Hindawi
Shock and Vibration
Volume 2019, Article ID 2756284, 10 pages
https://doi.org/10.1155/2019/2756284

Research Article

A Novel Hierarchical Algorithm for Bearing Fault Diagnosis Based on Stacked LSTM

Lu Yu, Jianling Qu, Feng Gao, and Yanping Tian

Qingdao Branch of Naval Aviation University, Qingdao, China

Correspondence should be addressed to Lu Yu; yulu_china@163.com

Received 8 April 2018; Revised 23 October 2018; Accepted 30 October 2018; Published 6 January 2019

Academic Editor: Emiliano Mucchi

Faced with severe operating conditions, rolling bearings tend to be one of the most vulnerable components in mechanical systems. Due to the requirements of economic efficiency and reliability, effective fault diagnosis methods for rolling bearings have long been a hot research topic of rotary machinery fields. However, traditional methods such as support vector machine (SVM) and backpropagation neural network (BP-NN) which are composed of shallow structures trap into a dilemma when further improving their accuracies. Aiming to overcome shortcomings of shallow structures, a novel hierarchical algorithm based on stacked LSTM (long short-term memory) is proposed in this text. Without any preprocessing operation or manual feature extraction, the proposed method constructs a framework of end-to-end fault diagnosis system for rolling bearings. Beneficial from the memorize-forget mechanism of LSTM, features inherent in raw temporal signals are extracted hierarchically and automatically by stacking LSTM. A series of experiments demonstrate that the proposed model can not only achieve up to 99% accuracy but also outperform some state-of-the-art intelligent fault diagnosis methods.

1. Introduction

Rolling bearings are vital elements in auto-manufacturing and heavy-load mechanical systems [1, 2]. Due to the harsh working conditions, any small fault that occurs to the bearings may cause fatal consequences to machines, which straightly leads to severe economic losses and casualties. Therefore, an urgent demand of detecting and recognizing faults automatically in rolling bearings as early as possible is practical and meaningful. During the past decades, with the rapid development of computer technology, scholars around the world have dedicated considerable efforts to bearing fault diagnosis, and many excellent intelligent algorithms have been proposed and utilized in practical applications.

Vibration analysis is one of the prevalent signal processing techniques which are widely used in fault diagnosis [3]. Machinery fault diagnosis with vibration signal analysis can be transformed into the framework of pattern recognition problem, which consists of three main steps: feature extraction, feature selection, and classification [4, 5]. Prevalent pattern recognition algorithms such as backpropagation neural network (BP-NN) and support vector machine (SVM) are representative ones used in rotating machinery fault diagnosis issues [6, 7]. However, signals collected by vibration sensors are usually non-stationary and complex, and what is worse, heavy background noise contributes to the difficulty of feature extraction. A labor-intensive workload or expertise knowledge is indispensable before an effective model is constructed for certain fault diagnosis issue. Therefore, it is a great challenge to learn discriminative fault features effectively and automatically.

As far as the authors are concerned, two main typical algorithms for feature extraction exist in bearing fault diagnosis: signal processing-based algorithms and artificial intelligence-based algorithms. The former, which take prior knowledge of signal and make full use of signal processing techniques, such as time-frequency analysis (TFA) [8], wavelet package transform (WPT) [9], and recently prevalent empirical mode decomposition (EMD) with its variants [10], have proven their effectiveness in many advanced achievements. Due to benefits from the maturity of signal

processing techniques, methods mentioned above have been widely used. However, lack of flexibility also limits its further improvement in recognition accuracy. Namely, selecting suitable parameters of the model, the model can be applicable in a certain fault diagnosis issue, and the variation of loads or other factors may affect the accuracy of the model. The latter, which are represented by SVM and BP-NN, have newly sprung up because of its accessibility. Without knowing internal mechanism, users can design a fault diagnosis system just based on the theory of pattern recognition without much effort. We hold the view that the easier for users, the more widespread an algorithm is used. Therefore, in this paper, we lay emphasis on intelligent methods for bearing fault diagnosis.

Recently, a large number of research achievements have promoted intelligent methods to a new level. Chine et al. [11] utilized several features and constructed an ANN-based voltaic system for fault diagnosis. 24-dimension parameters extracted by Jamadar and Vakharia [12] for describing bearing working conditions are sent to a BP neural network for fault recognition. Volterra series was utilized by Xia et al. [13] for recognizing different working conditions of rotor-bearing system, and backpropagation (BP) neural network was used as classifier for fault diagnosis. Batista et al. [14] took 13 statistical features in both time and frequency domains for description of different bearing conditions, and then employed radial-basis-based SVM for fault diagnosis. Zhang et al. [15] combined EEMD permutation entropy for input feature and SVM for fault recognition. Zheng et al. [16] employed multiscale fuzzy entropy for input features and SVM for classifier.

Although intelligent algorithms mentioned above have achieved an acceptable accuracy and been widely applied in practical engineering, two significant disadvantages cannot be avoided: (1) the diagnosis effectiveness largely depends on feature extraction which is obtained mainly by manual extraction according to the knowledge of mechanical engineering experts, and the quality of extracted and selected features plays a vital role in the performance of methods. The features that are either manually selected or handcrafted may not optimally characterize vibration signals and thus cannot fulfill a generic solution that can be used for any bearing fault data [17]. What is worse, the task of selecting the most sensitive features for different diagnosis issues is a time-consuming and labor-intensive work, which increases the burden of workers and researchers. (2) Intelligent diagnosis methods such as BP-NN and SVM are both shallow learning structures, that is to say, only one hidden layer is used for nonlinear transformation. Several research results have clearly illustrated that shallow architectures hinder the ability of learning complex nonlinear relationships among different fault diagnosis issues [18, 19]. Therefore, it is essential to establish a deep and hierarchical architecture for better feature learning in rolling bearing fault diagnosis issues.

Deep learning, also known as deep neural network (DNN), has attracted an increasing attention from scholars of various fields in recent years. The predominant superiority of deep learning is the capacity of learning complex nonlinear features, which can discover inherent structures and useful features from raw data by a layerwise learning procedure. A great number of research achievements have demonstrated its powerful potential in many fields, such as natural language processing (NLP) [20], computer vision (CV) [21], and mechanical fault diagnosis [22]. Chen et al. [23] introduced CNN-based deep learning to gearbox fault diagnosis. Several time-domain and frequency-domain features were extracted and sent to the framework of CNN, and a softmax-based classifier was used for fault diagnosis. Although CNN is used in fault diagnosis issue mentioned above, it is more like a classifier than a feature extractor. Therefore, the capacity of CNN has not been fully utilized. Guo et al. [24] took a similar approach, which made use of time, frequency, and time-frequency features as input of deep neural network, constructing an autoencoder-based DNN for whole life validation of bearings.

Long short-term memory (LSTM) [25], which is an important component of recurrent neural network (RNN), has become a hot spot recently. By utilizing spatial and temporal information inherent in raw temporal signal, which imitates brain memory of human beings, LSTM-based structure has the potential for higher accuracy in fault diagnosis issues. Also, with the advantage of selective memory mechanism, LSTM solves long-term dependency problems derived from RNN network.

In this paper, a novel fault diagnosis method named as hierarchical LSTM-based deep network is proposed for both feature learning and fault recognition of rolling bearings. Experiment results verify that the proposed method can obtain a higher accuracy without relying on manual feature extraction as well as advanced signal processing techniques. To the best of the authors' knowledge, this paper is the first attempt to perform hierarchical LSTM-based strategy in rolling bearing fault diagnosis issue, which is meaningful and pioneering. The rest of the paper is arranged as follows: Section 2 makes a brief review of LSTM theory, Section 3 illustrates proposed methods for bearing fault diagnosis, Section 4 is used for experiments, and Section 5 makes the conclusion.

2. LSTM Theory

2.1. The Origin of LSTM. Recurrent Neural Networks (RNNs) were firstly introduced to solve time sequence learning problems. Unlike traditional neural networks which are formed by multilayer perceptron that can only map input data to target vectors, RNNs have the capability of tracing back the whole history of previous inputs in principle. Like many other neural networks, backpropagation algorithm is used for training RNNs. However, faced with vanishing or exploding gradients during backpropagation period, the performance and potential of RNNs have greatly limited, which means that traditional RNNs cannot capture long-term dependencies. Therefore, LSTM is proposed to get rid of the limitation of RNNs mentioned above. Forget gates which dominate the flow of information among different cell states are utilized to avoid

long-term dependency problem [26]. To learn effective representation and nonlinear dynamic features in time-series data, LSTM is superior to traditional RNNs in that the former abandon vanishing or exploding gradient problem, which have the capability of capturing long-term dependencies.

2.2. Basic Theory of LSTM. The main idea behind LSTM lies in that a few gates that control the information flow along time axis can capture more accurate long-term dependencies at each time step. Specifically, at each time step t, hidden state h^t is updated by fusion of data at the same step x^t, input gate i^t, forget gate f^t, output gate o^t, memory cell c^t, and hidden state at last time step h^{t-1}. The updated equations are as follows:

$$\begin{aligned} i^t &= \sigma\left(W^i x^t + V^i h^{t-1} + b^i\right), \\ f^t &= \sigma\left(W^f x^t + V^f h^{t-1} + b^f\right), \\ o^t &= \sigma\left(W^o x^t + V^o h^{t-1} + b^o\right), \\ c^t &= f^t \odot c^{t-1} + i^t \odot \tanh\left(W^c x^t + V^c h^{t-1} + b^c\right), \\ h^t &= o^t \odot \tanh\left(c^t\right), \end{aligned} \tag{1}$$

where model parameters including $W \in \mathbb{R}^{d\times k}$, $V \in \mathbb{R}^{d\times d}$, and $b \in \mathbb{R}^d$ are learned during training and shared at each time step, σ is sigmoid activation function, $\odot$ means elementwise product, and k is a hyperparameter that characterizes the dimensionality of hidden layers.

Firstly, basic LSTM is utilized to deal with time-series data. And the final output, which is at endmost time step, is utilized to predict the output by a linear regression layer, as is shown in the following equation:

$$\overline{y}_i = W^r h_i^T, \tag{2}$$

where $W^r \in \mathbb{R}^{k\times z}$ and z are the dimensionality of output. In the phase of model training, the cross-entropy is used as loss function between the predicted label distribution $q(x)$ and the target label distribution $p(x)$. So the cross-entropy between $p(x)$ and $q(x)$ is

$$\text{loss} = H(p,q) = -\sum_x p(x)\log q(x). \tag{3}$$

Activation function enables the network to acquire a nonlinear representation of the input signal, which enhances the representation ability and makes the learned features more discriminative. In this text, rectified linear unit (ReLU) is adopted for fast convergence of our model. ReLU has the advantage of making the network sparser and weights more trainable during adjusting parameters. ReLU can be described in the following equation:

$$a_i^{l+1}(j) = f\left(y_i^{l+1}(j)\right) = \max\left\{0, y_i^{l+1}(j)\right\}, \tag{4}$$

where $y_i^{l+1}(j)$ and $a_i^{l+1}(j)$ represent the output of LSTM and activation value of $y_i^{l+1}(j)$, respectively.

The corresponding LSTM unit architecture is shown in Figure 1.

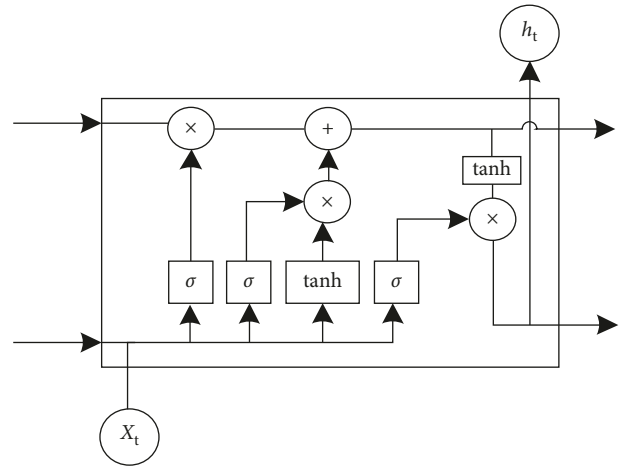

Figure 1: The internal structure of an LSTM unit.

2.3. Hierarchical LSTM. With the rapid development of computer hardware and a series of deep learning algorithms being put forward, deep architectures have shown their powerful capability in feature self-learning. Therefore, stacking several LSTM layers for a deep LSTM-based neural network is meaningful. The main idea of deep neural network is that many nonlinear mapping layers between inputs and outputs are utilized for hierarchically feature learning. As is shown in Figure 2, the output of hidden layer is not only propagated forward through time, but also used as one of inputs of next LSTM hidden layer. Therefore, the l-th layer can be updated by the following equations:

$$\begin{aligned} i_l^t &= \sigma\left(W_l^i h_{l-1}^t + V_l^i h_l^{t-1} + b_l^f\right), \\ f_l^t &= \sigma\left(W_l^f h_{l-1}^t + V_l^f h_l^{t-1} + b_l^f\right), \\ o_l^t &= \sigma\left(W_l^o h_{l-1}^t + V_l^o h_l^{t-1} + b_l^o\right), \\ c_l^t &= f_l^t \odot c_l^{t-1} + i_l^t \odot \tanh\left(W_l^c h_{l-1}^t + V_l^c h_l^{t-1} + b_l^c\right), \\ h_l^t &= o_l^t \odot \tanh\left(c_l^t\right). \end{aligned} \tag{5}$$

The input of 1st layer is raw temporal signals, i.e., $h_0^t = x^t$, while the output of the 1st layer is an abstraction of raw signals, which is regarded as a hierarchical feature. Other LSTM layers use the output of previous layer as input, and the output of last LSTM is sent to a full-connect layer for classification. The advantages of stacked LSTM are obvious: (1) stacking LSTM layers enables the model to learn characteristics of raw temporal signal from different aspects at each time step. (2) Model parameters are distributed over the whole space of the model without increasing memory capacity, which enables the model to accelerate convergence and refine nonlinear operations of raw data.

Note that LSTM neural network has the mechanism of recalling memory with time steps. As for one-dimensional signal processing, a signal with limited length can be reformed into a matrix with rows for input dimension and columns for time steps. It is intuitive that LSTM imitates the memory process as human beings do, which means it can memorize a signal line by line and catch important points

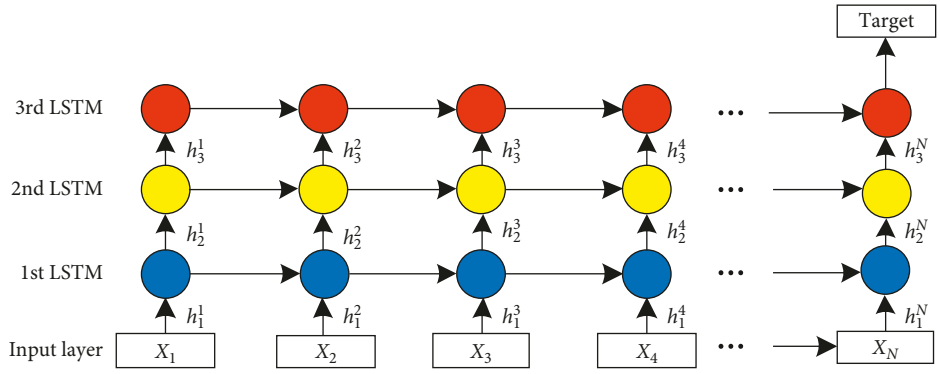

Figure 2: The hierarchical structure of LSTM.

inherent in raw temporal signal automatically. A deep LSTM neural network may help reinforce the whole process.

3. Structure of Proposed Method

Based on the study of rolling bearing fault diagnosis of this paper, a hierarchical structure of LSTM is proposed. Figure 3 depicts the flow chart of proposed method, which consists of three parts: (1) data augmentation: training dataset is one of the most important factors of all deep learning methods. Raw data of each condition is one dimension temporal signal, hence data augmentation strategy aims to enlarge training datasets by dividing raw signal with overlap, which helps to reduce computation cost and accelerate model convergence. (2) Model training: based on built model structure, the input data are divided into two groups with their corresponding labels: training dataset and testing dataset. In other words, the proposed method is a supervised training process with unsupervised feature learning. A dropout strategy [27] is adopted after each LSTM to avoid overfitting for better generalization. (3) Evaluation with testing dataset: after the model being trained, a test dataset is utilized to validate the effectiveness of the model, and the evaluation indicator is the accuracy of classification obviously.

All steps above form the main framework of the proposed hierarchical LSTM neural network. A series of experiments will be performed in the following section.

4. Experiment Analysis and Discussion

As mentioned above, rolling bearings are essential elements of rotating machinery, and recognizing their faults as timely as possible has great effect on the reliability and performance of machinery that they are mounted on. Therefore, a Case Western Reserve University (CWRU) dataset for rolling bearings with different fault rolling bearing conditions is adopted in our experiments [28]. The performance of proposed hierarchical LSTM neural network is compared with some existing state-of-the-art diagnosis algorithms, with details listed in the following subsections.

4.1. Introduction of CWRU Dataset. The CWRU dataset has been regarded as a benchmark for testing algorithms related to vibration signal analysis of rolling bearings. The CWRU datasets consist of vibration time-series of various rolling bearing conditions which are generated by a test rig, which is shown in Figure 4. The test rig is composed of a 2-horsepower (hp) motor for driving a shaft, a control circuit model for controlling various speeds to meet different requirements, and a torque converter for signal processing. The sampling frequency of the accelerometers is 12 kHz.

In our current experiment, the adopted vibration data are collected from accelerometers mounted on the housing with magnetic bases and installed at the 12 o'clock position for the bearings.

4.2. Experiment Setup. In our experiments, 13 kinds of health conditions with 1 hp are considered. All condition samples are segmented with length 256 and overlap 50%, and each condition has 300 samples, half for training and half for testing. Detailed information about experiment samples is listed in Table 1.

Also, some other key parameters used in proposed model are listed as follows: the input layer has 256 units which is equal to the dimension of input sample, the hidden units of 1st LSTM to 3rd LSTM are 64, 32, and 32 respectively, and a RMSprop optimization algorithm [29] is used to train the model. Mean square error (MSE) is an indicator for evaluation performance. In the output layer, a softmax classifier is used for classification.

4.3. Experiment Results. For better and fair experiment results, a random selection strategy has been adopted for all samples. Namely, 150 samples of each health condition are randomly selected for training, while the remaining for testing. Each raw temporal signal is 256 for balancing information coverage and computing efficiency. It is worth noting that an "early stopping" strategy has been introduced in the proposed method during training phase for better generalization performance. Even though we set a max iteration 100, the training will stop if the loss of training dataset does not change much for several iterations. Accuracy and loss of our model during training phase are plotted in Figure 5. From the curves, we can clearly see a convergence after 43 iterations; obviously it greatly reduces time and cost for training.

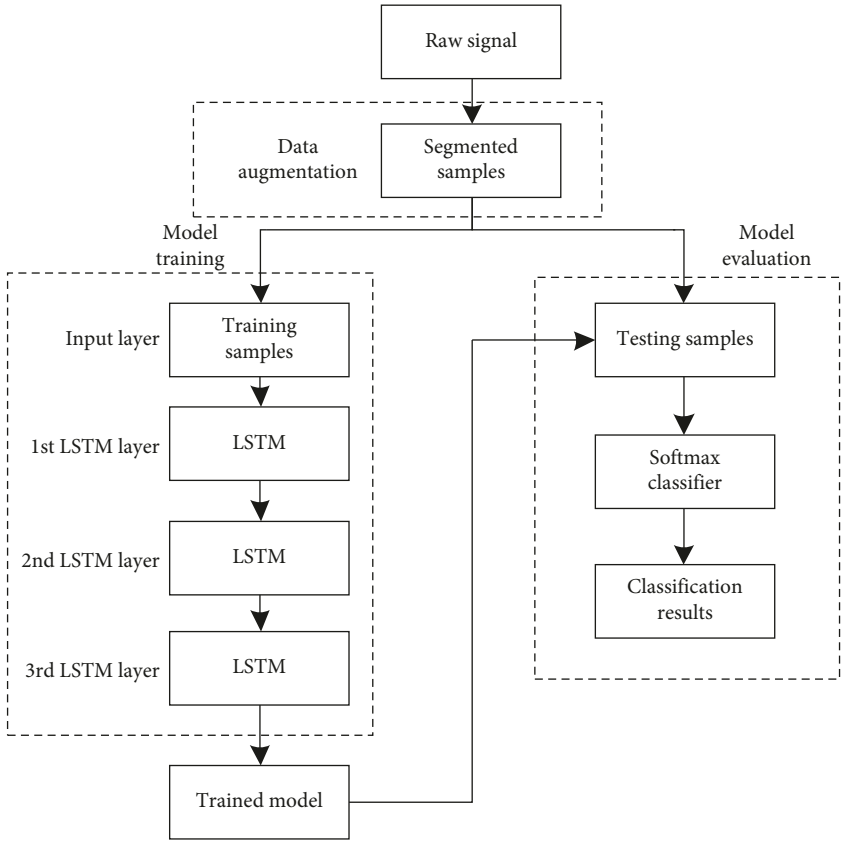

Figure 3: The flow chart of the proposed method.

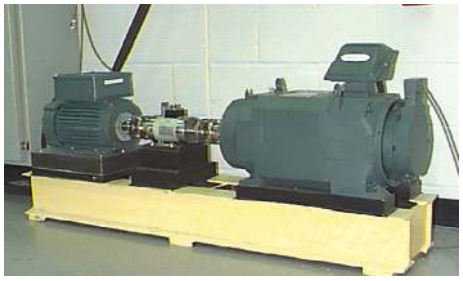

Figure 4: Test rig of rolling bearings.

To better illustrate experiment results, a multiclass confusion matrix for the third trail of proposed method is shown in Figure 6. The multiclass confusion matrix is an exhibition method for visualizing classification results of all conditions in detail, which consists of classification accuracy and misclassification error. The ordinate and horizontal axis of a confusion matrix refers to predicted label and true label, respectively.

Obviously, most faults have 100% accuracy in our model. The worst accuracy, 94%, occurs in outer ring fault with 21 inch at 6 o'clock position. The total average accuracy is up to 98.8%, which demonstrates the efficiency and feasibility of the proposed method.

4.4. Comparison with Other Methods. For comparison, four other methods which consist of 1-layer LSTM neural network, backpropagation neural network (BP-NN), SVM, and CNN have been considered in our experiments. The detailed parameter settings of other methods in the experiment are depicted as follows: (1) 1-layer LSTM neural network: the architecture is the same as the 1st layer of proposed method. (2) BP-NN: it is also an "end-to-end" neural network with a 32-unit hidden layer. The whole structure is 256-32-13. (3) SVM: the feature set includes time-domain features (RMS, kurtosis, skewness, variance, standard deviation, etc.), frequency-domain features (frequency

Table 1: Experiment samples.

Condition	Label	Sample length	Sample number
Normal	0	256	300
IF—7 inch	1	256	300
BF—7 inch	2	256	300
OF@6—7 inch	3	256	300
OF@3—7 inch	4	256	300
OF@12—7 inch	5	256	300
IF—14 inch	6	256	300
BF—14 inch	7	256	300
OF@6—14 inch	8	256	300
IF—21 inch	9	256	300
OF@6—21 inch	10	256	300
OF@3—21 inch	11	256	300
OF@12—21 inch	12	256	300

Note. IF, OF, and BF mean inner, outer, and ball faults, respectively. 7 inch means the diameter of faults, and so on. @6 means the location of faults in outer fault, and so on.

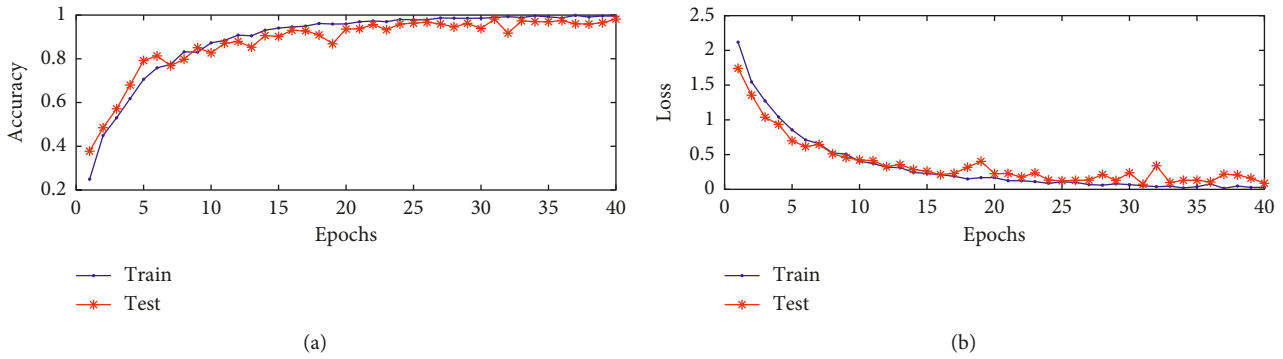

Figure 5: Training and testing curve for accuracy and loss.

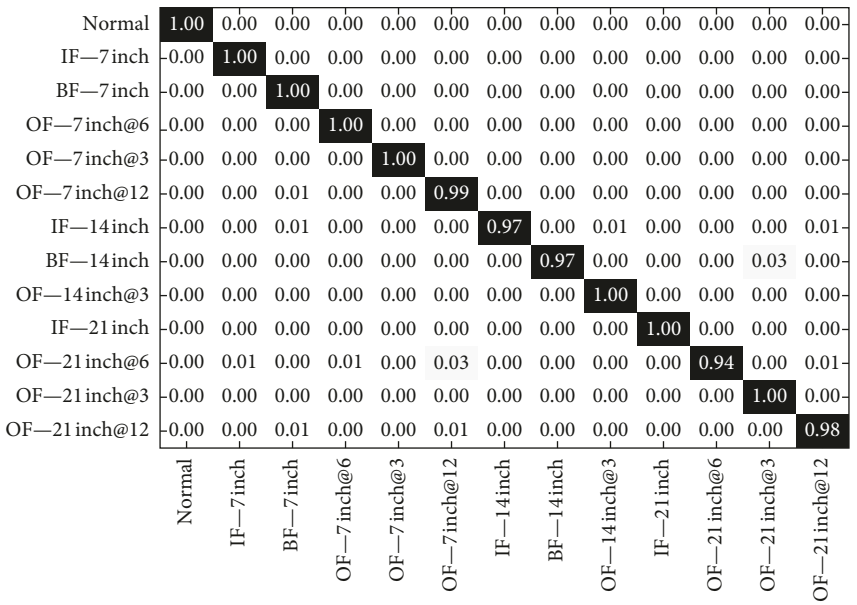

Figure 6: Confusion matrix for each category.

bands), and time-frequency domain features (3-level wavelet package coefficients) which are utilized as the input of SVM. Radius basis function is used for kernel function with penalty factor 50 and radius of kernel function 0.02. (4) CNN: the CNN method used in our experiment is that referred in Reference [23], which has a structure of input layer, 2 convolutional layers, and 2 pooling layers. Time-domain and frequency-domain statistical features are sent to input layer after transformed to 2D format, the shape of input feature map is 32×32 with 6 kernels, and max pooling size is set to 2 with learning rate 0.1 and maximum iteration 100.

Figure 7 shows diagnosis results of all methods in 10 trails. It is clear that the proposed method has the highest recognition accuracy among all methods with average accuracy up to 98.65%. 1-layer LSTM neural network has the second highest accuracy partly because it has memory mechanism, which is derived from LSTM. However, its shallow structure hinders its accuracy from improvement. It is worth noting that BP-NN has the worst accuracy among all methods. BP-NN does not own memory mechanism, and it just uses information for forward propagation, which lacks the capacity of learning useful information in previous data points. Also, a shallow structure limits its performance.

In order to graphically display the performance of our hierarchical LSTM neural network, T-SNE [30] has been utilized for visualizing each layer of our model. After selecting the first two important components obtained by T-SNE, the outputs of all layers are shown in Figure 8. From Figure 8, a clear and intuitive conclusion has arrived: from input layer to output layer, the distribution of each category has been shown more and more clearly. Namely, input layer mixed all categories together, in which we cannot distinguish any category. After the first layer of LSTM, categories of No. 0, No. 4, No. 10, and No. 12 have converged into their own spaces. With the progress of deeper layers, a clearer distinguishability of each category can be obtained. Finally, in the output of the third LSTM layer, each category almost gets its own space in the 2D image, which demonstrates the availability of our model.

4.5. Influence of Some Hyperparameters. In our proposed model, two hyperparameters need to be discussed, namely the

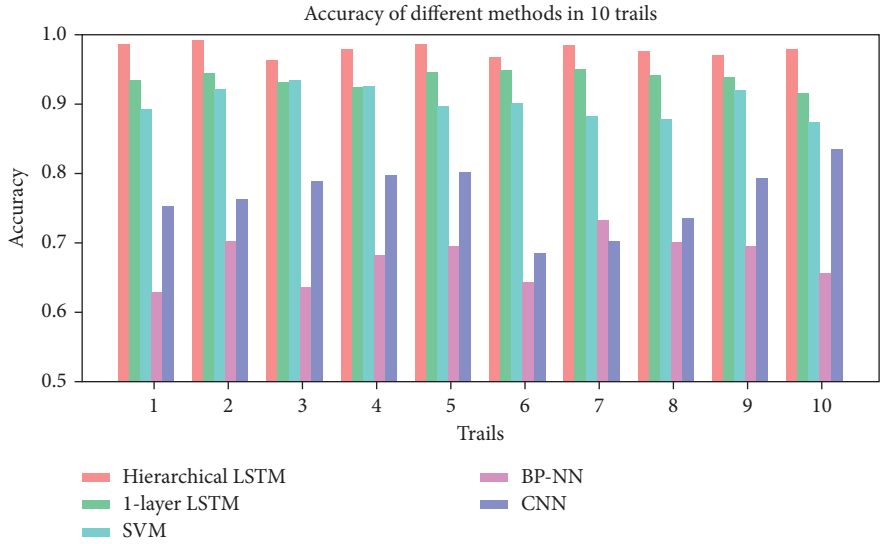

Figure 7: Experimental results of 10 trails.

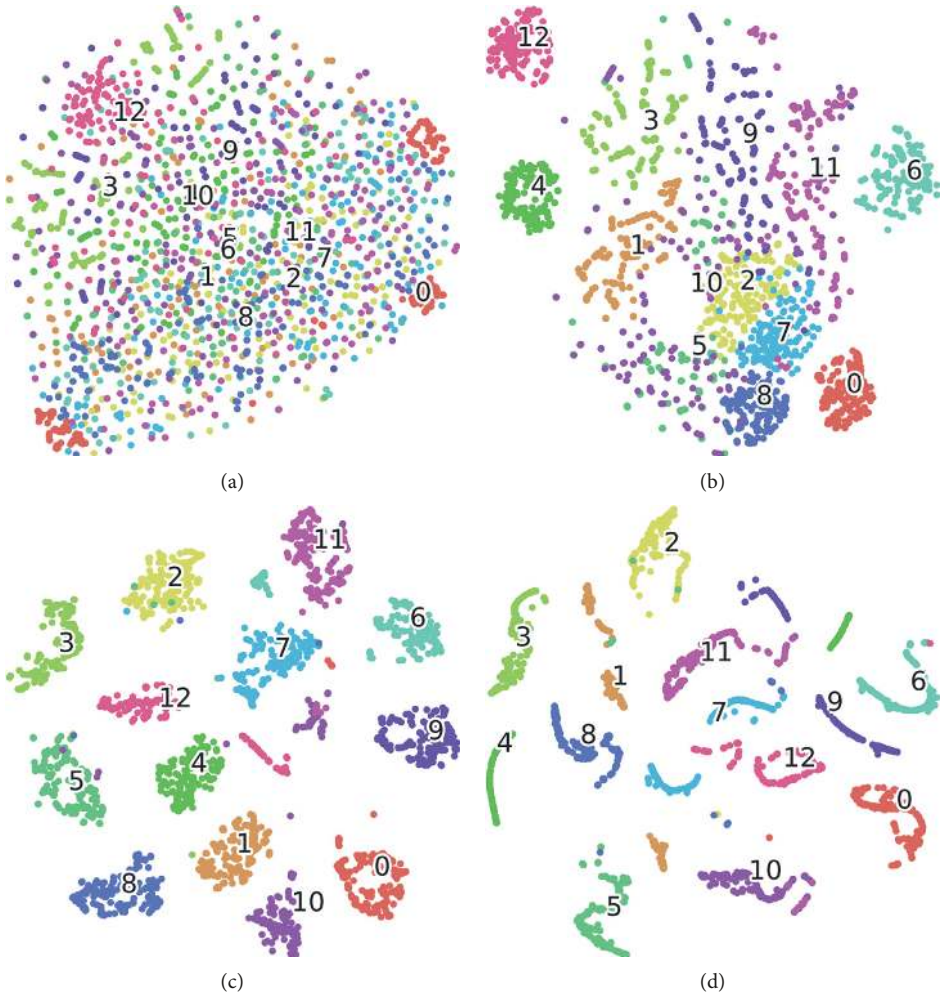

Figure 8: Layerwise visualization based on T-SNE. (a) Input layer, (b) 1st LSTM layer, (c) 2nd LSTM layer, and (d) 3rd LSTM layer.

input size of samples and the layer number of LSTM neural networks. We have conducted experiments for both hyper-parameters, and the experiment results are shown in Figure 9.

Figure 9(a) shows the influence of input size on the performance of proposed method. The size of the input units is set to 32, 64, 128, 256, 512, 768, and 1024, respectively. It is

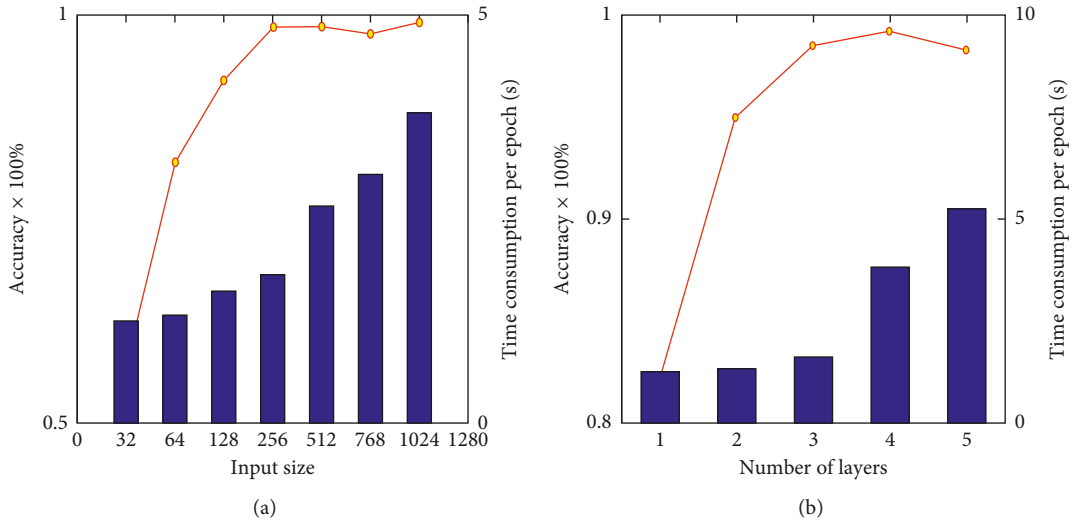

Figure 9: Experimental results of hyperparameters. (a) Input size. (b) Number of LSTM layers.

obvious that (1) samples with larger input size tend to accomplish better diagnosis performance than fewer ones. It is probably that samples with larger input size contain more fault information which is essential to feature learning, while fewer data points may easily ignore the local spatial information inherent in the raw data. (2) The time cost increases rapidly with input size, which is not suitable for online applications. We choose the input size 256 for compromise. A similar conclusion can be reached in the Figure 9(b), which shows the layer number of LSTM versus accuracy and computation time, and we choose layer number 3 in our model for the balance of accuracy and computation cost.

4.6. Generalization Experiments. In order to investigate generalization capacity of the proposed method, we form a testing dataset by taking samples under 2 hp load and 3 hp load corresponding to Section 4.1 samples. Similarly, we use confusion matrix to illustrate our results, which is shown in Figure 10. Although some accuracies fall in the variations of loads, the proposed method still achieves 97.8% and 98.4% accuracy in total. Mentioning that the worst fault recognition accuracy in our generalization experiments is 90.0%, it is still higher than some intelligent methods in 1 hp load such as SVM, BP-NN, and CNN in our comparative experiments conducted above, which clearly demonstrates the efficiency and superiority of the proposed method.

5. Discussion

Through various experiments conducted above, we can safely conclude that the proposed stacked LSTM neural network for self-learning method is able to adaptively mine inherent fault characteristics and effectively identify faults with high diagnosis accuracy. The prominent superiority of proposed method is that the features are extracted by deep structure in a more identifiable way than extracted by hand-engineered or prior knowledge, which makes it easier to apply to other diagnosis issues.

However, the proposed method also has some shortcomings which need to be improved in the near future. (1) The computation cost of proposed method is relatively higher than traditional ones such as SVM or BP-NN. Part of the reason is for the limitation of computer hardware used in our method. We believe that this defect can be perfectly solved by the hardware improvement in the future. (2) The parameter selection of our method needs consecutive trial-and-error experiments. Some necessary experiments need to be conducted before a suitable model constructed for a certain fault diagnosis issue. So far, no perfect solution to this problem has been proposed yet. We just follow a simple idea which has been introduced by many other scholars that the input length should contain several whole cycles of raw temporal signal and the number of hidden units should be no larger than the previous one. In practice, the principle works well in our model.

6. Conclusions

The proposed method of fault diagnosis for rolling bearings based on stacked LSTM neural networks is novel and promising. It has three main advantages that other traditional methods do not possess: (1) it gets rid of dependencies of handcrafted features or advanced signal processing techniques which are essential for traditional methods. (2) The learning process of LSTM is performed automatically based on raw temporal signal without any prior knowledge of signal types or inherent mechanism. Thanks to the memory capability of LSTM, the correlation within signal is further strengthened. (3) Based on stacked architecture, features are extracted hierarchically, and the deeper structure gives the learning model more potential for mining inherent characteristics.

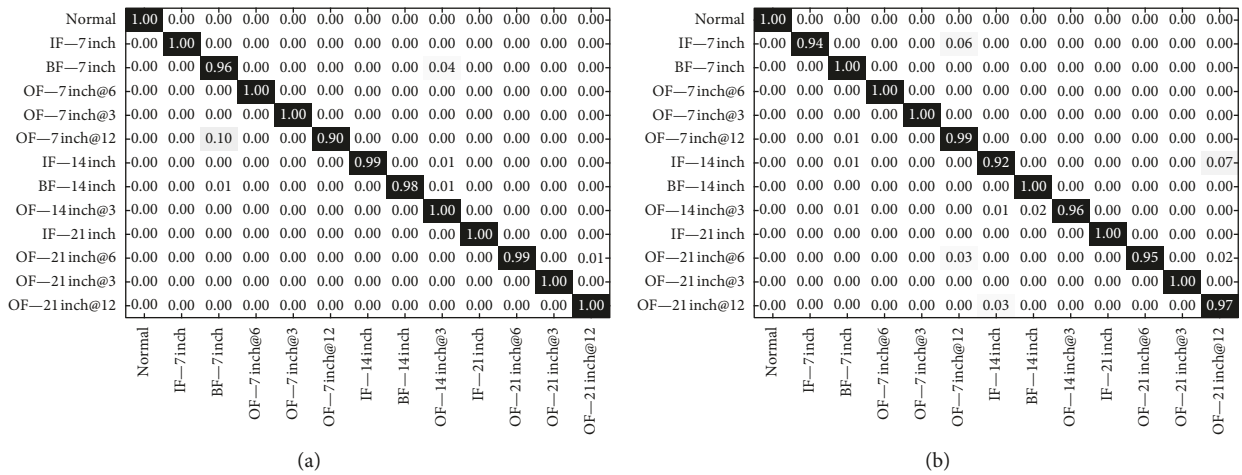

Figure 10: Confusion matrix for generalization experiments. (a) 2 hp results. (b) 3 hp results.

All of the above demonstrate the efficiency and availability of the proposed method. However, the computation cost still has room for improvement, which will be the focus of our future work.

Data Availability

All data used in this paper are from the open-source rolling bearing datasets of CWRU (http://csegroups.case.edu/bearingdatacenter/pages/download-data-file).

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 the National Natural Science Foundation of China (Grant no. 51505491) and Aeronautical Science Foundation of China (Grant no. 20165853040), and the authors would like to thank Professor K. A. Loparo of Case Western Reserve University for his kind permission to use their bearing data.

References

[1] H. Jiang, C. Li, and H. Li, "An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 36, no. 2, pp. 225–239, 2013.

[2] Y. Li, X. Liang, and M. J. Zuo, "Diagonal slice spectrum assisted optimal scale morphological filter for rolling element bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 85, pp. 146–161, 2017.

[3] Y. Lei, Z. He, and Y. Zi, "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9941–9948, 2009.

[4] M. D. Prieto, G. Cirrincione, A. G. Espinosa, J. A. Ortega, and H. Henao, "Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 8, pp. 3398–3407, 2013.

[5] X. Jin, M. Zhao, T. W. S. Chow, and M. Pecht, "Motor bearing fault diagnosis using trace ratio linear discriminant analysis," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 5, pp. 2441–2451, 2014.

[6] H. Pecht and P. Chen, "Intelligent diagnosis method for rolling element bearing faults using possibility theory and neural network," *Computers and Industrial Engineering*, vol. 60, no. 4, pp. 511–518, 2011.

[7] Y. Lei, Z. Liu, X. Wu, N. Li, W. Li, and J. Lin, "Health condition identification of multi-stage planetary gearboxes using a mRVM-based method," *Mechanical Systems and Signal Processing*, vol. 60-61, pp. 289–300, 2015.

[8] X. Ding and Q. He, "Time-frequency manifold sparse reconstruction: a novel method for bearing fault feature extraction," *Mechanical Systems and Signal Processing*, vol. 80, pp. 392–413, 2016.

[9] J. Chen, Z. Li, J. Pan et al., "Wavelet transform based on inner product in fault diagnosis of rotating machinery: a review," *Mechanical Systems and Signal Processing*, vol. 70-71, pp. 1–35, 2016.

[10] Z. He, J. Chen, Y. Zi, and J. Pan, "Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive," *Mechanical Systems and Signal Processing*, vol. 85, pp. 512–529, 2017.

[11] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, and A. Malek Pavan, "A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks," *Renewable Energy*, vol. 90, pp. 501–512, 2016.

[12] I. M. Jamadar and D. P. Vakharia, "A novel approach integrating dimensional analysis and neural networks for the detection of localized faults in roller bearings," *Measurement*, vol. 94, pp. 177–185, 2016.

[13] X. Xia, J. Zhou, J. Xiao, and H. Xiao, "A novel identification method of Volterra series in rotor-bearing system for fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 66-67, pp. 557–567, 2016.

[14] L. Batista, B. Badri, R. Sabourin, and M. Thomas, "A classifier fusion system for bearing fault diagnosis," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6788–6797, 2013.

[15] X. Zhang, Y. Liang, J. Zhou, and Y. zang, "A novel bearing fault diagnosis model integrated permutation entropy,

ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164–179, 2015.

[16] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mechanical Systems and Signal Processing*, vol. 85, pp. 746–759, 2017.

[17] H. Shao, H. Jiang, H. Zhang, W. Duan, T. Liang, and S. Wu, "Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing," *Mechanical Systems and Signal Processing*, vol. 100, pp. 743–765, 2018.

[18] E. D. L. Rosa and W. Yu, "Randomized algorithms for nonlinear system identification with deep learning modification," *Information Sciences*, vol. 364-365, pp. 197–212, 2016.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[20] Y. Goldberg, "A primer on neural network models for natural language processing," *Computer Science*, 2015, http://arxiv.org/abs/1510.00726.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[22] O. Janssens, V. Slavkovikj, B. Vervisch et al., "Convolutional neural network based fault detection for rotating machinery," *Journal of Sound and Vibration*, vol. 377, pp. 331–345, 2016.

[23] Z. Q. Chen, C. Li, and R. V. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock and Vibration*, vol. 2015, Article ID 390134, 10 pages, 2015.

[24] L. Guo, H. L. Gao, Y. W. Zhang et al., "Research on bearing condition monitoring based on deep learning," *Journal of Vibration and Shock*, vol. 35, no. 12, pp. 166–171, 2016.

[25] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] N. Srivastava, G. Hinton, A. Krizhevsky et al., "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] K. A. Loparo, *Case Western Reserve University Bearing Data Center*, 2012, http://csegroups.case.edu/bearingdatacenter/home.

[29] T. Tieleman and G. Hinton, "Lecture 6.5—RmsProp: divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, 2012.

[30] G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2, pp. 2579–2605, 2008.