Fault Diagnosis of Wheelset Bearings using Deep Bidirectional Long Short-term Memory Network

Zhixing Zhu[1], Huan Wang[2], Zhiliang Liu[2], Shaoyu Meng[1]
1: Glasgow College 2: School of Mechanical and Electrical Engineering
University of Electronic Science and Technology of China
Chengdu, PR China

***Abstract*—The wheelset bearing is the core component of a railway bogie, whose health status is critical to guarantee the safety of high-speed trains (HSTs). However, the traditional data-driven fault diagnosis methods are mostly founded on the feature extraction methods requiring extensive domain knowledge and experience, compared with deep learning approaches that can learn hierarchical representations automatically from data. Recurrent neural network (RNN) which can learn features from sequential data directly without any feature engineering has been proved to be effective in the research area of machine failure diagnosis. Therefore, an innovative fault diagnosis method for the wheelset bearings in the HSTs using Deep Bidirectional Long Short-term Memory Network (DBLSTM) is proposed in this paper. Long Short-term Memory Network (LSTM), as an improved framework of Recurrent Neural Network, is able to overcome the gradient vanishing or exploding problem, capturing long-term dependencies effectively. In the DBLSTM, the bidirectional structure is applied to enhance the performance of the LSTM Network by capturing temporal information from both future and past contexts of input sequential data. In addition, by stacking multiple Bidirectional LSTM layers to build the DBLSTM network, for the vibration signals measured under the complicated environment, more complex fault features can be effectively learned. With the increasement of the depth of the network, the regularization method using recurrent dropout technique is introduced to relieve the problem of overfitting and enhance the generalization ability of the deep network. The DBLSTM is assessed on the dataset of the wheelset bearings of the HSTs under different operating states. The experimental results indicate that the DBLSTM can accurately diagnose the fault mode of the wheelset bearings of the HSTs under strong noise, outperforming five existing deep learning methods used in fault diagnosis.**

Keywords- fault diagnosis; wheelset bearing ; Recurrent neural network; Bidirectional Long Short-term Memory network

I. INTRODUCTION

The past decades have witnessed the rapid development of high-speed trains (HSTs) throughout the world. However, with the increase of running speed, preserving the safety and reliability of the HSTs is facing enormous challenges. The wheelset bearing is the crucial but easily damaged component of a railway bogie. The failure of the bearing may lead to the malfunction of machine and cause huge economic loss or even heavy casualties [1]. Therefore, it is significant to carry out the fault diagnosis for wheelset bearings to ensure the safe operation of the HSTs.

Recently, with the development of sensor network and computer technology, the data-driven fault diagnosis methods for machine health monitoring have drawn many researchers' attention. In data-driven method, the vibration signals of wheelset bearings are collected by sensors in the form of discrete time sequence. The key steps in data-driven models are feature extraction and representation learning from these sensor data. However, the working environment of the HSTs is complex and variable, leading sensor data to be easily affected by noise and making the data-based fault diagnosis of these wheelset bearings become more challenging.

The conventional intelligent data-based fault diagnosis methods consist of three major parts: data acquisition, feature extraction and fault identification. More specifically, it firstly takes time-series sensor data as input, then feature selection and extraction are applied to obtain representation of machine condition features, including time domain features (e.g. variance, skewness, kurtosis [2], entropy [3] and so on), frequency domain features (e.g. spectral [4]) and time-frequency domain features (e.g. wavelet packet [5], Hilbert spectrum [6] and so on). Then, those machine condition features in the form of characteristic parameters are processed with machine learning algorithms for fault identification, like support vector machine (SVM) [6], k-Nearest neighbor distance analysis[2] and artificial neural network (ANN) [3].

However, these conventional intelligent data-based fault diagnosis methods mentioned above require intensive expert knowledge and complicated feature engineering. This means the performance of the fault diagnosis is heavily dependent on the feature extraction method, which usually requires rich domain knowledge and practical experience. Furthermore, designing the feature extraction methods artificially is generally time consuming and difficult, which also cannot guarantee good performance. This is because the complex nonlinear relationship exists between the vibratory signals of wheelset bearings with strong noise and the corresponding fault types, making the fault features become extremely complicated. In addition, the fault features are variable due to the working conditions and external environment changing all the time during the equipment's operation.

Corresponding Authors: Zhiliang Liu (zhiliang_liu@uestc.edu.cn) and Huan Wang (wh.huanwang@gmail.com).

978-1-7281-0860-5/19/$31.00 ©2019 IEEE

In recent years, deep learning methods, which can learn complicated representations automatically from data, have drawn many researchers' attention. Deep learning architectures like convolutional neural network (CNN) and recurrent neural network (RNN) have made remarkable advances in fields like computer vision [7] [8], natural language processing [9] and sequential data processing [1]. The RNN was proposed especially for the deep learning of sequence data. Long Short-term Memory network, as a significant improvement of the RNN to relieve the problem of gradient exploding or vanishing in the RNN, has been used in a wide range of fields of machine health fault diagnosis, showing its powerful feature learning ability to classify time series sequence. For instance, in [10] Yuan et al. applied LSTM network for the fault diagnosis and prognostics of aero engine. In [11], a one-layer LSTM neural network is employed in the fault diagnosis for the three-phase asynchronous motor. Zhao et al. [12] proposed the CBLSTM model combining CNN and RNN for the tool wear test to monitor the machine health.

Therefore, this paper proposes an original fault diagnosis method for the wheelset bearings in the HSTs using Deep Bidirectional Long Short-term Memory Network (DBLSTM) is proposed. In this method, a deeper neural network is constructed by stacking bidirectional long short-term memory network (BLSTM) to obtain more discriminant fault features. In this paper, as far as we know, the LSTM network is for the first time utilized in the fault diagnosis for the wheelset bearings of the HSTs and deep bidirectional structure is employed to improve the performance of the LSTM network, which is very promising to be applied to more complicated problems. The regularization for the RNNs using recurrent dropout method is introduced, which can effectively relieve the problem of overfitting, improving the generalization ability of the deep model.

The rest of this paper is organized as follows. In section 2, basic theory of the LSTM and BLSTM is introduced. Then, in section 3, the DBLSTM fault diagnosis method is presented in detail. In section 4, the experiment results obtained using the wheelset bearing datasets are presented and discussed to illustrate the effectiveness of the DBLSTM. Finally, the paper is concluded in Section 5 and the future work is introduced.

II. Basic and Bidirectional LSTMs

A. *Basic LSTM*

Recurrent neural networks(RNNs) are a family of neural networks especially designed for processing sequential data [13]. Unlike the multi-layer perceptron and the convolutional neural network which can effectively process spatial information, the recurrent neural network is designed to process temporal information better. The state variables are introduced to store information from the past, which are used to determine the current output with current input. The recurrent neural network can be trained by Back-propagation through time algorithm, which however can be very difficult because of the problem of gradients vanishing or exploding [14].

To overcome the problem of gradient vanishing or exploding of the recurrent neural network, the LSTM was introduced [15], which is capable of capturing long-term dependencies. The architecture of the LSTM unit is shown in Fig. 1. Three gates, the input gate, the forget gate and the output gate, as well as memory cell, are introduced into the LSTM to relieve the gradient vanishing or exploding problem of RNN and better capture the dependencies of large time step distance in the time series. For a given sequence $X_i = [x_1, x_2, \cdots x_t, \cdots x_n]$, $x_t \in R^d$, where *n* is the length and *d* is the dimension of the sequence, at each time step *t*, hidden state h_t is updated by the current input x_t, hidden state at previous time step h_{t-1}, input gate i_t, forget gate f_t, output gate o_t and a memory cell C_t. The updating equations of the LSTM are given as follows:

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = tanh(W_C \bullet [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh(C_t) \quad (6)$$

where $W_f, W_i, W_C, W_o \in R^{d \times h}$, $b_f, b_i, b_C, b_o \in R^h$ are learnable parameters during model training and shared by all timesteps. σ is the sigmoid activation function. The operator $*$ denotes the element-wise product and *h* is the number of hidden units. Assuming that the hidden layer activation function of the LSTM is φ , at time step *t*, the transformation from input to output of the LSTM hidden layer can be expressed in a brief form as shown in (7).

$$h_t = \varphi(x_t, h_{t-1}) \quad (7)$$

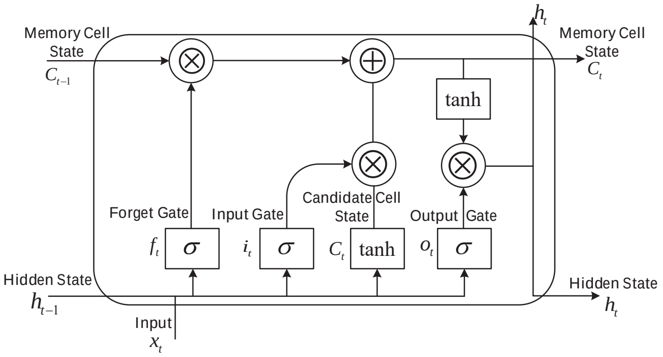


Figure 1. The architecture of the LSTM unit.

B. *Bidirectional LSTM*

The basic LSTM only considers the previous context of data without future context. This may be insufficient, considering that the time-series sensor data of bearing have strong temporal dependencies. Therefore, it is significant to apply BLSTM, which is capable of capturing temporal information from both future and past contexts of the input sequential data that might be ignored by the one direction basic LSTM [16]. To realize the BLSTM, one LSTM hidden layer deals with the time sequence from the beginning to the end in a forward time direction. Meanwhile, another LSTM hidden layer handles the sequence from the end to the beginning in a backward direction, then both the forward and backward temporal information will be fed into the same output layer. The structure of the BLSTM is shown in Fig. 2.

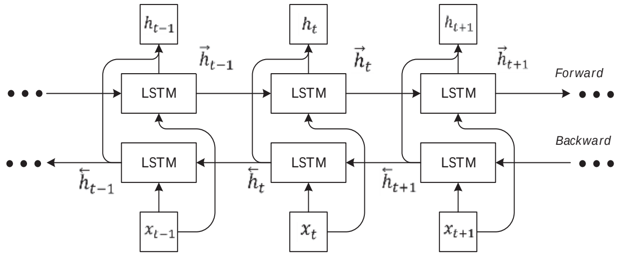


Figure 2. The structure of the BLSTM.

The $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ denote the forward and backward hidden states for time step *t* respectively, expressed as in (8) and (9).

$$\overrightarrow{h_t} = \varphi_f(x_t, \overrightarrow{h_{t-1}}) \tag{8}$$

$$\overleftarrow{h_t} = \varphi_b(x_t, \overleftarrow{h_{t+1}}) \tag{9}$$

Then, the hidden state h_t of BLSTM is obtained by concatenating the forward and backward hidden states as follows.

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} = \varphi_f(x_t, \overrightarrow{h_{t-1}}) \oplus \varphi_b(x_t, \overleftarrow{h_{t+1}}) \tag{10}$$

III. DBLSTM BASED FAULT DIAGNOSIS

The DBLSTM fault diagnosis model include two primary parts: the DLSTM network and fully connected classification layers. The original time series sequence is fed into the DBLSTM network to extract temporal features and learn meaningful representations. Then, fully connected layers are employed to process the extracted high-level features from the output of the DBLSTM network and make a classification to diagnose the fault type finally. Then, the regularization method with dropout technique is introduced into the model to prevent the problem of overfitting. The whole framework of the DBLSTM fault diagnosis model is shown in Fig. 3.

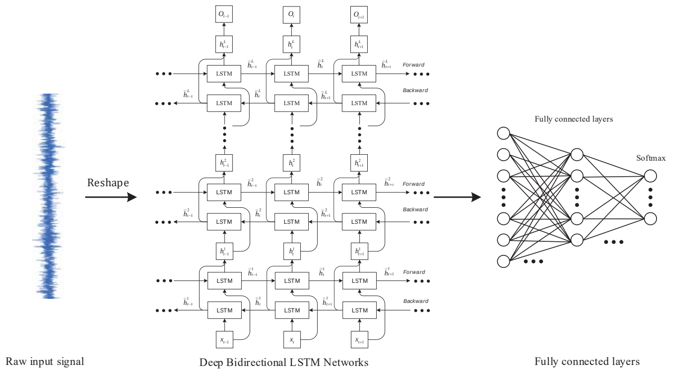


Figure 3. The architecture of the DBLSTM framework.

A. Deep Bidirectional LSTM network

Deep architectures of neural network have been shown to be powerful in representing a function more efficiently than shallow architecture, dramatically improving the automatic feature learning ability of neural network. Therefore, it is significant to build a DBLSTM neural network by stacking multiple BLSTM layers. This deep neural network can learn and abstract high-level feature representations as the input original time-series signal pass through multiple LSTM layers. The architecture of the DBLSTM neural network with *L* hidden BLSTM layers is shown in Fig. 3. Each hidden state of one BLSTM layer h_t^i is passed to the following time step of the present layer and used as the input to the following layer of the present time step. The hidden state of layer 1 and layer *l* can be computed as in (11) and (12).

$$h_t^1 = \overrightarrow{h_t^1} \oplus \overleftarrow{h_t^1} = \varphi_f(x_t, \overrightarrow{h_{t-1}^1}) \oplus \varphi_b(x_t, \overleftarrow{h_{t+1}^1}) \tag{11}$$

$$h_t^l = \overrightarrow{h_t^l} \oplus \overleftarrow{h_t^l} = \varphi_f(h_t^{l-1}, \overrightarrow{h_{t-1}^l}) \oplus \varphi_b(h_t^{l-1}, \overleftarrow{h_{t+1}^l}) \tag{12}$$

Finally, from the hidden state of hidden layer *L*, the output of the output layer can be expressed as in (13) with the output function *g*.

$$O_t = g(h_t^L) \tag{13}$$

The DBLSTM network has two benefits. Firstly, it can learn the most intrinsic high-level fault features than the shallow network in the process of addressing raw signals measured under different working conditions and strong noise environment. It is very promising to apply the DBLSTM network to solve more complicated problems. Secondly, it has good flexibility and expandability. By simply stacking multiple layers of the LSTM, the deeper network having more effective feature representation ability is obtained. Therefore, we can adjust the depth of the network flexibly to address different problems with varying complexity.

B. Regularization for Deep Bidirectional LSTM

The RNNs are particularly easy to overfit quickly despite their powerful learning ability, making it difficult to address small data [17]. Moreover, with the increase of the depth of neural network, it is easy to overfit due to the model complexity. In order to prevent the model from learning incorrect or irrelevant patterns from the training data due to overfitting, the optimal solution is to obtain more training data to improve the generalization ability. However, in real life, it tends to be hard to acquire plenty of training data. Therefore, successful model training requires good regularization, which means controlling the capacity of the neural network to relieve overfitting.

Dropout, as one of common methods for network model regularization, can effectively relieve the problem of overfitting and enhance the generalization ability of the model [18]. In [19], it points out that the proper way to use dropout with the RNNs is “to repeat the same dropout mask at each time step for both inputs, outputs, and recurrent layers”. Meanwhile, for the recurrent connections, no dropout is utilized due to the fact that the use of various masks with these connections causes deteriorated performance”. This means the constant network units are dropped at every time step, instead of sampling different dropout masks for the inputs and outputs alone. Compared with the temporally random dropout mask, which would disrupt the learning error signal and be harmful to the learning process, using the unchanged dropout mask at each timestep allows the network to properly propagate its learning error through time [20].

From [19], the new LSTM model with dropout can be represented as:

$$f_t = \sigma(W_f \bullet [h_{t-1} * Z_h, x_t * Z_x] + b_f) \tag{14}$$

$$i_t = \sigma(W_i \bullet [h_{t-1} * Z_h, x_t * Z_x] + b_i) \tag{15}$$

$$\tilde{C}_t = tanh(W_C \bullet [h_{t-1} * Z_h, x_t * Z_x] + b_C) \tag{16}$$

$$o_t = \sigma(W_o \bullet [h_{t-1} * Z_h, x_t * Z_x] + b_o) \tag{17}$$

where Z_x and Z_h are random masks repeated at all time steps, the operator $*$ denotes the element-wise product.

C. Fully connected classification layers

In this part, the output of DBLSTM network is fed into the multiple fully connected layers for classification. The expression of each layer is given by:

$$o^i = f(W_i h_i + b_i) \tag{18}$$

where o^i, W_i, h_i and b_i represent the output, the input, the weight matrix and the bias term of the *i*th fully-connected layer respectively. The activation function f is set to be ReLu function. The softmax activation function is used in the final layer of the whole classification network. For model training, the cross-entropy loss function is used to assess the error between the speculated softmax output probability distribution and the object class probability distribution. The Root Mean Square Propagation (RMSprop) optimization method is used to optimize model parameters over the loss function, which is usually a good choice for recurrent neural network [21].

D. The fault diagnosis system using the DBLSTM

In this part, the whole fault diagnosis system on the DBLSTM model for the wheelset bearing of the HSTs is displayed. The flow diagram of the system is illustrated in Fig. 4, with its overall workflow summed up as below.

1) *Assemble the dataset:* The raw signals of bearings under different working status with various faults categories are sampled by numberous acceleration sensors attached to the axle box of an HST.

2) *Determine the evaluation method:* The metric for successful fault diagnosis is determined, which will guide the selection of the loss function and the goal of model optimizing.

3) *Prepare the data:* The raw signal is divided into small signal segments through the data augmentation approach, acquiring training and testing samples in the appropriate proportion. Then the data is normalized to expedite the convergence speed of the model learning parameters while the distribution of data itself is maintained.

4) *Train the model:* Firstly, the appropriate network depth is selected according to the condition of the sample data. In the training process, the model will be repeatedly modified, trained and evaluated until the model becomes good enough. During this process, the regularization method like dropout can be utilized to avoid overfitting to enhance the performance of the model.

5) *Fault type diagnosis:* The testing samples are inputed into the well-trained fault diagnosis model to straightly diagnose the fault type of wheelset bearings of the HSTs through its effective feature representation ability.

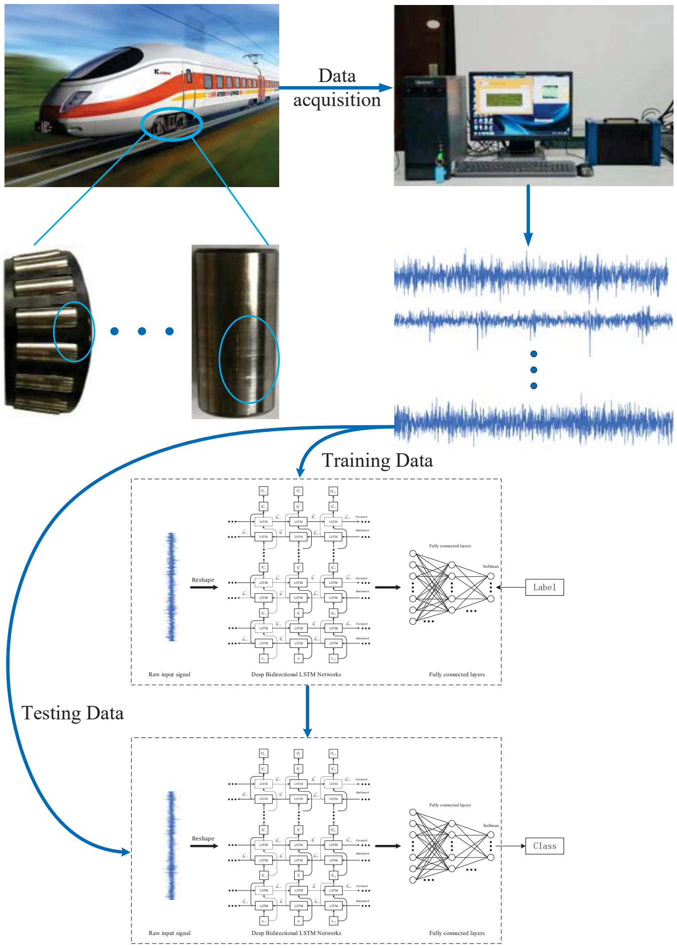


Figure 4. The fault diagnosis system based on the DBLSTM.

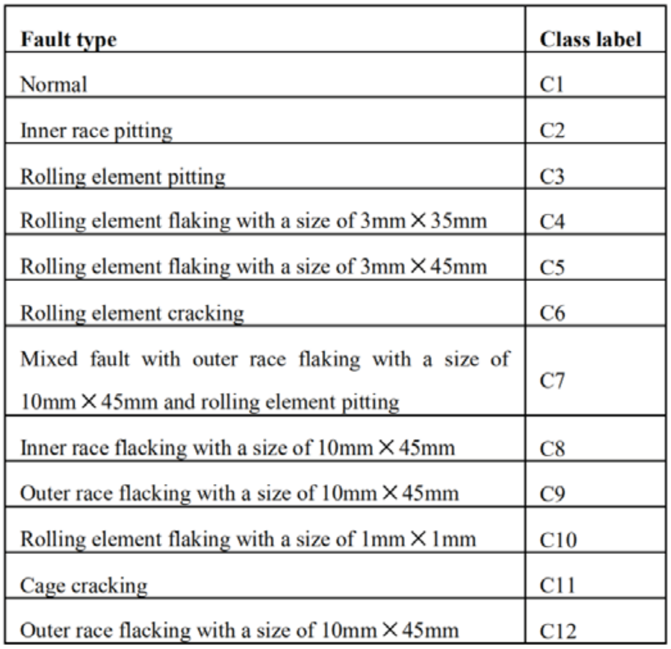
IV. EXPERIMENTS

In this section, the performance of the DBLSTM for the fault diagnosis of wheelset bearings is experimentally evaluated. Firstly, the descriptions of datasets are given. Then, in the part of experimental setup, the data augmentation methods, the compared methods and the implementation details are introduced. Finally, the experimental comparison results are shown. The effects of the depth of the DBLSTM and dropout technique are also discussed.

A. Dataset description

The experiment datasets are assembled by the wheelset bearing test rigs of the HSTs. The structure of wheelset bearing test rig is illustrated in [1]. Two acceleration sensors attached to the axle boxes of the HSTs are used to collect the raw signals from horizontal and vertical directions respectively. The sampling frequency is 5120Hz and 12 varying health conditions of wheelset bearings are simulated in the test rig. The detailed descriptions of all fault information are listed in Table I.

TABLE I. THE FAULT TYPE OF WHEELSET BEARINGS.



Fault type	Class label
Normal	C1
Inner race pitting	C2
Rolling element pitting	C3
Rolling element flaking with a size of 3mm × 35mm	C4
Rolling element flaking with a size of 3mm × 45mm	C5
Rolling element cracking	C6
Mixed fault with outer race flaking with a size of 10mm × 45mm and rolling element pitting	C7
Inner race flacking with a size of 10mm × 45mm	C8
Outer race flacking with a size of 10mm × 45mm	C9
Rolling element flaking with a size of 1mm × 1mm	C10
Cage cracking	C11
Outer race flacking with a size of 10mm × 45mm	C12

B. Experimental setup

1) Data preparation: Since a large number of data is required for the better performance of deep learning, it is very meaningful to use the data augmentation technique to expand the original data in the training of deeper models. After extending the acquired vibration signals through the data augmentation methods detailed in [1], a total of 329,752 samples are obtained. Then, these samples are separated at random into 284,260 training samples and 45,492 testing samples for model training respectively.

2) Compared methods: To demonstrate the effectiveness and superiority of the DBLSTM fault diagnosis method, the following methods will be compared:

a) CNN: A two-layer CNN model with wide first-layer kernels and three fully connected classification layers.

b) RNN: A one-layer original simple RNN model with the softmax classification layer.

c) LSTM: A one-layer LSTM network with the softmax classification layer.

d) Deep LSTM (DLSTM): A two-layer LSTM network with the softmax classification layer.

e) BLSTM: A one-layer BLSTM network with the softmax classification layer.

As to the performance metrics, for the problem of fault type classification, the accuracy evaluation method is adopted. In every experiment, the unchanged training samples and testing samples are used and 60 epochs are trained with the consistent training programme.

3) Implementation details: The DBLSTM model is constructed using the Keras library and Python 3.6. The training and testing of the model are carried out on a cloud server with an Intel Core i7-4770K CPU, 32GB RAM and a GTX 1080Ti GPU. After the data agumentation, the obtained input sequence data of each sample is a vector with the length of 2048. For the CNN, it can directly regard the vector as the input. However, training the RNN and the LSTM with long time steps can be very difficult. In the LSTM, with longer sample sequence, there will be more gradient iteration times of error backward propagation, and the computational load will be greater, which will slow down the convergence speed and reduce the learning efficiency. Therefore, the original input sequence data vector is reshaped into a matrix with the size of 128×16, where 128 is the length of time steps and 16 is the dimension of input data (input size). The reshaping process is shown in Fig. 5, where every 16 points of original time sequence form one time step in turn so that the features of one time step can be regarded as synchronous. For one LSTM layer, the number of hidden units (hidden size) is set to be 128.

During the training process, the learning rate of the RMSprop optimizer is determined to be 0.0001. The batch size is set to be 192 to accelerate the training process because of the large size of sample data.

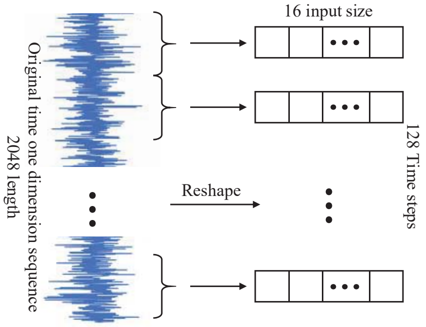


Figure 5. The reshaping process of the original data.

C. Experimental Results

In this part, we show the performance results of the DBLSTM and the comparison methods. Then, the influence of the network depth and dropout technique on the performance of the DBLSTM is displayed and discussed.

1) The comparison results in different noise environment: In this part, the white Gaussian noise is supplemented in the original signals with vrious signal noise ratios (SNRs) to imitate the strong noise environment of the practical operation of the HSTs. The effectiveness of the DBLSTM is verified in varying noise environments as the SNRs ranges from –5 dB to

5 dB. In Fig. 6 and Table II, the experimental results are displayed. Obviously, the DBLSTM outperforms the other five methods, achieving the best diagnostic performance under any noise environment. In particular, the DBLSTM model achieves nearly 94% diagnosis performance at SNR = –5 dB, showing its stronger anti-noise ability than other models under a strong noise environment.

TABLE II. THE COMPARISON RESULTS UNDER DIFFERENT NOISE ENVIRONMENT (SNR= –5DB, 0DB, 5DB).

	–5dB	**0dB**	**5dB**
CNN	0.7732	0.9066	0.9540
RNN	0.4174	0.5683	0.6148
LSTM	0.8628	0.9515	0.9787
DLSTM	0.8797	0.9650	0.9861
BLSTM	0.8830	0.9675	0.9856
DBLSTM	0.9389	0.9851	0.9937

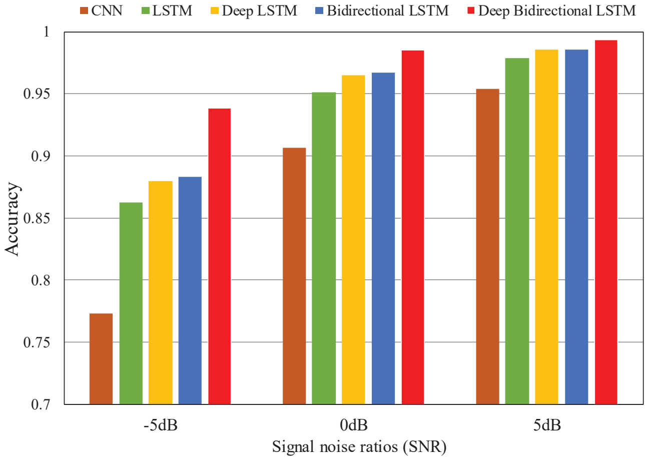


Figure 6. The comparison results under different noise environment (the RNN is not shown in the figure).

2) *The influence of the network depth:* The network depth of the DBLSTM can be adjusted by simply increasing or reducing the number of stacking BLSTM layers. The experiment is conducted under the noise of SNR = –5 dB with the results shown in Fig. 7. It is shown that the performance of the DBLSTM with two BLSTM layers is better than the one-layer BLSTM model with the improvement of about 6.4%, verifying that the deeper network structure having more powerful feature representation ability is capable of automatically learning the most intrinsic high-level fault features from original signals collected from the complicated environment. However, after adding the number of layers to be 3 and 4, the testing accuracy of the DBLSTM model decreases slightly in turn instead of increasing by a significant factor, but is still better than the one-layer BLSTM model. Therefore, diminishing returns or even performance degradation will emerge after increasing the network depth from one particular point. This may be caused by the overfitting problem as the network capacity increases. Therefore, the dropout techinique for preventing overfitting problem will be discussed in the next section.

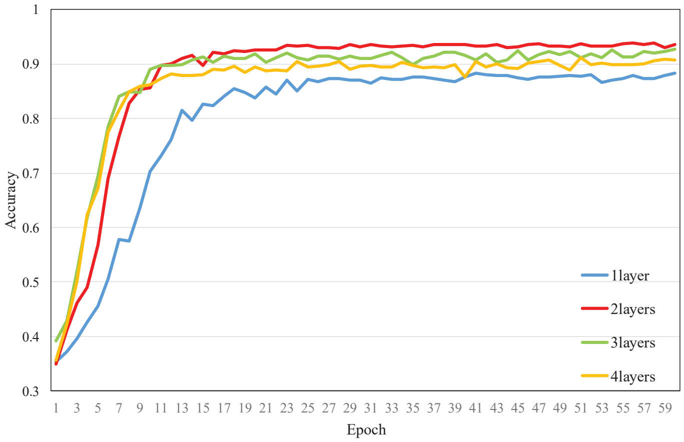


Figure 7. The test accuracy of the DBLSTM with four scenarios of network depth.

3) *The effects of dropout methods:* In this experiment, the dropout technique is adopted in the three-layer DBLSTM model with two parameters *dropout (the dropout rate for input units of the layer)* and *recurrent dropout (the dropout rate of the recurrent units)*. Additionally, the experiment is conducted under the noise of SNR = –5 dB with 60 training epochs. The performance of accuracy is shown in Fig. 8 and the training process is shown in Fig. 9. It can be seen that the dropout techinque for the LSTM network can effectively relieve the possible overfitting problem in strong noise environment. However, as shown in Fig. 9, with the dropout rate increaing, the rate of convergence of the DBLSTM slows down and the network capacity decreases, which may explain why the accuracy falls at the points of dropout rate being 0.4 and 0.5 in Fig. 8.

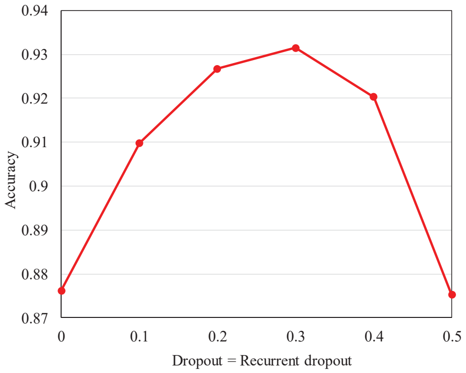


Figure 8. The test accuracy of the DBLSTM with six dropout rates (dropout = recurrent dropout).

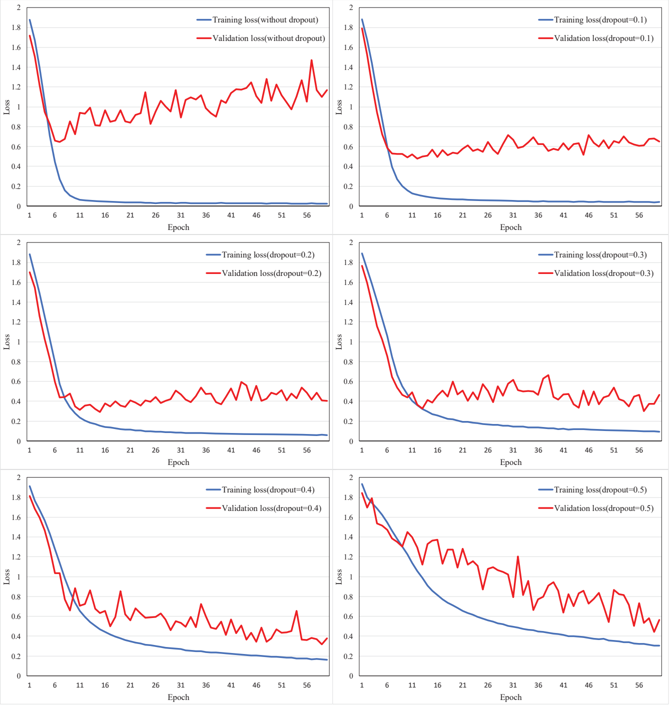


Figure 9. The loss of the DBLSTM with six dropout rates (dropout = recurrent dropout) in the training process.

V. Conclusions

In this paper, the DBLSTM network has been used to address the intelligent fault diagnosis for wheelset bearings of the HSTs with the original signals collected from complicated operating conditions. In the DBLSTM, the BLSTM is applied to enhance the performance of the LSTM network by capturing temporal information from both future and past contexts of the input sequential data. With the deeper network structure, the DBLSTM is capable of automatically learning the most intrinsic high-level fault features from original signals sampled from the complex environment. Additionally, we introduce the dropout method for regularization to prevent the overfitting problem and enhance the generalization ability of the DBLSTM. The experimental results indicate that the DBLSTM has better performance in accuracy than other five deep learning methods of bearing fault diagnosis, proving the effectiveness and advantage of the DBLSTM.

In future work, we consider introducing the convolutional neural network and the attention mechanism into the DBLSTM to improve its performance of the fault diagnosis for wheelset bearings in the HSTs.

VI. Acknowledgement

This work was supported in part by The National Key Research and Development Program of China (2016YFB1200401-106), National Natural Science Foundation of China (61833002), in part by Glasgow college, University of Electronic Science and Technology of China.

References

[1] D. Peng, Z. Liu, H. Wang, Y. Qin and L. Jia, "A Novel Deeper One-Dimensional CNN With Residual Learning for Fault Diagnosis of Wheelset Bearings in High-Speed Trains," in IEEE Access, vol. 7, pp. 10278-10293, 2019.

[2] J. Tian, C. Morillo, M. H. Azarian, and M. Pecht, ‘‘Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis,’’ IEEE Trans. Ind. Electron., vol. 63, no. 3, pp. 1793–1803, Mar. 2016.

[3] Y. Yang, D. Yu, and J. Cheng, ‘‘A roller hearing fault diagnosis method based on EMD energy entropy and ANN,’’ J. Sound Vib., vol. 294, nos. 1–2, pp. 269–277, Jun. 2006.

[4] Taylor JI. Identification of Bearing Defects by Spectral Analysis. ASME. J. Mech. Des. 1980;102(2):199-204.

[5] N. G. Nikolaou and I. A. Antoniadis, ‘‘Rolling element bearing fault diagnosis using wavelet packets,’’ NDT&E Int., vol. 35, no. 3, pp. 197–205, Apr. 2009.

[6] Y. Yang, D. Yu, and J. Cheng, ‘‘A fault diagnosis approach for roller bearing based on IMF envelope spectrum and SVM,’’ Measurement, vol. 40, nos. 9–10, pp. 943–950, Nov./Dec. 2007.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural network,’’ in Proc. Int. Conf. Learn. Represent., 2012, pp. 1097–1105.

[8] J. Donahue et al., "Long-term recurrent convolutional network for visual recognition and description," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2625-2634.

[9] T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," in IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55-75, Aug. 2018.

[10] M. Yuan, Y.Wu, and L. Lin, “Fault diagnosis and remaining useful life estimation of aero engine using lstm neural network,” in 2016 IEEE International Conference on Aircraft Utility Systems (AUS), Oct 2016, pp. 135–140.

[11] D. Xiao, Y. Huang, X. Zhang, H. Shi, C. Liu and Y. Li, "Fault Diagnosis of Asynchronous Motors Based on LSTM Neural Network," 2018 Prognostics and System Health Management Conference (PHM-Chongqing), Chongqing, 2018, pp. 540-545.

[12] R. Zhao, R. Yan, J. Wang and K. Mao. Learning to monitor machine health with convolutional bi-directional LSTM networks[J]. Sensors, 2017, 17(2): 273.

[13] I. Goodfellow, Y. Bengio, and A. C. Courville, Deep Learning, 1st ed. Cambridge, MA, USA: MIT Press, 2016, in press.

[14] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE transactions on neural network, 1994, 5(2): 157-166.

[15] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[16] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Network, 2005, 18(5-6): 602-610.

[17] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329, 2014.

[18] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural network from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[19] Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural network. In Advances in neural information processing systems (pp. 1019-1027).

[20] F. Chollet, Deep Learning With Python, Shelter Island, NY, USA:Manning, 2018, in press.

[21] Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA Neural Netw. Mach. Learn. 2012, 4, 2, unpublished.