

ROLLING BEARING FAULT DIAGNOSIS AND PREDICTION BASED ON VMD-CWT AND MOBILENET

Jing Zhu, Aidong Deng, Shuo Xue, Xue Ding and Shun Zhang

School of energy and environment, southeast university, Nanjing , China

ABSTRACT

When deep learning is used for rolling bearing fault diagnosis, there are problems of high model complexity, time-consuming, and large memory. In order to solve this problem. This paper presents an intelligent diagnosis method of rolling bearings based on VMD-CWT feature extraction and MobileNet, VMD is used to extract the signal features, and then wavelet transform is used to extract the time-frequency features. After the image is enhanced, the MobileNet network is trained. In order to accelerate the convergence speed, this paper adds transfer learning in the network training process, and migrates the weights of the first several layers pretrained to the corresponding network. Experimental results based on bearing fault data sets show that after adopting VMD-CWT, the accuracy of mobilenet increased from 68.7% to 94%, and its network parameters were reduced by an order of magnitude compared with CNN.

KEYWORDS

Mobilenet, Variational modal decomposition, Continuous wavelet transform, Rolling bearing.

1. INTRODUCTION

Rolling bearing is an important mechanical device, which has important practical significance for social and economic development[1]. However, due to the harsh operating environment and the high incidence of rolling bearing failures, rolling bearing failures often cause huge casualties and economic losses. Using machine learning and deep learning methods to carry out abnormal detection and research on rolling bearings to achieve intelligent fault diagnosis is of important practical significance for timely detection of faults, early warning and predictive maintenance, safe operation of units, improvement of unit operation efficiency, and avoidance of accidents [2-4].

V. Purashotham[5] presents a new method for detecting localized bearing defects based on wavelet transform. Bearing race faults have been detected by using discrete wavelet transform (DWT). Vibration signals from ball bearings having single and multiple point defects on inner race, outer race, ball fault and combination of these faults have been considered for analysis. Liu[6]proposes a new feature fusion method to extract new features using kernel joint approximate diagonalization of eigen-matrices (KJADE). Hoang[7] proposed a bearing fault diagnosis method based on deep convolutional neural network structure. Using vibration signal as input data directly, it has high accuracy and robustness in noise environment. Although machine learning algorithms could identify faults based on extracted features automatically, the shallow structures restrict its ability to learn more abstract and discriminative information from the input automatically.

Convolutional Neural Network has been widely used in the field of computer vision and industry for its excellent image classification effect. However, with the improvement of model classification accuracy, the depth and complexity of the model are increasing[8]. Taking the Deep Residual Network (RESNET) proposed at the end of 2017 as an example, The model size

of Resnet50 is 98MB, the number of model layers is 152 layers, the number of parameters is 25636712, and the top-5 Accuracy is 0.921. The model size of Inceptionresnetv2 is 215MB, the number of layers is 572, the number of parameters is 55873736, and the top-5 Accuracy is 0.953. A small increase in model precision brings about a huge increase in model parameters[9].

However, in real application scenarios, especially in the transmission system of a wind turbine, if a mechanical failure is not handled in time, it will bring serious consequences.

Fault diagnosis requires strong real-time, low latency, and fast response. The complexity of excessive model would lead to prolonged calculation time. In case of equipment failure, saving time can avoid personnel injuries and economic losses caused by larger accidents. Second, the lightweight model would reduce memory, can save server memory, reduce server size, and even the equipment of processing could evolve towards mobile or embedded devices[10].

Therefore, it is of great research significance to apply the small but efficient CNN model to the fault diagnosis of rolling bearings of wind turbines. At present, the main research direction is to train the model, and then compress the trained complex model. Or it could conduct the research in the process of model design and design smaller models for training. The aim of this method is to reduce the size of the model while improving the accuracy of the model[11]. The research idea of MobileNet is the latter, which is to build a smaller model. The size and parameters of the model are greatly reduced while the model precision is slightly reduced, the enhanced images are classified based on the MobileNet network to achieve the purpose of fault diagnosis.

At the same time, due to the high ambient noise of rolling bearings in engineering environment, this paper used the Variational modal decomposition(VMD) to denoise the signal firstly, then uses the continuous wavelet transform(CWT) to draw the time-frequency image of the signal, finally, the MobileNet network is used to classify the enhanced image.

2. MATERIALS

2.1. Continuous wavelet transform

In the field of fault diagnosis, many applications are based on feature extraction, When the unit has a fault or abnormal state, the signal is often accompanied by corresponding feature components. Therefore, the detection of these components in the signal has become an important content in fault diagnosis. For example, in the fault diagnosis of rolling bearings and gears, the appearance of periodic pulse components often indicates the occurrence of faults. Wavelet transform is the convolution operation between the original signal and the wavelet function, which actually measures the similarity degree between the signal and the wavelet function [12]. In this way, by selecting different wavelet basis functions, the content of the components that are close to the wavelet shape in the signal could be detected, which can be used to detect the characteristic components in the signal.

Expand any function $f(t)$ in $L^2(\mathbb{R})$ space under wavelet basis, and call it the continuous wavelet transform(CWT) of function $f(t)$. The wavelet transform adjusts the window size according to the frequency automatically, and could perform multi-resolution analysis. Through the wavelet time-frequency analysis, the frequency transformation law of the signal with time is obtained, which reflects the difference between the rolling bearing faults. In this experiment, CWT is used to generate a two-dimensional matrix of wavelet coefficients, and then the mobilenet is used to automatically extract the time-frequency characteristics of vibration signals. CWT have very loose requirements for wavelet basis functions, and the following conditions are sufficient:

$$C_{\varphi} = \int_{-\infty}^{+\infty} \left| \frac{\varphi(\omega)^2}{\omega} \right| d\omega < \infty \quad (1)$$

Where, ω —Angular frequency

$\varphi(\omega)$ —Fourier transform of wavelet function $\varphi(t)$

The definition of CWT is as follows:

$$W(a,b) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt \quad (2)$$

Where, a , b —The two parameters of wavelet transform are the expansion factor and the size factor. For example, when $b=5$ and $a=2$ are given, the wavelet basis function is shifted by 5 units and reduced by 2 times.

2.2. Variational modal decomposition

The VMD decomposes the input signal x into a specified number (K) of quasi-orthogonal band-limited intrinsic mode functions (BLIMFs) μ_k with unknown but separable spectral bands. The fundamental principle of VMD can be expressed as the solving of a constrained variational problem[13]:

$$\min_{\{u_{k_1}\}, \{\omega_{k_1}\}} \left\{ \sum_{k_1}^K \left\| \partial_t \left[\delta(t) + \frac{j}{\pi t} \right] * \mu_{k_1}(t) e^{-j\omega_{k_1} t} \right\|_2^2 \right\} \quad (3)$$

$$s.t. \quad \sum_{k_1}^K u_{k_1} = f$$

Where f is the original signal, μ_{k_1} is the k_1 th intrinsic mode function(IMF) component and $\delta(t)$ denotes pulse signal; j is the imaginary unit, ω_{k_1} denotes the center frequency of the k_1 th IMF component; $\|\cdot\|_2$ is 2-norm.

The penalty parameter α and LaGrangian multiplier $\lambda(t)$ are introduced to solve the above-constrained issue:

$$L(\{u_{k_1}\}, \{\omega_{k_1}\}, \lambda(t)) = \alpha \sum_{k_1=1}^K \left\| \partial_t \left[\delta(t) + \frac{j}{\pi t} \right] * u_{k_1}(t) e^{-j\omega_{k_1} t} \right\|_2^2 + \left\| f(t) - \sum_{k_1} u_{k_1}(t) \right\|_2^2$$

$$+ \langle \lambda(t), f(t) - \sum_{k_1=1}^K u_{k_1}(t) \rangle \quad (4)$$

IMFs u_{k_1} and their corresponding center frequencies ω_{k_1} , the LaGrangian multiplier λ , are subsequently updated as:

$$\hat{u}_{k_1}^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k_1} \hat{u}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_{k_1})^2} \quad (5)$$

$$\omega_{k_1}^{n+1} = \frac{\int_0^{\infty} \omega \left| \hat{u}_{k_1}(\omega) \right|^2 d\omega}{\int_0^{\infty} \left| \hat{u}_{k_1}(\omega) \right|^2 d\omega} \quad (6)$$

$$\lambda^{n+1}(\omega) = \lambda^n + \mathcal{G}[f(\omega) - \sum_{k_1=1}^K u_{k_1}^{n+1}] \quad (7)$$

The above update process is implemented repeatedly until the following convergence criterion is satisfied:

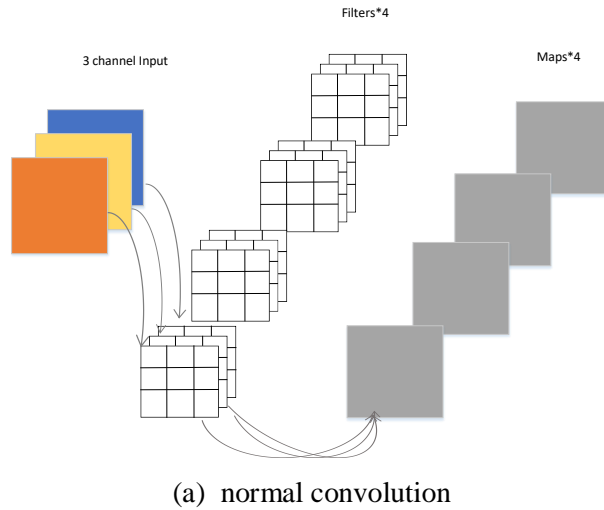
$$\sum_{k_1=1}^K \frac{\|u_{k_1}^{n+1} - u_{k_1}^n\|_2^2}{\|u_{k_1}^n\|_2^2} < \varepsilon \quad (8)$$

Where ε is set as 10^{-6} .

2.3. Mobilenet

Conventional Convolutional Neural Network has a large memory requirement, which makes it impossible to run on mobile devices and embedded devices, MobileNet is a lightweight CNN network proposed by Google. Compared with the traditional convolutional neural network, it greatly reduces the model parameters and calculation amount under the premise of slightly lower accuracy. The core of the model is Depthwise Convolution (DSC), which decomposes standard Convolution into depthwise(DW) Convolution and pointwise(PW) Convolution[14].

It can be seen from Figure 1 that in traditional convolution, channel of convolution kernel is equal to input eigenmatrix channel, and output eigenmatrix channel is equal to the number of convolution kernel. And in DW convolution, channel of the convolution kernel is equal to one, the input eigenmatrix channel is equal to the number of convolution cores is equal to the output eigenmatrix channel.



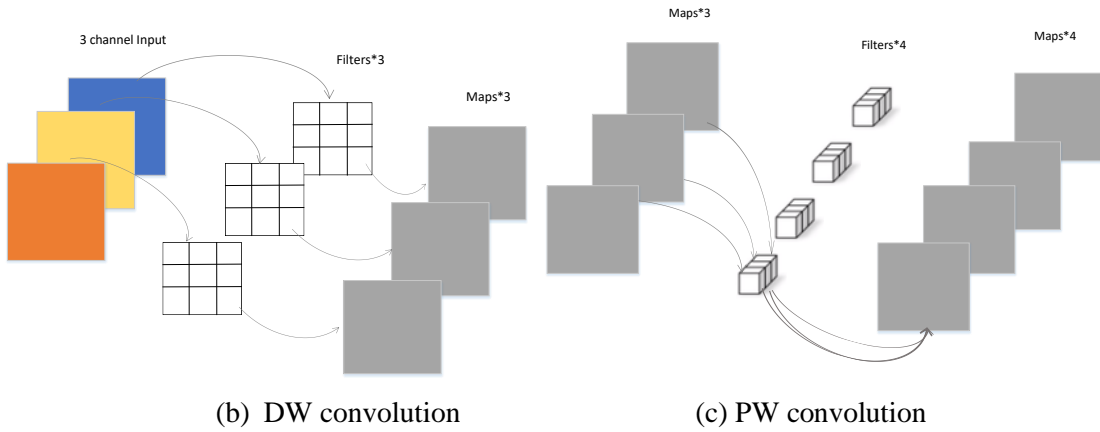


Fig 1 The structure diagram of normal convolution and depthwise Separable Convolution

Where, D_F represents the height and width of the input feature matrix, M represents the depth of the input feature matrix; D_K represents the depth of the input feature matrix, N represents the depth of the output feature matrix

Then, the calculation quantity of ordinary convolution is: $D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$, the calculation quantity of DW+PW is : $D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$ The ratio of calculation amount of MobileNet to that of traditional convolution is:

$$\frac{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{9}$$

The value of N is generally large, so the result is mainly determined. At present, the size of CNN convolution kernel is generally 3*3, so theoretically, the calculation amount of ordinary convolution is about 8~9 times that of DW+PW convolution.

Where, the value of N is generally large, So the result is mainly determined by D_K , at present, the size of the CNN convolution kernel is generally 3*3. In theory, the calculation amount of ordinary convolution is about 8-9 times that of DW+PW convolution.

MobileNetV2 uses the inverted residual structure, as shown in Fig2. The inverted residual structure increases the dimensionality of the input matrix by a 1*1 convolution, and then uses a 3*3 DW convolution kernel for convolution, and then uses a 1*1 convolution kernel for dimensionality reduction[15-17].

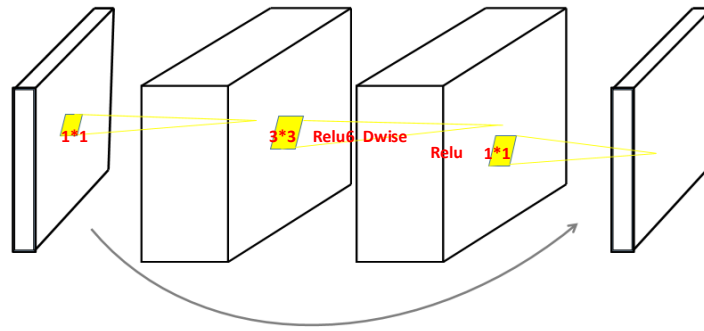


Fig 2 Inverted residual structure

The first two layers of the inverted residual structure use ReLU6 as the activation function. In the last convolutional layer, a linear activation function is used instead of the ReLU activation function to avoid causing a large loss of low-dimensional feature information. Fig 3 is the main flow chart of MobilenetV2.

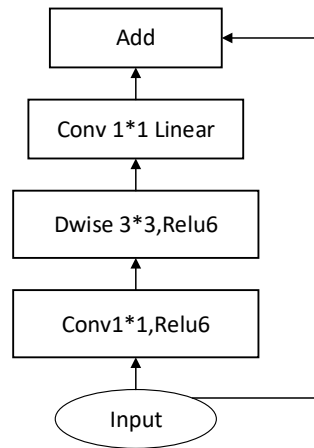


Fig 3 MobilenetV2 flowchart

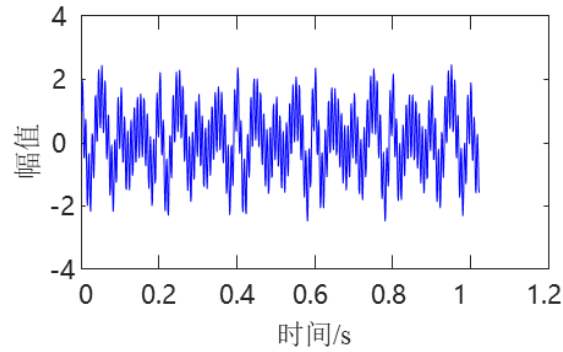
3.MAIN IDEA

3.1 Validity analysis of time-frequency graph of CWT

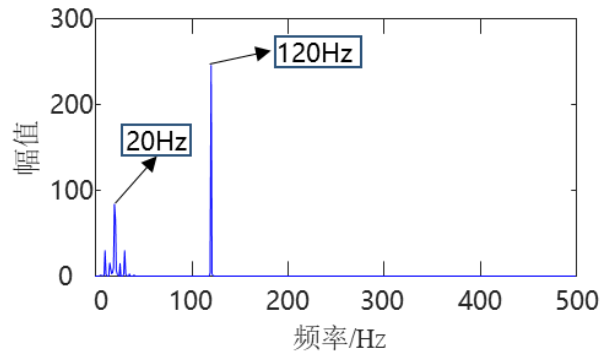
In order to prove the effectiveness of the CWT, The simulation of rolling bearing vibration signals were used to verify the results. A simulation signal with AM and FM characteristics is constructed. Time-frequency analysis based on PWVD and CWT are carried out for the simulation signal respectively. The simulation signal expression is as follows:

$$x(t) = (1+0.5\sin(2\pi \cdot 5.5t))\cos(2\pi \cdot 20t+0.8 \cdot \sin(2\pi \cdot 10t))+\sin(2\pi \cdot 120t) \quad (10)$$

Set the sampling frequency to $F_s = 1000$ Hz, the number of sample points $N = 1024$, and the waveform and power spectrum are shown in Fig4. It is known that the simulation signal is composed of an FM signal with a fundamental frequency of 20 Hz and a modulation frequency of 10 Hz, a 5.5 Hz amplitude modulation signal and a sine signal with a frequency of 120 Hz. Therefore, the 120 Hz frequency component in the signal always exists. In addition, There is also a frequency component that fluctuates with time around the fundamental frequency of 20Hz, with a frequency range of 12Hz~28Hz. These two frequency components can also be observed from the power spectrum.



(a) Simulation signal waveform diagram



(b) Power spectrum of simulated signal

Fig 4 Simulation signal waveform and power spectrum

Time-frequency analysis based on PWVD and continuous wavelet was carried out respectively for the simulation signals. As shown in Fig5, the time-frequency diagram is obtained, in which Fig5 (a) is the time-frequency diagram of wavelet continuous transformation. It can clearly observe the frequency component that always exists at 120Hz and the frequency component that is slightly blurred around the fluctuation of 20Hz. Fig5 (b) is the PWVD time-frequency diagram, and intermittent 120Hz and 20Hz frequency components could be observed, and they are not clear. In contrast, CWT time-frequency diagram has higher time-frequency resolution and can clearly present each frequency component.

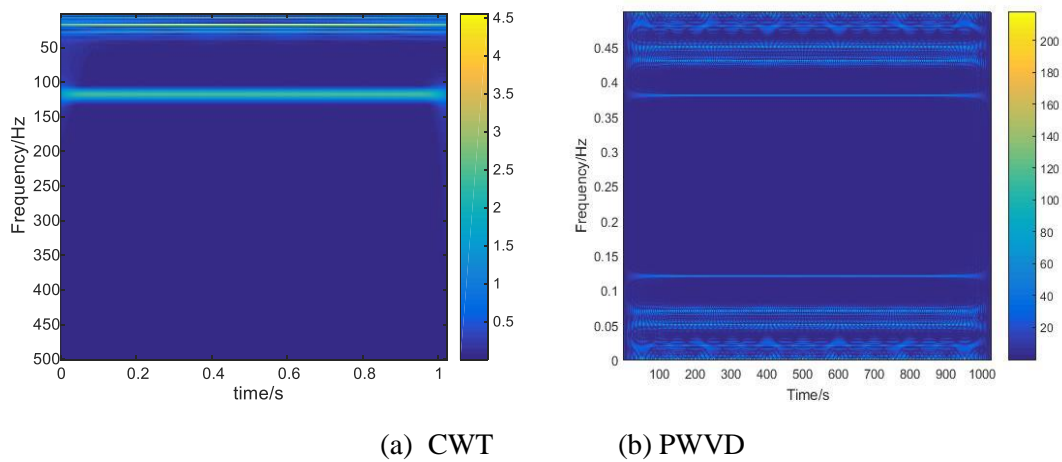


Fig 5 The time-frequency diagram of CWT and PWVD

3.2. Data enhancement

When classifying images with fault information, the lack of samples would lead to over-fitting of the model. Therefore, data enhancement can be performed on the images in the training set to improve the number of samples in the training set and the generalization ability of the neural network model. Image enhancement refers to generating new training samples by flipping, cropping, changing grayscale, contrast, and color. Thereby improving the invariance of the model such as scaling, and preventing the model from overfitting effectively.

3.3. Transfer learning

In order to accelerate the convergence speed of the model in the training process, the model used the transfer learning method to transfer the weights of the first several layers pretrained in other data sets of MobileNet to the corresponding network structures, and trains different neural network models on this basis.

3.4. Rolling Bearing Fault Diagnosis Based on VMD-CWT and MobileNet

In this paper, the VMD algorithm is used to extract the signal features, the CWT was used to extract the time-frequency features. After the data was enhanced, the MobileNet network was trained. In order to accelerate the convergence speed, in this paper, transfer learning was added in the network training process, and the weights of the first several layers which are pretrained on other data sets are transferred to the corresponding network.

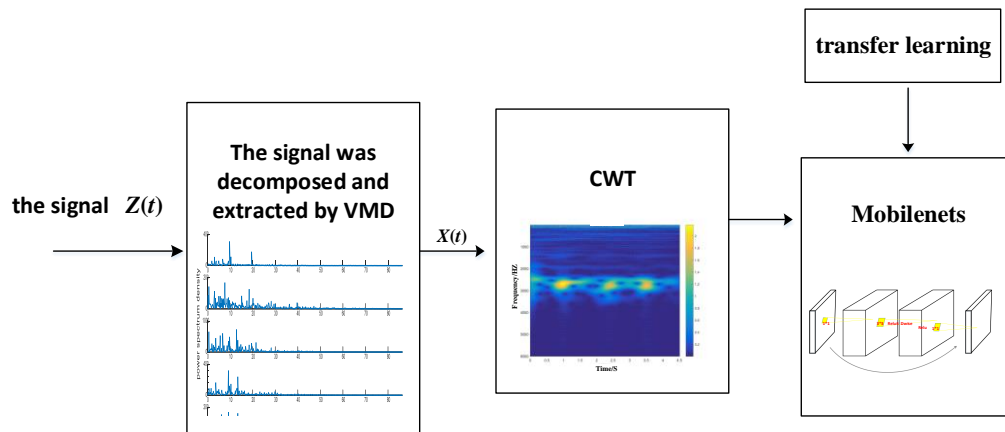


Fig 6 Flow chart of the proposed method

4. EXPERIMENTAL ANALYSIS OF PLANETARY GEAR BOX

4.1. The introduction of experimental platform

In this section, the data set of rolling bearing from Case Western Reserve University is used. Fig7 is the physical diagram and schematic diagram of the experimental platform.

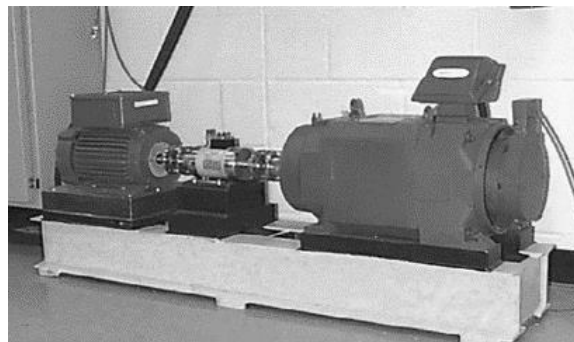


Fig 7 experimental platform

As shown in Fig7, the experimental platform consists of a 2 HP motor, torque sensor/encoder, power meter, and control electronics. They are respectively arranged on the left, middle and right side of the test bench. The test bearing supports the motor shaft, and a single point fault is introduced into the test bearing through EDM. The fault diameter is 0.007, 0.013 and 0.021 feet, all of which are SKF bearings. The specific parameters are shown in Table 1. In the experiment, an acceleration sensor was arranged to collect vibration data, and the sensor was arranged at the 12 o'clock position of the drive end and the fan end of the motor housing. The vibration signal is collected by a 16-channel DAT recorder, and later processed in the MATLAB environment. The sampling frequency of the signal sampling process is 12kHz and 48kHz. The fault of the outer ring is fixed, so the position of the fault relative to the bearing load area has a direct effect on the vibration response of the motor/bearing system. In order to make a quantitative study of this effect, the outer ring of the bearing at the drive end and the fan end were prepared with 3 o'clock, 6 o'clock and 12 o'clock faults respectively in the experiment.

Table 1 The parameters of test bench bearing

Installation location	Bearing designation	outer diameter	inner diameter	• Ball diameter	contact angle	Ball number
drive end	SKF6205	52	25	7.94	0	9
Fan end	SKF6203	40	17	6.75	0	8

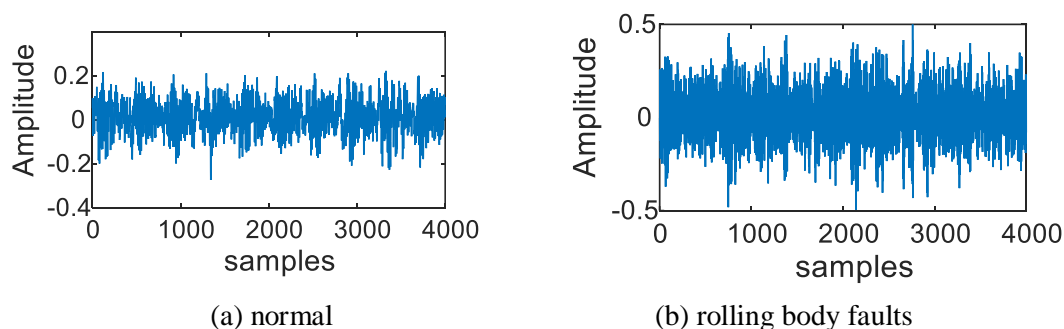
In this section, vibration signals at 1797r/min are selected to construct the data set. A total of 5 state samples are constructed, which are normal, 1 rolling body faults, 1 inner ring faults and 2 outer ring faults. The number of sampling points is 2000, and the specific number of sampling points and the categories corresponding to faults are shown in Table 2.

Table 2 Classification of rolling bearing datasets

sample number	Crack diameter	fault type	label
60	0	normal	0
30	0.007	rolling body faults	1
30	0.007	inner ring faults	2
30	0.007	outer ring faults@3	3
30	0.007	outer ring faults@6	4

4.2. signal analysis

Fig8 shows the waveforms of four vibration signals of rolling bearings under the condition of 0.007 inch crack



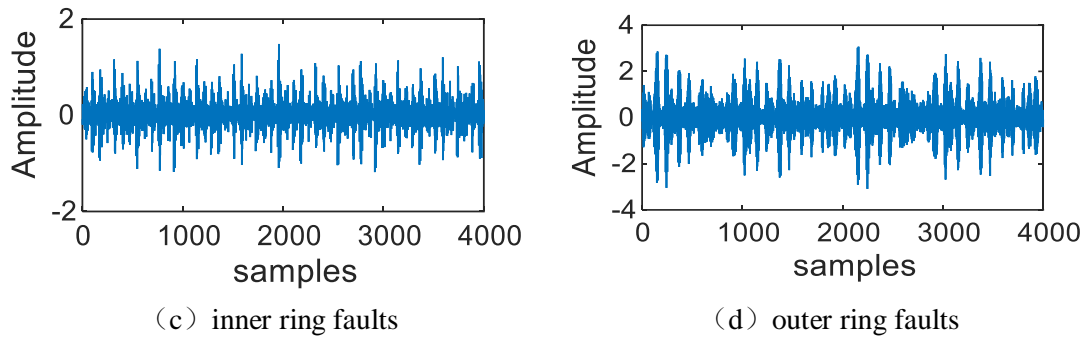


Fig 8 Time domain diagram of vibration signal

4.3. Experimental demonstration

As can be seen from the Fig8, the amplitude of different states varies greatly, the amplitude of normal state is the smallest, and the amplitude of outer ring fault is the largest. Among them, the impact of outer ring and inner ring faults is obvious. In order to improve the recognition rate of time-frequency images, VMD feature extraction was first carried out on the signals. The following is the comparison of time-frequency images before and after VMD feature extraction under normal and rolling body fault conditions.

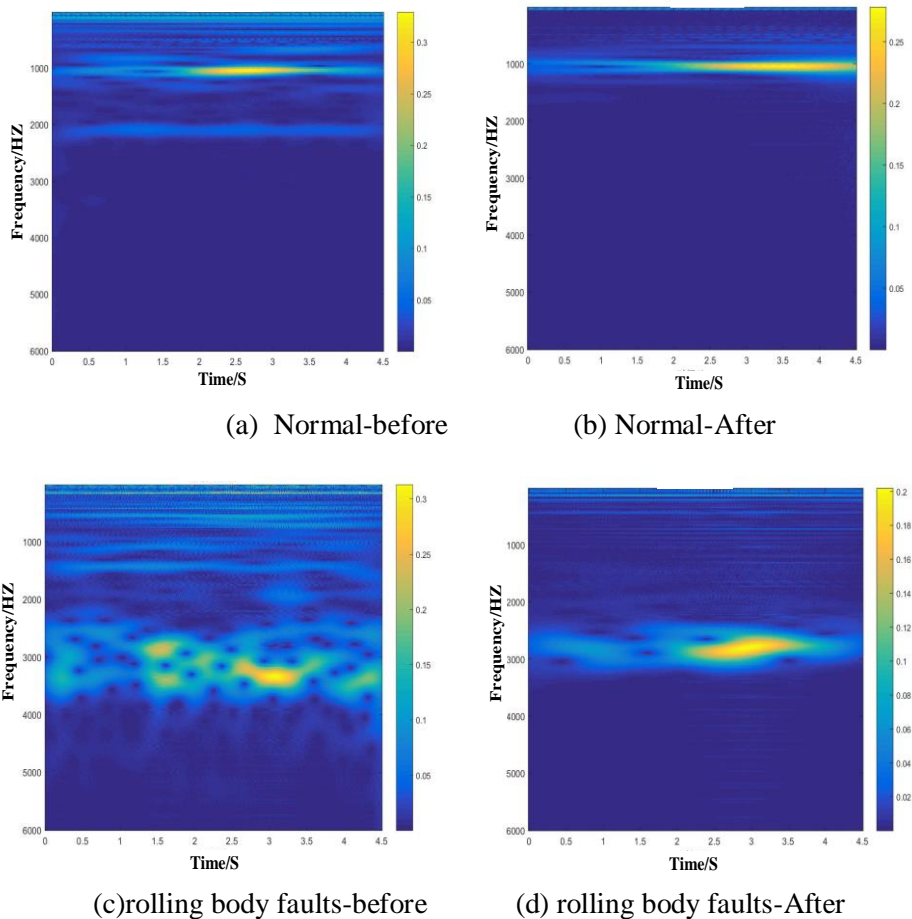


Fig 9 CWT time-frequency diagram before and after VMD processing

Fig9 (a) and (b) are the time-frequency diagrams before and after VMD processing under normal conditions, and Fig9 (c) and (d) are the time-frequency diagrams before and after VMD processing in the ball fault state. It is found by comparison that after VMD extracts the fault

features, the fault features of the time-frequency image are more obvious, the time-frequency focus of the time-frequency image is better, the impact is more obvious, and there are more noises before VMD processing, which is not conducive to fault identification.

Then, CWT is used to transform the signal into time-frequency diagram, and the processed time-frequency diagram is shown in Fig10

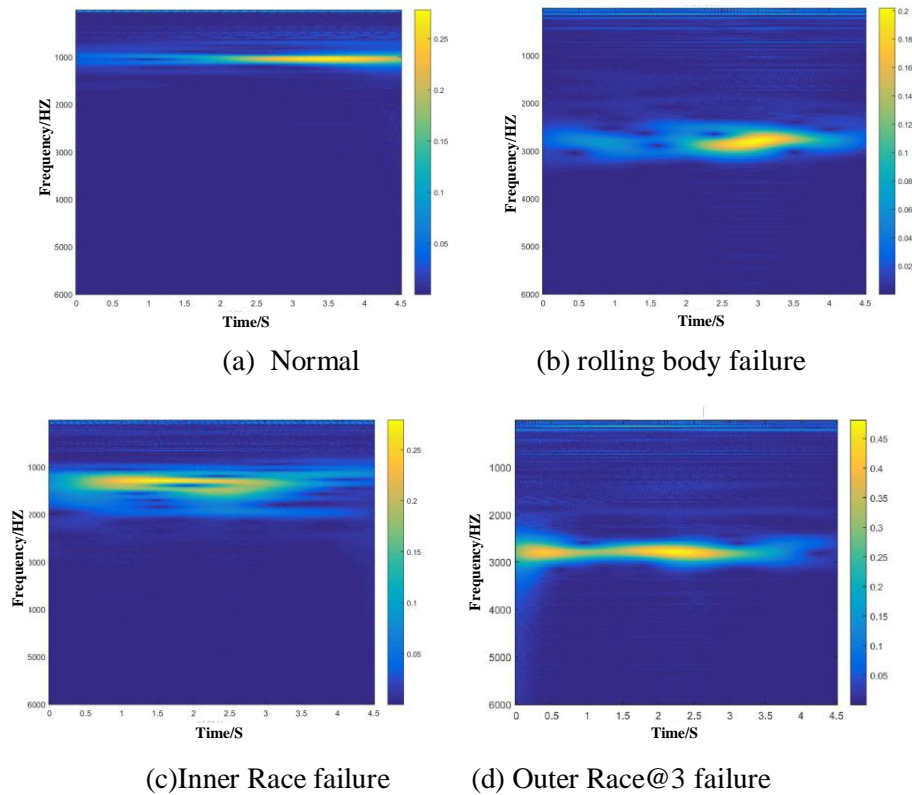


Fig 10 The time-frequency diagram of CWT

The processed time-frequency diagram is input into the migrated MobileNet network to train the network. Time-frequency image feature set includes 600 CWT time-frequency images of different parts of faults with a fault diameter of 0.007 inches. The dimension of each image is 224 by 224. The network training parameters were set as follows: batch_size = 16, epoch = 20. Under the condition of not adding noise, the rolling bearing fault identification based on vibration signal is realized through MobileNet.

Fig11 shows the error convergence curve of the MobileNet network in the training process, where the abscissor is the number of iterations of all batches of samples, and the ordinate is the recognition rate and loss value. As can be seen from the figure, the loss value of the model tends to 0 and remains stable when the iteration reaches the seventh time. At this point, the model has been trained to the convergence state. As can be seen from the figure, the test set has a recognition rate of 94%.

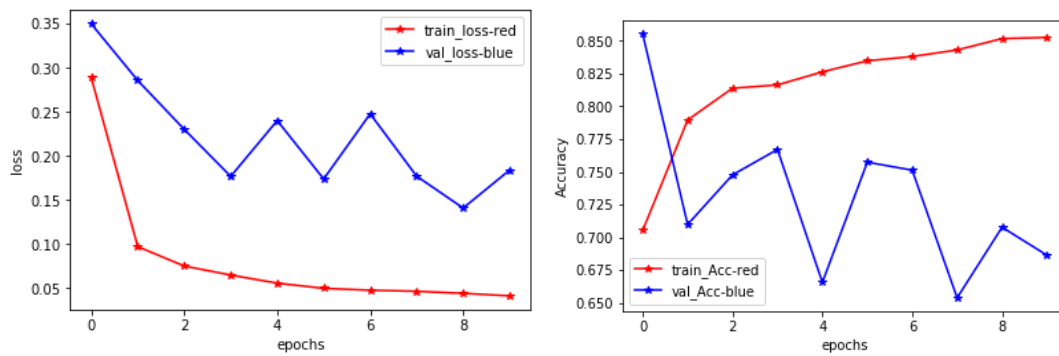


Fig 10 The error convergence curve of VMD-CWT-MobileNet

In order to further prove the effectiveness of the method presented in this paper, time-frequency images without VMD processing are used to train the MobileNet network, and the error convergence curve is shown in Fig11. It can be seen from the figure that the recognition rate of the test set only reaches 68.7%, which proves the superiority of the method in this paper.

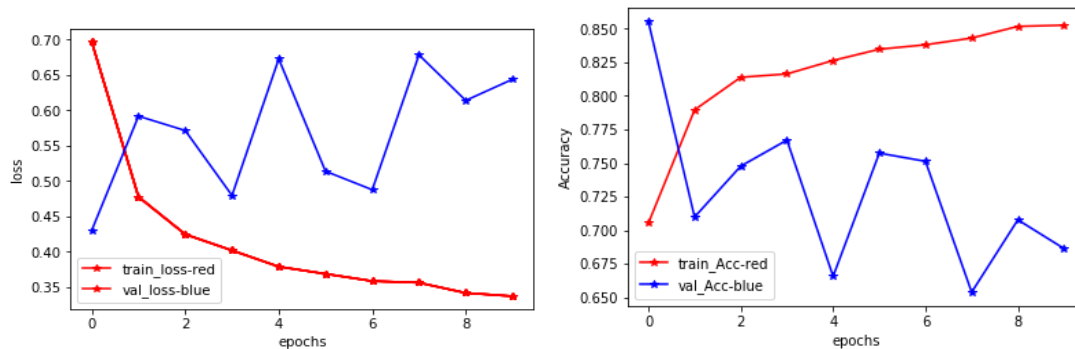


Fig 10 The error convergence curve of MobileNet

5. CONCLUSION

In this paper, the VMD algorithm is used to extract the signal features, and then the wavelet transform is used to extract the time-frequency features. After the data is enhanced, the MobileNet network is trained to accelerate the convergence speed. In this paper, transfer learning is added in the network training process, and the weights of the first several layers which are pretrained on other data sets are transferred to the corresponding network. The following conclusions can be drawn:

- (1). After the extraction of fault features by VMD, the fault features of time-frequency images are more obvious, the time-frequency focusing of time-frequency images is better, the impact is more obvious, and there are more noises before VMD processing, which is not conducive to fault identification.
- (2). The comparative experiment shows that the fault classification accuracy of the MobileNets network only is 68.7%, and the fault classification accuracy of the proposed method is 94%, which is a great improvement.
- (3). The comparison of CNN experiment shows that the training parameters of CNN network are 14,591,685 and the accuracy rate is 95%, while the network training parameters of the method in this paper are 2,264,389 and the accuracy rate is 94%. It is shown that the proposed method can reduce the network training time with a small reduction of accuracy.

REFERENCES

- [1] Zhang K , Xu Y , Liao Z , et al. A novel Fast Entrogram and its applications in rolling bearing fault diagnosis[J]. *Mechanical Systems and Signal Processing*, 2021, 154:107582.
- [2] Ma X , Zhou X , An F P . Bi-dimensional empirical mode decomposition (BEMD) and the stopping criterion based on the number and change of extreme points[J]. *Journal of ambient intelligence and humanized computing*, 2020, 11(2):623-633.
- [3] Liu G , Yang C , Liu S , et al. Feature Selection Method Based on Mutual Information and Support Vector Machine[J]. *International Journal of Pattern Recognition and Artificial Intelligence*, 2021:2150021.
- [4] Fault diagnosis of downhole drilling incidents using adaptive observers and statistical change detection[J]. *Journal of Process Control*, 2015, 30:90-103.
- [5] Wang H , Du W . Rolling bearing fault diagnosis based on Slice Energy Entropy Spectral Correlation Density-Continuous Hidden Markov Model[C]// 2019 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC). 2019.
- [6] LIU Y, HE B, LIU F, et al. Feature fusion using kernel joint approximate diagonalization of eigen-matrices for rolling bearing fault identification[J]. *Journal of Sound and Vibration*, 2016, 385: 389-401.
- [7] Hoang D T , Kang H J . Rolling element bearing fault diagnosis using convolutional neural network and vibration image[J]. *Cognitive Systems Research*, 2018, 53(JAN.):42-50.
- [8] Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
- [9] Meghana A S , Sengan S , Arumugam G , Srinivasan P , Kolla Bhanu Prakash. Age and Gender prediction using Convolution, ResNet50 and Inception ResNetV2[J]. *International Journal of Advanced Trends in Computer Science and Engineering*, 2020, 9(2):1328-1334.
- [10] Tian Y , Liu X . A Deep Adaptive Learning Method for Rolling Bearing Fault Diagnosis Using Immunity[J]. *Tsinghua Science and Technology*, 2019, 24(006):750-762.
- [11] Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
- [12] Wahab M F , O'Haver T C . Wavelet transforms in separation science for denoising and peak overlap detection[J]. *Journal of Separation Science*, 2020.
- [13] Liu Z , Chai T , Tang J , et al. Signal Analysis of Mill Shell Vibration Based on Variational Modal Decomposition[C]// 2020 39th Chinese Control Conference (CCC). IEEE, 2020.
- [14] Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
- [15] Koonce B . MobileNet v1[M]// *Convolutional Neural Networks with Swift for Tensorflow*. 2021.
- [16] Kumar K K R S , Subramani G , Thangavel S K , et al. A Mobile-Based Framework for Detecting Objects Using SSD-MobileNet in Indoor Environment[M]. 2021.
- [17] Saini R , Jha N K , Das B , et al. ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks[C]// IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020.