Hindawi
Advances in Materials Science and Engineering
Volume 2020, Article ID 6625273, 12 pages
https://doi.org/10.1155/2020/6625273

Research Article

Rolling Bearing Fault Diagnosis Based on Stacked Autoencoder Network with Dynamic Learning Rate

Hong Pan,[1] **Wei Tang**,[2] **Jin-Jun Xu**,[1] **and Maxime Binama**[2]

[1]*College of Energy and Electrical Engineering, Hohai University, Nanjing 211100, China*
[2]*College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 211100, China*

Correspondence should be addressed to Hong Pan; hongpan_00@163.com

Received 3 October 2020; Revised 4 December 2020; Accepted 18 December 2020; Published 28 December 2020

Academic Editor: José António Fonseca de Oliveira Correia

Fault diagnosis is of great significance for ensuring the safety and reliable operation of rolling bearing in industries. Stack autoencoder (SAE) networks have been widely applied in this field. However, the model parameters such as learning rate are always fixed, which have an adverse effect on the convergence speed and accuracy of fault classification. Thus, this paper proposes a dynamic learning rate adjustment approach for the stacked autoencoder network. First, the input data is normalized and enhanced. Second, the structure of the SAE network is selected. According to the positive and negative value of the training error gradient, a learning rate reducing strategy is designed in order to be consistent with the current operation of the network. Finally, the fault diagnosis models with different learning rate adjustment are conducted in order to validate the better performance of the proposed approach. In addition, the influence of quantities of labeled sample data on the process of backpropagation is analyzed. The results show that the proposed method can effectively increase the convergence speed and improve classification accuracy. Moreover, it can reduce the labeled sample size and make the network more stable under the same classification accuracy.

1. Introduction

As an important part of rotating machinery, bearing plays an important role in modern industry. Bearing faults could cause unbearable and unpredictable loss [1, 2]. Therefore, lots of artificial intelligent (AI) fault diagnosis methods have been applied to keep the bearings working properly and reliably. However, the traditional AI methods are primarily based on shallow machine learning theory, which work on the original feature representation without creating new features during the learning process. It is difficult to reveal the inherent nonlinear relationship of complex mechanical systems [3, 4]. In addition, with the development of condition monitoring, the data of the bearing working station become much wider than ever before, which brings new opportunities and challenges for bearing fault diagnosis. In view of the characteristics of big data such as imperfection, multisource, and low value density, the shallow machine learning method needs to be fundamentally improved. Consequently, the deep learning method has been carried out.

Deep learning was considered a top-ten breakthrough by MIT Technology Review. It attempts to assemble deep architectures with many layers and build more complex nonlinear functions to simulate the process of learning knowledge [5]. Due to the powerful feature extraction ability and unsupervised training pattern, the deep learning method has received a great deal of attention in many areas such as computer vision [6–8], speech recognition [9, 10], text processing [11, 12], medicine [13, 14], finance [15, 16], and driverless car [17–19]. Especially in the field of machine fault diagnosis, many scholars have done a lot of research which included bearings [20–23], wind turbines [24–26], pumps [27–29], and power transmission systems [30–32]. Until now, there are four typical architectures: stacked autoencoder (SAE), deep belief network (DBN), convolutional neural network (CNN), and recurrent neural network (RNN). Among all of these models, SAE can learn rich representation features and reduce data dimensionality and has received a great deal of attention in the field of bearings. Sun et al. [33] used sparse autoencoder to realize bearing

fault diagnosis and analyzed the influence of dropout on the fault recognition rate. Pang and Yang [34] developed a cross-domain stacked denoising autoencoders (CD-SDAE) with a new adaptation training strategy. Wang et al. [35] proposed a new activation function named RelTanh to solve the problem of vanishing gradient. This research makes SAE own the potential to be used to improve the diagnosis accuracy of bearing. However, there are still some questions that need to be solved. For example, the learning rate is an important parameter of iteration. The traditional SAE adopts a fixed learning rate, which requires a lot of experience to set the parameter. If the learning rate is set too high, it will be difficult to converge or skip the optimal value. Otherwise, if the learning rate is set too small, the convergence speed will be too slow. Some scholars have studied methods including Adam and AdaDec [36], but these methods do not really change the learning rate, and these methods will also be affected by the initial learning rate. In addition, the quantity analysis on the number of labeled sample data in the process of reverse fine-tuning of SAE is rarely reported. If the labeled sample data can be decreased, the cost of collecting and marking sample data would be substantially reduced.

In order to solve the above problems, this paper proposes a novel dynamic learning rate method to replace the original fixed learning rate in pretraining and reverse fine-tuning process of bearing fault diagnosis, making the following two main contributions: (1) according to the positive and negative value of training error gradient, a learning rate reducing strategy is designed to be consistent with the current operation of the network. The convergence rate and convergence accuracy had been accelerating significantly. (2) It studies the influence of the number of labeled sample data on the accuracy and iterations of SAE. The bearing data sets, provided by Case Western Reserve University's (CWRU) Bearing Data Center, were used to verify the performance of the proposed method. Compared with the fixed learning rate model, the results showed the proposed method took up less convergence time and had higher classification accuracy. Under the same accuracy, the proposed method needed less labeled sample data.

The organization of the remainder of the paper is as follows: Section 1 introduces the basic methods briefly and the proposed approach. Section 2 details the data source and the method of data processing. In Section 3, some experiments are conducted to evaluate the proposed method under different dynamic learning rates and different numbers of labeled samples; visualization about the proposed method is presented in Section 4 too. Finally, we present the conclusions and future work in Section 5.

2. Materials and Methods

2.1. Stacked Autoencoder

2.1.1. Autoencoder. Autoencoder (AE) is a single-hidden layer neural network proposed by Rumelhart in 1986, whose structure is shown in Figure 1 [37]. This kind of network keeps the input and output as consistent as possible in the way of unsupervised learning. Assume the input is n-dimensional data X and the output is n-dimensional data Y. The transfer process of raw data from the input layer to the hidden layer is called encoding, and the transfer process from the hidden layer to the output layer is called decoding, which can be described by equations (1) and (2) [5]. In reality, AE aims to learn an approximate function of input data through minimizing errors between the reconstructed data and the original data.

The mathematical expressions of the autoencoder are as follows:

$$H = \sigma_a\left(W_a \cdot X + b_a\right), \tag{1}$$

$$Y = \sigma_s\left(W_s \cdot H + b_s\right), \tag{2}$$

$$\min_{\theta} L\left(\theta\right) = \left\|Y - X\right\|_2^2, \tag{3}$$

$$\theta = \left[W_a, b_a, W_s, b_s\right], \tag{4}$$

where $W_a \in R^{n\times k}, W_s \in R^{k\times n}, b_a \in R^k, b_s \in R^k$ are the weights and bias that need to be optimized and $\sigma_a(\cdot), \sigma_s(\cdot)$ are the activation functions.

2.1.2. Stacked Autoencoder. SAE was improved by Hinton on the autoencoder machine, and its network structure is shown in Figure 2 [38]. The coding part of the autoencoder is stacked; that is, the input of the first layer of the AE machine is the original data, and the input of the lower layer is the hidden layer data of the upper layer. Finally, a classifier is added to the network. The training mode of SAE is consisting of the pretraining and the reverse fine-tuning process. It uses a large amount of unlabeled data for unsupervised learning, extracts features autonomously, and uses the labeled data to reverse fine-tuning the network. Both the pretraining and the reverse fine-tuning process are based on the gradient descent algorithm.

2.2. Dynamic Learning Rate. The essence of the gradient descent method is adjusting the weight of the network according to the partial derivative of the loss function iteratively. The updated calculation formula of the weight is as follows:

$$W_{i+1} = W_i - \eta\frac{\partial L_i}{\partial W_i}, \tag{5}$$

where W_i and W_{i+1} are the weights of the i^{th} iteration and the $(i+1)^{\text{th}}$ iteration in the calculation process. η is the learning rate and L_i is the loss function.

Learning rate is a very important parameter in the training process. If the learning rate is too large, it may lead to the difficulty of convergence or skip the optimal solution. On the contrary, if the learning rate is too small, it may lead to slow convergence speed, increase in the calculation time, and difficulty of efficiency improvement. In order to solve this problem, the learning rate should be nonconstant and self-adapting adjusted.

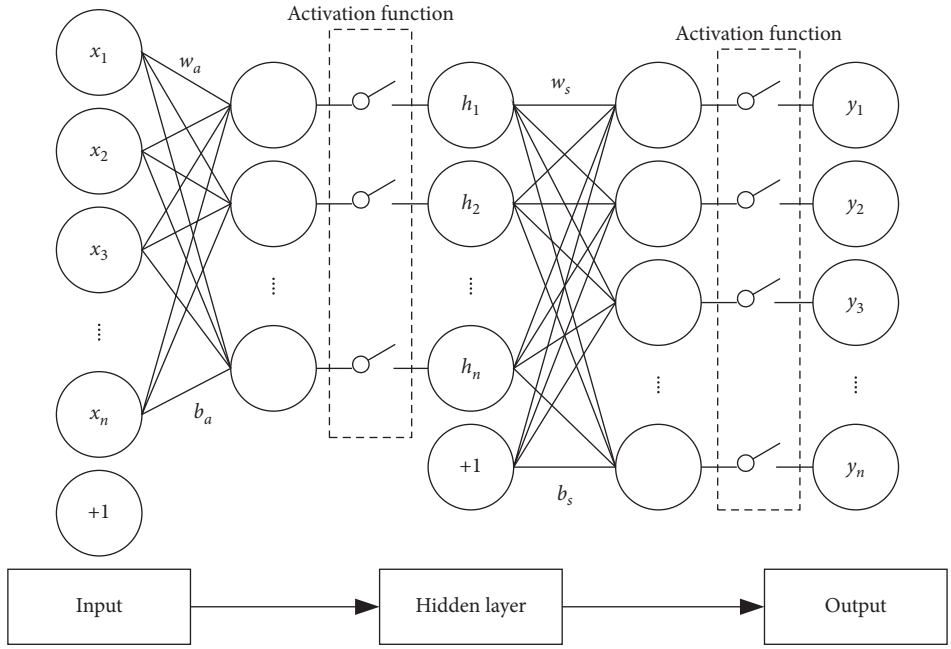

Figure 1: The network structure of the autoencoder.

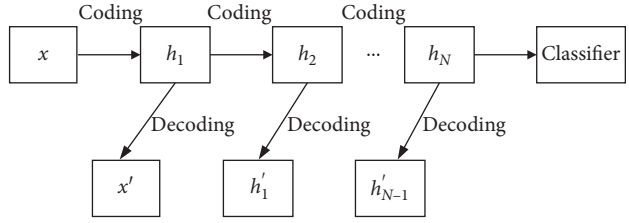

Figure 2: The network structure of SAE.

2.2.1. AdaDec Dynamic Learning Rate. AdaDec is an improved form based on AdaGrad; it was proposed by senior in 2013 [36]. The principle is shown in equations (6), (7), and (8). The gradient part of the denominator is determined by the gradient value of the previous round and the current round. It eliminates too many historical gradient data. At the same time, it uses a power index with a downward trend as a numerator to ensure the stable decline of the learning rate:

$$\eta(t) = \frac{p(t)}{\sqrt{K + G(t)}}, \tag{6}$$

$$G(t) = \varepsilon g^2(t-1) + g^2(t), \tag{7}$$

$$p(t) = \eta(0)\left(1 + \frac{t}{R}\right)^{-q}. \tag{8}$$

2.2.2. Improved Dynamic Learning Rate. The AdaDec method has the following disadvantages: it depends on the initial learning rate setting; it can only reduce the learning rate, and it cannot increase the learning rate.

Because the relationship between the partial derivative of the loss function and the weight is complex, the training error is selected as the basis of the learning rate adjustment. In this manuscript, the learning rate adjustment strategy is designed as follows:

$$h(i+1) = \begin{cases} \dfrac{h(i)}{1 + \left(\Delta L^i\right)^2}, & \Delta L^i > 0, \\ \dfrac{h(i)}{\sqrt{1 - \left(\Delta L^i\right)^2}}, & \Delta L^i \le 0, \end{cases} \tag{9}$$

where $h(i)$ is the learning rate at iteration i, $h(i+1)$ is the learning rate at iteration $i+1$, and ΔL^i is the gradient of training error at iteration i $(i > 2)$. The forward difference quotient method is adopted to calculate the gradient of reconstruction error, and the solution method is shown as follows:

$$\Delta L = \frac{L(i) - L(i-2)}{2}. \tag{10}$$

Through the calculation of the above equation, if the volatility of training error is obvious, ΔL^i is negative, so the learning rate decreases in a slow way. Otherwise, if the training error decreases smoothly, ΔL^i is positive, so the

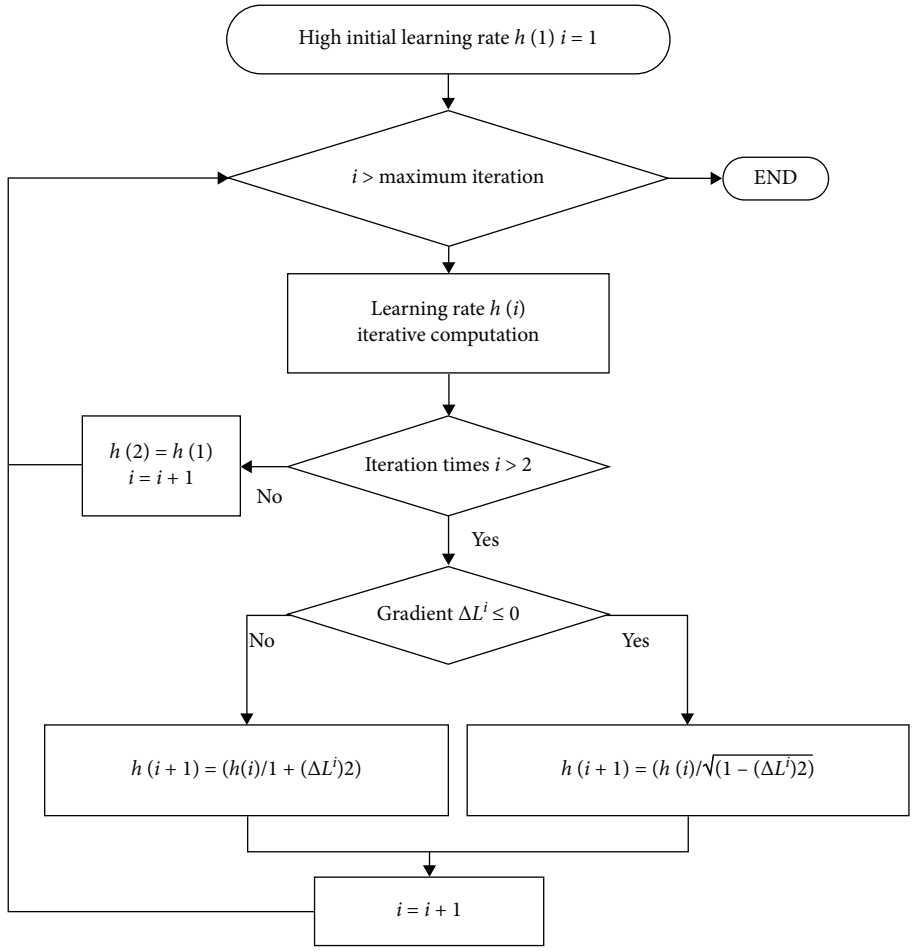

Figure 3: Training flow chart of dynamic learning rate.

learning rate decreases rapidly. Furthermore, in order to avoid the learning rate being too small, the range of the learning rate is limited to 0.01–5. The flowchart of the dynamic learning rate method is shown in Figure 3.

3. Data Source and Data Processing

3.1. Data Source. The proposed method was verified by the experimental data provided by Western Reserve University (CWRU) [39]. The test rig is shown in Figure 4. The bearing fault location is outer race, inner race, and rolling body, respectively. For each type of fault, there are three fault diameters, 0.18 mm, 0.36 mm, and 0.54 mm. So, there are 10 kinds of classifications in this data set. Two accelerometers are installed at the drive end and the fan end of the motor casing to collect vibration signals at a sampling frequency of 12 kHz. 120000 nodes were used for each state, among which the first 80800 nodes were taken as training data, and the rest were test data.

3.2. Data Processing

3.2.1. Data Normalization. The vibration data were normalized. The normalization formula is as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{11}$$

where X represents the sample, X' represents the normalized result of X, $X_{\max}$ represents the maximum value of X, and $X_{\min}$ represents the minimum value of X.

3.2.2. The Enhancement of the Data Set. The data length of each sample was set to 800. In order to obtain more training samples, the overlapping sampling technique by sliding window was adopted to enhance the data according to [40, 41], as shown in Figure 5. The offset was set to 50, and the sample size of the training data in each state was 1600, while the test data was not overlapped, and the sample size of the test data in each state was 49. So, the sum of the training data samples was 16000, and the sum of the test data samples was 490.

4. Results and Discussion

4.1. Network Structure. The constant pretraining learning rate was 0.1, the pretraining iteration times were 600, the reverse fine-tuning learning rate was 0.01, the labeled samples were set to 10% (after calculation, when the ratio is more than 10%, the change of accuracy is not obvious), the

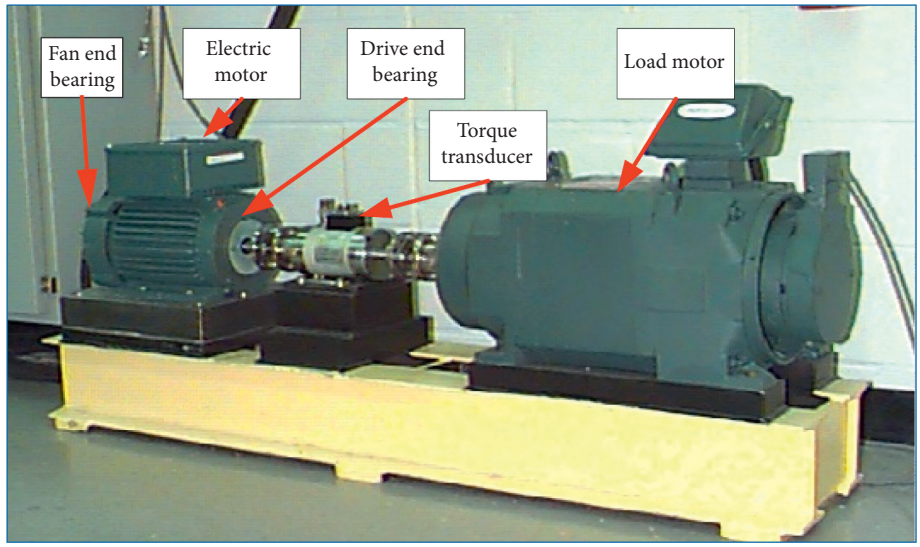

Figure 4: Fault simulation test rig of the CWRU data set.

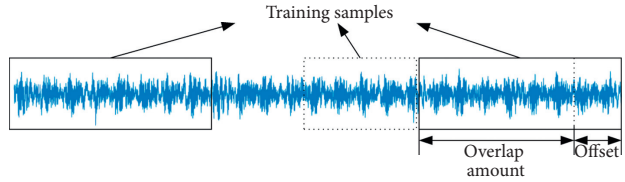

Figure 5: Schematic diagram of overlapping sampling.

iteration times were 600, and the batch size was 40. The number of hidden layer nodes was set to 95 initially according to the following empirical formula:

$$S = \sqrt{mn} + \frac{k}{2}, \tag{12}$$

where m is the length of the input data node, n is the length of the output data node, and k is a constant within [0, 10].

First, aiming to study the influence of different hidden layers on the bearing fault classification, a comparative experiment was conducted in which the hidden layers were set to 1, 2, 3, 4, 5, and 6, and the average value of ten results was taken as the final result.

As shown in Table 1, the number of hidden layers was three and the accuracy was the highest. If the number of hidden layers was too small, the network cannot extract the features of the input data effectively, so you can find that the accuracy was 77.08% when there was only one hidden layer. On the other hand, the network with lots of hidden layers also cannot achieve a good accuracy because it might lead to the information loss of data features in the iteration process. When the hidden layers ranged from three to six, the accuracy reduced to 86.62%. Thus, it is important to choose an appropriate number of hidden layers. In the next experiment, according to the above analysis, the number of hidden layers was set to 3 identically in order to improve the accuracy of the network.

Subsequently, the experiments with different hidden layer nodes were compared and analyzed. All the results were the average of ten experiments also. As can be seen from Table 2, the network structure (800-400-200-100-10) had the highest accuracy, and this network structure was selected in the follow-up study.

4.2. Dynamic Learning Rate Adjustment. The experiments with the fixed learning rates of 0.01, 0.1, 0.2, and 0.3 and dynamic learning rate were conducted simultaneously. In the pretraining process, the learning rate was set to 0.2 initially. The other network parameters were consistent with the network of fixed learning rate. The training error curve is shown in Figure 6.

As can be seen from Figure 6, the training error reduced with the increase of iterations. When the learning rate was 0.3, the training error appeared unstable and fluctuant at the beginning of the iterations and converged into a local optimal solution eventually. When the learning rate was 0.01, at the end of the iterations, the training error was still unstable (not convergent), and the training error showed an obvious downward trend. Different from the performance of the learning rate 0.3 and 0.01, the training error reduced quickly and smoothly when the learning rate was set to 0.1 and 0.2 and the dynamic learning rate. The training error of the dynamic learning rate, especially, was the smallest. Besides, the descent velocity of the training error with the dynamic

Table 1: Comparison of the results of different hidden layers.

Network structure	Accuracy (%)
800-95-10	77.08
800-95-95-10	90.58
800-95-95-95-10	95.47
800-95-95-95-95-10	94.96
800-95-95-95-95-95-10	93.83
800-95-95-95-95-95-95–10	86.62

Table 2: Comparison of accuracy of different grid structures.

Network structure	Accuracy (%)
800-400-400-400-10	95.73
800-400-300-100-10	97.86
800-600-250-100-10	96.99
800-400-200-100-10	98.97

learning rate was the fastest. It is obvious that the network with the dynamic learning rate requires minimal steps to reach the stable and convergent station. It can be seen from Figure 7 and Table 3 that the dynamic learning rate method has a faster convergence speed and better convergence accuracy than AdaDec.

Table 3 shows the iterations numbers, the time consumption, and the training error of various learning rates. Compared with the results of learning rates of 0.1 and 0.01, the performance of dynamic learning rate was much better, whereas when compared with the results of learning rate 0.2, the iteration number and convergence time of the dynamic learning rate were increased by 24.4% and 25.5%, respectively, but the training error decreased by 47.6% dramatically. The training error reduced from 0.2868 to 0.1504. So, from a comprehensive perspective, the proposed dynamic learning method can be regarded as a better choice to improve the effectiveness of bearing fault diagnosis. Figure 8 shows the change of dynamic learning rate with the iteration process. The learning rate increased at the beginning, reaching a maximum of 0.2863; then, the learning rate gradually decreased to 0.0573 eventually. The initial learning rate given by AdaDec is 0.02. When the initial learning rate is greater than 0.03, the training error is always around 3.6 and cannot converge. Therefore, the AdaDec method is also affected by the initial learning rate. It can be seen from Table 4 that dynamic learning can converge to better accuracy no matter what the initial learning rate is.

4.3. Different Number of Labeled Samples in the Process of Reverse Fine-Tuning. In order to explore the effect of dynamic learning rate in the process of reverse fine-tuning, three groups of experiments are set up, in which the fixed learning rate of the reverse fine-tuning process is 0.01, the weight value and bias obtained by the pretraining method of 0.1 fixed learning rate or dynamic learning rate. The comparison of the results obtained through reverse fine-tuning training is shown in Figure 9. It is obvious that the accuracy of the dynamic learning rate used in the process of reverse fine-tuning showed fluctuation also. But compared with the

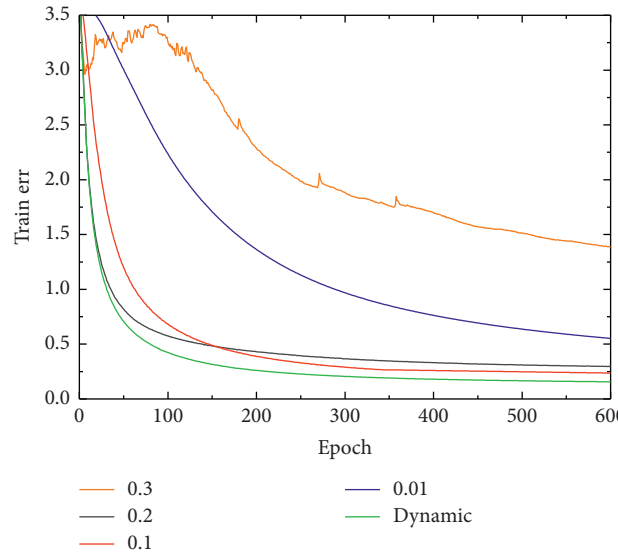

Figure 6: The training error of different learning rates.

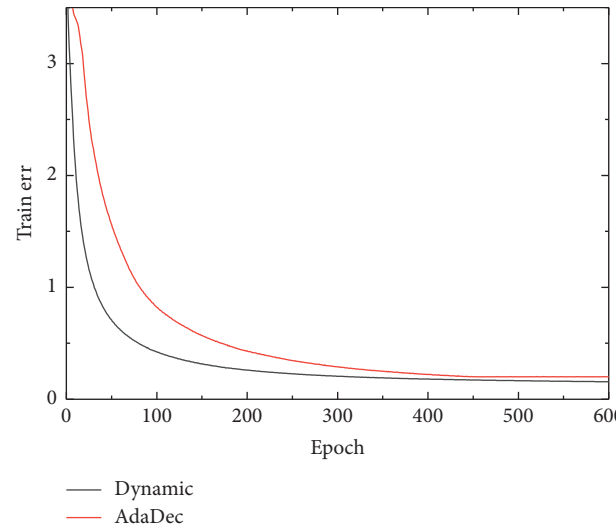

Figure 7: The training errors between dynamic and AdaDec.

fixed learning rate 0.01, the amplitude of variation of dynamic learning rate was bigger at the beginning of the iteration. While after about 100 iterations, the fluctuation of the dynamic learning rate became smaller than that of the fixed learning rate. It proved that the dynamic learning rate method not only had higher accuracy but also had better convergence. It can also be seen from the figure that the results obtained by using the fixed learning rate in both pretraining and reverse fine-tuning are the worst in terms of stability and accuracy.

For the purpose of exploring the influence of the labeled samples on the accuracy in the process of reverse fine-tuning, the experiments of different percentages of labeled samples were conducted. The labeled samples were set to 1%, 2%, 3%, 4%, 5%, 6%, 8%, and 9% of the training data, respectively. The rules for adding labeled samples are shown in

Table 3: The results of different learning rates.

Different learning rates	The number of iterations required to converge or reach the number of iterations	Time required to converge or reach the number of iterations (s)	Training error value in convergence
0.2	270	437.696	0.2868
0.1	401	667.442	0.2346
0.01	600	996.857	0.4909
AdaDec	450	705.688	0.2014
Dynamic	336	549.325	0.1504

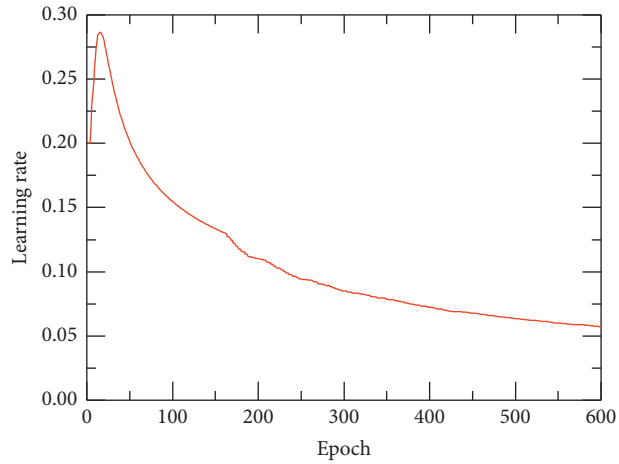

Figure 8: Dynamic learning rate curve.

Table 4: Training error of dynamic learning rate with different initial learning rates.

Initial learning rate	The number of iterations required to converge or reach the number of iterations	Training error value in convergence
0.3	313	0.1496
0.2	336	0.1504
0.1	456	0.1516
0.01	573	0.1531

Figure 10. 2% of the labeled samples should include the first 1% of the labeled samples. In the same way, 3% of the labeled samples should include the first 2% of the labeled samples and so on.

Experimental results of different percentages of labeled samples were shown in Figure 10. The reverse fun-tuning process was accomplished using the weight and bias obtained by the same pretraining process in which dynamic learning rate was used. The reverse fine-tuning process used a dynamic learning rate too. As can be seen from Figure 11, the more the labeled samples, the fewer the required iteration steps to achieve 90% accuracy. In general, the accuracy increased with the number of labeled samples.

In order to determine the most suitable number of labeled samples, the experiment of fixed learning rate 0.1 and dynamic learning rate used in pretraining process was conducted. The fault classification accuracy was the average value obtained by ten experiments. The results were shown in Figure 12. It is obvious that the network of dynamic learning rate was more accurate under the same labeled sample size. In other words, fewer labeled samples were needed for dynamic learning rate to achieve the same accuracy. When the percentage of the labeled date ranged from 1% to 8%, the accuracy increased rapidly. When the percentage of labeled data exceeded 8%, the accuracy increased slowly. Taking into account the iteration numbers and accuracy simultaneously, it was recommended to set the percentage of labeled data as 8%.

4.4. Visualization. In order to verify the above conclusions further, the visualization of the third hidden layer with different percentages of labeled samples is given in Figures 13–15. In this manuscript, the T-Distributed Stochastic Neighbor Embedding (t-SNE) method was adopted to extract two features for visualization. T-SNE was proposed by van der Maaten and Hinton in 2008 [42]. T-SNE has achieved good results in dimension reduction, clustering, and visualization. The horizontal axis and the vertical axis represent the first two principle components achieved by t-SNE. Figure 12 is the visualization of 1% labeled samples. It can be seen that only four kinds of faults can be

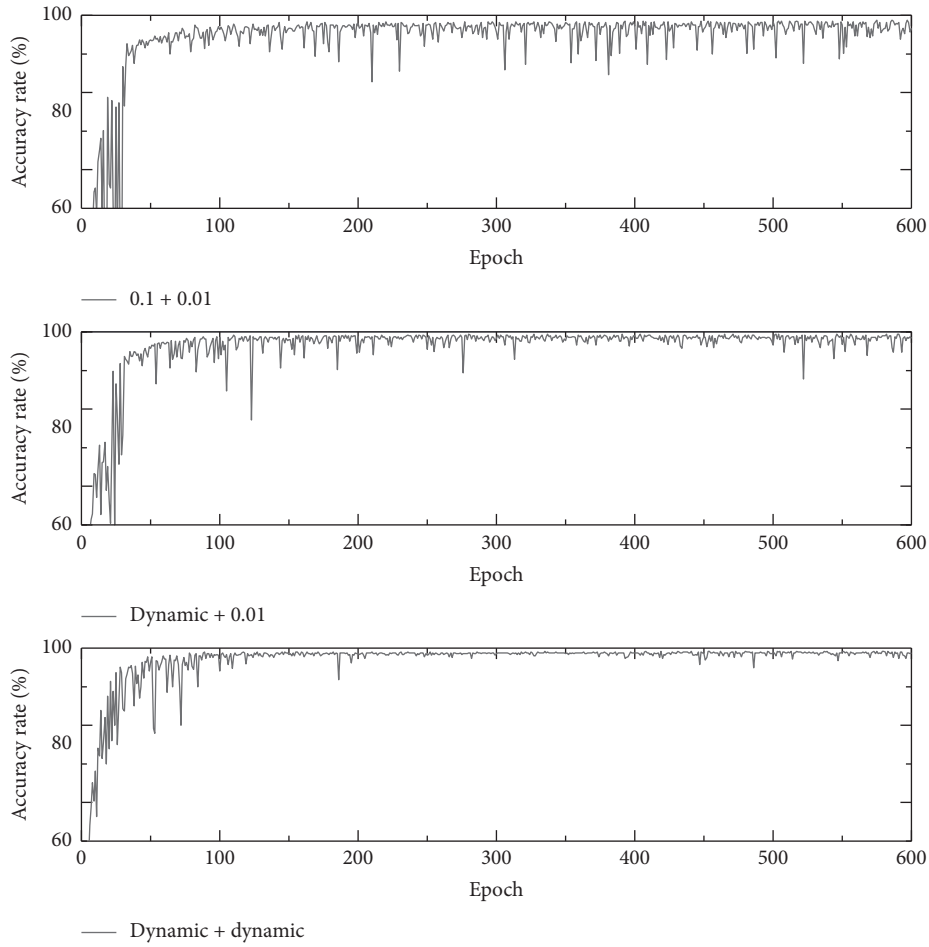

Figure 9: The accuracy fluctuation of two kinds of learning rate.

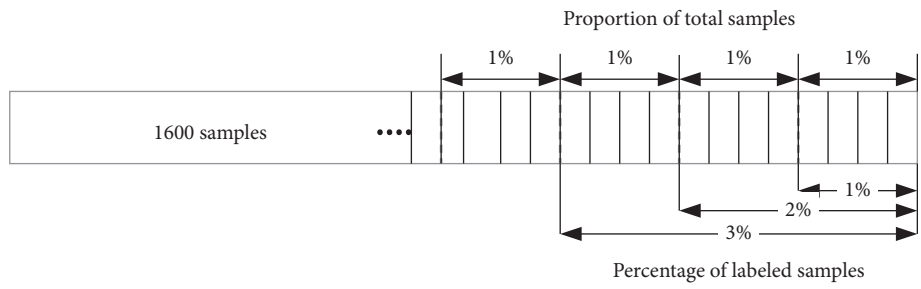

Figure 10: The additional rules for label samples.

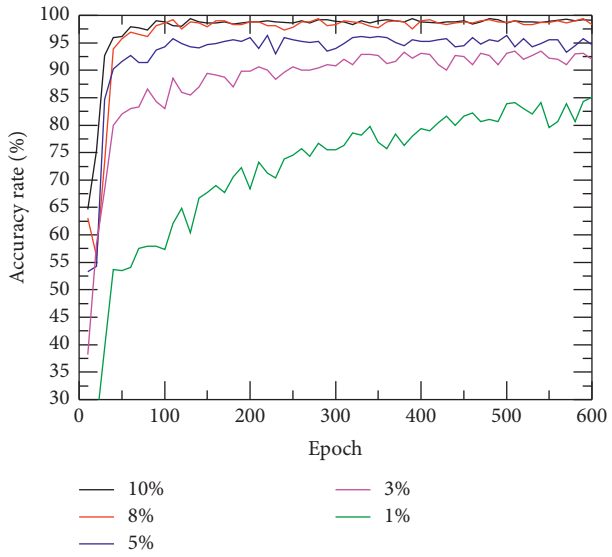

Figure 11: Accuracy of different labeled samples with iteration steps.

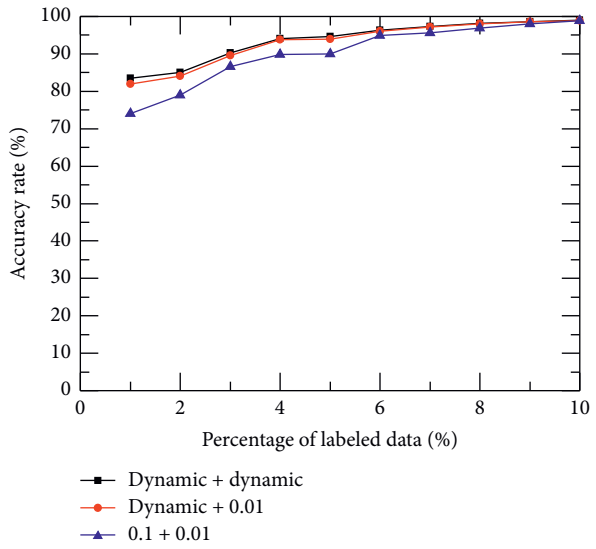

Figure 12: The accuracy curve of two kinds of learning rate accuracy with different percentages of labeled samples.

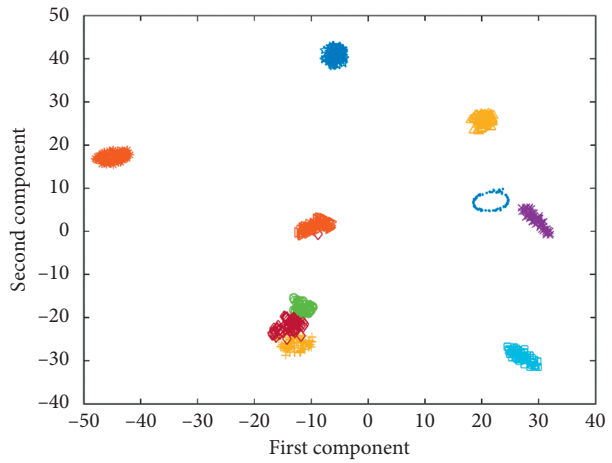

Figure 13: Visualization of 1% labeled sample.

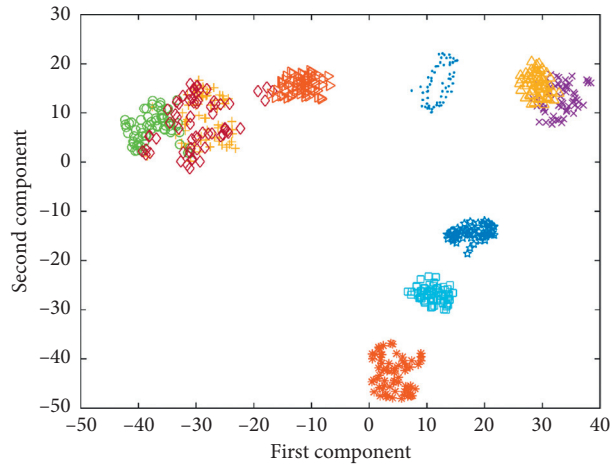

Figure 14: Visualization of 5% labeled sample.

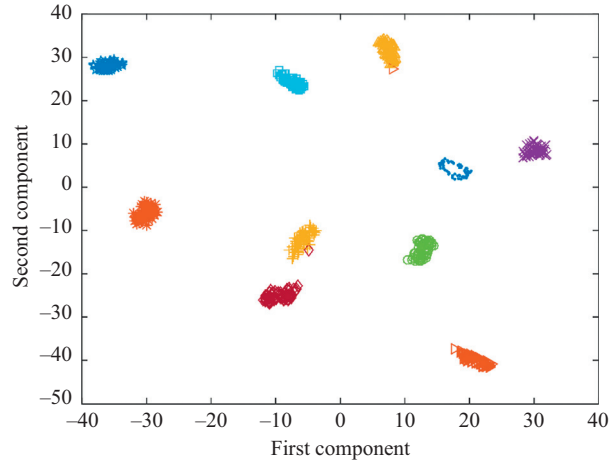

Figure 15: Visualization of 8% labeled sample.

distinguished. Figure 13 is the visualization of 5% labeled samples. Seven kinds of faults can be distinguished clearly. But there are no clear boundaries among the remaining three faults. Figure 14 is the visualization of 8% labeled samples. All of the ten kinds of faults can be easily distinguished.

5. Conclusions

In this manuscript, a novel SAE model using dynamic learning rate is developed for bearing fault diagnosis, which can effectively overcome the shortcomings of the fixed learning rate. In order to verify the performance, the proposed method is applied on a typical bearing fault data set. According to the positive and negative value of the training error gradient, different learning rates updating strategies are used. In addition, the optimal network structure and the optimal percentage of labeled samples are determined through comparative experiments. The results show that the dynamic learning rate method can improve the accuracy and convergence ability of the network, and the influence of the initial learning rate is very small.

Data Availability

The data used to support the findings of this study have been deposited in http://csegroups.case.edu/bearingdatacenter/home.

Conflicts of Interest

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

Authors' Contributions

All of the authors contributed equally to the conception of the idea, the design of experiments, the analysis and interpretation of results, and the writing of the manuscript. W.T. and H.P. wrote the original draft; H.P., J.X., and M.B

reviewed and edited the article. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant no. 51809082) and the National Key R&D Program of China (Grant no. 2019YFE0105200).

References

[1] E. Zio, "Reliability engineering: old problems and new challenges," *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 125–141, 2009.

[2] Y. Liu, X. Yan, C.-a. Zhang, and W. Liu, "An ensemble convolutional neural networks for bearing fault diagnosis using multi-sensor data," *Sensors*, vol. 19, no. 23, p. 5300, 2019.

[3] Y. Zhang, X. Li, L. Gao, and P. Li, "A new subset based deep feature learning method for intelligent fault diagnosis of bearing," *Expert Systems with Applications*, vol. 110, pp. 125–142, 2018.

[4] S. He, J. Chen, Z. Zhou, Y. Zi, Y. Wang, and X. Wang, "Multifractal entropy based adaptive multiwavelet construction and its application for mechanical compound-fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 76-77, pp. 742–758, 2016.

[5] Y. Jian, X. Qing, L. He, Y. Zhao, X. Qi, and M. Du, "Fault diagnosis of motor bearing based on deep learning," *Advances in Mechanical Engineering*, vol. 11, no. 9, pp. 8–14, 2019.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," pp. 4278–4284, 2016, https://arxiv.org/abs/1602.07261.

[8] P. Martinez, M. Al-Hussein, and R. Ahmad, "A scientometric analysis and critical review of computer vision applications for construction," *Automation in Construction*, vol. 107, Article ID 102947, 2019.

[9] H. Deng, L. Zhang, and X. Shu, "Feature memory-based deep recurrent neural network for language modeling," *Applied Soft Computing*, vol. 68, pp. 432–446, 2018.

[10] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs," *Computer Vision and Image Understanding*, vol. 176-177, pp. 22–32, 2018.

[11] R. Bhargava and Y. Sharma, "Deep extractive text summarization," *Procedia Computer Science*, vol. 176-177, pp. 22–32, 2017.

[12] T. Boudaa, M. E. Marouani, and N. Enneya, "Alignment based approach for Arabic textual entailment," *Procedia Computer Science*, vol. 148, pp. 246–255, 2019.

[13] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[14] V. Suárez-Paniagua, R. M. Rivera Zavala, I. Segura-Bedmar, and P. Martínez, "A two-stage deep learning approach for extracting entities and relationships from medical texts," *Journal of Biomedical Informatics*, vol. 99, Article ID 103285, 2019.

[15] C. Krauss, X. A. Do, and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500," *European Journal of Operational Research*, vol. 259, no. 2, pp. 589–702, 2017.

[16] O. Lachiheb and M. S. Gouider, "A hierarchical deep neural network design for stock returns prediction," *Procedia Computer Science*, vol. 126, pp. 264–272, 2018.

[17] A. Boloor, K. Garimella, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Attacking vision-based perception in end-to-end autonomous driving models," *Journal of Systems Architecture*, vol. 110, Article ID 101766, 2020.

[18] R. Sarcinelli, R. Guidolini, V. B. Cardoso et al., "Handling pedestrians in self-driving cars using image tracking and alternative path generation with Frenét frames," *Computers & Graphics*, vol. 84, pp. 173–184, 2019.

[19] B. Niu, C. Liang, Y. Lu et al., "Glioma stages prediction based on machine learning algorithm combined with protein-protein interaction networks," *Genomics*, vol. 112, no. 1, pp. 837–847, 2020.

[20] S. S. Zhao, S. K. Singh, and S. G. Bhirud, "A bearing data analysis based on kurtogram and deep learning sequence models," *Measurement*, vol. 145, pp. 665–677, 2019.

[21] D.-T. Hoang and H.-J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, 2019.

[22] S. Ma and F. Chu, "Ensemble deep learning-based fault diagnosis of rotor bearing systems," *Computers in Industry*, vol. 105, pp. 143–152, 2019.

[23] W. Liu, P. Guo, and L. Ye, "A low-delay lightweight recurrent neural network (LLRNN) for rotating machinery fault diagnosis," *Sensors-Basel*, vol. 19, no. 14, 2019.

[24] Y. Chang, J. Chen, C. Qu, and T. Pan, "Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels," *Renewable Energy*, vol. 153, pp. 205–213, 2020.

[25] L. Chen, G. Xu, Q. Zhang, and X. Zhang, "Learning deep representation of imbalanced SCADA data for fault detection of wind turbines," *Measurement*, vol. 139, pp. 370–379, 2019.

[26] R. Chen, X. Huang, L. Yang, X. Xu, X. Zhang, and Y. Zhang, "Intelligent fault diagnosis method of planetary gearboxes based on convolution neural network and discrete wavelet transform," *Computers in Industry*, vol. 106, pp. 48–59, 2019.

[27] Z. Sun, H. Jin, J. Gu, Y. Huang, X. Wang, and X. Shen, "Gradual fault early stage diagnosis for air source heat pump system using deep learning techniques," *International Journal of Refrigeration*, vol. 107, pp. 63–72, 2019.

[28] Y. Ding, L. Ma, J. Ma et al., "Intelligent fault diagnosis for rotating machinery using deep Q-network based health state classification: a deep reinforcement learning approach," *Advanced Engineering Informatics*, vol. 42, Article ID 100977, 2019.

[29] K. Kan, Y. Zheng, H. Chen et al., "Numerical simulation of transient flow in a shaft extension tubular pump unit during runaway process caused by power failure," *Renewable Energy*, vol. 154, pp. 1153–1164, 2020.

[30] C. Zhou, S. Wang, Y. Liu, and C. Liu, "A novel RNN based load modelling method with measurement data in active distribution system," *Electric Power Systems Research*, vol. 166, pp. 112–124, 2019.

[31] V. N. Nguyen, R. Jenssen, and D. Roverso, "Automatic autonomous vision-based power line inspection: a review of current status and the potential role of deep learning," *International Journal of Electrical Power & Energy Systems*, vol. 99, pp. 107–120, 2018.

[32] L. Yang, Y. Li, and Z. Li, "Improved-ELM method for detecting false data attack in smart grid," *International Journal of Electrical Power & Energy Systems*, vol. 91, pp. 183–191, 2017.

[33] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, 2016.

[34] S. Pang and X. Yang, "A cross-domain stacked denoising autoencoders for rotating machinery fault diagnosis under different working conditions," *IEEE Access*, vol. 7, p. 1, 2019.

[35] 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, pp. 88–98, 2019.

[36] A. Senior, G. Heigold, M. Ranzato et al., "An empirical study of learning rates in deep neural networks for speech recognition," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6724–6728, IEEE, Vancouver, BC, Canada, May 2013.

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[38] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[39] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: a benchmark study," *Mechanical Systems and Signal Processing*, vol. 64-65, pp. 100–131, 2015.

[40] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mechanical Systems and Signal Processing*, vol. 72-73, pp. 303–315, 2016.

[41] K. Chen, X. Zhou, J. Fang, P. Zheng, J. Wang, and T. Wu, "Fault feature extraction and diagnosis of gearbox based on EEMD and deep briefs network," *International Journal of Rotating Machinery*, vol. 2017, Article ID 9602650, 10 pages, 2017.

[42] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.