Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090207
ISSN 2371-9524 Vol. 9 Num. 2

Fault diagnosis method for lightweight gearboxes based on depth-separable cascaded residual block and feature-weighted module

Jiangfu Liu[1,a], Jianchao Zhang[1,b,*], Yihui Mo[2,c]

[1]Beijing Research Institute of Automation for Machinery Industry Co., Ltd, Beijing, China
[2]SINOPEC Maoming Petrochemical Company, Maoming, China
[a]liujiangfu@riambsoft.com, [b]zhangjianchao@riambsoft.com, [c]moyh.mmsh@sinopec.com
[*]Corresponding author

Keywords: Fault diagnosis; gearbox; deeply separable convolution; feature weighting module

Abstract: Aiming at the problem of insufficient feature extraction in some deep learning-based gearbox fault diagnosis models under small sample conditions leading to lower fault diagnosis accuracy and larger number of parameters, in this paper, a lightweight gearbox fault diagnosis method based on depth-separable cascade residual block and feature weighting module is proposed. Firstly, the one-dimensional original signal of the gearbox is used as the input of this model, which reduces the loss of information in data processing. Then the depth-separable cascade residual block is constructed, which utilizes the depth-separable convolution with a cascade residual structure to maximize the extraction of fault information while reducing the amount of feature parameters. Finally, the feature weighting module strengthened the model's identification and exploitation of key features by calculating the contribution of each channel and giving them weighting. The experimental validation is given by the gearbox dataset of Southeast University, and the experimental results show that the proposed method achieves 99.99% fault diagnosis accuracy under the original signal, and 99.60% under the SNR=6dB noise environment, which shows that the proposed method has high fault diagnosis accuracy and low complexity under the small sample condition.

1. Introduction

In modern industry and transport field, gearbox as the core power transmission and speed change device, its performance is directly related to the efficiency, reliability and economy of the whole mechanical system [1,2]. The gearbox can accurately regulate the speed and torque to satisfy meet the demands of different working conditions and to ensure the smooth operation of mechanical systems in a variety of complex environments. However, there are many challenges to the efficient and reliable operation of gearboxes [3,4]. Due to continuous load operation, wear and tear, lubrication failure, as well as improper operation and other factors, gearboxes may have various failures, such as gear wear, bearing damage, or lubricant deterioration, etc. These failures, if not

detected and handled in a timely manner, may lead to inefficiency of the entire system, or even huge casualties and property losses [5-7].

Traditional gearbox fault diagnosis methods [8,9] are able to obtain fault characteristics and complete fault diagnosis by using data samples to construct mapping relationships between input features and output types. Zhou et al [10] proposed a gearbox fault diagnosis method based on improved fine composite multi-scale inverse weighted entropy and support vector machine (SVM) in order to solve the problem that the coarse-graining processing of multi-scale weighted entropy is prone to information loss and unable to comprehensively extract the fault information of gearboxes. This method introduces fine composite multi-scale inverse weighted permutation entropy to effectively alleviate the shortcomings of the traditional coarse-grained processing and strengthen the quality of fault features, and then the SVM classifier was used to give recognition to the fault features. Xie et al [11] proposed an adaptive variational modal decomposition (VMD) based fault diagnosis method for gearboxes, which uses a comprehensive evaluation index to adaptively select the K-value in the variational modal decomposition (VMD), and reconstructs the original signals to achieve noise reduction and feature enhancement of the signals. Then the feature vectors of the moving signal are fully extracted using the fine composite multiscale scattering entropy. Finally, the extracted features are given recognition using the kernel limit learning machine optimized by particle swarm algorithm. The problem that the vibration signals collected in gearbox fault diagnosis contain complex noise interference and redundant components is effectively solved. Although these methods have achieved better diagnostic results, the traditional methods are overly dependent on artificial prior knowledge, resulting in poor diagnostic stability of these methods.

In recent years, with the continuous development of deep learning technology, more and more researches have begun to focus on the use of this technology to improve the accuracy and efficiency of gearbox fault diagnosis. Deep learning models, especially Convolutional Neural Networks (CNNs) [12,13], have been widely used to automatically detect and identify early fault signals in gearboxes due to their superior capabilities in feature extraction and pattern recognition. These models are able to learn complex fault features from large amounts of sensor data and optimize their parameters through training to adapt to different operating environments and fault types. Yan et al [14] reduced the interference of noise on the original signal by Gaussian filtering, and used Transformer and convolutional neural network to fully extract the fault information, and experimental validation was given using two gearbox datasets, which effectively solved the problem that the operating environment of some gears is complex, which leads to the problem of insufficient sample data collected. Gu et al [15], in order to solve the problems of poor model generalization ability and low diagnostic accuracy caused by sample distribution imbalance in planetary gearbox fault diagnosis, used Gram's angle field image coding technique to convert 1D vibration signals into 2D images, and then used the 2D images as inputs for deep convolutional generative adversarial network to achieve the expansion of the fault samples, and finally trained the original 2D images and generated image samples with an AlexNet convolutional neural network to achieve fault diagnosis of gearbox. Tan et al [16] proposed a gearbox fault diagnosis method based on graph attention network, which uses the spectrum of the original signal to define the nodes and edges, constructs the graph structure, then inputs the graph structure into the graph attention network, and embeds the neighbour self-attention mechanism in the graph attention network to adaptively extract the node features and structural features of the graph structure, and finally gives the fault identification using the classifiers, which effectively solves the problems such as non-smoothness, feature aliasing and low diagnostic correctness of wind turbine gearbox fault vibration signals. Xie et al [17] decomposed the original signal into a number of eigenmode components by variational modal decomposition and used wavelet thresholding to perform noise reduction on the eigenmode components and reconstruct the signal, and then the reconstructed signal was

transformed into grey scale map and then inputted into convolutional neural network to give fault diagnosis. Experimental validation is given through gearbox data, and the results show that the proposed method has a high fault diagnosis accuracy, which solves the problems of the complexity of the actual working environment of the gearbox, the traditional method of extracting features, and the lack of performance of grey-scale map extracted features.

The above methods have high fault diagnosis accuracy under different conditions by constructing deep learning models, however, there are also the following problems: (1) All of the above methods use sufficient data to train the network model, however, in the actual industrial process, it is very difficult to obtain sufficient data. (2) All of the above methods give processing to the original data through data processing methods, without considering the problem of model lightweighting, which may lose part of the fault information while increasing the model complexity. Therefore, to solve the above problems, in this paper, an end-to-end fault diagnosis method for lightweight gearboxes is proposed. Firstly, the original signal is directly input into the deep learning model to give fault diagnosis, which avoids the problem of losing some fault features. Second, the depth separable convolution and cascaded residuals are used to construct a depth separable cascaded residual block, which effectively extracts the fault features while having fewer model parameters, and finally, the feature weighting module is constructed to give weights to different channels and adjust the weight values, so that the model is able to mine more important feature information. The experimental validation is given by gearbox data from Southeast University, and the results show that the proposed method has a high fault diagnosis accuracy.

2. Basic Theory

Deep separable convolution is a technique commonly used in convolutional neural networks, which is rapidly developing with a unique step-by-step computation, it not only extracts and transforms rich feature information, but also reduces the number of parameters and computational complexity of the model. First, it applies a deep convolution operation to the input data, and then it applies a point-by-point convolution operation to the output of the deep convolution.

Deep convolution is a technique that performs convolution independently for each channel of the input data, suppose the input data has c channels, deep convolution uses c separate convolution kernels to perform convolution operations on each channel, deep convolution is the convolution operation is given independently on each channel, thus reducing the number of references.

Point-by-point convolution is an operation that gives a convolution of the output of the deep convolution using the convolution of a 1×1 convolution kernel to convert it to the desired output dimension. It serves to give a linear combination between channels of the output feature maps of the deep convolution to get the final output feature maps.

Suppose the size of the convolution kernel for convolution is $K_h \times K_w$, the number of input channels is C_i, and the number of output channels is C_o. Then the formula for the number of parameters obtained after the input signal is processed by ordinary convolution is shown in Eq. (1):

$$num = (K_h \times K_w \times C_i + 1) \times C_o \tag{1}$$

And the formula for calculating the amount of feature parameters and operations after deeply separable convolution is shown in Eq. (2):

$$num = (K_h \times K_w \times 1 + 1) \times C_i + (1 \times 1 \times C_i + 1) \times C_o \tag{2}$$

It can be seen that the number of parameters of the deep separable convolution is much smaller than that of the ordinary convolution, which shows that the deep separable convolution can

effectively improve the computational efficiency of the network model.

3. Fault diagnosis method for lightweight gearboxes based on depth-separable cascaded residual block and feature-weighted module

3.1. Depth separable cascade residual block

The real working environment of a gearbox is very harsh, and the signals generated by its own vibration and friction are often mixed with the noise generated by the surrounding environment, which makes it difficult to extract fault features. Extracting features using multiple convolutional layers and feature mapping is a common approach, however, this operation will not only increase the amount of feature parameters, but also make the deep features and the original features out of correlation, resulting in poor fault diagnosis. To address the above problems, in this paper, the depth separable cascade residual block is designed, and the structure of this module is shown in Fig. 1.

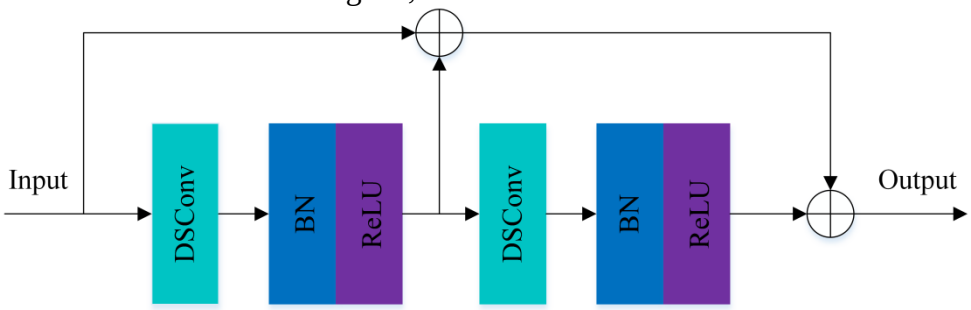


Figure 1: Depth separable cascade residual block

To address the problem of difficult feature extraction and large amount of feature parameters, the module chooses 2 layers of depth-separable convolution with different scales for feature extraction. The first layer of the module uses a depth-separable convolution with a convolution kernel size of 7. This convolution captures the global fault features hidden in the interference signal in a wide range, which can effectively resist the performance impact of noise on the proposed model. The second layer uses a depth-separable convolution with convolution size 3. This convolution maps the features extracted in the upper layer into a higher nonlinear space to generate a high-level feature representation. This operation amplifies weak fault features to the extent that these weak faults are not lost in the later layers of propagation.

To address the problem of deep features being uncorrelated with the original features, this module concatenates the original inputs in the second and third layers of inputs. This connection allows the model to learn the residual information by "skipping" certain layers, which allows the original features to be learned repeatedly at each layer. This operation not only helps to improve the model training and convergence speed, but also does not detach the original fault features as the network depth increases. To improve the computational efficiency and performance of the model, each depth-separable convolution of the module is followed by a Relu activation function and layer normalization.

3.2. Feature weighting module

After feature extraction with depth-separable cascaded residual block, the features of different fault types are distributed irregularly in different channels. To reduce the complexity of the model, the channels containing more fault features should be given larger weights, while the channels containing fewer fault features or no features should be given smaller weights. To address this problem, the feature weighting module is designed, and the structure of this module is shown in Fig.

2.

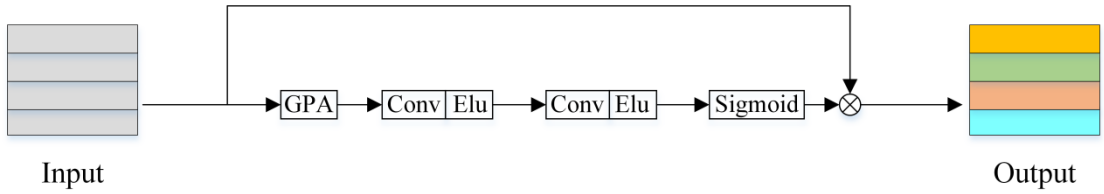


Figure 2: Feature weighting module

Firstly, this module uses a global average pooling layer to downscale and compress the features as shown in Eq. (3), which ultimately compresses the input features to 1×1×C. The benefit of doing so is that it reduces the number of parameters and computation, and improves the efficiency of the model. At the same time, global average pooling helps to capture the global importance of each channel in the feature map since it extracts the overall information of the feature map.

$$Out_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{3}$$

where c is the serial number of the compressed channel, Out_c is the compressed value of the first c-channel; and $u_c(i,j)$ is $i \times j$ dimensional two-dimensional matrix.

Secondly, the module uses a one-dimensional convolution to nonlinearly transform the compressed global description, thereby increasing the expressiveness of the feature map.This introduces more flexibility and non-linearity, allowing the model to learn more complex and abstract feature representations. Next, to increase the performance of the model, this module adds an ELU activation function behind the 1D convolution. This function is an improvement on the Relu function, on one hand, the ELU function is not a constant zero on the negative interval, thus avoiding the zero-gradient problem of Relu and making the neural network more stable during the training process. On the other hand, ELU is smooth and conductible over the entire range of real numbers, including the negative interval. In contrast, Relu is not derivable on negative intervals, which may lead to the problem of vanishing or exploding gradients, and the smoothness of ELU helps to improve the training speed and stability of the model. This step-by-step process is shown in Eq. (4):

$$s_c = f(W_2\sigma(W_1 Out_c)) \tag{4}$$

where s_c is the fault feature after activation, $f(\bullet)$ is the sigmoid activation function, W_1 and W_2 are the weight parameters of the two 1D convolutions, $\sigma(\bullet)$ is the ELU activation function.

Finally, the weights of the excited features are multiplied with the original feature map to obtain a weighted feature map. This weights the feature map according to the importance of each channel, which allows the model to focus more on important features and suppress unimportant ones. This helps to improve the performance and generalization of the model, enabling the model to adapt better to different tasks and data, the process is shown in Eq. (5).

$$F = s_c \bullet Out_c \tag{5}$$

where F is the weighted feature map.

3.3. Fault diagnosis method

To address the problem of insufficient feature extraction in some deep learning-based gearbox

fault diagnosis models under small sample conditions leading to lower fault diagnosis accuracy and larger number of parameters, a lightweight gearbox fault diagnosis method based on a depth-separable convolutional feature extraction module and a feature weighting module is proposed. The method mainly consists of an input layer, a depth-separable cascade residual block, a feature weighting module, a dimensionality reduction layer and an output layer, and the structure is shown in Fig. 3. In which, the task of the input layer is to normalize the vibration signals collected by the sensors. The Deep Separable Cascade Residuals module is used to maximize the extraction of fault information while reducing the number of feature parameters. The Feature Weighting module improves the identification and use of key features in the model by calculating the contribution of each channel and weighting it to remove redundant information.

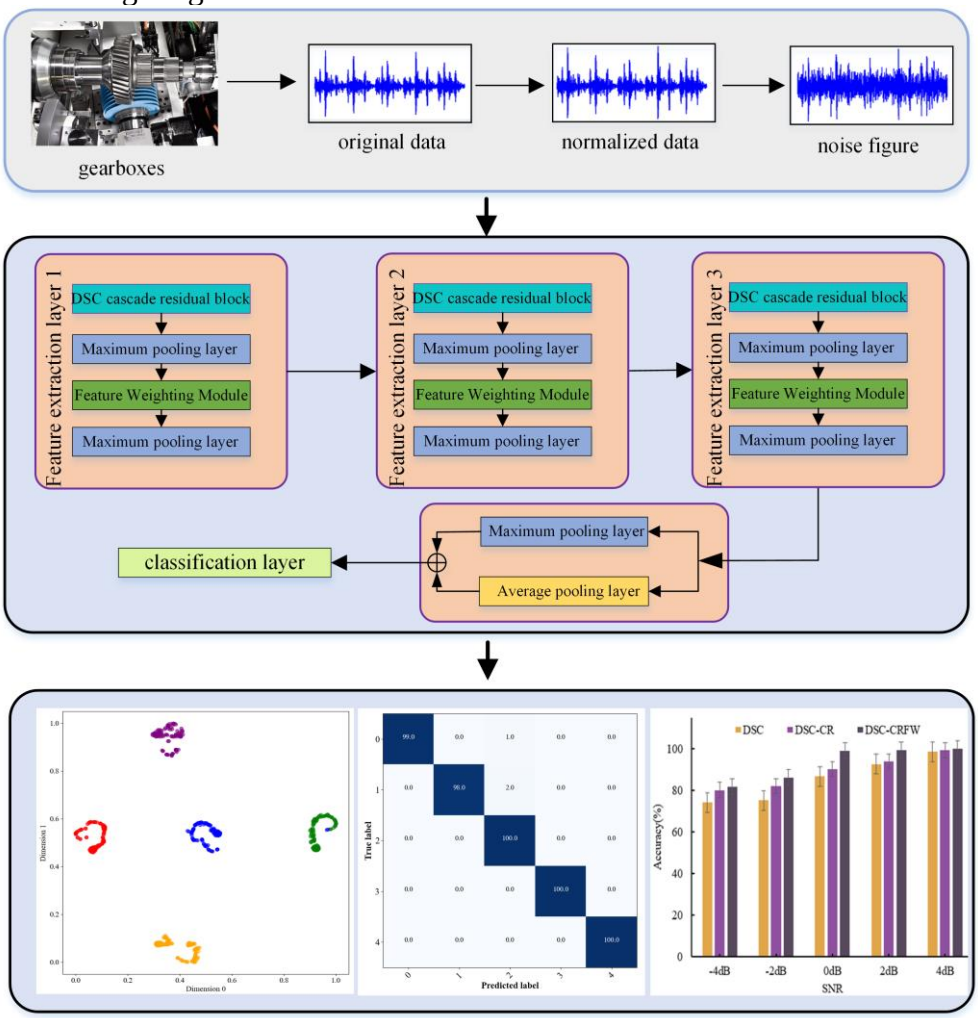


Figure 3: Structure of the proposed method

3.4. Fault Diagnosis Process

The fault diagnosis flowchart of the proposed method is shown in Fig. 4, and its diagnosis detailed steps are:

Step 1: Gearbox vibration signals are collected using sensors.

Step 2: Firstly, the vibration signals are normalized, secondly, the normalized data are divided into training set, verification set and test set according to a fixed ratio.

Step 3: Firstly, the model is initialized, and then the training set is fed into the proposed model

for feature extraction and learning, and the model parameters are updated using the back-propagation algorithm.

Step 4: Judge whether the number of model training times reaches the designed batch, if so, save the model, if not, continue training.

Step 5: Feed the test set data into the saved model for testing and output the fault diagnosis results.

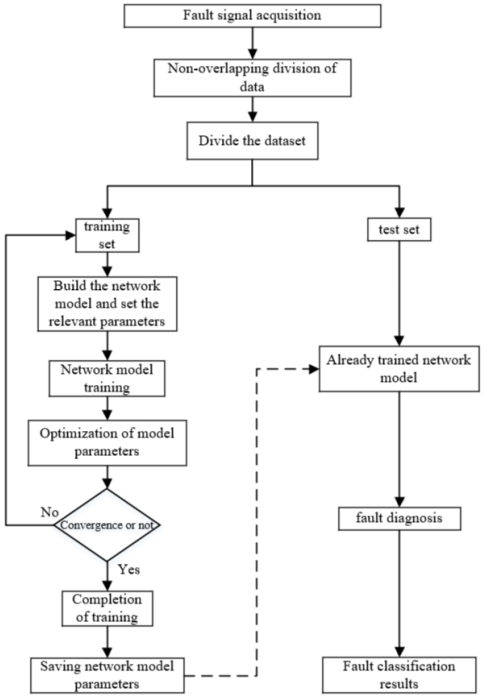


Figure 4: Fault diagnosis flow chart

4. Experimental validation

In this paper, the gearbox dataset is selected for the experiment from Southeast University, and the test rig is mainly composed of motor, motor controller, planetary gearbox, reduction gearbox, and load controller. The dataset is collected from the Driveline Dynamic Simulator (DDS) shown in Fig. 5, with a total of 8 channels of signals and the sampling frequency of 5120 Hz. The gears for different failure states are machined in advance and the variable speed can be achieved by the motor controller, while the change in load can be achieved by the load controller.

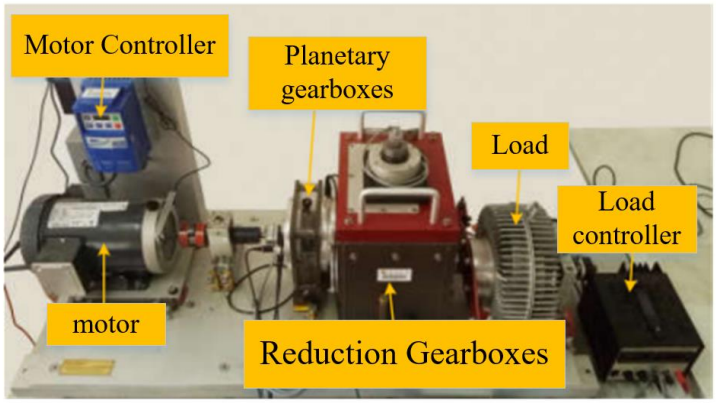


Figure 5: Installation diagram of the experimental platform

Two different operating conditions are primarily investigated, with the tacho system load set to 20HZ-0V or 30HZ-2V. The dataset contains five different operating conditions: four fault types and one health state. Therefore, fault diagnosis of the DDS is a 5-classification task, and data from 2 channels of the bearing dataset with the tacho load configuration set to 20Hz-0V are selected for verification, with 1,000 samples selected for each type of fault, and the length of each sample is 2,048 data points. The training, validation, and test sets are divided in the ratio of 8:1:1 and the dataset description is shown in Table 1.

Table 1: Division of the data set.

Type of fault	Sample training set	Sample Validation Set	Sample test set
defective	800	100	100
broken tooth	800	100	100
Tooth root cracks	800	100	100
Tooth surface wear	800	100	100
Normal condition	800	100	100

4.1. Data enhancement

To mitigate the potential negative impact of a small amount of anomalous data in a dataset on the accuracy of gearbox fault diagnosis, in this paper, data collected by sensors mounted on gearboxes are normalized. This process not only helps to improve the robustness of the model to anomalous data, but also when dealing with small samples of data, normalization enhances data consistency in the model learning process, thus improving the accuracy and reliability of fault diagnosis. Normalization ensures that data with different characteristics have a balanced impact on model training by adjusting the data scale, which is particularly important in small sample scenarios. Because each sample is extremely valuable, ensuring that they can be used to maximum effect during model training optimizes fault diagnosis performance, the mathematical expression for normalization is as follows.

$$z_i' = \frac{z_i - min(z_i)}{max(z_i) - min(z_i)} \tag{6}$$

where z_i' is the normalized data, z_i is the raw data, $min(z_i)$ is the minimum value and $max(z_i)$ is the maximum value.

In addition, the data enhancement method used in this paper is overlap sampling, which collects the original signal to select some samples as training samples, the length of each segment is 2048 samples, the overlap sampling points are 1984, the step size is 64, and the sampling process is shown in Figure 6. For the test set, there is no overlap sampling. Assuming that a segment of the signal has 10000 sampling points, the length of the training samples collected each time is 2048, and the step size is 1, a maximum of 7951 training samples can be produced, and the sampling formula is shown in Eq. (7).

$$\text{Sample} = \frac{total - point\_sample}{stride} + 1 \tag{7}$$

where *total* is the length of the sampling points, $point\_sample$ is the length of the signal sampling points per segment, and *stride* is the sampling step size.

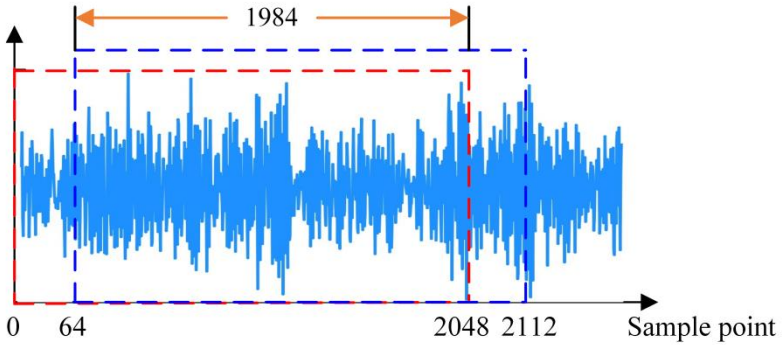


Figure 6: Overlapping sampling method

4.2. Comparison of methods

To prove the effectiveness of the methods and its advantages compared to other fault diagnosis methods, experiments with the same batch size, learning rate, and other hyperparameters are selected for comparison with other deep learning fault diagnosis methods, using accuracy as indicator. The results of different methods are shown in Table 2.

Table 2: Accuracy rate of different methods.

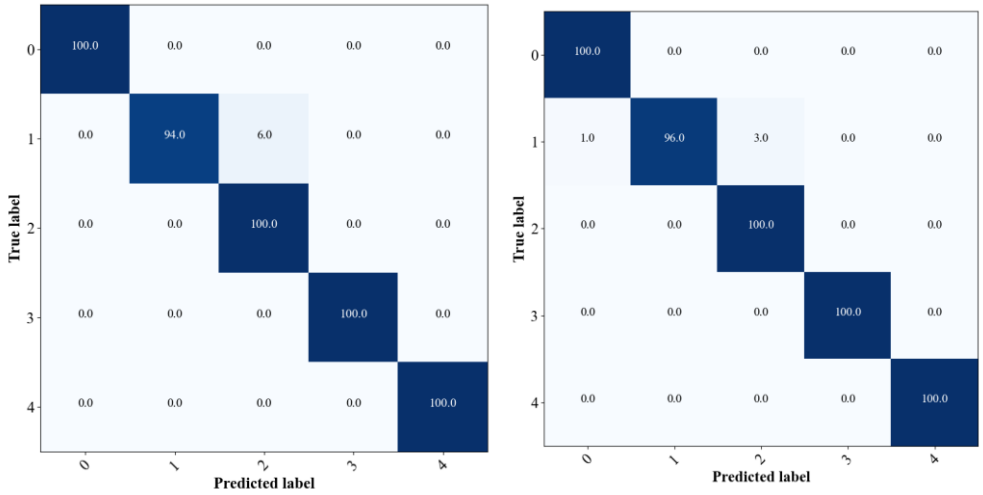
Method	Accuracy (%)
ANCEEMD [18]	97.26
DNN [19]	96.22
GAF-DCGAN [20]	98.57
MC-MSDARL [21]	97.00
Proposed method	99.99

As can be seen from Table 2, the DNN method has the lowest accuracy of 96.22%, the ANCEEMD, DPD-1DCNN and GAF-DCGAN methods have higher complexity and better results than DNN, with accuracies of 97.26%, 97.40% and 98.57%, respectively. Although ANCEEMD and MC-MSDARL methods can extract the features of the signal to some extent, for some complex failure modes, if the sample data is small, it may affect the effect of feature extraction, which requires further improvement and optimization of the algorithm model. Although GAF-DCGAN improves the image generalization by converting the data into two-dimensional image, the synthetic data generated by using a deep convolutional generative adversarial network for data enhancement may have a certain instability, which may introduce some noisy or unrealistic features and affect the accuracy of fault diagnosis. The proposed method uses deep separable convolution to effectively reduce the parameters and computational complexity, combined with the residual learning mechanism, which helps to solve the gradient vanishing problem when training the deep network. Meanwhile, it ensures the depth and efficiency of feature extraction, in addition, the proposed method strengthens the ability to recognize the key features by weighting the features in different channels.

4.3. Variable Noise Experiments

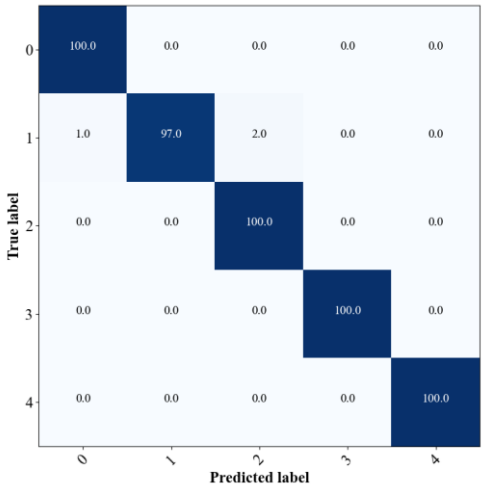
To compare the recognition accuracies of the methods under different noise levels more directly, the results are presented based on the confusion matrix, as shown in Fig. 7. Compared with the comparison methods, almost all methods achieve good results without noise. However, the increase in noise level decreases the accuracy of all the methods to varying degrees. In contrast, the proposed

method shows strong robustness to strong noise and an accuracy of 99.60%, which proves that the proposed method has strong noise immunity.

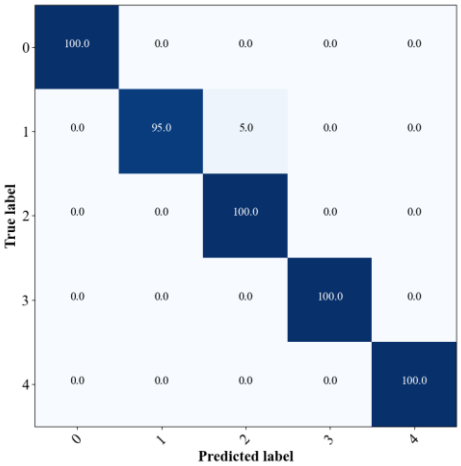


(a) DNN

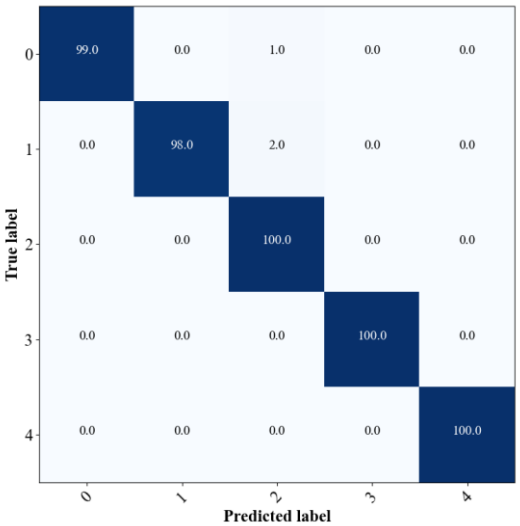
(b) ANCEEMD



(c) GAF-DCGAN



(d) MC-MSDARL

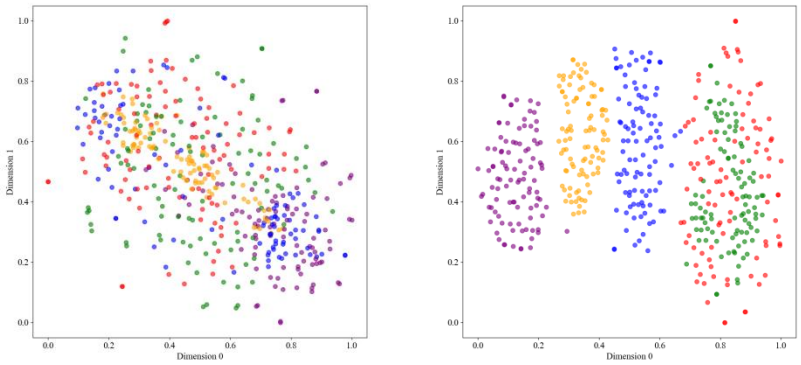


(e) Proposed method

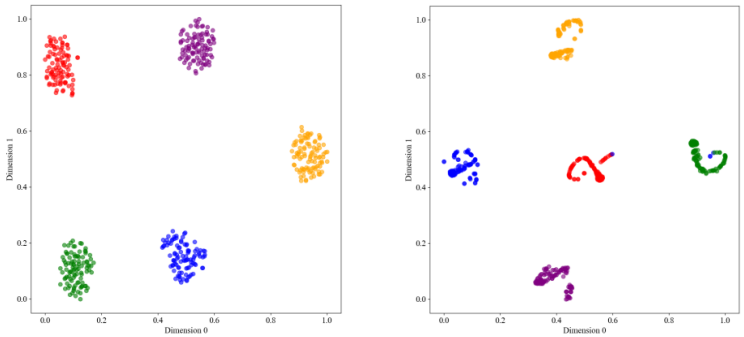
Figure 7: Diagnostic accuracy of different methods at 6dB

4.4. TSNE Visualization

To visualize the performance of this paper's method under different noises, Fig. 8 gives the visualization results of the different layers of the proposed method at 6 dB SNR.



(a)Input data (Left) (b) First depth separable joint residual block (Right)



(c)Third depth separable joint residuals block (Left) (d) Classification layer (Right)

Figure 8: Visualization results of different layers with SNR=6dB

In Fig. 8, the graph (a) shows the original signal visualization results, the five types of defects are disordered, overlapped with each other and indistinguishable from each other. The graph (b) shows the visualization results of the first depth separable joint residual block, the five types of faults have a tendency to separate from each other, but they still can't be effectively distinguished. The graph (c) shows the visualization results of the third depth separable joint residual block, the five types of faults are separated from each other, and the graph (d) shows the visualization results of the classification layer, the five types of faults are better behaved in terms of feature clustering, where there is no obvious misclassification phenomenon for the different states. The results show that the proposed method can cluster accurately without misclassification.

4.5. Comparison of model parameters

To verify the specific effect of the proposed method in lightweight, four current mainstream methods are cited to compare with the proposed method under the data of adding 0dB SNR, the comparison methods are ANCEEMD, DNN, GAF-DCGAN, and MC-MSDARL, the experimental results are shown in Table 3. DNN and ANCEEMD have smaller parameters among the four methods, GAF-DCGAN has the most parameters, and MC-MSDARL is a method with good running speed and parameters. The proposed method is superior to other networks in terms of

parameters and accuracy. The main reason is that the proposed method reduces the parameters by deep separable convolution, calculates the contribution of each channel by using the feature weighting module, and weights them, which strengthens the model's identification and utilization of key features. Therefore, the proposed method can achieve high robustness and high accuracy at the cost of limited neurons.

Table 3: Comparison of parameters.

Method	Parameters	Time (ms)	Average accuracy (%)
DNN	5.51×10^4	100	92.18
ANCEEMD	7.72×10^5	447	94.32
GAF-DCGAN	10.28×10^7	2280	93.64
MC-MSDARL	8.24×10^6	1295	97.21
Proposed method	1.17×10^5	189	99.13

4.6. Ablation Experiments

In this experiment, we set up three experiments to verify the effectiveness of the fault diagnosis of each module in the proposed method. The network structures are DSC, DSC-CR, and DSC-CRFW, where DSC, CRB-CR, and DSC-CRFW indicate that only one structure is changed.DSC is the method that does not use the cascaded residual module.CRB-CR replaces the standard convolution with the cascaded residuals, and DSC-CRFW is outputted directly after three feature extraction layers, which means that it does not have the feature-weighting module, and the rest of the network parameters are the same as those in the proposed method is the same. Different signal-to-noise ratios are added for experimental comparison and the results are shown in Table 4 and Fig. 9.

Table 4: Comparison of ablation experiments.

Method	Accuracy (%)				
	-4dB	-2dB	0dB	2dB	4dB
DSC	74.22	75.17	86.68	92.7	98.54
DSC-CR	80.19	82.12	90.28	94.00	99.27
DSC-CRFW	81.88	86.29	99.13	99.29	99.98

As shown in Table 4, the DSC has lower diagnostic accuracy in high noise environments because it does not use the CR module, which results in lower detection accuracy than the DSC-CRFW at different signal-to-noise ratios.

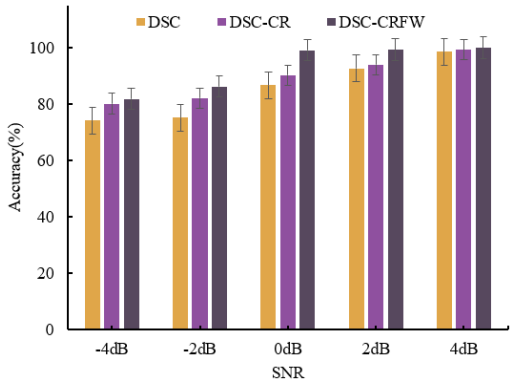


Figure 9: Ablation experiments at different SNRs

4.7. Noise resistance performance experiment under different number of samples

In the operating environment of bearings, the number of fault samples is uncertain, so the diagnostic method must be able to effectively accomplish fault identification with either more or fewer samples. However, when the number of samples is too large, it leads to increased model training time and memory consumption. Therefore, diagnostic methods need to accomplish fault identification with as few samples as possible.

To verify the effectiveness and generalization performance of the proposed method, a 4db white noise signal will be added to the original signal, the model will be trained using the different number of samples, and the sample selection will be repeated 5 times for each experiment to reduce the bias of randomly selecting a small training set, and the results will be averaged over 5 experiments to ensure that the results of the faults are reasonable. The experimental results are shown in Table 5.

Table 5: Fault diagnosis accuracy with different samples.

Total sample size	ANCEEMD	DNN	GAF-DCGAN	MC-MSDARL	Proposed method
120	72.58	68.74	80.55	82.62	90.27
180	77.49	70.88	84.33	88.53	93.22
240	80.28	75.68	88.29	90.27	94.58
300	85.44	80.55	90.18	93.33	96.55
360	88.96	88.74	93.21	94.28	98.88
420	94.52	93.20	96.66	97.24	99.90

As can be seen from Table 5, the accuracy of the proposed method gradually increases with the increase in the number of samples. When the number of samples is 360 and 420, the accuracy rates are 98.88% and 99.90%, respectively, and when the number of samples is increased by 60, the accuracy rate increases by 1.02%.

To verify the superiority of the proposed method with the small samples, we conducted a comparison experiment with the comparison method, and the results are shown in Fig. 10.

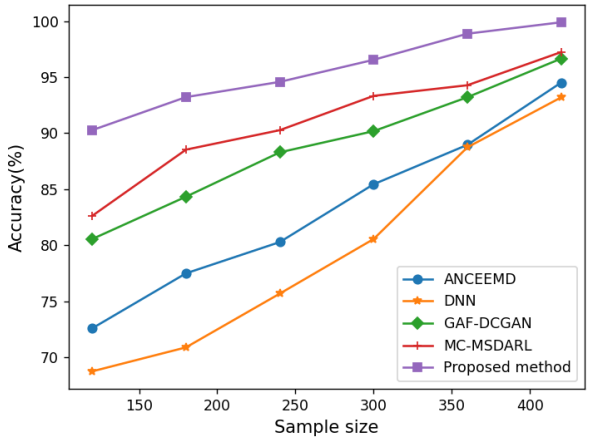


Figure 10: Fault category accuracy at 4 dB

As shown in Fig. 10, when the number of samples is 120, the accuracy of the comparison method is 72.58%, 68.74%, 80.55%, and 82.62%, respectively, while the proposed method is 90.27%. This shows that the method is still able to accomplish the fault diagnosis task with fewer samples in a less-sample environment.

5. Conclusions

To solve the problem of insufficient feature extraction in some deep learning-based transmission fault diagnosis models under small sample conditions, which leads to lower fault diagnosis accuracy and larger number of parameters, this paper proposes a lightweight transmission fault diagnosis method based on depth-separable cascade residual block and feature weighting module. The conclusions are as follows:

(1) The one-dimensional raw vibration signal of the transmission is used as the input of this model, which reduces the information loss in data processing.

(2) A depth-separable convolutional feature extraction module is constructed, which utilizes depth-separable convolution with cascaded residual structure to maximize the extraction of fault information while reducing the amount of feature parameters.

(3) The final feature weighting module strengthens the model's identification and utilization of key features by calculating the contribution of each channel and weighting them.

(4) Experimental validation is carried out using the Southeast University's gearbox dataset, and the experimental results show that the proposed method achieves a fault diagnosis accuracy of 99.99%, and the fault diagnosis accuracy is 99.60% in the SNR=6dB noise environment, which indicates that the proposed method has a high fault diagnosis accuracy and low complexity under small samples.

References

[1] Ye Zhuang, Wang Siyuan, Yue Shang, et al. Virtual-Real Fusion-Based Transfer Learning With Limited Data for Gearbox Fault Diagnosis [J]. IEEE Sensors Journal, 2024, 24(3): 3420-3430.

[2] Zhang Yongchao, Ding Jinliang, Li Yongbo, et al. Multi-modal data cross-domain fusion network for gearbox fault diagnosis under variable operating conditions[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108236.

[3] Ye Zhuang, Yue Shang, Yang Pu, et al. Deep Morphological Shrinkage Convolutional Autoencoder-Based Feature Learning of Vibration Signals for Gearbox Fault Diagnosis [J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 3510712.

[4] Liu Wenyu, Wang Qiang, Xu Feiyun. Multi-sensor gearbox fault diagnosis using generalized minimum entropy deconvolution and main frequency center extraction[J]. Measurement Science and Technology, 2024, 35(1): 015117.

[5] Xie Fengyun, Wang Gan, Shang Jiandong, et al. Gearbox Fault Diagnosis Method Based on Multidomain Information Fusion[J]. Sensors, 2023, 23(10): 4921.

[6] Xia Jingyan, Huang Ruyi, Chen Zhuyun, et al. A novel digital twin-driven approach based on physical-virtual data fusion for gearbox fault diagnosis[J]. Reliability Engineering and System Safety, 2023, 240: 109542.

[7] Jiang Fei, Lin Weiqi, Wu Zhaoqian, et al. Fault diagnosis of gearbox driven by vibration response mechanism and enhanced unsupervised domain adaptation[J]. Advanced Engineering Informatics, 2024, 61: 102460.

[8] Men Zhihui, Hu Chaoqun, Li Yonghua, et al. A hybrid intelligent gearbox fault diagnosis method based on EWCEEMD and whale optimization algorithm-optimized SVM[J]. International Journal of Structural Integrity, 2023, 14(2): 322-336.

[9] Jing Luyang, Wang Taiyong, Zhao Ming, et al. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox[J]. Sensors (Switzerland), 2017, 17(02): 414.

[10] Zhou Yuncheng, Pan Wei, Tan Qingsong. Application of Improved Fine Composite Multi-Scale Reverse Weighted Arrangement Entropy in Gearbox Fault Diagnosis[J]. Combined Machine Tool and Automatic Processing Technology, 2023 (5): 138-142,147.

[11] Xie Fengyun, Wang Gan, Appreciation Dong, et al. Gearbox fault diagnosis based on adaptive variational mode decomposition[J/OL]. Propulsion Technology, 1-11[2024-05-10].

[12] Zhang Jianqun, Zhang Qing, Qin Xianrong, et al. Robust fault diagnosis of quayside container crane gearbox based on 2D image representation in frequency domain and CNN[J]. Structural Health Monitoring, 2024, 23(01): 324-342.

[13] Xu Minmin, Han Yaoyao, Ding Xiaoxi, et al. Decision Self-Regulating Network for Imbalanced Working Conditions Identification in the Application of Gearbox Intelligent Fault Diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 3523711.

[14] Yan Huiyu, Zhang Chao. Gear fault diagnosis method based on Transformer and convolutional neural network[J]. Mechatronics Engineering, 2024, 41(03): 409-417.
[15] Gu Yingkui, Shi Changwu, Chen Jiafang. Fault Diagnosis of Planetary Gearbox Based on Gramm Angle Field and Deep Convolutional Generation Adversarial Network [J]. Noise and Vibration Control, 2024, 44(1): 111.
[16] Tan Qiyu, Ma Ping, Zhang Hongli, et al. Fault diagnosis of wind turbine gearbox based on graph attention network[J]. Acta Solar-Sinica, 2024, 45(01): 265-274.
[17] Xie Fengyun, Li Gang, Wang Linglan, et al. Gearbox fault diagnosis with improved timing grayscale map and deep learning [J/OL]. Computer Engineering and Applications, 1-9[2024-05-10].
[18] He Yan, Wei Xiuye, Cheng Haiji, et al. Intelligent fault diagnosis of planetary gearbox based on ANCEEMD sample entropy feature extraction[J]. Failure Analysis and Prevention, 2023, 18(03):155-163.
[19] Huipeng C, Cheng Z. Fault diagnosis method of planetary gearbox based on empirical mode decomposition and deep convolution neural network [J]. Journal of Mechanical Engineering, 2019, 55(7): 9-18.
[20] Zhang Jianqun, Zhang Qing, Feng Wenzong, et al. Gearbox fault diagnosis based on frequency-domain Gramian angular difference field and deep convolutional neural network[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2023, 237(21): 5187-5202.
[21] Chen Qi, Chen Changzheng, An Wenjie. Research on fault diagnosis of planetary gearbox based on multi-channel fusion multi-scale adaptive residual learning [J]. Mechatronic Engineering, 2023, 40(07): 1031-1038.