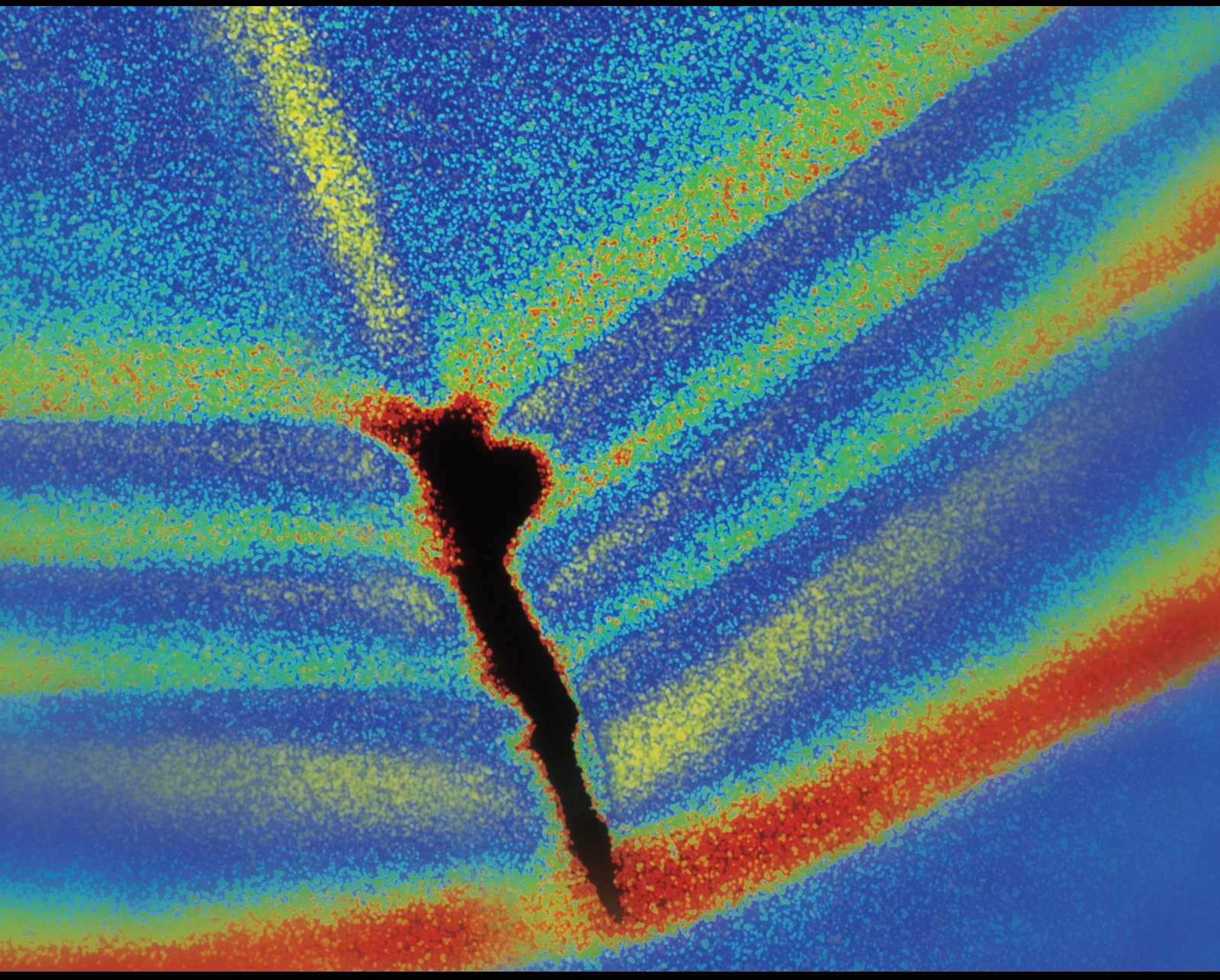
Shock and Vibration

Intelligent Feature Learning Methods for Machine Condition Monitoring 2021

Lead Guest Editor: Changqing Shen
Guest Editors: Jun Zhu and Min Xia



Intelligent Feature Learning Methods for Machine Condition Monitoring 2021

Shock and Vibration

Intelligent Feature Learning Methods for Machine Condition Monitoring 2021

Lead Guest Editor: Changqing Shen
Guest Editors: Jun Zhu and Min Xia

Contents

A Stochastic Learning Algorithm for Machine Fault Diagnosis
Zhipeng Dong, Yucheng Liu, Jianshe Kang, and Shaohui Zhang
Research Article (9 pages), Article ID 5790185, Volume 2022 (2022)

A Novel Approach of Label Construction for Predicting Remaining Useful Life of Machinery
Hailong Lin, Zihao Lei, Guangrui Wen, Xiaojun Tian, Xin Huang, Jinsong Liu, Haoxuan Zhou, and Xuefeng Chen
Research Article (14 pages), Article ID 6806319, Volume 2021 (2021)

A Simultaneous Fault Diagnosis Method Based on Cohesion Evaluation and Improved BP-MLL for Rotating Machinery
Yixuan Zhang, Rui Yang, Mengjie Huang, Yu Han, Yiqi Wang, Yun Di, Dongke Su, and Qidong Lu
Research Article (12 pages), Article ID 7469691, Volume 2021 (2021)

A New Transferable Fault Diagnosis Approach of Rotating Machinery Based on Deep Autoencoder and Dominant Features Selection under Different Operating Conditions
Fei Dong, Xiao Yu, Xinguo Shi, Ke Liu, Zhaoli Wu, and Wanli Yu
Research Article (21 pages), Article ID 7383255, Volume 2021 (2021)

Mechanical Efficiency of HMCVT under Steady-State Conditions
Guangqing Zhang, Hengtong Zhang, Yanyan Ge, Wei Qiu, Maohua Xiao, Xiaomei Xu, and Minghui Zhou
Research Article (14 pages), Article ID 4275922, Volume 2021 (2021)

Bearing Defect Detection with Unsupervised Neural Networks
Jianqiao Xu, Zhaolu Zuo, Danchao Wu, Bing Li, Xiaoni Li, and Deyi Kong
Research Article (11 pages), Article ID 9544809, Volume 2021 (2021)

Image Denoising Using Nonlocal Means with Shape-Adaptive Patches and New Weights
Chenglin Zuo, Jun Ma, Hao Xiong, and Lin Ran
Research Article (10 pages), Article ID 9532702, Volume 2021 (2021)

Hindawi
Shock and Vibration
Volume 2022, Article ID 5790185, 9 pages
https://doi.org/10.1155/2022/5790185

Research Article

A Stochastic Learning Algorithm for Machine Fault Diagnosis

Zhipeng Dong,[1] Yucheng Liu,[2,3] Jianshe Kang,[1] and Shaohui Zhang[2]

[1]*Army Engineering University of PLA, Shijiazhuang, China*
[2]*School of Mechanical Engineering, Dongguan University of Technology, Dongguan 523808, China*
[3]*College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China*

Correspondence should be addressed to Shaohui Zhang; zhangsh@dgut.edu.cn

Received 18 October 2021; Accepted 7 January 2022; Published 18 February 2022

Academic Editor: Changqing Shen

Industrial big data bring a large number of high-dimensional sample datasets. Although a deep learning network can well mine the internal nonlinear structure of the dataset, the construction of the deep learning model requires a lot of computing time and hardware facilities. At the same time, there are some nonlinear problems such as noise and fluctuation in industrial data, which make the deep architecture extremely complex and the recognition accuracy of the diagnosis model difficult to guarantee. To solve this problem, a new method, named stochastic learning algorithm (SL), is proposed in this paper for dimension reduction. The proposed method consists of three steps: firstly, to increase the computational efficiency of the model, the dimension of the high-dimensional data is reduced by establishing a random matrix; secondly, for enhancing the clustering influence of the sample, the input data are enhanced by feature processing; thirdly, to make the clustering effects more pronounced, the noise and interference of the data need to be processed, and the singularity value denoising method is used to denoise training data and test data. To further prove the superiority of the SL method, we conducted two sets of experiments on the wind turbine gearbox and the benchmark dataset. It can be seen from the experimental results that the SL method not only improves the classification accuracy but also reduces the computational burden.

1. Introduction

Deep neural networks have extensive applications in artificial intelligence mainly including computer vision [1–8], speech recognition [9–13], medical detection [14–19], and mechanical fault diagnosis [20–28]. Compared with human ability, the DNN model is more capable of solving this complicated problem. But it also has certain challenges. For example, to complete different tasks effectively, different DNN models need to be trained, tuning the parameters through repeating the trial and error, which optimizes the model structure [29]. Therefore, training a DNN model to effectively process specific tasks takes up days or even weeks of the entire computing cluster time [30]. In addition, the parameter optimization of the DNN model not only requires high-performance GPU, TPU, and other higher computer hardware environments but also has high requirements for datasets. Therefore, applications that require high real-time performance or data samples that lack markup are not suitable [31].

In addition to deep learning methods, shallow learning algorithms (PCA, KNN, LPP, etc.) are still largely applied in the artificial intelligence area [32–37]. Although this kind of shallow learning algorithm has the advantages of simple structure, low hardware environment configuration requirements, and relatively high computational efficiency, it also has limitations that are difficult to overcome, such as classifying data containing a large number of variables and a simplified sample set, and the problem of nonlinear nature [38]. In contrast, to overcome the shortcomings of the above algorithm, random forest (RF) was proposed and validated [39, 40]. In addition, meanwhile, RF also has many other advantages [41]. For example, it is easy to understand and simple to implement, tests fast, is highly able to handle outliers and nonlinearity, and shows good performance in parallel training and big data. As a result, it is widely used in medicine, computer vision, machine learning, and other fields, which achieved great results [42–50]. However, the number of decision trees is an important parameter of RF,

which will affect the RF's classification accuracy and computational efficiency. In this paper, a better machine learning algorithm-random learning (SL) is put forward to solve this problem. SL method uses a random mapping matrix to randomly decrease the high-dimensional data's dimensionality, improving the data after the dimensionality reduction. Moreover, the feature is denoising based on SVD to improve the classification rate. Therefore, this method has outstanding advantages: (1) the sample dimension is greatly reduced after processing; (2) the calculation efficiency of the random forest is improved; and (3) the recognition effect of random learning is ensured by the reinforcement process.

In this paper, the other sections are arranged as follows: Section 2 introduces the proposed method's theoretical research; Section 3 introduces the experimental setup and analyzes the results and the experiments; and Section 4 gives the conclusion.

2. The Proposed Method

This section describes the detailed information of the proposed stochastic learning method, and the structure of the proposed model is shown in Figure 1. The strategy of stochastic learning is introduced in Section 2.1. The basics of stochastic learning used for classification are explained in detail in Section 2.2. The implementation of the stochastic learning strategy is presented in Section 2.3. Section 2.4 introduces the proposed algorithm's specific processing process.

2.1. Strategies for the Present Stochastic Learning Method. Sample size and dimensionality would affect the machine learning method's computational efficiency directly. Inspired by the extreme learning machine from the operating mechanism, the input data is randomly reduced in dimensionality to obtain low-dimensional sample data. Unlike the extreme learning machine, the samples are not simply classified, but enhanced in two steps: firstly the sample characteristics were strengthened to obtain a good sample clustering effect; secondly, the feature denoising method is used to improve the cluster. Based on these steps, the recognition accuracy as well as the calculation efficiency of the model can be improved.

2.2. Stochastic Learning Strategy

(1) The research in this part is reference [51]. For an input dataset $\mathbf{X} = [x_1, x_2, \ldots, x_m] \in \mathfrak{R}^{n\times m}$, and if the random feature extraction layer has d nodes, then $\mathbf{H} = [h_1, h_2, \ldots, h_d] \in \mathfrak{R}^{n\times d}$ is the corresponding layer's features. The randomly extracted features are expressed by the formula

$$\mathbf{H} = (\mathbf{X} + \delta)\mathbf{W}, \tag{1}$$

where δ is the biases and W represents the input weights. $\mathbf{H} = [h_1, h_2, \ldots, h_d]$ and $\boldsymbol{W}$ are produced randomly before the start of training, and they remain fixed during the training state without any iterations.

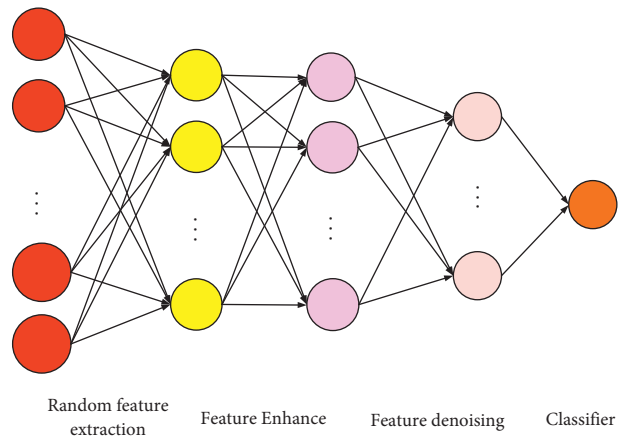


Figure 1: The proposed SL method.

(2) The research in this part is reference [51]. Since the input data seems to be nonlinear, in order to improve classification, an aggregate method of the activation function is introduced. For the random features $\mathbf{H} = [h_1, h_2, \ldots, h_d]$, through the activation function, the dataset can be transformed into

$$Q(\mathrm{H}) = \frac{1}{1 + e^{-\mathbf{H}}},$$
$$g = Q(\mathbf{H})\mathrm{W}_{im}, \tag{2}$$

where $\mathbf{W}_{im}$ represents a weight vector connecting the improved classification operation and the random feature output, g represents the better classification layer's output, and $\boldsymbol{Q}$ is an activation function. Assuming that the training samples contain c categories, q_{i} represents the *i-th* category's sample amount ($i = 1, 2, \ldots, \mathrm{c}$), and p_{i} represents the *i-th* category's centrality, then

$$p_i = \frac{1}{q_i}\sum_{j=1}^{N} d_j \begin{cases} d_j = h(j)x(j) \in i-th \text{ category}, \\ d_j = 0 \text{ otherwise}, \end{cases}$$
$$t(k) = \sum_{i=1}^{c} z_i \begin{cases} z_i = p_i t(k) \in i-th \text{ category}, \\ z_i = 0 \text{ otherwise}, \end{cases} \tag{3}$$

where the center matrix is represented by t, $t \in \mathrm{R}^{n\times d}$. In order to get the connection weight vector $\mathbf{W_{im}}$, the improved classification's output and the random features can be converted as

$$\mathbf{t} = QW_{im},$$
$$W_{im} = \left(Q^T Q\right)^{-1} Q^T \mathbf{t}. \tag{4}$$

To guarantee the mapping's stability which is between the input and the target while preventing the overfitting of the function, regularization processing

is performed. The regularization formula is as follows:

$$W_{im} = \left[Q(H)^T Q(H) + \lambda e\right]^{-1} Q(H)^T \mathbf{t}, \tag{5}$$

where e represents the unit diagonal matrix and λ represents the penalty coefficient.

Therefore, the output dataset $\boldsymbol{Y}$ is expressed as

$$Y = Q(H) W_{im}. \tag{6}$$

The above expression includes the model-building process. Regarding the test datasets, it is assumed that the input dataset is $Z = [z_1, z_2, \ldots, z_m]$. The random feature is expressed by the following formula:

$$P = (Z + \delta) W. \tag{7}$$

With regard to random features $P = [p_1, p_2, \ldots, p_d]$, through the activation function Q, the dataset can be transformed into

$$Q(\mathbf{P}) = \frac{1}{1 + e^{-\mathbf{P}}}. \tag{8}$$

For the test dataset, the output of the improved classification layer is expressed as follows:

$$\mathbf{T} = Q(\mathbf{P}) \mathbf{W_{im}}. \tag{9}$$

(3) Although the above process can reduce the impact of noise and data fluctuation on clustering results, it is impossible to avoid overfitting completely, and noise still exists in the transformation process. Therefore, training samples and test samples are denoised at the same time. After dimension reduction, the dimension of the sample is greatly reduced and the computational complexity is greatly reduced. Traditional data mining methods only rely on training samples to build models, and test samples cannot be processed.

First, based on the phase space reconstruction method, each feature vector of $\boldsymbol{Y}$ can be converted into an $n \times m$ matrix $\boldsymbol{F}$ where m represents the phase space's embedding dimension, which can be decided by the mutual information and false nearest neighbor approaches, respectively [19, 20]. Then, the SVD-based is applied for exact factorization of the matrix $\boldsymbol{F}$.

$$F = U \Lambda V^T, \tag{10}$$

where $\boldsymbol{V^T}$ represents the complex unitary matrix or an $m \times m$ real, U represents the complex unitary matrix or an $n \times n$ real, and $\boldsymbol{\Lambda}$ represents a diagonal matrix with dimension $n \times m$ in which the diagonal entries of the matrix $\boldsymbol{F}$ are singular value.

The energy concentration of features and noise is reflected by the distribution of singular values. Among them, the useful feature corresponds to a larger value, and the first k useful features are retained, while the noise corresponds to a smaller value, and the noise is set to zero, a new diagonal matrix $\boldsymbol{\Lambda'}$ can be obtained, and then no-free features can be obtained through the inverse singular value transformation.

$$F' = U \Lambda' V^T. \tag{11}$$

Finally, the no-noise features $\boldsymbol{F'}$ is transformed into the new feature by inverse space reconstruction and the new features space can be rewritten as $\boldsymbol{Y'}$.

Likewise, the clustering effect of test samples can be greatly improved by denoising test samples directly here.

2.3. Proof. Assuming the input dataset can be represented as follows:

$$\mathbf{X} = \overline{\mathbf{X}} + \Delta \mathbf{X}, \tag{12}$$

where $\overline{\mathbf{X}}$ represents the difference ideal dataset while $\Delta \mathbf{X}$ represents the perturbation dataset. Then,

$$t = \overline{\mathbf{X}}. \tag{13}$$

For the random features $\mathbf{H} = [h_1, h_2, \ldots, h_d]$, the dataset is represented as

$$\mathbf{H} = (\overline{\mathbf{X}} + \Delta \mathbf{X} + \delta) \mathbf{W} = \overline{\mathbf{H}} + \Delta \mathbf{H}, \tag{14}$$

where $\overline{\mathbf{H}}$ represents the difference ideal dataset while ΔH represents the perturbation dataset.

For the random features $\mathbf{H} = [h_1, h_2, \ldots, h_d]$, through the activation function, the dataset can be converted to

$$Q(\mathbf{H}) = Q(\overline{\mathbf{H}} + \Delta \mathbf{H}). \tag{15}$$

As described in (5), the relationship is expressed by the formula as follows:

$$t = Q(\mathbf{H}) \mathbf{W_{im}} = \overline{\mathbf{X}}. \tag{16}$$

Thus,

$$\mathbf{X} = \overline{\mathbf{X}} + \Delta \mathbf{X} \xrightarrow{W} \mathbf{H} \xrightarrow{Q} Q(\mathbf{H}) \xrightarrow{\mathrm{W_{im}}} \overline{\mathbf{X}}. \tag{17}$$

That is

$$\Delta \mathbf{X} \xrightarrow{W} \mathbf{H} \xrightarrow{Q} Q(\mathbf{H}) \xrightarrow{\mathrm{W_{im}}} 0. \tag{18}$$

End.

2.4. Detailed Description of the Proposed Algorithm. First, reduce the dimensionality of the input dataset, then optimize the transformation by improving the classification, and then improve the internal class. Relying on the artificial intelligence network driven by big data can effectively increase the fault diagnosis's recognition accuracy and reduce the calculation time, improve the calculation efficiency, and overcome the shortcomings of depending on the physical field's knowledge to extract features manually.

On the basis of the above ideas, the schematic diagram of the proposed SL method in actual fault diagnosis is shown in Figure 2. Here are the concrete steps:

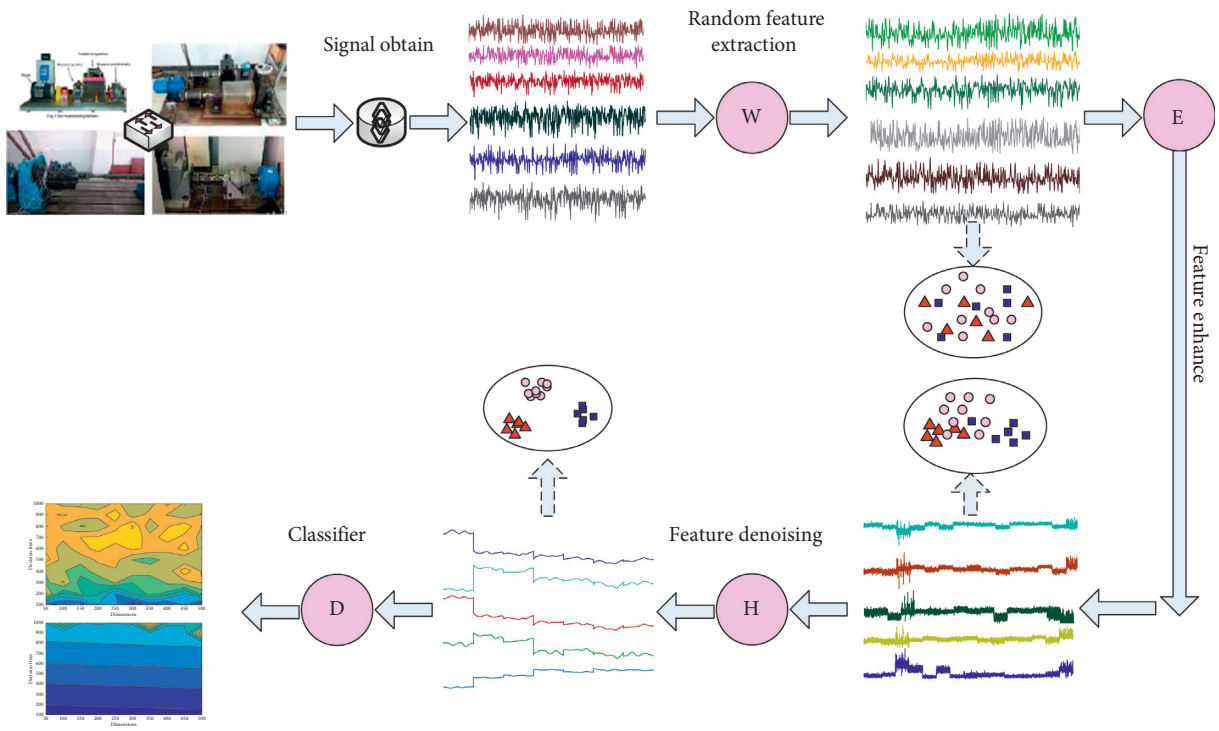


Figure 2: Overview of the proposed SL method.

Step 1: install the sensor at the position where the vibration is most direct, conduct experiments under different working conditions, and collect raw data

Step 2: process the sensor's full-channel data to generate sample dataset and corresponding labels required for training and testing

Step 3: the training dataset is used to train the SL model

Substep 1: reduce the dimensionality of the input dataset to a low dimension by using the method of random feature extraction

Substep 2: Aggregate nonlinear mapping matrices using random features extracted from samples and sample informative labels to shorten the distance between samples of the same type

Substep 3: the training samples after improve classification was denoised by SVD

Substep 4: input the information labels and training samples in order to build the classifiers model

Step 4: the testing dataset is used to test the trained SL model

Step 5: finally, the fault diagnosis accuracy is outputted

3. Stochastic Learning Algorithm for Condition Recognition

Fault diagnosis experiments are carried out based on the wind turbine gearbox and bearing benchmark dataset, respectively, to verify the proposed SL method's effectiveness. The experimental results of the two experiments are analyzed to verify whether the proposed method can increase the calculation efficiency and diagnosis accuracy.

In the experiments, determine the effect of model parameters (such as the number, step size, and dimension of decision trees) of the proposed stochastic learning method on computation time and classification accuracy is determined. The decision trees range from 100 to 500, the step size by 50, and the dimensions are set according to the input dataset.

3.1. Fault Diagnosis for Wind Turbine Gearbox

3.1.1. Failure Experiment Setup. We conduct experiments on the transmission platform that is named DDS which is designed by SpectraQuest Inc. (company website "http://www.pinxuntech.com/"), and the experimental transmission is shown in Figure 3. The experimental device mostly consists of a drive motor, a two-stage planetary gearbox, two grade parallel shaft gearboxes, and a magnetic powder brake. In the experiment, four typical gear failures were studied, namely, surface wear, tooth cracks, chipped tooth, and missing tooth. At the same time, in order to ensure the unity of the experiment, the transmission platform adopts ordinary gears for the experiment. The four most typical gear failures are discussed, including surface wear, cracked teeth, chipped teeth, and missing teeth.

The main component of the drivetrain diagnostics simulator is a gearbox, and it is also the place where the drivetrain diagnosis simulator is prone to failure. In the experiment, the secondary sun gear of the planetary gearbox was diagnosed by using the acceleration sensor to collect the

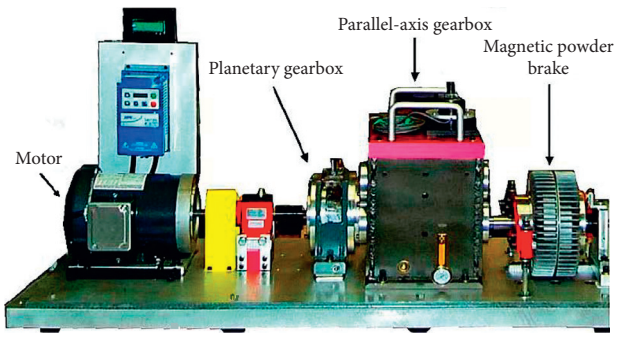


Figure 3: The experiment platform.

vibration signals of different fault types of faults in the transmission process. Table 1 shows the fault state settings and working condition settings of the gearbox.

During the experiment, a total of 5 patterns were set up, 1 normal pattern and 4 different failure patterns. The vibration signal in each pattern was collected by the acceleration sensor. We performed a wavelet transform on the collected vibration signals to extract impulse signals. According to the literature [52], features are extracted from the original signal and impulse signal, respectively, and a total of 50 features are counted. We can obtain 2000×50 feature samples in these experiments. During model training, the dataset is divided, 50% of the samples are applied for model training, and 50% of the samples are applied for testing.

The experimental outcomes of the SL algorithm are indicated in Figure 4. We conclude from Figure 4(a) that the data dimensions and the decision trees quantity will influence the SL algorithm's classification accuracy, but the sensitivity of the recognition accuracy to the dimensions is below the number of decision trees. Figure 4(b) illustrates that decision trees quantity mainly influences the calculation time of the SL algorithm, and the data dimension has a small effect on the calculation time. Therefore, in order to ensure that the algorithm model has high classification accuracy and, at the same time, high computational efficiency, the parameters of the SL algorithm should be set to multiple decision trees and fewer dimensions. From the yellow dot area in Figure 4(b), we can see that the SL algorithm can classify up to 92%, and its corresponding calculation time is relatively low, which is 15 seconds.

3.1.2. Comparison of Different Methods of Diagnosis. The data collected in the experiment is used for different algorithm models for fault diagnosis, the recognition accuracy of different methods is compared, and the diagnostic performance of the proposed SL method is further discussed. The classification accuracy and calculation time of different methods are shown in Table 2. In Table 2, the average recognition accuracy rates of SVM, ESN, SAE + ESN, and SAE + Softmax are 22.10%, 76.98%, 44.73%, and 62.00%, respectively. It is unacceptable in actual engineering applications. The corresponding std is 0, 0.679, 1.988, and 2.770, respectively. In addition, as a shallow learning algorithm, the SVM method must have a lower computational efficiency than the deep learning algorithm SAE. The calculation time of SAE + ESN and SAE + Softmax exceeds 12 s, while the calculation time of SVM and ESN is 0.96 s and 0.68 s, respectively. The classification accuracy and calculation time that are about the SL method are shown in Figure 5. The blue in the picture shows the recognition accuracy, and the red shows the calculation time. With a different number of decision trees, the identification accuracy obtained by the SL method is more than 90%, ranging from 90.08% to 90.28%, far exceeding the recognition accuracy of other algorithms. The greater the number of decision trees, the longer the calculation time. The calculation time of the SL method ranges from 3.25 s to 14.55 s. From Table 2 and Figure 5, compared with other methods, the SL method has the highest computational efficiency. The proposed SL method is preceded by other methods in recognition accuracy and calculation time in Table 2 and Figure 5.

3.2. Fault Diagnosis for the Benchmark Dataset

3.2.1. Failure Experiment Setup. To further approve not only the effectiveness but also the superiority of the SL method, the benchmark data is used to evaluate the proposed SL model. The benchmark dataset is a rolling bearing fault dataset offered with the Case Western Reserve University Bearing Data Center (CWRU) [53]. For the benchmark dataset, we consider the bearings under normal conditions and the bearings with 3 different faults, including outer ring faults, memory faults, and ball faults. Each fault includes 3 defect levels (0.18 mm, 0.36 mm, and 0.53 mm wide grooves). During the course of this experiment, the sensor sampling frequency was set to 48 kHz, and the data were collected under a load of "1," "2," and "3" HP and the acquisition time of each group of vibration signals last for 10 s. A total of 10 sets of bearing data in different states were obtained in the experiment, including 1 set of healthy bearings and 9 sets of faulty bearings, called NR and fault 1 to fault 9.

The dataset is extracted from the original signal, 800 small sample sets are collected for each load, each small sample set has a total of 300 data, each group of bearing data has 2400 small sample sets, and the benchmark dataset uses a total of 24,000 small samples set. Therefore, the size of the original data in the high-dimensional space is 24000×300, and the data plan used to train the model is the same as in Section 3.1.

Figure 6 has displayed the experimental results of the SL algorithm. As shown in Figure 6, the SL algorithm has a recognition accuracy of over 98% under the conditions of different numbers of decision trees and diverse data dimensions, and the algorithm is not sensitive to changes in dimensions and decision trees. When the amount of decision trees is 400 and the dimension is 100, the classification accuracy of the SL method almost reaches 100% at the highest. Generally speaking, the identification accuracy of the algorithm is influenced by input dimensions and decision trees quantity, but the former is more influential, and the recognition accuracy is positively correlated with the amount of decision trees.

Table 1: Planetary gearbox conditions.

Failure pattern	Failure type	Load (Nm)
A	Health	0/1.4/2.8/25.2
B	Surface wear	0/1.4/2.8/25.2
C	Cracked tooth	0/1.4/2.8/25.2
D	Chipped tooth	0/1.4/2.8/25.2
E	Missing tooth	0/1.4/2.8/25.2

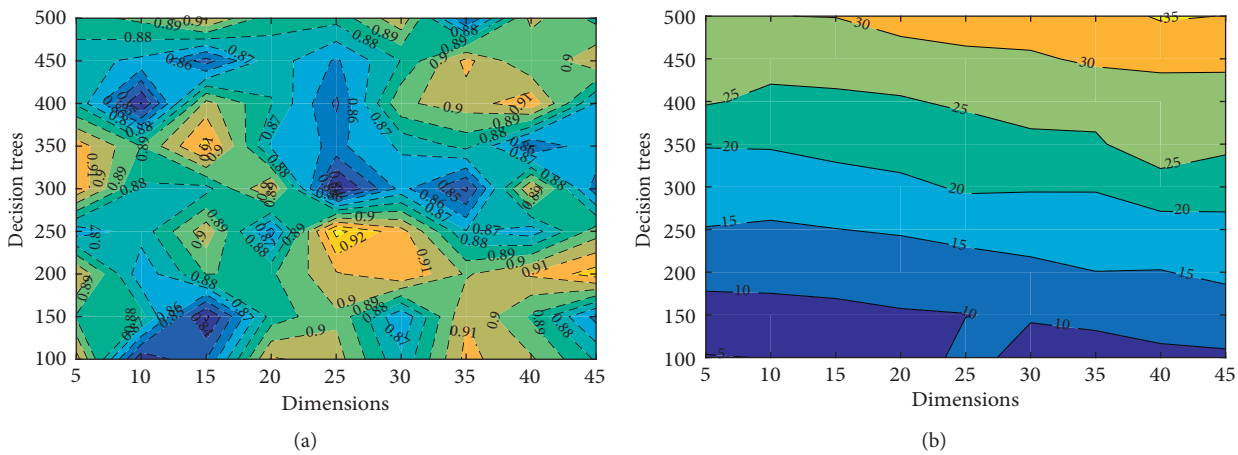


Figure 4: The experiment results for wind turbine gearbox based on SL classifier. (a) Recognition accuracy rate; (b) computing time.

Table 2: Diagnosis results of different approaches.

Approach	Accuracy		Time (s)
	avg (%)	std	
SVM	22.10	0	0.96
ESN	76.98	0.679	0.68
SAE + ESN	44.73	1.988	12.64
SAE + Softmax	62.00	2.770	12.15

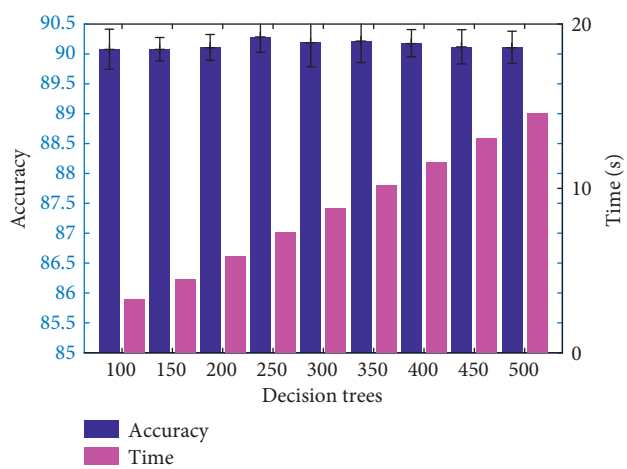


Figure 5: The experiment results for the wind turbine gearbox based on the SL classifier.

3.2.2. Comparison of Different Methods of Diagnosis. As described above, the corresponding comparison results are exhibited in Table 3 and Figure 7. The average classification accuracy rate of SVM, ESN, SAE + ESN, and SAE + Softmax is 26.28% ± 0.673, 39.07% ± 1.27, 37.14% ± 0.902, and 10.02% ± 0.0289, respectively. Furthermore, the SVM and ESN method must have a lower computational efficiency than the deep learning algorithm SAE. The computing times of SAE + ESN and SAE + Softmax are about 290 s, while that of SVM and ESN is 171.55 s and 7.95 s, respectively. As

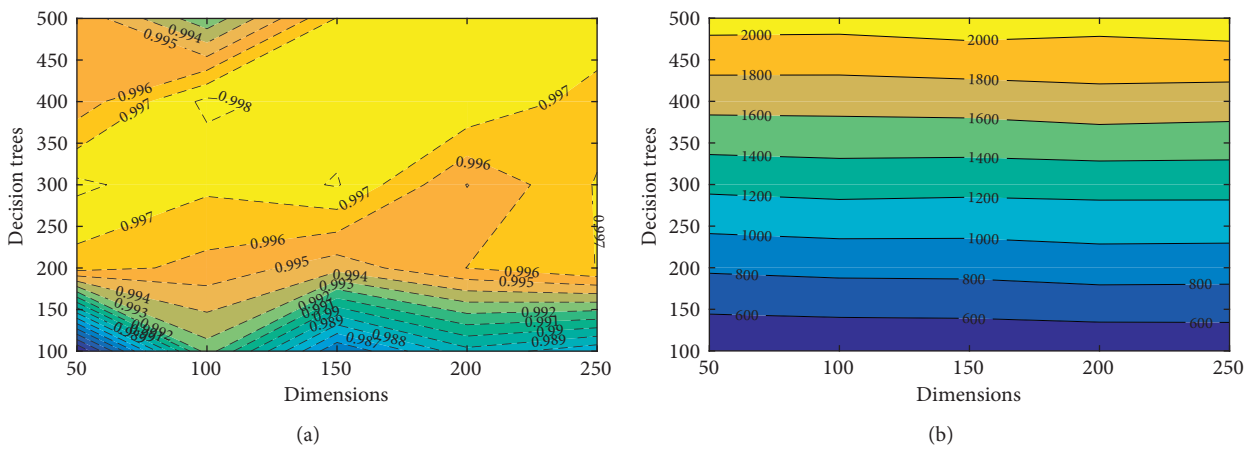


Figure 6: The experiment results for benchmark dataset based on SL classifier. (a) Recognition accuracy rate; (b) computing time.

Table 3: Diagnosis results of different approaches.

Approach	Accuracy		Time (s)
	avg (%)	std	
SVM	26.28	0.673	171.55
ESN	39.07	1.27	7.95
SAE + ESN	37.14	0.902	297.53
SAE + Softmax	10.02	0.0289	292.90

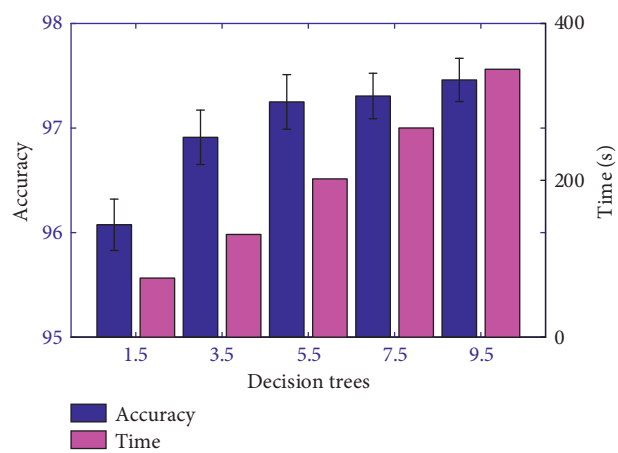


Figure 7: The experiment results for benchmark dataset based on RF classifier.

shown in Figure 7, the SL method classification accuracy and calculation time under different amounts of decision trees. The blue in the figure shows the recognition accuracy, and the red shows the calculation time. Figure 7 shows that the average recognition accuracy rate is 96.08% to 97.46% and increasing the number of decision trees can improve the classification accuracy. Meanwhile, the calculation time is also increasing. The calculation time of the SL method ranges from 75.36 s to 341.7 s.

4. Conclusions

This paper proposes a random learning dimensionality reduction algorithm method and uses it for machine fault state recognition. The classification accuracy and operation efficiency of the dimension reduction algorithm is affected by the size of the dimensions. In the stochastic learning method (SL), random feature extraction is performed on the input high-dimensional data through a random mapping matrix. After random feature extraction, low-dimensional feature data will be obtained for model training. Therefore, after feature extraction, the input SL model sample dimension is largely decreased, and the calculation efficiency is greatly improved. The use of information-enhanced data for training guarantees the classification effect of the SL algorithm.

Data Availability

The data used to support the findings of this study are obtained from previously reported studies. These prior

studies (and datasets) are cited at relevant places within the text as references.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported in part by the special projects in Key Fields of Ordinary Colleges and Universities in Guangdong Province (2020ZDZX3029) and Dongguan Science and Technology Commissioner Project (20201800500212and 20201800500282).

References

[1] Q.-K. Tran and S.-k. Song, "Computer vision in precipitation nowcasting: applying image quality assessment metrics for training deep neural networks," *Atmosphere*, vol. 10, no. 5, p. 244, 2019.

[2] R. Patel and S. Patel, "A comprehensive study of applying convolutional neural network for computer vision," *International Journal of Advanced Science and Technology*, vol. 29, no. 6s, pp. 2161–2174, 2020.

[3] A. Z. da Costa, H. E. H. Figueroa, and J. A. Fracarolli, "Computer vision based detection of external defects on tomatoes using deep learning," *Biosystems Engineering*, vol. 190, pp. 131–144, 2020.

[4] C.-T. Wu, P. van Beek, P. Schmidt, J. P. Moreira, and T. R. Gardos, "Evaluation of semi-frozen semi-fixed neural network for efficient computer vision inference," *Electronic Imaging*, vol. 2021, no. 17, p. 213, 2021.

[5] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable bayesian deep learning methods for robust computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 318-319, Seattle, Washington D. C., USA, June 2020.

[6] R. Yang, S. K. Singh, M. Tavakkoli et al., "CNN-LSTM deep learning architecture for computer vision-based modal frequency detection," *Mechanical Systems and Signal Processing*, vol. 144, Article ID 106885, 2020.

[7] N. Hussain, M. A. Khan, M. Sharif et al., "A Deep Neural Network and Classical Features Based Scheme for Objects Recognition: An Application for Machine inspection," *Multimedia Tools and Applications*, pp. 1–23, 2020.

[8] G. Mukherjee, B. Tudu, and A. Chatterjee, "A convolutional neural network-driven computer vision system toward identification of species and maturity stage of medicinal leaves: case studies with Neem, Tulsi and Kalmegh leaves," *Soft Computing*, vol. 25, pp. 1–20, 2021.

[9] G. Li, S. Liang, S. Nie, W. Liu, and Z. Yang, "Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition," *Neural Networks*, vol. 141, pp. 225–237, 2021.

[10] C. H. H. Yang, J. Qi, S. Y. C. Chen et al., "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6523–6527, IEEE, Toronto, ON, Canada, June 2021.

[11] X. Cui, W. Zhang, U. Finkler, G. Saon, M. Picheny, and D. Kung, "Distributed training of deep neural network acoustic models for automatic speech recognition: a comparison of current training strategies," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 39–49, 2020.

[12] Z. Wu, D. Zhao, Q. Liang, Y. Jiahui, G. Anmol, and P. Ruoming, "Dynamic sparsity neural networks for automatic speech recognition," in *Proceedings of the ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6014–6018, IEEE, Toronto, ON, Canada, June 2021.

[13] J. Guglani and A. N. Mishra, "DNN based continuous speech recognition system of Punjabi language on Kaldi toolkit," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 41–45, 2021.

[14] J. Li, Y. Pei, A. Yasin, A. Yasin, S. Ali, and T. Mahmood, "Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network," *Sensors*, vol. 21, no. 14, p. 4854, 2021.

[15] K. M. Black, H. Law, A. Aldoukhi, J. Deng, and K. R. Ghani, "Deep Learning Computer Vision Algorithm for Detecting Kidney Stone composition," *BJU Int*, vol. 125, 2020.

[16] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine*, vol. 196, Article ID 105581, 2020.

[17] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, and K. C. Santosh, "Colorectal histology tumor detection using ensemble deep neural network," *Engineering Applications of Artificial Intelligence*, vol. 100, Article ID 104202, 2021.

[18] S. Albahli, "A deep neural network to distinguish covid-19 from other chest diseases using x-ray images," *Current medical imaging*, vol. 17, no. 1, pp. 109–119, 2021.

[19] S. Sabut, O. Pandey, B. S. P. Mishra, and M. Monalisa, "Detection of ventricular arrhythmia using hybrid time–frequency-based features and deep neural network," *Physical and Engineering Sciences in Medicine*, vol. 44, no. 1, pp. 135–145, 2021.

[20] H. Han, L. Xu, X. Cui, and Y. Fan, "Novel chiller fault diagnosis using deep neural network (DNN) with simulated annealing (SA)," *International Journal of Refrigeration*, vol. 121, pp. 269–278, 2021.

[21] C. D. Nguyen, A. E. Prosvirin, C. H. Kim, and J. Kim, "Construction of a sensitive and speed invariant gearbox fault diagnosis model using an incorporated utilizing adaptive noise control and a stacked sparse autoencoder-based deep neural network," *Sensors*, vol. 21, no. 1, p. 18, 2021.

[22] F. Zhou, T. Sun, X. Hu, and T. Wang, "A sparse denoising deep neural network for improving fault diagnosis performance," *Signal, Image and Video Processing*, vol. 15, pp. 1–10, 2021.

[23] D. T. Hoang, X. T. Tran, M. Van, and H. J. Kang, "A deep neural network-based feature fusion for bearing fault diagnosis," *Sensors*, vol. 21, no. 1, p. 244, 2021.

[24] K. Zhou, C. Yang, J. Liu, and Q. Xu, "Dynamic Graph-Based Feature Learning with Few Edges Considering Noisy Samples for Rotating Machinery Fault diagnosis," *IEEE Transactions on Industrial Electronics*, 2021.

[25] C. Yang, K. Zhou, and J. Liu, "SuperGraph: Spatial-Temporal Graph-Based Feature Extraction for Rotating Machinery diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 69, 2021.

[26] X. Li, J. Cheng, H. Shao, and B. Cai, "A fusion CWSMM-based framework for rotating machinery fault diagnosis under strong interference and imbalanced case," *IEEE Transactions on Industrial Informatics*, 2021.

[27] Z. He, H. Shao, Z. Ding, and H. Jiang, "Modified Deep Auto-Encoder Driven by Multi-Source Parameters for Fault Transfer Prognosis of aero-engine," *IEEE Transactions on Industrial Electronics*, vol. 69, 2021.

[28] C. Wang and Z. Xu, "An Intelligent Fault Diagnosis Model Based on Deep Neural Network for Few-Shot Fault diagnosis," *Neurocomputing*, vol. 456, 2021.

[29] H. Tong, R. C. Qiu, D. Zhang, H. Yang, Q. Ding, and X. Shi, "Detection and classification of transmission line transient faults based on graph convolutional neural network," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 3, pp. 456–471, 2021.

[30] M. A. Asghar, M. J. Khan, M. Rizwan, R. M. Mehmood, and S.-H. Kim, "An innovative multi-model neural network approach for feature selection in emotion recognition using deep feature clustering," *Sensors*, vol. 20, no. 13, p. 3765, 2020.

[31] X. Zhao, M. Jia, and Z. Liu, "Semisupervised graph convolution deep belief network for fault diagnosis of electormechanical system with limited labeled data," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5450–5460, 2020.

[32] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," *IETE Technical Review*, vol. 38, no. 4, pp. 377–396, 2021.

[33] T. R. Gadekallu, D. S. Rajput, M. P. K. Reddy et al., "A novel PCA-whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1383–1396, 2021.

[34] M. P. Uddin, M. A. Mamun, M. I. Afjal, and M. A. Hossain, "Information-theoretic feature selection with segmentation-based folded principal component analysis (PCA) for hyperspectral image classification," *International Journal of Remote Sensing*, vol. 42, no. 1, pp. 286–321, 2021.

[35] R. Ran, Y. Ren, S. Zhang, and B. Fang, "A novel discriminant locality preserving projections method," *Journal of Mathematical Imaging and Vision*, vol. 63, no. 5, pp. 541–554, 2021.

[36] Y.-L. He, Y. Zhao, X. Hu, X.-N. Yan, Q.-X. Zhu, and Y. Xu, "Fault diagnosis using novel AdaBoost based discriminant locality preserving projection with resamples," *Engineering Applications of Artificial Intelligence*, vol. 91, Article ID 103631, 2020.

[37] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," *Knowledge-Based Systems*, vol. 205, Article ID 106270, 2020.

[38] Y. Zhou, K. Xu, F. He, and D. He, "Nonlinear fault detection for batch processes via improved chordal kernel tensor locality preserving projections," *Control Engineering Practice*, vol. 101, Article ID 104514, 2020.

[39] B. P. O. Lovatti, M. H. C. Nascimento, Á. C. Neto, E. V. R. Castro, and P. R. Filgueiras, "Use of Random forest in the identification of important variables," *Microchemical Journal*, vol. 145, pp. 1129–1134, 2019.

[40] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134, pp. 93–101, 2019.

[41] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, Article ID 102051, 2020.

[42] C. Iwendi, A. K. Bashir, A. Peshkar et al., "COVID-19 patient health prediction using boosted random forest algorithm," *Frontiers in Public Health*, vol. 8, p. 357, 2020.

[43] J. Li, Y. Tian, Y. Zhu et al., "A multicenter random forest model for effective prognosis prediction in collaborative clinical research network," *Artificial Intelligence in Medicine*, vol. 103, Article ID 101814, 2020.

[44] G. L. Watson, D. Xiong, L. Zhang et al., "Fusing a Bayesian case velocity model with random forest for predicting COVID-19 in the US," 2020.

[45] D. V. Urista, D. B. Carrué, I. Otero et al., "Prediction of antimalarial drug-decorated nanoparticle delivery systems with random forest models," *Biology*, vol. 9, no. 8, p. 198, 2020.

[46] Y. Chen, W. Zheng, W. Li, and Y. Huang, "Large group activity security risk assessment and risk early warning based on random forest algorithm," *Pattern Recognition Letters*, vol. 144, pp. 1–5, 2021.

[47] D. Sun, S. Shi, H. Wen, J. Xu, X. Zhou, and J. Wu, "A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping," *Geomorphology*, vol. 379, Article ID 107623, 2021.

[48] L. Dong, H. Du, F. Mao et al., "Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—subtropical area for example," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 113–128, 2019.

[49] O. Guehairia, A. Ouamane, F. Dornaika, and A. Taleb-Ahmed, "Feature fusion via Deep Random Forest for facial age estimation," *Neural Networks*, vol. 130, pp. 238–252, 2020.

[50] J. Upadhyay, A. Rawat, D. Deb, V. Muresan, and M.-L. Unguresan, "An RSSI-based localization, path planning and computer vision-based decision making robotic system," *Electronics*, vol. 9, no. 8, p. 1326, 2020.

[51] J. Luo, Y. Liu, S. Zhang, and J. Liang, "Extreme random forest method for machine fault classification," *Measurement Science and Technology*, vol. 32, no. 11, Article ID 114006, 2021.

[52] X. Wang, Y. Qin, Y. Wang, S. Xiang, and H. Chen, "Reltanh: an activation function with vanishing gradient resistance for sae-based dnns and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 363, 2019.

[53] K. A. Loparo, "Bearing Data center, Case Western Reserve University," 2014, http://csegroups.case.edu/bearingdatacenter/pages/download-data-file.

Hindawi
Shock and Vibration
Volume 2021, Article ID 6806319, 14 pages
https://doi.org/10.1155/2021/6806319

Research Article

A Novel Approach of Label Construction for Predicting Remaining Useful Life of Machinery

Hailong Lin,[1] **Zihao Lei**,[2] **Guangrui Wen**,[2] **Xiaojun Tian**,[1] **Xin Huang**,[2] **Jinsong Liu**,[1] **Haoxuan Zhou**,[2] **and Xuefeng Chen**[2]

[1]*SDIC Biotechnology Investment Co., Ltd., Beijing 100034, China*
[2]*School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China*

Correspondence should be addressed to Zihao Lei; zihao_lei@163.com

Received 14 September 2021; Accepted 6 December 2021; Published 18 December 2021

Academic Editor: Changqing Shen

Rolling bearings are key components of rotating machinery, and predicting the remaining useful life (RUL) is of great significance in practical industrial scenarios and is being increasingly studied. A precise and reliable remaining useful life prediction result provides valuable information for decision-makers, which is essential to ensure the safety and reliability of mechanical systems. Generally, the RUL label is considered to be an ideal life curve, which is the benchmark for RUL prediction. However, the existing label construction methods make more use of expert experience and seldom mine knowledge from data and combine experience to assist in constructing a health index (HI). In this paper, a novel and simple approach of label construction is proposed for predicting the RUL accurately. More specifically, the degradation index of the multiscale frequency domain is first extracted. Furthermore, the fuzzy C-means (FCM) algorithm is innovatively used to divide the degradation data into several stages to obtain the turning point of degradation. Then, a nonlinear degradation index, the RUL label with the turning point, was constructed based on principal component analysis (PCA). Finally, the recurrent neural network (RNN) is used for prediction and verification. In order to verify the effectiveness of the proposed approach, two different bearing lifecycle datasets are gathered and analyzed. The analysis result confirms that the proposed method is able to achieve a better performance, which outperforms some existing methods.

1. Introduction

Rolling bearings are widely used in various mechanical systems as one of the most critical components. Among them, the failure of rolling bearings is one of the most important causes of mechanical system failure [1]. Therefore, the diagnosis and prognosis of bearings play an important role in the performance of mechanical equipment [2–4]. Predicting the RUL of bearings is of great importance to prevent sudden failures in mechanical systems and has also received much attention as a key issue in prognostics and health management (PHM) [5–8].

In general, RUL prediction methods could be mainly classified into model-based, data-driven, and hybrid methods [9]. In recent years, more and more data-driven methods have been proposed for RUL prediction. Lei et al. divided the data-driven RUL prediction into four main steps, including data acquisition, HI construction, health stage division, and RUL prediction [10].

The RUL label is considered to be the ideal life curve of the equipment, which means the remaining useful life and corresponds to each operating cycle [11]. As for RUL prediction of rolling bearings, the RUL label is commonly regarded as the benchmark of accuracy evaluation for prediction results [12]. As research on data-driven prediction approaches continues to advance, the RUL label leaves more significant impacts on the model training process. Since the prediction model is also a kind of neural network,

its parameters are also obtained by satisfying the following conditions: First, the parameters are initialized, and then the predicted label is obtained by forward propagation. Next, the loss between the predicted label and the RUL label is calculated, and finally, the parameters are updated by back-propagation gradient descent. Therefore, reasonable label plays a critical role in RUL prediction of rolling bearings and greatly affects the accuracy and generalization ability of prediction models [13, 14].

According to previous studies, for the construction of the RUL label, the following three methods are mainly classified: (1) Failure determination based on the fault threshold. As the most common method in early RUL prediction studies, the maximum allowable vibration value or experienced value is usually considered the failure threshold for real situations [15–19], and it is clear to determine the failure time with fixed thresholds. However, it lacks tolerance to sudden noise in operating conditions and is difficult to obtain an ideal fault threshold in advance for a new component. (2) The RUL label based on the ideal degradation curve. In [11, 12, 20, 21], linear function and improved piecewise function were used to fit the degradation curve as much as possible. Among them, when a linear function is used, the RUL value decreases linearly with the time period. When the piecewise function is used, the RUL value remains unchanged during the previous period and then linearly decreases. In different degradation stages, the degradation rate will be different, which is then reflected in the slope of the degradation curve. The above two labels are the most widely used in the methods of RUL prediction. The key point is to find the transition time of different failure degrees, that is, the turning point of degradation, and set a more ideal RUL value. (3) The RUL label based on the degradation index. Some recent research studies [22–25] use one specific index to construct the life curve and evaluate the bearing degradation state, such as kurtosis, skewness, and other common statistical features. However, the single indicator is difficult to contain enough information and can easily be affected by fluctuations, and this is apparently contrary to the stable and monotonous performance of RUL label.

There is no doubt that the construction of RUL labels is crucial for RUL prediction because model training and result evaluation of RUL prediction require ideal life curves as a benchmark, yet most of the actual industrial field rolling bearing-monitoring data are unlabeled data. As an important part of RUL prediction, it relies more on expert experience, which represents the current linear function or piecewise linear function, rather than constructing the RUL label based on the knowledge mined by the data itself [26]. And this is obviously insufficient for PHM in the era of big data for the mining of knowledge in the data, especially when predictive maintenance is driven by the rapid development of big data [27, 28]. In addition, it can impose limitations on the prediction accuracy in practical applications. Despite these strong motivations for research, there are few studies and certain gaps on how to construct a more reasonable RUL label, which is completely incompatible with its important status of predictive maintenance.

Considering the problems mentioned above, a novel and simple RUL label construction approach is proposed in this paper and the accurate RUL prediction of rolling bearings is achieved based on the RNN. First, ensemble empirical mode decomposition (EEMD) is used to decompose signals into several intrinsic mode functions (IMFs). The frequency-domain features and energy of the IMFs are extracted and fused to obtain the degradation index using PCA. Then, the FCM algorithm is applied to detect the turning points between normal condition, slight fault, and heavy fault. It is rather remarkable that the new RUL label is constructed based on the degradation index and the known turning points. To guarantee the robustness of the RUL label, the anomalous jump points of the degradation index are further eliminated using the linear regression method. Finally, to verify the effectiveness and superiority of the proposed method, the simplest recurrent neural network is constructed for RUL prediction. And the major contributions of this work can be summarized as follows:

(1) An improved life label is proposed based on the health index and turning points of failure stages, which provides a new idea to improve the existing methods of constructing the RUL label based on human experience or single health index.

(2) The proposed approach is capable of adaptively constructing the novel label that reflects changes in the rate of degradation at different stages while being more suitable for practical applications. Moreover, the construction of new label relies on the knowledge mined from the data rather than just the experience of experts.

(3) It provides the solution for researchers to construct the RUL label for new equipment in time once the fault occurs. Therefore, it is also promising to be applied in online RUL prediction. Some experimental results also confirm the effectiveness and superiority of the proposed method.

The rest of this paper is organized as follows: Basic methods including the RNN model and FCM algorithm are introduced in Section 2. In Section 3, the principle and schematic diagram of the proposed method are presented. Section 4 shows the results of experiments on two datasets, XJTU-SY bearing data and IMS bearing data. Finally, the summary and conclusions are given in Section 5.

2. Theoretical Background

2.1. RNN. Because of the outstanding ability to handle time-series data, the RNN model is suitable to be used in RUL prediction. There are many developed versions of RNN, but the basic principles of these networks remain unchanged. Thus, the classical structure of the RNN model is applied for the prediction of RUL in this work.

The classical RNN model is shown in Figure 1. It is assumed that there is a certain part of the equipment, which

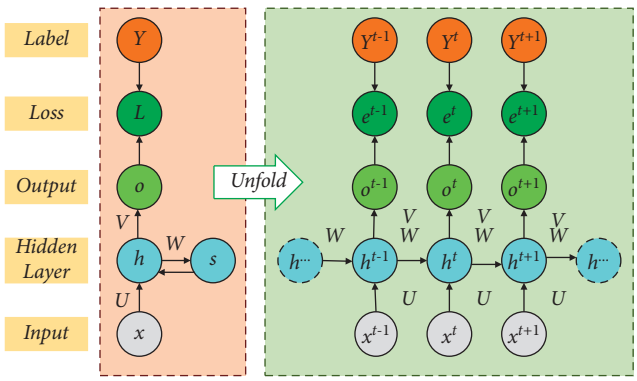


Figure 1: Standard and unfolded RNN.

provides a set of time-series data of sensors like vibration. The data collected from the equipment can be presented as $X = [X^1, X^2, \ldots, X^t, \ldots, X^T], X \in R^{1\times T}$, where X^t is the sampling signal at time $t\,(t = 1, 2, \ldots, T)$. And X is often replaced by features extracted from raw data in application. The RNN takes time-series data X as model input, and then the RUL is obtained as model output. Parameters of the RNN model are optimized via implicit function mapping and error backpropagation through time (BPTT) algorithm. Taking time order into consideration in error propagation, the RNN shows great superiority in time-series data processing.

The RNN model is described in Figure 1 at time t with its fold form on the left and unfold form on the right, where X^t denotes the input at time t, h^t denotes the hidden state at time t, O^t denotes the output at time t, L denotes the loss function, e^t denotes the error, and Y^t denotes the real RUL at time t, which is often replaced by the provided RUL label. Besides, matrixes U, V, W are passing parameters of the RNN model from input nodes to hidden nodes, from hidden nodes to output nodes, and from hidden nodes at time $t-1$ to hidden nodes at time t, respectively.

In Step 1, signals propagate forward along the arrows. h^t and O^t are given as follows:

$$\begin{aligned} h^t &= \psi\left(UX^t + Wh^t + b\right), \\ O^t &= \sigma\left(Vh^t + c\right), \end{aligned} \tag{1}$$

where ψ and σ are activation functions and b and c are the deviation of the input layer and output layer, respectively.

In Step 2, calculation errors propagate back forward through time at every iteration. While tanh is used in the hidden layer and softmax is used in the output layer, the error e^t, total loss L, and partial derivatives of $U, V, W, b,$ *an d* c are calculated as follows:

$$\begin{aligned} e^t &= Y^t - O^t, \\ L &= \sum_{t=1}^{T} e^t, \\ \nabla U &= \frac{\partial L}{\partial U} = \sum_{t=1}^{T} \frac{\partial e^t}{\partial h^t}\frac{\partial h^t}{\partial U} \\ &= \sum_{t=1}^{T} \operatorname{diag}\left(1 - \left(h^t\right)^2\right)\frac{\partial L}{\partial h^t}\left(X^t\right)^T, \\ \nabla V &= \frac{\partial L}{\partial V} = \sum_{t=1}^{T} \frac{\partial e^t}{\partial O^t}\frac{\partial O^t}{\partial V} = \sum_{t=1}^{T}\left(\hat{Y}^t - Y^t\right)\left(h^t\right)^T, \\ \nabla W &= \frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial e^t}{\partial h^t}\frac{\partial h^t}{\partial W} \\ &= \sum_{t=1}^{T} \operatorname{diag}\left(1 - \left(h^t\right)^2\right)\frac{\partial L}{\partial h^t}\left(h^t - 1\right)^T, \\ \nabla b &= \frac{\partial L}{\partial b} = \sum_{t=1}^{T} \frac{\partial e^t}{\partial h^t}\frac{\partial h^t}{\partial b} = \sum_{t=1}^{T} \operatorname{diag}\left(1 - \left(h^t\right)^2\right)\frac{\partial L}{\partial h^t}, \\ \nabla c &= \frac{\partial L}{\partial c} = \sum_{t=1}^{T} \frac{\partial e^t}{\partial O^t}\frac{\partial O^t}{\partial c} = \sum_{t=1}^{T}\left(\hat{Y}^t - Y^t\right). \end{aligned} \tag{2}$$

Through the circles of Step 1 and Step 2, parameters of the RNN model are optimized and the objective function of total loss is minimized; thus, the predicted RUL is as close as possible to its real value. The training process will be stopped when there is no significant improvement on prediction accuracy or when iteration times reach the default number.

2.2. FCM. Clustering is a suitable technique for handling the prognosis tasks without labels, since it aims to organize a set of samples into the corresponding groups based on similarity. Therefore, clustering is actually described as the method for grouping unlabeled data. As the extended version of K-means algorithm, the FCM algorithm was proposed by Bezdek and possesses the ability of assigning each sample to each cluster in a certain degree.

Due to its superiority in dealing with the uncertainty and independence of labels, the FCM algorithm has been widely applied to fault diagnosis of rotating machinery. The FCM algorithm defines and assembles samples into certain classes via minimizing the objective function and calculating the membership degree to clustering centers. The set

of samples can be presented as $X = [x_1, x_1, \ldots, x_1]$, including $c\,(c > 1)$ hidden subsets totally. The clustering center V_i, membership degree V_{ij}, and objective function J are calculated as follows:

$$V_i = \frac{\sum_{j=1}^{n} V_{ij}^m x_j}{\sum_{j=1}^{n} V_{ij}^m}, \tag{3}$$

$$V_{ij} = \frac{1}{\sum_{k=1}^{c} \left(d_{ij}/d_{ik}\right)^{2/(m-1)}}, \tag{4}$$

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} V_{ij}^m d_{ij}^2, \tag{5}$$

where V_{ij}, m, and d_{ij}, respectively, represent the membership degree, the weighted index, and the Euclidean distance from the $j-$th sample to the $i-$th clustering center. The FCM algorithm can be summarized into three steps as follows.

Step 1. Initialize parameters including the number of subsets $c\,(c > 1)$, fuzzy degree q_{ij}, membership degree matrix V, number of iterations, and threshold.

Step 2. Update clustering centers according to the membership degree matrix V and then recalculate the membership degree V_{ij}.

Step 3. Estimate the deviation of function J.

All these steps will be stopped if the deviation or the absolute value of function J is less than the threshold; otherwise, Step 2 and Step 3 will be repeated.

3. Proposed Method

In this section, an RNN-based approach is proposed for RUL prediction. Frequency-domain and time-frequency domain features are used as the input of PCA to obtain the degradation index, and the turning points obtained from the FCM algorithm are used to reconstruct the degradation index into normal condition, slight fault, and heavy fault. The ideal RUL label is then generated to train the RNN model accurately and synchronously. The proposed method mainly consists of three steps, as shown in Figure 2.

3.1. Data Preprocessing. Vibration energy of frequency domain contains information of spectral distribution and position change for the main frequency band. And related frequency-domain features are sensitive to bearing degradation since an imperceptible change produces a spectrum line in the corresponding frequency spectrum. Therefore, it is vital for fault prognosis to extract some indicators in the frequency domain. Besides, as a self-adaptive signal processing technique, EEMD decomposes a signal into several IMFs and one residual. The number and selection of these IMFs depend on the signal itself. Thus, the internal oscillatory modes imbedded in the signal are denoted by IMFs.

The first twelve-dimensional frequency-domain features and the energy of the first N IMFs are extracted for each sample. The default energy of single IMF is zero when the decomposition result is less than default N. And N is set as ten in this study as the number of obtained components from EEMD is about ten for all bearing vibration signals. In order to ensure the stable performance of the training model, the min-max normalization is applied for data preprocessing as in

$$x_i' = \frac{x_i - \min x_i}{\max x_i - \min x_i}. \tag{6}$$

3.2. RUL Label Construction. The FCM algorithm provides the membership degree value of each sample for each clustering center. As the total membership degree value of one sample for all groups is 1, the maximum corresponds to the most likely group that the sample belongs to. Besides, the FCM algorithm has no requirements on the amount of data. Once the obtained data are provided, the algorithm clusters samples into default groups based on similarity. Considering that the bearing's life circle is often divided into three stages of normal condition, slight fault, and heavy fault, the default value of hidden clusters is set as three. With the objective function being optimized in the FCM algorithm, the membership degree values of each sample for each cluster are shown in Figure 3.

As seen in Figure 3, the membership degree from sample 1 to sample $n1$ is 0.85 and that of the other samples is below 0.5 in the first cluster. Thus, the samples from the first to $n1$ belong to the first group. Similarly, samples from $n1 + 1$ to $n2$ belong to the second group, and samples from $n2 + 1$ to the last belong to the third group. So, $n1 + 1$ is the turning point between normal condition and slight fault, and $n2 + 1$ is the turning point between slight fault and heavy fault.

One degradation index is generated by dimensionality reduction of PCA, which is recently regarded as the RUL label in other research studies. Since the linear decreasing relationship is the fatal trend of real life, the new RUL label of three degradation phases is constructed by linear regression based on the degradation index, as shown in Figure 4.

As shown in Figure 4, there are several abnormal points such as the drastic rise point B and straight decrease point A. Obvious deteriorations of bearing condition appear at these moments since the vibration or noise is rapidly increasing. Thus, the value of the degradation index is usually near the extremum, which causes conflicts with the reality of rapid reduction in remaining useful life. Therefore, these abnormal points are eliminated here. One point will be regarded as abnormal once its deviation is three times larger than the average deviation of several adjacent points as in (7) and (8). $2a$ adjacent points are used for calculation, and linear regression is applied to fit the degradation index when all abnormal points are eliminated. And $a = 5$ in this work:

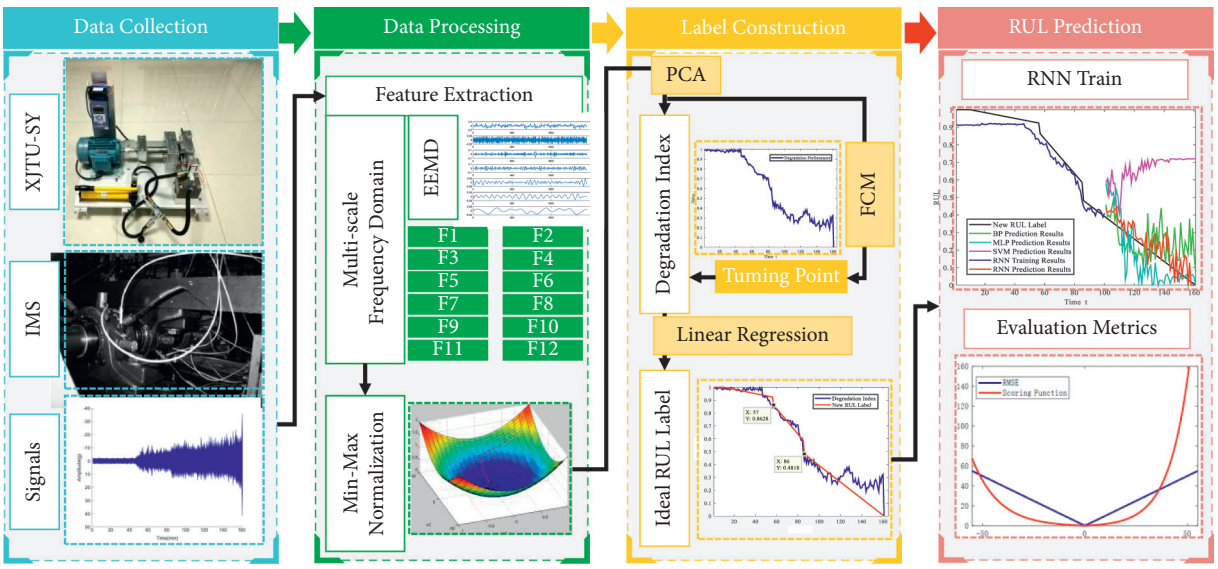


Figure 2: Schematic diagram of the proposed method.

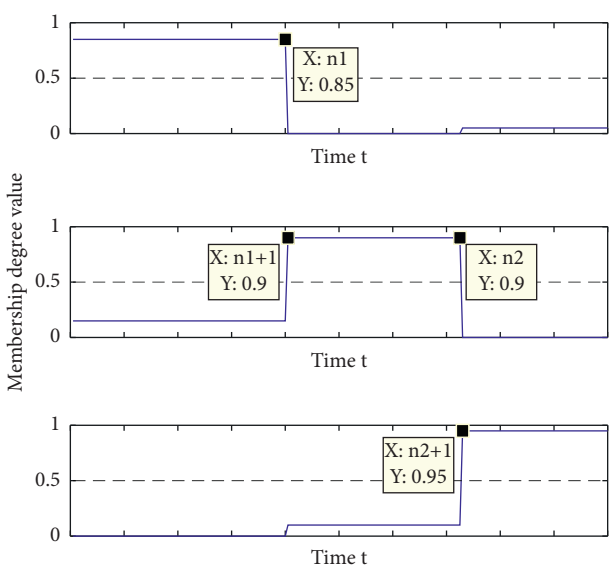


Figure 3: Clustering results of the FCM algorithm.

$$f(t) = f_T(t) + f_R(t), \tag{7}$$

$$M(t) = N \times \left[\frac{\left(\sum_{t-a}^{t+a} f_R(t) - f_R(t)\right)}{2a}\right], \tag{8}$$

where $f(t)$ is the degradation index at time t.,$f_T(t)$ is the mean trend of the degradation index at time t, $f_R(t)$ is the random part of the degradation index at time t, and $M(t)$ is N times the average deviation of $2a$ adjacent points at time t. One point named "$f(t)$" will be replaced by $f_T(t)$ when $f_R(t)$ is larger than $M(t)$.

3.3. RUL Prediction. In order to simulate the online RUL prediction procedure for rolling bearings, the batch size of model training is set as one. Thus, the RNN model will be updated for each new sample. Once the first turning point is detected by the FCM algorithm, the model is going to learn degradation modes and obtain the ability of prediction. And the operation data of previous moments are used to predict the RUL in the future. The demarcation point of training and prediction is determined by an experienced ratio of two to one, when the first turning point between normal condition and slight fault appeared in the first two-thirds of samples. Otherwise, the training set should be expanded until a set of failure samples are also learned by the RNN model.

As shown in Figure 5, the RNN model is constructed based on the frequency-domain features and energy features of the first ten IMFs. With loop iteration for parameter optimization, two hidden layers are established in the RNN model. The key work in the model is the update of weight matrixes for all nodes, which are trained by the BPTT algorithm. The square root error is used as the loss function for partial derivative calculation in the RNN model. Besides, tanh is used as the activation function in the input layer, and *relu* is applied in the output layer.

Suppose X^t is a feature vector extracted from a single sample at time t in the input layer. Outputs of the input layer, the two hidden layers, and the output layer at time t are, respectively, shown as follows:

$$I^t = U * X^t, \tag{9}$$

$$H_1^t = \tanh\left(I^t + W * H_1^{t-1} + b_1\right), \tag{10}$$

$$H_2^t = \tanh\left(H_1^t + W * H_2^{t-1} + b_2\right), \tag{11}$$

$$O^t = \mathrm{relu}\left(V * H_2^t + c\right), \tag{12}$$

where $U, V, W, b, an\ d\ c$ are as the same as mentioned above, and original values of these parameters are initialized randomly from −1 to 1.

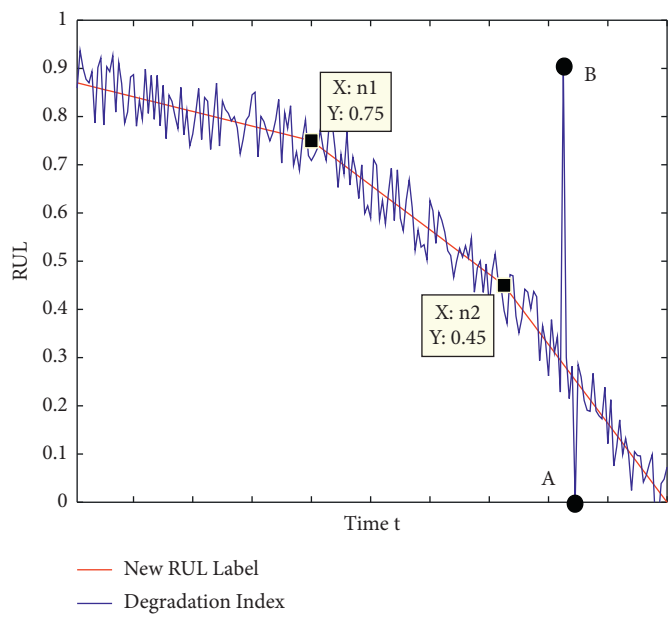


Figure 4: Reconstructed RUL label.

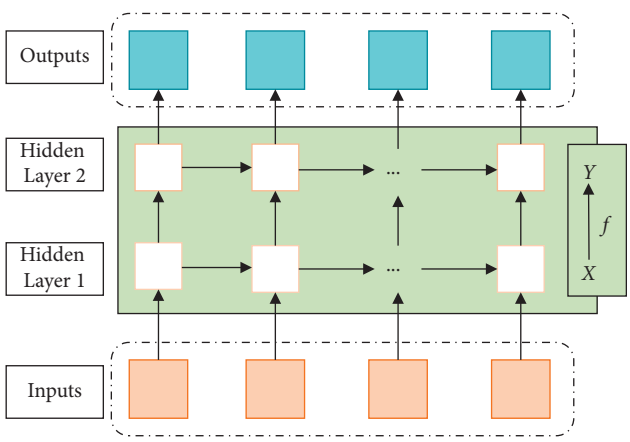


Figure 5: Structure of the RNN model.

The ideal RUL at time t is obtained as Y^t according to the new RUL label. The loss of output is calculated with the ideal RUL, and the partial derivative of model parameters is then updated. Through the repeated forward calculation and BPTT algorithm, the training process of the RNN model is completed. Compared with traditional training algorithms based on gradient descent, the decoupled extended Kalman filter algorithm shows significant advantages on computation and performance. And it is applied in prediction model training. Evaluation is finally conducted based on the prediction results and the ideal RUL of test data.

4. Experimental Study

In experiments, XJTU-SY bearing data and IMS bearing data are used. Compared with other approaches including the backpropagation (BP) network, support vector machine (SVM), and multilayer perceptron (MLP) and traditional functions including linear RUL function and piecewise RUL function, the prediction results of the proposed method show a significant superiority.

4.1. Case Study 1: XJTU-SY Datasets

4.1.1. Data Description. XJTU-SY Bearing Datasets are provided by the Institute of Design Science and Basic Component, Xi'an Jiaotong University (XJTU), and the Changxing Sumyoung Technology (SY). They contain complete run-to-failure data of fifteen rolling bearings by conducting several accelerated degradation experiments [12]. Bearings of type LDK UER204 were operated in totally three conditions, and five bearings were tested under each operating condition. The sampling rate was kept at 25.6 kHz, and the sampling interval was equal to one minute. The bearing test bed is shown in Figure 6.

4.1.2. Evaluation Metrics. In this section, the prediction performance of the proposed method is evaluated quantitatively by employing scoring function and root mean square error (RMSE), which are described, respectively, as follows.

(1) Scoring Function. This is different when the measurement runs ahead of the real value or when the predicted RUL value lags behind the real value. The definition of scoring function is shown as follows:

$$\text{score} = \begin{cases} \sum_{i=1}^{n}\left(e^{h_i/13} - 1\right), & h_i < 0, \\ \sum_{i=1}^{n}\left(e^{h_i/10} - 1\right), & h_i \gg > 0, \end{cases} \tag{13}$$

where h_i is the difference between x_i and h_i as mentioned above.

Furthermore, a smaller score means a better prediction result. Besides, scoring function gives a different penalty

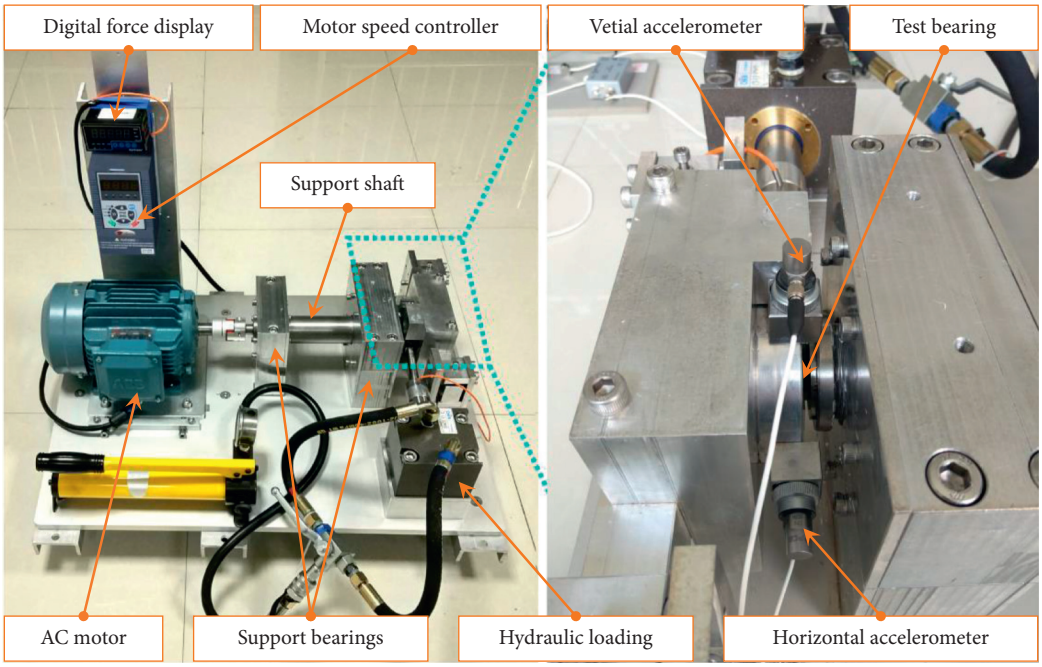


Figure 6: Test bed of XJTU-SY bearing datasets.

when the model underestimates the RUL and when the model overestimates the RUL because there remains little time for maintenance if the predicted RUL is larger than the real one. Faults or disasters are going to happen soon, and prediction gets no significance at all.

(2) RMSE. This is widely used for RUL prediction, which is the root of error square divided by the number of samples. The definition of RMSE is shown as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum (x_i - d_i)^2}{n}}, \tag{14}$$

where x_i represents the measurement, d_i is the real value, and n refers to the number of samples.

This statistical index reflects the extent from measurements to real values and gives the same value no matter the error is positive or negative. Thus, a smaller RMSE value means a better prediction result.

4.1.3. Experimental Results and Analysis. Operation data in the vertical axis of bearing 2-2 are extracted for this experiment. The operating conditions include a rotation speed of 2250 RPM and hydraulic loading of 11 KN. After a constant operation of two hours and forty-one minutes, fault occurred in the outer race, and 161 sampling files are obtained. The former 100 sampling files are used for model training to predict the RUL for the next 61 test samples. The degradation index obtained by PCA is shown in Figure 7. In previous studies, based on expert experience, the degradation stages of bearings are generally divided into three stages: normal condition, slight fault, and severe fault [4]. Therefore, the degradation process can be divided into stages according to the FCM algorithm mentioned in Section 2.2,

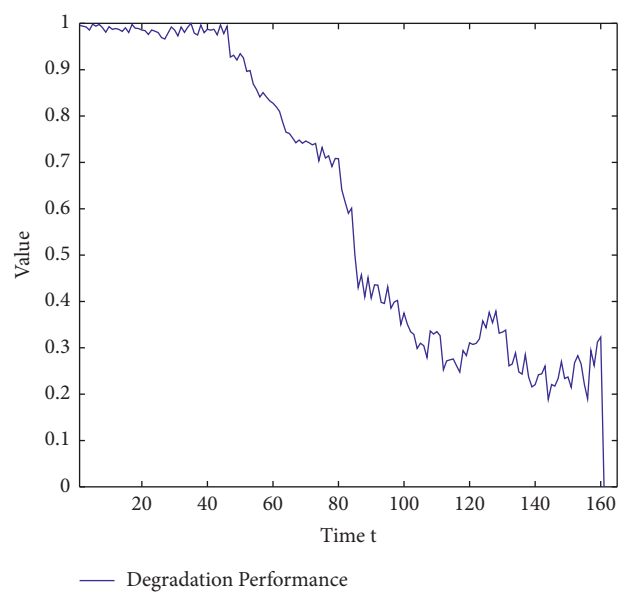


Figure 7: Degradation index of bearing 2-2.

and then the membership degree of each sample to each cluster center can be obtained.

As mentioned above, three clustering centers are set in the FCM algorithm with the maximum iteration as 100 and the error as $1e-6$. The result of membership degree function is shown in Figure 8. The membership degree from sample 1 to sample $n1$ is 0.85, and that of the other samples is below 0.5 in the first cluster. Thus, the samples from the first to $n1$ belong to the first group. Similarly, samples from $n1+1$ to n2 belong to the second group, and samples from $n2+1$ to the last belong to the third group. So, $n1+1$ is the turning point between normal condition

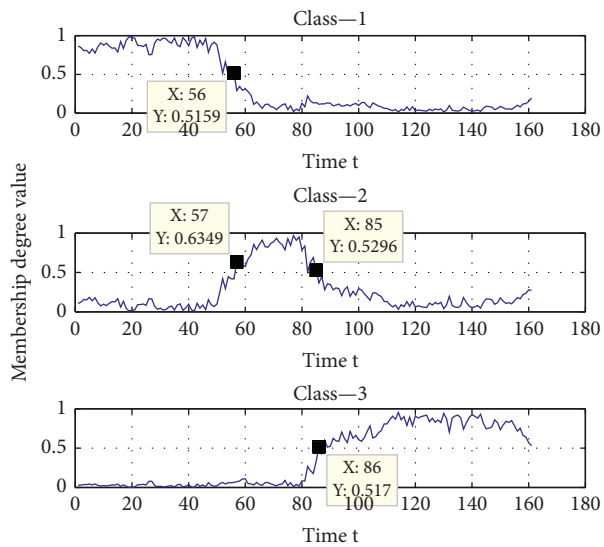


Figure 8: Clustering results of bearing 2-2.

and slight fault, and $n2+1$ is the turning point between slight fault and heavy fault. It can be found that bearing 2-2 is operating normally from the beginning to the 56th sample. The slight fault occurred at the 57th sample, and the heavy fault started at the 86th sample. Based on the turning points, linear regression analysis is applied to construct the new RUL label, as shown in Figure 9.

With the novel RUL label constructed, the RNN model is applied for RUL prediction. Through 10 test runs, the model structure and parameters corresponding to the best prediction result are obtained. The RNN model with two hidden layers is able to discover the hidden patterns from the inputs, and the suitable maximum iteration number is 2000. In Figure 10, the results of RUL prediction at any time step of the test set are given. The prediction results of the test set are given by the red line, while the results of model training are given by the blue line. The results of RUL prediction are fluctuating around the ideal RUL, which proves the prediction ability of the RNN model.

Different models including the classic RNN, back-propagation (BP) network, multilayer perceptron (MLP), and support vector regression (SVR) are compared in this paper in Figure 11. Considering the unpredictable performance of BP networks and MLP, the best network structure is obtained by 10 loop iterations. Finally, the three-layered node 22-27-1 is used in the BP network, which means 22 nodes in the input layer, 27 nodes in the hidden layer, and 1 output node. Two hidden layers with the structure of 22-25-7-1 are used in the MLP. In the SVR model, RBF kernel function is used with gamma value 4, cost value 1.5, and p value $1e-5$ in the loss function.

Linear and piecewise RUL functions are both discussed. Bearing life decreases linearly from one to zero between the first and the last sample in linear function. As shown in Figure 12, the turning point of piecewise function is set at the 70th sample and the original RUL is determined at 130 finally after repeated experiments in the experienced range. The structures of models remain the same as before.

The RMSE and scoring function are adopted to evaluate the performances of models. The comparison of prediction results with different labels and models is shown in Table 1.

As shown in Table 1, the novel RUL label shows a better performance than linear and piecewise RUL functions among all four models, which strongly proves the superiority of the proposed method for RUL label construction. Besides, the RNN obtained the best result of RUL prediction with both the novel RUL label and the piecewise RUL label, followed by the MLP and BP network. And SVR shows little potential for prediction in this experiment.

Turning points are the key information of the proposed method. As the degradation of rolling bearings is usually continuous, the adjacent samples in the time series are usually clustered into the same group until the turning point appears. The influence on clustering results of fluctuations in input data is discussed in this part. Since the first turning point has been known, 10, 20, and 30 samples are added, respectively. Clustering results with three training sequence lengths are shown in Figure 13.

As seen in Figure 13, misjudgment takes place when there exist only 67 samples. And the first turning point comes earlier than the 57th sample. It is because that the clustering centers are farther away from the correct positions when there are fewer samples for training. And the small fluctuations in clustering centers will cause large changes in calculated membership degree values according to the principles of the FCM algorithm. Samples that actually belong to the first group but are far away from the calculated center are misjudged. Thus, the situation will be improved with more samples provided. Through experiments, it is noted that about twenty samples are required for generating accurate turning points, which is not a serious problem with the reality of high-frequency sampling rate and relatively slow degradation rate at the two turning points.

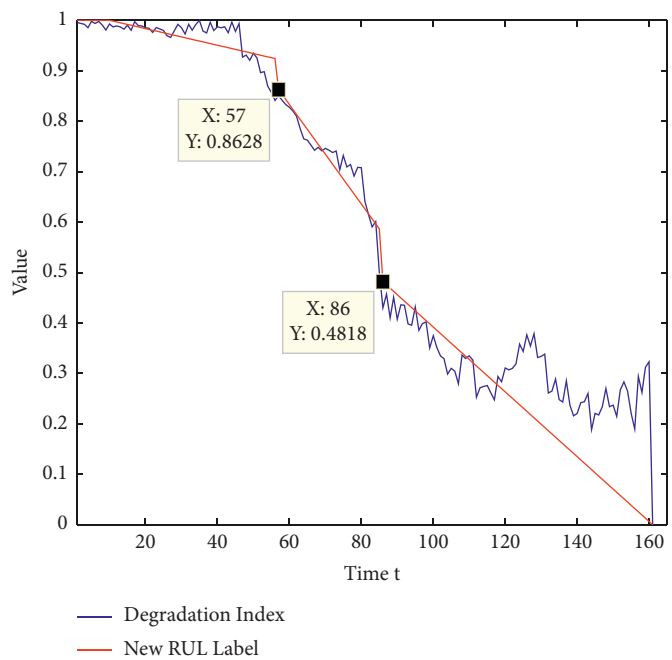


Figure 9: Novel RUL label of bearing 2-2.

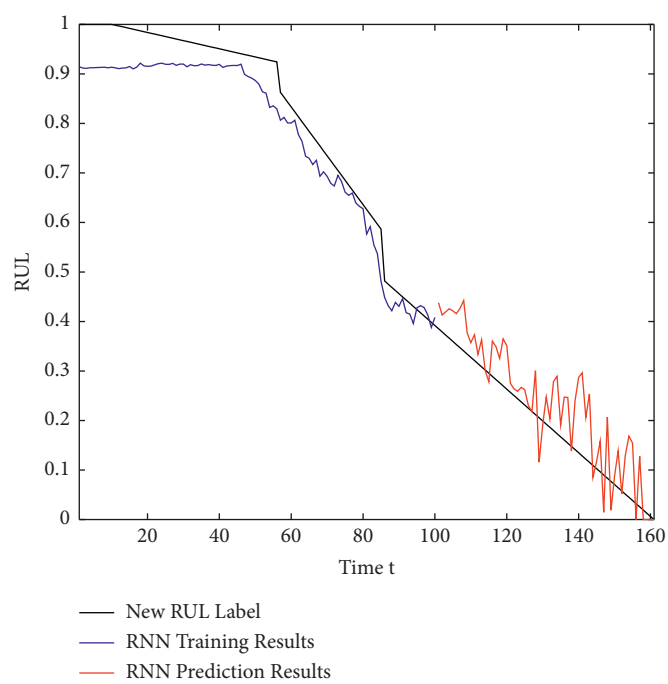


Figure 10: Prediction results of the RNN model.

4.2. Case Study 2: IMS Bearing Data

4.2.1. Data Description. IMS (Intelligent Maintenance Systems of NSF I/UCR Center) bearing data include three sets of degradation experiments among four rolling bearings. These experiments were carried out using bearings of type Rexnord ZA-2115. The test bench structure is shown in Figure 14. As the most commonly used dataset in bearing life prediction research studies, Set 2 is applied for verification of the proposed method.

Set 2 describes a situation that the outer race failure occurred in bearing 1 at the end of experiment, containing 984 sampling files in total. The rotation speed was kept constantly at 2000 RPM, and a radial load of 6000 lbs was applied onto the shaft and bearing by a spring mechanism. Besides, eight accelerometers were mounted, respectively, in vertical and horizontal directions of four bearings with a sampling rate of 20 kHz and a sampling interval of 10 minutes.

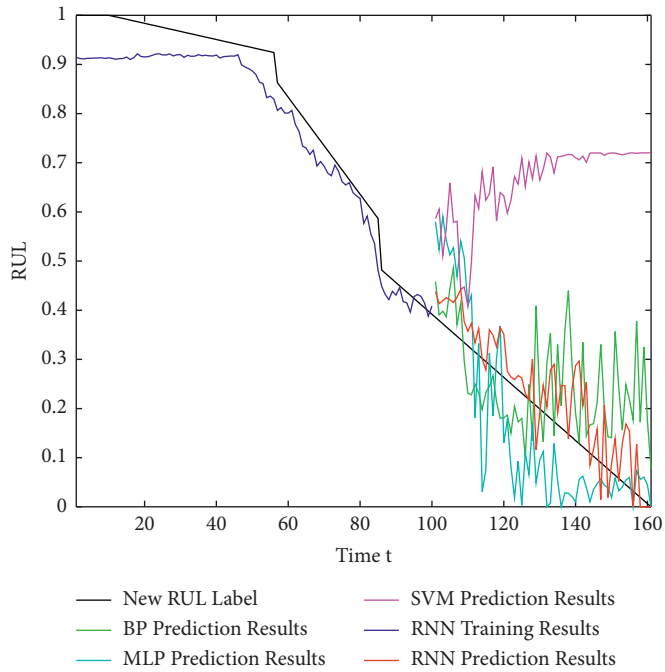


Figure 11: Prediction results of four models.

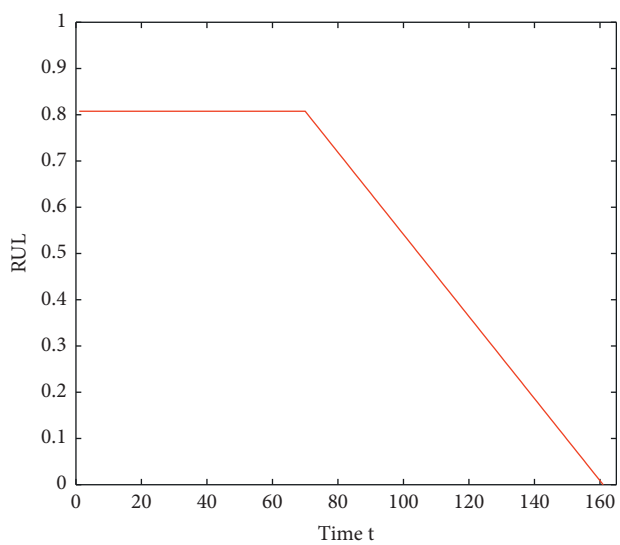


Figure 12: Piecewise RUL function.

Table 1: Evaluation results of three labels and four models (XJTU-SY).

Models	New RUL label		Linear RUL label		Piecewise RUL label	
	Score	RMSE	Score	RMSE	Score	RMSE
RNN	0.4173	0.2250	0.4492	0.5724	1.2704	1.6062
BP network	0.6192	0.4582	1.6552	1.9789	1.4326	1.7585
MLP	0.5369	0.2647	1.5889	1.9935	1.4882	1.8286
SVR	2.9445	3.6695	3.1515	3.9295	3.3848	4.2084

4.2.2. Experimental Results. 700 samples are used for model training to predict the next 284 outputs. PCA is applied for the feature's dimensionality reduction. The FCM algorithm is then applied for clustering, as shown in Figure 15. As seen in Figure 15, the bearing operates normally from the beginning to the 533rd sample and gets into slow

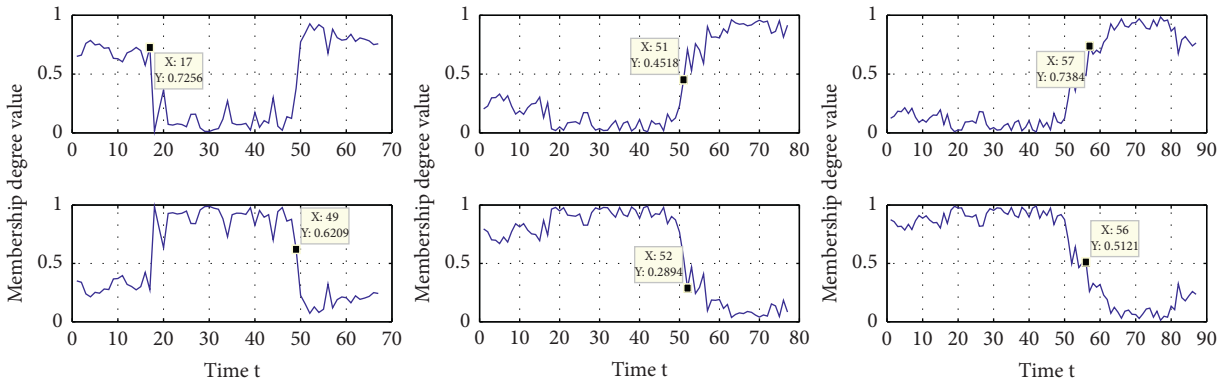


Figure 13: Clustering results with different training lengths.

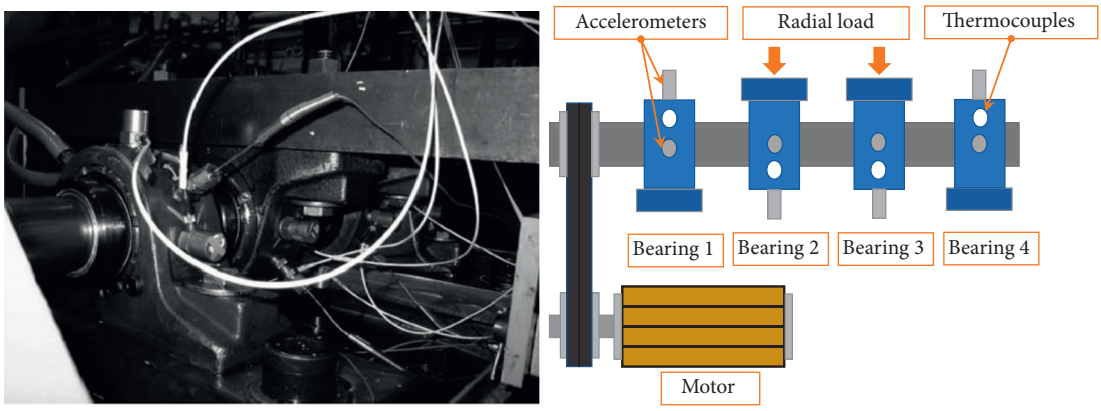


Figure 14: IMS test rig.

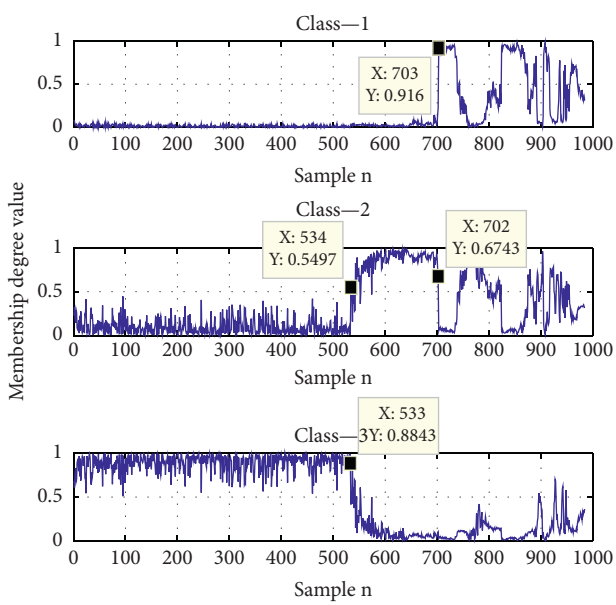


Figure 15: Clustering results of bearing 1.

degradation at the 534th sample and rapid degradation at the 703rd sample. With the turning points known, the linear regression algorithm is applied, and a novel RUL label appears in Figure 16.

It is noted that the RUL label at a training sequence length shorter than 703 is different from that in Figure 16, since the FCM algorithm has merely detected the first turning point. And then the RUL label is constructed

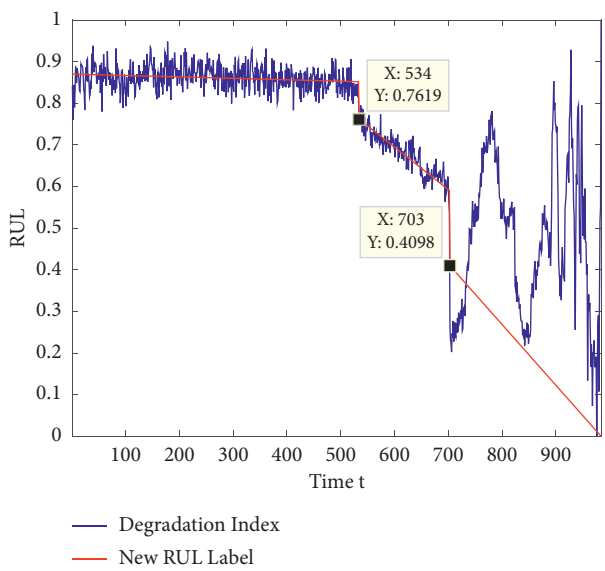


Figure 16: Novel RUL label of bearing 1.

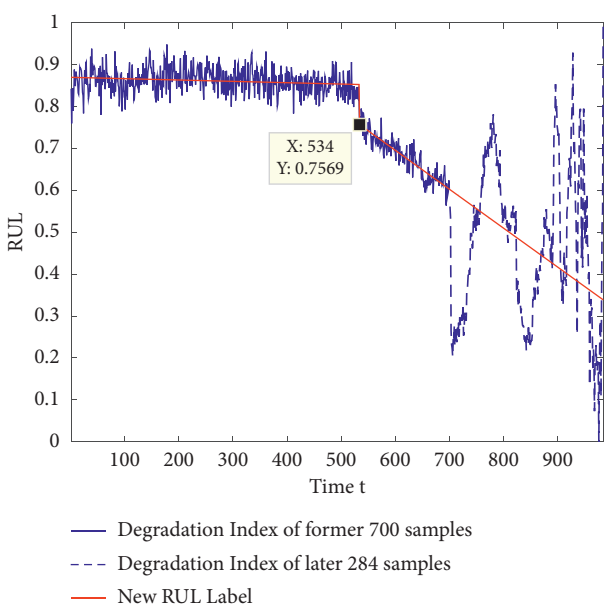


Figure 17: RUL label at a training length of 700.

through operation data before the next turning point. When there are 700 samples for model training, the RUL label is shown in Figure 17.

Four models of RNN, BP network, MLP, and SVR are compared. Here, the BP network with nodes 22-27-1 is constructed. The MLP with two hidden layers of 22-25-7-1 is obtained. A standard RNN structure is used with two hidden layers and the maximum iterations of 2000. In the SVR model, a polynomial is used as the kernel function with a cost value 1.5 and p value $1e-4$ in loss function. These models run for 10 times, respectively, and compared in the best results.

Besides, linear and piecewise RUL functions are used for prediction, where the turning point of piecewise function is set at the 450th sample and the original RUL is 750. For evaluation, the RMSE and scoring function of different models and labels are calculated and shown in Table 2. Similar conclusions are easily achieved to the first experiment in XJTU-SY bearing data. The proposed label better agrees with the real RUL, and the RNN performs the best among four models.

Table 2: Evaluation results of three labels and four models (IMS).

Models	New RUL label		Linear RUL label		Piecewise RUL label	
	Score	RMSE	Score	RMSE	Score	RMSE
RNN	3.2199	0.6838	5.7834	2.9054	4.1641	1.8924
BP network	3.5863	3.0051	8.3087	5.1212	4.9563	3.2139
MLP	3.9781	3.5016	6.1778	5.3195	5.1935	4.7358
SVR	4.6046	1.2288	5.8675	5.9114	6.1047	3.1722

5. Conclusions

An adaptive method of RUL training label construction based on the FCM algorithm is proposed in this paper, and the whole proposition is demonstrated by comparing the new constructed RUL label and the current linear RUL function and piecewise RUL function. Experiments carried out on one recent bearing degradation dataset, XJTU-SY bearing data, and a widely used dataset, IMS bearing data, strongly proved the superiority of the proposed method. The conclusions of this work are as follows:

(1) The RUL label constructed by the proposed method better fits the hidden degradation mode than linear and piecewise RUL functions through a comparison of four common models for remaining useful life estimation. And the recurrent neural network performs the best in these models with its excellent capability of time-series data processing.

(2) With the FCM algorithm for turning point detection, the fault of bearings is found synchronously, which provides a solution for online RUL prediction with no experience of failure threshold and ending time of life to use. Besides, it takes several samples for the algorithm to identify the turning points accurately.

(3) The influence of the training sequence length is discussed in both experiments. As there exists more degradation information in the larger training set, the FCM algorithm shows higher accuracy on cluster and the RNN provides better performance on prediction.

The proposed method based on the FCM algorithm is proved to be powerful and accurate for RUL label construction, but the computing performance of the RNN model still needs to be improved. For example, the RUL can be predicted by the long short-term memory (LSTM) network, gated recurrent unit (GRU), bidirectional LSTM network, bidirectional GRU, transformer, and so on. And more improvements on algorithms should be developed. Besides, the linear degradation mode based on current

research studies is applied in this work. And nonlinear degradation curves could be used in the future research studies, as many degradation progresses are nonlinear in application.

As for further work, it is not idealistic to divide the bearing degradation process into several stages since it is commonly considered continuous and gradual. Another thing is that imbalance of degradation process division and rare failure data may lead to the poor performance of the assessment model. Thus, further work will pay attention to the continuous label of degradation process, and the imbalance of faulty data will also be considered.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hailong Lin and Zihao Lei contributed equally to this work.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (no. 2020YFB1710002), in part by the National Natural Science Foundation of China (no. 51775409), and in part by the Equipment Pre-Research Fund of China (no. 61420030301).

References

[1] H. Zhou, J. Chen, G. Dong, and R. Wang, "Detection and diagnosis of bearing faults using shift-invariant dictionary learning and hidden Markov model," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 65–79, 2016.

[2] Z. Huang, Z. Lei, G. Wen et al., "A multi-source dense adaptation adversarial network for fault diagnosis of machinery," *IEEE Transactions on Indutrial Electronics*, 2021.

[3] Z. Lei, G. Wen, S. Dong et al., "An intelligent fault diagnosis method based on domain adaptation and its application for bearings under polytropic working conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[4] S. Dong, G. Wen, Z. Lei, and Z. Zhang, "Transfer learning for bearing performance degradation assessment based on deep hierarchical features," *ISA Transactions*, vol. 108, pp. 343–355, 2021.

[5] W. T. Sui, D. Zhang, X. M. Qiu, W. Zhang, and L. Yuan, "Prediction of bearing remaining useful life based on mutual information and support vector regression model," in *Proceedings of the 2019 the 5th International Conference on Electrical Engineering, Control And Robotics*, IOP Conference Series-Materials Science and Engineering, Guangzhou, China, 2019.

[6] Y. Hu, H. Li, P. Shi et al., "A prediction method for the real-time remaining useful life of wind turbine bearings based on the Wiener process," *Renewable Energy*, vol. 127, pp. 452–460, 2018.

[7] N. Li, Y. Lei, J. Lin, and S. X. Ding, "An improved exponential model for predicting remaining useful life of rolling element bearings," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 12, pp. 7762–7773, 2015.

[8] D. Wang and K.-L. Tsui, "Two novel mixed effects models for prognostics of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 99, pp. 1–13, 2018.

[9] K. Javed, R. Gouriveau, and N. Zerhouni, "State of the art and taxonomy of prognostics approaches, trends of prognostics applications and open issues towards maturity at different technology readiness levels," *Mechanical Systems and Signal Processing*, vol. 94, pp. 214–236, 2017.

[10] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018.

[11] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in *Proceedings of the 2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 88–95, IEEE, Dallas, TX, USA, 19 June 2017.

[12] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 401–412, 2020.

[13] F. O. Heimes, "Recurrent neural networks for remaining useful life estimation," in *Proceedings of the 2008 International Conference on Prognostics and Health Management*, pp. 59–64, IEEE, Denver, CO, USA, 6 October 2008.

[14] G. S. Babu, P. Zhao, and X.-L. Li, Eds., in *Proceedings of the International conference on database systems for advanced applications*, Springer, Cham, 25 March 2016.

[15] N. Gebraeel, M. Lawley, R. Liu, and V. Parmeshwaran, "Residual life predictions from vibration-based degradation signals: a neural network approach," *IEEE Transactions on Industrial Electronics*, vol. 51, no. 3, pp. 694–700, 2004.

[16] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing*, vol. 240, pp. 98–109, 2017.

[17] W. Mao, J. He, J. Tang, and Y. Li, "Predicting remaining useful life of rolling bearings based on deep feature representation and long short-term memory neural network," *Advances in Mechanical Engineering*, vol. 10, no. 12, 2018.

[18] Y. Zhang, R. Xiong, H. He, and Z. Liu, Eds., in *Proceedings of the 2017 Prognostics and System Health Management Conference (PHM-Harbin)*, IEEE, Harbin, China, 9 July 2017.

[19] J. Wu, C. Wu, S. Cao, S. W. Or, C. Deng, and X. Shao, "Degradation data-driven time-to-failure prognostics approach for rolling element bearings in electrical machines," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 529–539, 2018.

[20] M. Yuan, Y. Wu, and L. Lin, Eds., in *Proceedings of the 2016 IEEE international conference on aircraft utility systems (AUS)*, IEEE, Beijing, China, 10 October 2016.

[21] Y. Wang, Y. Peng, Y. Zi, X. Jin, and K.-L. Tsui, "A two-stage data-driven-based prognostic approach for bearing degradation problem," *IEEE Transactions on industrial informatics*, vol. 12, no. 3, pp. 924–932, 2016.

[22] J. Wu, C. Wu, Y. Lv, C. Deng, and X. Shao, "Design a degradation condition monitoring system scheme for rolling

bearing using EMD and PCA," *Industrial Management & Data Systems*, vol. 117, 2017.

[23] J. Zhang, P. Wang, R. Yan, and R. X. Gao, "Long short-term memory for machine remaining life prediction," *Journal of Manufacturing Systems*, vol. 48, pp. 78–86, 2018.

[24] L. Cao, Z. Qian, and Y. Pei, Eds., in *Proceedings of the 2018 Prognostics and System Health Management Conference (PHM-Chongqing)*, IEEE, Chongqing, China, 26 October 2018.

[25] Y. Zhang, B. Tang, Y. Han, and L. Deng, "Bearing performance degradation assessment based on time-frequency code features and SOM network," *Measurement Science and Technology*, vol. 28, no. 4, Article ID 045601, 2017.

[26] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. Kamath, "Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," *Knowledge-Based Systems*, vol. 190, Article ID 105321, 2020.

[27] Y. Lei, *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery*, Butterworth-Heinemann, Oxford, UK, 2016.

[28] X. Zhao and M. Jia, "A novel deep fuzzy clustering neural network model and its application in rolling bearing fault recognition," *Measurement Science and Technology*, vol. 29, no. 12, Article ID 125005, 2018.

Hindawi
Shock and Vibration
Volume 2021, Article ID 7469691, 12 pages
https://doi.org/10.1155/2021/7469691

Research Article

A Simultaneous Fault Diagnosis Method Based on Cohesion Evaluation and Improved BP-MLL for Rotating Machinery

Yixuan Zhang,[1] Rui Yang,[1] Mengjie Huang,[1] Yu Han,[1] Yiqi Wang,[1] Yun Di,[1] Dongke Su,[1] and Qidong Lu[2]

[1]*Xi'an Jiaotong-Liverpool University, Suzhou, China*
[2]*Weihai Beiyang Electric Group, Weihai, China*

Correspondence should be addressed to Rui Yang; r.yang@xjtlu.edu.cn

Received 6 June 2021; Revised 22 August 2021; Accepted 20 October 2021; Published 8 November 2021

Academic Editor: Jun Zhu

In this paper, an improved simultaneous fault diagnostic algorithm with cohesion-based feature selection and improved backpropagation multilabel learning (BP-MLL) classification is proposed to localize and diagnose different simultaneous faults on gearbox and bearings in rotating machinery. Cohesion evaluation algorithm selects high sensitivity feature parameters from time and frequency domain in high-dimensional vectors to construct low-dimensional feature vectors. The BP-MLL neural network is utilized for fault diagnosis by classifying the feature vectors. An effective global error function is proposed in BP-MLL neural network by modifying distance function to improve both generalization ability and fault diagnostic ability of full-labeled and nonlabeled situations. To demonstrate the effectiveness of the proposed method, simultaneous fault diagnosis experiments are conducted via wind turbine drivetrain diagnostics simulator (WTDDS). The experiment results show that the proposed method has better overall performance compared with conventional BP-MLL algorithm and some other learning algorithms.

1. Introduction

Rotating machinery is a power transmission device in various mechanical equipment and also has been an indispensable part in industrial applications. Components in the rotation machinery including rotor, rotating shaft, bearing, and gearbox are all under arduous work and, thus, are subject to performance degradations and mechanical failures [1, 2]. However, any failure of key components in rotating machinery will cause serious accidents with high economic losses [3]. Therefore, the accurate detection of mechanical fault locations and types in rotating machinery is highly needed.

Currently, there are two types of quantitative analysis fault diagnostic methods for rotating machinery: model-based and data-driven diagnosis [4]. Model-based methods implement dynamic process models in the form of mathematical formulas and parameters; however, describing models by mathematical structure can be difficult and inefficient because of the more and more complicated industrial processes [5]. Compared with model-based fault diagnosis, data-driven methods tend to transfer diagnostic problems to the pattern recognition problems. The data-driven methods are mainly composed of multivariate statistical analysis, such as regression [6] and principal component analysis [7], and machine learning methods, such as support vector machine (SVM) [8], random forest [9], neural networks [10–14], and transfer learning [15, 16].

Neural networks have been commonly used for intelligent fault diagnosis due to the powerful capabilities of pattern classification and function approximation [17]. On the basis of learning strategies, diagnostic algorithms based on neural networks can be classified into supervised and unsupervised learning. Backpropagation neural network is one of the most popular supervised learning strategies; in the 20th century, researches in [18–20] all proved the effectiveness of backpropagation neural networks in fault diagnosis. Additionally, compared with SVM, neural networks have a higher classification accuracy for fault diagnosis of rotor bearing systems [17]. Modifications to conventional backpropagation neural networks are also recommended to

address the problem of fault diagnosis. Meireles et al. pointed out that radial basis function (RBF) networks offer advantages of higher training speed and an easier optimization of performance over conventional neural networks for fault diagnosis [21]. Wu and Chow developed a RBF network-based system to induct machine faults and propose the cell-splitting grid algorithm so that the architecture of RBF network is automatically determined [22]. This proposed system can detect unbalanced electrical and mechanical faults under different working environment.

Unsupervised networks have different architectures such as self-organizing neural networks whose structures are adaptively determined to realize that all nodes in a neighborhood have similar output to an input when stable. The method based on self-organizing maps (SOM) proposed in [23] is not only able to detect the bearing faults, but also locates them and evaluates the failure extent. Jounela et al. developed a process monitoring system based on SOM associated with heuristic rules to detect machine malfunctions [24]. All in all, neural networks are capable of classifying arbitrary regions in space which makes it a good choice for fault diagnosis [25]. The recognition of simultaneous multiple faults is also discussed in [26]; the diagnostic performance by using single-fault recognition techniques may be limited: (1) fault isolation operations can be difficult since noise in the measured signals can obscure a particular fault feature; (2) a large training set is required, which is difficult and time-consuming to collect; (3) the choice of the most suitable classifier is still vague in engineering practice.

Multilabel learning methods are usually adopted during the detection and diagnosis of simultaneous fault in rotating machinery. Three main groups of multilabel learning strategies are data transformation, adaptation, and ensemble of classifiers [26]. The basic idea of data transformation methods is to turn the multilabel problems to other known learning problems according to [27]; one of the representative algorithms is the binary relevance which converts the original multilabel dataset to binary dataset [28]. Adaptation methods improve the conventional classification algorithms and directly employ the adapted algorithms for learning on multilabel data [29]. The kernel-dependent SVM in [8] is utilized to select features and to realize simultaneous fault detection of continuous processes. Zhang et al. extended k nearest neighbor (KNN) to a multilabel learning approach, named ML-KNN [30]. In detail, maximum a posteriori (MAP) principle is employed to determine the label set after k nearest neighbors are recorded for each instance in the training set. In terms of an ensemble of classifiers, Zhong et al. advanced a new probabilistic framework that combines multiple classifiers with a new ensemble method to realize simultaneous fault diagnosis with only single-fault data trained [31]. The first multilabel learning algorithm derived from the feed-forward neural network is proposed in [32], named back-propagation multilabel learning (BP-MLL). This neural network is optimized by minimizing the differences between the actual outputs and desired outputs on each training example. One of the most popular error functions is the sum-of-squares error functions, but BP-MLL applies a novel error function that does an exponential operation on the differences between the outputs of labeled units and unlabeled units to capture characteristics of multilabel learning, i.e., yield output of labeled unit larger than that of unlabeled unit, and then a threshold function is used to determine a label set associated with each instance. BP-MLL neural network has been applied to assist medical syndrome diagnosis [33,34]. Multilabel text categorization systems based on BP-MLL neural networks are developed to classify multilabel documents [32]. Moreover, the prediction model based on BP-MLL in [35] is applied to estimate the types of sustainable flood retention basins.

However, the computation in BP-MLL neural networks is complex; according to [32], the total training cost of BP-MLL is $O(W \cdot I \cdot n)$, where W represents the total number of weights and bias, I is the number of training instances, and n is the total number of training epochs. Furthermore, the distance between relevant labels and irrelevant ones in conventional BP-MLL is represented by subtraction, which may be not obvious enough to be observed. Thus, two new distance functions that enhance pairwise labels discrimination to improve BP-MLL algorithm are proposed in [36]. Besides the problems mentioned above, the conventional BP-MLL algorithm is also not applicable for scenarios with full-labeled or nonlabeled situations. Multilabel classification assigns each instance with multiple categories that reflect properties of a data-point such as topics relevant to a document. A text might be about any of politics, education, specialties, or finance at the same time or none of these. Assume that a set of labels is organized and associated with each instance; if an instance is relevant to all the labels, then all labels in the set will be marked, so called full-labeled situation. Similarly, if an instance does not have connection to any labels in the set, it will be considered as nonlabeled situation. Modifications for BP-MLL algorithm made in [37] avoid failures under these two situations by taking differences between the rank values and the thresholds into account; besides, experimental performance of the modified algorithm is better shown on the same dataset as that in [32]. However, none of the existing literature specifically addresses full-labeled or nonlabeled situations, which may cause serious problems in practical application as computational errors may happen during network learning process in current approaches. As a result, the normal working rotating machinery would be misdiagnosed as a faulty one. Therefore, in this paper, we proposed an improved BP-MLL algorithm with a novel global error cost function and regularization term enhancing the generalization ability. Additionally, the cohesion evaluation algorithm based on standard deviation analysis is applied to obtain more comprehensive signal information and improve the adaptive ability of dynamic models.

Based on the above literature review and discussions, the main contributions of this paper are declared as follows: (1) a new global error function is proposed to deal with the problem of full-label and nonlabel learning situations; (2) a fault diagnosis method based on the improved BP-MLL and cohesion evaluation is proposed; (3) the problem of multilabel gearbox and bearing fault diagnosis in rotating machine under different working and environmental conditions is investigated.

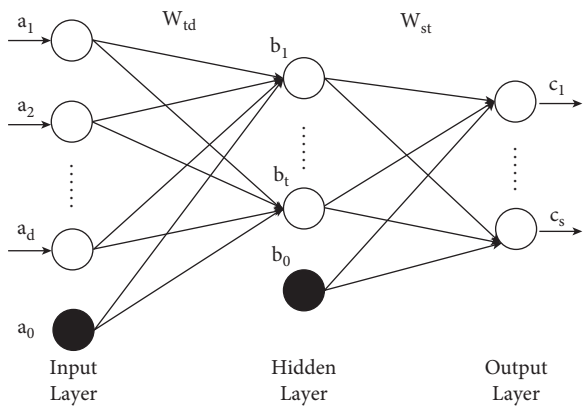


Figure 1: The architecture of BP-MLL model.

The structure of this paper is as follows. Section 2 discusses the preliminaries and formulates the problem. In Section 3, the proposed method is introduced. In Section 4, hardware experiments and comparative studies are carried out to verify the effectiveness of the method. Section 5 concludes this paper.

2. Preliminaries and Problem Formulation

2.1. BP-MLL. Suppose the training set is composed of I multilabel instances, i.e., $\{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots, (X_i, Y_i)\}$, $i = 1, 2, \ldots, I$. Each X_i is a d-dimension feature vector and Y_i is the associated set of labels. The BP-MLL architecture is shown in Figure 1, where d input neurons correspond to a feature vector, s output neurons represent s labels in Y_i, and the hidden layer has t hidden units. Each layer is fully connected with the next layer, with the weights $W = [W_{td}, W_{st}]$ $(d = 1, 2, \ldots, N; t = 1, 2, \ldots, M; s = 1, 2, \ldots, J)$. The number of hidden layers may be more than one in different neural network structures.

The error function proposed in [32] is

$$E = \sum_{i=1}^{I} E_i = \sum_{i=1}^{I} \frac{1}{|Y_i||\overline{Y}_i|} \sum_{(k,l)\in Y_i \times \overline{Y}_i} e^{-\left(c_k^i - c_l^i\right)}. \tag{1}$$

This error cost function reflects the relationship between relevant labels and irrelevant ones by calculating the difference between them:

$$\text{distance} = \text{func}\left(c_k^i, c_l^i\right) = c_k^i - c_l^i. \tag{2}$$

Specifically, $\overline{Y}_i$ is the complementary set of Y_i and $|\cdot|$ measures the cardinality of a set. c_k^i represents the output of the network on one label belonging to the instance $(k \in Y_i)$ and c_l^i represents the one not belonging to it $(l \in \overline{Y}_i)$. Apparently, the larger the difference is, the smaller value of the error function of BP-MLL algorithm is, so that labels in Y_i will get greater neural network outputs than those not in. Therefore, when the training set covers sufficient information to disseminate the learning problem, the trained neural network will eventually distinguish the relevant labels from irrelevant ones.

2.2. Problem Formulation. Consider the following uncertain cases in the diagnostic system:

(1) There are not any labels for one instance, indicating all components in the rotating machinery run perfectly such that $|Y_i| = 0$.

(2) All the components are broken down such that $|\overline{Y}_i| = 0$, where $\overline{Y}_i$ is the complementary set of Y_i.

When the diagnostic system applies error function Eq. (1), either uncertain case would cause mathematical failures, in the full-labeled case:

$$E = \sum_{i=1}^{I} E_i = \sum_{i=1}^{I} \frac{1}{|Y_i||\overline{Y}_i|} \sum_{k\in Y_i} e^{-c_k^i}. \tag{3}$$

Firstly, because of $|\overline{Y}_i| = 0$, the denominator $|Y_i||\overline{Y}_i| = 0$; and furthermore, the value of c_k^i would toward infinity based on the property of exponential function. Similarly, when there do not exist any labels for a specific instance, $|Y_i| = 0$ leads to an unreasonable denominator and besides, c_l^i is approaching infinity as well:

$$E = \sum_{i=1}^{I} E_i = \sum_{i=1}^{I} \frac{1}{|Y_i||\overline{Y}_i|} \sum_{l\in \overline{Y}_i} e^{c_l^i}. \tag{4}$$

According the chain rule and the gradient descent rule for updating weights, the mathematical formulations are shown as below:

$$\begin{aligned} W_{st} &= W_{st} + \Delta W_{st}, \\ \Delta W_{st} &= -\alpha \frac{\partial E_i}{\partial W_{st}}, \\ &= -\alpha \frac{\partial E_i}{\partial Sc_s} \frac{\partial Sc_s}{\partial W_{st}}, \\ &= -\alpha \frac{\partial E_i}{\partial c_s} \frac{\partial c_s}{\partial Sc_s} \frac{\partial Sc_s}{\partial W_{st}}, \end{aligned} \tag{5}$$

where α is the learning rate, W_{st} represents the weights from the hidden layer to the output layer, Sc_s represents the weighted sum, and c_s is the actual output of s-th output unit.

Apparently, there exists nondifferentiability in Eq. (5) since the value of c_s tends to infinity. Therefore, our approach is to optimize the error function to develop a fault diagnostic system such that it would not be affected by uncertain cases such as nonlabeled and full-labeled situation. In the next section, the improved error function that tolerates uncertain cases with high generalization ability is introduced.

3. Proposed Fault Diagnosis Method

Figure 2 illustrates the steps of fault diagnosis in rotating machinery. Firstly, the signals from the time-frequency domain are collected from multiple channels under different working conditions. Secondly, to fully grasp the characteristics of the signal and enhance the recognition ability of the fault diagnosis system, the cohesion evaluation algorithm is employed to pick out feature parameters with high

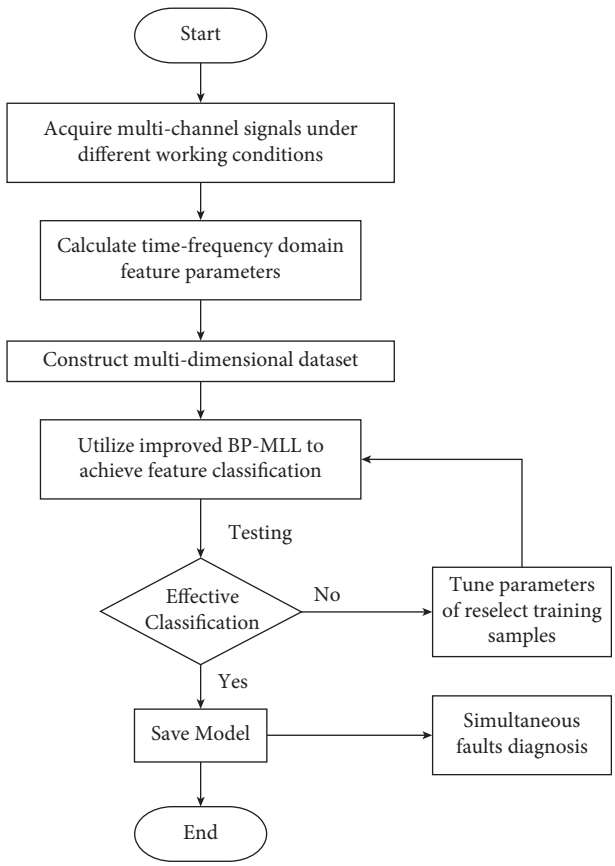


Figure 2: Training flow chart of the proposed method.

sensitivity to form the sensitive feature vector. Finally, an improved BP-MLL neural network is trained and utilized to classify the constructed feature vectors to make the system have dynamic model adaptability.

3.1. Feature Selection. Compared with conventional algorithms, the cohesion evaluation algorithm based on standard deviation analysis can combine multiple signals to obtain more comprehensive signal information and achieve the purpose of improving the accuracy of fault diagnosis [10]. The distance assessment technique is described as follows:

Assume a set of d-dimensional feature vector has J different classes and the index of samples for each category is i:

$$\left\{q_{s,m,i}, \quad s = 1, 2, \ldots, J;\ m = 1, 2, \ldots, M; i = 1, 2, \ldots, I\right\}, \tag{6}$$

where s, m, and i are positive integers, and $q_{s,m,i}$ reflects the m-th feature parameter of the i-th instance of the s-th category.

Table 1 is the specific operational steps of the cohesion evaluation algorithm where steps 1–3 reflect the intracategory standard deviation computation. c is a characteristic parameter, in which different characteristic parameters represent different practical meanings and u reflects the weight of the corresponding position neuron. The classification can be improved by reducing average intracategory standard deviation $\mathrm{clt}_J^{\mathrm{inner}}$ and the intracategory standard deviation f_J^{inner}. Steps 4–8 represent the standard deviation computation of the feature distance, where the larger standard deviation of the feature distance $\mathrm{stc}_{s,J}$ and the smaller imparity measure of intercategory cohesion difference f_J^{outer} are more favorable for classification. Steps 9–10 determine the sensitivity of each feature parameter.

Figure 3 represents the cohesion evaluation process using the parameters in Table 1 as the horizontal and vertical coordinates, where $\mathrm{clt}^{\mathrm{inner}}$ and stc, respectively, represent the size of circle radius and the position of circle center. In Figure 3, the intracategory standard deviation in class 3 and class 4 are easily overlapped and the distances between the points in each class are similar, resulting in a small average intercategory cohesion difference $\mathrm{clt}_J^{\mathrm{outer}}$, which is not conducive for distinguishing; in contrast, classes 1 and 2 belong to the easy classification feature parameter class. Overall, the cohesion evaluation algorithm can reflect the internal dispersion of the data and compare the detail of data differences. According to the steps of the cohesion evaluation algorithms in Table 1, the sensitivity weighting factor can be calculated as

$$\beta_m = \frac{1}{\left[f_m^{\mathrm{inner}}/\max\left(f_g^{\mathrm{inner}}\right) + f_m^{\mathrm{outer}}/\max\left(f_g^{\mathrm{outer}}\right) + e_m^{\mathrm{inner}}/\max\left(e_g^{\mathrm{inner}}\right) + e_m^{\mathrm{outer}}/\max\left(e_g^{\mathrm{outer}}\right)\right]}. \tag{7}$$

The sensitivity factor is

$$\eta_m = \beta_m \frac{\mathrm{clt}_m^{\mathrm{outer}} + \mathrm{pro} \cdot d_m^{\mathrm{outer}}}{\mathrm{clt}_m^{\mathrm{inner}} + \mathrm{pro} \cdot d_m^{\mathrm{inner}}}, \tag{8}$$

where pro is the proportional adjustment coefficient.

The input feature vector of classification neural network is constructed by selecting parameters with large sensitivity factor according to equation (7):

$$\begin{aligned} \nu &= \mathrm{sort}(\eta), \\ X_i &= [\nu_1, \nu_2, \ldots, \nu_d], \end{aligned} \tag{9}$$

where $\mathrm{sort}(\cdot)$ sorts the m feature parameters in descending order, and first d ($d < m$) high sensitive parameters construct a d-dimension input feature vector.

3.2. Feature Classification. The conventional BP-MLL algorithm captures correlation between relevant labels and irrelevant ones by using distance function which calculates the difference between them. The error function accumulates the differences in each instance and then normalize the summation by the total number of pairwise labels, i.e., $|Y_i||\overline{Y}_i|$. As a result, with the increase in distances, the value

Table 1: Cohesion evaluation.

Step	Process parameter	Expression
1	Intracategory standard deviation	$\sigma_{m,i} = \sqrt{\sum_{s=1}^{J} (q_{s,m,i} - u_{m,i})^2 / J - 1}$
2	Average intracategory standard deviation	$\mathrm{clt}_i^{\mathrm{inner}} = 1/d \sum_{m=1}^{d} \sigma_{m,i}$
3	Difference of intracategory standard deviation	$f_i^{\mathrm{inner}} = \max(\mathrm{clt}_{m,i})/\min(\mathrm{clt}_{c,i})$
4	Distance of each feature	$cd_{s,r,m,i} = \lvert q_{s,m,i} - q_{r,m,i} \rvert$
5	Quadratic sum of feature distance	$qs = \sum_{s,r=1}^{J} (cd_{s,r,m,i} - d_{m,i})^2$
6	Standard deviation of feature distance (intracategory cohesion)	$\mathrm{stc}_{m,i} = \sqrt{q_m / J(J-1) - 1}$
7	Average intercategory cohesion difference	$\mathrm{clt}_i^{\mathrm{outer}} = \sum_{m,c=1}^{d} \lvert stc_{m,i} - stc_{c,i} \rvert / d(d-1)$
8	Imparity measure of intercategory cohesion difference	$f_i^{\mathrm{outer}} = \max(\lvert \mathrm{stc}_{m,i} - \mathrm{stc}_{c,i} \rvert)/\min(\lvert \mathrm{stc}_{k,i} - \mathrm{stc}_{z,i} \rvert)$
9	Cohesion weighting factor	$c\omega_l = 1/f_i^{\mathrm{inner}}/\max(f_g^{\mathrm{inner}}) + f_i^{\mathrm{outer}}/\max(f_g^{\mathrm{outer}})$
10	Cohesion factor	$\gamma_i = c\omega_i \mathrm{clt}_i^{\mathrm{outer}}/\mathrm{clt}_i^{\mathrm{inner}}$

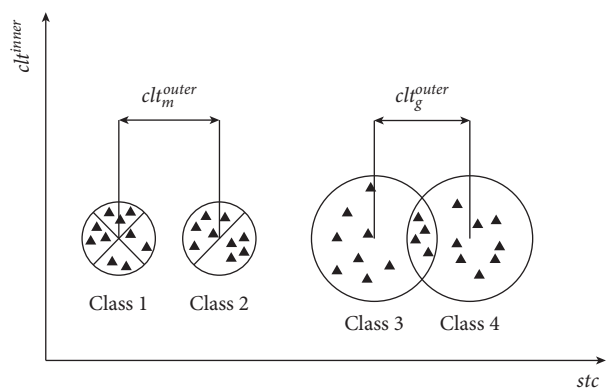


Figure 3: Cohesion evaluation.

of error function equation. (1) in BP-MLL algorithm will be smaller and smaller, which helps to rank labels belonging to an instance higher than those not belonging to.

$$\text{distance} = \text{func}\left(c_k^i, c_l^i\right) = c_k^i - c_l^i, \tag{10}$$

where c_k^i represents the output of relevant labels of i-th instance and c_l^i is the output of those irrelevant ones.

Nevertheless, as mentioned in Section 2, the conventional algorithm does not take full-labeled or nonlabeled situations into account. Full-label in an instance is shown in Figure 4 and the nonlabeled situation implies an instance does not have any marked labels.Mathematical failures such as an unreasonable denominator would occur during training or the trained neural network cannot attain an acceptable classification result on unseen cases if not considering those two situations. The most direct way is to modify the distance function so that the algorithm can let labels be as close to targets as possible while considering the characteristics of multilabel learning. Therefore, in this paper, modifications are made on the distance function to handle this problem. If an instance is not marked by any labels, that is, $|Y_i| = 0$, then there only exist irrelevant labels so that the distance function would be modified as

$$\text{distance} = \text{func}\left(-1, c_l^i\right) = -1 - c_l^i, \quad \text{if } \left|Y_i\right| = 0. \tag{11}$$

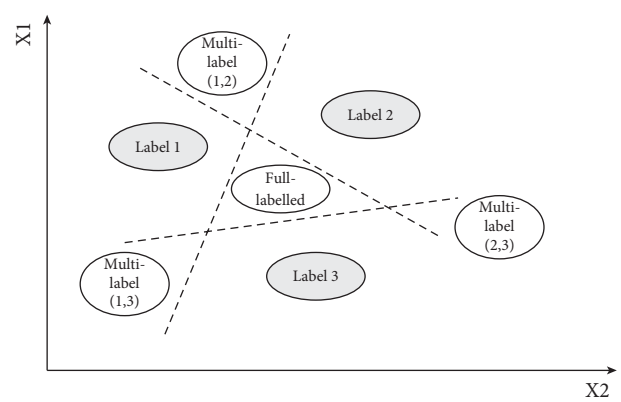


Figure 4: Labels for different instances.

Similarly, if all labels are marked in an instance, then these labels are all relevant so that the distance between 1 and these labels should be as small as possible:

$$\text{distance} = \text{func}\left(c_k^i, 1\right) = c_k^i - 1, \quad \text{if } |\pi| = 0. \tag{12}$$

The error function in BP-MLL algorithm is visualized in Figure 5, along with the modifications marked as red solid line and blue dotted line. From the perspective of image, two new distances functions based on full-labeled and

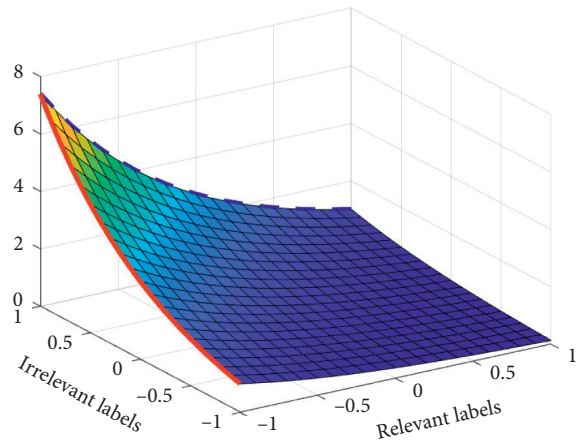


Figure 5: The error function of BP-MLL algorithm.

nonlabeled situations are derived from original error function, which indicates these would not generate conflicts in different types of training samples. Therefore, the improved error function in BP-MLL algorithm preserves the ability of discriminating relevant labels and irrelevant ones; meanwhile, it has capability of dealing with full-labeled or nonlabeled situations. Apart from the modifications on distance function, a regularization term is added to enhance the generalization ability. The main contribution in this paper is the improvement of BP-MLL algorithm to let it concentrate on both the correlations between different labels and the occurrence of empty sets by the following global error function:

$$E = \sum_{i=1}^{I} \frac{\sum_{(k,l)\in Y_i\times\overline{Y}_i} e^{-\left(\bar{c}_k^i - \bar{c}_l^i\right)}}{\max\left(1, \left|Y_i\right|\right)\max\left(1, \left|\overline{Y}_i\right|\right)} + \frac{\beta}{2}\sum_{d=0}^{N}\sum_{t=0}^{M}\sum_{s=1}^{J}\left(W_{td}^2 + W_{st}^2\right), \tag{13}$$

where β is the regularization coefficients and

$$\bar{c}_k^i = \begin{cases} c_k^i, & \left|Y_i\right| \neq 0, \\ -1, & \left|Y_i\right| = 0, \end{cases}$$
$$\bar{c}_l^i = \begin{cases} c_l^i, & \left|\overline{Y}_i\right| \neq 0, \\ 1, & \left|\overline{Y}\right|_i = 0 \end{cases} \tag{14}$$

Remark c_k represents the actual output of k-th output unit and this label belongs to this instance ($k \in Y_i$), while c_l represents the output of one label that does not belong to this instance ($l \notin Y_i$). Due to the property of the exponential term, the bigger the difference between c_k and c_l is, the smaller the global error is.

In conventional BP-MLL algorithm, saturation may occur due to the choice of activation function: sigmoid or hyperbolic tangent (tanh). Moreover, due to the exponential computation in the error function, using these two functions can also lead to high time complexity. Therefore, to avoid vanishing gradient problem and to reduce the difficulty of calculation, the improved method proposed in this paper adopts leaky rectified linear unit (Leaky ReLU) as the activation function:

$$h(x) = \begin{cases} x, & \text{if } x > 0, \\ \alpha_0 x, & \text{if } x \leq 0, \end{cases} \tag{15}$$

where $\alpha_0 \in (0, 1)$.

Let Sc_s represent the weighted sum; then the actual output of s-th output unit is

$$c_s = h\left(Sc_s\right). \tag{16}$$

In this paper, the gradient descent rule is adopted to adjust weights and bias until the global error converges to an acceptable value so that the updated W_{st} is

$$\begin{aligned} W_{st} &= W_{st} + \Delta W_{st}, \\ \Delta W_{st} &= -\alpha\left(\frac{\partial E_i}{\partial W_{st}} + \beta W_{st}\right), \\ &= -\alpha\frac{\partial E_i}{\partial Sc_s}\frac{\partial Sc_s}{\partial W_{st}} - \alpha\beta W_{st}, \end{aligned} \tag{17}$$

where α is the learning rate, and define temp_s as

$$\text{temp}_s = -\frac{\partial E_i}{\partial Sc_s} = -\frac{\partial E_i}{\partial c_s}\frac{\partial c_s}{\partial Sc_s}. \tag{18}$$

Let $h' = \partial c_s / \partial Sc_s$; then

$$h' = \frac{\partial c_s}{\partial Sc_s} = \begin{cases} 1, & \text{if } Sc_s > 0, \\ \alpha_0, & \text{if } Sc_s \leq 0. \end{cases} \tag{19}$$

Substituting equations (9) and (14) into (13):

$$\text{temp}_s = -\frac{\partial\left[\sum_{(k,l)\in Y_i\times\overline{Y}_i}\exp\left(-\left(\bar{c}_k - \bar{c}_l\right)\right)/\max\left(1, \left|Y_i\right|\right)\max\left(1, \left|\overline{Y}_i\right|\right)\right]}{\partial c_s}h'.$$

$$\text{temp}_s = \begin{cases} \dfrac{1}{\left|Y_i\right|}e^{-\left(c_s - 1\right)}h', & \text{if } \left|\overline{Y}_i\right| = 0, \\ -\dfrac{1}{\left|\overline{Y}_i\right|}e^{-\left(-1 - c_s\right)}h', & \text{if } \left|Y_i\right| = 0, \\ \left[\dfrac{1}{\left|Y_i\right|\left|\overline{Y}_i\right|}\sum\limits_{l\in\overline{Y}_i} e^{-\left(c_s - c_l\right)}\right]h', & \text{if } s \in Y_{,i} \\ \left[-\dfrac{1}{\left|Y_i\right|\left|\overline{Y}_i\right|}\sum\limits_{k\in Y_i} e^{-\left(c_k - c_s\right)}\right]h', & \text{if } s \in \overline{Y}_{\cdot,i} \end{cases} \tag{20}$$

So equation (12) can be rewritten as

$$\begin{aligned} \Delta W_{st} &= \alpha\text{temp}_s\left[\frac{\partial\sum_{t=1}^{M} b_t \cdot W_{st}}{\partial W_{st}}\right] - \alpha\beta W_{st}, \\ &= \alpha\text{temp}_s b_t - \alpha\beta W_{st}. \end{aligned} \tag{21}$$

A preset threshold ε is used for classification of each instance and fault detection, represented by res:

$$\text{res} = \begin{cases} 1, & \text{if } c_s > \varepsilon, \\ -1, & \text{if } c_s \leq \varepsilon. \end{cases} \tag{22}$$

The proposed algorithm for the multifault diagnosis of rotating machinery is summarized as Table 2 and the main contributions of the proposed algorithm are as follows: (1) optimize the BP-MLL algorithm by improved error cost function and regularization term; (2) propose a fault diagnostic algorithm based on the improved BP-MLL and cohesion evaluation; (3) perform an experimental study on the multilabel gearbox and bearing fault diagnosis in rotating machine under various working and environmental conditions.

4. Experimental Results

4.1. Experimental Platform. In this paper, wind turbine drivetrain diagnostics simulator (WTDDS) produced by SpectraQuest, USA, is used as the experimental platform. Figure 6 illustrates its operation diagram, in which label **a** represents the torque sensor set on the shaft, **b** and **d** are the vibration sensors, fixed above and to the left of the parallel shaft gearbox, respectively, and **c** is the pressure sensor. These sensors are connected to a multichannel signal acquisition device which can aggregate signals to a computer and convert them into voltage signals.

In Figure 7, the reference number 4 is a single-phase motor to power the entire system. The reference number 3 is a parallel shaft gearbox which can transmit the kinetic energy of the motor to the planetary gearbox by coupling with the motor. In the planetary gearbox (referred to by 2), four planetary gears are rotated under the traction of the driving wheel that can transmit kinetic energy to the load brake referred to by 1. The windmill (referred to by 6) then can be driven by the next stage of rotating shaft after the braking action of the load brake.

4.2. Experimental Setup. In reality, the motor frequency is affected by the mechanical structure and wind speed, and the load is related to the generator structure and voltage. To collect accurate and effective data, the combination of different motor frequency and load voltage is used to simulate different working conditions. Through the software LabView, the input voltage of the load controller and the speed controller is adjusted to realize manual control of shaft load and motor speed. Table 3 summarizes the six operating conditions considered and set in this experiment:

The proposed method is applied to classify the five types of faults in gears, which are ball bearing, outer bearing, inner bearing, chipped tooth, and missing tooth, respectively, as shown in Figure 8. A label of five-digit binary code represents the expected output value. In the experiment, the length of the feature vector for the signal segment is 2048, and the sampling time is 6.4s, and the sampling frequency is 5120 Hz. The categories of simultaneous faults and the samples assigned to training and testing procedures are shown in Table 4.

4.3. Experimental Results

4.3.1. Classification of Different Simultaneous Faults. The multilabel algorithms (improved BP-MLL, BP-MLL, and ML-KNN) and the conventional classification technique (BP neural network) were performed on the same data sets. In this paper, all experiments have been performed on a computer with 8G RAM and Intel® Core^{TM} i5-7200U CPU @ 2.70 GHz. The improved BP-MLL neural network has two hidden layers, the numbers of neurons in input, first hidden, second hidden, and output layers are 32, 72, 12, and 5, respectively. To approach the optimal solution and make the algorithm converge, the learning rate in improved BP-MLL neural network is set as $0.95^{\text{iter}} * \alpha_0$.

Table 5 summarizes the classification accuracy of different types of simultaneous faults and the total training time in terms of various algorithms. The ML-KNN algorithm required the least time to train and obtained a relatively high classification accuracy, but the classification accuracy of certain cases cannot be guaranteed, such as 00000 and 00101. Although the basic BP technique trained the neural network fast, it is limited to nonlabeled situation discriminations and fails to classify the second faulty types. As mentioned in Section 2, conventional BP-MLL algorithm is unable to deal with nonlabeled situations. Additionally, conventional BP-MLL spent more time than improved BP-MLL to train the same network. This is because the proposed method applied Leaky ReLU as the activation function that helps to reduce time consumption for calculation during the gradient descent process by judging the weighted sum in error function equation (9) firstly. The faults on outer bearing with chipped tooth confuses BP, BP-MLL, and ML-KNN. However, all algorithms for comparison can detect the faults on outer bearing with missing tooth. In general, the proposed method can achieve the accuracy of 100% with less training time than conventional BP-MLL method. Although the training time of the proposed method is longer than ML-KNN method, this is acceptable as the training process is conducted offline and the classification accuracy of the proposed method is consistently higher than ML-KNN method.

As a nonlinear dimensionality reduction algorithm, T-distribution stochastic neighbor embedding (t-SNE) technique proposed in [38] uses conditional probability to express the similarity of distance between data points, which is very applicable for dimensionality reduction. To visualize classification results directly, the three-dimensional mapping results are shown in Figure 9. Although data points in conventional BP-MLL have clear borders, the third type of simultaneous faults cannot be detected correctly due to the missing classification of one specific fault. Additionally, there are only a few points in ML-KNN; this is because the value of each output unit is the probability of each label. The predicted data instances after training in improved BP-MLL algorithm are basically distributed around their centers.

Table 2: Proposed fault diagnosis algorithm.

Training stage	
1	Obtain original sampling signals from multichannel sensors as training data.
2	Compute feature parameters for all the channels to construct a high-dimensional feature vector.
3	Compute sensitivity weighting factor β and sensitivity factor η using equations (6) and (7) and Table 1. Select parameters of high sensitivity factor to construct the sensitive feature vector.
4	Use the feature vector as input vector and modify weights and bias by using equations (10), (13), and (14) until the trained model can meet the test requirement of high accuracy or the maximum number of training epochs has been reached
Diagnosis Stage	
1	Construct the sensitive feature vector of new data.
2	Compute the outputs for each testing instance by epoch as shown in equations (11) and (10).
3	Use a preset threshold to classify each instance for fault diagnosis using equation (17)

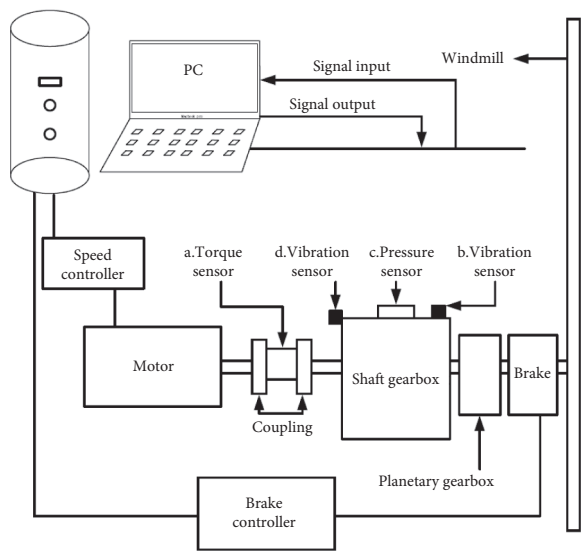


Figure 6: The operation diagram of WTDDS.

4.3.2. Comparison on Different Algorithms. To compare the performance of the proposed method with other conventional ones, the following six evaluation metrics are used to measure the classification results [39].

F1-score is also known as balanced F score, which is defined as the harmonic mean of precision and recall:

$$F1-\text{score} = \frac{1}{I}\sum_{i=1}^{I}\frac{2\left|Z_i \cap Y_i\right|}{\left|Z_i\right| + \left|Y_i\right|}, \tag{23}$$

where Z_i denotes the predicted labels and Y_i is the desired value of s-th label in i-th instance. Recall is the fraction of the correct labels expected from the actual labels, while precision is the fraction of labels correctly classified from the expected positive labels, averaged on all instances:

$$\begin{aligned}\text{Recall} &= \frac{1}{I}\sum_{i=1}^{I}\frac{\left|Z_i \cap Y_i\right|}{\left|Y_i\right|},\\ \text{Precision} &= \frac{1}{I}\sum_{i=1}^{I}\frac{\left|Z_i \cap Y_i\right|}{\left|Z_i\right|}.\end{aligned} \tag{24}$$

Hamming loss is used to investigate the misclassification of an instance on a single label; i.e., the correlation label does not appear in the predicted label set or the irrelevant label appears in the predicted label set. The smaller the value of Hamming loss is, the better the system performance is:

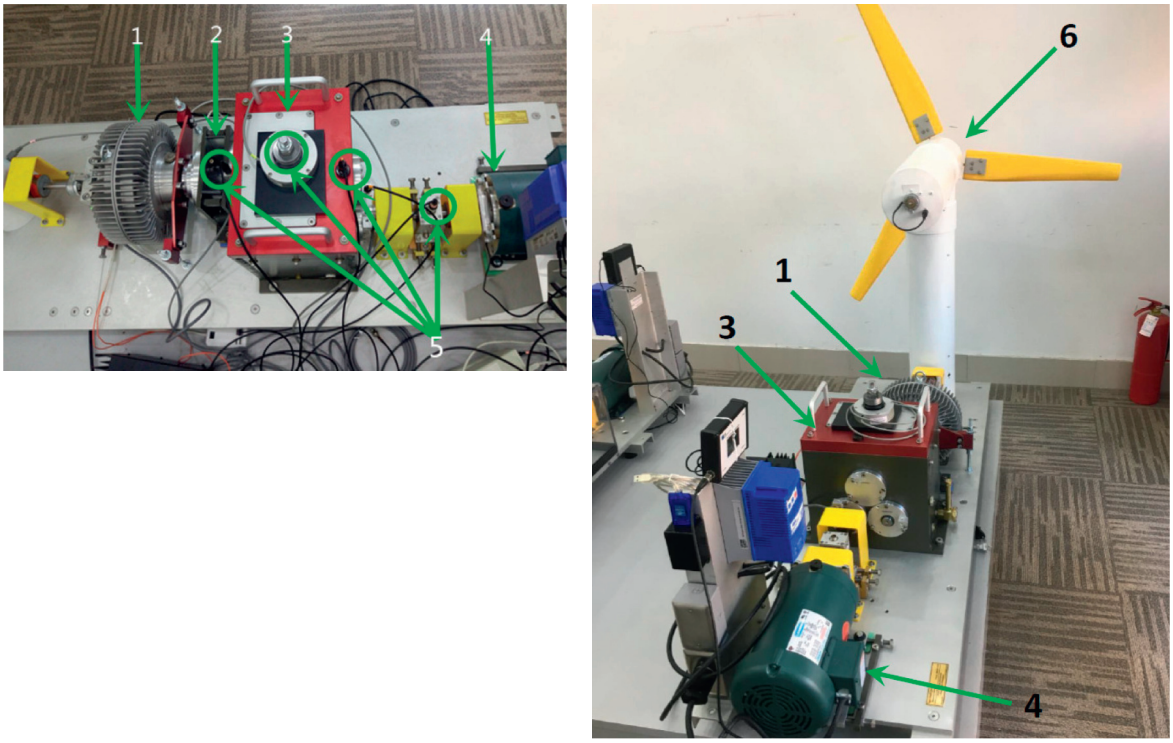
$$\text{Hamming loss} = \frac{1}{I}\sum_{i=1}^{I}\text{XOR}\left(Y_i, Z_i\right), \tag{25}$$

where XOR represents exclusive or.

Ranking loss evaluates the fraction of pairs of labels that are misclassified for the instance. The lower the values of this metric are, the better the performance is:

$$\text{Ranking loss} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{\left|Y_i\right|\left|\overline{Y}_i\right|}\left|E\right|, \tag{26}$$

where $|E|$ is the size of error-set and $E = \{(\lambda, \lambda')|\tau_i(\lambda) \le \tau_i(\lambda'), (\lambda, \lambda') \in Y_i \times \overline{Y_i}\}$, and $\tau_i(\cdot)$ is the true value of output before being labeled.



(a) (b)

Figure 7: The structure of WTDDS (1, brake; 2, planetary gearbox; 3, shaft gearbox; 4, motor; 5, four sensors (from left to right: vibration, force, vibration, and torque); 6, windmill).

Table 3: Working conditions settings.

Motor frequency (Hz)	Load voltage (V)	Condition category
6	8	A
6	5	B
10	8	C
10	5	D
14	8	E
14	5	F

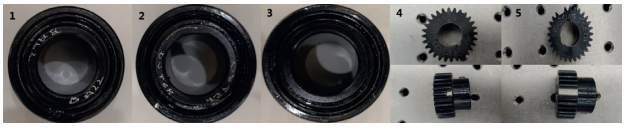


Figure 8: Five types of gears.

Table 4: Samples and labels.

Fault category	Number of training samples	Number of testing samples	Category code
Normal	48	10	00000
Ball bearing + missing tooth	48	10	10010
Outer bearing + missing tooth	48	10	00110
Outer bearing + chipped tooth	48	10	00101
Inner bearing + chipped tooth	48	10	01001

Table 5: Accuracy for different simultaneous faults.

Fault category	Algorithms			
	Improved BP-MLL	BP	BP-MLL	ML-KNN
00000	100.00%	66.67%	—	89.58%
10010	100.00%	0.00%	100%	100.00%
00110	100.00%	100.00%	100%	100.00%
00101	100.00%	83.33%	18.75%	83.33%
01001	100.00%	95.83%	100%	100.00%
Grand total	100.00%	69.17%	79.69%	94.58%
Training time (s)	105.1355	25.4463	787.2271	3.3972

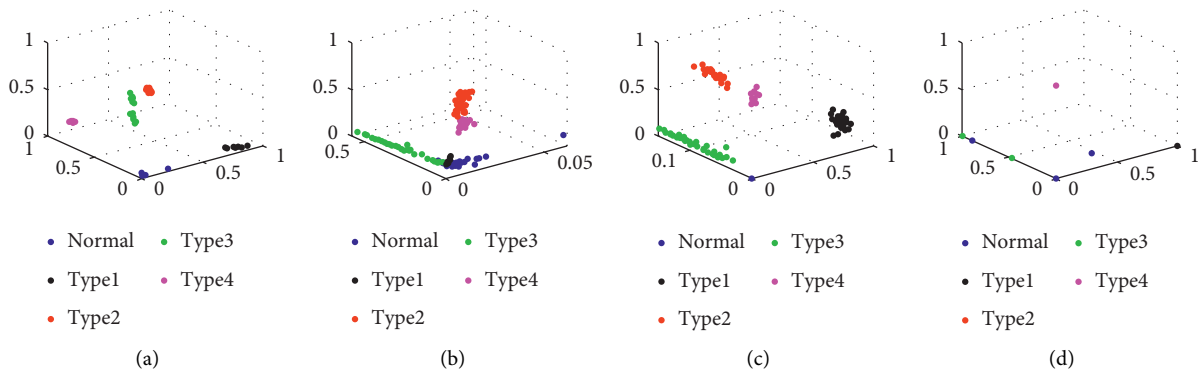


Figure 9: The classification results of different algorithms.

Table 6: Performance comparison of different algorithms.

Evaluation metrics	Algorithms			
	Improved BP-MLL	BP	BP-MLL	ML-KNN
Average precision	1	0.9155	0.9905	1
Hamming loss	0	0.1033	0.0406	0.0217
One error	0	0	0	0
F1-score	1	0.8967	0.9594	0.9783
Coverage	0.6	1.2292	1.0625	0.6
Ranking loss	0	0.1311	0.0104	0

Average Precision, which is also called classification accuracy or exact match ratio, computes the percentage of instances whose predicted labels are exactly the same as the actual corresponding set of labels:

$$\mathrm{AP} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{|Y_i|}\sum_{\lambda\in Y_i}\frac{\lambda' \in Y_i \left|\tau_i(\lambda') \le \tau_i(\lambda')\right|}{\tau_i(\lambda)}. \quad (27)$$

One error describes the possibility that top-ranked labels in one instance are not the actual labels in the proper set. The smaller the value of one error is, the better the system performance is

$$\mathrm{One-error} = \frac{1}{I}\sum_{i=1}^{I}\left[\arg\max_{\lambda\in y}\tau_i(\lambda) \notin y_i\right]. \quad (28)$$

Coverage measures how far the traversal all the labels is in the ranking averagely associated with an instance. The smaller the value is, the better the performance is

$$\mathrm{Coverage} = \frac{1}{I}\sum_{i=1}^{I}\max_{\lambda\in Y_i}\tau_i(\lambda) - 1. \quad (29)$$

Table 6 summarizes the performance of each algorithm based on the evaluation metrics. All four algorithms get high f1-score: 1, 0.8967, 0.9594, and 0.9783, separately; particularly, improved BP-MLL obtained the highest one. The values of average precision are all above 0.9. Additionally, all these classifiers can recognize the relevant labels in each instance which leads to one error at 0. Improved BP-MLL and ML-KNN shared the same figures in coverage and ranking loss. Nevertheless, values of coverage for basic BP and conventional BP-MLL algorithms are twice as high as those for both improved BP-MLL and ML-KNN. In comparison, the Hamming loss of improved BP-MLL is slightly lower than ML-KNN, which is because ML-KNN shows misclassifications on single labels in some instances. Although values of different metrics vary, multilabel learning

algorithm can predict the simultaneous faults more effectively, among which improved BP-MLL outperforms the other compared two methods.

5. Conclusion

In this paper, an improved fault diagnostic method based on cohesion evaluation and improved BP-MLL classification is proposed. Compared with the conventional single-fault diagnosis, the problem of simultaneous faults occurring on gearbox and bearings of rotating machinery under different environmental conditions is investigated. On the basis of BP-MLL, this paper proposes a new global error function to deal with full-label and nonlabel learning situations through modifying its distance function and enhancing the generalization ability. Experiments conducted on WTDDS show that the proposed method is superior to conventional methods under six performance evaluation metrics. Although this paper has achieved good experimental results, there are still limitations such as the current algorithm being supervised learning only. Therefore, further studies can be focused on improvements such as semisupervised learning based on partial labels and transfer learning based on different working conditions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

An earlier version of this manuscript has been presented in 2020 IEEE 29th International Symposium on Industrial Electronics as An Improved Simultaneous Fault Diagnosis Method Based on Cohesion Evaluation and BP-MLL for Rotating Machinery.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (61603223), Jiangsu Provincial Qinglan Project, Research Development Fund of XJTLU (RDF-18-02-30, RDF-20-01-18), Key Program Special Fund in XJTLU (KSF-E-34), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (20KJB520034).

References

[1] I. El-Thalji and E. Jantunen, "A summary of fault modelling and predictive health monitoring of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 60-61, pp. 252–272.

[2] Y. Lei, J. Lin, M. J. Zuo, and Z. He, "Condition monitoring and fault diagnosis of planetary gearboxes: a review," *Measurement*, vol. 48, pp. 292–305, 2014.

[3] D. Cabrera, A. Guamán, S. Zhang et al., "Bayesian approach and time series dimensionality reduction to LSTM-based model-building for fault diagnosis of a reciprocating compressor," *Neurocomputing*, vol. 380, pp. 51–66, 2020.

[4] J.-H. Zhong, J. Zhang, J. Liang, and H. Wang, "Multi-fault rapid diagnosis for wind turbine gearbox using sparse bayesian extreme learning machine," *IEEE Access*, vol. 7, pp. 773–781, 2019.

[5] R. Maamouri, M. Trabelsi, M. Boussak, and F. M'Sahli, "Mixed model-based and signal-based approach for open-switches fault diagnostic in sensorless speed vector controlled induction motor drive using sliding mode observer," *IET Power Electronics*, vol. 12, no. 5, pp. 1149–1159, 2019.

[6] D. Lieftucht, M. Völker, C. Sonntag, U. Kruger, G. W. Irwin, and S. Engell, "Improved fault diagnosis in multivariate systems using regression-based reconstruction," *Control Engineering Practice*, vol. 17, no. 4, pp. 478–493, 2009.

[7] M. Pirra, E. Gandino, A. Torri, L. Garibaldi, and J. M. Machorro-López, "PCA algorithm for detection, localisation and evolution of damages in gearbox bearings," *Journal of Physics: Conference Series*, vol. 305, Article ID 012019, 2011.

[8] M. Onel, C. A. Kieslich, Y. A. Guzman, and E. N. Pistikopoulos, "Simultaneous fault detection and identification in continuous processes via nonlinear support vector machine based feature selection," in *Computer Aided Chemical Engineering*, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds., vol. 44pp. 2077–2082, 2018.

[9] Z. Wang, M. Zhong, R. Yang, and Y. Liu, "An improved random forest algorithm of fault diagnosis for rotating machinery," in *Proceedings of the 2019 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, pp. 12–17, IEEE, Xiamen, China, July 2019.

[10] Q. Lu, R. Yang, M. Zhong, and Y. Wang, "An improved fault diagnosis method of rotating machinery using sensitive features and RLS-BP neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1585–1593, 2019.

[11] Y. Xue, D. Dou, and J. Yang, "Multi-fault diagnosis of rotating machinery based on deep convolution neural network and support vector machine," *Measurement*, vol. 156, Article ID 107571, 2020.

[12] R. Yang, P. V. Er, Z. Wang, and K. K. Tan, "An RBF neural network approach towards precision motion system with selective sensor fusion," *Neurocomputing*, vol. 199, pp. 31–39, 2016.

[13] R. Yang, M. Huang, Q. Lu, and M. Zhong, "Rotating machinery fault diagnosis using long-short-term memory recurrent neural network," *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 228–232, 2018.

[14] R. Yang, K. K. Tan, A. Tay et al., "An RBF neural network approach to geometric error compensation with displacement measurements only," *Neural Computing & Applications*, vol. 28, no. 6, pp. 1235–1248, 2017.

[15] Z. Wan, R. Yang, and M. Huang, "Deep transfer learning-based fault diagnosis for gearbox under complex working conditions," *Shock and Vibration*, vol. 2020, Article ID 8884179, 13 pages, 2020.

[16] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, 2021.

[17] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: a review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33–47, 2018.

[18] L. Ungar, B. Powell, and S. Kamens, "Adaptive networks for fault diagnosis and process control," *Computers & Chemical Engineering*, vol. 14, no. 4, pp. 561–572, 1990.

[19] M. A. Kramer and J. A. Leonard, "Diagnosis using back-propagation neural networks-analysis and criticism," *Computers & Chemical Engineering*, vol. 14, no. 12, pp. 1323–1338, 1990.

[20] G. M. Knapp and H.-P. Wang, "Machine fault classification: a neural network approach," *International Journal of Production Research*, vol. 30, no. 4, pp. 811–823, 1992.

[21] M. R. G. Meireles, P. E. M. Almeida, and M. G. Simoes, "A comprehensive review for industrial applicability of artificial neural networks," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 3, pp. 585–601, 2003.

[22] S. Wu and T. W. S. Chow, "Induction machine fault detection using som-based RBF neural networks," *IEEE Transactions on Industrial Electronics*, vol. 51, no. 1, pp. 183–194, 2004.

[23] S. Haroun, A. Nait Seghir, and S. Touati, "Feature selection for enhancement of bearing fault detection and diagnosis based on self-organizing map," in *Recent Advances in Electrical Engineering and Control Applications*, pp. 233–246, Springer International Publishing, New York, NY, USA, 2017.

[24] S.-L. Jämsä-Jounela, M. Vermasvuori, P. Endén, and S. Haavisto, "A process monitoring system based on the Kohonen self-organizing maps automation in Mining, Mineral and Metal Processing," *Control Engineering Practice*, vol. 11, no. 1, pp. 83–92, 2003.

[25] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 327–346, 2003.

[26] A. Dineva, A. Mosavi, M. Gyimesi, I. Vajda, N. Nabipour, and T. Rabczuk, "Fault diagnosis of rotating electrical machines using multi-label classification," *Applied Sciences*, vol. 9, no. 23, p. 5086, 2019.

[27] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[28] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[29] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pp. 681–687, Granada, Spain, December 2002.

[30] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[31] J.-H. Zhong, P. K. Wong, and Z.-X. Yang, "Fault diagnosis of rotating machinery based on multiple probabilistic classifiers," *Mechanical Systems and Signal Processing*, vol. 108, pp. 99–114, 2018.

[32] M. Zhi-Hua Zhou and Z. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[33] K. Zhang, H. Ma, Y. Zhao, H. Zan, and L. Zhuang, "The comparative experimental study of multilabel classification for diagnosis assistant based on Chinese obstetric emrs," *Journal of Healthcare Engineering*, vol. 2018, Article ID 7273451, 9 pages, 2018.

[34] G.-P. Liu, G.-Z. Li, Y.-L. Wang, and Y.-Q. Wang, "Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning," *BMC Complementary and Alternative Medicine*, vol. 10, no. 1, p. 37, 2010.

[35] Q. Yang, J. Shao, M. Scholz, C. Boehm, and C. Plant, "Multi-label classification models for sustainable flood retention basins," *Environmental Modelling & Software*, vol. 32, pp. 27–36, 2012.

[36] W. Long, K. Zhang, H. Ma, D. Yue, and L. Zhuang, "Neural network multi-label learning based on enhancing pairwise labels discrimination for obstetric auxiliary diagnosis," in *Proceedings of the 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp. 297–2977, Zhengzhou, China, October 2018.

[37] R. Grodzicki, J. Mańdziuk, and L. Wang, "Improved multi-label classification with neural networks," *Parallel Problem Solving from Nature - PPSN X*, vol. 5199, no. 9, pp. 409–416, 2008.

[38] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, 2012.

[39] E. Gibaja and S. Ventura, "A tutorial on multi-label learning," *ACM Computing Surveys*, vol. 47, p. 04, 2015.

Hindawi
Shock and Vibration
Volume 2021, Article ID 7383255, 21 pages
https://doi.org/10.1155/2021/7383255

Research Article

A New Transferable Fault Diagnosis Approach of Rotating Machinery Based on Deep Autoencoder and Dominant Features Selection under Different Operating Conditions

Fei Dong,[1] **Xiao Yu**,[2,3] **Xinguo Shi**,[4] **Ke Liu**,[4] **Zhaoli Wu**,[5] **and Wanli Yu**[6]

[1]*School of Internet, Anhui University, Hefei 230039, China*
[2]*IOT Perception Mine Research Center, China University of Mining and Technology, Xuzhou 221000, China*
[3]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221000, China*
[4]*Center of Information Technology, Zibo Mining Group Co, Ltd., Shandong, Zibo 255000, China*
[5]*Jiangsu Collaborative Innovation Center for Building Energy Saving and Construction Technology, Xuzhou 221116, China*
[6]*Institute of Electrodynamics and Microelectronics, University of Bremen, Bremen 28359, Germany*

Correspondence should be addressed to Xiao Yu; yxcumt2006@163.com

Received 29 July 2021; Accepted 15 September 2021; Published 30 October 2021

Academic Editor: Jun Zhu

In the actual industrial scenarios, most existing fault diagnosis approaches are faced with two challenges, insufficient labeled training data and distribution divergences between training and testing datasets. For the above issues, a new transferable fault diagnosis approach of rotating machinery based on deep autoencoder and dominant features selection is proposed in this article. First, maximal overlap discrete wavelet packet transform is applied for signals processing and mix-domains statistical feature extraction. Second, dominant features selection by importance score and differences between domains is proposed to select dominant features with high fault-discriminative ability and domain invariance. Then, selected dominant features are used for pretraining deep autoencoder (source model), which helps in enhancing the fault representative ability of deep features. The parameters of the source model are transferred to the target model, and normal state features from target domain are adopted for fine-tuning the target model. Finally, the target model is applied for fault patterns classification. Motor and bearing fault datasets are used for a series of experiments, and the results verify that the proposed methods have better cross-domain diagnosis performance than comparative models.

1. Introduction

With the prompt progress of modern industry, rotating machinery (RM) is developing towards integration and complexity [1]. RM usually operates under complex and harsh scenes, such as variable heavy loads, high temperature and speed, and strong impact [1]. Once a component fault occurs, it may further lead to the damage of other components and huge economic loss. Therefore, it is meaningful and practical to study intelligent fault diagnosis models towards real industrial scenes [1, 2].

In the last several years, intelligent fault diagnosis field has received many studies of traditional-machine-learning- (TML-) based framework, deep- learning (DL-) based framework, and transfer-learning- (TL-) based framework, which can achieve automatically fault recognition and classification by analysing massive signals collected from mechanical equipment [1–4]. TML-based framework is often constructed by using traditional machine learning algorithms that mainly include k-nearest neighbour (KNN) [5], support vector machine (SVM) [6], artificial neural network (ANN) [7], extreme learning machine (ELM) [8], decision tree (DT) [9], and some variations of them. Generally, TML-based framework consists of three steps: signal process and features extraction, features selection or reduction, and fault classification [2, 4, 8]. Vibration signals are the most commonly used for fault diagnosis, due to the strong nonlinearity and nonstationarity. Time-frequency analysis method is widely

applied for signal process and feature extraction, such as empirical mode decomposition (EMD) [10–12], short-time Fourier transform (STFT) [13–15], wavelet packet transform (WPT) [16–19], and some variations of them. References [20–25] and [26–28] applied the variations or improvement of EMD, STFT, and WPT for fault signals processing and feature extraction. The above-mentioned time-frequency analysis method can effectively help to extract fault features, but it often leads to a high dimensional feature set which contains interference and redundancy features. Thus, feature reduction and selection are a crucial step before the fault patterns classification [3, 6, 10, 18]. In [3], the extreme gradient promotion is used for the dimensional reduction and sensitive features selection, which applies the importance of features to refine a high quality feature subset. In [6], an ant colony algorithm is applied to select features, and the selected feature subset is combined with parameter optimized SVM for enhancing the generalization of the fault diagnosis model. In [10], indexes of the cohesion and class discriminative of features are used for evaluating features, and through the combination of these two indexes, a new index, ASR (the ratio of the adjusted rand index and standard deviation), was proposed to refine the original feature set. In [18], a sensitive feature selection method and modified features dimensionality reduction method are combined to obtain a low-dimensional features subspace, which improves the diagnosis accuracy. In TML-based framework, the traditional KNN, SVM, and ANN are widely used for constructing fault classification models by researchers. For example, references [29–31], [6, 10, 18], and [7, 32, 33] applied the KNN, SVM, and ANN for fault classification, respectively. Moreover, many variations of KNN, SVM, and ANN have been studied and applied for rotating machinery fault diagnosis. In [34], an enhanced KNN (EKNN) was designed to get embedded in a dimension-reduction stage; then, by using sparse filtering, some fault-discriminative features can be extracted. In [35], in order to determine the parameter K of KNN, an improved binary particle swarm optimization was proposed to select this parameter, which construct the IBPSO-KNN for bearing fault diagnosis. In [36], due to the fact that it is difficult for the traditional squares SVM to deal with complex imbalanced data, an improved SVM, a moth-flame optimization-based LS-SVM, was proposed for bearing fault diagnosis with complex imbalanced data. In [37], SVM was optimized by intercluster distance in the feature space, which was combined with improved symplectic geometry mode decomposition to design a novel fault diagnosis scheme for rotating machinery. In [38], a perceptron multilayer ANN (MLP-ANN) is used for detecting the bearing faults. However, the main limitation of TML-based framework relies heavily on expert knowledge when the diagnosis models need to be customized for different operating states and machines [4, 39].

Deep learning algorithms have received more and more attentions because they possess a powerful hidden features automatic mining ability [4, 40]. Therefore, DL-based framework has been widely studied in the intelligent fault diagnosis field. In references [13, 19, 39, 41], convolutional neural network (CNN) [13], deep belief network (DBN) [19], deep neural networks (DNN) [39], and deep autoencoders (DAE) [41] are used for constructing fault diagnosis models, respectively. However, some main limitations exist in most DL-based frameworks [3, 4, 40, 42, 43]. (1) Most conventional DL-based frameworks have insufficient generalization ability towards the engineering practical scenes; the reason is that an assumption of having the same distribution between the training and testing sets was widely used. In practical scenes, the real fault signals collected from machineries are inconsistent under variable operating states, which brings the distribution divergence of datasets. (2) They rely heavily on a mass of labeled data. When the labeled training data are not enough, the overfitting phenomenon may easily appear, which can lead to the reduction of diagnosis accuracy and stability. Facing at the engineering practical scenes, the sufficient labeled data are difficult to obtain due to the changeable and complex working conditions of machineries. Thus, how to enhance the fault representative ability of deep features and the stability of fault diagnosis models across different working conditions is still a challenging task.

Aiming at the limitations of DL-based frameworks mentioned above, a recent developed technique, domain adaptation under TL-based framework, intends to promote classifier learning by using labeled source domain data. At present, TL-based framework has became a study hot topic and employed for fault diagnosis of machinery [2–4, 20, 39, 42]. In the fault diagnosis, one working condition (a specific speed or load) can compose a domain. Source domain is a labeled dataset under one working condition, and the target domain is unlabeled data under another working condition. The object of domain adaptation under TL-based framework is that labeled source domain and unlabeled target domain are used to learn a cross-domain diagnosis model which can achieve desirable fault classification results of the target domain [4, 42]. Considering that deep learning methods have powerful ability to mine hidden features from original data, recently, deep TL models have been researched for cross-domain fault diagnosis of rotating machinery by many researchers. In [44], an enhanced DAE was designed by modifying the loss function, which improves the reconstruction performance of a decoder. Sufficient labeled data from source domain were employed for training the enhanced DAE model, and the corresponding parameters were transferred to target DAE model. In [45], a deep CNN with an attention mechanism was adopted for feature extraction, and a domain transformation algorithm was designed to match the distributions between source and target domains. In [46], a novel DAE model, deep transfer multiwavelet AEs, was designed for gearbox fault diagnosis by using little training samples. In this model, important features were learned by very few samples, and parameters of source model are directly migrated to target model. Although deep TL models have achieved many successful applications on cross-domain fault diagnosis of rotating machinery, how to enhance the fault representative ability of deep features and the stability of fault diagnosis models across different working conditions is still a challenging task [43]. For this issue, in this article, we propose a new transferable fault diagnosis approach of rotating machinery based on DAE and dominant features selection under different operating conditions (TFDD). In TFDD, the first step

is vibration signals processing and statistical features extraction, maximal overlap discrete wavelet packet transform (MODWPT) is used to decompose raw signals, and MODWPT is a time-frequency analysis method based on wavelet. Its advantages include two aspects: (1) it can overcome the limitation of discrete wavelet transform (DWT); that is, DWT requires the sample size to be exactly a power of 2 for the full transform because of the downsampling step; (2) MODWPT can overcome another problem that the DWT has very poor frequency resolution at low frequencies. Considering the above advantages of MODWPT, in our previous study [18, 19], MODWPT has been used for bearing fault diagnosis and compared with WPT; the performance of MODWPT is better than WPT. The second step is dominant features selection; a new feature selection method, dominant feature selection by importance score and domain differences (DSID), is proposed to evaluate the fault-discriminative ability and domain invariance of feature, which can help in enhancing the fault representative ability of deep hidden features obtained by DAE. The third step is to construct deep transfer autoencoder (DTAE) model. A DAE model (source model) is trained by feature data from source domain and the learned parameters transferred to the target model that has the same architecture as source model. Then, the normal state feature data from target domain are applied for fine-tuning the target model. The last step is that the learned DTAE model is applied to diagnose the unlabeled fault features from target domain and output fault identification accuracies. The main contributions of this article are organized as follows.

(1) A new dominant feature selection method DSID: firstly, based on the raw feature dataset, the sufficient labeled feature data from source domain under all fault states is used to evaluate the fault-discriminative ability of features by random forest, and the importance score of features can be obtained to quantify the fault-discriminative ability. Secondly, the normal state feature data from the source and target domains are used to evaluate the domain invariance of features by computing the maximum mean discrepancy. Finally, the proposed new dominant features selection index, RIM, is constructed.

(2) A DTAE model is learned by dominant features. For enhancing the fault representative ability of deep features and the stability of fault diagnosis models, we apply the DSID to select dominant features with high fault-discriminative ability and domain invariance to train a DTAE model, which is expected to enhance the fault representative ability of deep features.

(3) A series of experiments are performed by using a motor and bearing faults datasets sampled from SQI-MFS test platform. The experimental results prove the availability, flexibility, and advantages of the TFDD.

The remaining contents of this article are organized as follows. Section 2 discusses the introduction of preliminary knowledge. Section 3 presents the proposed DSID and fault diagnosis framework TFDD. Experimental verification is given in Section 4. Section 5 concludes this paper.

2. Preliminaries

2.1. Deep Autoencoder (DAE). DAE is an unsupervised deep neural network [44], which is constructed by stacking several basic autoencoders (AE). Each AE has two steps: encoder and decoder; the structure of AE is presented in Figure 1. There are three layers: input layer $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, hidden layer $\mathbf{h} = \{h_1, h_2, \ldots, h_n\}$, and output layer $\mathbf{z} = \{z_1, z_2, \ldots, z_m\}$.

In the step of encoder, the input data $\mathbf{x}$ is mapped into the data of hidden layer $\mathbf{h}$ by the activation function $\varphi_{\text{act}}(\bullet)$; the mapping process is shown as the following expression:

$$\mathbf{h} = \varphi_{\text{act}}(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}), \tag{1}$$

where W and b are weight matrix and bias vector of encoder, respectively.

In the step of decoder, the data of hidden layer h are mapped into the output data z by the activation function $\varphi_{\text{act}}(\bullet)$; the mapping process is presented as the following expression:

$$\mathbf{z} = \varphi_{\text{act}}\left(\mathbf{W}' \cdot \mathbf{h} + \mathbf{b}\prime\right), \tag{2}$$

where $\mathbf{W}'$ and $\mathbf{b}'$ are weight matrix and bias vector of decoder, respectively. The hidden layer is the new feature representation and the output data are the reconstruction of the input data. The parameters of an AE model are learned by minimizing reconstruction error between the input and output layers; the reconstruction error is expressed as follows:

$$J_{\text{AE}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left\|x_i - z_i\right\|^2, \tag{3}$$

where N is the number of samples and $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{W}', \mathbf{b}',\}$.

The hidden layer features of an AE are used as the input of the next AE, which can stack multiple AEs to construct a DAE. The structure of a DAE is presented in Figure 2. DAE has the strong ability to mine deep features of input feature data so that it can improve accuracy of fault classification [41].

2.2. Random Forest (RF)- Based Feature Selection. Random forest (RF), firstly proposed by Breiman [47], is one of ensemble classifiers that obtained wide attentions by researchers. RF can achieve desirable performance for classification and regression tasks with high dimensional and ill-posed feature dataset [48]. The mainly idea of RF is to construct some unbiased decision trees (DT) by using the randomly selected samples, where each tree votes for a class and the forest chooses the classification having the most votes over all the trees [47, 48].

Given a dataset $\mathbf{S} = \{(\mathbf{X}, \mathbf{Y}), \mathbf{X} = \{x_i\} \in \mathbf{R}^D, \mathbf{Y} = \{y_i\} \in \{1, 2, \ldots, C\}\}_{i=1}^N$, where x_i is a feature sample with D dimension and y_i represents the class label, C and N are respectively the number of classes and training samples. RF algorithm is usually described as follows [49].

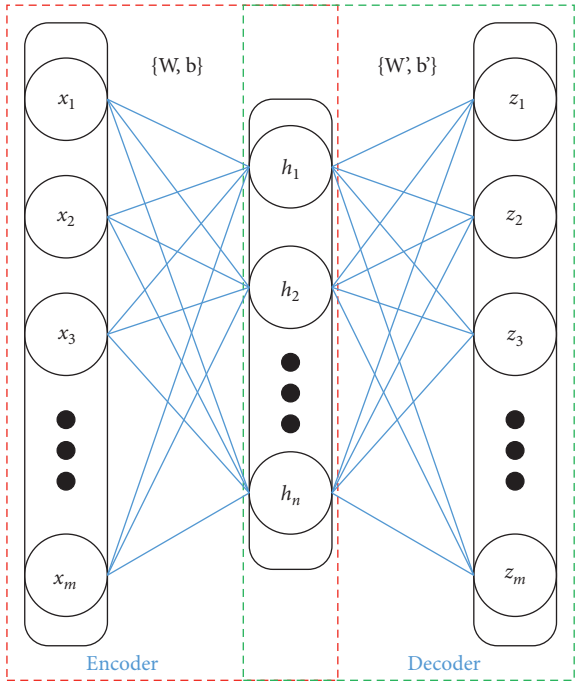


Figure 1: The structure of an AE.

(1) The M (the number of decision trees) bootstrap datasets are drawn from the training sample **S** by using bagging [50].

(2) For each bootstrap datasets, a decision tree is constructed by employing the Classification and Regression Tree (CART) algorithm [51]. At each node of DT, a subspace including p features is sampled and split points based on this subspace are computed. Then, the best split, for example, by the maximum Gini parameters impurity decrease, is applied to segment the data and grow the tree. That is, all data are pure with regard to the class.

(3) M DT are combined into a RF ensemble, and a majority vote manner is used to make the classification decision.

In RF algorithm, the performance and the diversity of DT can affect the performance of RF. A generalization error of RF is defined as

$$\text{error} = \frac{\overline{\rho}\left(1 - s^2\right)}{s^2}, \tag{4}$$

where $\overline{\rho}$ and s^2 represent the average correlation between DT and the average strength of DT, respectively. In addition, one of the important properties of RF is that the importance score (IS) of feature can be measured. For the high dimensional feature data, IS can be used to select relevant, compact, and discriminative features, which can help to improve the performance of classification. The Gini index (GI) is used to construct DT and determine the class in each tree [47, 49]. The GI at node t, GI(t), is applied to quantify the impurity of node t; the expression is defined as follows:

$$\text{GI}(t) = \sum_{i=1}^{I} \text{frac}_i \left(1 - \text{frac}_i\right), \tag{5}$$

where frac_i represents the fraction of category and i records at node t. Based on the GI, the GI information gain (GIG) of feature x_i, which is used to separate node t, is expressed as follows:

$$\text{GIG}\left(x_i, t\right) = \text{GI}(t) - \left(W_L \text{GI}\left(t^L\right) + W_R \text{GI}\left(t^R\right)\right), \tag{6}$$

where t^L and t^R are respectively the left and right child nodes of node t and W_L and W_R are respectively the corresponding fraction. Moreover, the IS of feature x_i can be obtained by calculating the following expression:

$$\text{IS}\left(x_i\right) = \frac{1}{n_{\text{DT}}} \sum_{k \in S_{x_i}} \text{GIG}\left(x_i, t\right), \tag{7}$$

where n_{DT} represents the number of DT in RF and $k \in S_{x_i}$ represents the set of split nodes. Finally, in the original high dimensional feature set, according to the IS of each feature, the features with high IS value can be selected to construct feature subset, so that many features with small IS are eliminated and improve the performance of classification.

2.3. Maximum Mean Discrepancy (MMD). Given two feature datasets $D_S = \left\{x_1, x_2, \ldots, x_{n_S}\right\}$ (source domain) and $D_T = \left\{y_1, y_2, \ldots, y_{n_T}\right\}$ (target domain) drawn from two different probability distributions, $P_S(D_S) \neq P_T(D_T)$, where n_S and n_T are respectively the number of D_S and D_T. For the purpose of estimating the distance between two distributions, MMD [52] was introduced by Gretton et al. for measuring distance of distributions based on reproducing

kernel Hilbert space (RKHS). The empirical distance estimate of distributions between D_S and D_T is defined as the following expression [53]:

$$\mathrm{Dist}\left(D_S, D_T\right) = \left\| \frac{1}{n_S} \sum_{x_i \in D_S} \phi\left(x_i\right) - \frac{1}{n_T} \sum_{x_j \in D_T} \phi\left(x_j\right) \right\|_H^2, \quad (8)$$

where $\|\bullet\|_H$ represents the RKHS norm and ϕ is the kernel-induced feature map. For the issue that the inconsistent feature distribution exists in fault diagnosis across variable operating conditions, based on the above-mentioned description, the MMD can be applied to estimate the discrepancy of two distributions and align these two distributions.

3. Proposed Method and System Framework

3.1. Dominant Feature Selection by Importance Score and Domain Differences (DSID). In order to reduce redundant features in the high dimensional raw feature set (HRFS) and select dominant features (fault-discriminative but operating-condition-invariant (FDOCI) features), we suppose that features should be evaluated from two aspects: fault-discriminative ability and domain invariance. Therefore, a new feature selection approach, dominant feature selection by importance score and domain differences (DSID), is proposed in this article. In DSID, firstly, the RF is employed to quantify the fault-discriminative ability of each feature based on labeled source domain data. Secondly, the MMD is used to evaluate the domain invariance of each feature based on normal state data of source and target domain data. Finally, a new selection index, the ratio of IS and MMD (RIM), is constructed to select dominant features for enhancing the performance of fault diagnosis across different operating conditions. The specific description of the DSID is summarized as follows.

(a) *Compute Importance Score of Features.* Given a raw feature set (RFS) $[f_1, f_2, \ldots, f_p]^T$ of source domain that contains p feature samples, each sample has q features; that is, $f_i = \{f_i^1, f_i^2, \ldots f_i^q\}$, $i \in [1, p]$. Let IS(k) denote the importance score of the k-th feature, according to the introduction of RF-based feature selection in Section 2.2. IS (k) can be calculated by (5)–(7); thus, the sequence $\mathrm{IS} = \{\mathrm{IS}(1), \mathrm{IS}(2), \ldots, \mathrm{IS}(q)\}$ of q features can be obtained. In this paper, we suppose that the fault-discriminative ability of feature is greater when the value of IS is larger.

(b) *Evaluate the Domain Invariance of Features.* MMD is employed to estimate the distribution discrepancy of the same feature in different domains, and the value of MMD is used as the quantitative index of domain invariance of feature. Let nf_S and nf_T denote normal state feature data from source and target domains, respectively. Both nf_S and nf_T consist of p feature samples, and each sample has q features. The expressions of nf_S and nf_T are presented as follows:

$$nf_S = \begin{bmatrix} nf_S^{11} & nf_S^{12} & \cdots & nf_S^{1q} \\ nf_S^{21} & nf_S^{22} & \cdots & nf_S^{2q} \\ \vdots & \vdots & \ddots & \vdots \\ nf_S^{p1} & nf_S^{p2} & \cdots & nf_S^{pq} \end{bmatrix},$$
$$nf_T = \begin{bmatrix} nf_T^{11} & nf_T^{12} & \cdots & nf_T^{1q} \\ nf_T^{21} & nf_T^{22} & \cdots & nf_T^{2q} \\ \vdots & \vdots & \ddots & \vdots \\ nf_T^{p1} & nf_T^{p2} & \cdots & nf_T^{pq} \end{bmatrix}, \quad (9)$$

where the j-th column elements of nf_S and nf_T represent the p samples of the j-th feature from source and target domains ($j \in [1, 2, \ldots, q]$), respectively. The expression of them is presented as

$$nf_S^j = \left[nf_S^{1n}, nf_S^{2n}, \cdots, nf_S^{pj}\right]^T,$$
$$nf_T^j = \left[nf_T^{1n}, nf_T^{2n}, \cdots, nf_T^{pj}\right]^T. \quad (10)$$

The MMD between nf_S^j and nf_T^j can be computed by (8). Therefore, a MMD sequence of q features can be further obtained, $\mathrm{MMD} = \{\mathrm{MMD}(1), \mathrm{MMD}(2), \ldots, \mathrm{MMD}(q)\}$. In this article, we suppose that the domain invariance of feature is greater when the value of MMD is smaller.

(c) *Construct the Selection Index RIM.* Based on the IS and MMD of features obtained in the previous two steps, a new selection index, RIM, is constructed for selecting dominant features from RFS. The RIM of the j-th feature is defined as follows:

$$\mathrm{RIM}(j) = \frac{\mathrm{IS}(j)}{\mathrm{MMD}(j)}. \quad (11)$$

Thus, for q features, the corresponding *RIM* values construct a *RIM* sequence, $\mathrm{RIM} = \{\mathrm{RIM}(1), \mathrm{RIM}(2), \ldots, \mathrm{RIM}(q)\}$. We suppose that the feature with higher value of RIM is more beneficial to cross-domain fault diagnosis, because the feature has great fault-discriminative ability and domain invariance at the same time. Finally, we can select the feature with higher value of RIM from the sorted RIM sequence that is sorted in descending mode to perform cross-domain fault diagnosis model training.

3.2. Transferable Fault Diagnosis Framework Based on DSID and DAE (TFDD)

3.2.1. The Mechanism of Deep Transfer AE Model. The structure of the deep transfer AE (DTAE) model is presented in Figure 3. In model, 1 input layer, 3 hidden layers, and 1 softmax layer are designed. The softmax layer is used to classify deep feature representations. The construction steps of a DTAE are shown in Figure 3 and stated as follows. There are 4 steps:

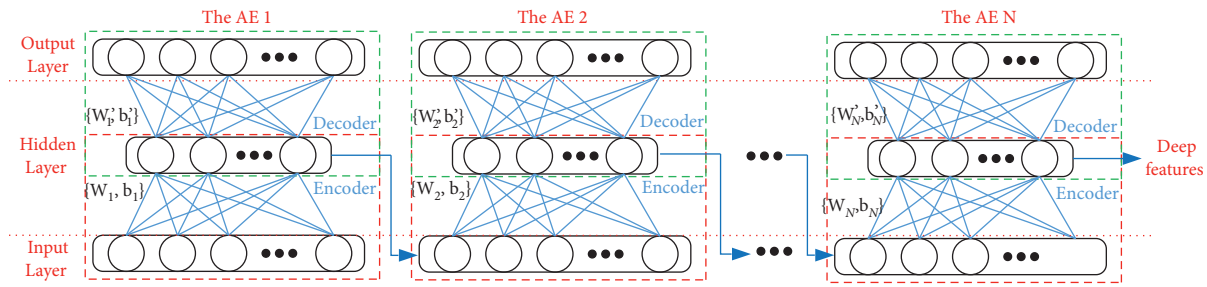


Figure 2: The structure of DAE.

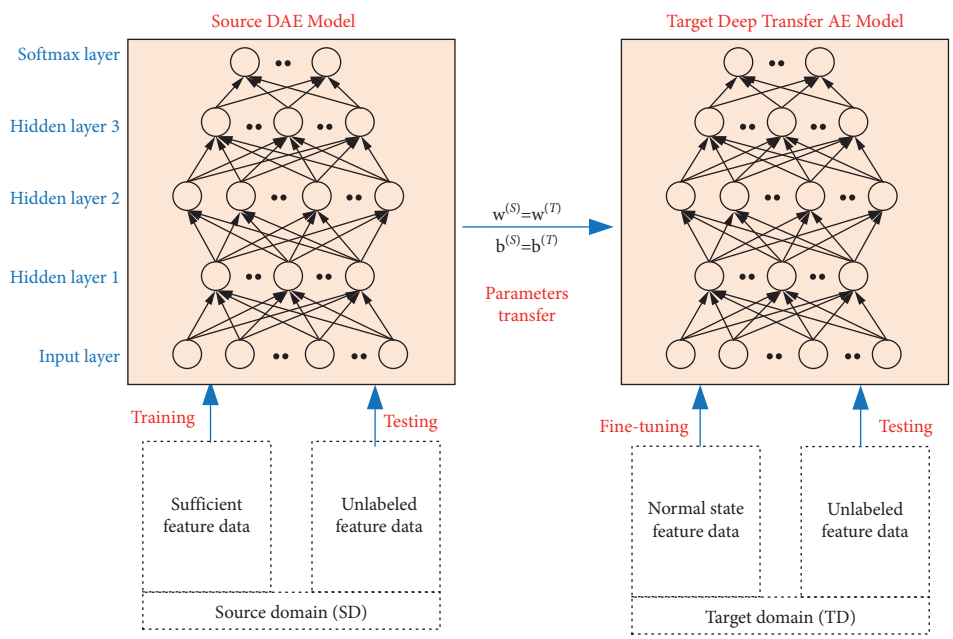


Figure 3: The mechanism of the DTAE model.

(1) Train a DAE model by using sufficient feature data from SD; this is source model. The parameters weight matrix $W^{(S)}$ and bias vector $b^{(S)}$ can be obtained.

(2) Construct another DAE model (target model), which has the same architecture as source model; that is, the number of layers and modes are the same as source model.

(3) Parameters transfer: the parameters weight matrix $W^{(S)}$ and bias vector $b^{(S)}$ learned from the procedure of source model training are transferred to target model; that is, $W^{(S)} = W^{(T)}$ and $b^{(S)} = b^{(T)}$.

(4) Fine-tuning target model: the normal state feature data from TD are employed to fine-tune the target DAE model. Finally, the fine-tuned target model is used to test remaining unlabeled target domain data.

3.2.2. Step Description of the Proposed Framework. In this article, TFDD, a novel transferable fault diagnosis framework based on DSID and DAE, is proposed for cross-domain fault diagnosis of rotating machinery. The framework TFDD is given in Figure 4, and specific descriptions are organized as the following four steps.

Step 1: *Signal Process and Feature Extraction.* The original vibration signals sampled from rotating machinery by acceleration sensors under operating conditions 1 and 2 are respectively the source and target domains data. Then, the signal process and statistical features extraction are performed by using MODWPT and calculating statistical parameters. The mixed domains statistical characteristics are generated to construct a raw feature set (RFS).

Step 2: *Dominant Features Selection.* Firstly, based on the RFS obtained in Step 1, the sufficient labeled feature data from source domain under all fault states are used to evaluate the fault-discriminative ability of features by RF, and the IS of features can be obtained to quantify the fault-discriminative ability. Secondly, the normal state feature data from two domains are used to evaluate the domain invariance of features by MMD. Finally, the proposed new dominant features selection index, RIM, is constructed. The features with high value of RIM can construct feature

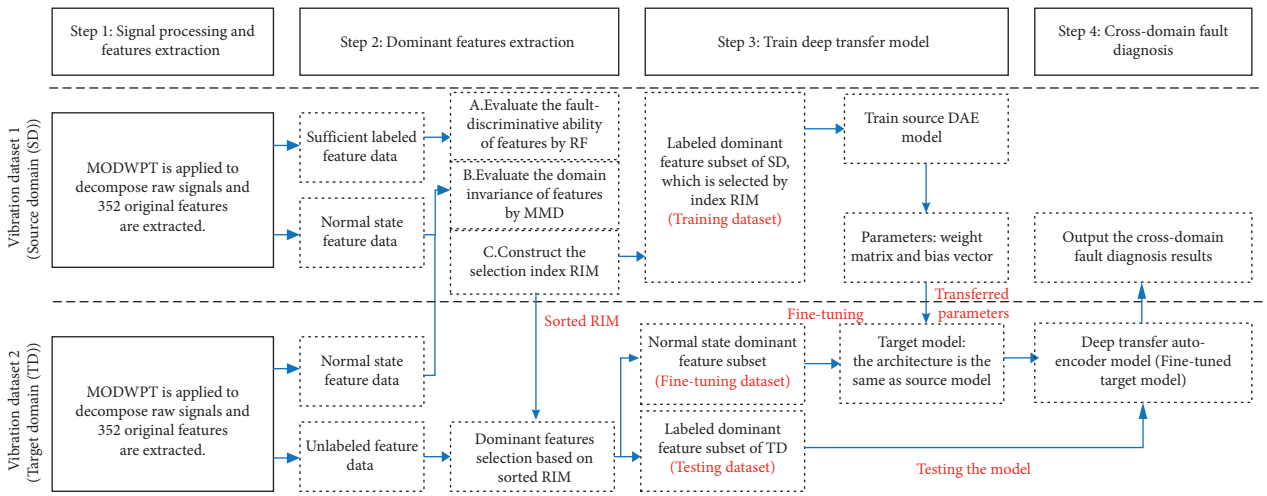


Figure 4: The proposed transferable fault diagnosis framework for rotating machinery across different operating conditions.

subset that is beneficial to cross-domain fault diagnosis. Thus, the sorted RIM sequence in descending mode is used for dominant features selection.

Step 3: *Construct Deep Transfer Autoencoder Model.* Firstly, a DAE model (source model) is trained by feature data from source domain and the parameters weight matrix $W^{(S)}$ and bias vector $b^{(S)}$ are obtained. Secondly, the parameters $W^{(S)}$ and $b^{(S)}$ are transferred to target the model that has the same architecture as the source model. Thirdly, the normal state feature data from target domain is applied to fine-tune the target model. Finally, the construction of DTAE model is completed.

Step 4: *Output the Fault Diagnosis Results.* Based on the learned DTAE model, the unlabeled feature data from target domain are used to test the performance of DTAE model and output diagnosis results.

4. Experimental Verification

In this article, motor and bearing fault datasets obtained from the SQI-MFS test platform [10, 18, 20, 54] are employed for experimental verification. The test platform is shown in Figure 5, and fault bearings and motors are presented in Figures 6 and 7. The vibration signals are sampled by acquisition cards and acceleration sensors installed at the drive end and fan end of the motor, and the sampling frequency is 16 kHz. Aiming at proving the availability and flexibility of the proposed transferable fault diagnosis framework across variable operating conditions, we collected faulty motor and bearing vibration data under different operating speeds, the experimental verification of two cases is carried out, and the detailed description is as follows.

4.1. Case 1: Transfer Diagnosis of Fault Motors under Different Operating Speeds

4.1.1. Introduction of Motor Dataset and Tasks. In this section, motor vibration data under two speeds of 1730 rmp and 1750 rmp are used for experimental verification. The main parameters of motor are shown in Table 1. Four faulty motors, including broken rotor bar fault (BF), winding fault (WF), rotor bowed fault (RF), and single phase voltage unbalance fault (SF), and a normal state motor (NM) are used in experiments. Thus, there are 5 motor conditions that correspond to 5 patterns. For each pattern, 30 and 60 vibration data samples are respectively random selected as the training and testing data. Each sample contains 5000 continues sampling points. More specific introduction of motor dataset is presented in Table 2. Based on the vibration data under speeds of 1730 rmp and 1750 rmp, we set up 2 cross-domain fault diagnosis tasks, as shown in Table 3. According to the details in Table 3, the vibration data under speeds of 1730 rmp and 1750 rmp are respectively chosen as the source datasets of tasks 1 and 2. The vibration data under speeds of 1750 rmp and 1730 rmp are respectively used as the target datasets of tasks 1 and 2. Source and target domains contain 150 and 300 samples, respectively.

4.1.2. Transfer Diagnosis Results of the Proposed TFDD Framework. According to the steps of the proposed framework TFDD, firstly, the raw vibration signals are processed by MODWPT, and statistical features are generated by calculating statistical parameters of single branch reconstruction signals of wavelet packet nodes. In this article, we apply the "dmey" as the mother wavelet in MODWPT, and the layer of wavelet decomposition is set to 4. Therefore, 16 terminal wavelet packet nodes (TWPN) are generated and the corresponding reconstruction signals (RS) are used for calculating 11 statistical parameters; thus, 176 time-domain statistical features are generated. Moreover, The Hilbert envelope spectra (HES) of 16 reconstruction signals are also used for generating 176 frequency-domain statistical features by 11 statistical parameters. These 11 statistical parameters are range, mean value, standard deviation, kurtosis, energy, energy entropy, skewness, crest factor, impulse factor, shape factor, and latitude factor, respectively [10, 18, 20, 29, 54, 55]. Therefore, 352 statistical

Figure 5: SQI-MFS test platform.

Figure 6: SER205 fault bearings.

characteristics are generated from a vibration sample to construct a raw feature set (RFS). The sampled vibration signals of 5 motor conditions under rotating speeds of 1730 rmp and 1750 rmp are shown in Figure 8, and the RS of TWPN that are obtained by decomposing normal state vibration signals are shown in Figure 9. Moreover, 352 statistical features extracted from NM and BF vibration signals under 1730 rmp and 1750 rmp are shown in Figure 10; these features have been normalized. From Figure 8, it is obvious that the distribution discrepancy existed between vibration signals from different operating speeds.

Based on the RFS obtained from signal process and features extraction, the proposed dominant feature selection method DSID is performed to evaluate the fault-discriminative ability and domain invariance of features, and the selection index RIM of each feature can be calculated by the equation (11). In this article, we suppose that the feature with higher value of RIM is more beneficial to cross-domain fault diagnosis. Then, the RIM sequence that includes RIM of 352 statistical features is obtained, the RIM sequence is sorted in descending mode, and the sorted RIM sequence is used to select dominant features for constructing feature subset. The IS, MMD, and RIM of 352 features are respectively shown in Figures 11–13.

According the sorted RIM sequence, some dominant features are chosen to construct feature subset. Then, the DTAE model training is performed; based on the source domain, the selected dominant features are used for training source DAE model, and the learned parameters $W^{(S)}$ and $b^{(S)}$ are directly transferred to initialize the target DTAE model that has the same architecture as source DAE model. Next, the normal state feature data from target domain are used for fine-tuning the target DTAE model. Finally, the testing data (unlabeled feature data from target domain) are inputted to the DTAE model, and the softmax layer of DTAE model can achieve fault classification for testing data. In this article, some parameters used in DTAE model training are as follows: the number of hidden layers is 4 and the sizes of hidden layers are respectively 400, 100, 50, and 50. The iteration is set to 200.

The experimental results of the proposed TFDD framework are respectively given in Figure 14 and Table 4. From the details of Table 4, when dominant features number is set as 352, that is, all 352 statistical features from source domain are applied for training DATE model, the diagnosis results of tasks 1 and 2 are only 69.00% and 66.67%, respectively. However, when the proposed DSID is performed before training DTAE model, it can significantly improve diagnosis accuracy. The maximum average accuracy of tasks 1 and 2 can attain 81.67% (*dfn*:101) and 82.67%(*dfn*:140), respectively. Figure 14 presents the diagnosis accuracies of tasks 1 and 2 when the *dfn* is from 40 to 352. We can

conclude that the proposed TFDD framework using DSID can enhance the performance of cross-domain diagnosis when a suitable *dfn* is selected.

4.1.3. Comparisons with Other Models. In order to further prove the advantages of TFDD framework on cross-domain fault diagnosis, we chose some common and competitive methods for comparison. Based on these methods, some comparative models are constructed, as shown in Table 5. These comparative models can be divided into two categories. (1) The model is not combined with transfer learning method; for example, the model RFS-KNN is a common model that the RFS is directly inputted to the KNN classifier. (2) The model is combined with transfer learning method; for example, RFS-TCA is a transfer learning-based model such that RFS is directly inputted to the TCA and the SVM classifier is applied to classify the fault features. For the RFS-DSID-TCA model, it is based on the RFS-TCA model, and the proposed DSID method is employed to select dominant features from RFS for the subsequent transfer learning.

The experimental results of comparative models are given in Table 6, Figures 15 and 16. According to the details in Table 6, the diagnosis accuracies of comparative models are obviously smaller than the accuracies of TFDD. The transfer learning-based models, RFS-TCA, RFS-JDA, RFS-DSID-TCA, and RFS-DSID-TCA, can achieve better diagnosis performance than other models. When the DSID is embedded in transfer learning-based model, the diagnosis performance can be further enhanced; the diagnosis accuracies of RFS-DSID-TCA and RFS-DSID-JDA models for task 1 are 71.67% and 77.00%, which are 5.67% and 4.67% higher than RFS-TCA and RFS-JDA models, respectively. The diagnosis accuracies of RFS-DSID-TCA and RFS-DSID-JDA models for task 2 are 77% and 78.00%, which are 7.33% and 8.00% higher than RFS-TCA and RFS-JDA models, respectively. However, the models using TCA and JDA do not outperform the proposed TFDD model, and the maximum average accuracy of TFDD is higher than RFS-DSID-TCA and RFS-DSID-JDA models. These comparison results can validate the advantages of the TFDD, which includes two aspects:

(1) For the cross-domain diagnosis tasks 1 and 2, the proposed TFDD model can effectively classify 5 motor conditions, and the maximum average accuracy can attain over 80%. The proposed dominant features selection method DSID can help in selecting features that have high fault-discriminative and domain invariance, which can significantly improve cross-domain diagnosis performance.

(2) According to the comparison results, it reveals that the diagnosis performance of TFDD is obviously better than comparative models shown in Table 6. Moreover, the diagnosis model combined transfer learning strategy can help in enhancing diagnosis accuracy across different domains.

4.2. Case 2: Transfer Diagnosis of Fault Bearings under Different Operating Speeds

4.2.1. Introduction of Bearing Dataset and Tasks. In this section, bearing vibration data under two speeds of 1200 rmp and 1600 rmp are used to further prove the availability, flexibility, and advantages of the TFDD. Three kinds of faulty bearings (inner race fault (IRF), outer race fault (ORF), and ball fault (BF)) are manufactured by laser machining, and three kinds of fault diameters (0.05 mm, 0.1 mm, and 0.2 mm) are set for each fault type for experiments. These faulty bearings are given in Figure 6. In addition, a normal bearing is also used for experiments; thus, there are 10 bearing states that correspond to 10 patterns. For each pattern, 30 and 60 vibration data samples are respectively random chosen as the training and testing samples. Each sample contains 5000 sampling points. More details of bearings dataset are presented in Table 7. Based on the vibration data under speeds of 1200 rmp and 1600 rmp, we set up 2 cross-domain fault diagnosis tasks, as shown in Table 8.

4.2.2. Transfer Diagnosis Results of the Proposed TFDD Framework. In this section, the process of experiment is the similar to that of Section 4.1.2. The fault diagnosis results of tasks 1 and 2 obtained by TFDD framework are shown in Table 9 and Figure 17. From the experimental results, it is obvious that the TFDD can effectively diagnose bearing faults across different operating speeds, and the highest diagnosis accuracies of tasks 1 and 2 can respectively reach 90.33% (*dfn*: 150) and 90.00% (*dfn*: 152), which are 8.83% and 9% higher than the models without using DSID. From Figure 17, when a suitable *dfn* is selected according to the sorted RMI sequence, the performance of model can obtain an obvious enhancement. This further proves the availability of the DSID.

4.2.3. Comparisons with Other Models. For the comparative experiments, these are the same as Section 4.2.3. The models used for comparison are shown in Table 5. The experimental results are presented in Table 10, Figures 18 and 19. The diagnosis results obtained by TFDD are significantly better than other models, and the maximum accuracies of tasks 1 and 2 can respectively reach 90.33% and 90%. For task 1, the diagnosis results of RFS-SVM, RFS-KNN, RFS-DAE, RFS-DBN, RFS-CNN, RFS-TCA, RFS-JDA, RFS-DSID-TCA, and RFS-DSID-JDA are 70.83%, 65.17%, 62.83%, 75.67%, 65.00%, 56.83%, 65.50%, 69.50%, and 85.67%, respectively. For task 2, the diagnosis results of RFS-SVM, RFS-KNN, RFS-DAE, RFS-DBN, RFS-CNN, RFS-TCA, RFS-JDA, RFS-DSID-TCA, and RFS-DSID-JDA are 50.83%, 56.33%, 58.33%, 52.67%, 57.50%, 52.17%, 61.33%, 61.83%, and 83.50%, respectively. This further proves the advantages of the TFDD. According to the diagnosis results given in Figures 18 and 19, the transfer learning-based models can obtain an improvement on diagnosis accuracy by combining the DSID; thus, the availability of the DSID is also verified.

Figure 7: Motor fault data collection and fault motors.

Table 1: Main parameters of motor.

Name	Value
Poles	1
Number of stator slots	34
Number of rotor slots	24
Stator bore	82.1 mm
Rotor bore	80.5 mm
Rated power	370 W

Table 2: Details about the motor vibration dataset.

Motor conditions	Sampling frequency (kHz)	Rotating speeds 1730 rmp		Rotating speeds 1750 rmp		Class label
		Size of training samples	Size of testing samples	Size of training samples	Size of testing samples	
NM	16	30	60	30	60	1
BF		30	60	30	60	2
WF		30	60	30	60	3
RF		30	60	30	60	4
SF		30	60	30	60	5

Table 3: 2 cross-domain fault diagnosis tasks for 5 motor conditions.

	Task 1			Task 2		
	Rotating speeds (rmp)	Motor conditions	Size of samples	Rotating speeds (rmp)	Motor conditions	Size of samples
Source domain	1730	Classes 1–5	150	1750	Classes 1–5	150
Target domain	1750	Classes 1–5	300	1730	Classes 1–5	300

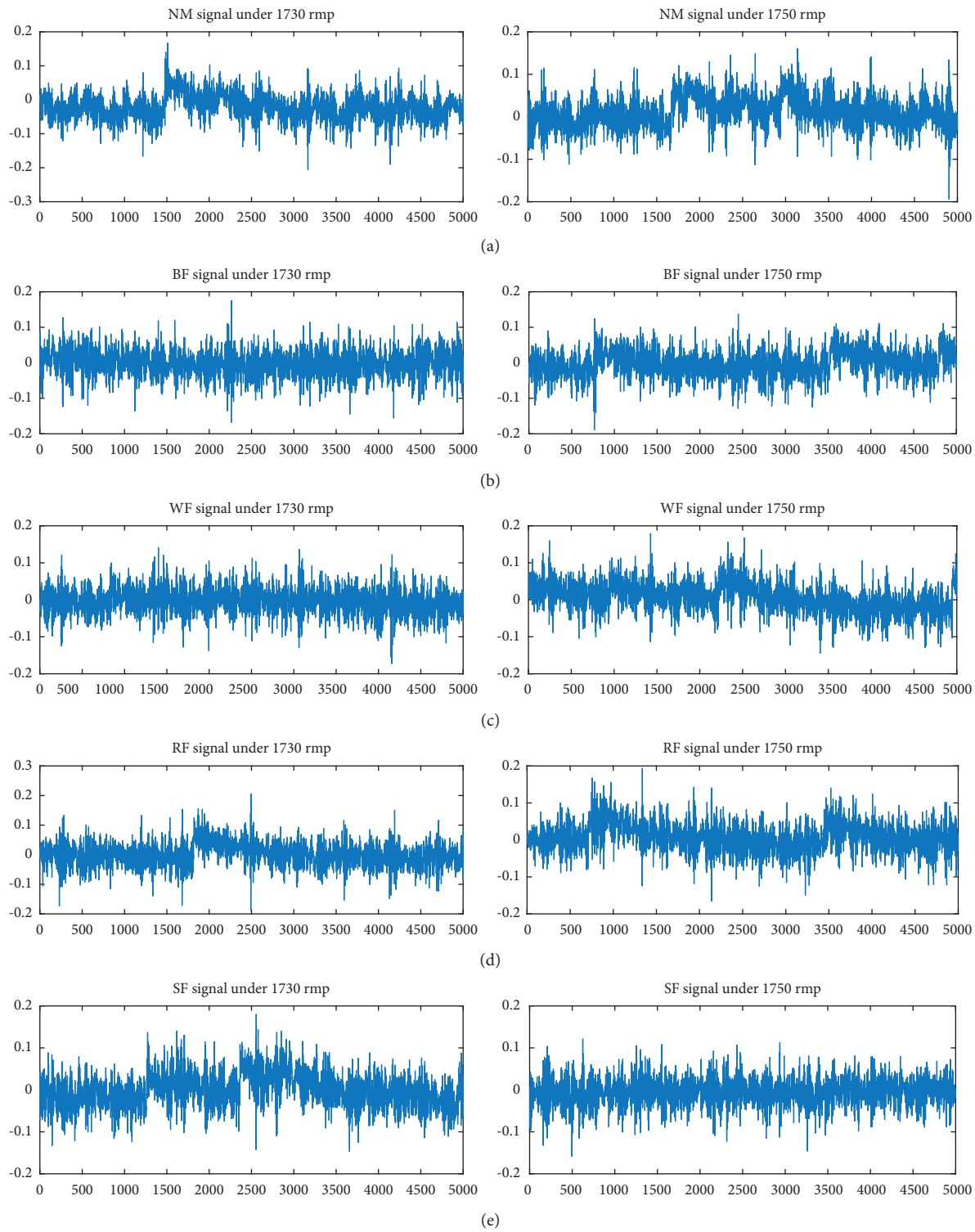


Figure 8: Raw signals of the 5 kinds of motor conditions. (a) Raw signal under normal state. (b) Raw signal under broken rotor bar fault. (c) Raw signal under winding fault. (d) Raw signal under rotor bowed fault. (e) Raw signal under single phase voltage unbalance fault.

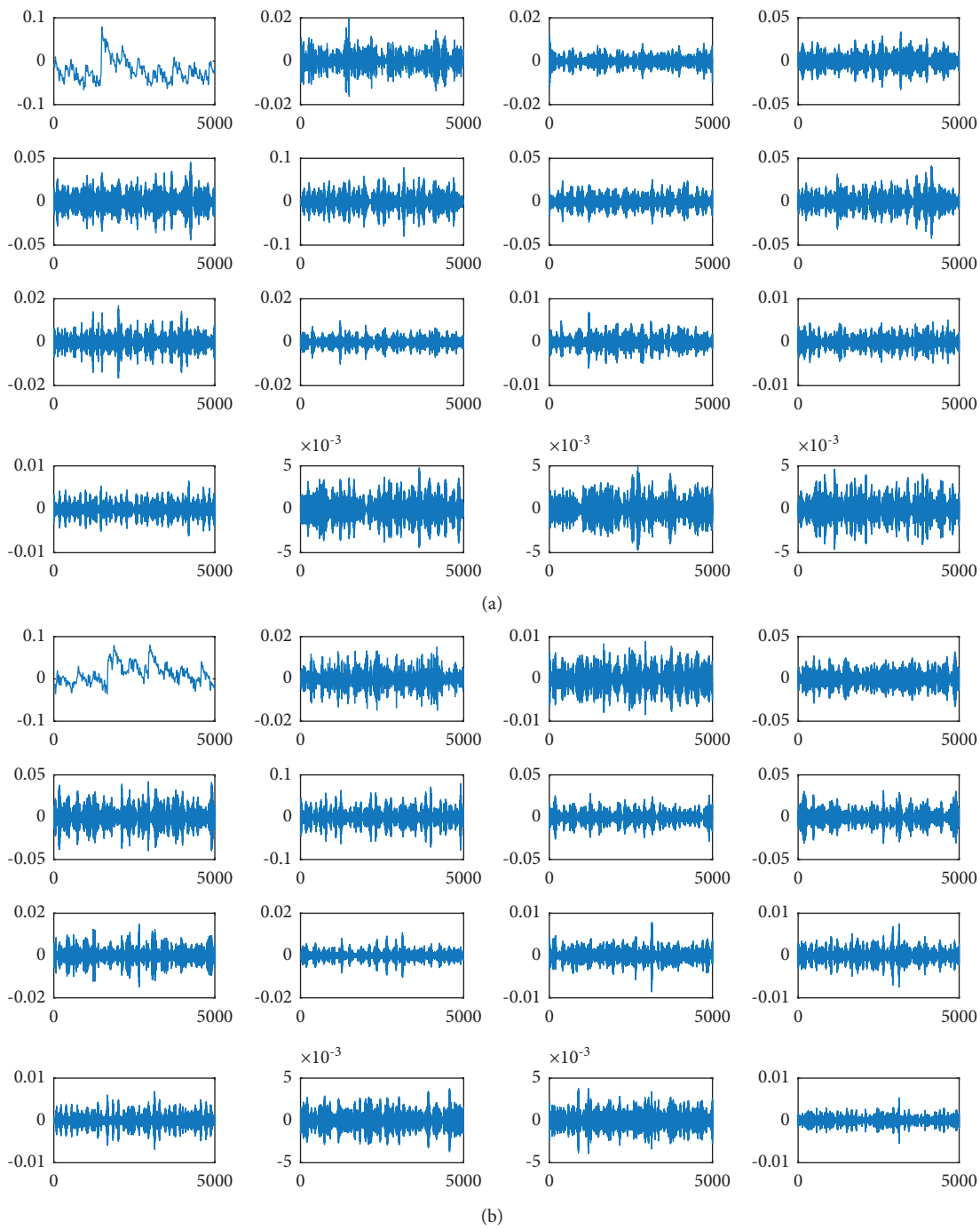


Figure 9: Reconstruction signals for wavelet packet nodes obtained by decomposing normal state vibration signals. (a) The reconstruction signals of NM signal under 1730 rmp by MODWPT. (b) The reconstruction signals of NM signal under 1750 rmp by MODWPT.

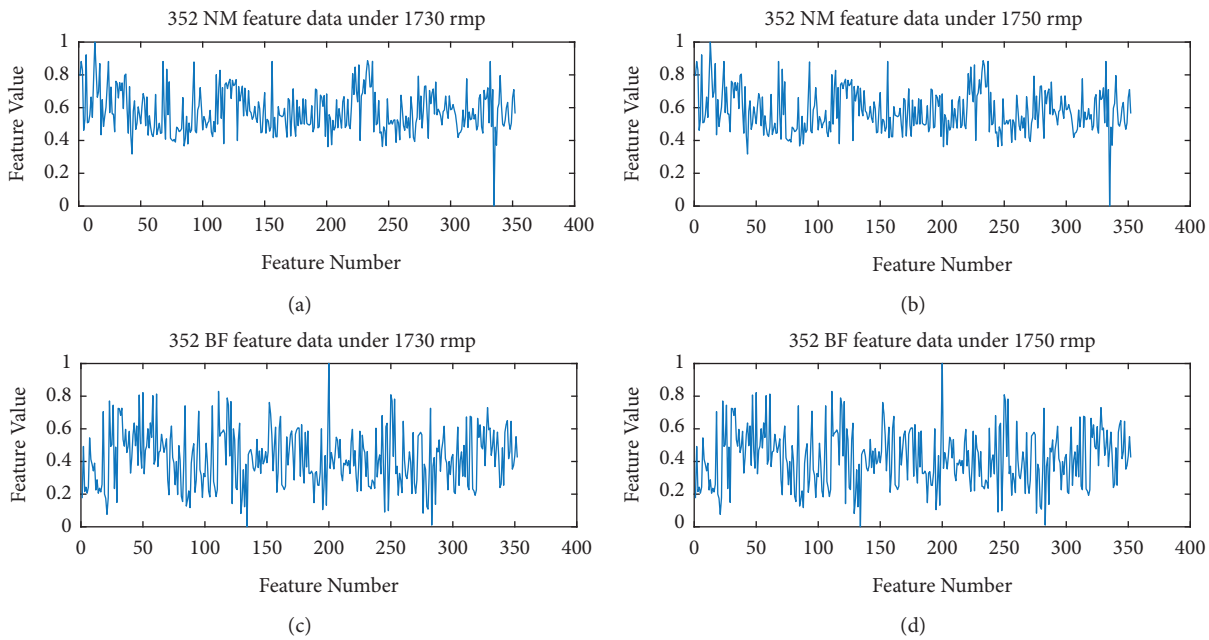


Figure 10: 352 statistical features extracted from NM and BF vibration signals under 1730 rmp and 1750 rmp (the feature data are normalized). (a) NM feature data under 1730 rmp. (b) NM feature data under 1750 rmp. (c) BF feature data under 1730 rmp. (d) BF feature data under 1730 rmp.

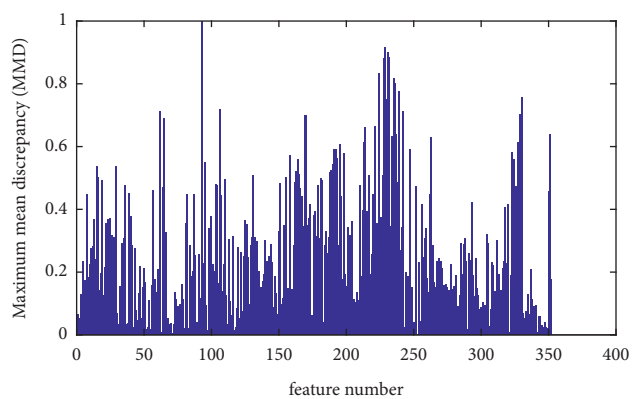


Figure 11: The maximum mean discrepancies of 352 features.

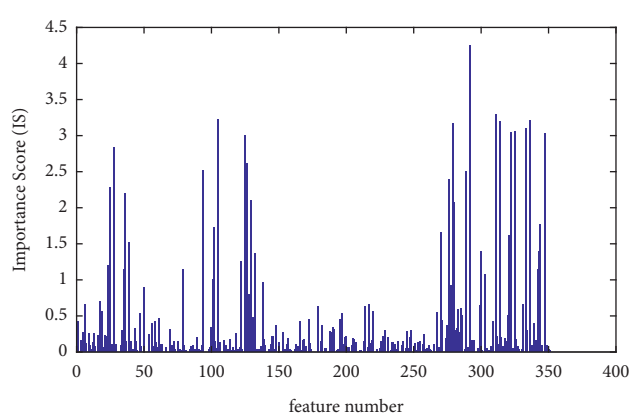


Figure 12: The importance scores of 352 features.

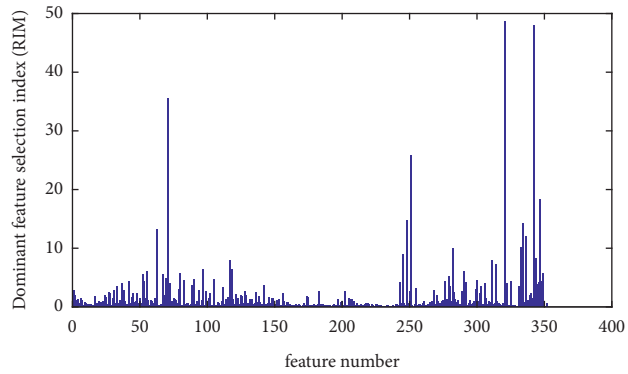


Figure 13: The dominant feature selection indexes of 352 features.

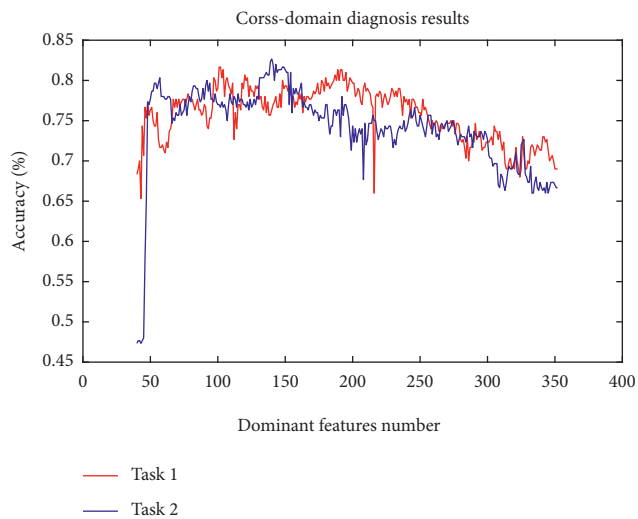


Figure 14: The diagram of cross-domain fault diagnosis results of TFDD framework.

Table 4: Cross-domain fault diagnosis results of TFDD framework.

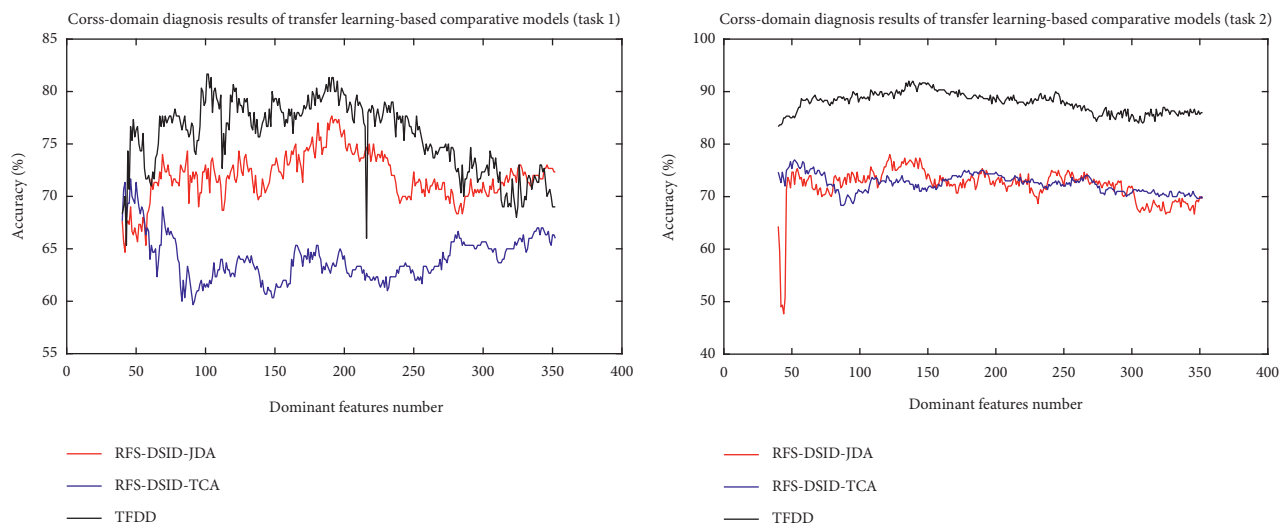
Dominant features number (*dfn*)	Results of task 1 (%)	Results of task 2 (%)
40	68.33	47.33
50	76.33	77.67
70	76.67	75.67
90	77.00	77.67
110	78.33	77.33
130	79.33	77.67
150	79.33	**81.33**
170	78.00	76.33
190	**80.00**	76.33
210	78.67	72.00
230	78.67	71.67
250	76.00	75.67
270	74.67	75.00
290	72.33	73.67
310	72.33	68.33
330	70.00	67.33
352	**69.00**	**66.67**

Table 5: Comparative models based on common and competitive methods.

The model is not combined with the transfer learning method	The model is combined with the transfer learning method
RFS-SVM	RFS-TCA
RFS-KNN	RFS-JDA
RFS-DAE	RFS-DSID-TCA
RFS-DBN	RFS-DSID-JDA
RFS-CNN	TFDD

Table 6: Cross-domain fault diagnosis results of comparative models.

Models	Results of task 1 (%)	Results of task 2 (%)
RFS-SVM	60.33%	72.33%
RFS-KNN	48.00%	51.00%
RFS-DAE	46.33%	63.33%
RFS-DBN	36.33%	64.67%
RFS-CNN	69.67%	64.33%
RFS-TCA	66.00%	69.67%
RFS-JDA	72.33%	70.00%
RFS-DSID-TCA	**71.67% (*dfn*:46)**	**77.00% (*dfn*:52)**
RFS-DSID-JDA	**77.00% (*dfn*:181)**	**78.00% (*dfn*:122)**
TFDD	**81.67% (*dfn*:101)**	**82.67% (*dfn*:140)**



(a) (b)

Figure 15: The diagram of cross-domain fault diagnosis results of transfer learning-based comparative models.

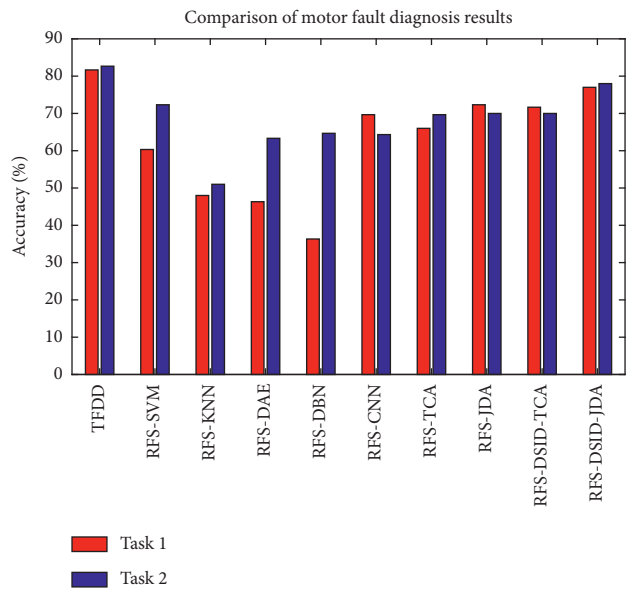


Figure 16: Comparison of fault diagnosis results of comparative models.

Table 7: Details about the bearing vibration dataset.

Bearing conditions	Fault diameters (mm)	Rotating speeds 1200 rmp		Rotating speeds 1600 rmp		Class label
		Size of training samples	Size of testing samples	Size of training samples	Size of testing samples	
IRF	0.05	30	60	30	60	1
	0.1	30	60	30	60	2
	0.2	30	60	30	60	3
ORF	0.05	30	60	30	60	4
	0.1	30	60	30	60	5
	0.2	30	60	30	60	6
BF	0.05	30	60	30	60	7
	0.1	30	60	30	60	8
	0.2	30	60	30	60	9
Normal	0	30	60	30	60	10

Table 8: 2 cross-domain fault diagnosis tasks for 10 bearing conditions.

	Task 1			Task 2		
	Rotating speeds (rmp)	Bearing conditions	Size of samples	Rotating speeds (rmp)	Bearing conditions	Size of samples
Source domain	1200	Classes 1–10	150	1600	Classes 1–10	150
Target domain	1600	Classes 1–10	300	1200	Classes 1–10	300

Table 9: Cross-domain fault diagnosis results of TFDD framework.

Dominant features' number (*dfn*)	Results of task 1 (%)	Results of task 2 (%)
50	85.33	83.50
70	86.33	84.00
90	85.67	83.33
110	86.83	85.17
130	86.50	84.67
150	**90.33**	**89.33**
170	87.50	86.50
190	86.00	85.83
210	82.50	81.33
230	81.67	80.83
250	81.50	80.83
270	81.50	80.83
290	81.50	80.83
310	81.67	81.00
330	81.50	80.83
352	81.50	81.00

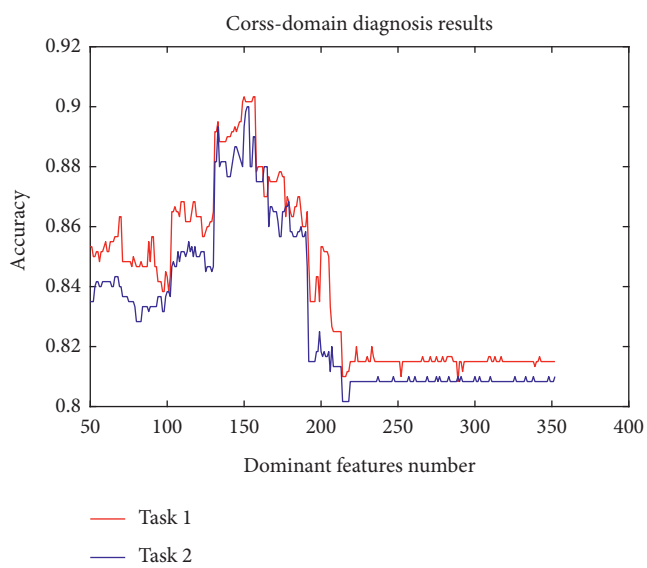


Figure 17: The diagram of cross-domain fault diagnosis results of TFDD framework.

Table 10: Cross-domain fault diagnosis results of comparative models.

Models	Results of task 1 (%)	Results of task 2 (%)
RFS-SVM	70.83	50.83
RFS-KNN	65.17	56.33
RFS-DAE	62.83	58.33
RFS-DBN	75.67	52.67
RFS-CNN	65.00	57.50
RFS-TCA	56.83	52.17
RFS-JDA	65.50	61.33
RFS-DSID-TCA	69.50% (*nkf*:142)	61.83% (*nkf*:231)
RFS-DSID-JDA	85.67% (*nkf*:175)	83.50% (*nkf*:271)
TFDD	**90.33% (*nkf*:150)**	**90.00% (*nkf*:152)**

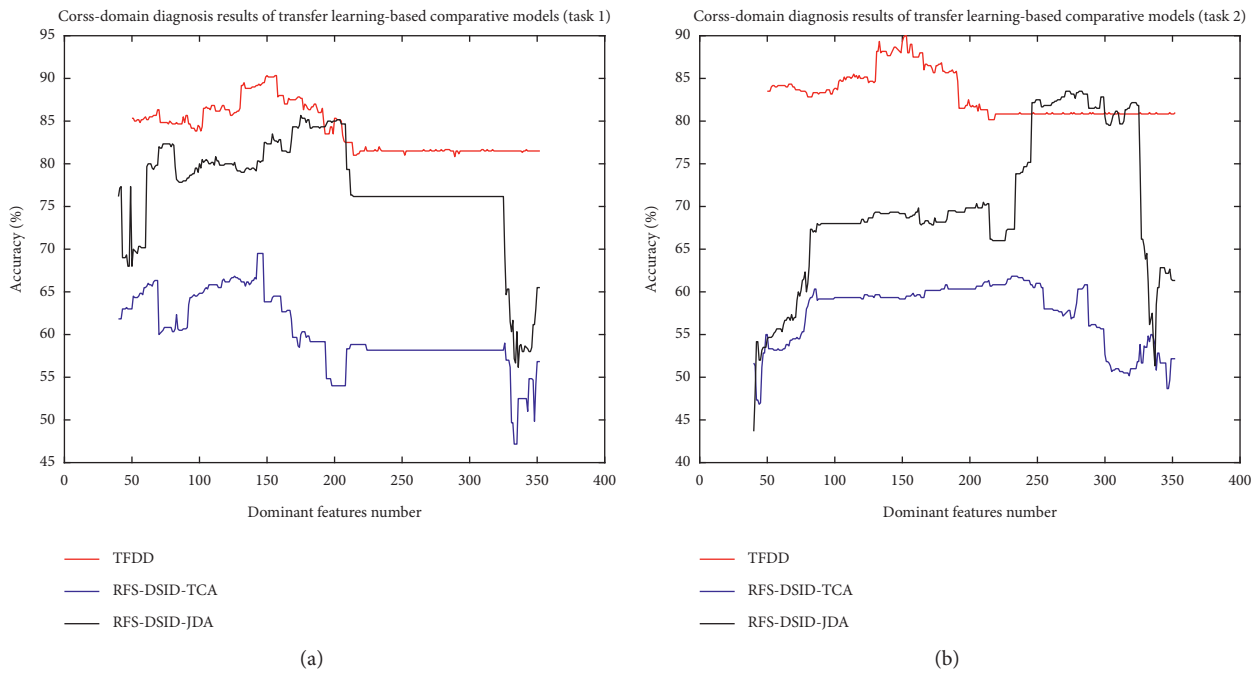


Figure 18: The diagram of cross-domain fault diagnosis results of transfer learning-based comparative models.

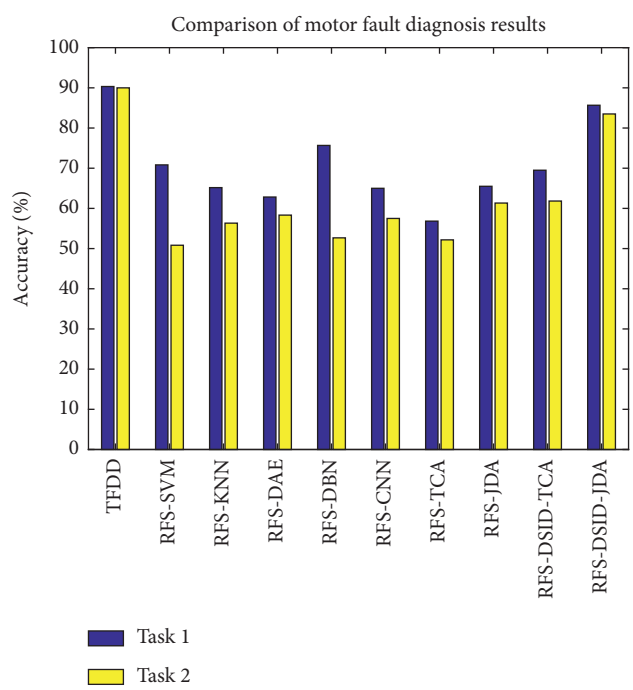


Figure 19: Comparison of fault diagnosis results of comparative models.

5. Conclusions

A new transferable fault diagnosis approach of rotating machinery based on deep autoencoder and dominant features selection, TFDD, is proposed. Firstly, the signal process and features extraction are performed. Then, based on the sufficient labeled source feature data and normal state target feature data, the proposed DSID is performed to evaluate the features, the new selection index, RIM, is used to selected dominant features for training DTAE model. Next, by using labeled feature subset of source domain, a source DAE model can be learned and the corresponding parameters are transferred to the target DTAE model. Finally, this DTAE model classifies the unlabeled data from target domain.

A series of experiments are carried out by using motor and bearing fault datasets sampled from SQI-MFS test

platform. The experimental results prove the availability, flexibility, and advantages of the TFDD. The details are as the following aspects: (1) the proposed TFDD model can effectively diagnose faulty motors and bearings across different operating speeds, and the diagnosis performance significantly outperforms comparative models. (2) The proposed DSID can help to select features that have high fault-discriminative and domain invariance, when a suitable *dfn* is chosen, which can significantly improve cross-domain diagnosis performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Youth Science and Technology Fund of China University of Mining and Technology, Basic Scientific Research Business (no. 2021QN1093), the "Smart Mine" Key Technology R&D Open Fund of China University of Mining and Technology and Zibo Mining Group Co., Ltd. (no. 2019LH08), the National Key R&D Program of China (nos. 2017YFC0804400 and 2017YFC0804401), and the fund project of JiangSu Collaborative Innovation Center for Building Energy Saving and Construction Technology (no. SJXTY1603).

References

[1] A.-S. Qin, H.-L. Mao, and Q. Hu, "Cross-domain fault diagnosis of rolling bearing using similar features-based transfer approach," *Measurement*, vol. 172, no. 1, Article ID 108900, 2021.

[2] D. Liu, W. Cheng, and W. Wen, "Intelligent cross-condition fault recognition of rolling bearings based on normalized resampled characteristic power and self-organizing map," *Mechanical Systems and Signal Processing*, vol. 153, no. 11, Article ID 107462, 2021.

[3] Z. Lei, G. Wen, S. Dong et al., "An intelligent fault diagnosis method based on domain adaptation and its application for bearings under polytropic working conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.

[4] D. Wei, T. Han, F. Chu, and M. J. Zuo, "Weighted domain adaptation networks for machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 158, Article ID 107744, 2021.

[5] Z. Zhou, C. Wen, and C. Yang, "Fault isolation based on k-nearest neighbor rule for industrial processes," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 4, pp. 2578–2586, 2016.

[6] X. Zhang, W. Chen, B. Wang, and X. Chen, "Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization," *Neurocomputing*, vol. 167, pp. 260–279, 2015.

[7] R. S. Gunerkar, A. K. Jalan, and S. U. Belgamwar, "Fault diagnosis of rolling element bearing based on artificial neural network," *Journal of Mechanical Science and Technology*, vol. 33, no. 2, pp. 505–511, 2019.

[8] J. Wang, Y. Zhang, F. Zhang et al., "Accuracy-improved bearing fault diagnosis method based on AVMD theory and AWPSO-ELM model," *Measurement*, vol. 181, Article ID 109666, 2021.

[9] J. S. L. Senanayaka, H. Van Khang, and K. G. Robbersmyr, "Towards online bearing fault detection using envelope analysis of vibration signal and decision tree classification algorithm," in *Proceeding of the 2017 20th International Conference on Electrical Machines and Systems (ICEMS)*, pp. 1–6, IEEE, Sydney, NSW, Australia, August 2017.

[10] X. Yu, F. Dong, E. Ding, S. Wu, and C. Fan, "Rolling bearing fault diagnosis using modified LFDA and EMD with sensitive feature selection," *IEEE Access*, vol. 6, pp. 3715–3730, 2017.

[11] T. Guo and Z. Deng, "An improved EMD method based on the multi-objective optimization and its application to fault feature extraction of rolling bearing," *Applied Acoustics*, vol. 127, pp. 46–62, 2017.

[12] X. Ye, Y. Hu, J. Shen, R. Feng, and G. Zhai, "An improved empirical mode decomposition based on adaptive weighted rational quartic spline for rolling bearing fault diagnosis," *IEEE Access*, vol. 8, Article ID 123813, 2020.

[13] D. Zhong, W. Guo, and D. He, "An intelligent fault diagnosis method based on STFT and convolutional neural network for bearings under variable working conditions," in *Proceedings of the 2019 Prognostics and System Health Management Conference (PHM-Qingdao)*, pp. 1–6, IEEE, Qingdao, China, October 2019.

[14] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, "An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks," *Journal of the Franklin Institute*, vol. 357, no. 11, pp. 7286–7307, 2020.

[15] Y. Du, A. Wang, S. Wang, B. He, and G. Meng, "fault diagnosis under variable working conditions based on STFT and transfer deep residual network," *Shock and Vibration*, vol. 2020, Article ID 1274380, 18 pages, 2020.

[16] G. Li, C. Deng, J. Wu, Z. Chen, and X. Xu, "Rolling bearing fault diagnosis based on wavelet packet transform and convolutional neural network," *Applied Sciences*, vol. 10, no. 3, 2020.

[17] S. Xiong, H. Zhou, S. He, L. Zhang, and T. Shi, "Fault diagnosis of a rolling bearing based on the wavelet packet transform and a deep residual network with lightweight multi-branch structure," *Measurement Science and Technology*, vol. 32, no. 8, Article ID 085106, 2021.

[18] D. Fei, Y. Xiao, D. Enjie, S. Wu, C. Fan, and Y. Huang, "Rolling bearing fault diagnosis using modified neighborhood preserving embedding and maximal overlap discrete wavelet packet transform with sensitive features selection," *Shock and Vibration*, vol. 2018, Article ID 5063527, 29 pages, 2018.

[19] X. Yu, X. Ren, H. Wan, S. Wu, and E. Ding, "Rolling bearing fault feature extraction and diagnosis method based on MODWPT and DBN," in *Proceedings of the 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, October 2019.

[20] I. I. E. Amarouayache, M. N. Saadi, N. Guersi, and N. Boutasseta, "Bearing fault diagnostics using EEMD processing and convolutional neural network methods,"

International Journal of Advanced Manufacturing Technology, vol. 107, no. 9, pp. 4077–4095, 2020.

[21] H. Li, T. Liu, X. Wu, and Q. Chen, "Application of EEMD and improved frequency band entropy in bearing fault feature extraction," *ISA Transactions*, vol. 88, pp. 170–185, 2019.

[22] J. Hou, Y. Wu, H. Gong, A. S. Ahmad, and L. Liu, "A novel intelligent method for bearing fault diagnosis based on EEMD permutation entropy and GG clustering," *Applied Sciences*, vol. 10, no. 1, Article ID 5063527, 2020.

[23] A. F. Aimer, A. H. Boudinar, N. Benouzza, A. Bendiabdellah, and M. E. A. Khodja, "Bearing fault diagnosis of a PWM inverter fed-induction motor using an improved short time Fourier transform," *Journal of Electrical Engineering & Technology*, vol. 14, no. 3, pp. 1201–1210, 2019.

[24] L. Xu, S. Chatterton, P. Pennacchi, and C. Liu, "A t order tracking method based on inverse short time fourier transform and singular value decomposition for bearing fault diagnosis," *Sensors*, vol. 20, no. 23, 2020.

[25] W. Zhao, Z. Wang, J. Ma, and L. Li, "Fault diagnosis of a hydraulic pump based on the CEEMD-STFT time-frequency entropy method and multiclass SVM classifier," *Shock and Vibration*, vol. 2016, Article ID 2609856, 8 pages, 2016.

[26] H. Shao, J. Lin, L. Zhang, and M. Wei, "Compound fault diagnosis for a rolling bearing using adaptive DTCWPT with higher order spectra," *Quality Engineering*, vol. 32, no. 3, pp. 342–353, 2020.

[27] H. Shao, H. Jiang, F. Wang, and Y. Wang, "Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet," *ISA Transactions*, vol. 69, pp. 187–201, 2017.

[28] A. Afia, C. Rahmoune, and D. Benazzouz, "New gear fault diagnosis method based on modwpt and neural network for feature extraction and classification," *Journal of Testing and Evaluation*, vol. 49, no. 2, pp. 1064–1085, 2019.

[29] X. Yu, W. Chen, C. Wu et al., "Rolling bearing fault diagnosis based on domain adaptation and preferred feature selection under variable working conditions," *Shock and Vibration*, vol. 2021, Article ID 8843124, 27 pages, 2021.

[30] R. V. Sánchez, P. Lucero, R. E. Vásquez, M. Cerrada, J. C. Macancela, and D. Cabrera, "Feature ranking for multi-fault diagnosis of rotating machinery by using random forest and KNN," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 6, pp. 3463–3473, 2018.

[31] J. Xiong, Q. Zhang, G. Sun, X. Zhu, M. Liu, and Z. Li, "An information fusion fault diagnosis method based on dimensionless indicators with static discounting factor and KNN," *IEEE Sensors Journal*, vol. 16, no. 7, pp. 2060–2069, 2016.

[32] Z. Yang, C. Kong, Y. Wang, X. Rong, and L. Wei, "Fault diagnosis of mine asynchronous motor based on MEEMD energy entropy and ANN," *Computers & Electrical Engineering*, vol. 92, Article ID 107070, 2021.

[33] T. A. Shifat and J.-W. Hur, "ANN assisted multi sensor information fusion for BLDC Motor fault diagnosis," *IEEE Access*, vol. 9, no. 99, pp. 9429–9441, 2021.

[34] J. Lu, W. Qian, S. Li, and R. Cui, "Enhanced K-nearest neighbor for intelligent fault diagnosis of rotating machinery," *Applied Sciences*, vol. 11, no. 3, 2021.

[35] S. Fei, "The hybrid method of VMD-PSR-SVD and improved binary PSO-KNN for fault diagnosis of bearing," *Shock and Vibration*, vol. 2019, Article ID 4954920, 7 pages, 2019.

[36] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New imbalanced fault diagnosis framework based on Cluster-MWMOTE and MFO-optimized LS-SVM using limited and complex bearing data," *Engineering Applications of Artificial Intelligence*, vol. 96, Article ID 103966, 2020.

[37] X. Zhang, C. Li, X. Wang, and H. Wu, "A novel fault diagnosis procedure based on improved symplectic geometry mode decomposition and optimized SVM," *Measurement*, vol. 173, Article ID 108644, 2021.

[38] S. L. Souad, B. Azzedine, and S. Meradi, "Fault diagnosis of rolling element bearings using artificial neural network," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 5, 2020.

[39] R. Meng and J. Z. Ming, "A new strategy for rotating machinery fault diagnosis under varying speed conditions based on deep neural networks and order tracking," in *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, December 2018.

[40] P. Lei, C. Shen, D. Wang, L. Chen, Z. Zhou, and Z. Zhu, "A new transferable bearing fault diagnosis method with adaptive manifold probability distribution under different working conditions," *Measurement*, vol. 173, Article ID 108565, 2021.

[41] X. Kong, G. Mao, Q. Wang, H. Ma, and W. Yang, "A multi-ensemble method based on deep autoencoders for fault diagnosis of rolling bearings," *Measurement*, vol. 151, Article ID 107132, 2020.

[42] P. Ma, H. Zhang, W. Fan, and C. Wang, "A diagnosis framework based on domain adaptation for bearing fault diagnosis across diverse domains," *ISA Transactions*, vol. 99, pp. 465–478, 2020.

[43] W. Mao, W. Feng, Y. Liu, D. Zhang, and X. Liang, "A new deep autoencoder method with fusing discriminant information for bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 150, no. 12, Article ID 107233, 2021.

[44] H. Zhiyi, S. Haidong, J. Lin, C. Junsheng, and Y. Yu, "Transfer fault diagnosis of bearing installed in different machines using enhanced deep autoencoder," *Measurement*, vol. 152, Article ID 107393, 2020.

[45] G. B. Jang and S. B. Cho, "Feature space transformation for fault diagnosis of rotating machinery under different working conditions," *Sensors*, vol. 21, no. 4, 2021.

[46] Z. He, H. Shao, P. Wang, J. Lin, J. Cheng, and Y. Yang, "Deep transfer multi-wavelet autoencoder for intelligent fault diagnosis of gearbox with few target training samples," *Knowledge-Based Systems*, vol. 191, Article ID 105313, 2020.

[47] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[48] M. T. Uddin and M. A. Uddiny, "A guided random forest based feature selection approach for activity recognition," in *Proceedings of the 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1–6, IEEE, Savar, Bangladesh, May 2015.

[49] Y. Wang and S. T. Xia, "A novel feature subspace selection method in random forests for high dimensional data," in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 4383–4389, IEEE, Vancouver, BC, Canada, July 2016.

[50] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[51] L. I. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression Trees.Wadsworth," *Biometrics*, vol. 40, no. 3, 1984.

[52] A. Gretton, K. Borgwardt, and M. Rasch, "A kernel method for the two-sample-problem," *Advances in Neural Information Processing Systems*, vol. 19, pp. 513–520, 2006.

[53] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.
[54] Y. Feng, J. Chen, Z. Yang et al., "Similarity-based meta-learning network with adversarial domain adaptation for cross-domain fault identification," *Knowledge-Based Systems*, vol. 217, Article ID 106829, 2021.
[55] A. S. Minhas and S. Singh, "A new bearing fault diagnosis approach combining sensitive statistical features with improved multiscale permutation entropy method," *Knowledge-Based Systems*, vol. 218, no. 17, Article ID 106883, 2021.

Hindawi
Shock and Vibration
Volume 2021, Article ID 4275922, 14 pages
https://doi.org/10.1155/2021/4275922

Research Article

Mechanical Efficiency of HMCVT under Steady-State Conditions

Guangqing Zhang,[1] Hengtong Zhang,[2] Yanyan Ge,[2] Wei Qiu,[2] Maohua Xiao,[2] Xiaomei Xu,[3] and Minghui Zhou[4]

[1]*Xuzhou Carter Agricultural Equipment Co., Ltd., Xuzhou 221011, China*
[2]*College of Engineering, Nanjing Agricultural University, Nanjing 210031, China*
[3]*College of Automobile and Traffic Engineering, Nanjing Forestry University, Nanjing 210042, China*
[4]*Weichai Power Hydraulic Transmission Research Institute, Weifang, 261001, China*

Correspondence should be addressed to Wei Qiu; qiuwei@njau.edu.cn

Received 19 June 2021; Accepted 11 August 2021; Published 7 September 2021

Academic Editor: Jun Zhu

Hydromechanical continuously variable transmission (HMCVT) technology has been widely used due to its advantages of ride comfort and fuel economy. The relatively uniform efficiency expression of HMCVT is obtained by studying torque and transmission ratios to reveal steady-state characteristics and predict the output torque. Mathematical models of torque ratios are derived by analyzing the HMCVT system power flow and calculating the equivalent meshing power of epicyclic gear train and efficiency for the hydraulic system. The relationship between mechanical system transmission and hydraulic system parameters is established using the torque ratios, and a mechanical system demanding surface is proposed. Two numerical examples of the HMCVT system with single and dual variable units are demonstrated to establish an effective and convenient method. The method is validated through a physical prototype TA1-02 test.

1. Introduction

An epicyclic gear train (EGT) with two DOFs is closed by a hydrostatic transmission system (HST), so that only one DOF exists overall [1]. The mechanical hydraulic components are organically combined to form a simple hydromechanical continuously variable transmission (HMCVT) [2, 3]. When the input speed of HMCVT is stable, with displacement ratio e of HST changing continuously in a certain range, the output speed of HMCVT continuously changes from the minimum speed (e.g., 0) to the maximum speed. Multistage EGTs or two EGTs with multiple segments (fixed transmission ratio) are often used in HMCVT design to obtain a large transmission ratio range. By reasonably setting the system parameters of EGTs and HST, the multistage shifting transmission ratios are almost the same to reduce the impact. In the shifting process, shift quality is improved by using the control strategy of nonspeed difference. HMCVT technology, with its characteristics of ride comfort and fuel economy, has been widely studied and applied [4–8].

Wang et al. [9], in his research on the mechanical efficiency of HMCVT, performed a theoretical reasoning and calculation of the mechanical efficiency of HMCVT and made a comparative analysis with an improved scheme. Cheng et al. [10] studied, based on the improved simulated annealing algorithm, the efficiency model of the hybrid continuously variable transmission. Li et al. [11, 12] analyzed the power distribution of compound and closed planetary gear transmissions and calculated the overall transmission efficiency with and without considering the power loss. The results showed that graphical representation is a practical power analysis and efficiency calculation method that can provide a theoretical reference for the design of complex closed planetary gear transmissions. Awadallah et al. [13] established a simulation model based on the mathematical model of conventional and mild hybrid power systems and studied its transmission efficiency through simulation.

Although simulation studies have high work efficiency and reliability, they lack experimental verification. In addition, transmission efficiency has consistently been the focus of research on EGT [14–17]. HMCVT has important features and advantages because of the organic combination of EGTs and HST. In the HST system [18, 19], when the working pressure Δ_p is improved, the mechanical efficiency η_{mh} of the system is gradually increased, whereas the volumetric efficiency η_v of the system is reduced. This changing regularity directly affects the working characteristics of HMCVT. Mechanical efficiency η_{mh} affects the output torque ratio of the hydraulic system, and volumetric efficiency η_v affects the output angular velocity ratio. Maintaining HMCVT's continuously variable transmission ratio in the HST system involves two aspects, namely, mechanical efficiency η_{mh} and volumetric efficiency η_v, both of which affect the overall efficiency of HMCVT. The effects of mechanical efficiency η_{mh} and volumetric efficiency η_v on the overall efficiency of HMCVT are studied to obtain a relatively uniform efficiency expression. Through numerical examples, this study provides a method to analyze the overall efficiency of HMCVT.

2. Parameter Matching

In accordance with the role of EGTs in HMCVT, HMCVTs with three active shafts [1] are divided into input coupled planetary (summing planetary) and output coupled planetary (divider planetary). The single variable hydraulic unit of HST is selected as the foundation of this research to simplify the derivation of the relationship between system parameters. In the system efficiency analysis (Section 5.2), the double-variable hydraulic units of HST are adopted to establish a method of improving system efficiency. As shown in Figure 1, an HMCVT with three active shafts and input coupled planetary is investigated in this work. The power splitting point lies in the intersection of link 1 and shaft I. In accordance with the difference in torques, the shaft throughout transmission system is divided into three links, namely, shaft I, link 1, and link 5. The power merging point is located between links 4 and 8. Angular velocities ωL ($\omega 6$) and ωH ($\omega 7$) present a low output speed and a high output speed of HMCVT, respectively. In Figure 1, Zi ($i = 1, 2, 3, 4, s, s', p, p'$, and r) is the tooth number for each meshing gear. Z_c is the equivalent tooth number of planetary carriers according to the installation diameter of the planetary gears.

Links 4, 5, 6, 7, and 8 comprise two-stage EGTs (the mechanical path). Link 6 is the planet carrier of EGTs, and link 7 is the sun gear of second-stage EGT. HST (the variable path) is composed of constant displacement (q_c) hydraulic unit 2 and variable displacement (q_v) hydraulic unit 1, with the former being connected to link 3. Links 1, 2, 3, and 4 comprise the internal transmissions with a fixed ratio.

2.1. Transmission Ratio Calculation. In the transmission scheme shown in Figure 1, the angular velocities of links 1 and 5 are equal to the input angular velocity.

$$\omega_{\mathrm{I}} = \omega_1 = \omega_5, \tag{1}$$

where ω_i denotes the actual angular velocity of the ith link. Displacement ratio e is equal to the displacement of the variable hydraulic unit (q_v) divided by the displacement of the constant unit (q_c), i.e., $e = q_v/q_c$. The angular velocity ratio of link 3 to link 2 is called the transmission ratio τ_H of the variable path. Given that the constant hydraulic unit acts as a motor or a pump, two expressions can be established as follows [1]:

$$\tau_H = \frac{\omega_3}{\omega_2} = \begin{cases} \dfrac{q_v \eta_v}{q_c} = e\eta_v, & \text{III (motor)}, \\[2ex] \dfrac{q_v}{(q_c \eta_v)} = \dfrac{e}{\eta_v}, & \text{I (pump)}. \end{cases} \tag{2}$$

The angular velocity ratio of link 2 to link 1 is called the fixed transmission ratio τ_{F1}, i.e., $\tau_{F1} = -Z_1/Z_2$. The angular velocity ratio of link 4 to link 3 is called the fixed transmission ratio τ_{F2}, that is, $\tau_{F2} = -Z_3/Z_4$. By using τ'_H to represent the total transmission ratio of parameters τ_{F1}, τ_{F2} and τ_H, we obtain

$$\tau'_H = \tau_H \tau_{F1} \tau_{F2} = \begin{cases} \dfrac{e\eta_v Z_1 Z_3}{(Z_2 Z_4)}, & \text{III}, \\[2ex] \dfrac{e Z_1 Z_3}{(Z_2 Z_4 \eta_v)}, & \text{I}. \end{cases} \tag{3}$$

The relative angular velocity of the ith link, ω_i^j, in a reference frame (the jth link) rotating with angular velocity ω_j is defined as

$$\omega_i^j = \omega_i - \omega_j. \tag{4}$$

In the EGTs, τ_{PG} and τ'_{PG} are defined as the ratio of ring gear angular velocity to sun gear angular velocity when the planetary carrier is fixed for the first and second EGTs, respectively,

$$\tau_{PG} = \frac{\omega_4^6}{\omega_5^6} = -\frac{Z_s Z_{p'}}{Z_p Z_r}, \tag{5}$$

$$\tau'_{PG} = \frac{\omega_4^6}{\omega_7^6} = -\frac{Z_{s'}}{Z_r}. \tag{6}$$

In the HMCVT, τ_{HM} and τ'_{HM} are defined as the ratio of output angular velocity to input angular velocity when ω_L and ω_H are required, respectively,

$$\tau_{HM} = \frac{\omega_L}{\omega_I} = \frac{\tau_{PG} - \tau'_H}{\tau_{PG} - 1}, \tag{7}$$

$$\tau'_{HM} = \frac{\omega_H}{\omega_I} = \frac{(\tau'_{PG} - \tau_{PG})\tau'_H + \tau_{PG}(1 - \tau'_{PG})}{\tau'_{PG}(1 - \tau_{PG})}, \tag{8}$$

where equations (3), (5), and (6) are utilized. We let $k = (1 - \tau_{PG}) - 1$ and $k' = (\tau'_{PG^-} - \tau_{PG})$ $(\tau'_{PG} - \tau'_{PG}\tau_{PG}) - 1$; then, equations (7) and (8) can be rewritten as

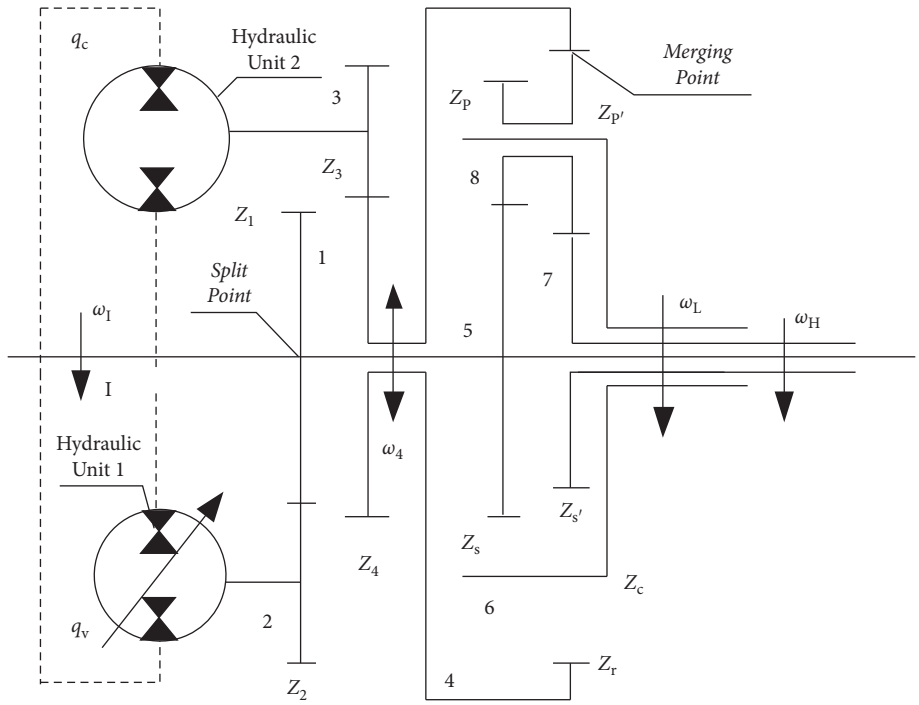


Figure 1: Sketch of the transmission system.

$$\tau_{HM} = k\tau'_H + 1 - k, \tag{9}$$

$$\tau_{HM} = k\tau'_H + (1 - k\prime). \tag{10}$$

2.2. Transmission Ratio Requirement. With parameter k_H to show the range of variable τ'_H and parameter λ to represent the asymmetry coefficient of τ'_H ($k_H > 0$, $\lambda > 0$), when τ'_H increases from $-k_H$ to λk_H, τ_{HM} increases gradually within $[0, \tau^0_{HM}]$, where τ^0_{HM} is the maximum of τ_{HM}.

When $\tau'_H = -k_H$ and $\tau_{HM} = 0$, we can deduce from equation (7) that $\tau_{PG} = -k_H$. When $\tau'_H = \lambda k_H$ and $\tau_{HM} = \tau^0_{HM}$, we can deduce from equation (7) that $\tau^0_{HM} = (1 + \lambda)k_H/(1 + k_H)$.

When τ'_H decreases from $[-k_H, \lambda_{kH}]$, τ'_{HM} increases gradually within $[\tau^0_{HM}, \tau^{\max}_{HM}]$, where $\tau^{\max}_{HM}$ is the maximum of τ'_{HM} ($\tau^0_{HM} < \tau^{\max}_{HM}$). When $\tau'_{HM} = \tau^0_{HM}$, we can deduce from equation (8) that

$$\begin{cases} \tau^0_{HM} = 1\,, \\ k_H = \dfrac{1}{\lambda}. \end{cases} \tag{11}$$

We can obtain the expression of τ'_{HM} from equation (8) by using equations (10) and (11). $t'_{HM} = t^{\max}_{HM}$, and $\tau'_{PG} = -1/\tau^{\max}_{HM}$.

When $\tau_{PG} = -1/\lambda$ and $\tau'_{PG} = -1/\tau^{\max}_{HM}$, τ'_H gradually increases in the range of $[-1/\lambda, 1]$. Correspondingly, transmission ratio τ_{HM} increases from 0 to 1, and transmission ratio τ'_{HM} decreases from $\tau^{\max}_{HM}$ to 1. We can deduce from equations (9) and (10) that

$$\begin{cases} k = \dfrac{-1}{\tau_{PG} - 1} = \dfrac{\lambda}{\lambda + 1}, & 0 < k < 1, \\ k\prime = \dfrac{\tau'_{PG} - \tau_{PG}}{\tau'_{PG}(1 - \tau_{PG})} = -\dfrac{\lambda}{\lambda + 1}\left(\tau^{\max}_{HM} - 1\right) & k' < 0. \end{cases} \tag{12}$$

An example is provided throughout the paper. When $\tau_{F1}\tau_{F2} = 1$, $\lambda = 1$ and $\tau^{\max}_{HM} = 2.7$, with the above mentioned equations, we can obtain

$$\begin{cases} \tau_{PG} = -1, \\ \tau'_{PG} = -0.37, \end{cases} \tag{13}$$

$$\begin{cases} \tau_{HM} = 0.5\tau'_H + 0.5, \\ \tau'_{HM} = -0.85\tau'_H + 1.85. \end{cases} \tag{14}$$

The tooth numbers of the EGTs fitting equations (13) and (14) are summarized in Table 1.

3. Power Flows in the Absence of Losses

3.1. Angular Velocity Ratios in the Absence of Losses. In the example mentioned, considering the absence of losses, $\tau'_H = \tau_H = e$. By using equations (9) and (10), the relationship between output angular velocity ratio τ_{HM} of HMCVT and displacement ratio e is shown in Figure 2(a).

3.2. Direction of Power Flows. According to Macmillan [20], the direction of power flow in EGT cannot be modified by meshing friction. Torque equilibrium and power conservation equations are widely used to determine the efficiency

Table 1: Tooth numbers of the planetary gear train.

Z_s	Z_r	Z_c	Z_p	$Z_{s'}$	$Z_{p'}$
48	92	63	15	34	29

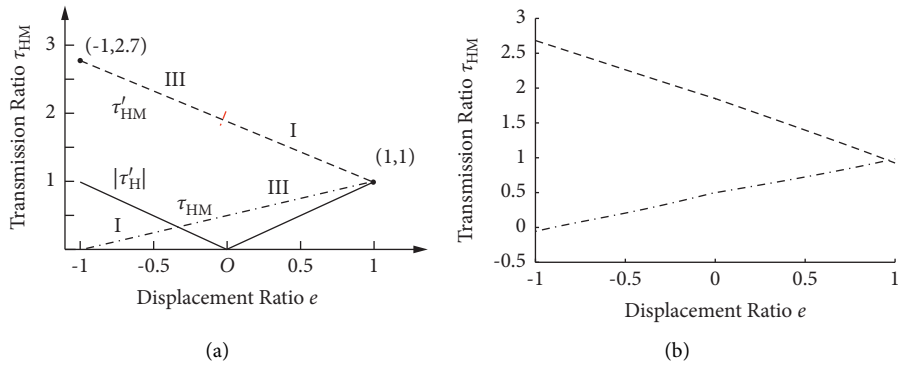


Figure 2: Relationship between output transmission ratio τ_{HM} and displacement ratio e. (a) In the absence of losses. The working area of τ_{HM} is formed by dot-dashed and double-dot-dashed lines, and the working area of τ'_{HM} is formed by dashed and long dashed lines. The solid line is $|\tau'_H|$. HMCVT realizes continuously variable transmission in the range of [0, 2.7], shifting with no speed difference in point (1,1). (b) With the losses calculated by Sts1 in Section 5.2.

of EGT [1, 21]. In this study, the power that flows into EGT is positive, and the outflow is negative.

On steady state, directions of the possible power flow [22–24] are those shown in Figure 3 as type I, type II, and type III flow.

When power conservation and torque equilibrium equations are applied to EGT, we can obtain the following from Figure 3:

$$T_O\omega_O + T_4\omega_4 + T_5\omega_5 = 0, \tag{15}$$

$$T_O + T_4 + T_5 = 0, \tag{16}$$

where T_i is the torque applied on the ith link. Without loss of generality, we can assume that the output shaft of the EGTs is link 7. In this case, by using equations (1) and (10), we obtain

$$\omega_O = \omega_I\tau'_{HM} = \omega_4 k' + \omega_5(1-k'), \tag{17}$$

$$\frac{T_7}{-1} = \frac{T_4}{k'} = \frac{T_5}{1-k'}. \tag{18}$$

In type I power flow with output shaft 7 [22], equation (15) is rewritten as

$$|T_7\omega_7| + |T_4\omega_4| - |T_5\omega_5| = 0. \tag{19}$$

By substituting equations (10), (18), and (17) in equation (19), we obtain

$$|\tau'_H| = \left|\frac{1-k'}{k'}\right| - \frac{1}{|k'|}|\tau'_{HM}|. \tag{20}$$

As indicated in the first quadrant of Figure 2(a), equation (10) is rewritten as

$$|\tau'_{HM}| = |1-k'| - |k'\tau'_H| \quad k' < 0 \ \tau'_H > 0. \tag{21}$$

In consideration of equations (18) and (21), equation (20) is identical. When the output shaft is link 7, the work area for generating type I power flow is shown in the first quadrant of Figure 2(a) by long dashed lines.

The direction of power flow relative to every working condition can be obtained. Figure 2(a) shows every type of power flow with symbols I and III.

PM is the HST system (the variable path). PG is the two-stage EGTs (the mechanical path). F1 and F2 are the internal transmissions with a fixed ratio. 1, 2, 3, 4, 5, I, and O are the links of HMCVT shown in Figure 1.

4. Efficiency of HMCVT

4.1. Torque Ratio of Links with Losses. The pressure efficiency ηp of the HST system is approximately expressed as $\eta p = p_P/p_M$, where p_P and p_M are pump and motor operating pressures, respectively. In the HST system, total mechanical efficiency η_{mh} is expressed as $\eta_{mh} = \eta_P\eta_P\eta_M$, where η_M and η_P are pump and motor mechanical efficiencies, respectively. The output torque of motor T_M and the input torque of pump T_P are expressed as

$$T_P = \frac{q_P p_P}{2\pi\eta_P}, \qquad T_M = \frac{q_M p_M \eta_M}{2\pi}, \tag{22}$$

where q_P and q_M are the displacements of the pump and motor, respectively.

In type III power flow, hydraulic unit 1 acts as the variable pump with a torque of T_P, and hydraulic unit 2 acts

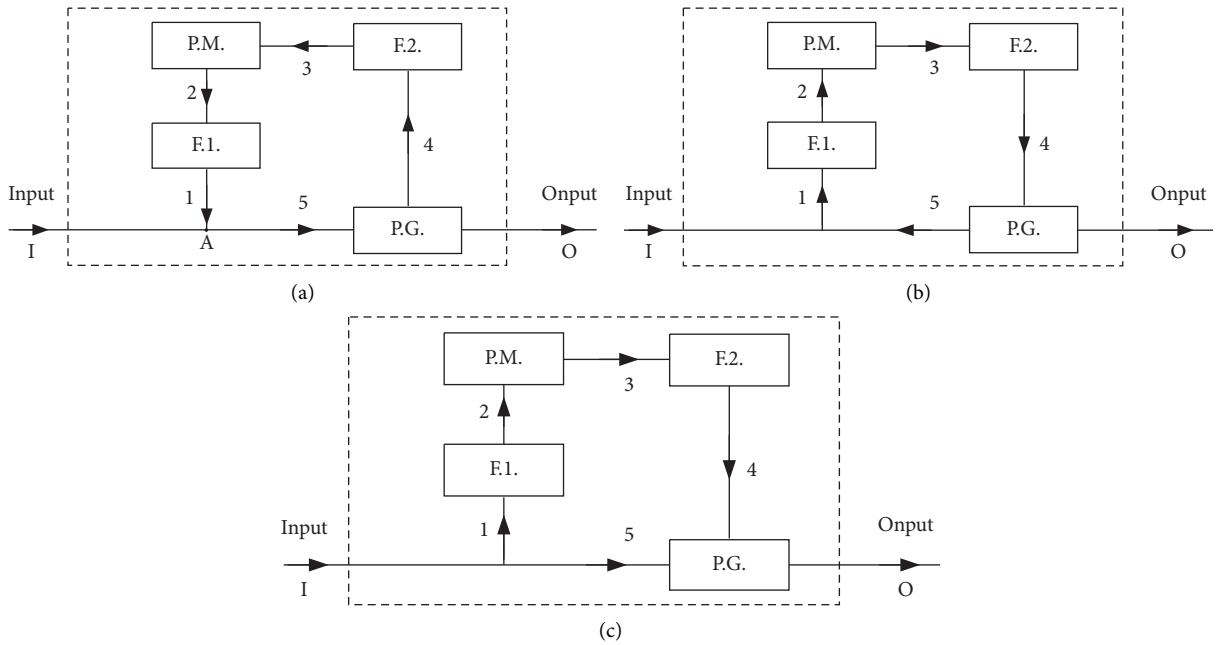


Figure 3: Power flows in HMCVT. (a) Type I power flow. (b) Type II power flow. (c) Type III power flow.

as the motor with a torque of T_M. The torque ratio Δ_H of HST in type III power flow is defined as T_P divided by T_M.

$$\Delta_H = \frac{T_P}{T_M} = \left|\frac{T_2}{T_3}\right| = \left|\frac{q_P p_P}{q_M p_M \eta_M \eta_P}\right| \approx \frac{|e|}{\eta_M \eta_P \eta_p} = \frac{|e|}{\eta_{mh}}. \quad (23)$$

The torque ratio Δ_{12} of F1 to F2 (shown in Figure 3; internal transmissions) in type III power flow is defined as

$$\Delta_{12} = \frac{\tau_{F1}\tau_{F2}}{\eta_{F1}\eta_{F2}}. \quad (24)$$

In type I power flow, hydraulic unit 1 acts as the variable motor with a torque of T_M, and hydraulic unit 2 acts as the pump with a torque of T_P. Δ'_H is defined as T_P divided by T_M.

$$\Delta'_H = \frac{T_P}{T_M} = \left|\frac{T_3}{T_2}\right| = \left|\frac{q_P p_P}{q_M p_M \eta_M \eta_P}\right| \approx \frac{1}{|e|\eta_M \eta_P \eta_p} = \frac{1}{|e|\eta_{mh}}. \quad (25)$$

The torque ratio Δ'_{12} of F1 to F2 in type I power flow is defined as

$$\Delta'_{12} = \frac{1}{\eta_{F1}\eta_{F2}\tau_{F1}\tau_{F2}}. \quad (26)$$

Given that the transmission losses and the torques applied to the links are unrelated to the observer's motion, the method of epicyclic inversion was used by Pennestri and Freudenstein [14] and Chen and Liang [21] in different ways.

This study uses the transforming mechanism model [21, 25], in which the interaction forces and the relative angular velocity are all the same as the EGTs shown in Figure 1, but the planet carrier is fixed. Thus, the losses in the two models are considered to be the same. The torque ratio of the links based on meshing power [21, 25] is applied to calculate the efficiency of the two-DOF EGTs. Meshing power p_i^c (the ith link relative to planet carrier 6) is a virtual power that is independent of the actual power p_i. Meshing power ratio ψ_i is defined as the ratio of meshing power p_i^c through the ith link to actual power p_i through the same link. If power pi flows into EGT, it is defined as positive (negative otherwise). The expressions of ω_i^6, p_i^6, p_i and ψ_i are

$$\begin{cases} \omega_i^6 = \omega_i - \omega_6, \\ p_i^6 = T_i(\omega_i - \omega_6) = T_i\omega_i^6, \\ p_i = T_i\omega_i, \\ \psi_i = \dfrac{p_i^6}{\substack{i \\ p}} = \dfrac{T_i(\omega_i - \omega_6)}{T_i\omega_i} = \dfrac{\omega_i^6}{\omega_i}. \end{cases} \quad (27)$$

In the transforming mechanism, the direction of p_i^c can be determined in accordance with the sign of p_i and ψ_i. If the sign of $p_i^c$is positive, power p_i^c flows into the transforming mechanism; otherwise, it flows out of the system.

Without loss of generality, we assume that the output shaft of the HMCVT is link 7 in type I power flow, and the relationship among ω_i, ψ_i and p_i between links is$\omega_7 > \omega_5 > \omega_6 > \omega_4 > 0$.

$$\begin{cases} \psi_5 = \dfrac{p_5^6}{p_5} = \dfrac{T_5(\omega_5 - \omega_6)}{T_5\omega_5} > 0, & p_5 < 0, p_5^6 > 0, \\ \psi_4 = \dfrac{p_4^6}{p_4} = \dfrac{T_4(\omega_4 - \omega_6)}{T_4\omega_4} < 0, & p_4 < 0, p_4^6 > 0, \\ \psi_7 = \dfrac{p_7^6}{p_7} = \dfrac{T_7(\omega_7 - \omega_6)}{T_7\omega_7} > 0, & p_7 < 0, p_7^6 < 0. \end{cases} \tag{28}$$

Under this condition, the meshing power is from links 5 and 4 to link 7 in the transforming mechanism. When the power conservation equation and the torque equilibrium condition are applied to the EGT, we have

$$\begin{cases} T_4\omega_4^6\eta_{48}^6\eta_{87}^6 + T_5\omega_5^6\eta_{58}^6\eta_{87}^6 + T_7\omega_7^6 = 0, \\ T_4 + T_5 + T_7 = 0, \end{cases} \tag{29}$$

where η_{ij}^6 is the efficiency of the path, in which the meshing power flows from the ith link to the jth link with the planet carrier being fixed in the transforming mechanism.

To obtain the relationship between output torque and input torque, by using equations (5) and (6), we can reform equation (29) to

$$\begin{cases} T_7 = T_5\dfrac{\eta_{48}^6\eta_{87}^6\tau'_{PG}\tau_{PG} - \eta_{58}^6\eta_{87}^6\tau'_{PG}}{\left(1 - \eta_{48}^6\eta_{87}^6\tau'_{PG}\right)\tau_{PG}}, \\ T_4 = -T_5\left(1 + \dfrac{\eta_{58}^6\eta_{87}^6\tau'_{PG}\tau_{PG} - \eta_{58}^6\eta_{87}^6\tau'_{PG}}{\left(1 - \eta_{48}^6\eta_{87}^6\tau'_{PG}\right)\tau_{PG}}\right). \end{cases} \tag{30}$$

From equation (28), torque ratios Δ_{75} and Δ'_{45} of EGT (shown in Figure 3; the mechanical path) in type I power flow are defined as

$$\begin{cases} \Delta_{75} = \dfrac{\eta_{48}^6\eta_{87}^6\tau'_{PG}\tau_{PG} - \eta_{58}^6\eta_{87}^6\tau'_{PG}}{\left(1 - \eta_{48}^6\eta_{87}^6\tau'_{PG}\right)\tau_{PG}}, \\ \Delta'_{45} = -\left(1 + \dfrac{\eta_{48}^6\eta_{87}^6\tau'_{PG}\tau_{PG} - \eta_{58}^6\eta_{87}^6\tau'_{PG}}{\left(1 - \eta_{48}^6\eta_{87}^6\tau'_{PG}\right)\tau_{PG}}\right). \end{cases} \tag{31}$$

When the output shaft of the EGT is link 7 in type III power flow, in reference to the actual angular velocity ω_i, the relationship between links is $\omega_7 > \omega_5 > \omega_6 > 0 > \omega_4$. The same expression of Δ_{75} and Δ'_{45} as that in equation (29) can be obtained because the sign of p_i^c remains similar to that in type I power flow. The other case is that the output shaft of HMCVT is link 6 in types III and I power flows. In this condition, we can obtain the expressions of Δ_{65} and Δ_{45} in a similar manner.

$$\begin{cases} \Delta_{45} = -\dfrac{\eta_{58}^6\eta_{84}^6}{\tau_{PG}}, \\ \Delta_{65} = \dfrac{\eta_{58}^6\eta_{84}^6}{\tau_{PG}} - 1. \end{cases} \tag{32}$$

4.2. Efficiency of HMCVT. In the steady state of HMCVT, we assume that output torque *To* of HMCVT is the result of the calculation and not of the actual load, such that the input torque and angular velocity remain similar to those for diesel working under rated conditions. When the power pi contributed to T_i flows into the Split point in Figure 1, T_i is defined as positive; otherwise, it is negative. Here, T_5 is negative in type I and III power flows. To maintain consistency with the signs of T_5 and T_I for EGTs, when the torque equilibrium condition is applied to the Split point, we have

$$T_I - T_5 - T_1 = 0. \tag{33}$$

Depending on the type of power flow with output shaft 6, two kinds of relationship exist between T_1 and T_5.

$$T_1 = \begin{cases} \Delta_H|\Delta_{45}|\Delta_{12}T, & \text{III}, \\ -\dfrac{|\Delta_{45}|}{\Delta'_H\Delta'_{12}}T, & \text{I}. \end{cases} \tag{34}$$

The relationship between T_5 and T_I is

$$T_5 = \begin{cases} \dfrac{T_I}{1 + \Delta_H|\Delta_{45}|\Delta_{12}}, & \text{III}, \\ \dfrac{T_I}{1 - |\Delta_{45}|/\Delta'_H\Delta'_{12}}, & \text{I}. \end{cases} \tag{35}$$

We obtain the transmission efficiency of the HMCVT with output shaft 6 by using the relationship expressions of T_I with T_5 and equation (7) as follows:

$$\eta_{HM} = -\frac{T_6\omega_6}{T_I\omega_I} = \begin{cases} -\dfrac{\Delta_{65}T_I\omega_6/\left(1 + \Delta_H|\Delta_{45}|\Delta_{12}\right)}{T_I\omega_I} = \dfrac{-\Delta_{65}\tau_{HM}}{1 + \Delta_H|\Delta_{45}|\Delta_{12}}, & \text{III}, \\ -\dfrac{\Delta_{65}T_I\omega_6/\left(1 - \left(|\Delta_{45}|/\Delta'_H\Delta'_{12}\right)\right)}{T_I\omega_I} = \dfrac{-\Delta_{65}\tau_{HM}}{1 - \left(|\Delta_{45}|/\Delta'_H\Delta'_H\right)}, & \text{I}. \end{cases} \tag{36}$$

The working pressure of the hydraulic system is represented by the working pressure of the fixed displacement hydraulic unit 2.

$$\Delta p = \frac{2\pi|T_3|}{q_c} = \begin{cases} \dfrac{2\pi|\tau_{F2}\Delta_{45}|T_5}{q_c\eta_{F2}} = \dfrac{2\pi|\tau_{F2}\Delta_{45}|T_I}{q_c\eta_{F2}\left(1+|\Delta_{45}|\Delta_{12}\Delta_H\right)}, & \text{III}, \\ \dfrac{2\pi|\tau_{F2}\Delta_{45}|T_5\eta_{F2}}{q_c} = \dfrac{2\pi|\tau_{F2}\Delta_{45}|\eta_{F2}T_I}{q_c\left(1-\left(|\Delta_{45}|/\Delta'_{12}\Delta'_H\right)\right)}, & \text{I}. \end{cases} \tag{37}$$

When the output shaft is link 7, equations (32), (34), and (35) remain the same, but Δ_{45}, Δ_{65} and τ_{HM} are replaced by Δ'_{45}, Δ_{75}, and τ'_{HM}, respectively:

$$T_1 = \begin{cases} \Delta_{\mathrm{H}}|\Delta'_{45}|\Delta_{12}T_5, & \text{III}, \\ -\dfrac{|\Delta'_{45}|}{\Delta'_H\Delta'_{12}}T_5, & \text{I}, \end{cases} \tag{38}$$

$$\eta'_{HM} = -\frac{T_7\omega_7}{T_{\mathrm{I}}\omega_{\mathrm{I}}} = \begin{cases} \dfrac{-\Delta_{75}\tau'_{HM}}{1+\Delta_{\mathrm{H}}|\Delta'_{45}|\Delta_{12}}, & \text{III}, \\ \dfrac{-\Delta_{75}\tau'_{HM}}{1-\left(|\Delta'_{45}|/\Delta'_H\Delta'_{12}\right)}, & \text{I}, \end{cases} \tag{39}$$

$$\Delta' p = \frac{2\pi|T_3|}{q_{\mathrm{c}}} = \begin{cases} \dfrac{2\pi|\Delta'_{45}\tau_{F2}|T_I}{q_c\eta_{F2}\left(1+|\Delta'_{45}|\Delta_{12}\Delta_H\right)}, & \text{III}, \\ \dfrac{2\pi|\Delta'_{45}\tau_{F2}|\eta_{F2}T_I}{q_c\left(1-\left(|\Delta'_{45}|/\Delta'_{12}\Delta'_H\right)\right)} & \text{I}. \end{cases} \tag{40}$$

When the losses from the churning of the lubricating oil and shaft bearing friction are not considered, once the characteristics of the HMCVT have been determined, torque ratios Δ_{12}, Δ'_{12}, Δ_{75}, Δ'_{45}, Δ_{65}, and Δ_{45} become constant. However, firstly, torque ratios Δ_H and Δ'_H vary with the change in mechanical efficiency η_{mh} and displacement ratio e for the hydraulic system. Secondly, τ_{HM} and τ'_{HM} vary with the change in volumetric efficiency η_v and displacement ratio e. Thirdly, volumetric efficiency η_v and mechanical efficiency η_{mh} change with displacement ratio e and system working pressure Δ_p. According to equation (37) or (40), the curved surface of η_{mh} relative to e and Δ_p is called the mechanical system demanding surface.

5. Efficiency Analysis

5.1. Numerical Example. The mechanical parameters of the example are displayed in equations (13) and (14) and Table 1. The loss of a conventional gear pair is estimated via the following formula [13]:

$$\psi^{\mathrm{C}}_{mn} = 0.23\left(\frac{1}{Z_1} \pm \frac{1}{Z_2}\right), \tag{41}$$

where ψ^{C}_{mn} is the meshing loss of the power flowing through the mth and nth links when the gear carrier is fixed and Z_1 and Z_2 are the teeth numbers of the small and large gears of this pair, respectively.

By using the equations above, we can obtain constant torque ratios. The parameters of the example are listed in Table 2.

By using the hydraulic parameters η_H, η_v and η_{mh} relative to Δ_p and e [26] listed in Tables 1 and 2, the expressions of the fitting surfaces are obtained, with the displacement of hydraulic unit 2 being fixed to the maximum value.

$$\begin{aligned} \eta_H &= \left(\Delta p^2\ \Delta p\ 1\right)\cdot\begin{pmatrix} -0.000203 & 0.000244 & -0.000402 \\ 0.010453 & -0.011201 & 0.022677 \\ -0.376406 & 0.618912 & 0.267755 \end{pmatrix}\left(e^2\ e\ 1\right)^T, \\ \eta_v &= \left(\Delta p^2\ \Delta p\ 1\right)\cdot\begin{pmatrix} -0.000145 & 0.000263 & -0.000143 \\ -0.000594 & 0.001655 & -0.002349 \\ 0.001936 & -0.008338 & 0.998824 \end{pmatrix}\left(e^2\ e\ 1\right)^T, \\ \eta_{mh} &= \left(\Delta p^2\ \Delta p\ 1\right)\cdot\begin{pmatrix} -0.000039 & -0.000055 & -0.000251 \\ 0.008926 & -0.008787 & 0.023435 \\ -0.377170 & 0.621347 & 0.263924 \end{pmatrix}\left(e^2\ e\ 1\right)^T. \end{aligned} \tag{42}$$

The analysis steps (Sts1) for obtaining the HMCVT efficiency with the single variable unit are shown below, and the results are listed in Figure 4 and Table 3.

(1) Two curved surfaces in the same coordinate system are established. One is the curved surface of the hydraulic system's mechanical efficiency, and the other is the mechanical system demanding surface. To obtain the relationships between hydraulic system mechanical efficiency η_{mh} and system pressure Δ_p with displacement ratio e, the data on the intersection of surfaces are fitted by a polynomial.

(2) Torque ratios Δ_H and Δ'_H are obtained from the equations above. By substituting the polynomial of

Table 2: Parameters of HMCVT.

Δ_{12}	Δ'_{12}	Δ_{75}	Δ'_{45}	Δ_{65}
1.02	1.02	−0.528	−0.472	−1.975
Δ_{45}	τ_{F2}	T_I (N·m)	q_c ($cm^3 \cdot r^{-1}$)	Δ_{45}
0.975	0.729	702.5	107	0.975

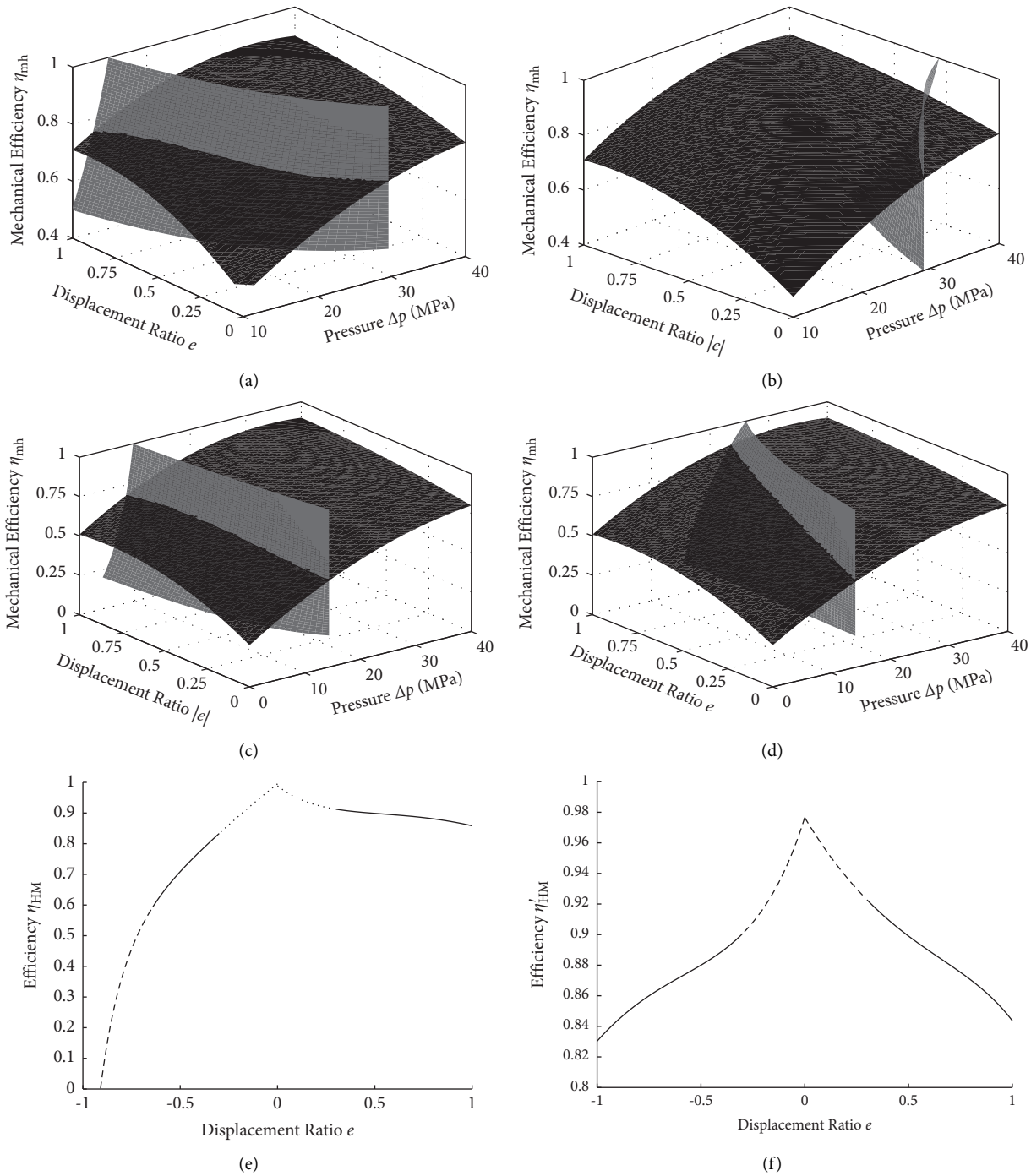


Figure 4: Results of the efficiency analysis. (a) Intersection of surfaces in type III power flow with output shaft 6. (b) Intersection of surfaces in type I power flow with output shaft 6. (c) Intersection of surfaces in type III power flow with output shaft 7. (d) Intersection of surfaces in type I power flow with output shaft 7. (e) Overall efficiency of HMCVT. (f) Overall efficiency of HMCVTwith output shaft 7.

Table 3: Polynomial coefficients of ratios.

Type	Δ_H or $1/\Delta'_H$	τ_H
6 III Δ_H	[0.3515, −0.4770, 1.4554, −0.0042]	[−0.1093, 0.2033, 0.8792, −0.0027]
6 I $1/\Delta'_H$	[−0.0908, 0.1886, 0.8034, −0.0028]	[−0.4537, 0.6241, −0.0107, −1.2757, 0.0027]
7 III Δ_H	[0.0112, 0.6150, −0.9826, 1.8224, 0.0002]	[0.0324, −0.0628, −0.9496, 0.0005]
7 I $1/\Delta'_H$	[−0.1881, 0.5067, 0.5532, −0.0005]	[0.0497, −0.0399, −0.0171, 1.0616, 0.0002]

Δ_p relative to |e| in the fitting formula for η_v, the relationship between η_v and τ_H with $|e|$ is obtained.

(3) The overall angular velocity ratios τ_{HM} and τ'_{HM} of the HMCVT are obtained.

(4) The overall efficiencies η_{HM} and η'_{HM} of the HMCVT are obtained.

By observing the overall efficiency of output shaft 6 in Figure 4(e), we find that the least efficiency appears at the start of the vehicle, and $-1 < e < -0.9$. The actual numerical result is negative. As shown in Figure 5(b), when $-1 < e < -0.9$, $\tau_H < -1$. In this condition, the direction of output shaft 6 is reversed. Therefore, we can assume that the vehicle should be started at an e of about −0.9, and |e| should be gradually reduced to improve the speed of the vehicle.

(1) The dotted line indicates that the curve is out of the fitting data source. (2) Denoted by a line of long dashes, this working area without the full input power is called traction limiting, indicating that the efficiency curve is highly inaccurate at the start of the entire machine. It is calculated with a peak pressure of 40 MPa of the hydraulic system.

When the vehicle speed is lower than 6 km/h, the working condition without the full input power is called traction limiting. When $\Delta_p > 40$ MPa in type I power flow with output shaft 6, the efficiency is calculated using the peak pressure (40 MPa) of the hydraulic system, as shown in Figures 4(b) and 5(a). This kind of calculation reflects the carrying capacity of HMCVT, which is independent of the weight of the whole vehicle and the ground adhesion performance.

In the steady state, with the vehicle speed being larger than the maximum speed in type I power flow with output shaft 6 (about 7.4 km/h in the example), the maximum working pressure of the hydraulic system is 29.57 MPa when the system output shaft is link 6 and $e = 0$. Considering that the coefficient of torque reservation for the diesel engine is about 40%, when the system is working under rated conditions, the suitable hydraulic system pressure is about 30 MPa.

By substituting the polynomials for τ_H (in Table 3) in equation (14), the relationship between angular velocity ratio τ_{HM} of HMCVT and displacement ratio e is obtained, as shown in Figure 2(b). When the system output shaft is link 7, the hydraulic system pressure is generally lower than that for output link 6 because torque ratio Δ'_{45} is smaller than Δ_{45}. The output angular velocity of link 7 is greater than that of link 6, as shown in Figure 2(b).

5.2. HMCVT with Double-Variable Hydraulic Units. In the HMCVT system with double-variable hydraulic units, hydraulic unit 2 is the variable displacement (q'_vq'v) hydraulic unit connected to link 3 in Figure 1. We can obtain a relatively stable e through different combinations of variable qv and variable q'_v, which allow hydraulic units 1 and 2 to work in a relatively high-efficiency area. Without consideration of fuel consumption and vehicle dynamic characteristics, the problem of HMCVT efficiency changes. How variables qv and q'_v should be varied to obtain the maximum system efficiency in the steady state under the rated conditions of the diesel engine is the new problem to be solved.

The analysis steps (Sts2) for obtaining the curve of variables q'_v ($q_v = q'_v e$) are shown as follows, and the results obtained with the example above are listed in Figure 6:

(1) When variable $e = \text{constant}$, $e' = q'_v / q'_{v\max}$, where $q_{v\max}{}'$ is the maximum value of variable q'_v. The fitting formulas (curved surfaces) of η_v and η_{mh} relative to Δ_p and e' are obtained with the changed displacement of hydraulic unit 2.

(2) To obtain the relationship between overall efficiency η and variable e', the steps of Sts1 are completed with q'_v replacing q_c. Attention is given to the difference between e and e'.

(3) In the range of variable e, processes (1) and (2) are repeated. In the three-dimensional coordinate system, the curved surface that represents η relative to e and e' is obtained, and the points (e, e', η) that have the highest efficiency for each e in step (b) are connected.

By substituting the polynomial Δ'_H or Δ_H, equations (30) and (31) in equation (34), we obtain the relationship between output torque T_o of HMCVT and the foreword speed of the vehicle, whilst assuming that the minimum speed is 0, and the maximum speed is 39.96 km/h, as shown in Figure 6(h).

When the machine is started with type I power flow with output shaft 6, most of the time, it works on $e' = 1$ or $\Delta' p = 40$ MPa, and the efficiency curve is almost the same as that of the single variable unit system. In addition, from the comparisons of the efficiency curves with outputs 6 and 7 in Figures 6(e) and 6(f), we find that the overall efficiency of HMCVT with double-variable hydraulic units is improved. The maximum value of q'_v is about 91 $\text{cm}^3 \cdot \text{r}^{-1}$, and the system variable q'_v is effectively reduced.

5.2.1. The Solid Curves Belong to the HST with a Single Variable. As shown by the curves of Δ_p relative to e, the HST of HMCVT always works at the medium-pressure stage, in which the relative maximum efficiency of the hydraulic system can be obtained. However, the working pressure is unstable, mainly because the maximum values

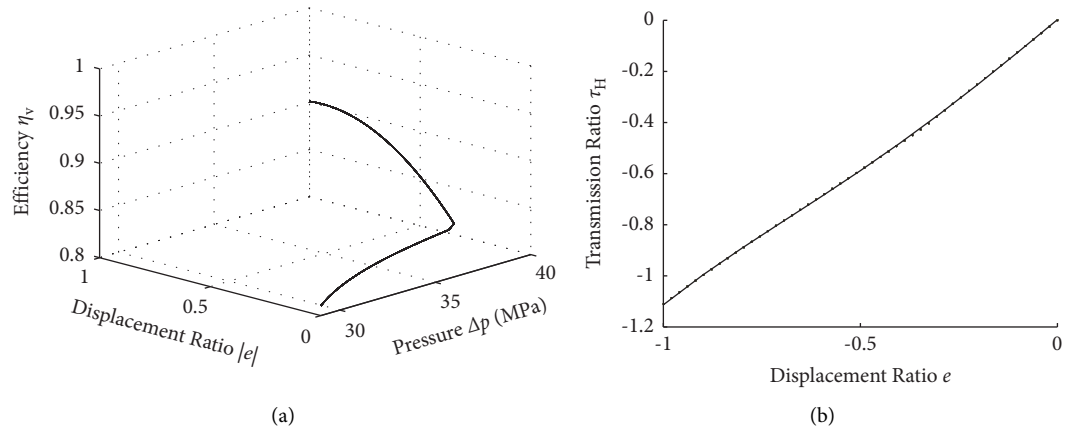


Figure 5: Curves η_v and τ_H of the numerical results for type I power flow with output shaft 6.

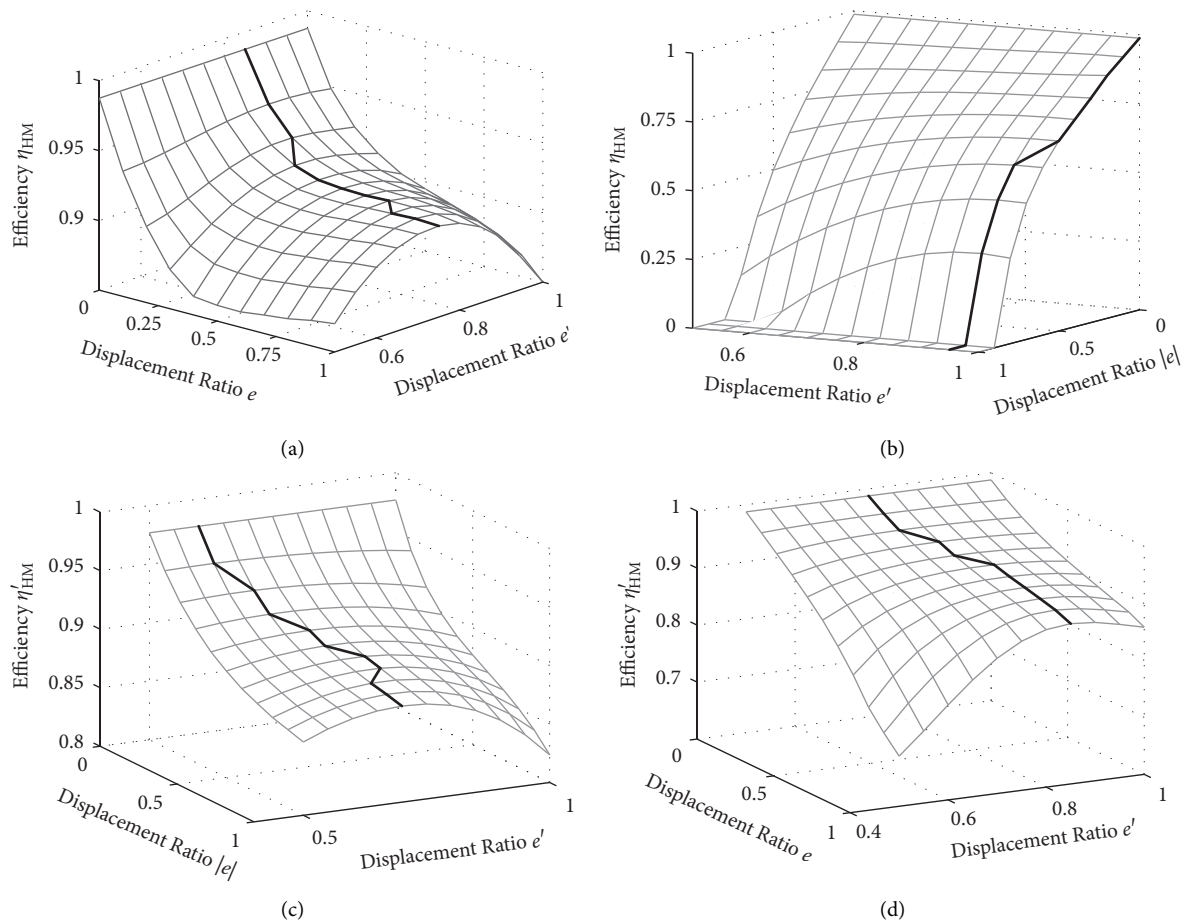


Figure 6: Continued.

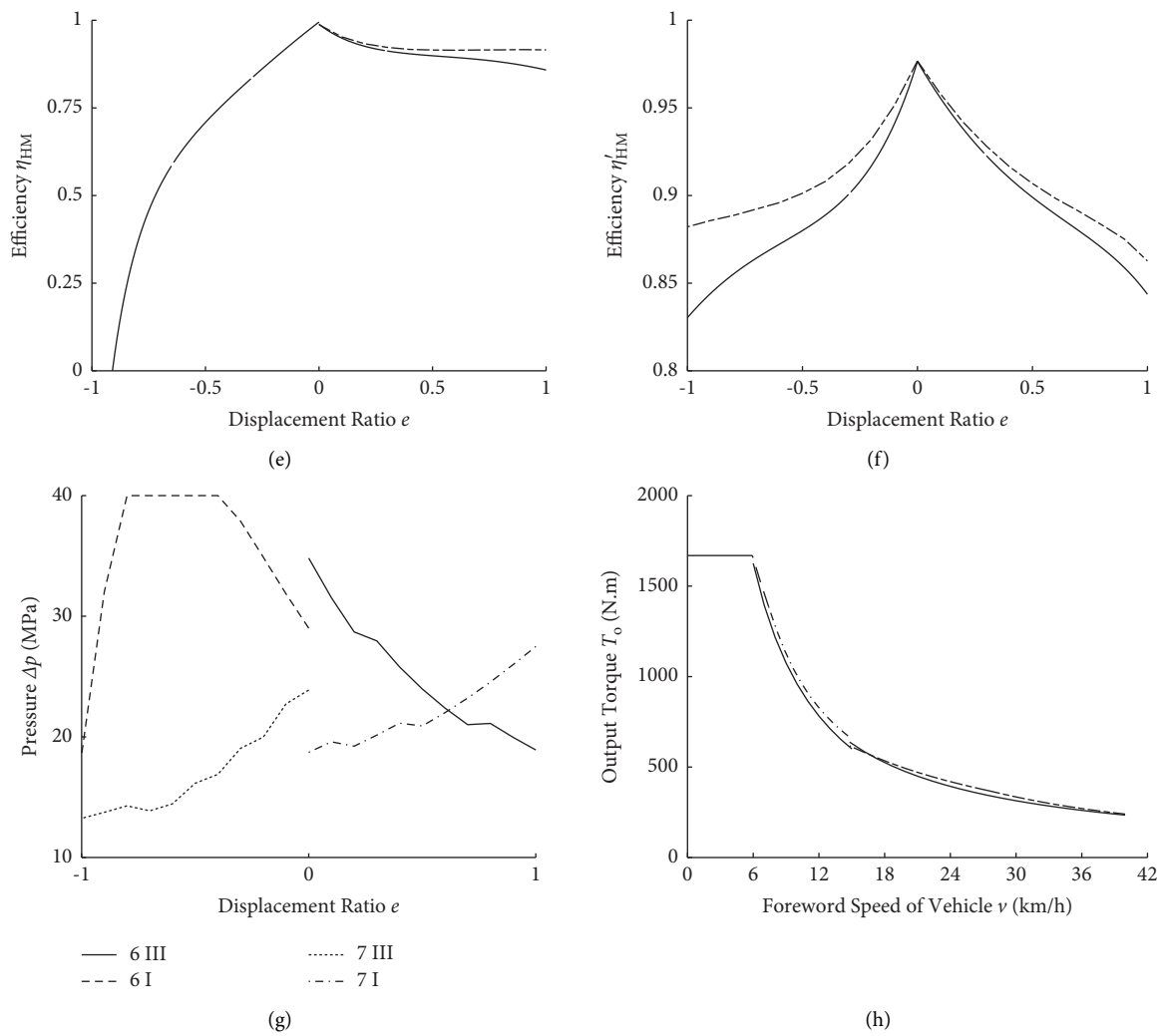


Figure 6: Results of Sts2 with double-variable hydraulic units. (a) In type III with output link 6. (b) In type I with output link 6. (c) In type III with output link 7. (d) In type I with output link 7. (e) Comparison of efficiency with 6. (f) Comparison of efficiency with 7. (g) Curves of Δ_p relative to e. (h) Comparison of T_o relative to v.

of mechanical efficiency η_{mh} and volumetric efficiency η_v cannot be obtained at the same time, and because the positions of η_v and η_{mh} in the efficiency expression are different. As shown by the curves of Δ_p relative to e, when $e = 0$, the mutations of working pressure are mainly due to the mutations of e'.

At the start of the operation of the entire machine in type I power flow with output shaft 6, an effective way to improve system efficiency is to increase the q'vmax of hydraulic unit 2, as shown in Figure 7, by using the same curved surfaces of η_v and η_{mh} without consideration of the influence of maximum displacement $q'_{v\max}$ on system efficiency η_v and η_{mh}.

To increase the maximum speed of the vehicle with dual variable units, a maximum speed greater than 60 kph can be obtained by setting variable $e < -2.6$ in the type III power flow with output shaft 7. Let $e'' = q_v/q_{v\max}$, where $q_{v\max}$ is the maximum value of variable q_v. To obtain the curved surface that represents η relative to e and e'', the processes of Sts2 are repeated, with e'' ($=ee'$) replacing e'. The points (e, e'', η) that have the highest efficiency for each e are connected.

When $e < -1.8$, $e'' = 1$, as shown in Figure 8. This condition indicates that the transmission efficiency of HMCVT can be improved when the value of qvmax is increased and $e < -1.8$, but the improvement is not significant.

6. Test Verification

The effectiveness of the HMCVT transmission efficiency calculation method proposed in this study is verified, and an HMCVT gearbox with a multisegment univariate unit similar to that in Figure 1 is selected as the test object. The test site of the physical prototype TA1-02 is shown in Figure 9. Torque and speed sensors HBM3000 and HBM5000 with ranges of 3000 and 5000 N m, respectively, are installed on the output shaft of the DC variable frequency motor, the output shaft of the HMCVT, and the hydraulic motor of the energy recovery device. System pressure test

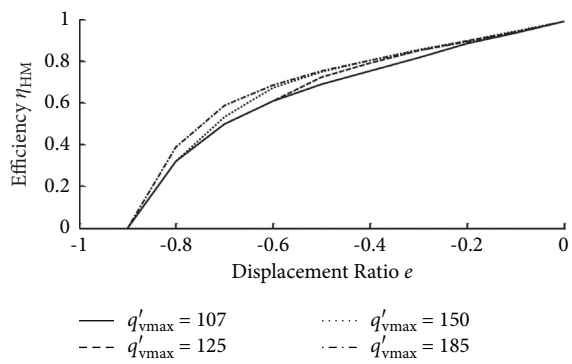


Figure 7: Relationship of η_{HM} with e and $q'_{v\max}$ in type I power flow with output shaft 6.

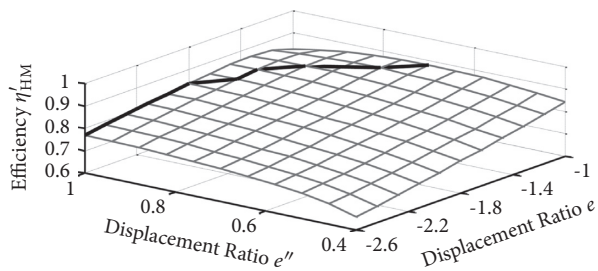


Figure 8: Curved surfaces of η'_{HM} relative to e and e'' in type III power flow with output shaft 7 when $e < -1$.

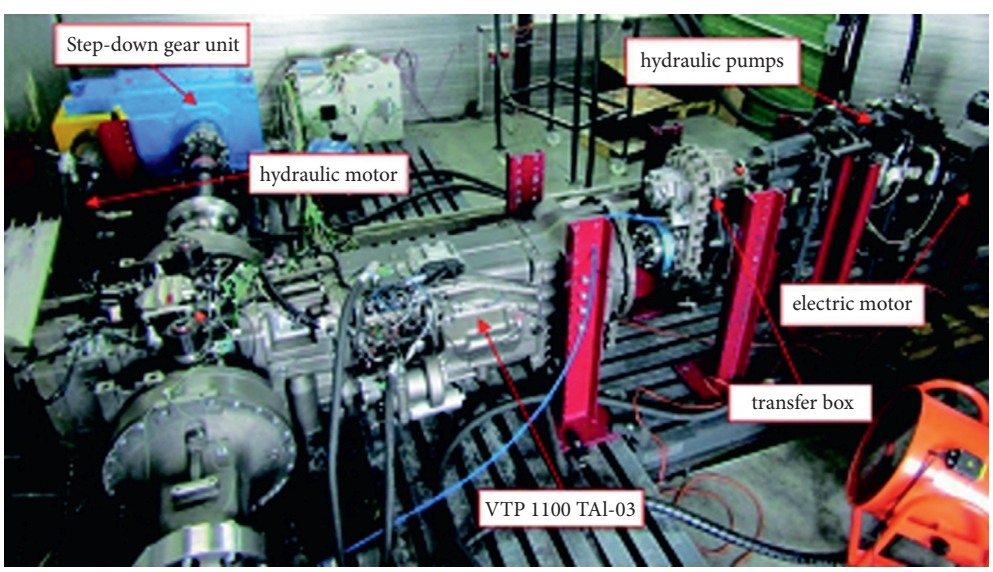


Figure 9: Graph of systematic experiments.

data from the transmission TCU are read via CAN bus. The test arrangement is applied in accordance with the gear shift strategy of the tractor's traction conditions. The input torque is stable at 1030 N m, and the input speed is stable at 1700 r·min^{-1}. These values are gradually increased according to the transmission ratio, and A, B, C, D, E, and F are selected for a total of six working conditions for the bench test.

Some of the data collected on site are shown in Figure 10. The input torque is represented by a blue line, the pressure of the B channel between the hydraulic units is represented by the red line, and the high-pressure-side system pressure is represented by a yellow line.

The transmission efficiency comparison is shown in Figure 11, and the system pressure comparison is shown

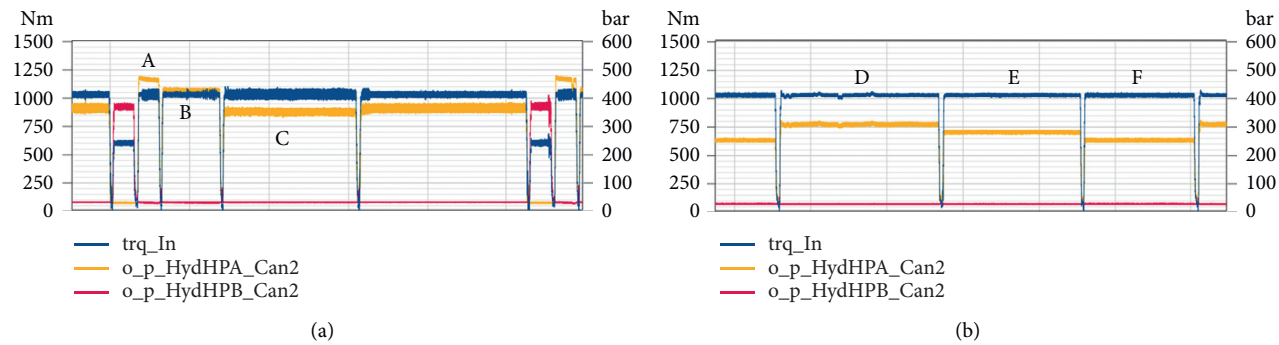


Figure 10: Measurement test program.

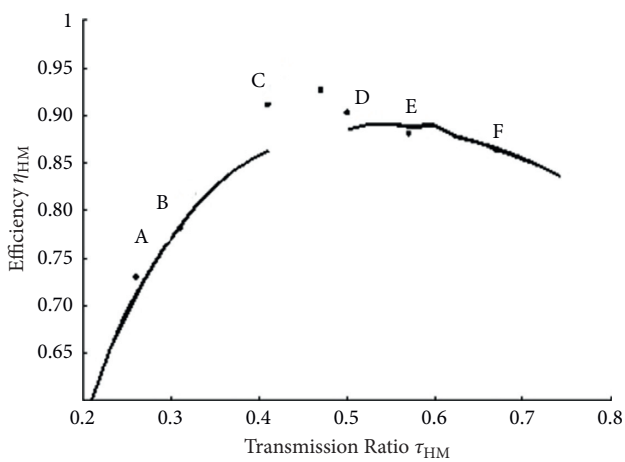


Figure 11: Data on system efficiency.

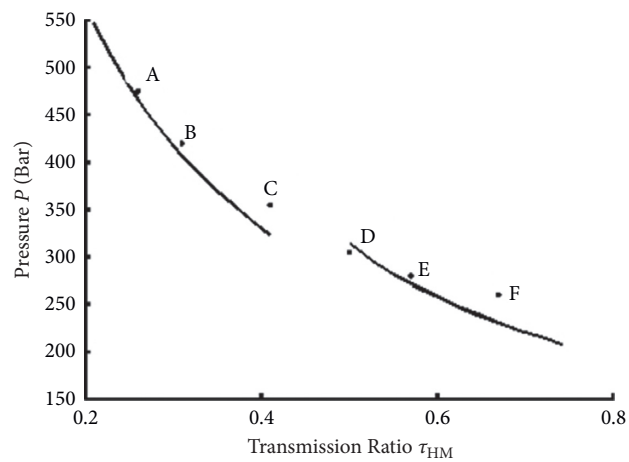


Figure 12: Data on system pressure.

in Figure 12. In both figures, the star point denotes the test data of the bench under six working conditions, and the solid line is the numerical calculation result of the 22 point positions. The maximum point of the simulation efficiency error is at point C of the minimum displacement ratio of the type I power flow, and the error is 4.7%. The error for the other operating conditions is less than 2%. Working conditions C and D are at the minimum displacement ratio, and the torque of hydraulic units 1 and 2 is small. The efficiency simulation error is large due to the relatively large simulation torque control error and the relatively large influence of the nonlinear fitting effect of the HST hydraulic system. The efficiency simulation values of the operating conditions, except for operating conditions A, C, and D, are greater than the actual measured values, because the simulation does not consider transmission churning loss and bearing loss. In Figure 12, the simulated pressure curve effectively reflects the working characteristics of the hydraulic system. As the transmission ratio increases, the HST system pressure gradually decreases with a hyperbolic curve, and the system works in type I and III power flow states. The high-pressure side of the system shows no changes. The simulated pressure curve is lower than the actual measured value, and the transmission power of the HST system is lower than that in the actual working conditions, which is the main reason for the high efficiency of the simulation.

7. Conclusions

Two parameters, namely, hydraulic transmission asymmetry λ and maximum angular velocity ratio $\tau_{HM}^{\max}$, are introduced to express the structural characteristics of hydromechanical continuously variable transmission (HMCVT) with a variable speed range of 0 to the maximum. This approach provides a means to solve the classification problem of HMCVT.

The relationship between mechanical system transmission and the parameters of the hydraulic system is established by the mechanical system demanding surface. The theory of meshing power is applied to determine the torque ratio coefficient Δ_i of an epicyclic gear train (EGT) with two DOFs. The expression of HMCVT overall efficiency is simplified and unified by torque ratio coefficient Δ_i in the steady state under the rated conditions of the diesel engine.

The control theory of parameters e' and e'' is revealed to obtain the maximum transmission efficiency of HMCVT. The application method of this theory is introduced through numerical examples, so that hydraulic system components can be selected properly, and the output torque can be predicted.

The physical prototype bench test conducted using TA1-02 multisection univariate-unit HMCVT and the simulation show that the maximum efficiency simulation error is at point C of the minimum displacement ratio of the type I power flow, and the error is 4.7%. The error for the other operating conditions is less than 2%. The simulated pressure curve effectively reflects the working characteristics of the hydraulic system. As the transmission ratio increases, the HST system pressure gradually decreases in a hyperbolic curve, which verifies the effectiveness of the HMCVT efficiency analysis steps.

For the study of transmission efficiency of future transmissions, the influence of diesel engine working characteristics, overall slip rate, and changes in traction force on transmission efficiency will be comprehensively considered.

Data Availability

The data described in this study could be fully provided by the authors.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the Jiangsu International Science and Technology Cooperation Project (BZ2020061), Jiangsu Provincial Agricultural Science and Technology Independent Innovation Fund (CX (19)3071), and Development Program of Jiangsu Province (BE2019337).

References

[1] J. H. Kress, "Hydrostatic power-splitting transmissions for wheeled vehicles- classification and theory of operation," *SAE International Journal of Engines*, vol. 77, pp. 2282–2306, 1968.

[2] J. Wang, C. Xia, X. Fan, and J. Cai, "Research on transmission characteristics of hydromechanical continuously variable transmission of tractor," *Mathematical Problems in Engineering*, vol. 2020, Article ID 6978329, 14 pages, 2020.

[3] L. Xu, Z. Zhou, M. Zhang, and F. Cao, "Speed ratio matching strategies of hydro-mechanical continuously variable transmission system of tractor," *Journal of China Agricultural University*, vol. 2, no. 4, pp. 94–98, 2006.

[4] H. T. Xue, D. Y. Ding, Z. M. Zhang, M. Wu, and H. Q. Wang, "A fuzzy system of operation safety assessment using multi-model linkage and multi-stage collaboration for in-wheel motor," *IEEE Transactions on Fuzzy Systems*, 2021.

[5] K. T. Renius, "Trends in tractor design with particular reference to europe," *Journal of Agricultural Engineering Research*, vol. 57, no. 1, pp. 3–22, 2013.

[6] H. T Xue, M. Wu, Z. M. Zhang, and H. Q. Wang, "Intelligent diagnosis of mechanical faults of in-wheel motor based on improved artificial hydrocarbon networks," *ISA Transaction*, 2021.

[7] M. Xiao, J. Zhao, Y. Wang, F. Yang, J. Kang, and H. Zhang, "Research on system identification based on hydraulic pump-motor of HMCVT," *Engineering in Agriculture, Environment and Food*, vol. 12, no. 4, pp. 420–426, 2019.

[8] W. Wei, R. Cong, T. Xue, A. D. Abraham, and C. Yang, "Surface roughness and chip morphology of wood-plastic composites manufactured via high-speed milling," *Bioresources*, vol. 16, no. 3, pp. 5733–5745, 2021.

[9] G. M. Wang, S. H. Zhu, S. H. Wang, L. Shi, X. D. Ni, and D. Ouyang, "Speed ratio control of tractor hydraulic mechanical continuously variable transmission," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 29, no. 7, pp. 17–23, 2013.

[10] Z. Cheng, Z. Lu, and F. Dai, "Research on HMCVT efficiency model based on the improved SA algorithm," *Mathematical Problems in Engineering*, vol. 2019, Article ID 2856908, 10 pages, 2019.

[11] J. Y. Li, Q. C. Hu, C. F. Zong, and T. J. Zhu, "Power analysis and efficiency calculation of multistage micro-planetary transmission," *Energy Procedia*, vol. 141, pp. 654–659, 2017.

[12] J. Y. Li and Q. C. Hu, "Power analysis and efficiency calculation of the complex and closed planetary gears transmission," *Energy Procedia*, vol. 100, pp. 423–433, 2016.

[13] M. Awadallah, P. Tawadros, P. Walker, and N. Zhang, "Dynamic modelling and simulation of a manual transmission based mild hybrid vehicle," *Mechanism and Machine Theory*, vol. 112, pp. 218–239, 2017.

[14] E. Pennestri and F. Freudenstein, "The mechanical efficiency of epicyclic gear trains," *Journal of Mechanical Design*, vol. 115, no. 3, pp. 645–651, 1993.

[15] G. Mantriota and E. Pennestrì, "Theoretical and experimental efficiency analysis of multi-degrees-of -freedom epicyclic gear trains," *Multibody System Dynamics*, vol. 9, no. 4, pp. 389–408, 2003.

[16] J. M. del Castillo, "The analytical expression of the efficiency of planetary gear trains," *Mechanism and Machine Theory*, vol. 37, no. 2, pp. 197–214, 2002.

[17] F. Yang, J. Feng, and H. Zhang, "Power flow and efficiency analysis of multi-flow planetary gear trains," *Mechanism and Machine Theory*, vol. 92, pp. 86–99, 2015.

[18] W. H. Wei, Y. L. Li, and Y. T. Li, Y. Li, C. Yang, Research on tool wear factors for milling wood-plastic composites based on response surface methodology," *Bioresources*, vol. 16, no. 1, pp. 151–162, 2021.

[19] X. M. Xu and P. Lin, "Parameter identification of sound absorption model of porous materials based on modified particle swarm optimization algorithm," *PLOS ONE*, vol. 16, no. 5, pp. 1–16, 2021.

[20] R. H. Macmillan, "Power flow and loss in differential mechanisms," *Journal of Mechanical Engineering Science*, vol. 3, no. 1, pp. 37–41, 1961.

[21] C. Chen and T. T. Liang, "Theoretic study of efficiency of two-dof of epicyclic gear transmission via virtual power," *Journal of Mechanical Design*, vol. 133, no. 3, pp. 1–7, 2011.

[22] L. Mangialardi and G. Mantriota, "Power flows and efficiency in infinitely variable transmissions," *Mechanism and Machine Theory*, vol. 34, no. 7, pp. 973–994, 1999.

[23] G. Mantriota, "Performances of a series infinitely variable transmission with type I power flow," *Mechanism and Machine Theory*, vol. 37, no. 6, pp. 579–597, 2002.

[24] G. Mantriota, "Performances of a parallel infinitely variable transmissions with a type II power flow," *Mechanism and Machine Theory*, vol. 37, no. 6, pp. 555–578, 2002.

[25] C. Chen and J. Angeles, "Virtual-power flow and mechanical gear-mesh power losses of epicyclic gear trains," *Journal of Mechanical Design*, vol. 129, no. 1, pp. 107–113, 2007.

[26] Y. I. Xiao, S. J. Jiao, Z. F. Liu, T. Q. Zhang, and S. G. Long, "Research on key technical parameters of full hydraulic bulldozer," *China Journal of Highway and Transport*, vol. 17, no. 23, pp. 119–123, 2004.

Hindawi
Shock and Vibration
Volume 2021, Article ID 9544809, 11 pages
https://doi.org/10.1155/2021/9544809

Research Article

Bearing Defect Detection with Unsupervised Neural Networks

Jianqiao Xu,[1] Zhaolu Zuo,[2] Danchao Wu,[3] Bing Li,[3] Xiaoni Li,[4] and Deyi Kong[2,5]

[1]*Department of Information Security, Naval University of Engineering, Wuhan 430033, China*
[2]*Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China*
[3]*Hefei Xiaobu Intelligent Technology Co., Ltd., Hefei 230011, China*
[4]*Shaanxi Aerospace Times Navigation Equipment Co., Ltd., Baoji 721000, China*
[5]*Innovation Academy for Seed Design, CAS, Beijing 10000, China*

Correspondence should be addressed to Zhaolu Zuo; zuozl@iim.ac.cn

Received 16 June 2021; Revised 9 July 2021; Accepted 19 July 2021; Published 20 August 2021

Academic Editor: Jun Zhu

Bearings always suffer from surface defects, such as scratches, black spots, and pits. Those surface defects have great effects on the quality and service life of bearings. Therefore, the defect detection of the bearing has always been the focus of the bearing quality control. Deep learning has been successfully applied to the objection detection due to its excellent performance. However, it is difficult to realize automatic detection of bearing surface defects based on data-driven-based deep learning due to few samples data of bearing defects on the actual production line. Sample preprocessing algorithm based on normalized sample symmetry of bearing is adopted to greatly increase the number of samples. Two different convolutional neural networks, supervised networks and unsupervised networks, are tested separately for the bearing defect detection. The first experiment adopts the supervised networks, and ResNet neural networks are selected as the supervised networks in this experiment. The experiment result shows that the AUC of the model is 0.8567, which is low for the actual use. Also, the positive and negative samples should be labelled manually. To improve the AUC of the model and the flexibility of the samples labelling, a new unsupervised neural network based on autoencoder networks is proposed. Gradients of the unlabeled data are used as labels, and autoencoder networks are created with U-net to predict the output. In the second experiment, positive samples of the supervised experiment are used as the training set. The experiment of the unsupervised neural networks shows that the AUC of the model is 0.9721. In this experiment, the AUC is higher than the first experiment, but the positive samples must be selected. To overcome this shortage, the dataset of the third experiment is the same as the supervised experiment, where all the positive and negative samples are mixed together, which means that there is no need to label the samples. This experiment shows that the AUC of the model is 0.9623. Although the AUC is slightly lower than that of the second experiment, the AUC is high enough for actual use. The experiment results demonstrate the feasibility and superiority of the proposed unsupervised networks.

1. Introduction

With the continuous development and progress of manufacturing industry, the demand for bearings is increasing as a basic component widely used. Performance and life of the machine itself often have a great relationship with the quality of the bearings [1], so the requirements for the quality of the bearings in industrial production continue to increase. In the process of manufacturing and assembly of bearings, defects on the bearing surface are often caused by various reasons. Common defects include pull marks, dark spots, pits, scratches, rust, and yellow spots. These surface defects will cause the corrosion resistance, elasticity, wear resistance, and lubricity of the bearing to decrease, resulting in a greatly reduced service life of the machine, and even serious safety accidents. Therefore, it is essential to detect the defects of the bearing.

For the detection of bearing surface defects, there are methods such as manual inspection, physical inspection, and machine vision inspection [2]. At this stage, the most important method is manual detection. However, manual inspection is very subjective, and it is often determined by the experience of the inspection operators based on their practice, which is time-consuming and labor-intensive. In

addition, when the operation is performed under continuous light, the inspecting staff are prone to misdetection or missed inspection due to visual fatigue, and it will cause serious harm to the health of the inspector. The common methods of testing in physics are eddy current testing, ultrasonic testing, magnetic particle testing, and so on. These physics-oriented inspection methods are widely used to detect the defects of bearing rollers, but this type of inspection method also has its own shortcomings; that is, it also requires operators to determine the defects of the bearing, but the inspection is not accurate. If the performance is still too low, it will cause missed detection or false detection.

With the continuous development and progress of modern science and technology, when we need to detect defects, machine vision begins to be more and more used. Ye and Hsu designed a new lighting system to collect images in a darkroom, avoiding the influence of external factors and light sources, and developed a rule-based local mask sensor algorithm to achieve high-precision detection of metal defects [3]. Shen et al. designed a new type of lighting and image acquisition system. By taking three photos of the bearing, the left and right photos are used to detect the deformation on the sealing ring, and other defects are detected by the central illumination image to correct the deformation on the sealing ring. Defects have high accuracy and efficiency [4]. Tao proposed a multi-threshold segmentation image based on OSTU to quickly detect defects on the bearing surface. After denoising the collected images, use OSTU to perform threshold segmentation to obtain two thresholds before detecting and locating defects [5].

Traditional surface detection algorithms obtain detected images through image preprocessing and then use statistical machine learning methods to extract image features to achieve the goal of defect detection. These algorithms have achieved good results in some specific applications, but there are still many shortcomings. For example, there are many image preprocessing steps and strong pertinence, with poor robustness; a variety of algorithms have an amazing amount of calculation and cannot accurately detect the size and shape of defects. Deep learning directly updates parameters through learning data, avoids manual design of complex algorithm processes, and has extremely high robustness and accuracy. Zhao et al. [6] proposed a new defect detection framework based on positive sample training, which combines GAN and autoencoder to reconstruct defect image, and LBP is used for image local contrast to detect defects. Wen et al. [7] proposed a multitask convolutional neural network to detect defects. Instead of using a large convolution kernel, a smaller convolution kernel is used to convolve the input data, and the shared neural network is used to classify and locate the defects after extracting the defect features of the sample data. Cha et al. [8] used a sliding window-based convolutional neural classification network to realize the location of crack surface defects, and the combination of two sliding window redundant paths to achieve full image coverage. Wang et al. [9] used a deep convolutional neural network to classify samples of defects when detecting defects in cloth and then detect defects after classification. Chen et al. [10] use DCNNs combined with SSD, Yolo, and other network methods to build a cascaded detection network from coarse to fine, including firmware positioning, defect detection, and classification. DCNNs have good robustness and adaptability, which means that this method has a good application prospect in the defect detection and classification of fasteners. Mei et al. [11, 12] adopt the idea of image pyramid hierarchy and convolutional denoising autoencoder network to realize defect detection of cloth texture images. The results show that full use of unsupervised learning and multimodal result fusion strategy can improve the robustness and accuracy of defect detection. Bergmann et al. [13] propose an improving unsupervised defect segmentation by applying structural similarity to autoencoders, and the proposed method achieves significant performance gains on a challenging real-world dataset of nanofibrous materials. Yang et al. [14] propose an end-to-end surface quality detection method based on deep convolutional neural networks (CNNs) to improve the accuracy and efficiency of VDR surface quality detection. Essid et al. [15] develop a new machine vision framework for efficient detection and classification of manufacturing defects in metal boxes. The results show that the proposed autoencoder deep neural network (DNN) architecture can not only classify manufacturing defects, but also localize them with high accuracy. Wu et al. [16] propose a high-sensitivity magnetic flux leakage method based on magnetic induction head for the detection of tiny cracks in bearing rings. Xu et al. [17] propose a new multidefect detection method based on a combination of an improved visual attention model and image partitioning-weighted eigenvalue for surface defects of explosive cartridge in the automatic sorting process that are of small area, irregular shape, and random distribution. Kong et al. [18] propose a unified framework for detecting defects in planar industrial products or planar surfaces of nonplanar products based on a template-matching strategy. Tao et al. [19] propose an algorithm for pixel-level segmentation and classification of defects. The entire network can be divided into two stages: defect detection stage and defect classification stage. Fang et al. [20] propose an SLIC head of object instance segmentation in proposal regions (Mask R-CNN) containing a network block to learn the quality of the predict masks. Park et al. [21] propose a convolutional neural network (CNN) based method that inspects nonpatterned welding defects (craters, pores, foreign substances, and fissures) on the surface of the engine transmission using a single RGB camera. Ming et al. [22] propose a combined classifier with dynamic weights (CCDW) to classify the LPG samples considering both feature extraction diversity and base classifiers diversity after image segmentation and enhancement. Martínez et al. [23] propose a machine vision system, performing the detection of flaws on textured surfaces, and multiple images under different lighting conditions are processed and merged into one, which is used to extract features with a supervised classifier. Peng et al. [24] propose a precision measurement and inspection of O-rings with good accuracy and efficiency.

This research is to use the deep neural network to realize the defect detection of the bearing. The main content of this work focuses on the following topics: (1) how to increase the number of samples, (2) how to improve the AUC of the model, and (3) how to enhance the feasibility of the method. The organization of this paper is as follows. Section 2 describes the defect representation and data acquisition system, and Section 3 introduces the methodology. Experiment and results are illustrated in Section 4, and Section 5 gives some discussion. Finally, Section 6 summarizes this paper.

2. Defect Representation and Data Acquisition

2.1. Data Acquisition System. The data acquisition system is composed of cameras, lighting systems, and computers, as shown in Figure 1. The image capture device can capture images of the inner end surface, outer diameter, inner diameter, and lower end surface separately. Basler industrial camera as A1300-60gm with resolution of 1282×1026 pixels is selected, and the lens is PCHI012. Different field of view sizes can be obtained by adjusting the focal length, so as to match the inner diameter, outer diameter, upper end surface, and lower end surface size. By adjusting the exposure time to obtain the largest signal-to-noise ratio, the light source is uniformly illuminated by the ring LED (the light source model is HZN DRL-70-60-W). The final images obtained are shown in Figure 2.

2.2. Defect Representation. Bearing defects mainly include the following types: outer diameter defects (stretch marks, dark spots, pits, scratches, rust, and yellow spots); lower end surface defects (dents, convex deformation, scratches, and embroidery); inner diameter defects (dimples, scratches, and embroidery); inner end surface defects (dents, convex deformation, rust, and yellow spots). There are many types of defects, and the characteristics of defects are not obvious, as shown in Figure 3.

3. Methodology

Carefully observe the samples obtained by the above-mentioned devices, and you can find that, in addition to useful information, there is some useless redundant information in the samples. In order to ensure the accuracy of detection, a series of pretreatments are required on the samples. Although the defects of the inner end surface, inner diameter, outer diameter, and lower end surface are different, their distributions are similar. They are all distributed along the circumference of the bearing, but the position is different. Therefore, this article selects the inner diameter sample with more complicated appearance and more interference factors. Processing: samples from other parts can be processed in the same way.

3.1. Normalized Sample Method. Since the bearing is taken on the liner, in addition to the bearing, images of other parts are also taken. To solve this problem, we first find the contours of the outer and inner edges and perform ellipse fitting on the contours. Then, based on the center position of the fitted ellipse, move the bearing to the center of the image, and use perspective transformation to transform the ellipse into a circle based on the parameters of the ellipse. Finally, remove all the parts outside the outer edge and inside the inner edge after the transformation. The captured bearing image and the processing algorithm schematic are shown in Figure 4.

3.2. Sample Split Based on Normalized Sample Symmetry. After the sample is normalized, the inner diameter part of the bearing is converted into a standard ring, which satisfies the characteristics of stacking based on the center of the image. Since the defect part is generally very small and only occupies a small part of the ring, the symmetry can be used to split the sample into a large number of fan-shaped rings, as shown in Figure 5. The 12 samples obtained will be labelled, and the classifier will be trained based on the divided samples.

3.3. Supervised Neural Networks Using ResNet Neural Networks. Deep convolutional neural networks have already shined in image classification problems. Recent studies have also shown that the depth of the network plays a crucial role in accuracy. However, as the network deepens, there is a problem worth noting. As the network continues to stack and deepen, will the effect of the network always get better and better? Obviously, you will encounter the problem of gradient disappearance or gradient explosion, and this problem can already be solved by normalizing the input during initialization, but when the network finally converges, there will be a "degradation" problem, resulting in a decrease in accuracy (not overfitting), so although the number of network layers can be continuously stacked to allow it to train and converge, there is still no way to encounter degradation problems [25].

He et al. [25, 26] build a new network structure (ResNet) to solve the above problem that when the number of network layers is too high, the effect of the deeper network is not as good as the shallower network, and a proper explanation is made. ResNet uses the input of one layer and the output of another layer as the output of a block. Assuming that x is the input of a block, and one block is composed of two layers, then he first passes through a convolutional layer and activates relu to obtain $F(x)$, and then the result of $F(x)$ after the convolutional layer is added to the previous input x to obtain a result, and the result is activated by relu as the output of the block. For ordinary convolutional networks, we output $F(x)$, but in ResNet, we output $H(x) = F(x) + x$, but we still use $F(x) = H(x) - x$. This changed the learning goal, changing the original learning to make the objective function equal to a known constant value to make the residual between the output and the input 0, which is the identity mapping. The result is that after the residual is introduced, the output is mapped to the output. The changes are more sensitive.

Based on the samples obtained with Sections 3.1 and 3.2, supervised neural networks can be trained with ResNet neural networks as the following process, as shown in

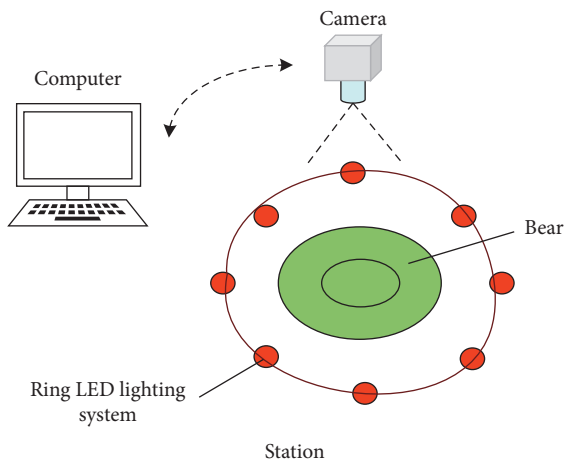


Figure 1: Schematic of measurement device.

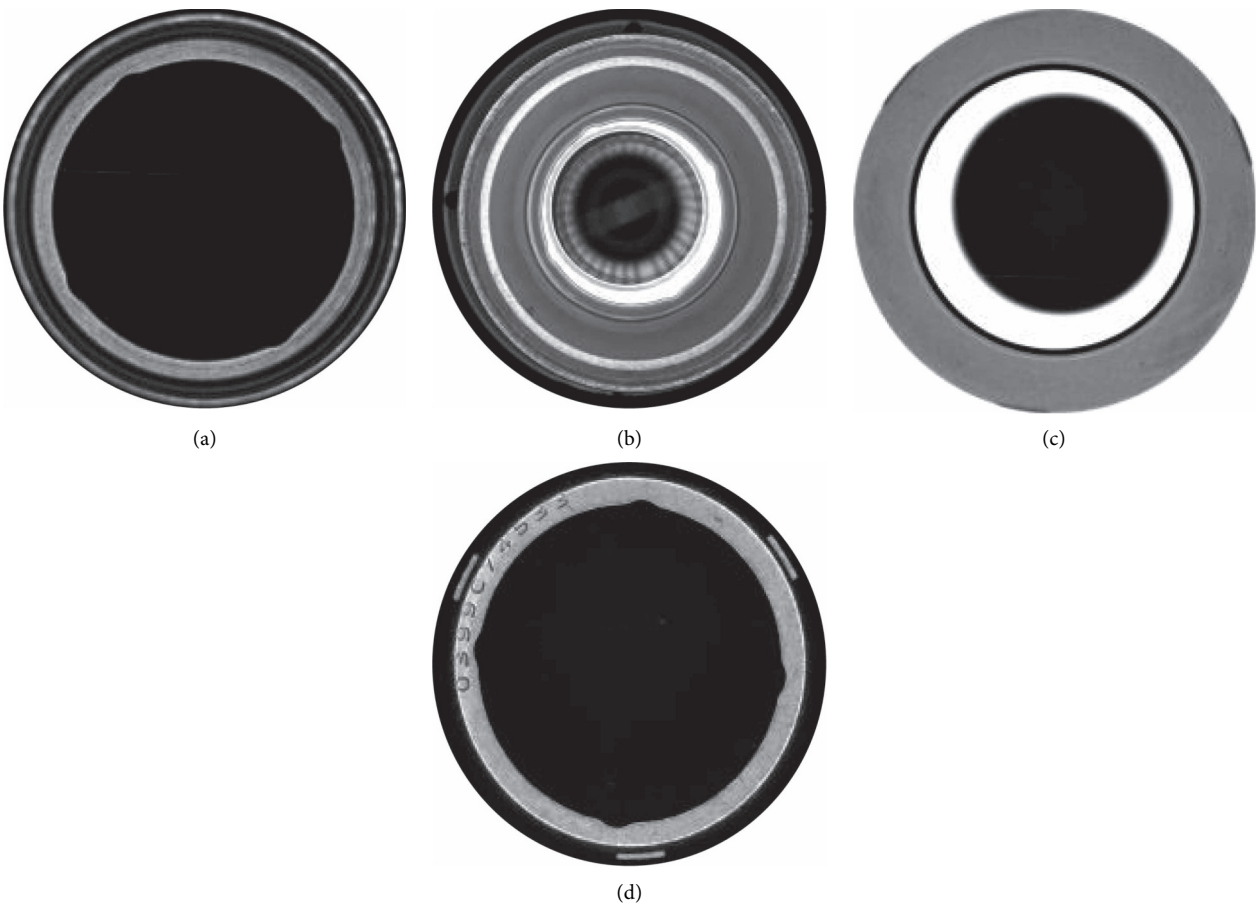


Figure 2: Bearing images of different parts. (a) Inner end surface. (b) Outer diameter. (c) Inner diameter. (d) Lower end surface.

Figure 6. Also, more details can be found in our previous work [27].

3.4. Autoencoder Neural Networks Implemented with U-Net. In the field of image generation, there is a very important network structure called Autoencoder [28]. An autoencoder neural network architecture is a feedforward network composed of one or multiple connected hidden layers. It uses a nonlinear mapping function between the original data as input and output specific learned features. The feature of autoencoder is that the first half is the downsampling part, which is generally implemented by CNN; the second half is the upsampling part, which is generally implemented by

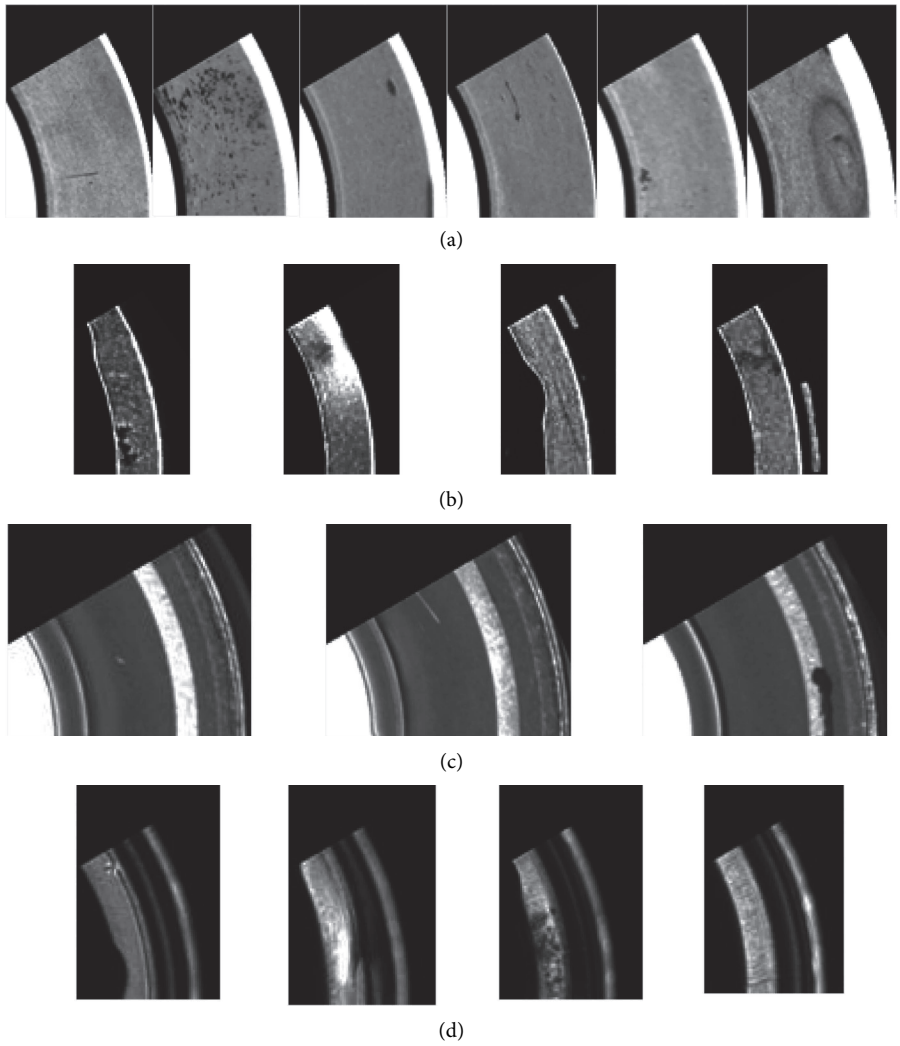


Figure 3: Defect classification. (a) Defects of outer diameter. (b) Defects of lower end surface. (c) Defects of inner diameter. (d) Defects of inner end surface.

inverse convolution. The most amazing thing about the entire autoencoder is that even if we only have the features of the middle layer, we can recover a picture that is very close to the original picture through the second half. Therefore, the entire autoencoder has at least two attractive applications: (1) use the first half for feature extraction; (2) use the second half for image generation.

U-Net itself is not used for autoencoder; it first appeared in the segmentation of medical images [29]. On the one hand, its structure is very similar to the traditional structure of autoencoder. On the other hand, its unique feedforward structure allows the network to capture a lot of spatial information. So recently, a lot of image synthesis and generation work are based on U-Net. In this paper, U-Net is used to extract feature map from the original image firstly, and then feature map is used to generate gradient image.

3.5. The Proposed Unsupervised Neural Network. Lighting attenuation or batches will affect the classification effect of the supervised network; therefore, an unsupervised neural network is proposed to solve the disturbing factors, as shown in Figure 7. Based on the samples obtained with Sections 3.1 and 3.2, the proposed unsupervised neural networks can be trained with AE neural networks implemented with U-Net as the following process.

Step 1: raw bearing samples are normalized using Algorithm 1 in Section 3.1

Step 2: normalized samples are split based on normalized sample symmetry using Algorithm 2 in Section 3.2

Step 3: the gradient of the samples is extracted as label data, and Sobel operator is selected to calculate the gradient of the samples

Step 4: AE neural networks implemented with U-Net are used to predict the gradient of the samples

Step 5: the loss function is defined with the argmax of the difference between the label data and the predict data

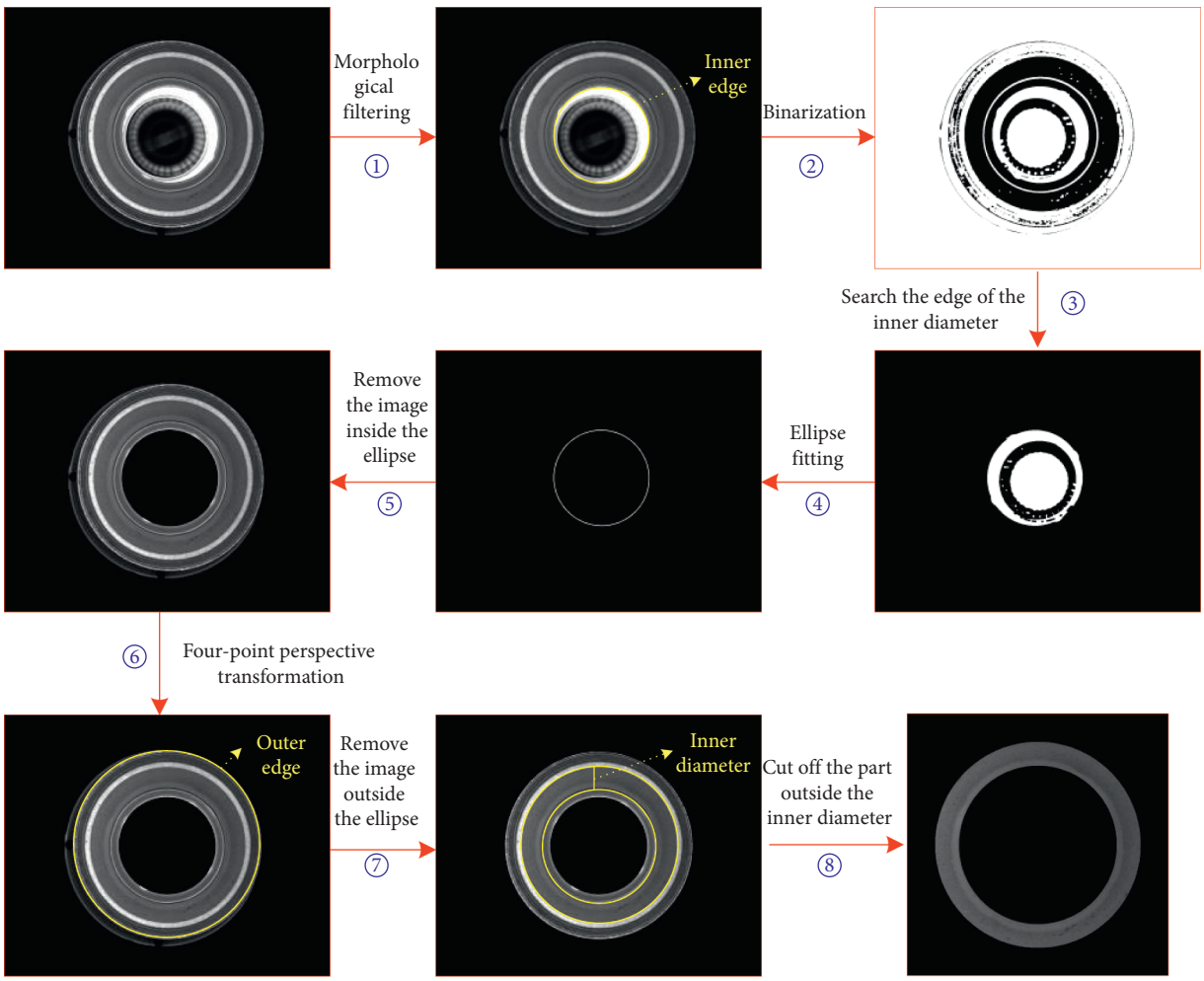


Figure 4: Normalization algorithm for bearing inner diameter samples.

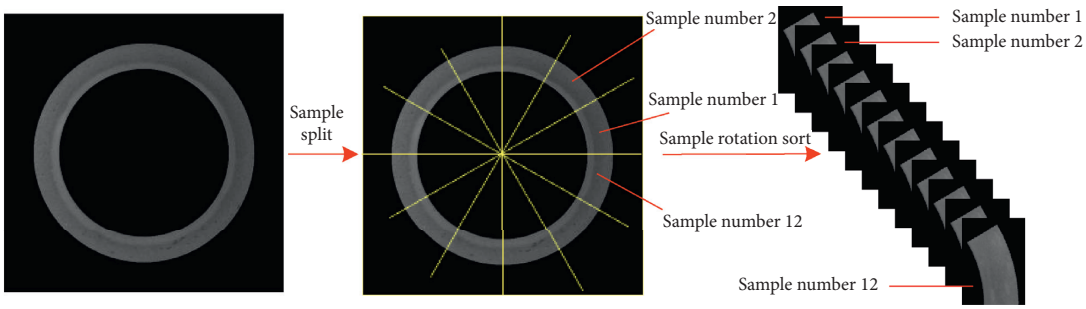


Figure 5: Sample splitting algorithm based on symmetry.

Step 6: new data can be updated to online train and online modify the model

4. Experiment and Results

The image processing algorithm in this article is trained and tested on the server. The server's processor is Intel(R) Xeon(R) CPU E5-2678v3@2.5 GHz, the graphics card is 2 GeForce GTX 1080 Ti from NVIDIA, and the deep learning architecture uses TensorFlow.

4.1. Model Training Method. Different datasets are made for different workstation training, and different defect classifiers are trained through the datasets of different workstations. This article selects the inner diameter sample as an example of the algorithm display. Three experiments are conducted.

The first experiment is the supervised neural networks using ResNet neural networks. The bearing inner diameter samples are divided into training set, validation set, and test set with numbers 16760, 2490, and 2076 separately. The numbers of positive samples and negative samples of the

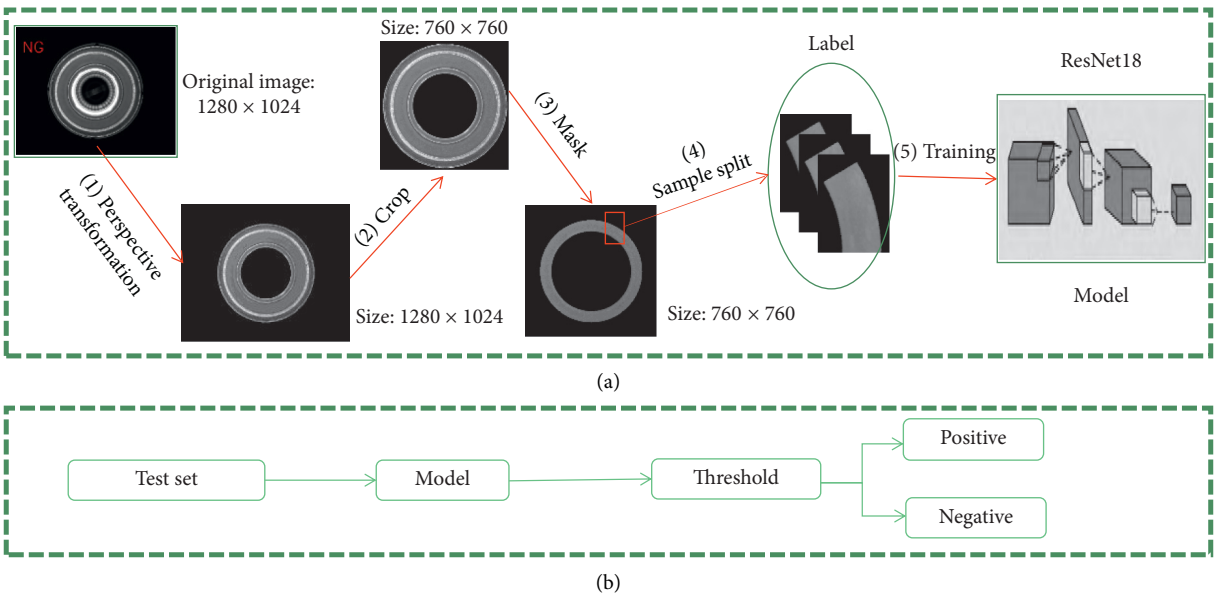


Figure 6: Classifier training with supervised neural networks. (a) Train stage. (b) Test stage.

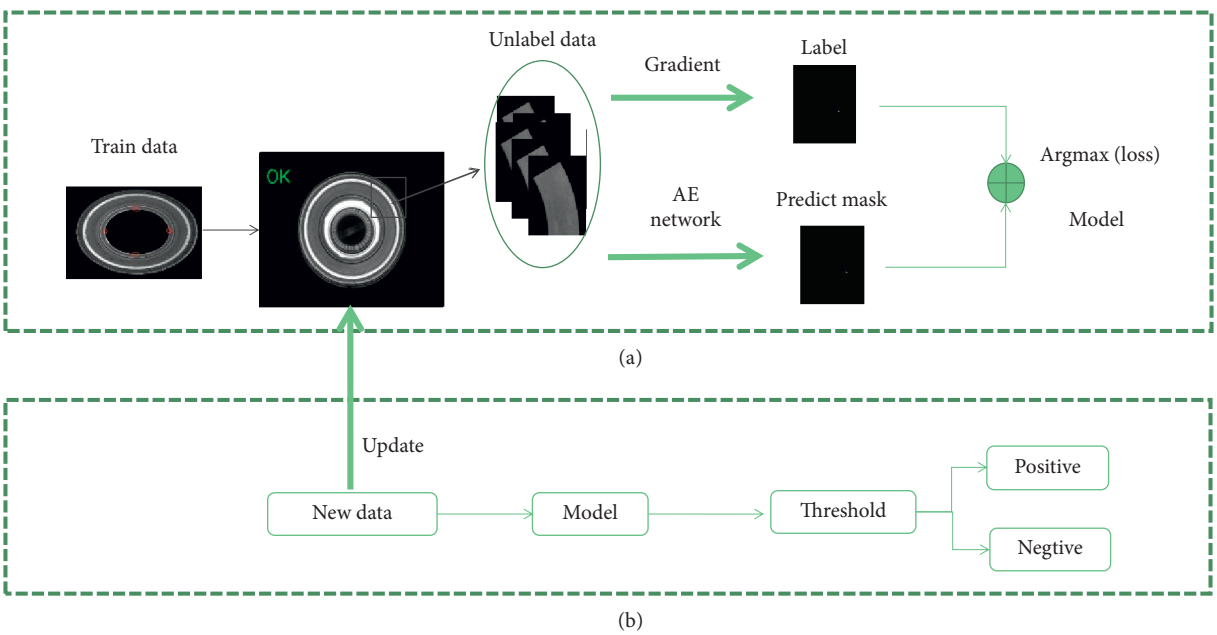


Figure 7: Classifier training with the proposed unsupervised neural network. (a) Train stage. (b) Test stage.

training set are 13440 and 3320, respectively. The numbers of positive samples and negative samples of the validation set are 2076 and 414, respectively.

The second experiment is the proposed unsupervised neural networks with AE neural networks. In this experiment, all the negative samples in training set and validation set are discarded. The bearing inner diameter samples are divided into training set, validation set, and test set with numbers 13440, 2076, and 2076 separately. The numbers of training set and validation set in this experiment are less than those of the first experiment. No negative samples are contained in this experiment.

The third experiment is also the proposed unsupervised neural networks with AE neural networks. The difference between this experiment and the second experiment is the training set and validation set. The samples sets are the same with the first experiment, but without any labels. All the positive samples and negative samples in training set and validation set are integrated together, respectively. The numbers of training set and validation set are also as 16760 and 2490 separately.

In order to evaluate the model trained with supervised neural network and the proposed unsupervised neural networks effectively, all the experiments share the same test

Input: inner diameter sample with 1280×1024 pixels.
Output: normalized samples of inner diameter sample with 760×760 pixels.
(1) Morphological denoising: the original image is corroded and expanded, and the template is a $5 * 5$ rectangular morphological structural element;
(2) Binarize the original image, take the maximum gray value and minimum gray value of the inner diameter area as the threshold, set the image greater than the maximum threshold and less than the minimum threshold to 255, and the inner diameter area becomes 0;
(3) Search the inner edge contour, and then fit the inner edge with an ellipse;
(4) Use the ellipse fitted in step 3 to remove the extra part of the image;
(5) Search the four points at the top, bottom, left, and right of the inner edge;
(6) Map the above ellipse to a circle. Take the four points of the top, bottom, left, and right of the circle with the center of the image as the center and the radius of 290 as the target points to establish a projection transformation mapping matrix, and then use this transformation matrix to transform the image in step 4;
(7) Search for the outer edge contour, fit the ellipse, and cut off the outside of the ellipse;
(8) Search the area of the inner diameter, and cut off the outer part of the inner diameter area.

Algorithm 1: A normalized samples method to obtain the effective part of inner diameter sample from the raw bearing samples.

Input: normalized samples of inner diameter sample with 760×760 pixels.
Output: 12 shares samples along the center of the circle with labels from 1 to 12.
(1) Divide the inner diameter sample into 12 shares evenly along the center of the circle;
(2) Label the 12 shares with numbers 1–12
(3) Rotate the samples 2–12 by a certain angle to the position of sample number 1.

Algorithm 2: Sample split based on normalized sample symmetry.

Table 1: Classification confusion matrix.

Ground truth	Predictive value	
	Positive	Negative
True	TP	FN
False	FP	TN

set. The numbers of positive samples and negative samples of the test set are 1980 and 96, respectively.

4.2. Model Evaluation Method. Generally, the parameters of the classification confusion matrix in the following table are used for statistical calculation. Table 1 shows the classification confusion matrix.

In this paper, the accuracy rate ACC, accuracy rate P, and recall rate R of the training model on the black box set are used to evaluate the pros and cons of the model. The accuracy rate ACC is defined as follows: the proportion of the correct result of the classification model to the total observation sample, that is, the proportion of all the predicted results that is correctly predicted. The accuracy rate P is defined as follows: among the samples that are identified as positive samples, the model predicts the correct proportion. From the perspective of prediction, one type of prediction result is taken out to evaluate the prediction accuracy rate. The recall rate R is defined as the ratio of correctly identified samples in all positive categories, reflecting the sensitivity of the model.

The accurate rate, accuracy rate, and recall rate are defined as

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}},$$
$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{1}$$
$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

The accuracy rate can better represent the accuracy of the model. Accuracy and recall rate are better performance evaluation indicators than correct rate, which is an evaluation of a certain category. Accuracy and recall are a pair of contradictory measures. Generally speaking, when the accuracy is high, the recall is often low; when the recall is high, the accuracy is often low.

Another more comprehensive evaluation index is receiver operating characteristic (ROC) curve. The ROC curve is used to describe the performance of the two classification systems (the threshold of the classifier is

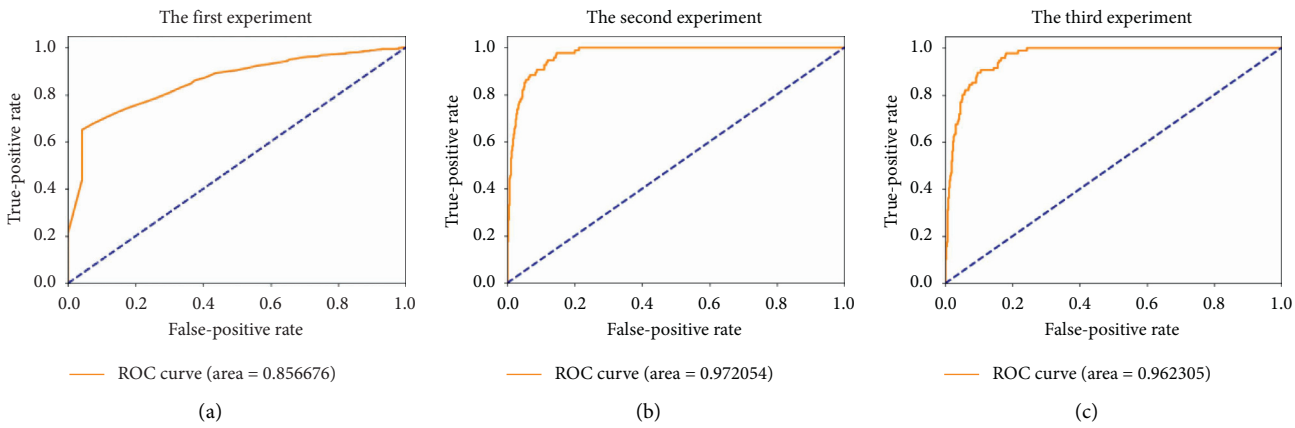


Figure 8: ROC curves of different models. The ROC of (a) the supervised neural networks (the first experiment), (b) the unsupervised neural networks (the second experiment), and (c) the unsupervised neural networks (the third experiment).

Table 2: Statistics of the above indicators for the supervised neural networks and the unsupervised neural networks.

	TP	FN	FP	TN	ACC	P	R	AUC
The first experiment	1980	0	96	0	0.9538	0.9538	1	0.8567
The second experiment	1980	0	25	71	0.9879	0.9875	1	0.9721
The third experiment	1980	0	47	49	0.9774	0.9768	1	0.9623

variable), a comprehensive index of continuous changes in response sensitivity and specificity, and the points on the ROC curve reflect the susceptibility of the same signal stimulus. ROC curve and AUC are indicators to evaluate the pros and cons of the two-class model as a whole, where AUC is the area between the ROC curve and its horizontal axis. The ROC curve is generally above $y = x$. The larger the AUC value, the better the model. The ROC curve is drawn by two indicators, the true-positive rate (TPR) and the false-positive rate (FPR). The true-positive rate (TPR) is defined as follows: the true label is the proportion of the positive sample, in which the prediction is also the positive sample. The false-positive rate (FPR) is defined as the proportion of positive samples, whose true labels are negative.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}. \tag{2}$$

4.3. Results. Train the three experimental models and test them on the same test set, draw the ROC curve, and calculate the AUC, as shown in Figures 8(a)~8(c). It is easy to find that the model of Figure 8(b) has the best performance, while Figure 8(a) has the worst.

Statistics of the above indicators are shown in Table 2. The R indicator of all the three networks is 100%. From the ACC, P and AUC indicators, the unsupervised networks have better performance than supervised network. The AUC of the three models is 0.8567, 0.9721, and 0.9623 separately. Though the indicators of the third model are slightly less than those of the second model, the third model is still good enough for actual use. What is more, the third model is totally an unsupervised model, which is very convenient in actual use and can update the model online.

5. Discussion

Some experiments about the supervised neural networks with ResNet networks and unsupervised neural networks with AE networks for bearing defect detection have been carried out in Section 4. According to the results, some points should be discussed further:

(1) Why does the unsupervised network have better performance than the supervised network? We think the supervised network can have good performance if the defect characteristics are obvious. However, the defects of the bearing are very small and very inconspicuous. The unsupervised networks are good at identifying small defects. Thus, the unsupervised network has better performance.

(2) Training process: in experiment 2, the unsupervised networks are trained with positive samples, which have the best performance; however, the samples have to be selected manually. In experiment 3, the unsupervised networks are trained with positive samples and negative samples; that is to say, the process of selecting samples is not necessary, which will be of great convenience for industrial site processing.

(3) Automatic networks update process: the environment of the industrial site may change over time; in this condition, the networks should be updated

automatically for good performance of the networks. The proposed networks can update the networks with the update samples.

6. Conclusions

This paper proposes new unsupervised neural networks based on AE networks for bearing defect detection. Sample preprocessing algorithm based on normalized sample symmetry of bearing is adopted to greatly increase the number of samples. Gradients of the unlabeled data are used as labels, and AE networks are created with U-net to predict the output. Three experiments, one with supervised network and the other two with the unsupervised network, are conducted. The AUC of the three models is 0.8567, 0.9721, and 0.9623 separately. Though the indicators of the third model are slightly less than those of the second model, the third model is still good enough for actual use. What is more, the third model is totally an unsupervised model, which is very convenient in actual use and can update the model online. The experiment results demonstrate the feasibility and superiority of the proposed unsupervised networks. It can be expected that, with the widespread application of visual inspection systems in bearing automation production lines, the proposed method can greatly improve production efficiency and make a certain contribution to the improvement of bearing production quality.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

References

[1] J. Xingxing, S. Changqing, S. Juanjuan, and Z. Zhu, "Initial center frequency-guided VMD for fault diagnosis of rotating machines," *Journal of Sound and Vibration*, vol. 435, no. 24, pp. 36–55, 2018.

[2] K. Wang, *Application of Image Fusion in Detection of Micro Bearing Surface Defect*, Jilin University, Changchun, China, 2011.

[3] J.-H. Ye and Q.-C. Hsu, "Automatic optical apparatus for inspecting bearing assembly defects," *Sensors and Materials*, vol. 30, no. 11, pp. 2637–2652, 2018.

[4] H. Shen, S. Li, D. Gu, and H. Chang, "Bearing defect inspection based on machine vision," *Measurement*, vol. 45, no. 4, pp. 719–733, 2012.

[5] Q. Tao and X. Wu, "Rapid detection method of the bearing surface defects," *Microelectronics & Computer*, vol. 28, no. 10, pp. 98–100, 2011.

[6] Z. Zhao, L. Bo, D. Rong, and P. Zhao, "A surface defect detection method based on positive samples," in *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Springer, Nanjing, China, July 2018.

[7] S. Wen, Z. Chen, and C. Li, "Vision-based surface inspection system for bearing rollers using convolutional neural networks," *Applied Sciences*, vol. 8, no. 12, Article ID 2565, 2018.

[8] Y. J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, 2017.

[9] T. Wang, Y. Chen, M. Qiao, and H. Snoussi, "A fast and robust convolutional neural network-based defect detection model in product quality control," *International Journal of Advanced Manufacturing Technology*, vol. 94, 2017.

[10] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 2, pp. 257–269, 2018.

[11] S. Mei, H. Yang, and Z. Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1266–1277, 2018.

[12] S. Mei, Y. Wang, and G. Wen, "Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model," *Sensors*, vol. 18, no. 4, Article ID 1064, 2018.

[13] P. Bergmann, S. Lwe, M. Fa User et al., "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," in *Proceedings of the 14th International Conference on Computer Vision Theory and Applications*, January 2018.

[14] T. Yang, S. Peng, and L. Huang, "Surface defect detection of voltage-dependent resistors using convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 1–16, 2020.

[15] O. Essid, H. Laga, and C. Samir, "Automatic detection and classification of manufacturing defects in metal boxes using deep neural networks," *PLoS One*, vol. 13, 2018.

[16] J. Wu, Y. Yang, E. Li et al., "A high-sensitivity MFL method for tiny cracks in bearing rings," *IEEE Transactions on Magnetics*, vol. 54, 2018.

[17] L. Xu, H. Xu, X. Li et al., "A defect inspection for explosive cartridge using an improved visual attention and image weighted eigenvalue," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 99, p. 1, 2019.

[18] H. Kong, J. Yang, and Z. Chen, "Accurate and efficient inspection of speckle and scratch defects on surfaces of planar products," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1855–1865, 2017.

[19] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu, "Automatic metallic surface defect detection and recognition with convolutional neural networks," *Applied Sciences*, vol. 8, no. 9, Article ID 1575, 2018.

[20] X. Fang, W. Jie, and T. Feng, "An industrial micro-defect diagnosis system via intelligent segmentation region," *Sensors*, vol. 19, no. 11, Article ID 2636, 2019.

[21] J.-K. Park, W.-H. An, and D.-J. Kang, "Convolutional neural network based surface inspection system for non-patterned welding defects," *International Journal of Precision Engineering and Manufacturing*, vol. 20, no. 3, pp. 363–374, 2019.

[22] W. Ming, F. Shen, H. Zhang et al., "Defect detection of LGP based on combined classifier with dynamic weights," *Measurement*, vol. 143, pp. 211–225, 2019.

[23] S. S. Martínez, C. O. Vázquez, J. G. García, and J. G. Ortega, "Quality inspection of machined metal parts using an image fusions technique," *Measurement*, vol. 111, pp. 374–383, 2017.

[24] G. Peng, Z. Zhang, and W. Li, "Computer vision algorithm for measurement and inspection of O-rings," *Measurement*, vol. 94, pp. 828–836, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision*, pp. 630–645, Amsterdam, The Netherlands, September 2016.

[27] J. Xu, J. Wu, X. Chen, D. Wu, and B. Li, "Bearing defects detection based on standardized sample split," *Journal of Applied Optics*, vol. 42, no. 2, pp. 327–333, 2021.

[28] C. Shen, Y. Qi, J. Wang, G. Cai, and Z. Zhu, "An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 170–184, 2018.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Munich, Germany, November 2015.

Hindawi
Shock and Vibration
Volume 2021, Article ID 9532702, 10 pages
https://doi.org/10.1155/2021/9532702

Research Article

Image Denoising Using Nonlocal Means with Shape-Adaptive Patches and New Weights

Chenglin Zuo, Jun Ma, Hao Xiong, and Lin Ran

Low Speed Aerodynamics Institute, China Aerodynamics Research and Development Center, Mianyang 621000, China

Correspondence should be addressed to Jun Ma; majunttt@sina.com

Received 5 June 2021; Accepted 20 July 2021; Published 27 July 2021

Academic Editor: Jun Zhu

Digital images captured from CMOS/CCD image sensors are prone to noise due to inherent electronic fluctuations and low photon count. To efficiently reduce the noise in the image, a novel image denoising strategy is proposed, which exploits both nonlocal self-similarity and local shape adaptation. With wavelet thresholding, the residual image in method noise, derived from the initial estimate using nonlocal means (NLM), is exploited further. By incorporating the role of both the initial estimate and the residual image, spatially adaptive patch shapes are defined, and new weights are calculated, which thus results in better denoising performance for NLM. Experimental results demonstrate that our proposed method significantly outperforms original NLM and achieves competitive denoising performance compared with state-of-the-art denoising methods.

1. Introduction

Digital imaging devices such as digital cameras and camera phones are ubiquitous in our daily life, which use complementary metal oxide semiconductors (CMOS) or charged coupled devices (CCD) image sensors to acquire images. However, since the CMOS and CCD image sensors are subject to noise from two notable sources, i.e., electronic instruments and the photo-sensing devices [1, 2], the quality of captured images is usually not satisfactory, especially when images are taken in low light condition, which leads to degraded imaging results. Hence, denoising has become a fundamental image restoration problem in image signal processor (ISP).

During the past decades, image denoising has been widely studied. However, until now, how to remove the noise efficiently while preserving significant image details has remained a challenge. Early smoothing methods, such as Gaussian filter [3], anisotropic filter [4], total variation [5], and bilateral filter [6], perform noise removal solely based on the information provided in a local neighborhood, thereby resulting in disturbing artifacts around edges. Later, transform domain-based denoising methods have been developed and extensively studied as well [7–19]. Wavelet transform (WT) [7] decomposes the image into multiple frequency components, where the noise is removed with thresholding [8, 9] or statistical modeling [10–12]. By transforming back the processed wavelet coefficients into spatial domain, denoising is accomplished. Late development of WT denoising includes ridgelet [13] and curvelet [14] methods. In [15], adaptive principal components-based denoising method was proposed. Compared with WT that uses a fixed wavelet basis, it computes the locally fitted basis to decompose the image. In [16, 17], the highly overcomplete dictionary was trained by using K-SVD algorithm for sparse and redundant image representation. In [18, 19], discrete cosine transform (DCT) was applied to the local neighborhood, which achieves very sparse representation of the image and hence leads to effective denoising performance.

Recently, nonlocal methods, that exploit the image nonlocal self-similarity, have achieved outstanding denoising performance. In [20, 21], Buades et al. first proposed the nonlocal principle-based denoising method, called nonlocal means (NLM). In this method, noise-free pixel is estimated as a weighted average of all pixels in the image, where the weights are determined based on the similarity between the patch centered at the pixel being estimated and the patches centered at other pixels. Since NLM exploits the fact that

similar patches appear abundantly in the image and can contribute for denoising, it obtains high quality denoising performance. Subsequently, numerous extensions of NLM have been developed. In [22], the optimal neighborhood for each pixel was chosen during the iteration procedure to balance the accuracy of approximation and the stochastic error. In [23], the noisy image was classified into several region types, according to which the patch size was adaptively adjusted to match the local property. In [24], adaptive patch size and bandwidth were selected pixel-wise, relying on the feature metric that can provide a quantitative measure of local geometric structures. In [25], the smoothing parameter was chosen automatically based on noise estimation, and then the two-stage NLM with adaptive smoothing parameter was performed. In [26], quadtree-based NLM was proposed, which employs quadtree decomposition on each image patch to obtain more subpatches of various sizes. In addition to considering the patch sizes, some methods try to handle variable patch shapes. In [27], the adaptive binary shape for each patch was estimated by thresholding the difference between the central pixel and other patch pixels. In [28], shape-adaptive patches were constructed to match more homogeneous pixels successfully, especially in textured areas. In [29], several types of patches with various shapes were predefined and applied in NLM, respectively. Then, local estimates associated with these shapes were combined using Stein's unbiased risk estimate (SURE). To improve the accuracy of similarity measure, nonlocal similarity of residual image structures in method noise was further exploited in [30, 31]. Besides, rotation invariant patch comparison, that can handle rotational similarity existing in the image, was also studied in [32–36]. Analogously in [37], affine invariant similarity measure was applied to find more similar patches. Although NLM and its extensions have achieved significant denoising results, only exploiting the spatially nonlocal redundancy still limits their performance. Therefore, some methods combine the nonlocal principle with other techniques [38, 39], resulting in state-of-the-art denoising performance [40–43].

In this paper, we address these issues and propose an efficient denoising method, as shown in Figure 1. First, the original NLM is employed to obtain an initial estimate of the noisy image. However, due to inaccurate weight computation with noise interference, the initial estimate does not contain complete image details, which means the method noise still contains residual image information. For well preserving the residual image in method noise, the wavelet thresholding is used to smooth noise as much as possible. Then, the preserved residual image is combined with initial estimate to obtain a basic denoising result, based on which spatially adaptive patch shapes are defined using LPA-ICI and new weights are calculated. Finally, NLM denoising is implemented again but with the shape-adaptive patches and new weights.

The remainder of this paper is structured as follows: in Section 2, original NLM is briefly reviewed. In Section 3, our proposed denoising method is described in detail. In Section 4, we present and analyse the comparative experimental results. Finally, Section 5 concludes this paper.

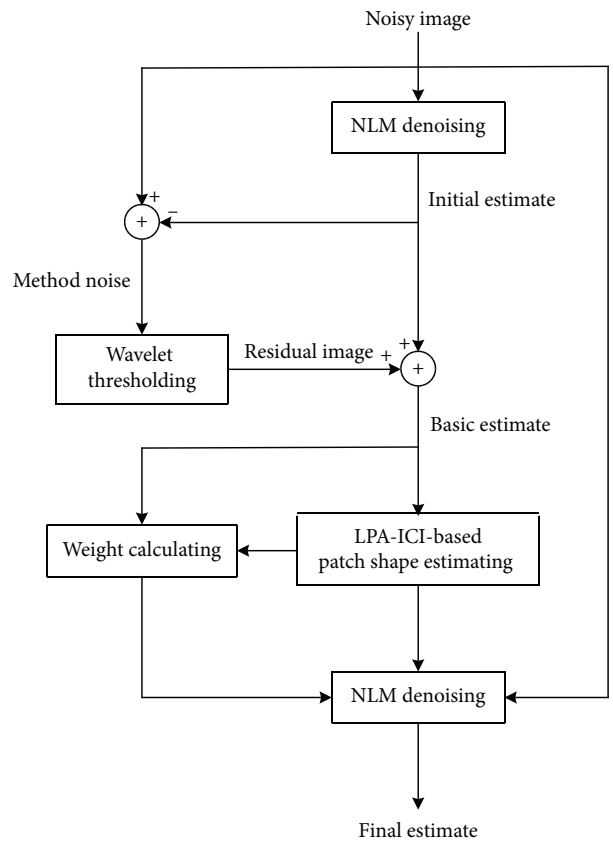


Figure 1: Flowchart of the proposed image denoising method.

2. Nonlocal Means

Given a noise-free image u defined on a discrete grid I, the noisy observation of u at pixel $i \in I$ is defined as

$$v(i) = u(i) + n(i), \tag{1}$$

where $n(i)$ is the zero-mean white Gaussian noise perturbation at pixel i. Let N_i denote the patch centered at pixel i, and its noisy observation is defined as

$$v(N_i) = \{v(j) | j \in N_i\}. \tag{2}$$

In classical NLM [20], for a pixel i, the estimated value of its noise-free version, $\hat{u}(i)$, is calculated as the weighted average of all noisy observations of the pixels in the image:

$$\hat{u}(i) = \frac{\sum_{j \in I} w_{i,j} v(j)}{\sum_{j \in I} w_{i,j}}. \tag{3}$$

The weight $w_{i,j}$ depends on the similarity of patch N_i and patch N_j, which is defined as

$$w_{i,j} = e^{-\left(\left\|v(N_j) - v(N_j)\right\|_2^2 / \lambda^2\right)}, \tag{4}$$

where $\|\cdot\|_2$ denotes the Euclidean norm to measure the similarity between the patches, and the parameter λ acts as a degree of filtering, therefore controlling the decay of the weight. Since the classical NLM only estimates a single pixel, it is referred as the pixel-wise NLM.

Later, the patch-wise NLM does also exist [21]. Here, the whole patch is estimated as follows:

$$\hat{u}(N_i) = \frac{\sum_{j\in I} w_{i,j} v(N_i)}{\sum_{j\in I} w_{i,j}}. \quad (5)$$

As the patches are overlapped, each pixel in a patch achieves multiple estimates. Therefore, for a pixel i, its different estimates are aggregated to obtain the final result:

$$\hat{u}(i) = \frac{1}{|A_i|} \sum_{j\in A_i} \hat{u}(N_j, i), \quad (6)$$

where $A_i = \{j | i \in N_j\}$ and $\hat{u}(N_j, i)$ denote the estimated value of noise-free patch N_j at pixel i. Thanks to the multiple estimations for each pixel, the patch-wise NLM achieves better denoising performance than the pixel-wise one. In this paper, when NLM is mentioned, it refers to the patch-wise NLM.

Based on the denoising result, method noise is defined as follows:

$$\hat{n} = v - \hat{u}. \quad (7)$$

It can be inferred that, if NLM performs well, the method noise must look like a noise and should contain as little structure as possible. In Figures 2(a) and 2(b), an example image and its noisy version ($\sigma = 20$) are shown, respectively. Figure 2(c) shows the NLM denoising result, while Figure 2(d) shows the corresponding method noise. As can be seen, obvious image structure appears in the method noise, which means that some image details are removed from the denoised image. Therefore, we make use of the residual image in method noise to exploit nonlocal self-similarity further.

3. Proposed Denoising Method

3.1. Method Noise Thresholding. Since the method noise contains obvious image structure, the residual image in it is estimated firstly. Here, we use the BayesShrink wavelet thresholding method [44] to suppress the noise as much as possible.

BayesShrink is an adaptive, data-driven thresholding strategy via soft-thresholding which derives the threshold in a Bayesian framework, assuming a generalized Gaussian distribution for the wavelet coefficients. This method is adaptive to each sub-band because it depends on data-driven estimates of the parameters. The threshold for a given sub-band is derived by minimizing Bayesian risk as follows:

$$T = \frac{\sigma_n^2}{\sigma_w}, \quad (8)$$

where σ_n^2 is the noise variance estimated from sub-band HH_1 by a robust median estimator, given by

$$\hat{\sigma}_n = \frac{\text{Median}\left(\left|Y_{i,j}\right|\right)}{0.6745}, \quad Y_{i,j} \in \{HH_1\}, \quad (9)$$

and σ_w^2 is the variance of wavelet coefficients in that sub-band, whose estimate is computed using

$$\hat{\sigma}_w^2 = \max\left(\hat{\sigma}_y^2 - \hat{\sigma}_n^2, 0\right), \quad (10)$$

where $\hat{\sigma}_y^2 = (1/MN)\sum_{i,j=1}^{M,N} Y_{i,j}^2$.

Figure 2(e) shows the filtered method noise thresholding with wavelet. As can be seen, the residual image is preserved well while the noise is smoothed efficiently. Then, we combine the initial estimate, denoted as $\hat{u}_{\text{initial}}$, and the filtered method noise, denoted as $\hat{n}_f$, together to obtain a basic estimate:

$$\hat{u}_{\text{basic}} = \hat{u}_{\text{initial}} + \hat{n}_f. \quad (11)$$

In Figure 2(f), we show the combined result. It can be seen that the basic estimate preserves more image details than the initial estimate, such as the regions marked by green boxes, which means that it will be more accurate to estimate the spatially adaptive patch shapes and to calculate the weights based on the basic estimate.

3.2. Spatially Adaptive Patch Shape Estimating. The anisotropic local polynomial approximation- (LPA-) intersection of confidence intervals (ICI) technique is used to estimate the spatially adaptive shape for each patch in the image.

Figure 3 shows the implementation of the LPA-ICI-based patch shape estimating. For a pixel i, eight directions are first predefined. For every specified direction $\theta_k = ((k-1)/4)\pi$, $k = 1, \ldots, 8$, a varying-scale family of narrow "linewise" directional LPA convolution kernels $\{g_{h,\theta_k}\}_{h\in H}$ is used to obtain a corresponding set of directional varying-scale estimates $\{\hat{u}_{h,\theta_k}\}_{h\in H}$, $\hat{u}_{h,\theta_k} = v \otimes g_{h,\theta_k}$, $h \in H$, where $H \in R^+$ is the set of scales. Then, for each estimate, a confidence interval is built as follows:

$$D_{h,\theta_k} = \left[\hat{u}_{h,\theta_k} - \Gamma\sigma, \hat{u}_{h,\theta_k} + \Gamma\sigma\right], \quad (12)$$

where $\Gamma > 0$ is a tuning parameter and σ is the noise standard deviation. Based on ICI rule, an adaptive scale $h^+(i, \theta_k) \in H$ is defined for every direction θ_k. Finally, the shape-adaptive patch N_i^+ is constructed as the polygonal hull of $\{\text{supp}\, g_{h^+(i,\theta_k),\theta_k}\}_{k=1}^8$.

In Figure 4, we show some examples of the estimated shape-adaptive patches in the noise-free, noisy, initially estimated, and basically estimated images, respectively. It can be seen that, due to the influence of strong noise, the estimated patch shapes in the noisy image are incorrect. The same goes for those in the initially estimated image but because of the loss of image details during NLM denoising. By contrast, patch shapes in the basically estimated image are more accurate.

3.3. New Weight Calculating. Based on the basic estimate, weights between the patches are calculated again. Since we estimate the spatially adaptive shape for each patch, the new weight is calculated as follows:

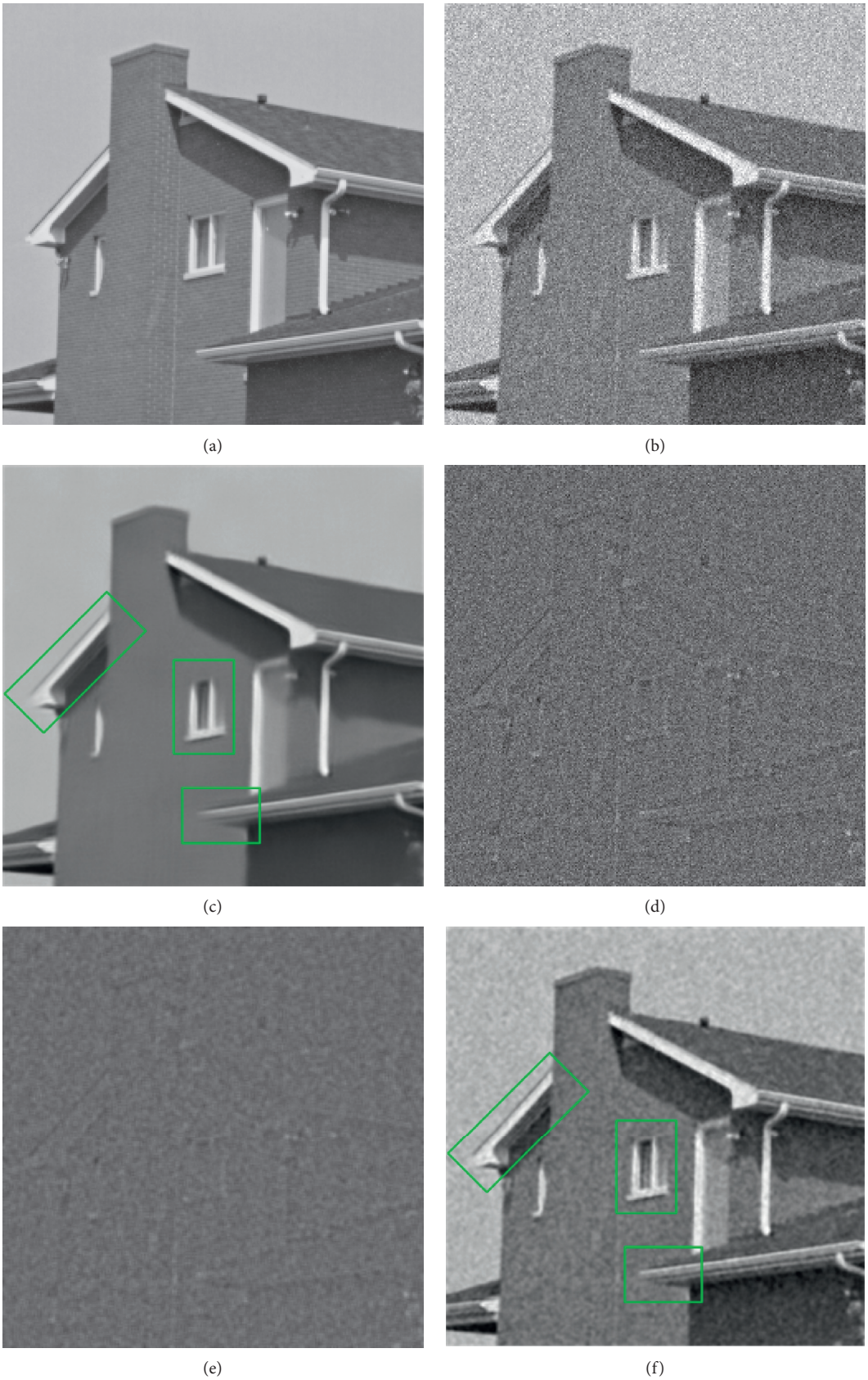


Figure 2: NLM denoising performance. (a) Original image; (b) noisy image ($\sigma = 20$); (c) initial estimate; (d) method noise; (e) method noise thresholding with wavelet; (f) basic estimate combining initial estimate and filtered method noise.

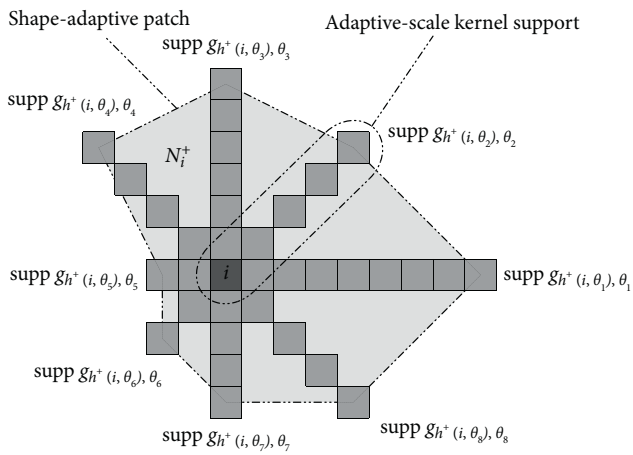


Figure 3: Implementation of the LPA-ICI-based patch shape estimating. "Linewise" one-dimensional directional LPA kernels are used for 8 directions. The shape-adaptive patch N_i^+ is constructed as the polygonal hull of the adaptive-scale kernel supports.

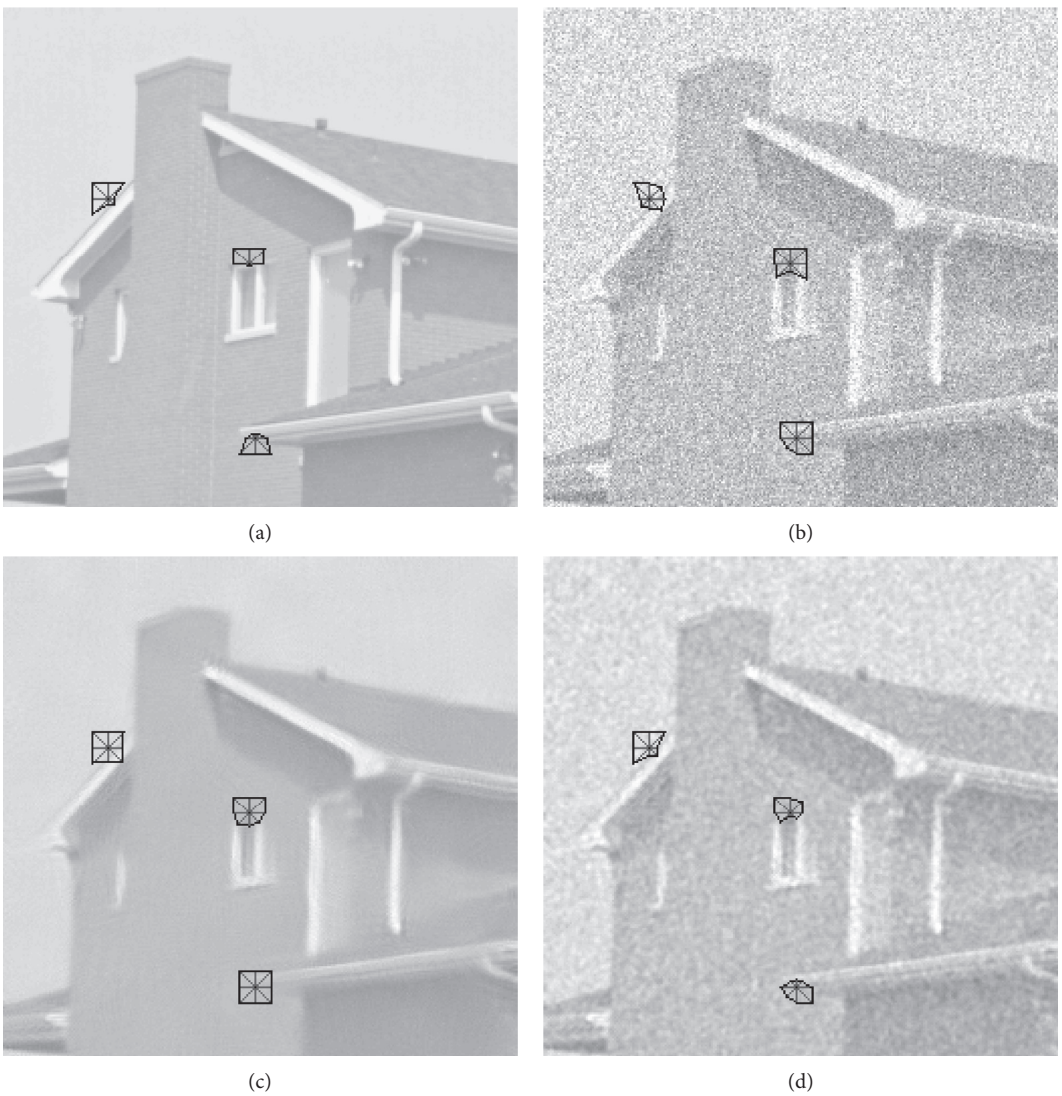


Figure 4: Some examples of the estimated shape-adaptive patches in (a) noise-free image, (b) noisy image ($\sigma = 50$), (c) initially estimated image by NLM, and (d) basically estimated image, respectively.

Table 1: The PSNR results by different denoising methods. In each cell, the results of the four denoising methods are presented in the following order: top left, NLM [21]; top right, NLM-SAP [29]; bottom left, BM3D-SAPCA [41]; bottom right, our proposed denoising method.

σ	20		35		50		100	
C. man	29.85	29.75	27.09	26.98	25.37	24.82	21.42	20.86
	30.91	29.97	**28.17**	27.54	**26.59**	26.02	**22.87**	22.50
House	32.40	32.62	29.79	29.00	27.74	26.37	23.16	22.79
	33.89	33.28	**31.37**	30.90	**29.52**	28.97	**25.08**	24.77
Peppers	30.17	30.64	27.08	27.50	25.16	25.26	21.02	21.02
	31.57	30.82	**28.75**	28.22	**26.98**	26.51	**23.24**	22.90
Lena	31.58	31.97	28.95	29.00	27.39	27.12	23.93	23.99
	33.20	32.62	**30.72**	30.28	**29.06**	28.67	25.36	**25.37**
Barbara	30.47	30.50	27.60	26.83	25.72	24.69	22.14	22.02
	32.37	31.36	**29.61**	28.34	**27.68**	26.40	**23.22**	22.65
Boats	29.80	29.64	26.99	26.80	25.28	25.04	22.19	22.31
	31.02	30.34	**28.51**	27.96	**26.88**	26.47	**23.68**	23.42
Man	29.75	29.58	27.03	26.92	25.41	25.39	22.54	22.95
	30.83	30.32	**28.38**	28.04	**26.93**	26.59	23.96	**23.98**
Hill	29.77	29.45	27.14	26.92	25.49	25.55	22.84	23.28
	30.85	30.32	**28.61**	28.21	**27.19**	26.86	24.26	**24.29**
Average	30.47	30.51	27.70	27.49	25.94	25.53	22.40	22.40
	31.83	31.12	**29.26**	28.68	**27.60**	27.06	**23.95**	23.73

The bold values represent the best results among the four methods, which has been explained in the second paragraph of Section 4.

Table 2: The SSIM results by different denoising methods. In each cell, the results of the four denoising methods are presented in the following order: top left, NLM [21]; top right, NLM-SAP [29]; bottom left, BM3D-SAPCA [41]; bottom right, our proposed denoising method.

σ	20		35		50		100	
C. man	0.840	0.845	0.763	0.773	0.714	0.706	0.541	0.493
	0.886	0.854	**0.827**	0.801	**0.787**	0.758	**0.643**	0.587
House	0.831	0.849	0.775	0.787	0.733	0.715	0.553	0.531
	0.876	**0.870**	**0.838**	**0.836**	**0.807**	0.788	**0.676**	0.639
Peppers	0.840	0.863	0.759	0.794	0.703	0.730	0.518	0.560
	0.886	**0.877**	**0.833**	**0.826**	**0.792**	0.772	**0.666**	0.631
Lena	0.830	**0.918**	0.764	**0.858**	0.724	**0.801**	0.574	0.644
	0.880	0.879	0.837	0.819	**0.801**	0.772	**0.674**	0.652
Barbara	0.861	**0.925**	0.779	0.843	0.711	0.768	0.510	**0.608**
	0.912	0.894	**0.863**	0.815	**0.811**	0.737	0.600	0.531
Boats	0.786	**0.876**	0.689	**0.789**	0.625	**0.720**	0.471	0.561
	0.828	0.810	0.761	0.737	0.708	0.679	**0.578**	0.531
Man	0.793	**0.871**	0.692	**0.785**	0.625	**0.721**	0.473	**0.582**
	0.840	0.825	0.763	0.745	0.710	0.684	0.579	0.555
Hill	0.765	**0.851**	0.659	**0.757**	0.589	**0.694**	0.448	**0.572**
	0.809	0.790	0.729	0.709	0.675	0.648	0.549	0.535
Average	0.818	**0.874**	0.735	0.798	0.678	0.731	0.511	0.568
	0.864	0.849	**0.806**	0.786	**0.761**	0.729	**0.620**	0.582

The bold values represent the best results among the four methods, which has been explained in the second paragraph of Section 4.

$$w_{i,j}^{+} = e^{-\left(\left\|\widehat{u}_{\text{basic}}\left(N_i^{+}\right)-\widehat{u}_{\text{basic}}\left(N_j^{i,+}\right)\right\|_2^2/\gamma^2\right)}, \tag{13}$$

where N_i^{+} denotes the estimated shape-adaptive patch, $N_j^{i,+}$ denotes the patch using the same shape with N_i^{+}, and γ is the filtering factor.

3.4. NLM with Shape-Adaptive Patches and New Weights. With shape-adaptive patches and new weights, we implement the NLM denoising again to remove the noise in the noisy image. For a shape-adaptive patch N_i^{+}, the estimated value of its noise-free version, $\widehat{u}(N_i^{+})$, is calculated as follows:

Figure 5: Denoising performance on the C. man image with moderate noise corruption. (a) Original image; (b) noisy image ($\sigma = 20$); denoised image (c) by NLM [21]; (d) NLM-SAP [29]; (e) BM3D-SAPCA [41]; (f) our proposed method.

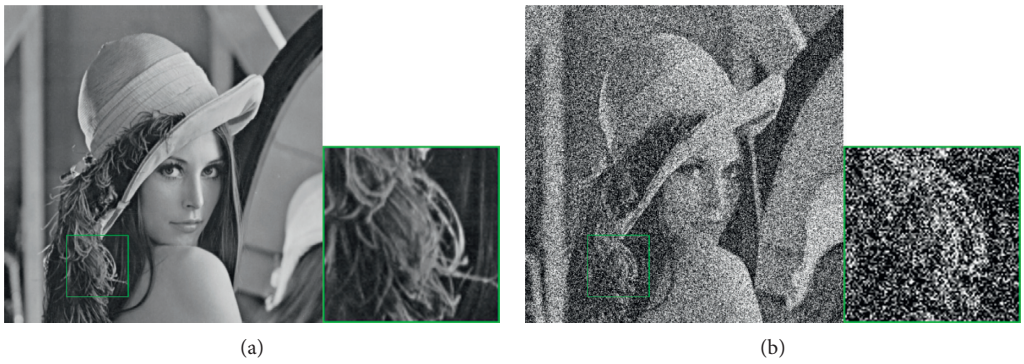


Figure 6: Continued.

(c) (d)

(e) (f)

Figure 6: Denoising performance on the Lena image with strong noise corruption. (a) Original image; (b) noisy image ($\sigma = 100$); denoised image (c) by NLM [21]; (d) NLM-SAP [29]; (e) BM3D-SAPCA [41]; (f) our proposed method.

$$\hat{u}\left(N_j^+\right) = \frac{\sum_{j\in I} w_{i,j}^+ v\left(N_j^{i,+}\right)}{\sum_{j\in I} w_{i,j}^+}. \tag{14}$$

Similarly, for a pixel i, its different estimates are aggregated to obtain the final result:

$$\hat{u}_{\text{final}}(i) = \frac{1}{\left|A_i^+\right|} \sum_{j\in A_i^+} \hat{u}\left(N_j^+, i\right), \tag{15}$$

where $A_i^+ = \left\{j | i \in N_j^+\right\}$ and $\hat{u}(N_j^+, i)$ denote the estimated value of noise-free patch N_j^+ at pixel i.

3.5. Computational Complexity. For the image of size $\sqrt{N} \times \sqrt{N}$, the computational complexity of the original NLM for initial estimate is $O(n_p N^2)$, where n_p denotes the patch size used by NLM. In our practical implementation, a limited window of size $\sqrt{w_1} \times \sqrt{w_1}$ is used to restrict the search of similar patches, which reduces the complexity to $O(n_p w_1 N)$. Besides, by using the moving average filter together with weight symmetry, the complexity can be further brought down to $O(w_1 N)$. Then, the computational complexity of method noise thresholding with BayesShrink wavelet is usually $O(N)$. For the spatially adaptive patch shape estimation with LPA-ICI technique, since it is based on convolutions against one-dimensional kernels for a very limited number of directions, its computational overhead for the whole noise removal processing is negligible. In the final estimating procedure, for the search window of size $\sqrt{w_2} \times \sqrt{w_2}$, the computational complexity is $O(w_2 N)$.

4. Results and Discussion

In this section, we compare our proposed method with original NLM [21] and other two state-of-the-art denoising methods: shape-adaptive patches-based NLM (NLM-SAP) [29] and shape-adaptive PCA-based BM3D (BM3D-SAPCA) [41]. In the experiments, a set of 8 natural images commonly used in the literature of image denoising are used for the comparison, and their noisy versions are simulated by adding independent white Gaussian noise with varying noise levels. The results of our proposed method are generated using the scales $H = \{1, 2, 3, 4, 5, 6\}$ and a search window of 21×21. The parameter γ is selected experimentally as $\gamma = 0.4\sigma$. The results of other three methods are obtained by using the codes available online with recommended parameters. To evaluate the quality of denoised images, the popular peak signal to noise ratio (PSNR) and the structural similarity index (SSIM) are calculated.

The results of the experiments are shown in Tables 1 and 2, where the best results among the four methods are highlighted. It can be seen that our proposed method significantly outperforms NLM and NLM-SAP and achieves competitive denoising performance compared with BM3D-SAPCA. Particularly, in some cases of high noise levels, our method performs even slightly better than BM3D-SAPCA. In terms of SSIM results, our proposed method is quite close to BM3D-SAPCA, and its superiority exists in all cases with respect to NLM.

Let us then focus on the visual quality of the denoised images by the four methods. In Figures 5 and 6, we show the denoising results on two typical images with moderate and

strong noise corruption, respectively. It can be seen that our proposed method is very effective in reconstructing both the smooth and the texture/edge regions. When the noise level is not very high, as shown in Figure 5, our proposed method performs better than NLM and NLM-SAP on edge preservation, and BM3D-SAPCA achieves the best visual output. When the noise level is high, as shown in Figure 6, however, details in the denoised images by NLM and NLM-SAP become blurred, and BM3D-SAPCA tends to generate many visual artifacts. By contrast, our proposed method performs much better, which preserves the image details well and generates much less artifacts.

5. Conclusions

In this work, we have presented an efficient image denoising method. By exploiting the residual image in the method noise, spatially adaptive patch shapes are defined, and new weights are calculated to improve the denoising performance of NLM further. Experimental results demonstrate that our proposed method is effective in noise removal and texture/edge preservation and can achieve competitive denoising performance compared with state-of-the-art denoising methods.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 52006235.

References

[1] N. Kawai and S. Kawahito, "Noise analysis of high-gain, low-noise column readout circuits for CMOS image sensors," *IEEE Transactions on Electron Devices*, vol. 51, no. 2, pp. 185–194, 2004.

[2] M. Cho and B. Javidi, "Three-dimensional photon counting imaging with axially distributed sensing," *Sensors*, vol. 16, no. 8, p. 1184, 2016.

[3] M. Lindenbaum, M. Fischer, and A. Bruckstein, "On Gabor's contribution to image enhancement," *Pattern Recognition*, vol. 27, no. 1, pp. 1–8, 1994.

[4] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[5] L. Rudin and S. Osher, "Total variation based image restoration with free local constraints," in *Proceedings of the 1st International Conference on Image Processing*, pp. 31–35, Austin, TX, USA, November 1994.

[6] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 839–846, Bombay, India, January 1998.

[7] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, NewYork, 1998.

[8] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[9] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," *Wavelets and Statistics*, Springer, New York, US, pp. 125–150, 1995.

[10] M. Kivanc Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 300–303, 1999.

[11] S. G. Chang, B. Bin Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1522–1531, 2000.

[12] A. Pizurica, W. Philips, I. Lemahieu, and M. Acheroy, "A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising," *IEEE Transactions on Image Processing*, vol. 11, no. 5, pp. 545–557, 2002.

[13] G. Y. Chen and B. Kégl, "Image denoising with complex ridgelets," *Pattern Recognition*, vol. 40, no. 2, pp. 578–585, 2007.

[14] J. L. Jean-Luc Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 670–684, 2002.

[15] D. D. Muresan and T. W. Parks, "Adaptive principal components and image denoising," in *Proceedings of the 2003 International Conference on Image Processing (Cat. No.03CH37429)*, pp. 1101–1104, Barcelona, Spain, September 2003.

[16] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[17] M. Aharon, M. Elad, and A. Bruckstein, "$rm K$-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[18] G. Yu and G. Sapiro, "DCT image denoising: a simple and effective image denoising algorithm," *Image Processing On Line*, vol. 1, pp. 292–296, 2011.

[19] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images," *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1395–1411, 2007.

[20] A. Buades, B. Coll, and J. M. Morel, "A non local algorithm for image denoising," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 60–65, San Diego, CA, USA, June 2005.

[21] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling and Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[22] C. Kervrann and J. Boulanger, "Optimal spatial adaptation for patch-based image denoising," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2866–2878, 2006.

[23] W. L. Zeng and X. B. Lu, "Region-based non-local means algorithm for noise removal," *Electronics Letters*, vol. 47, no. 20, pp. 1125–1127, 2011.

[24] J. Hu and Y. Luo, "Non-local means algorithm with adaptive patch size and bandwidth," *Optik*, vol. 124, no. 22, pp. 2639–5645, 2013.

[25] S. Zhu, Y. Li, and Y. Li, "Two-stage non-local means filtering with adaptive smoothing parameter," *Optik*, vol. 125, no. 23, pp. 7040–7044, 2014.

[26] C. Zuo, L. Jovanov, B. Goossens et al., "Image denoising using quadtree-based nonlocal means with locally adaptive principal component analysis," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 434–438, 2016.

[27] A. A. Tahmouresi, S. Saryazdi, and S. R. Seydnejad, "Non-local means denoising using an adaptive Kernel," in *Proceedings of the 20th Iranian Conference on Electrical Engineering (ICEE2012)*, pp. 1436–1441, Tehran, Iran, May 2012.

[28] S. Peng, W. Changcheng, G. Han, and Z. Jianjun, "An adaptive nonlocal mean filter for polsar data with shape-adaptive patches matching," *Sensors*, vol. 18, no. 7, Article ID 2215, 2018.

[29] C.-A. Deledalle, V. Duval, and J. Salmon, "Non-local methods with shape-adaptive patches (NLM-SAP)," *Journal of Mathematical Imaging and Vision*, vol. 43, no. 2, pp. 103–120, 2012.

[30] C. Yang, X. Zhang, and H. Zhong, "A new weight for nonlocal means denoising using method noise," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 535–538, 2012.

[31] B. K. Shreyamsha Kumar, "Image denoising based on nonlocal means filter and its method noise thresholding," *Signal, Image and Video Processing*, vol. 7, no. 6, pp. 1211–1227, 2013.

[32] S. Grewenig, S. Zimmer, and J. Weickert, "Rotationally invariant similarity measures for nonlocal image denoising," *Journal of Visual Communication and Image Representation*, vol. 22, no. 2, pp. 117–130, 2011.

[33] S. Zimmer, S. Didas, and J. Weickert, "A rotationally invariant block matching strategy improving image denoising with non-local means," in *Proceedings of the International Workshop Local Non-local Approx. Image Process*, pp. 135–142, Lausanne, Switzerland, August 2008.

[34] C. Zuo, L. Jovanov, H. Q. Luong et al., "Rotation invariant similarity measure for non-local selfsimilarity based image denoising," in *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*, pp. 1618–1622, Quebec City, Canada, September 2015.

[35] R. Yan, L. Shao, S. D. Cvetkovic, and J. Klijn, "Improved nonlocal means based on pre-classification and invariant block matching," *Journal of Display Technology*, vol. 8, no. 4, pp. 212–218, 2012.

[36] O. Kleinschmidt, T. Brox, and D. Cremers, "Nonlocal texture filtering with efficient tree structures and invariant patch similarity measures," in *Proceedings of the International Workshop On Local And Non-local Approximation In Image Processing*, pp. 103–113, Lausanne, Switzerland, August 2008.

[37] V. Fedorov and C. Ballester, "Affine non-local means image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2137–2148, 2017.

[38] C. Shen, Y. Qi, J. Wang, G. Cai, and Z. Zhu, "An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 170–184, 2018.

[39] X. Jiang, C. Shen, J. Shi, and Z. Zhu, "Initial center frequency-guided VMD for fault diagnosis of rotating machines," *Journal of Sound and Vibration*, vol. 435, no. 24, pp. 36–55, 2018.

[40] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[41] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "BM3D image denoising with shape-adaptive principal component analysis," in *Proceeding Signal Process. Adapt. Sparse Struct. Represent*, Saint Malo, France, April 2009.

[42] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 2272–2279, Kyoto, Japan, September 2009.

[43] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, 2013.

[44] S. G. Chang, B. Bin Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000.