal{F}_y, \mathcal{F}_{SF(y)}^{(\max)}, \mathcal{F}_{SF(y)}^{(\min)}) = D^{(B)}(\mathcal{F}_y, \mathcal{F}_{SF(y)}^{(\min)}) - D^{(B)}(\mathcal{F}_y, \mathcal{F}_{SF(y)}^{(\max)}), \quad (20)$$

$$y_{SF}^{(\min)} = \arg \min_{i \in SF(y)} D^{(B)}(\mathcal{F}_y, \mathcal{F}_i), \quad (21)$$

$$y_{SF}^{(\max)} = \arg \max_{i \in SF(y)} D^{(B)}(\mathcal{F}_y, \mathcal{F}_i), \quad (22)$$

$$y_{SF}^{(\min)} = \arg \min_{i \in SF(y)} D^{(B)}(\mathcal{F}_y, \mathcal{F}_i). \quad (23)$$

Algorithm 1 describes the training algorithm based on the network framework shown in Figure 4 of formula (15).

4. Experiments

4.1. Data Description. The experimental data used in this paper is collected on a planetary gearbox for a particular type of vehicle mentioned in Section 1. And the physical diagram of fault simulation experimental platform is shown in Figure 6.

The planetary gearbox data samples are divided into five health states: normal state (Normal), K1 planetary gear failure (fault 1), K1 planetary gear failure (fault 2), K2 planetary gear failure (fault 3), and K3 planetary sun gear failure (fault 4). The labels correspond to five health states

denoted as 0–4, respectively. Each health state samples is collected under 4 working conditions, which correspond to the gear positions 1 to 4. The data sampling frequency is 20 kHz and each working condition has four types of input speed: 600 r/min, 900 r/min, 1200 r/min, and 1500 r/min. The load torque is 900 N · m. A total of 4,800 samples are obtained for each health state, and 80% of samples are used for training. Each sample contains 4 measurement points, which are the vibration signal of the gearbox, and the length of each measurement point is 2000 which is longer than one rotation period. So the sample data length is 4*2000. There are 12,000 samples in total. All the sample data are shown in Table 1.

For the fault diagnosis problem described in this paper, the traditional method (a) needs to establish a model according to each independent condition; a total of four models can be completely processed; each model has 5 states; each state has 600 tags, while the traditional method (b) ignores the working conditions to use all the samples, a total of 5 states, 2400 samples each. The fine granularity method used in this paper also has the same label form as the traditional method (a) while using all the working condition samples uniformly. And the method (c) focuses on fault diagnosis while avoiding losing the working condition information at the same time. There are 4 conditions as the coarse class and 4 × 5 faults as the fine class. As for every method, the total number of samples is 12,000.

4.2. Experiment Settings. The CNN models used in this paper have the same network structure and hyperparameters except for the different output structures, that is, 5 convolutional layers and 2 fully connected layers, each of which includes a pooling layer and a BN layer. The output layer is a softmax layer. The activation function of the convolution layer is Relu. The activation function of the fully connected layer is Sigmoid. The cross-entropy loss is used as the loss function. And the optimization method is the Adam algorithm. The regularization function is L2. The learning rate is 0.0001 and the batch size is 32. The hyperparameter λ used in the loss function (formula (2)) of the proposed method is 0.1.

4.2.1. Traditional CNN Fault Diagnosis Method Using Separate Modeling of Multiple Working Conditions. The framework of the traditional CNN model using separate modeling of multiple working conditions is illustrated in Figure 1(a), and this method is abbreviated as CNN-S. It has four CNN models in total and each model corresponds to one working condition. The output of each CNN model contains 5 health states; the working condition information of the input sample needs to be clarified during testing stage. The 5 health states are K1 planetary row large planet wheel failure (fault 1), K1 planetary row asteroid gear failure (fault 2), K2 planetary gear failure (fault 3), K3 planetary gear solar gear failure (fault 4), and normal status. The composition of the data samples used in the traditional CNN method using separate modeling of multiple working conditions is shown in Table 2. It shows only the data used by one model, and each model has the same data structure.

Input: fault signal training set Γ , hyperparameter λ , α_1 , and α_2 , the maximum number of iterations $T_{\max}$, and counter $t = 0$,
 Output: parameter $\mathcal{W}$ of the CNN model

- (1) Select a mini-batch sample from the fault signal training set Γ
- (2) Execute the forward propagation of the CNN model, for each input signal; calculate the activation values of each layer
- (3) Compute the softmax loss error flows of the fc7; then calculate the backpropagated error flows from fc7 to fc6 and the error flows from fc7 to fc5
- (4) Compute the softmax loss of the fc6
- (5) Compute the overall error flows of fc6, which consist of the softmax loss of fc7 and softmax loss of itself. Then use the BP algorithm to calculate the error flows backpropagated from fc6 to fc5
- (6) Compute the GLM loss error flows of the backpropagation to fc5, and then multiply hyperparameter λ
- (7) Compute the overall error flows of fc5 which consist of fc6, fc7, and the GLM loss
- (8) Execute the backpropagation from fc7 to conv1 layer, and use the BP algorithm to compute the error flows of these layers
- (9) Based on the activation values and error flow values, calculate $\partial \mathcal{L} / \partial \mathcal{W}$ through BP algorithm
- (10) Update $\mathcal{W}$ according to the gradient descent method
- (11) $t \leftarrow t + 1$. If $t < T_{\max}$, back to step 1

ALGORITHM 1: The training algorithm of the proposed method.

FIGURE 6: Physical diagram of fault simulation experimental platform.

TABLE 1: Number of samples for different training methods.

Sample number	Condition 1	Condition 2	Condition 3	Condition 4	Ignore conditions
Normal	600	600	600	600	2400
Fault 1	600	600	600	600	2400
Fault 2	600	600	600	600	2400
Fault 3	600	600	600	600	2400
Fault 4	600	600	600	600	2400
Total					12000

4.2.2. Traditional CNN Fault Diagnosis Method Using Simultaneous Modeling of Multiple Working Conditions.

The framework of traditional CNN model using simultaneous modeling of multiple working conditions is illustrated in Figure 1(b), and this method is abbreviated as CNN-M. The output of the traditional CNN model contains 5 health states; it cannot recognize the working condition of the input sample. The 5 health states are K1 planetary row large planet wheel failure (fault 1), K1 planetary row asteroid gear failure (fault 2), K2 planetary gear failure (fault 3), K3 planetary gear solar gear failure (fault 4), and normal status. The composition of the data samples used in the traditional CNN method using simultaneous modeling of multiple working conditions is shown in Table 3.

TABLE 2: Data description of CNN-S.

Health state	Sample number	Test/train sample proportion	Label
Normal	600	20%/80%	0
Fault 1	600	20%/80%	1
Fault 2	600	20%/80%	2
Fault 3	600	20%/80%	3
Fault 4	600	20%/80%	4

4.2.3. *Fine-Grained Fault Classification Algorithm.* The framework of the fine-grained defect classification algorithm is illustrated in Figure 1(c), and our method is abbreviated as CNN-FG. Compared to CNN's traditional methods, the

TABLE 3: Data description of CNN-M.

Health state	Working condition number	Sample number	Test/train sample proportion	Label
Normal	4	2400	20%/80%	0
Fault 1	4	2400	20%/80%	1
Fault 2	4	2400	20%/80%	2
Fault 3	4	2400	20%/80%	3
Fault 4	4	2400	20%/80%	4

TABLE 4: Data description of CNN-FG.

Working condition	Health state	Sample number	Test/train sample proportion	Health state label	Fine-grained class label	Coarse class label
Condition 1	Normal	600	20%/80%	0	0	1
	Fault 1	600	20%/80%	1	1	
	Fault 2	600	20%/80%	2	2	
	Fault 3	600	20%/80%	3	3	
	Fault 4	600	20%/80%	4	4	
Condition 2	Normal	600	20%/80%	0	5	2
	Fault 1	600	20%/80%	1	6	
	Fault 2	600	20%/80%	2	7	
	Fault 3	600	20%/80%	3	8	
	Fault 4	600	20%/80%	4	9	
Condition 3	Normal	600	20%/80%	0	10	3
	Fault 1	600	20%/80%	1	11	
	Fault 2	600	20%/80%	2	12	
	Fault 3	600	20%/80%	3	13	
	Fault 4	600	20%/80%	4	14	
Condition 4	Normal	600	20%/80%	0	15	4
	Fault 1	600	20%/80%	1	16	
	Fault 2	600	20%/80%	2	17	
	Fault 3	600	20%/80%	3	18	
	Fault 4	600	20%/80%	4	19	

fine-grained classification algorithm uses two levels of label hierarchical structure, which are coarse class level and fine-grained class level. So, it has two classifiers (softmax layer) after the fully connected layer of CNN; one recognizes the fine-grained class level and the other recognizes the coarse level. The coarse class level corresponds to working conditions and the fine-grained class level represents health conditions. Working conditions are divided into condition 1, condition 2, condition 3, and condition 4, which correspond to coarse labels 1/4. Data samples under each working condition are divided into 5 health conditions: normal state and fault 1–fault 4; and corresponding state tags are 0–4. So we have a total of 20 fine-grained class labels, corresponding to fine-grained class labels 0–19. The composition of the data samples used in the fine-grained classification model is shown in Table 4.

5. Results and Discussion

5.1. Experimental Results

5.1.1. The Results of CNN-S. CNN-S is the most common method. During the model training, when the training set accuracy reaches 100% and the loss function no longer decreases, the average fault diagnostic accuracy of each working condition diagnostic model on the test set is only

95.5%, 88.5%, 94.5%, and 95.8%. It shows that the network structure is reasonable and the functional operation is preliminarily effective. However, due to the limitation of the number of samples (each model only has 600 samples per health state in total), it has not been fully activated in training, which appears to be a typical overfitting. In order to understand the diagnosis results in more detail, Figure 7 shows the confusion matrix of the CNN-S model on each working condition test set and the overall statistics. On the confusion matrix plot, the rows correspond to the predicted class and the columns correspond to the true class. Numbers 0~4 mean the health state labels as shown in Table 2. The diagonal cells correspond to observations that are correctly classified. The number of observations and the percentage of the total number of observations are indicated in each cell. The column on the far right of the plot shows the precision and false discovery rate, respectively. The row at the bottom of the plot shows the recall and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy. The blocks (a)~(d) show the confusion of each separated condition and their statistical values of merging conditions are shown in block (e).

In order to analyze the fault diagnosis results more intuitively, t-distributed stochastic neighbor embedding (t-SNE) algorithm is used to visualize the extracted features. The results are shown in Figure 8.

FIGURE 7: Confusion matrix for CNN-S. (a) Condition 1. (b) Condition 2. (c) Condition 3. (d) Condition 4. (e) Statistics of all single conditions.

Figure 8(a) shows the classification results of the extracted features of the corresponding working condition 1 data set. From the figure, we can see that the class spacing between the

scattered points of different health states is relatively small, and some scattered points even overlap each other. It is easy to cause the two health states to be incorrectly identified,

FIGURE 8: The t-SNE figure of feature extracted by CNN-S: (a) condition 1, (b) condition 2, (c) condition 3, (d) condition 4.

which will lead to misdiagnosis. For the scattered points of the same health state, the inner distance is relatively large in some classes, and some scattered points of the health state may even be distributed across classes. This will result in the inability to effectively identify the health state and cause failure to diagnose. The above analysis results show that the CNN-S model has no obvious effect on the classification of the health status under working condition 1 and cannot accurately diagnose the fault status. The visualized classification results of the extracted features of the data sets under working condition 2, working condition 3, and working condition 4 are shown in Figures 8(b)–8(d), respectively. Similarly, the classification results of these three working conditions show the following: the cases where the distance between classes of different health states is relatively small, and the distance between classes in the same health state is relatively large.

5.1.2. The Results of CNN-M. In order to prove the influence of working conditions on gearbox fault diagnosis, and because the sample size is not enough to support CNN-S, another common method of CNN-M is used here. The final diagnostic accuracy of CNN-M is 79.9% on test set. Although data from all working conditions are based on a single model, the number of samples increases, but the accuracy of the final diagnosis is lower than that of one of CNN-S's four unique work status models.

Figure 9 shows the confusion matrix of the extended sample data set shown in Table 3. From Figure 9, the same CNN network structure with the same network parameters can be seen, the diagnostic accuracy for the extended sample data set is 66% to 93%, and the average accuracy is only 79%, which is much lower than any of the four single working condition models in CNN-S. Further, the feature extraction visualization of CNN-M is shown in Figure 10. It shows that

Confusion matrix

0	321 13.4%	38 1.6%	6 0.3%	32 1.3%	90 3.8%	65.9% 34.1%
1	56 2.3%	316 13.2%	1 0.0%	7 0.3%	9 0.4%	81.2% 18.8%
2	1 0.0%	1 0.0%	462 19.3%	0 0.0%	0 0.0%	99.6% 0.4%
3	68 2.8%	92 3.8%	6 0.3%	440 18.3%	3 0.1%	72.2% 27.8%
4	34 1.4%	33 1.4%	5 0.2%	1 0.0%	378 15.8%	83.8% 16.2%
	66.9% 33.1%	65.8% 34.2%	96.3% 3.7%	91.7% 8.3%	78.8% 21.3%	79.9% 20.1%
	0	1	2	3	4	

Target class

FIGURE 9: Confusion matrix for CNN-M.

FIGURE 10: The t-SNE visualization figure of features extracted from CNN-M.

there is a serious overlap between the scattered points of different health status features, which is more serious than any of the cases in Figure 8. In addition, the intraclass spacing between the scattered points of the same health status feature becomes larger, which indicates that the accuracy of fault diagnosis on the expanded sample data set is very low. The results show that the working condition cannot be ignored as the key element of gearbox fault diagnosis. An appropriate global modeling method needs to distinguish all faults under different working conditions.

5.1.3. The Results of CNN-FG. The final diagnostic accuracy of CNN-FG is 98.8% on test set. Figure 11 shows the fault diagnosis accuracy of the proposed method, both the coarse class (working condition) and the fine-grained class (health state). It

FIGURE 11: Fine-grained classification (health state) accuracy and coarse classification (working condition) accuracy of CNN-FG.

Confusion matrix

0	1200 25.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
1	0 0.0%	1200 25.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	0 0.0%	1200 25.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	0 0.0%	1200 25.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%
	0	1	2	3	

Target class

FIGURE 12: Confusion matrix of coarse class recognition using CNN-FG.

can be seen that the specific working condition information of the input signal and the specific health state can be accurately identified. This also reflects the effectiveness of the hierarchical label structure proposed in this paper. It combines the working condition information with the health state information, and these two complement each other, providing more reliable information for signal fault diagnosis.

The confusion matrix of CNN-FG is shown in Figure 12, which shows the ability of CNN-FG model to recognize the coarse class of the data samples. From the figure, it can be seen that, for any data sample under working conditions 1~4, the CNN-FG model is able to identify the working conditions accurately. The numbers 0~3 are coarse class labels defined in Table 4.

0	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
1	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
2	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
3	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
4	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	238 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	99.6% 0.4%	
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	99.2% 0.8%	
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	99.6% 0.4%	
10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	238 5.0%	0 0.0%	0 0.0%	11 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	95.6% 4.4%	
11	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	237 4.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
12	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
13	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	3 0.1%	0 0.0%	229 4.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	97.9% 2.1%	
14	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%	
15	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	222 4.6%	3 0.1%	0 0.0%	0 0.0%	21 0.4%	
16	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	18 0.4%	236 4.9%	0 0.0%	0 0.0%	0 0.0%	
17	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	0 0.0%	
18	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	240 5.0%	0 0.0%	
19	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	219 4.5%	
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	99.2% 0.8%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	99.2% 0.8%	98.8% 1.2%	100% 0.0%	95.4% 4.6%	100% 0.0%	92.5% 7.5%	98.3% 1.7%	100% 0.0%	100% 0.0%	91.3% 8.7%	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
	Target class																				

FIGURE 13: Confusion matrix of fine-grained class recognition using CNN-FG.

Figure 13 shows the confusion matrix of fine-grained class recognition using fine-grained classification model. It can be seen from the right column that all accuracies are almost 100%, which means that the fine-grained classification model can accurately identify the 20 fine-grained classes.

In order to compare the two common methods (a) and (b) more intuitively, the diagnosis results of CNN-FG method are merged according to the working conditions. The statistical results are shown as confusion matrix, as

shown in Figure 14. Obviously, this method is much better than (a) and (b).

Figure 15 shows the t-SNE visualization of the overall feature extraction of the fine-grained classification model. It is obvious that the fine-grained classification model can clearly identify all 20 health states from the joint feature distribution of states across working conditions. The above results clearly show the superiority of the fine-grained classification algorithm for fault diagnosis under the coupling of working conditions and health states.

Confusion matrix

0	938 19.5%	3 0.1%	0 0.0%	11 0.2%	21 0.4%	96.4% 3.6%
1	18 0.4%	954 19.9%	0 0.0%	0 0.0%	0 0.0%	98.1% 1.9%
2	0 0.0%	0 0.0%	960 20.0%	0 0.0%	0 0.0%	100% 0.0%
3	4 0.1%	3 0.1%	0 0.0%	949 19.8%	0 0.0%	99.3% 0.7%
4	0 0.0%	0 0.0%	0 0.0%	0 0.0%	939 19.6%	100% 0.0%
	97.7% 2.3%	99.4% 0.6%	100% 0.0%	98.9% 1.1%	97.8% 2.2%	98.8% 1.2%
	0	1	2	3	4	

Target class

FIGURE 14: Statistics of CNN-FG.

FIGURE 15: The t-SNE visualization of the overall feature extraction of CNN-FG.

5.2. Discussion. The comparison of the traditional CNN methods (CNN-S and CNN-M) with the CNN-FG proposed in this paper is shown in Figure 16. Figure 16(a) is the comparison of test loss drop curves of those three methods. We can see that the proposed algorithm can achieve lower loss on the test set. It indicates that the proposed method has better generalization performance. As mentioned above, the network output of CNN-S and CNN-M model has only five classes, and the network output of CNN-FG contains a total number of 20 fine-grained classes that belong to 4 coarse classes, so the latter loss function value is larger at the beginning. However, as the training progresses, the loss function value of the latter decreases rapidly and is lower than the traditional CNN methods. It indicates that the fine-grained classification algorithm based on the working condition is more suitable to describe the problem of fault

diagnosis for variable working conditions. Figure 16(b) shows the fault diagnosis accuracy comparison between the traditional CNN methods and the fine-grained fault diagnosis method proposed in this paper. It can be seen that the method used in this paper is significantly better than the traditional CNN methods in fault diagnosis recognition accuracy. It indicates that the traditional CNN methods which directly use the CNN model to extract fault signal features without considering the working conditions have obvious defects. The use of fine-grained classification algorithm to extract the common features of different health states under different working conditions more fully explains the problems addressed in this paper. When the health status features are significantly affected by the working conditions, accurately identifying the health states means that the working conditions must be accurately identified at

FIGURE 16: The comparison of traditional methods and fine-grained classification algorithm. (a) The comparison of test loss drop curves. (b) The comparison of faults diagnosis accuracies.

TABLE 5: Comparison of fault diagnosis accuracies between traditional CNN models and fine-grained CNN model.

Working condition	Traditional CNN models		Fine-grained CNN model	
	Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)
Single working condition (CNN-S)				
Condition 1	95.5	93.6	—	—
Condition 2	88.5		—	
Condition 3	94.5		—	
Condition 4	95.8		—	
Multiple working conditions				
Conditions 1–4	79.9 (CNN-M)	98.8 (CNN-FG)		

the same time, and the implementation of the two should ultimately be synchronous.

Table 5 shows the comparison of fault diagnosis accuracies between traditional CNN models and fine-grained CNN models. The advantages of fine-grained classification methods in diagnostic accuracy can be seen very intuitively from the table.

6. Conclusion

The innovations of the proposed algorithm are twofold: (1) modify the CNN structure, introduce the skipping connection, and use softmax cascade loss to train the network, so that the CNN model has the ability to extract the common characteristics of the types of failures and working conditions, and use these common characteristics. (2) The hierarchical label structure allows CNN to simultaneously identify the types of failures and working conditions and at the same time improve the accuracy of the diagnosis of the types of failures. Compared with traditional CNN methods, the fine-grained classification model can not only effectively expand the number of original samples by combining the

samples of different working conditions. Considering the influencing factors of working condition coupling, the coupling between working conditions and health status is unified into the same CNN model, so as to extract features from all samples to the maximum and complete the feature extraction across the working conditions. Finally, the decoupling of working conditions and health status is achieved, and the fault diagnosis accuracy is effectively improved. The problem of the sample expansion of the planetary gearbox under a limited sample situation and the problem that the working conditions and the health states are coupled to reduce the diagnosis accuracy are solved. The results show that this method is superior to existing methods.

It should be noted that this method is very suitable for the gearbox, or other systems whose characteristics are significantly affected by working conditions, and that it is determined that it is difficult to obtain sampling on a large scale. When the influence of working conditions is limited, the SIMPLE CNN-M method can be used. If the system is heavily affected by working conditions, with large sample size, CNN-S can achieve the same effect. However, where

both exist, the methods provided in this paper are irreplaceable.

Data Availability

The data from military equipment that cannot be disclosed.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant no. 51875576.

References

- [1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [2] L. Wen, L. Gao, X. Li, L. Wang, and J. Zhu, "A jointed signal analysis and convolutional neural network method for fault diagnosis," *Procedia CIRP*, vol. 72, pp. 1084–1087, 2018.
- [3] G. Sheng, Y. Tao, G. Wei et al., "A novel fault diagnosis method for rotating machinery based on a convolutional neural network," *Sensors*, vol. 18, no. 5, p. 1429, 2018.
- [4] 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.
- [5] D. T. Hoang and H. J. Kang, "Rolling element bearing fault diagnosis using convolutional neural network and vibration image," *Cognitive Systems Research*, vol. 53, pp. 42–50, 2018.
- [6] F. Jia, Y. Lei, J. Lin, X. 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.
- [7] T. Wang, H. Wu, M. Ni et al., "An adaptive confidence limit for periodic non-steady conditions fault detection," *Mechanical Systems and Signal Processing*, vol. 72–73, pp. 328–345, 2016.
- [8] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 8, pp. 1926–1935, 2017.
- [9] M. Cerrada, G. Zurita, D. Cabrera et al., "Fault diagnosis in spur gears based on genetic algorithm and random forest," *Mechanical Systems and Signal Processing*, vol. 70–71, pp. 87–103, 2015.
- [10] K. Yan and D. Zhang, "Calibration transfer and drift compensation of e-noses via coupled task learning," *Sensors and Actuators B: Chemical*, vol. 225, pp. 288–297, 2016.
- [11] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2017.
- [12] S. Shao, S. McAleer, R. Yan et al., "Highly-accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2018.
- [13] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 539–546, San Diego, CA, USA, June 2005.
- [14] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1988–1996, 2014.
- [15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815–823, Boston, MA, USA, June 2015.
- [16] L. Wen, X. Li, and L. Gao, "A new two-level hierarchical diagnosis network based on convolutional neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, pp. 1–9, 2019.
- [17] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," *Computer Vision-ECCV 2016*, pp. 499–515, 2016.
- [18] S. Weiwei, G. Yihong, T. Xiaoyu et al., "Fine-grained image classification using modified DCNNs trained by cascaded softmax and generalized large-margin losses," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 99, pp. 1–12, 2018.