Research on gearbox fault detection model based on ridge regression and decision tree

Jiacong Zhou[1, #, *], Qianqian Chen[2, #], Nan Zhang[2, #], Hanqing Hua[3], Yangke Ni[4]

[1]Department of Information and Electronic Engineering, Zhejiang University of Science and Technology Hangzhou, Zhejiang, 310023

[2]Department of Science, Zhejiang University of Science and Technology, Hangzhou, Zhejiang, 310023

[3]Department of Economics and Management, Zhejiang University of Science and Technology, Hangzhou Zhejiang, 310023

[4]Department of Automation and Electrical Engineering, Zhejiang University of Science and Technology Hangzhou, Zhejiang, 310023

*Corresponding author: z2474087637@163.com

#These authors contributed equally

Abstract. Gearbox is widely used in mechanical equipment and plays an important role in mechanical transmission. Therefore, it is necessary to detect and diagnose the fault of the gearbox in time. This paper needs to establish a fault detection model of the gearbox to detect whether the gearbox is in a fault state. Because the result can only be yes or no, the ridge regression model is first established. However, because the difference between the original sample data is not obvious, the accuracy of the obtained ridge regression model is low. Therefore, this study extracts the features of the data of the four parts, and defines five indicators: effective value, pulse index, skewness index, margin index and kurtosis index. The decision tree model is established with 70 % of the sample data. Firstly, the depth of the largest tree is set to 5. Secondly, the importance of the feature is determined according to the size of the Gini value, and the fault detection decision tree model is constructed. Finally, the model is tested with 30 % of the test data, and the accuracy is 91.17 %. The precision rate is 88.9 %, and the recall rate is 94.12 %. It is considered that the model is more reliable.

Keywords: Gearbox, Ridge regression model, Fault detection, Decision tree model.

1. Introduction

Gearbox is one of the important components of many modern mechanical equipment. It is mainly used for variable speed, changing the direction and torque of rotation, and distributing power[1,2]. The gearbox is composed of several gears. Through the rotation of the gear, the force can be transmitted from the crankshaft timing gear to the axle gear. The running state of the gearbox has a great influence on the performance of the engine, which will cause the output torque and speed to deviate. Therefore, it is necessary to carry out fault detection and diagnosis of the gearbox regularly. However, most of the fault detection and treatment methods of the gearbox now adopt regular inspection and maintenance, which has the problems of low efficiency, high labor cost and low certainty of future fault prediction[3]. Therefore, it is particularly important to study the detection and diagnosis methods of gearboxes. By comparing the vibration signals collected under different fault conditions with those under normal conditions, a detection and diagnosis model is established to analyze and solve the fault conditions of gearboxes[4, 5].

This study needs to further diagnose the gearbox, determine the specific fault state of the gearbox in the fault state, establish a fault diagnosis model, and evaluate the established model.

1.1. Analysis of the problem

The purpose of this study is to establish a model to determine whether the gearbox has a fault. Since the fault result can only be a yes / no label, we give the data a value of 1 / 0.Since the number of 1 label data is only 29400, the corresponding number of 0 label data is randomly selected to avoid the result deviation caused by the imbalance of the number of samples. Then we establish a ridge regression model based on 70 % of the sample data, and use 30 % of the sample data to test the accuracy of the model. However, due to the small difference of sample data, the data can not reflect the difference of each fault machine, resulting in low prediction accuracy. Therefore, we sliced the characteristic data of vibration signals generated by four different parts according to the cycle, and extracted five characteristics respectively : effective value, pulse index, skewness index, margin index and kurtosis index. According to the above five characteristics, a decision tree model was established to determine whether the gear machine was faulty, and 30 % of the sample data was used to test the accuracy of the model, and the precision, recall and accuracy of the model were analyzed. The results were further tested, and the decision tree model obtained by the final analysis (Figure 1).

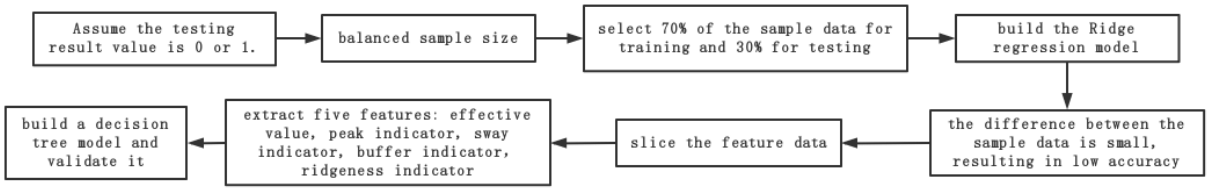


Figure 1. Technology roadmap

2. Development of fault detection model for gearboxes

2.1. Data pre-processing

Assign a label 1 to each entry of a normal gear machine and a label 0 to each entry of a faulty gear machine.

$$\begin{cases} 1, & x = Normal\ Data \\ 0, & x = Fault\ data \end{cases} \tag{1}$$

25% of the data from each of the four faulty machines are extracted and added to the normal machine sample set.

2.2. Development and solution of ridge regression model

3.2.1 Ridge regression modeling

Ridge regression is a modification and extension of the least squares method, which is an effective method specifically designed to solve the problem of data covariance in multiple linear regression. If the coefficients of the independent variables are estimated, the coefficients lack stability and lose the explanatory effect on the dependent variable[6-8]. The coefficients of the independent variables are estimated by ridge regression, which can reflect the relationship between the independent variables and the dependent variables by giving up the advantage of unbiasedness at the expense of losing some information and reducing the accuracy of the fit, and increasing the objectivity, stability, and reliability of the regression coefficients[9].

Multiple linear regression model

$$Y = X\beta + \varepsilon \tag{2}$$

Where Y is the dependent variable, X is the independent variable (in the form of a multivariate matrix), β is the regression coefficient, and ε is the error.

Let s_1, s_2, s_3, s_4, period, and mean values be X_1, X_2, X_3, X_4, X_5, and X_6, respectively, where the values of some of the periods are too large, and in order to avoid the influence of too large data on the results, this question performs a natural logarithmic transformation of the periods:

$$X_5 = \log(X_5) \tag{3}$$

The ridge regression method for solving the regression coefficient β is

$$\beta = (X^T X + kI)^{-1} X^T Y \tag{4}$$

Where k is the ridge regression parameter, and k is taken as 1 in this question[10].

3.2.2 Solution of the ridge regression model

The flow chart for solving the ridge regression model is shown in Figure 2.

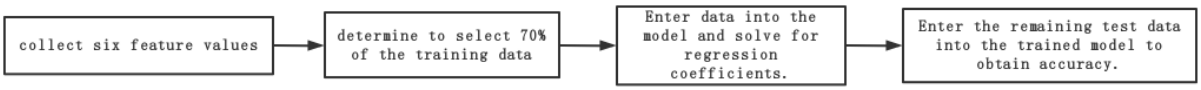


Figure 2. Flow chart of ridge regression model solution

Step1: The values of s_1, s_2, s_3, and s_4 are extracted directly in their respective states, and their periods and means are calculated as eigenvalues, respectively.

Step2: The values of the periods are processed and substituted into equation (3).

Step3: Solve the regression coefficient β of the ridge regression and substitute into equation (4) to get the result.

Step4: Substitute the remaining test data into the equation and perform the test of accuracy.

3.2.3 Results and analysis of the ridge regression model

In this study, we need to solve the regression equation for the gearbox in five states, and regress the five sets of data separately.

The regression equation for the normal gearbox was obtained as

$$Y = 1.72 - 0.025X_1 - 0.395X_2 + 0.164X_3 - 0.003X_4 - 0.270X_5 - 4.317X_6 \tag{5}$$

The results were checked for accuracy and completeness. Based on the prediction results, the confusion matrix of normal normal gearboxes was established as shown in Table 1.

Table.1. Confusion Matrix

Real situation	Number of predicted versus true scenarios	
	Normal	Fault
Normal	8854	4422
Fault	0	4364

According to the definition of recall rate P, precision rate R and measurement index F_1, the relevant indexes of the traditional method can be calculated as follows: P = 66.7 %, R = 1, F_1 = 75.9 %. The accuracy is low, so we began to change our thinking, modify and filter the selected features, and establish a new model.

2.3. Establishment and solution of decision tree model

Because the difference between the original data is not obvious, the accuracy of the model is low, so the four feature data are sliced according to the cycle of S_4, the features are extracted respectively, and the decision tree model is established(Figure 3).

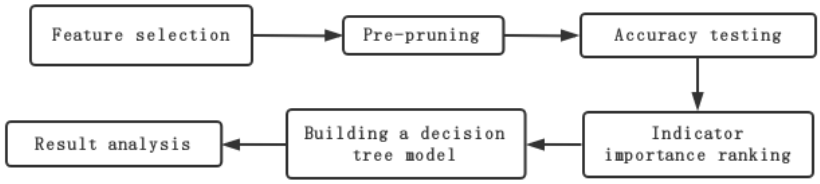


Figure 3. Decision tree flow chart

3.3.1 Establishment of decision tree model

Feature selection

According to the period of s 4, s 1, s 2 and s3 are segmented, and relevant papers are consulted. The effective value, pulse index, skewness index, margin index and kurtosis index of the five features are extracted respectively. See below:

Effective value:

$$X_{\text{rms}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{6}$$

(2) Pulse indicators:

$$C_f = X_p/\mu \tag{7}$$

Margin index:

$$C_e = X_{rms}/\mu \tag{8}$$

Skewness index:

$$C_w = [\frac{1}{N}\sum_{i=1}^{N}(|x_i|-\mu)^3]/X_{rms}^3 \tag{9}$$

Kurtosis index:

$$C_q = [\frac{1}{N}\sum_{i=1}^{N}(|x_i|-\mu)^4]/X_{rms}^4 \tag{10}$$

In the above formula, xi is the amplitude of the vibration signal sample data i ; n is the number of vibration signal sample data.

Step1: Pre-pruning of decision tree

By limiting the width and depth of the tree, the decision tree model is prevented from overfitting. Based on the analysis, the depth is limited to 5.

Step2: The establishment of node hierarchy

Gini value G (p)

The Gini value is used as the basis for node classification, and the attribute that minimizes the GINI value of the child node is selected as the splitting scheme.

Firstly, the Gini value of each index is calculated and sorted. The smaller the Gini value is, the higher the purity is, and the more it is located above the tree. The index with smaller Gini value is selected on the upper layer of the tree to construct the decision tree.

$$G(p) = \sum_{k=1}^{K} P_k\,(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{11}$$

Where pk represents the probability that the sample belongs to the kth state.

$$p_k = \frac{N_k}{N} \tag{12}$$

N_k is the number of states k.

Importance of indicators

$$\frac{N_t}{N} * (\boldsymbol{G(p)} - \frac{N_t_R}{N_t} * \boldsymbol{right_G(p)} - \frac{N_t_L}{N_t} * \boldsymbol{left_G(p)}) \tag{13}$$

Among them, N is the total number of vibration signals, N _ t is the number of fault vibration signals of the current node, N _ t _ L is the number of vibration signals of the lower left eigenvalue of the current eigenvalue, and N _ t _ R is the number of vibration information of the lower right eigenvalue of the current eigenvalue.

(1) The samples are randomly sampled (Bagging sampling) to train the decision tree.

(2) Among the samples containing M features, m features (m < < M) are selected and selected as the features of split nodes according to ' GiniImpurity '. The calculation method is as follows:

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{14}$$

Where represents the number of classifications, and the probability that a data is of the class is p (i).

(3) Repeat step (2) for each node until the depth of the tree reaches the set maximum depth to generate a decision tree.

Step3 : Establishment of decision tree model(Figure 4).

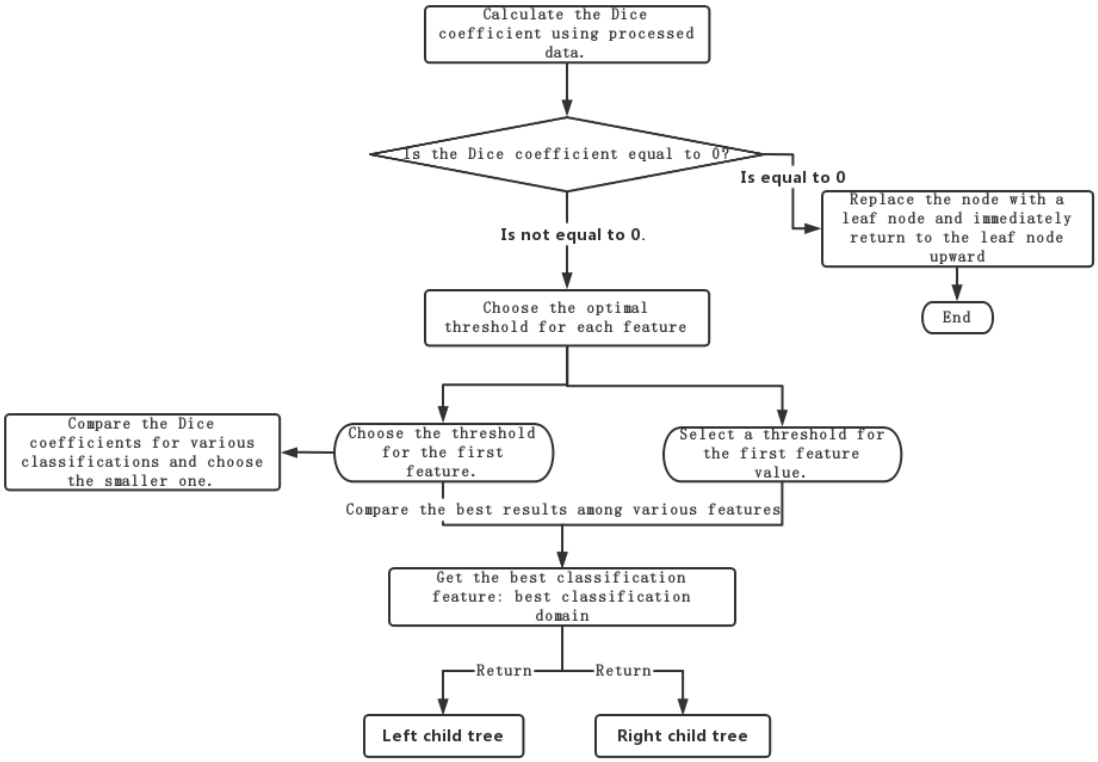


Figure 4. Flow chart of decision tree establishment

The method of selecting the 0th eigenvalue: first, all the eigenvalues of the 0th feature are sorted, and the average value between each two eigenvalues is used as a threshold classification.

3.3.2 Solution of the model

Step1: Because there are 'xyx3 ' and ' yd3 ' indicators that can clearly distinguish whether the gear machine is faulty, the boundary of segmentation is shown in table 2 below.

Table 2. Index partition table

	0	1
xyx_3	P(i) < 4.972	P(i) > 4.972
yd_3	P(i) < 1.006	P(i) > 1.006

In order to observe the influence of other indicators on the results, we eliminate these two indicators and rebuild the decision tree for the other 18 indicators.

Step2: Solving the importance coefficient.

Substitute the processed data into equation (11) (12), there are 58,800 sample numbers, and classify them with 5 indicators.

Calculate the proportion of all samples of the kth category in the customer experience dataset according to formula (13), substitute it into formula (11) to get the Gini value, then substitute the Gini value into formula (14) to solve the importance coefficient, and sort it from highest to lowest, and choose 5 importance indicators because of the limitation of the depth of the tree. Get the feature weights see the results of the model.

Step3: Solve the model to draw the decision tree (see the result of the model).

3.3.3 Results of the model and analysis

The results of the importance coefficients are shown in Table 3.

Table 3. Table of feature importance coefficients

mc _ 2	*mc* _ 3	*qd* _ 2	*wd* _1	*xyx* _ 0	*xyx* _1	*xyx* _ 2	*yd* _1
0.0221	0.0489	0.4303	0.3160	0.1331	0.0060	0.0243	0.01931

mc_2 and mc_3 represent the pulses of fault 2 and fault 3, qd_2 represents the cliff of fault , xyx_1 and xyx_2 represent the valid values of fault 1 and fault 2. yd_1 represents the margin of fault 1. The order from highest to lowest is:

$$qd_2 > wd_1 > yd_1 > xyx_0 > mc_3 > xyx_2 > mc_2 > xyx_1$$

Analysis of importance coefficient results. From the above characteristic importance table can be seen fault 2 cragginess indicates the degree of waveform smoothing, sensitivity to shock in the signal, the higher the importance of cragginess, indicating that there is a more obvious signal difference in the fault 2 gearbox; high importance of skewness indicates high asymmetry of the data, fault 1 gear machine there is a certain direction of friction on the data; margin shows the amplitude of the data, indicating that the magnitude of change in fault 1 is large; The effective value indicates the energy indicator of the vibration signal, and the high importance of the energy indicator of the normal gearbox indicates that the five gearboxes have some differences in the mean value.

The result of the decision tree is shown in Figure 5. If qd_2<=1161976768, it is true that the vibration signal data bifurcated to the left is no fault, and vice versa, it is bifurcated to the right for the presence of fault, and then the first split after the vibration signal data for the fault is restricted. The first layer forked to the left to whether xyx_0 <= 0.032 to divide, continue the above bifurcation rules; the first layer tends to bifurcate to the right to whether yd_1 <= 17.979 for the second division. The second layer first forked to the left is divided by mc_2<=91.314 as the division criterion, and those forking to the right are divided by xyx_0<=7.521; the second layer first to the left of the right with 11 normal samples, and the right with yd_1<=19.042 as the limiting factor. The third layer from left to right, in order, is limited by wd_0<=-167662.281, the second has 4 normal condition samples, the third limiting factor is qd_3<=69.342, the fourth has 1 faulty condition sample, the fifth limiting condition is wd_2<=566940.047, and the division continues with 65 faulty samples in the sixth. The fourth layer from left to right, except the fourth needs to be divided again, other than the results, in order, are 2 samples of normal condition, 10 samples of fault, 53 samples of normal, 1 sample of normal, 2 samples of fault; the fourth part needs to continue to divide qd_0<=286164160, to the left is there are 6 samples of normal, to the right is 1 sample of fault.

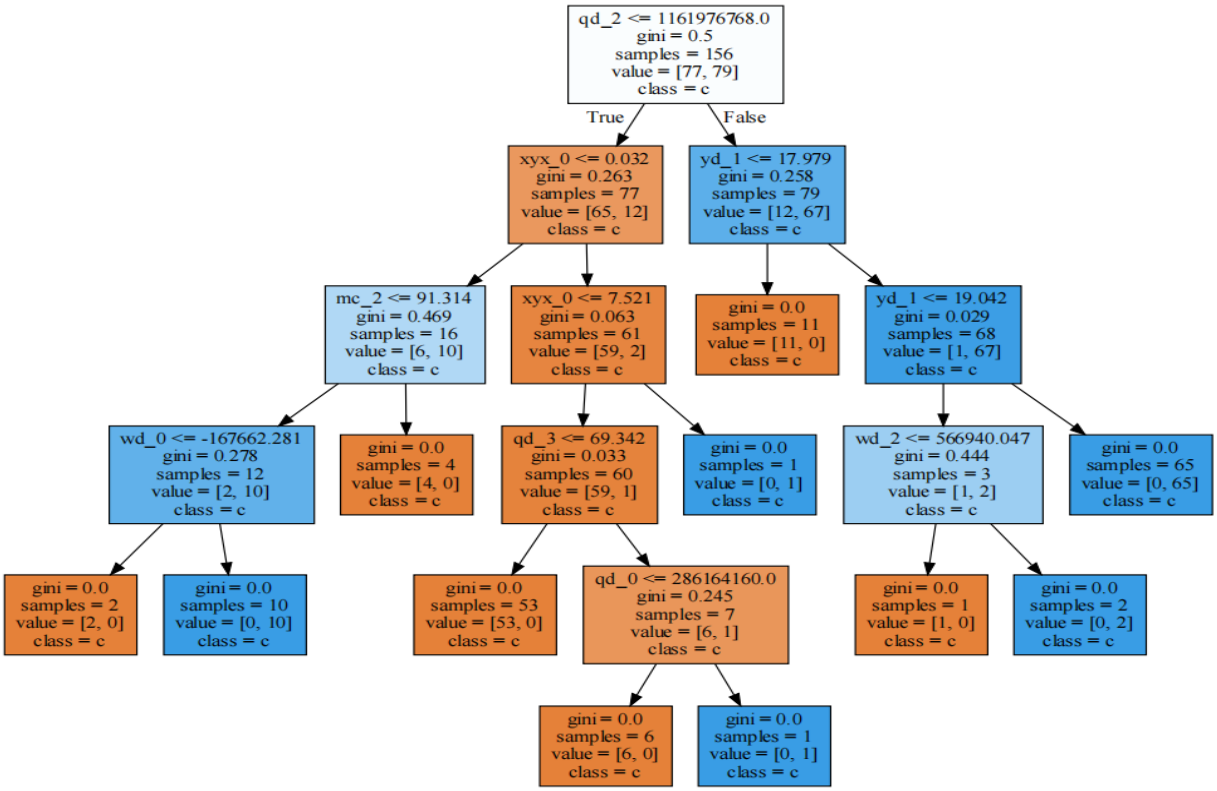


Figure 5. Fault detection decision tree

Note: Gini stands for Gini value. Simple represents the number of vibration signals of this class. Value[1,0] represents the number of normal vibration signals with a result value of 1 on the left side of the parentheses, and the number of faulty vibration signals with a result value of 0 on the right side.

2.4. Evaluation of the model

The results are checked for accuracy and completeness. Based on the prediction results, the confusion matrix of normal normal gearboxes is established. According to the definitions of the check-all rate P, check-accuracy rate R and the metric F1, the correlation indexes of the traditional method can be calculated as P = 88.9%, R = 94.12%, F1= 91.7% , with higher accuracy, and more accuracy improvement compared with the ridge regression model. Therefore, the performance of this model is better.

Table.4. Confusion Matrix

Real situation	Number of predicted versus true scenarios	
	Normal	Fault
Normal	32	4
Fault	2	30

3. Conclusions

In this paper, a decision tree model for gearbox fault detection is developed.

Compared to the ridge regression model, the decision tree model can better handle the case of insignificant differences between the original sample data because it does not require linear fitting of the data. The decision tree model is able to perform branch selection based on the importance of the features, thus better capturing the nonlinear relationships between the data.

The process of decision tree feature extraction can increase the distinguishability of the data, thus improving the accuracy of the model. Feature indicators such as RMS, impulse indicator, skewness

indicator, margin indicator and cliffness indicator are defined to help better describe the fault state of the gearbox. It is also a highly interpretable model, and the judgment process of the model can be understood by observing the branches and rules of the decision tree, which helps engineers and maintenance personnel to analyze and troubleshoot gearbox faults.

Compared with traditional manual fault diagnosis methods, automated fault detection methods based on machine learning are more efficient and cost effective. Once the model is established, rapid fault detection and diagnosis can be achieved, reducing human resource investment and maintenance costs.

In summary, the project based on gearbox fault detection has a clear relevance and feasibility. It can improve equipment reliability, reduce maintenance costs and downtime, enhance safety, and is feasible in terms of data acquisition, algorithm feasibility, technical support availability, and cost effectiveness.

4. References

[1] Li Cheng, Wang Yi,Zhang Guangyao,Qin Yi,Tang Baoping. An enhanced instantaneous angular speed estimation method by multi-harmonic time–frequency realignment for wind turbine gearbox fault diagnosis [J]. Measurement Science and Technology,2023,34(8).

[2] Lu Xi,Li Pan. Research on gearbox temperature field image fault diagnosis method based on transfer learning and deep belief network [J]. Scientific Reports,2023,13(1).

[3] Tong Shuiguang,Fu Zilong,Tong Zheming,Li Junjie,Cong Feiyun. Fault diagnosis for gearboxes based on Fourier decomposition method and resonance demodulation[J]. Journal of Zhejiang University-SCIENCE A,2023,24(5).

[4] Men Zhihui,Hu Chaoqun,Li Yong Hua,Bai Xiaoning. A hybrid intelligent gearbox fault diagnosis method based on EWCEEMD and whale optimization algorithm-optimized SVM[J]. International Journal of Structural Integrity,2023,14(2).

[5] Yin Na,Sun Dengyun,Meng Zong,Li Jimeng,Cao Lixiao,Fan Fengjie. Gearbox fault diagnosis based on generalized multivariate logarithmic regularization[J]. Measurement Science and Technology,2023,34(6).

[6] Zhang Zongyan,Weng Haohan,Zhang Tong,Chen C L Philip. A Broad Generative Network for Two-Stage Image Outpainting.[J]. IEEE transactions on neural networks and learning systems,2023,PP.

[7] Li Cheng,Wang Yi,Zhang Guangyao,Qin Yi,Tang Baoping. An enhanced instantaneous angular speed estimation method by multi-harmonic time–frequency realignment for wind turbine gearbox fault diagnosis[J]. Measurement Science and Technology,2023,34(8).

[8] Lu Xi,Li Pan. Research on gearbox temperature field image fault diagnosis method based on transfer learning and deep belief network[J]. Scientific Reports,2023,13(1).

[9] Tong Shuiguang,Fu Zilong,Tong Zheming,Li Junjie,Cong Feiyun. Fault diagnosis for gearboxes based on Fourier decomposition method and resonance demodulation[J]. Journal of Zhejiang University-SCIENCE A,2023,24(5).

[10] He Xi,Liu Jinde,Yang Chen,Jiang Gang. Predicting thermodynamic stability of magnesium alloys in machine learning[J]. Computational Materials Science,2023,223.