Mechanical Systems and Signal Processing 142 (2020) 106740

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Gear pitting fault diagnosis with mixed operating conditions based on adaptive 1D separable convolution with residual connection

Xueyi Li [a,c], Jialin Li [a], Chengying Zhao [a], Yongzhi Qu [b], David He [c,*]

[a] School of Mechanical Engineering and Automation, Northeastern University, Shenyang 110819, China
[b] School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, China
[c] Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA

ARTICLE INFO

Article history:
Received 31 May 2019
Received in revised form 11 January 2020
Accepted 13 February 2020
Available online 28 February 2020

Keywords:
Gear pitting fault diagnosis
Depthwise separable convolution
Residual connection
Vibration signals

ABSTRACT

Gear pitting fault diagnosis has always been an important subject to industry and research community. In the past, the diagnosis of early gear pitting faults has usually been carried out under single gear health state. In order to diagnose the early gear pitting faults with mixed operating conditions and reduce the number of training parameters, a new method is proposed in this paper. The proposed method uses an adaptive 1D separable convolution with residual connection network to classify gear pitting faults with mixed operating conditions. Compared to the traditional convolutional neural network, the separable convolution with residual connection network can carry out the channel convolution with point-by-point convolution to effectively reduce the number of network parameters. The residual connection can solve the representational bottleneck problem of the features in the model. Moreover, the method proposed in this paper applies the search algorithm to select better hyperparameters of the model. The raw vibration signals of the gear pitting faults at different speeds collected in a gear test rig are used to validate the effectiveness of the proposed method. The results show that the proposed method can accurately diagnose the early gear pitting faults with mixed speeds. In comparison with other machine learning models, the proposed method has provided a better diagnostic accuracy with fewer model parameters.

1. Introduction

Gears are an essential component of mechanical transmission systems and one of the most prone to pitting. Pitting is one of the most common gear faults and is often difficult to detect. Undetected pitting failure during gear operation can lead to catastrophic failure of the machine [1]. The diagnosis of gear pitting faults has always been a popular research subject. In recent decades, with the development and popularity of machine learning, many advanced methods have been developed for gear pitting fault diagnosis.

Over the years, feature selection has been still an essential aspect of machine learning based diagnostics in order to achieve good performance in diagnostic models [2]. Some researchers tried to propose an effective indicator to realize feature extraction [3,4]. Many researchers commonly use methods for transforming time-domain signals into frequency domain

* Corresponding author.
E-mail address: davidhe@uic.edu (D. He).

https://doi.org/10.1016/j.ymssp.2020.106740
0888-3270/

signals to extract features. Researchers tried to use time–frequency analysis for fault diagnosis [5 6]. Feng et al. [7] proposed a joint amplitude and frequency demodulation method for fault diagnosis of planetary gearboxes. For frequency demodulation analysis, fast Fourier transform was applied to the estimated instantaneous frequency to reveal its fluctuating frequency. Lei et al. [8] extracted the characteristic parameters in time domain, specifically for gear damage diagnosis and used frequency domain features to characterize the gear conditions. A two-stage feature selection and weighting technique (TFSWT) was proposed based on the Euclidean distance evaluation technique (EDET) to select sensitive features and eliminate features that are not related to the faults. Liang et al. [9] explained the vibration signals of the planet gears at the tooth flanks. Fault symptoms caused by a single split tooth of the planet gear could be extracted. The health of the planetary gears could be evaluated by comparing the differences between the signals of all the teeth of the planetary gears. Although these methods have a certain role in the fault diagnosis of rotating machinery such as gears, they still have some problems. First of all, these methods require extensive manual extraction of features and much prior knowledge on signal processing techniques and diagnostic experience. Besides, these features are extracted based on specific diagnostic conditions and may not apply to other conditions. Up today, few methods use the raw vibration signals directly for gear pitting fault diagnosis.

Machine learning based method fault diagnosis methods can be divided into two major types: supervised learning and unsupervised learning. In practical applications, the majority of the methods is still dominated by supervised learning. Common supervised learning methods include k-nearest neighbors, linear regression, logistic regression, support vector machines (SVM), decision trees, random forests, etc. Some people prefer SVM because of its elegant expression and high-quality performance. Other may like decision tree because of its simplicity and interpretability. Many researchers have conducted a lot of constructive discussion and research on gear fault diagnosis using the above-mentioned methods, such as artificial neural network (ANN) and SMV was used for binary classification of gear faults [10]. A genetic algorithm (GA) was used to optimize the number of nodes in the hidden layer in ANN. Li et al. [11] used deep random forest fusion (DRFF) technique to improve the fault diagnosis performance by simultaneously using acoustic emission (AE) and accelerometer measurements. A practical method based on hidden Markov model (HMM) was proposed for mechanical fault diagnosis [12]. Although many machine learning based classifiers have been used in industrial environments, in the case of more and more data collected in the industrial field, the capability of the traditional machine learning methods is somehow limited. Big data samples pose challenge to train the traditional machine learning models. Besides, the artificial neural network structure adopted by these methods is shallow, which limits the ability of ANN to learn complex nonlinear relationships in fault diagnosis problems.

Fortunately, as a breakthrough in solving the gradient disappearance problem, deep neural networks (DNN) with a deep architecture rather than a shallow architecture can be built to mine useful information from complex nonlinear data. An unsupervised learning method was developed based on a deep sparse autoencoder [13]. Li et al. [14] proposed augmented convolution sparse autoencoder (ACSAE) for gear pitting fault diagnosis with raw AE signals. The ACSAE method was used to diagnose gear pitting faults with fewer raw AE signal data. In supervised learning, Jia et al. [15] proposed an intelligent diagnosis method based on DNN. Li et al. [16] proposed a method of combining the convolutional neural network (CNN) and the gated recurrent unit (GRU) network with vibration and AE signals to diagnose the gear pitting faults. The proposed method provides more accurate diagnosis than using CNN only or GRU networks only. Jing et al. [17] developed a CNN variation that learns features directly from the frequency data of the vibration signals and validated the effectiveness of the method using data collected from a planetary gearbox test bench. Lee et al. [18] proposed a combination of CNN and GRU for anomaly detection. The deep learning method can effectively avoid the problem of shallow learning and manual extraction of features. The method has been widely used in the field of intelligent fault diagnosis due to its powerful feature learning and classification capability. Deep learning methods are playing an increasingly important role in gear pitting fault diagnosis.

In deep learning, the most widely used method is CNN. Two-dimensional (2D) CNN has made great progress in visual signal processing. However, in many industrial applications, most signals are one-dimensional (1D). Many researchers process 1D data by changing the dimension of the data into 2D. However, this may lead to worse performance than directly using 1D data because 1D data doesn't produce an image that humans can recognize. Instead of converting 1D data into 2D and processing it using 2D CNN, using 1D CNN may present as a good alternative. An accurate rotational machine diagnostic method based 1D CNN was proposed that works directly on the raw signals without any pre-processing [19]. Liu et al. [20] proposed a method combining 1D denoising convolution autoencoder and 1D CNN for fault diagnosis. Huang et al. [21] constructed a 1D CNN to use directly the raw signals for intelligent mechanical fault diagnosis and ensure the authenticity of signal samples. Because a 1D CNN has good interpretability, one can explain the convolutional feature extraction mechanism and analyze the convolution kernel's interoperability by its output in time domain and frequency domain. Waziralilah et al. [22] discussed the application of the CNN in fault diagnosis of rolling bearing bearings.

Techniques for data-driven model training are receiving more and more attention because of their good diagnostic performance and ease of implementation, such as residual learning, data enhancement techniques, transfer learning, batch standardization, dimensionality reduction, and other emerging deep learning training techniques have effectively eliminated the phenomenon of overfitting. Because of the vanishing gradient problem, simply increasing the number of layers in a network is often not a good solution [23]. The key to the design is to use convolutional residual networks to train extremely deep models [24]. Inspired by residual learning, Wand et al. [25] used a group of connected jumps to transfer multi-scale spatial features to subsequent high-level operations. Zhang et al. [26] proposed a residual learning algorithm. The proposed architecture significantly improved the flow of information throughout the network and was well suited for processing mechanical vibration signals with variable sequence lengths. Since accurately labeled data is often difficult to obtain in real applications, Li et al. [27] proposed a data enhancement technique to create additional valid samples for manual model

training. The proposed method seeks to achieve better diagnostic performance with a small original training dataset. Transfer learning is an approach to retrain complex neural networks for similar problems and save a lot of training time. Han et al. [28] proposed a transfer learning framework based on pre-trained CNN that uses knowledge learned from training data to facilitate the diagnosis of new but similar tasks. Instead of training the network from scratch, the pre-trained CNN architecture and weights are transferred to new tasks with appropriate fine-tuning. In general, a diagnostic model that uses data from finite conditions for training may not generalize well to what is not observed during training. To address this challenge, Han et al. [29] introduced resistance learning. They proposed a new deep anti-convolution neural network (DACNN). Principal component analysis (PCA) was used to reduce the dimensionality of the data matrix [30]. A 1×1 filter can significantly reduce the number of network parameters, thus speeding up the training and testing process [31]. Hou et al. [32] proposed to use multiple 1×1 filters to reduce the model size effectively. Tang et al. [33] also used 1×1 filter to compress the feature graphs of all convolutional layers except the top convolutional layer. These methods have played a significant role in the application of deep learning for gear pitting fault diagnosis.

Utilizing the advantages of 1D CNN for processing 1D data and the latest training techniques for deep learning, this paper proposes a novel method for gear pitting fault diagnosis. The method effectively reduces about 50% training parameters than traditional 1D CNN while ensuring a good diagnostic performance. The proposed method also shortens the training time. The most significant difference between the proposed method and the traditional 1D CNN is that the channel convolution is separated from the spatial convolution process in the proposed method. First, the spatial convolution is performed on the raw vibration signals, and then the channel convolution is extracted with point-by-point convolution. This would effectively reduce the number of model parameters of the 1D CNN and reduce the training time. A residual connection is integrated into the deep separable CNN so that the features obtained in the lower layers can be transmitted to the upper neural network layers. This thereby would effectively avoid the bottleneck problem of the features in upper layer. At the same time, grid search algorithm and random search algorithm are applied to adaptively select the hyperparameters for the best network performance.

The method proposed in this paper has the following significant advantages for gear pitting fault diagnosis:

(1) The proposed method can effectively detect the faults of different pitting degrees of gears under mixed conditions.
(2) The proposed method can reduce the model parameters by approximately 50% in comparison to the traditional 1D CNN, while maintaining excellent diagnostic performance.
(3) The proposed method uses grid search algorithm and the random search algorithm to optimize the hyperparameters of the model and avoid manual adjustment of the hyperparameters.
(4) The proposed method uses directly the raw vibration signals for training without pre-processing, reduces the manual operation workload, effectively and expands the scope of the model for gear pitting fault diagnosis.
(5) The proposed method passes the features through the residual connection and can effectively solve the representational bottleneck problem of the features in the model.

The rest of this paper is organized as follows. Section 2 describes the method proposed in this paper. It mainly introduces the method of adaptive 1D depth separable convolution. In this section, the differences between traditional CNN and the deep separable CNN are also emphasized. In Section 3, twenty experiments designed to validate the effectiveness of the proposed method for gear pitting fault diagnosis under mixed conditions are provided. Section 4 analyzes and discusses the results of the proposed method and other mainstream machine learning and deep learning methods under mixed conditions. Present research results and future research directions are summarized in Section 5.

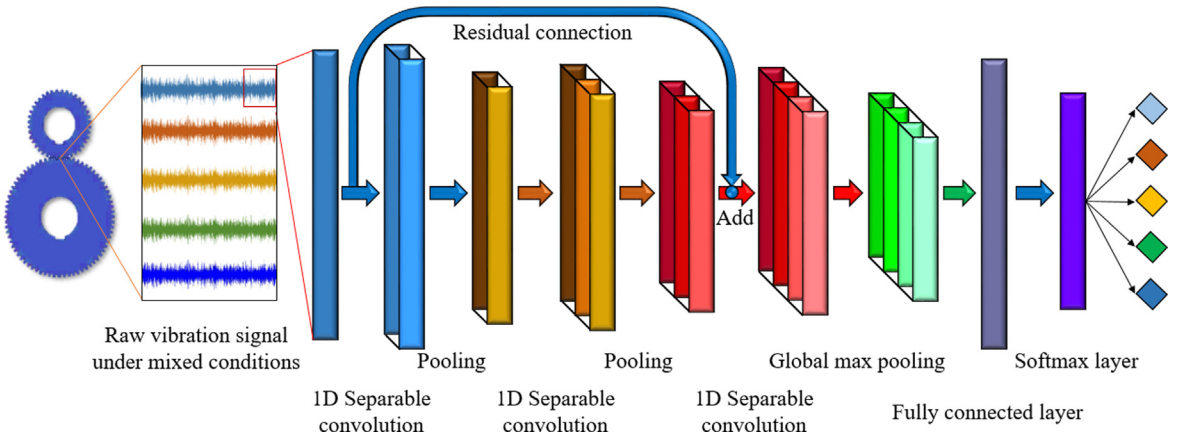


Fig. 1. The overall framework of the method proposed in this paper.

2. The method

The overall framework of the adaptive 1D separable convolution combined with residual connection proposed in this paper is shown in Fig. 1. The acceleration sensor collected the raw acceleration signal of the gears. Then inputting the raw vibration signal under the mixed operating conditions to the network. The raw vibration signal is first passed through stacked 1D separable convolution layers and pooling layers, then input to the fully connected layer through global max pooling. Finally, the diagnosis of gear pitting failure by the softmax classifier. In the stacked 1D separable convolution layers, the lower feature output is passed to the upper layer through the residual connection. Therefore, the feature bottleneck in the dimension reduction process is avoided. The optimal hyperparameters are adaptively searched by the parameter search algorithm to achieve an accurate diagnosis of gear pitting failure under the mixed operating conditions. The main reasons for proposing model based on 1D separable convolution with residual connection to detect gear fault in this paper are reflected in the following two aspects: dimensionality reduction and adding nonlinearity into the model. The 1×1 convolution adds non-linear activation into the learning representation of the previous layer to improve the representation ability of the network. A 1×1 convolution can realize cross-channel interaction and information integration while reducing the dimension and the number of convolution kernel channel [34]. It breaks the symmetry of the network and improves its representational ability. The main reason for adding residual connection is that the residual network can learn more features of the raw vibration signals. Hence, it helps to simplify the learning process and enhance the gradient propagation. The residual network can be connected by jumping layers to improve the flow for gradients, thus making it possible to train a large number of layers networks [35]. Residual connection can also enhance the generalization ability of the network [36].

2.1. 1D separable convolution

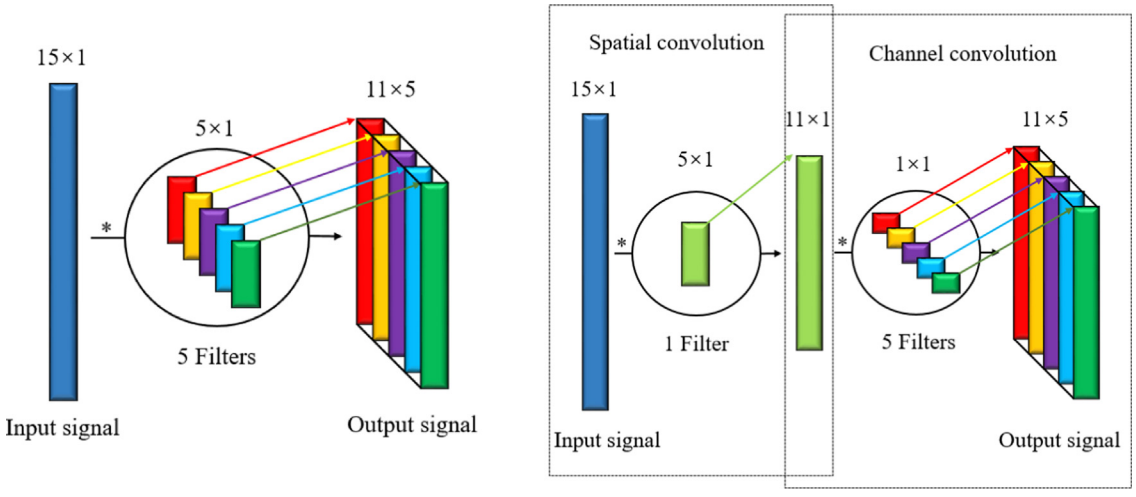
For the first time, the principle of the traditional convolutional neural network is briefly explained. The principle of a traditional convolutional neural network is shown in Fig. 2(a). A conventional convolution operation is a joint mapping of spatial convolution and channel convolution. Both the input variable X and the model parameter W_i are represented by a matrix, $\times$ represents a convolution operation, $b_i$is a bias, and σ is the activation function. The process for 1D CNN can be defined as the following equation:

$$Y_i = \sigma(X \times W_i + b_i) \tag{1}$$

Given an input signal sequence X_i, i = 1, . . ., n, and filters W_j, j = 1, . . ., m. The output of the convolution is defined as:

$$y_t = \sum_{k=1}^{m} W_j \times X_{i-j+1} \tag{2}$$

In the convolutional layer, each neuron in the L layer is connected to neurons in a part of the L-1 layer to form a local connection network. The convolutional layer requires an activation function f(x) for nonlinear feature mapping, and W^L is an m-dimensional filter that is the same for all neurons. $a^L_{(i+m-1):i} = [a^L_{(i+m-1)}, \cdots, a^L_{(i+m-1):i}]$, $b_i$is partial set the function, i = 1, 2, . . ., n. Then the output of the neuron of the L layer is defined as:



(a) Traditional 1D convolution (b) 1D separable convolution

Fig. 2. A comparison of traditional 1D CNN with 1D separable convolution.

$$a_i^L = f\left(\sum_{j=1}^{m} W_j^L \times a_{i-j+m}^{L-1} + b^L\right) = f\left(W^L \times a_{(i+m-1):i}^{L-1} + b_i\right) \tag{3}$$

For a feature map a_i obtained by the convolutional layer, it is divided into a plurality of regions R_k, $k = 1, \ldots, n$. Taking the maximum value of each region as:

$$pool_{max}(R_k) = \max_{i \epsilon R_k} a_i \tag{4}$$

Compared to traditional CNN, 1D separable convolution divides this process into two parts: spatial convolution and channel convolution. Firstly, the spatial convolution is performed on the raw gear pitting signal, and then the channel convolution is extracted by point-by-point convolution, which effectively reduces the number of model parameters of the 1D CNN and shortens the training time of the model. It can be seen from Fig. 2(b) that the use of 1D separable convolution reduces the required parameters compared to conventional convolution. It is crucial that the separable convolution changes the traditional operation of convolution into consideration of both the spatial convolution and channel convolution, and the convolution first considers spatial convolution and then the channel convolution. The separation of channels and channel is achieved.

The mathematical calculation is used to compare the number of parameters of 1D separable convolution with that of traditional 1D convolution. The calculation amount of the 1D standard convolution operation is:

$$D_K \times M \times N \times D_F \tag{5}$$

The calculation amount of the 1D separable convolution operation is:

$$D_K \times M \times D_F + M \times N \times D_F \tag{6}$$

In Eqs. (5) and (6), D_K is the size of the convolution kernel, M is the number of channels input, N is the number of convolution kernels, D_F is the size of the input data.

The ratio of the depth separable convolution calculation to the traditional convolution calculation is:

$$\frac{D_K \times M \times D_F + M \times N \times D_F}{D_K \times M \times N \times D_F} = \frac{1}{N} + \frac{1}{D_K} < 1 \tag{7}$$

It can be seen from Eq. (5)–(7) that the 1D separable convolution reduces a large number of calculations and improves the performance of the system. If the size of the convolution kernel is 3 and the number of convolution kernels is 5, According to Eq. (3), the 1D separation convolution can reduce the parameters to about 50% of the original parameters. Therefore, the 1D separable convolution model has been improved in speed.

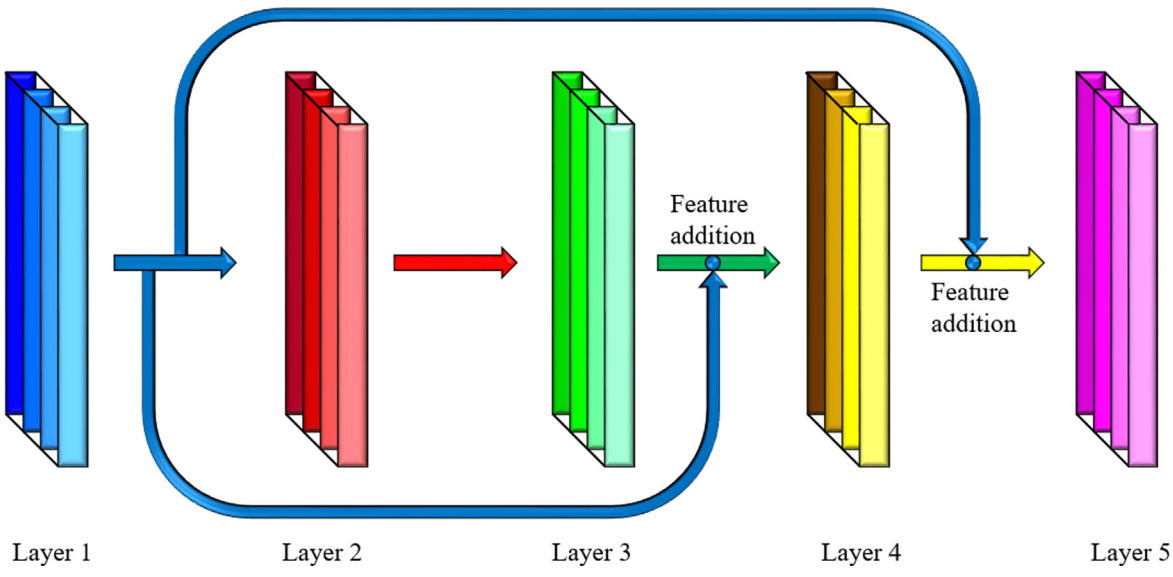


Fig. 3. The principle of residual connection.

2.2. Residual connection

The residual connection [37] allows the original feature information in the gear vibration signal to be obtained for any deeper layer. The basic structure of the residual connection is shown in Fig. 3. The calculation process of the residual unit is as follows:

$$y_l = h(x_l) + F(x_l, W_l) \tag{8}$$

where x_l represents the input of residual unit, $y_l$represents the output of the residual unit, $F(x_l, \mathbf{W}_l)$ represents nonlinear transformation layers of the neural network, $h(x_l)$ represents transfer function the bypass. In order to ensure the consistency of the dimensions, network usually requires dimensionality reduction operations, such as pooling operation.

2.3. Adaptive hyperparameter search

Usually, researchers use a large number of experimental comparisons and historical experience to determine appropriate hyperparameters for different known training samples. This method of repeated experiments and historical experience has great blindness and complicated process. It is not only difficult to determine hyperparameters but also computational expensive. The main idea of the grid search algorithm developed in the empirical method is to divide the space into the parameter spaces to be searched according to the set search step and to perform parameter group value for each node in the grid. Then, each parameter group is brought into the model and evaluated by the parameter evaluation method. Finally, the parameter group capable of achieving the optimal performance of the model is decided as the final parameters.

The grid search has many advantages that many other algorithms do not have. First of all, in the case that the number of parameters to be determined is small, the computational complexity of the grid search algorithm is often excellent, compared with many heuristic algorithms. Secondly, the parallelism of the grid search algorithm is high, because each a set of parameters are independent of each other and can be searched from multiple aspects simultaneously from the initial position within the determined interval. The determination of the number of subspaces is mainly based on the value range of the superparameter. For hidden units and layers as an example: Hidden units usually take an integer power of 2, such as 16, 32, 64, 128, 256, and 512 [38]. In this paper, these 6 values are used in the grid search. The range for the number of layers is affected by the over-fitting and under-fitting results. As a rule of thumb, values between 4 and 8 are usually sufficient [39].

The grid search step is shown in Fig. 4: for each grid of the parameter detection area, an evaluation function of the cross-validation model is performed. The optimal hyperparametric model is obtained by comparing all the mesh parameters. Grid search is mainly for the selection of discrete hyperparametric variables. The random search is primarily for the selection of continuous hyperparametric variables. In the random search, the initial values of the inputs are randomly generated in the given range. It usually includes interval boundaries. For learning rate as an example: learning rate is searched with the method of log scale search. The minimum value of learning rate is set as 0.00001 and set 10^a = 0.00001. In this case, a is calculated as -5. The maximum value of learning rate is set as 1 and set10^b = 1. In this case, b is calculated as 0. In interval $[-5, 0]$, a random number $r \in [-5, 0]$ is generated and one can set learning rate $\alpha = 10^r$.

Random search is a set of numerical optimization methods that do not require optimization of the gradient of the problem. So random search can be used for non-continuous or differentiable functions. This optimization method is also known as direct search, no derivative or black box method. The random search was first proposed by Rastrigin [40] in 1963, and he gave an early introduction to random search and conducted basic mathematical analysis. The way the random search works is to iteratively move to better locations in the search space, which are sampled from a hypersphere that surrounds the current location. In 2012, the random search was used for hyperparameter optimization by Bergstra et al. [41].

In Table 1, *Cost*() represents the diagnostic accuracy, *NumIterations* is the number of iterations, *SearchSpace* is the search space, and *Best* is the optimal value. The candidate within the number of iterations updates the optimal value if the cost of the current candidate is less than the cost of the current optimal value.

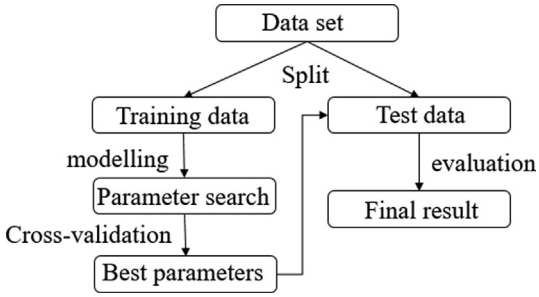


Fig. 4. Overview of the process of parameter selection and model evaluation with grid search.

Table 1
Pseudocode for the random search.

Pseudocode for Random Search

Input: *NumIterations*, *SearchSpace*
output: *Best*
Best←Φ
For *i* in *NumIterations*:
  i ← *RandonSolution* (*SearchSpace*)
  If *Cost* (*i*) < *Cost* (*Best*)
  Best ← *i*
  End
End
Return(*Best*)

3. Gear test experiments

In order to validate the effectiveness of the proposed method, this paper designs 20 sets of different degrees of gear pitting conditions. The acceleration sensor collects the raw gear vibration signals into an adaptive 1D separable convolution with the residual connection. The sampling rate was set as 10.24 kHz, the file extension was TXT, and the sampling time was 6 s. The acceleration sensor was mounted on the gearbox housing. The test rig is shown in Figs. 5 and 6. The experimental device consists mainly of two 45 kW Siemens motors and a gearbox and its cooling and lubrication device. The main parameters of the gearbox are shown in Table 2.

Different degrees of gear pitting are shown in Fig. 7 and Table 3. Health state 1 is normal gear. The intermediate teeth of Health state 2 and 3 have about 10% and 30% pitting, respectively. Health state 4 and 5 have 50% pitting for the middle teeth, and adjacent gears also have varying degrees of pitting.

All experiments in this paper were carried out with a load of 50 Nm and a rotational speed of 400, 500, 600, and 700 rpm, respectively. The raw vibration signal in the z-axis direction of the gear is shown in Fig. 8. Each row represents a vibration signal of the different gears pitting at the same speed, and each column represents the same gear pitting health state and different rotational speeds. It can be seen from the figure that the raw vibration signal of the gear in Health state 4 has a large amplitude, which may be caused by the mutual vibration between the two gears. The gear signals of the other four cases are not too obvious.

A major challenge for machine learning is that if poor data is entered into the system, the performance of the system will gradually decline. Before providing data to a machine learning model, first, check for the existence of data outliers. If there is any need to perform a data cleanup operation, remove the outliers in advance. In order to understand the raw vibration signal of the gear, the range and the number of amplitudes are visually displayed through the histogram. Fig. 9 is a histogram of the frequencies of four different loads, each of which takes 5000 points to plot the frequency histogram of the amplitude. It can be seen from the histogram that the four loads all show a normal distribution, and some of them are farther away from the zero point, especially in the case of 700r/min, the normal gear also has some amplitudes between 1 and 2. point. The data collected is not particularly perfect. We use these raw data for fault detection of gear pitting.

The data of the training set, the verification set, and the test set are not duplicated. The amount of data in each sample signal is 1536. The number of samples for each speed training set, verification set, and test set is 1600, 300, and 100, respectively. So the sample size input to the network is 6400, 1200, and 400, respectively. The number of channels of the separable convolution network was 32, 64, 128, and 128, respectively. Strides were 1, pooling size was 2, the batch size was 512,

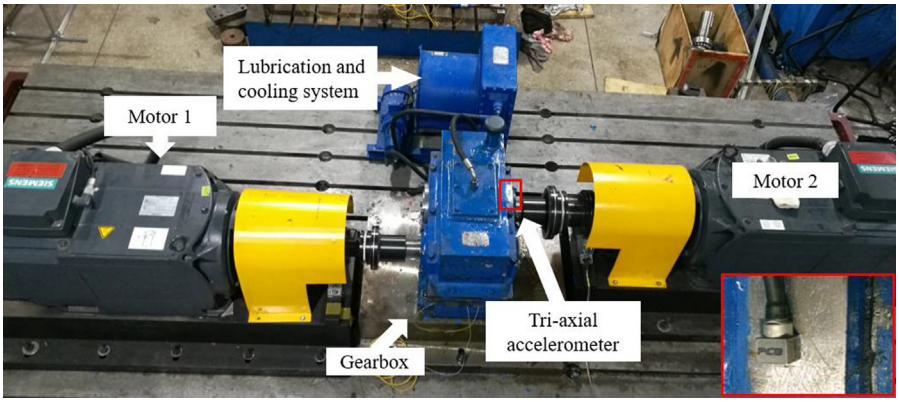


Fig. 5. Photo of the experimental equipment.

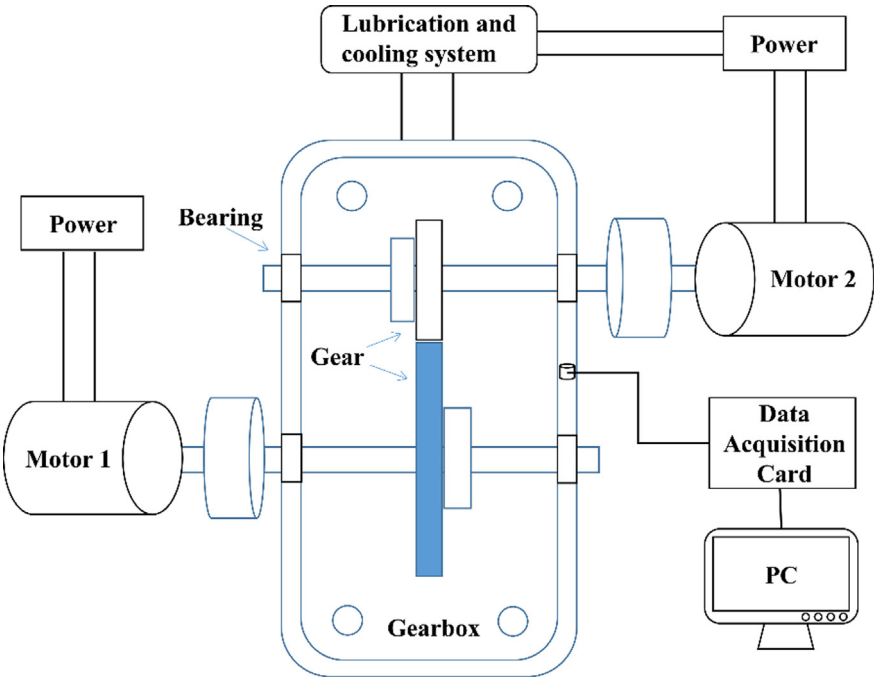


Fig. 6. Experimental test device schematic (blue color represents changed components).

Table 2
Main parameters of the gearbox.

Gear parameter	Driving gear	Driven gear
Tooth number	40	72
Module (mm)	3	3
Base circle diameter (mm)	112.763	202.974
Pitch diameter (mm)	120	216
Pressure angle (°)	20	20
Addendum coefficient	1	1
Coefficient of top clearance	0.25	0.25
Diametric pitch	8.4667	8.4667
Engaged angle (°)	19.7828	19.7828
Circular pitch (mm)	9.42478	9.42478
Addendum (mm)	4.5	3.588
Dedendum (mm)	2.25	3.162
Addendum modification coefficient	0.5	0.196
Addendum modification (mm)	1.5	0.588
Fillet radius (mm)	0.9	0.9
Tooth thickness (mm)	5.8043	5.1404
Tooth width (mm)	85	85
Theoretical center distance (mm)	168	168
Actual center distance (mm)	170.002	170.002

activation was relu, and the padding operation is to keep the size of the data the same as before and after the convolution. The loss used categorical cross entropy, the optimizer used RMSprop, the original learning rate was 2e-5, and the NVIDIA 1080Ti was used for training the model. Table 4 shows the method proposed in this paper, the connection mode of each corresponding layer, the size and the number of corresponding parameters.

4. Results and discussions

4.1. Binary classification

The method proposed in this paper and other machine learning methods are used to classify the gear pitting faults with mixed operating health state. Selected Health state 1 and Health state 3 to binary classification. The corresponding gears are

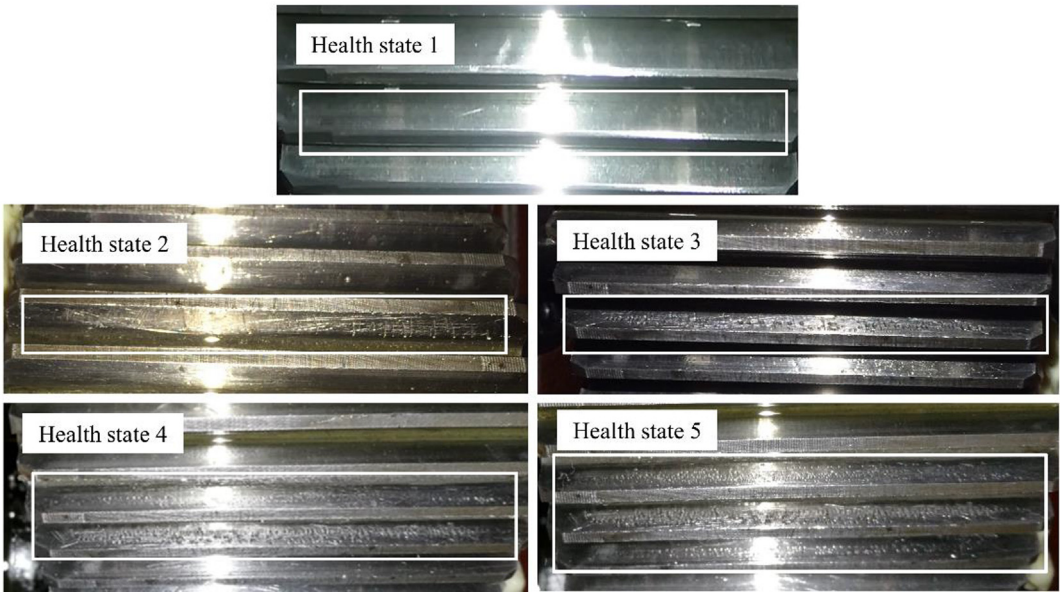


Fig. 7. Different degrees of gear pitting.

Table 3
The approximate percentage of wearing area.

Pitting percentage	Upper tooth	Middle tooth	Lower tooth
Health state 1	Normal	Normal	Normal
Health state 2	Normal	10%	Normal
Health state 3	Normal	30%	Normal
Health state 4	10%	50%	Normal
Health state 5	30%	50%	10%

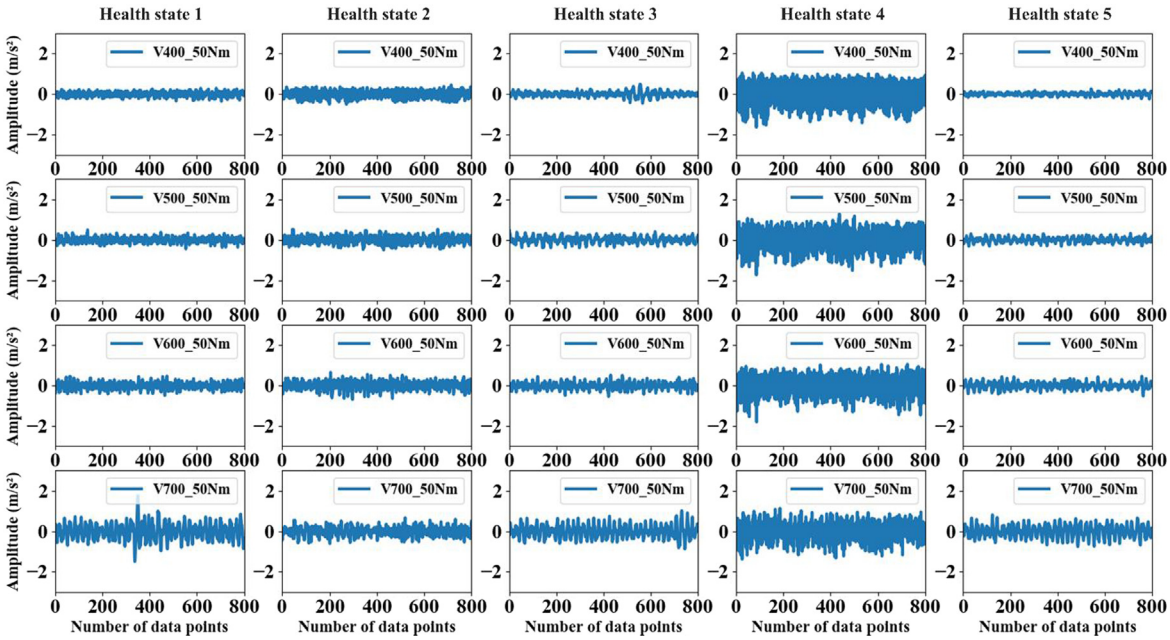


Fig. 8. Raw vibration signals of gears.

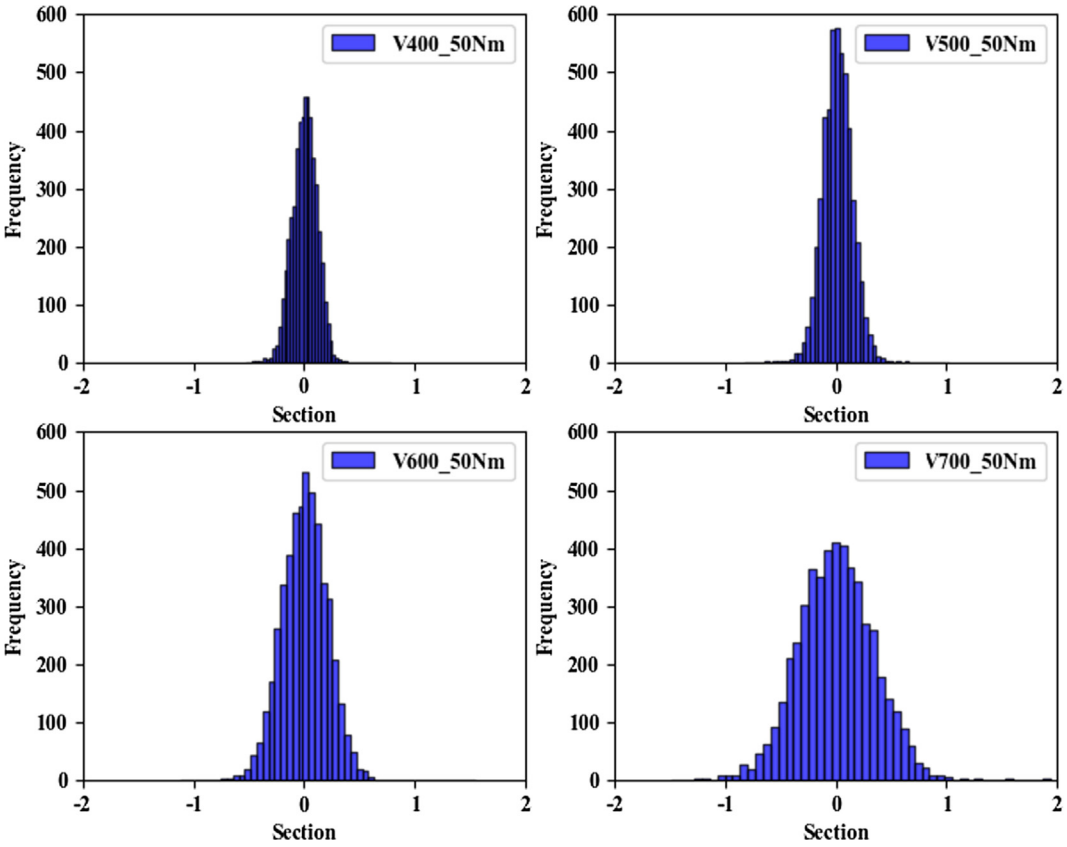


Fig. 9. The frequency histogram of the amplitude.

Table 4
The layers of the network and the main attributes.

Layer	Output Shape	Parameters	Connected to
InputLayer	(None, 1536, 1)	0	
Separable_Conv1D_1	(None, 1536, 32)	67	InputLayer
Max_Pooling1D_1	(None, 768, 32)	0	Separable_Conv1D_1
Separable_Conv1D_2	(None, 768, 64)	2208	Max_Pooling1D_1
Max_Pooling1D_2	(None, 384, 64)	0	Separable_Conv1D_2
Separable_Conv1D_3	(None, 384, 64)	131	InputLayer
ResidualLayer_1	(None, 384, 64)	0	Max_Pooling1D_2 Separable_Conv1D_3
Separable_Conv1D_4	(None, 384, 128)	8512	ResidualLayer_1
Separable_Conv1D_5	(None, 384, 128)	16,896	Separable_Conv1D_4
Separable_Conv1D_6	(None, 384, 128)	259	InputLayer
ResidualLayer_2	(None, 384, 128)	0	Separable_Conv1D_5 Separable_Conv1D_6
Global_Max_Pooling1D	(None, 128)	512	ResidualLayer_2
DenseLayer	(None, 100)	12,900	Global_Max_Pooling1D
SoftmaxLayer	(None, 5)	505	DenseLayer

the normal gears and pitting fault gears with 30% pitting of the intermediate teeth. Indicators for measuring the results of binary classifications in machine learning: accuracy and recall, which are:

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

where, TP is the number of true positives, and FP is the number of false positives. FN is the number of false negatives. The recall, also known as sensitivity or true positive rate (TPR), is the ratio of positive instances that the classifier correctly detects.

Unfortunately, we can't increase accuracy and reduce the recall rate at the same time, and vice versa. This is called the precision/recall rate trade-off. Therefore, we can easily combine the accuracy and recall rate into a single indicator called F_1 score. This is a perfect indicator when you need an easy way to compare two classifiers. The F_1 score is the harmonic mean of accuracy and recall. The normal average treats all values equally, while the harmonic mean gives a lower value a higher weight. Therefore, the classifier can get a higherF_1score only when the recall rate and accuracy are high.

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TF + \frac{FN+FP}{2}} \tag{11}$$

Fig. 10 is the curve of the accuracy and recall of the proposed method for gear pitting fault diagnosis. It can be seen from the figure that the accuracy of the proposed method has been relatively high, and the recall rate is also good, indicating the proposed method works well for the diagnosis of gear faults.

There is also a tool that is often used with binary classifiers, called the receiver operating characteristic curve (ROC). It is very similar to the accuracy/recall rate curve, but not the accuracy and recall rate, but the *true positive rate* and the *false positive rate* (FPR). FPR is a negative class instance ratio that is incorrectly divided into positive classes. It is equal to 1 minus the true negative rate (TNR), which is the negative class instance ratio that is correctly classified as a negative class, also known as *specificity*. Therefore, the ROC curve plots the relationship between sensitivity and 1-*specificity*.

Fig. 11 is a ROC curve for the diagnosis of gear pitting failure using the proposed method and other common machine learning and deep learning methods. It can be seen from the roc curve in the figure that the proposed method has the highest accuracy, which is better than 1D CNN, and the long-short-term memory unit (LSTM) and GRU also have excellent performance. The classification effect of k-nearest neighbors in machine learning is also good.

The advantage of using a random search algorithm is that it is necessary to set the corresponding code computer to run without the need to modify the parameters over and over again manually. Taking the random forest and 1D CNN as an example, the random search algorithm and grid search algorithm for hyperparameter selection used in this paper are introduced. The most important of the decision tree parameters include the max *features* and the *nestimators*. *Max features* are the maximum number of features that a random forest allows for a single decision tree to use. *Nestimators* is the number of subtrees you want to create. In this experiment, max *features* is set to an integer with a minimum value of 1 and a value of 100. *Nestimators* is set to an integer with a minimum value of 1 and a value of 300. After passing the algorithm on 100 steps of data and 5-fold cross-validation, the final optimal value is *nestimators* = 88 and *maxfeatures* = 75. The learning rate of the neural network is essential for 1D CNN. In this paper, the hyperparameter of learning rate is searched by the grid search method, and the experience of training a neural network is measured every 0.1 degrees, and finally, 1.5 is selected. The ultimate optimal learning rate.

Fig. 12 is the accuracy of the different methods for the diagnosis of gear pitting with mixed operating conditions. It can be seen from the figure that the proposed method, 1D CNN, LSTM, GRU, and the k-nearest neighbor method all achieve an accuracy of more than 80%. Generally speaking, we use two kinds of conditions to carry out different classification model experiments to select better models and use this model for multi-category classifications, which can save a lot of time. For models with a lower classification accuracy of less than 80%, multiple classifications are no longer used. In the following, the method of classification results exceeding 80% will be used to diagnose gear pitting in multi-category classifications.

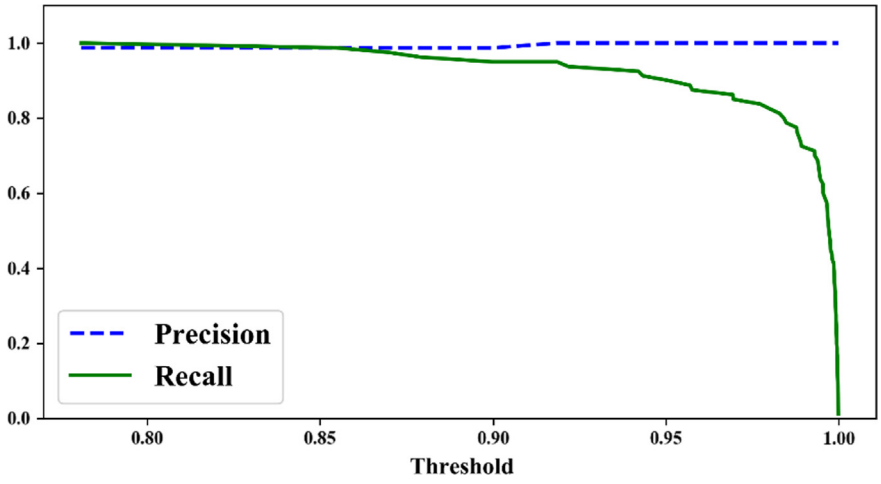


Fig. 10. The accuracy and recall rate curves of gear pitting classification.

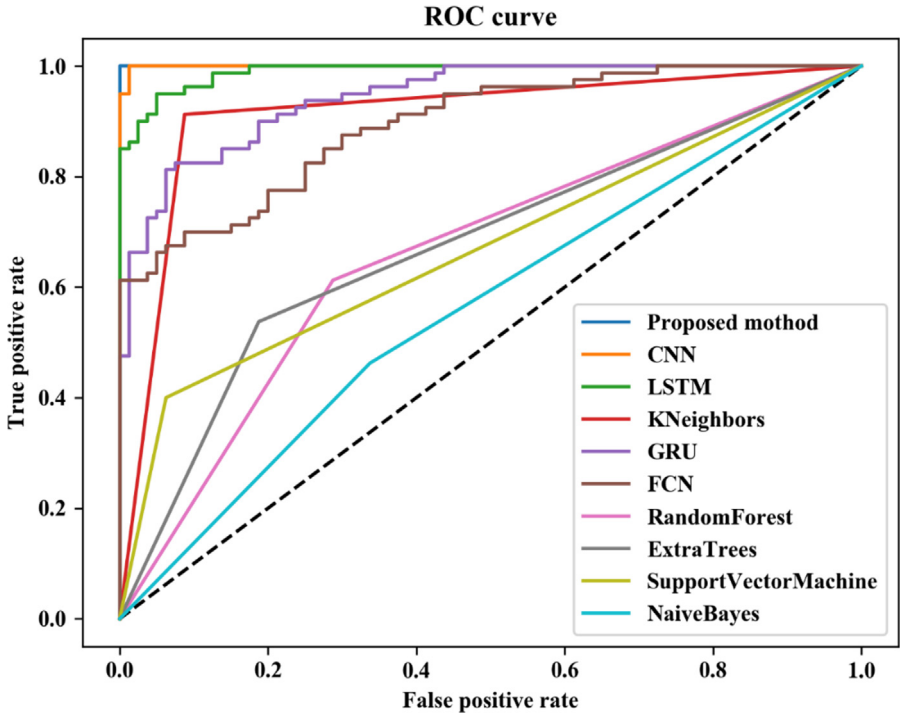


Fig. 11. ROC curve.

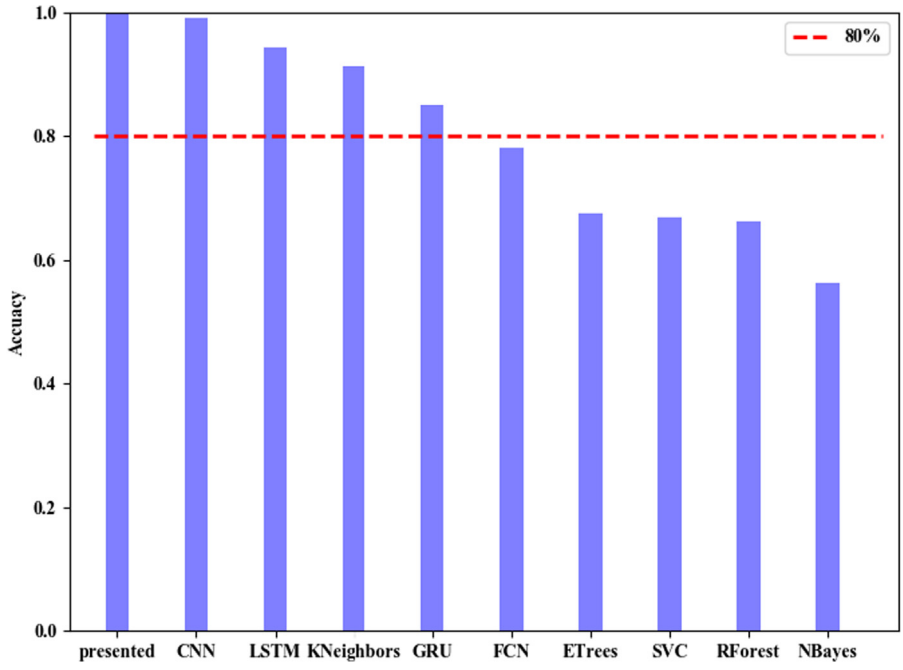


Fig. 12. Accuracy of different methods for the diagnosis of gear pitting.

4.2. Multi-category classification

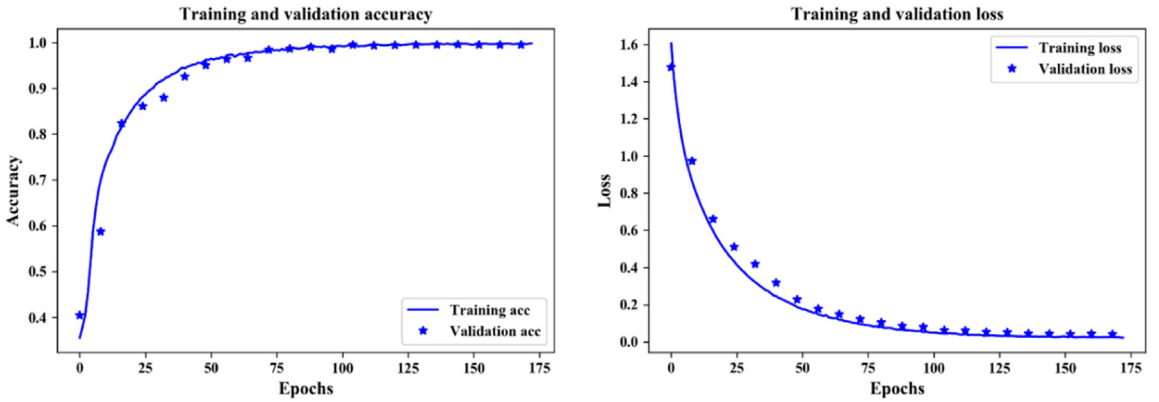
For the multi-category classification diagnosis of gear pitting failure with mixed operating conditions, the main difference is that the activation function is changed from sigmoid to softmax. The predicted value of softmax is:

$$\hat{p}_k = \sigma(\mathbf{s}(\mathbf{x}))_k = \frac{\exp(s_k(\mathbf{x}))}{\sum_{j=1}^{K} \exp(s_j(\mathbf{x}))} \tag{12}$$

where, K is the number of categories, $\mathbf{s}(\mathbf{x})$ is the vector of the scores of each category of instance $\mathbf{x}$, $\sigma(\mathbf{s}(\mathbf{x}))_k$ is the probability that instance $\mathbf{x}$ belongs to category k under a given category score.

The accuracy and loss curves of training and verification are shown in Fig. 13. It can be seen from the figure that the method proposed in this paper has better processing results for the original gear vibration signal. At the same time, there is no visible overfitting phenomenon.

The confusion matrix is shown in Fig. 14. As can be seen from the results, the proposed method has a good effect on the raw vibration signal of the gear with mixed operating conditions. The test set seems to have an excellent effect on the gear pitting recognition of different speed conditions. Only one of the 400 test samples had a misjudgment.



a) Accuracy curves. b) Loss curves.

Fig. 13. The accuracy and loss curves of training data and verification data.

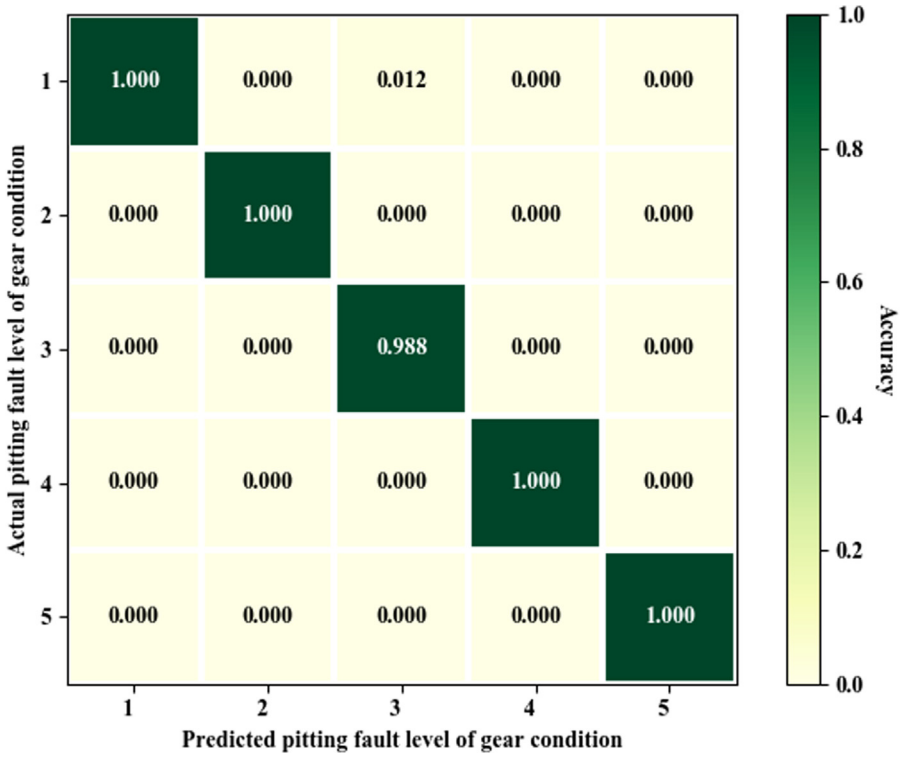


Fig. 14. The confusion matrix for the testing set.

Table 5
The accuracy comparison between the proposed method in this paper and the other methods.

Accuracy	Training set	Testing set
Proposed method	0.9981	0.9975
Traditional CNN	0.9948	0.9950
LSTM	0.9996	0.9950
KNeighbors	1.0000	0.6975
GRU	0.9996	0.9950

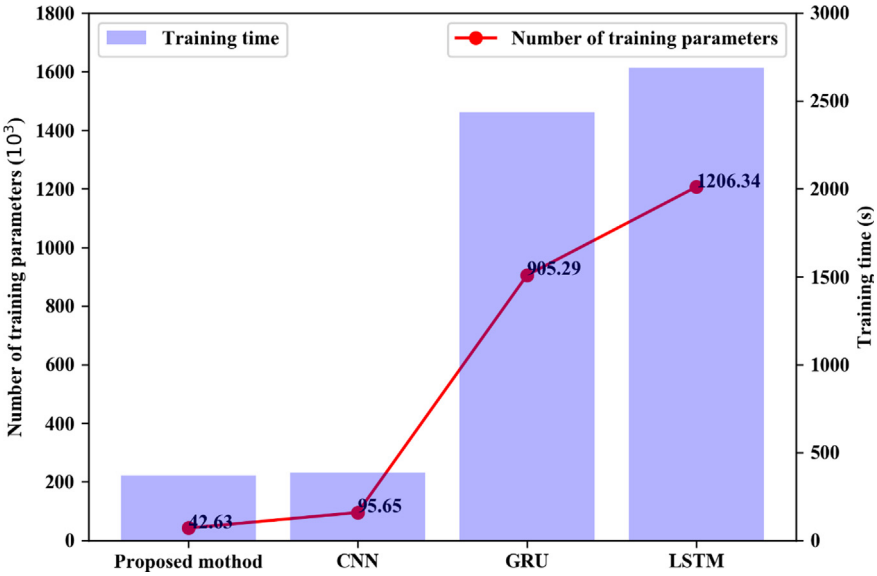


Fig. 15. The number of training parameters and the training time of the four methods.

Compared with the traditional 1D convolutional neural network, the adaptive 1D separable convolution with residual connection proposed in this paper can not only reduce the parameters of the network, but also the classification result of the gears with different pitting conditions with mixed operating conditions is as better as the other art of states methods. The structure of the comparison network is the same as that of the method proposed in this paper. As shown in Table 5, the test accuracy of the proposed method reaches 99.75%. It can be seen from Table 5 that the method of deep learning is better for multi-fault diagnosis of gear pitting with mixed operating conditions. The method proposed in this paper is slightly better than other deep learning methods on the testing set.

The number of training parameters and the training time of the four best methods in Table 5 was compared. It can be seen from Fig. 15 that the training parameters of the proposed adaptive 1D separable convolution with a residual connection method are only about 50% less than those of the traditional 1D CNN. Training parameters of the proposed method are far below the LSTM and GRU methods. For the training time, the proposed method is the shortest than the others. The method proposed in this paper can reduce the number of training parameters and reduce the training time based on excellent diagnostic performance.

In the future, the author will do some research on the multi-label classification of gear faults, and detect the number of pitting teeth under mixed conditions, not only the one kind category of gear pitting faults. The authors would also compare it with other up-to-date large networks.

5. Conclusions

Gear pitting fault diagnosis with mixed operating conditions is important. Compared with the traditional convolutional neural network, an adaptive 1D separable convolution with a residual connection network was proposed in this paper. It can greatly reduce the number of network parameters and can extract the features from the raw vibration signals more effectively and efficiently. Moreover, the method proposed in this paper applies the search algorithm to select better hyperparameters of the model. Therefore, better pitting fault diagnosis results can be obtained under mixed operating conditions. In comparison with other machine learning models, the proposed method obtained a gear pitting fault diagnosis accuracy of 99.75% at different speeds and avoided the over-fitting effectively with fewer model parameters.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymssp.2020.106740.

References

[1] Y.Z. Qu, M. He, J. Deutsch, D. He, Detection of pitting in gears using a deep sparse autoencoder, Appl. Sci. (5) (2017) 7.
[2] M. Cerrada, G. Zurita, D. Cabrera, R.V. Sanchez, M. Artes, C. Li, Fault diagnosis in spur gears based on genetic algorithm and random forest, Mech. Syst. Signal. Process. 70–71 (2016) 87–103.
[3] H. Ma, R. Song, X. Pang, B. Wen, Time-varying mesh stiffness calculation of cracked spur gears, Eng. Fail. Anal. 44 (2014) 179–194.
[4] H. Ma, J. Zeng, R. Feng, X. Pang, Q. Wang, B. Wen, Review on dynamics of cracked gear systems, Eng. Fail. Anal. 55 (2015) 224–245.
[5] G. Yu, Z. Wang, P. Zhao, Multisynchrosqueezing transform, IEEE Trans. Ind. Electron. 66 (7) (2018) 5441–5455.
[6] K. Yu, T.R. Lin, H. Ma, H. Li, J. Zeng, A combined polynomial chirplet transform and synchroextracting technique for analyzing nonstationary signals of rotating machinery, IEEE Trans. Instrum. Meas. (2019).
[7] Z.P. Feng, M.J. Zuo, J. Qu, T. Tian, Z.L. Liu, Joint amplitude and frequency demodulation analysis based on local mean decomposition for fault diagnosis of planetary gearboxes, Mech. Syst. Signal Process. 40 (1) (2013) 56–75.
[8] Y.G. Lei, M.J. Zuo, Gear crack level identification based on weighted K nearest neighbor classification algorithm, Mech. Syst. Signal. Process. 23 (5) (2009) 1535–1547.
[9] X.H. Liang, M.J. Zuo, L.B. Liu, A windowing and mapping strategy for gear tooth fault detection of a planetary gearbox, Mech. Syst. Signal. Process. 80 (2016) 445–459.
[10] B. Samanta, Gear fault detection using artificial neural networks and support vector machines with genetic algorithms, Mech. Syst. Signal. Process. 18 (3) (2004) 625–644.
[11] C. Li, R.V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, R.E. Vasquez, Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals, Mech. Syst. Signal. Process. 76–77 (2016) 283–293.
[12] H.T. Zhou, J. Chen, G.M. Dong, R. Wang, Detection and diagnosis of bearing faults using shift-invariant dictionary learning and hidden Markov model, Mech. Syst. Signal. Process. 72–73 (2016) 65–79.
[13] X.Y. Li, J.L. Li, D. He, Y.Z. Qu, Semi-supervised gear fault diagnosis using raw vibration signal based on deep learning, Chin. J. Aeronaut. (2019).
[14] X.Y. Li, J.L. Li, D. He, Y.Z. Qu, Gear pitting fault diagnosis using raw acoustic emission signal based on deep learning, Maintenance Reliab. 21 (3) (2019) 403–410.
[15] F. Jia, Y.G. Lei, J. Lin, X. Zhou, N. Lu, Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, Mech. Syst. Signal. Process. 72–73 (2016) 303–315.
[16] X.Y. Li, J.L. Li, Y.Z. Qu, He D. Gear, Pitting fault diagnosis using integrated CNN and GRU network with both vibration and acoustic emission signals, Appl. Sci. Basel (4) (2019) 9.
[17] L.Y. Jing, M. Zhao, P. Li, X.Q. Xu, A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox, Measurement 111 (2017) 1–10.
[18] K. Lee, J.-K. Kim, J. Kim, K. Hur, H. Kim, CNN and GRU combination scheme for Bearing Anomaly Detection in Rotating Machinery Health Monitoring, in: 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), 2018, pp. 102–105.
[19] X. Liu, Q. Zhou, H. Shen, Real-time fault diagnosis of rotating machinery using 1-D convolutional neural network, in: 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2019, pp. 104–108.
[20] X.C. Liu, Q.C. Zhou, J. Zhao, H.H. Shen, X.L. Xiong, Fault diagnosis of rotating machinery under noisy environment conditions based on a 1-D convolutional autoencoder and 1-D convolutional neural network, Sensors-Basel (4) (2019) 19.
[21] S. Huang, J. Tang, J. Dai, Y. Wang, Signal status recognition based on 1DCNN and its feature extraction mechanism analysis, Sensors-Basel 19 (9) (2019) 2018.
[22] N.F. Waziralilah, A. Abu, M. Lim, L.K. Quen, A. Elfakharany, A review on convolutional neural network in bearing fault diagnosis, MATEC Web Conf. (2019) 06002.
[23] L. Zhao, X. Qiu, Q. Zhang, X. Huang, Sequence labeling with deep gated dual path CNN, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (12) (2019) 2326–2335.
[24] A. Elboushaki, R. Hannane, K. Afdel, L. Koutti, MultiD-CNN: a multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences, Expert Syst. Appl. 139 (2020) 112829.
[25] M. Wang, S. Hou, H. Li, F. Li, Generative image deblurring based on multi-scaled residual adversary network driven by composed prior-posterior loss, J. Visual Commun. Image Representation 65 (2019) 102648.
[26] W. Zhang, X. Li, Q. Ding, Deep residual learning-based fault diagnosis method for rotating machinery, Isa T (2018).
[27] X. Li, W. Zhang, Q. Ding, J.-Q. Sun, Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation, J. Intell. Manuf. (2018) 1–20.
[28] T. Han, C. Liu, W. Yang, D. Jiang, Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions, Isa T (2019).
[29] T. Han, C. Liu, W.G. Yang, D.X. Jiang, A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults, Knowl.-Based Syst. 165 (2019) 474–487.
[30] T.H. Loutas, D. Roulias, E. Pauly, V. Kostopoulos, The combined use of vibration, acoustic emission and oil debris on-line monitoring towards a more effective condition monitoring of rotating machinery, Mech. Syst. Signal. Process. 25 (4) (2011) 1339–1352.
[31] H. Gao, Y. Yang, C. Li, X. Zhang, J. Zhao, D. Yao, Convolutional neural network for spectral–spatial classification of hyperspectral images, Neural Comput. Appl. 31 (12) (2019) 8997–9012.
[32] Y. Hou, H. Luo, W. Zhao, X. Zhang, J. Wang, J. Peng, Multilayer feature descriptors fusion CNN models for fine-grained visual recognition, Comput. Anim. Virtual Worlds 30 (3–4) (2019) e1897.
[33] G. Tang, R. Liang, Y. Xie, Y. Bao, S. Wang, Improved convolutional neural networks for acoustic event classification, Multimedia Tools Appl. 78 (12) (2019) 15801–15816.
[34] W. Shang, K. Sohn, D. Almeida, H. Lee, Understanding and improving convolutional neural networks via concatenated rectified linear units, in: International Conference on Machine Learning, 2016, pp. 2217–2225.
[35] A.E. Orhan, X. Pitkow, Skip connections eliminate singularities, International Conference on Learning Representations, 2018.
[36] A. Veit, M.J. Wilber, S. Belongie, Residual networks behave like ensembles of relatively shallow networks, Adv. Neural Inf. Process. Syst. (2016) 550–558.
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[38] L. Wen, X. Li, L. Gao, Y. Zhang, A new convolutional neural network-based data-driven fault diagnosis method, IEEE Trans. Indus. Electron. 65 (7) (2017) 5990–5998.

[39] B. Zhang, W. Li, X.L. Li, S.K. Ng, Intelligent fault diagnosis under varying working conditions based on domain adaptive convolutional neural networks, IEEE Access 6 (2018) 66367–66384.
[40] L. Rastrigin, The convergence of the random search method in the extremal control of a many parameter system, Autom. Remote Control 24 (1963) 1337–1342.
[41] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.