A Gradient-Based Optimizer Method for Improving the Feature Extraction Capability of Network Architecture and Its Application to Condition Monitoring of Rotating Machinery

Fengshuo Cao

College of Mechanical and Electrical Engineering, Harbin Engineering University, Harbin, China, 150001

18645116502@163.com

Abstract. Aiming at the problem of inaccurate feature extraction under the massive data of rotating machinery, this paper proposes a feature extraction capability enhancement method based on Gradient-Based Optimizer for trusted network architecture to monitor the conditions of bearings, which firstly transforms the original signal into a wavelet time-frequency image and visualizes the high-dimensional and difficult-to-visualize data in low dimensions. Then, different convolutional neural network architectures are tested by implanting different noise intensities to determine the most robust network architecture. Secondly, this paper adopts the grid method as well as the swarm optimization algorithm to guide the parameters of the model to achieve the best feature extraction effect, compares the optimization effect of several swarm optimization algorithms horizontally, and determines the most effective GBO algorithm among them. Finally, to improve the credibility of the best network model, this paper combines the grad-cam method with the trusted fusion method to improve the accuracy and reliability of the model. Validated by the bearing data of Case Western Reserve University, it shows that the proposed model in this paper has good credibility and accuracy.

Keywords: Gradient-Based Optimizer, Condition monitoring, Rotating machinery, CNN.

1. Introduction

Rolling bearings, as an important part of the rotating machinery, play a pivotal role in the manufacturing accuracy and reliability of the stable operation of the equipment to ensure the stable operation of the equipment and have a direct impact on the performance and stability of the machinery. Therefore, we need a method to monitor the health of bearings to ensure the smooth operation of rotating parts and avoid accidents [1].

The mechanical engineering field has been applied to common heuristic algorithms, such as BP neural networks [2]. In earlier years. However, the BP neural network is easy to fall into the local optimum. It has the disadvantage of too long a training time and too poor a nonlinear carrying capacity to meet the case of a large amount of data training.

In addition, the support vector machine (SVM) method has also been widely used. Zhao et al. [3] based on a reinforcement learning (RL) optimized support vector machine (SVM) model. proposed a data-driven early fault diagnosis method in view of vibration signals to distinguish abnormal vibration of mechanical equipment from motion information mixed with noise signals.

Azadeh Gholaminejad et al [4] compared the performance of shallow and deep neural networks and concluded that shallow neural networks limit their ability to learn complex and nonlinear expressions. Therefore, deep artificial neural networks are demanded to learn nonlinear models

Agahi H M A [5] researched the establishment of intelligent fault diagnostic systems, dividing it into five parts: namely signal segmentation, feature extraction and reduction, fault classification, and decision fusion, to explore how to obtain credible bearing detection results. This paper refers to Agahi H M A's method to expand the discussion.

Convolutional neural networks can be useful in Condition monitoring. However, noise immunity testing, parameter optimization, and decision-making plausible fusion are required to arrive at the final result as to which network is better or worse. Based on this, in this paper, we tested the GoogLeNet [6] ,Alexnet [7], and mobilenet [8], with noise at 15 dB based on gradient optimization for enhancing the feature extraction capability of the network architecture and concluded that GoogleNet is optimal. At the same time, the design of algorithms for parameterization of network feature extraction capability enhancement and model interpretable decision-making are also carried out.

2. Theory part

2.1. GoogleNet

There are many common deep-learning networks such as GoogLeNet [6], Alexnet [7], mobilenet [8], and so on. Each network has advantages and disadvantages due to differences in various aspects such as depth, breadth, feature fusion method, feature extraction method, number of parameters, etc.

The depth and width of convolutional neural networks are crucial to improve their performance, however, blindly increasing the depth will introduce too many parameters leading to overfitting on limited datasets. besides, the expansion of the network size makes the computation complex and limits its practical applications

At the same time, the increase in depth exacerbates the gradient dispersion phenomenon, posing a challenge to the optimization of the model. To overcome these difficulties, researchers have proposed various methods, such as model pruning to increase sparsity. However, this approach is often difficult to balance the relationship between model sparsity and accuracy in practical applications, and the improvement of algorithm operation speed is not significant.

As a result, the Google team came up with an innovative Inception architecture[6], which meets the requirements of a convolutional neural network with a depth of only 22 layers, and about one million parameters.

This structure allows the network to process information in parallel at multiple scales, thereby retaining key information in the input signal more fully. The network structure of GoogleNet specifically adds two auxiliary classifiers, which can back-propagate the gradient during training, effectively preventing the phenomenon of gradient dispersion.

At the core of the Inception architecture are nine Inception modules. These modules use three 1*1, 3*3, and 5*5 convolutional kernels to extract features at different scales of the image and fuse these features with a pooling module connected in parallel as an output to extract features at different scales simultaneously and fuse these features through deep stitching.

To further reduce the amount of computation and the number of parameters, the researchers added a 1×1 convolutional layer to the Inception structure to reduce the dimensionality and reduce the number of parameters to simplify the computation. Therefore, GoogleNet networks will be of great advantage when memory or computational resources are limited. With the continuous optimization and development of the Inception architecture, from Inception-V1 to Inception-V4, the performance has been significantly improved. The GoogleNet network is built on this structure, which achieves excellent performance under limited memory and computational resources through efficient feature extraction and fusion.

2.2. GBO optimization algorithm

After determining the network architecture, the problem of how to set the parameters, such as how many samples to choose and how many variables to be given for searching, is still to be solved. If the parameters are given too subjectively, it will lead to too much randomness or fall into the local optimum. So this paper adopts the swarm optimization algorithm to optimize the parameters of the network.

In recent years, researchers and scholars have proposed a wide variety of swarm intelligence optimization algorithms, such as the Whale Optimization Algorithm, (WOA) [9], Particle Swarm Optimization, (PSO) [10], Ant Colony Optimization, (ACO) [11], Gorilla Troop Optimizer (GTO) [12], Grey Wolf Optimizer (GWO) [13], Light Spectrum Optimizer (LSO) [14] and Gradient-Based

Optimizer (GBO) [15], In this paper, we compared these optimization algorithms above and finally reached the best one, GBO.

Gradient-based heuristic optimization (GBO) algorithms [15] have use of two main operations: the gradient search rule (GSR) and the local escape operator (LEO), as well as a set of vectors to search for.

GSR is suggested to enhance the search trend and accelerate the convergence for a better position in the search space. LEO ensures the GBO algorithm does not fall into the local optimum. The main process is as follows [15].

$$GSR = rand(0,1) \times \rho_1 \times \frac{2\Delta x \cdot x_n}{yp_n - yq_n + \varepsilon} \tag{1}$$

Where ρ_1, yp_n, yq_n are different stochastic parameters designed to create the right balance between search and exploration.

The stochastic parameter ρ_1 varies continuously with each iteration. It has a large value in early iterations to enhance population diversity and decreases with an increasing number of iterations to accelerate convergence.

To better explore the region around, a moving bi-directional DM is introduced to move x_n in the direction of $x_{best} - x_n$. This process creates reasonable local search trends and facilitates the convergence of the GBO algorithm. The DM can be represented as follows [15]:

$$DM = rand(0,1) \times \rho_2 \times (x_{best} - x_n) \tag{2}$$

$$X1_n = x_n - GSR + DM \tag{3}$$

$$X2_n = x_{best} - GSR + DM \tag{4}$$

Where $X1_n$ is more suitable for global search, $X2_n$ is more suitable for local search, and the GBO algorithm combines the two.

Finally, the new solution can be shown as [15]:

$$X3_n = x_n - \rho_1 \times (X2_n - X1_n) \tag{5}$$

$$X_n = r_a \times r_b \times X1_n + (1 - r_b) \times X2_n + (1 - r_a) \times X3_n \tag{6}$$

When rand<0.5:

$$\begin{aligned} X_{LEO} &= x_n + f_1 \times (\mu_1 \times x_{best} - \mu_2 \times x_k) + \\ &f_2 \times \rho_1 \times (\mu_3 \times (X2_n - X1_n) + \mu_2 \times (x_{r1} - x_{r2})) / 2 \end{aligned} \tag{7}$$

When rand>0.5:

$$\begin{aligned} X_{LEO} &= x_{best} + f_1 \times (\mu_1 \times x_{best} - \mu_2 \times x_k) + \\ &f_2 \times \rho_1 \times (\mu_3 \times (X2_n - X1_n) + \mu_2 \times (x_{r1} - x_{r2})) / 2 \end{aligned} \tag{8}$$

2.3. Grad-CAM

Convolutional neural networks are considered black boxes, we do not know the focus of their predictions and the reasons for the various predictions. So, they are not very interpretable.

The Grad-CAM algorithm, proposed by Selvaraju R R, [16] is based on the CAM (Class Activation Map) algorithm to improve, without using the attention mechanism and changing the network structure of the original model, achieve the neural convolutional network prediction process visualization by the heat maps.

It uses the gradient information of the feature map obtained by back-propagation calculation of the category confidence scores output from the classification model as weights.

The first step of the Grad-CAM model is to extract the highly abstract, semantically rich image features from the feature extraction network, which is the output of the last convolutional layer of the convolutional neural network.

Secondly, by back-propagating the predicted values for a specific category, the gradient information back-propagated to the feature layer can be obtained. Each value in this gradient information represents the contribution of an image feature to a specific category.

The network determines the importance of the corresponding feature to the category by the size of the contribution value. The final visualization can be obtained by overlaying the results obtained from Grad-CAM with the original map. This method can produce heat maps for specific categories without modifying the model and without retraining. Then, introduce the computational steps of this algorithm. Firstly, the partial derivatives of each channel feature map are computed, and global average pooling is performed to obtain the importance weight of the channel, which is computed by the formula [16].

$$\alpha_k^C = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^C}{\partial A_{ij}^k} \tag{9}$$

Then, to obtain the feature map with weights, the obtained weight matrix is multiplied by the original feature map, which is calculated by the formula [16].

$$L_{Grad\_CAM}^c = ReLU\left(\sum\nolimits_k \alpha_k^c A^k\right) \tag{10}$$

Finally, the results of Grad-CAM are measured with the voting method.

3. Method steps

The flow chart of the proposed method is shown in Fig.1, and the specific steps are as follows:

Step1: Based on the bearings data from Case Western Reserve University, this paper converts these data into a series of time-frequency images that can be recognized based on the two-dimensional wavelet transform

Step 2: Add Gaussian noise to the time-frequency images to form several groups of images to be tested, and select several neural networks to recognize the above time-frequency images with noise for the noise immunity test. Finally use the signal-to-noise ratio, the network architecture as the independent variable, and the accuracy as the target value to get the best neural network architecture.

Step 3: Test the parameters of the obtained network architecture with grid method, progressive grid method, and multiple swarm optimization algorithms, respectively, for parameter optimization. The test method is to set the range of the parameters and find the best solution in which the accuracy is the target value.

Step 4: Based on the optimized parameters and network architecture in the previous step, run the same algorithms with the same signal-to-noise ratio several times. Then combining the results of several runs, use the mean, voting, and probability methods to get the model accuracy after decision fusion, respectively.

Step 5: Based on the Grad-CAM method, assess the credibility of the results of the model identification, generate the corresponding heat maps, and calculate the identification accuracy according to each map. Finally, prove the credibility of the model by incorporating the voting method.

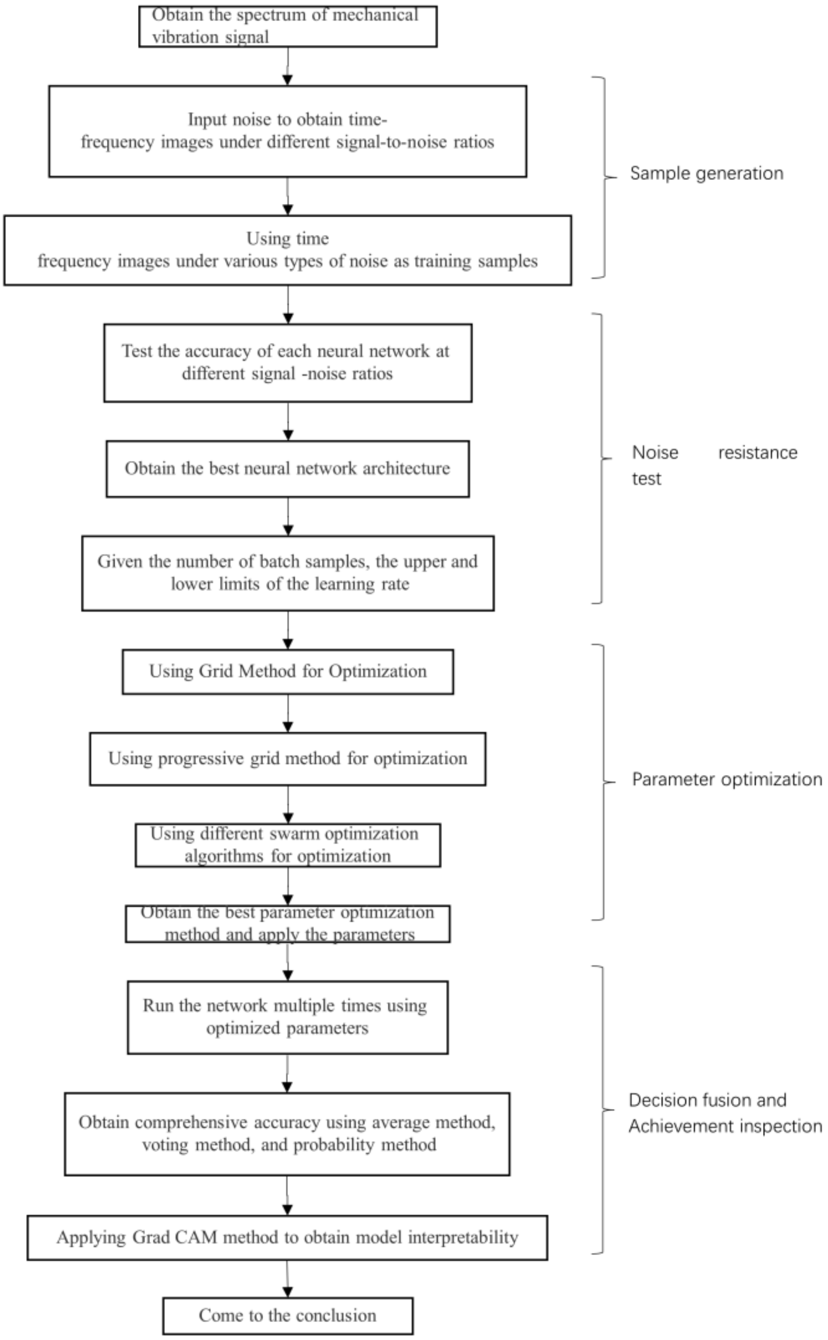


Figure 1. The flow chart of the proposed method

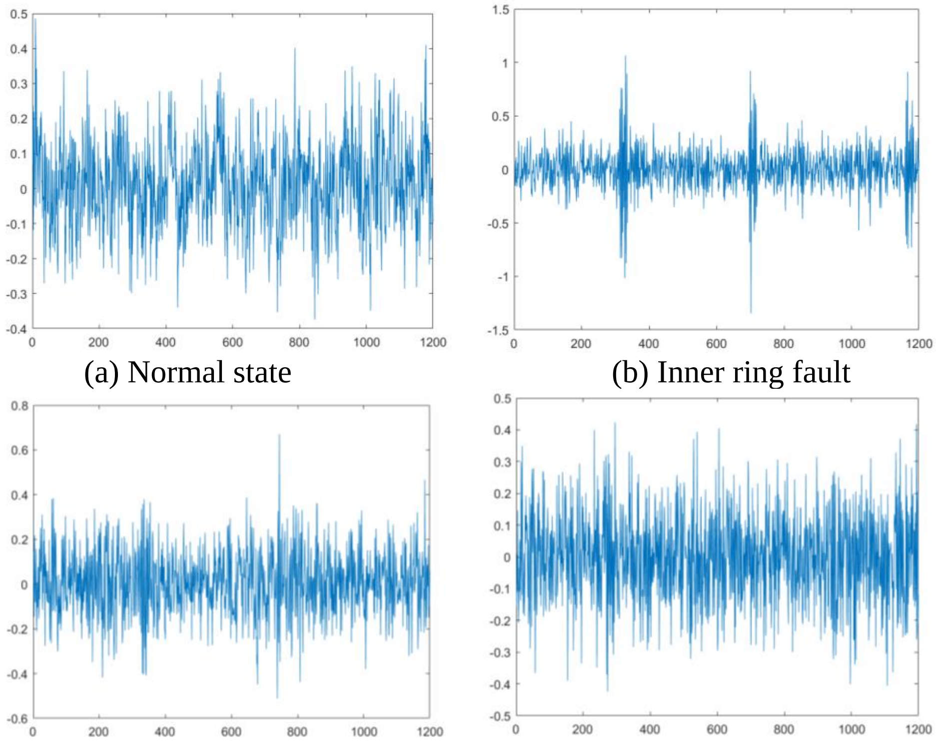
4. Case study

4.1. Source of data set

These test data originated from the bearing data level of Case Western Reserve University in the United States. During the testing, electric sparks were used to create a single point of failure in the rolling bearing ball, outer ring, and inner ring. In this experiment, electric sparks were used to create single-point failures in the balls, outer rings, and inner rings of rolling bearings. The fault sizes were categorized into four scenarios: 0.028in, 0.021in, 0.014in, and 0.007in. The motor loads were divided into four conditions: 0, 1hp/1797rpm (1hp equals to 745.700W), 2hp/1750rpm, and 3hp/1730rpm. The sampling frequency was set at 12kHz. The test bench collected vibration data from rolling bearings, encompassing four types: outer ring, inner ring, ball fault, and normal state. This paper selects some parts of the data above to complete the study.

4.2. Experimental comparison and analysis

In this study, the experimental data were collected under the following conditions: a motor with load of 3hp/1730rpm and a sampling frequency of 12KHz. The data encompassed bearing fault information for three different fault conditions, all with a fault size of 0.014in. Additionally, the normal state was also included, resulting in a total of four fault states being considered.



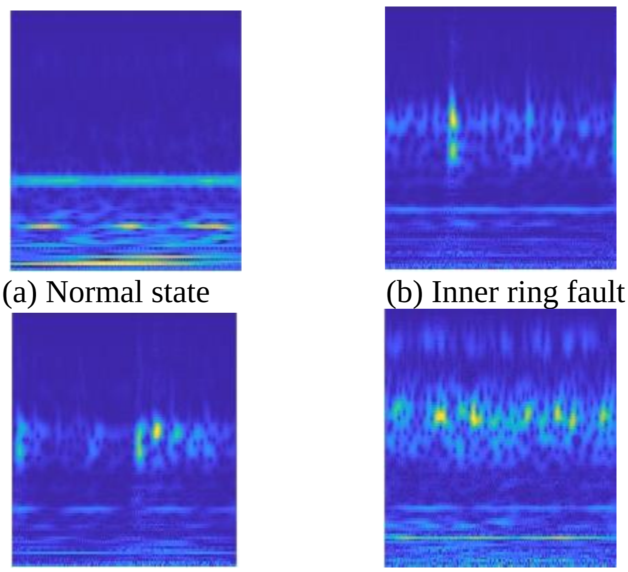
(a) Normal state (b) Inner ring fault

(c) Ball fault (d) Outer ring

Figure 2. The original signals of different states

In fig.2, the original data in this paper generated several sets of time-frequency images using the method of two-dimensional wavelet transform. The following four images as shown in Fig.3 represent the time-frequency images under different fault conditions.

Thus the original one-dimensional classification problem of monitoring whether the bearing frequency is abnormal can be transformed into a two-dimensional image recognition problem.



(a) Normal state (b) Inner ring fault

(c) Ball fault (d) Outer ring

Figure 3. The time-frequency diagram of different states

But for the actual data, if the trained neural network can only identify the clear experimental data of Case Western Reserve University, it is not enough. So this paper adds Gaussian white noise of different strengths to the original data to test whether different networks can identify the experimental data with mixed noise, and here introduce a concept of signal-to-noise ratio, which can be explained by the following equation.

$$\frac{BB}{BB' - BB} = SNR \tag{11}$$

Where: BB is the signal from the original experimental data, BB' is the new signal after mixing the noise, SNR is the signal-to-noise ratio.

Finally, with the same number of batch samples and learning rate, set the output layer of each neural network to 4, and run the above data into the neural network to get the corresponding accuracy data.

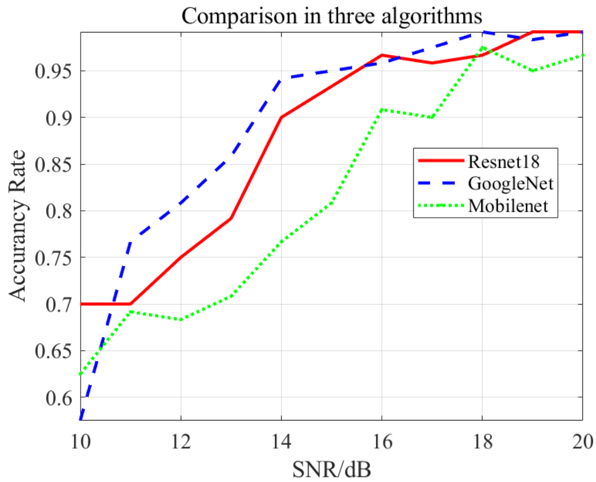


Figure 4. Noise resistance curves of different methods

In Fig.4, taking SNR as the independent variable and the accuracy of different networks at different signal-to-noise ratios as the dependent variable, this paper gives a set of curves to judge the performance of different neural networks, to get the results of their noise immunity tests. Comparing the accuracy of several neural networks shown by the graph above, and combining their running times, we get the conclusion that the GoogleNet network is the best choice.

After determining the architecture of the neural network, it is also necessary to determine exactly how the number of batch samples and the learning rate in the neural network should be given. In this paper, the most primitive grid method is firstly used for traversal, and the result is shown in Fig.5.

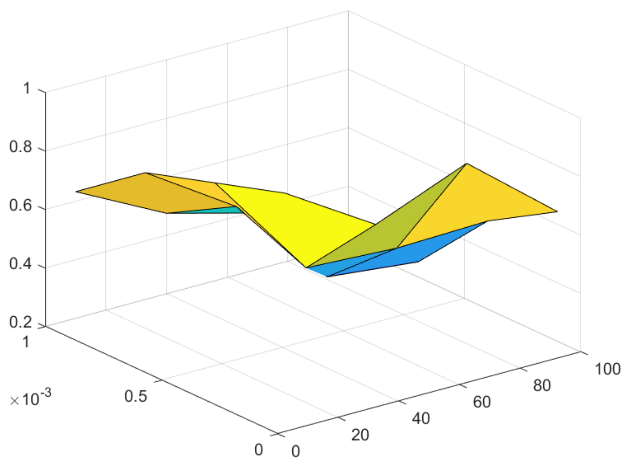


Figure 5. Precision grid diagram

As is shown in the fig.5, although the results of the grid method run are very intuitive, it takes too long time to run and tends to be very time-consuming when the amount of data is too large. So we introduced a progressive mesh method to improve it. But it do not have a figure to show the result intuitively, so here just give the solution. Finally we reached the solution that number of batch samples =8, learning rate =0.14167. This method runs significantly faster compared to the latter, but it tends to fall into a local optimum, leading to failure to find the exact solution This requires the use of swarm optimization algorithms. As table 1 shown, in this paper, 9 different swarm optimization algorithms such as LSO, SSA...etc. are selected, all with the same data input, and set the upper and lower limits of the number of initial batch samples and the learning rate. In Fig.6, the Iterative curve of 3 better algorithms in 9 different swarm optimization algorithms is visible. Comparing the accuracy rates obtained by each of them, and integrating their convergence speeds and runtimes, to obtain the conclusion that the GBO algorithm is the best algorithm.

Table 1. The optimal objective function of different optimization algorithms

Methods	Optimal objective function
LSO	0.1250
SSA	0.0917
SGBO	0.0500
RUN	0.1412
WOA	0.0583
GTO	0.0833
AVOA	0.0917
EO	0.1083
GWO	0.0667

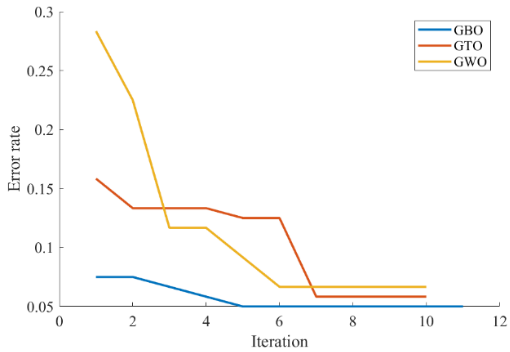


Figure 6. Graphs made from three better algorithms

Finally, this paper optimizes the neural network according to the parameters given by the best GBO algorithm. In fact, for a single run, it is random in the single run neural network due to the randomness of its initial value selection and sample selection. Therefore, in order to make the model more accurate, multiple runs are needed to ensure that the chance errors can be reduced, so this paper adopts the average value method, the voting method, and the probability method to solve the problem respectively. The comparison of the accuracy obtained by the three methods is shown in Fig.7.

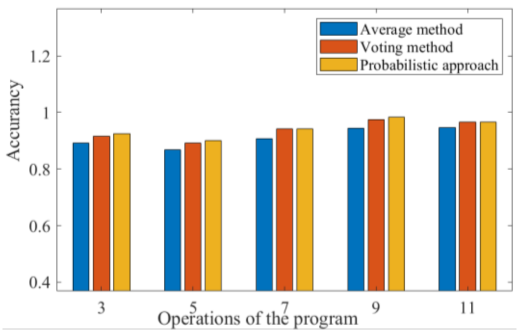
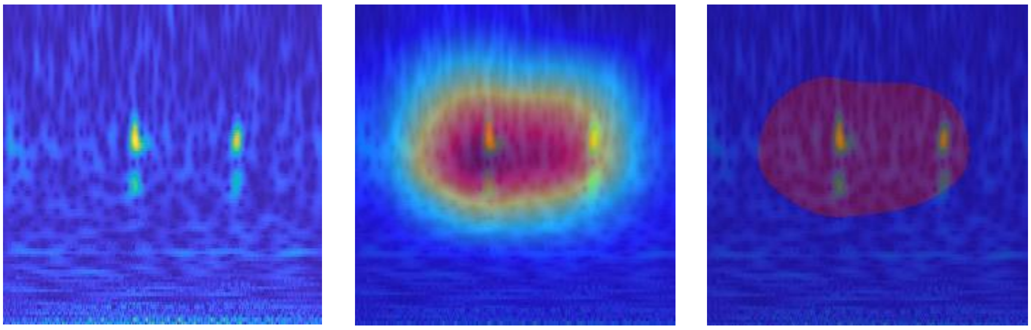


Figure 7. Accuracy of the three methods in each base network

After the previous steps, it has been possible to get the purpose we want to identify the type of bearings failure, but there is still a problem that whether the results of the neural network operation is reliable or not. what we can see is that the frequency should be aggregated in the normal situation below 2000Hz, while the abnormal situation should be aggregated in the above 2000Hz. However, it is still not known whether the neural network inference is reliable or not, so the final step is to verify its interpretability.

The method used in this paper to verify the interpretability of the model is the Grad-CAM method (Gradient based -Class Activation Mapping). After classification in the previous step, we need to import the image, crop it, and apply the Grad-CAM method to the cropped image to get the heat map, as shown in Fig.8

Subsequently, a thresholding process is applied to the heatmaps in this paper. Values of dots which is greater than 0.4 are converted to 1, while the rest are converted to 0, resulting in a set of 0-1 matrices. Finally, the trustworthiness of the model is determined by calculating the area of the matrix within a specific frequency range. If the area of the matrix in the region corresponding to 0-2000Hz is relatively large for bearing data that has been classified as normal, it is considered that the network is accurate. Conversely, if the area is small, the results of the network operation are deemed untrustworthy.



(a). origin graph (b). heat graph (c). thresholding process graph

Figure 8. The process of trusted decision making

As shown in the Fig.8, the highlights of the anomalous portion of this heat map are identified and thresholder to maintain the original main identifying features. So the network is considered to be credible. This method is then applied in a batch process to calculate the trustworthiness of each image based on the overlap area between the fault frequency band and the class activation mapping. Instead of assigning one vote based solely on the identified type for each image, the trustworthiness of each image is used as the number of votes in the voting method. This means that the number of votes for this method is a decimal number less than 1, instead of the original 1. Decision-making is then fused based on the trustworthiness scores. To achieve this, the nine best-performing networks identified in the previous step are selected and rerun to calculate the trustworthiness.

After fusing the trustworthiness scores, a precision of 0.9417 is achieved, meeting the expected requirements for trustworthiness.

5. Conclusion

Aiming at the problem of inaccurate feature extraction under the massive data of rotating machinery, this paper proposes a feature extraction capability enhancement method based on gradient optimization for trusted network architecture, the main conclusions are as follows:

1) Briefly describes the importance of bearing monitoring to mechanical engineering and the current status of neural network application, based on the imperfect monitoring system, puts forward a bearing image recognition method based on trusted decision fusion and deep network optimization.

2) The method based on GBO-Googlenet, trusted decision fusion and other theories proposed method can adaptively extract the fault information contained in the health condition signal spectrum, get rid of the dependence on a large number of signal processing knowledge and diagnostic

engineering experience, and achieve high monitoring and diagnostic accuracy as well as model credibility.

3) After the data verification of Case Western Reserve University, the method is feasible and robust, and can complete the task of monitoring the health condition of bearings under the situation of large noise.

References

[1] Wang H, Zhang W, Han C, et al. A bearing fault diagnosis method with an improved residual Unet diffusion model under extreme data imbalance [J]. Measurement Science and Technology, 2024, 35 (4).

[2] Qu H, Zhang W, Zhang G. A novel fault diagnosis scheme for a ballast system [J]. Automatica, 2023.

[3] Zhao W, Lv Y, Liu J, et al. Early fault diagnosis based on reinforcement learning optimized-SVM model with vibration-monitored signals [J]. Quality Engineering, 2023 (4): 35.

[4] Gholaminejad A, Jorkesh S, Poshtan J. A comparative case study between shallow and deep neural networks in induction motor's fault diagnosis [J]. IET Science, Measurement & Technology, 2023.

[5] Agahi H M A. Decision fusion scheme for bearing defects diagnosis in induction motors [J]. Electrical engineering, 2020, 102 (4).

[6] Liu X, Aldrich C. Recognition of flotation froth conditions with k-shot learning and convolutional neural networks [J]. Journal of Process Control, 2023: 128.

[7] Shen J, Song D, Ma T, Xu F. Blade crack detection based on domain adaptation and autoencoder of multidimensional vibro-acoustic feature fusion. Structural Health Monitoring. 2023 Sep; 22 (5): 3498-513.

[8] Shen J, Ma T, Song D, Xu F. Quantitative blade damage detection based on multisource domain and multistage joint transfer. Structural Health Monitoring. 2024 Feb 12: 14759217231226381.

[9] Mirjalili S, Lewis A. The Whale Optimization Algorithm [J]. Advances in Engineering Software, 2016, 95: 51-67.

[10] Van N L, Jeyaprakash N, Yang C H. Nondestructive characterization of elastic modulus under different tensile stresses through laser ultrasound combined inversion technique [J]. International Journal of Modern Physics B, 2024, 38 (04).

[11] Jia Y H, Mei Y, Zhang M. Confidence-Based Ant Colony Optimization for Capacitated Electric Vehicle Routing Problem with Comparison of Different Encoding Schemes [J]. IEEE Transactions on Evolutionary Computation, 2022, 26: 1394-1408.

[12] Abdel-Basset M, El-Shahat D, Sallam K M, et al. Parameter extraction of photovoltaic models using a memory-based improved gorilla troops optimizer [J]. Energy conversion & management, 2022 (Jan.): 252.

[13] GAI J B, SHEN J X, HU Y F, et al. An integrated method based on hybrid grey wolf optimizer improved variational mode decomposition and deep neural network for fault diagnosis of rolling bearing [J]. Measurement, 2020, 162 (10): 1-12.

[14] Mohamed A, Reda M, M. K S, et al. Light Spectrum Optimizer: A Novel Physics-Inspired Metaheuristic Optimization Algorithm [J]. Mathematics, 2022, 10 (19): 3466-3466.

[15] Ahmadianfar I, Bozorg-Haddad O, Chu X. Gradient-Based Optimizer: A New Metaheuristic Optimization Algorithm [J]. Information Sciences, 2020, 540. DOI: 10.1016/j.ins.2020.06.037.

[16] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [J]. International Journal of Computer Vision, 2020, 128 (2): 336-359.