Deep Learning based End-to-End Rolling Bearing Fault Diagnosis

Yongjie Li, Bohua Qiu, Muheng Wei, Wenqiushi Sun, Xueliang Liu
Oceanic Intelligent Technology Innovation Center
CSSC Systems Engineering Research Institute
Beijing, P. R. China
fromzerotoone@126.com (Yongjie Li)

***Abstract*—Rolling bearings play an important part in rotating machinery. As they work in complex conditions, faults will occur sometimes. Therefore, it is necessary to detect the faults early. Traditional bearing fault diagnosis methods are often based on mechanism analysis and feature selection, and the process is relatively complicated. Deep learning methods, however, have the ability to extract and select features automatically, which greatly reduces the workload. In recent years, deep learning-based methods have been successfully used in many fields, such as computer vision, voice recognition, medical diagnosis. In this paper, the end-to-end fault methods based on deep learning are proposed. The Long Short-Term Memory (LSTM) network, Gated Recurrent Unit (GRU) network and One-Dimensional Convolutional Neural Network (1D CNN) are used to build the deep learning network architecture respectively. A methodology is proposed for rolling bearing fault diagnosis, including data preprocessing, network modeling, training, validation and testing. Test bench data is used for fault diagnosis and the results show that deep learning based end-to-end methods are effective for the fault diagnosis of rolling bearings and that the model based on 1D CNN has the best performance.**

Keywords-Deep Learning; one-dimensional CNN; GRU; LSTM; Fault diagnosis

I. INTRODUCTION

In the field of electromechanical fault diagnosis, traditional methods based on mechanism analysis and feature selection have been successfully used [1-3]. However, these methods strongly rely on expert knowledge and need to extract features manually [4, 5]. Since 2006, deep learning has entered a new era, and since 2012, deep learning has entered a period of vigorous development. Nowadays, deep learning methods have been successfully used in fields such as computer vision [6, 7], voice recognition [8, 9], natural language processing [10, 11], etc. The outstanding advantage of deep learning compared to traditional methods is that deep learning methods have the ability to extract and select features automatically from data, which greatly reduces the workload [12]. As the amount of industrial data increases, sufficient sample data is provided for deep learning, which means more and more knowledge can be discovered automatically. Nowadays, in the field of fault diagnosis, deep learning-based methods are getting more and more attention [13]. CNNs and RNNs are the two main networks used in fault diagnosis in the literature. For CNNs, there are mainly two ideas for fault diagnosis. One idea is that the vibration data is firstly converted to spectrum or vibration images, and then these images can be taken as the inputs of CNN, just like image identification. The other idea is that the vibration data is directly used as the inputs of 1D CNN. For RNNs, traditional RNN is usually combined with other networks, such as LSTM, RGU, Simple Recurrent Unit (SRU), autoencoder (AE), etc.

Chen, et al. [14] proposed a fault identification method for rolling bearings based on discrete wavelet transform (DWT) and CNN. The time-frequency feature of the fault bearing signal is fully presented by DWT, and then used as the inputs of CNN. Hoang, et al. [15] proposed a fault diagnosis method base on the idea of data dimension transformation. The 1D signal data is firstly converted to 2D image data, and then the images are used as the inputs of the CNN. In the whole process, features are extracted automatically. Levent, et al. [16] proposed a lightweight adaptive 1D CNN classifier. The classifier uses original data (usually time series data) as the inputs, and after appropriate training, the classifier is able to learn the most important and excellent features automatically and no pre-transformation required. Zhang, et al. [17] proposed a novel adversarial adaptive 1D CNN method, which consists of two domains, the source domain and the target domain. Each domain consists of a feature extractor and the two extractors work partially together. This method has the ability to solve fault diagnosis problems in rolling bearing under time-varying working conditions. Peng, et al. [18] proposed a new deeper one-dimensional CNN based on residual learning, which can overcome the training difficulty and performance problem, which are unavoidable for traditional deep networks. The method is able to learn abstract features effectively with high accuracy under normal conditions, abnormal conditions and varying conditions.

Liu, et al. [19] proposed a deep learning method for bearing fault diagnosis based on RNN combined with an autoencoder. This method predicts operation values of the next period from the previous period using GRU-based denoising autoencoder. The errors between predicted values and true values are used for anomaly detection and fault type identification. Bruin, et al. [20] proposed a novel fault diagnosis method based on LSTM RNN network using easily available measurement signals, and concluded that compare with CNN, the LSTM network works better in real-time fault detection in the field of railway track. Yuan, et al. [21] applied LSTM network to aero engine fault location and remaining useful life (RUL) prediction under complicated working conditions and in noisy environment. Hinchi, et al. [22] proposed a deep learning framework for RUL estimation. The framework is based on convolutional network and LSTM units. The convolutional layer extracts and selects

This work is supported by the R&D project of intelligent ship 1.0.

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

features from raw data collected by sensors, and then the LSTM layer determines the degradation process.

As mentioned, deep learning-based methods have been widely applied to fault identification and diagnosis and achieved quite good results in recent years. On this basis, in this paper, a unified multi-channel end-to-end architecture with different networks is proposed. The word 'unified' means the architecture has certain versatility. The word 'multi-channel' means the networks can process multi-channel data. The word 'end-to-end' means the network can be treated as a black box from raw data inputs to result outputs and needs no pre-determined transformation. Since LSTM, GRU and 1D CNN networks are suitable for processing one-dimensional signals and have achieved good results in other related fields, these three networks are used to build the deep learning network structure respectively with a unified architecture. This paper is organized as follows. Section I is the introduction and paper organization. Section II describes the methodology of deep learning-based fault diagnosis. Section III introduces the experimental setup of data acquisition. Section IV presents the testing results and results analysis of different deep learning- based methods and section V gives the conclusions.

II. METHODOLOGY

As a special machine learning method, deep learning is also a data-driven method and the main steps include data acquisition and preprocessing, model building, training, validation and testing. The flowchart of deep learning-based rolling bearing fault diagnosis is shown in Fig. 1.

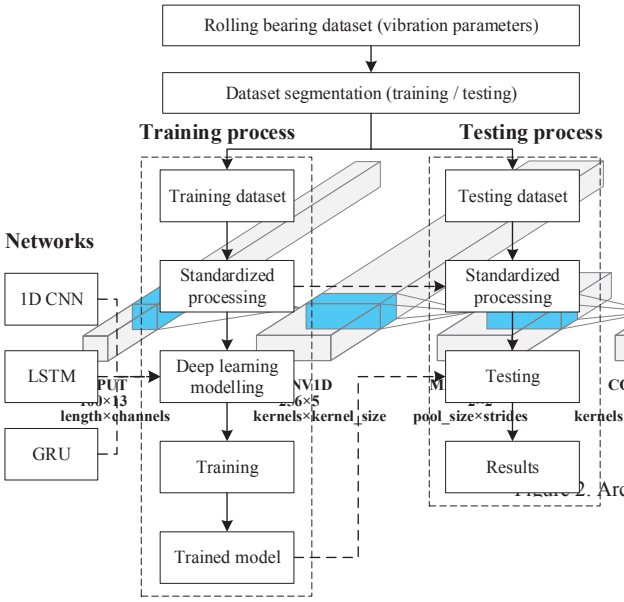


Figure 1. Flowchart of deep learning-based fault diagnosis

The overall process is as follows: Step I, acquire the rolling bearing dataset that consists of vibration parameters. Step II, split the dataset into training dataset and testing dataset. Step III is the training process. Firstly, standardize the training dataset. Secondly, build the deep learning model with 1D CNN, LSTM and GRU, respectively. Thirdly, train the built model with training dataset and get the trained model. Step IV is the testing process. Firstly, standardize the testing dataset according to the rules of the training dataset. Secondly, test the testing dataset with the trained model and get results.

A. Data acquisition and preprocessing

The data acquisition process will be described in Section III in detail. In the data preprocessing process, the dataset is split into two datasets, for training and testing respectively. And the two datasets are standardized with the same rules. Besides, the training dataset is split into two parts, namely training part and validation part, so that we can see the training process with these two parts of data.

B. Deep learning modelling

In the deep learning modelling process, three types of networks, namely 1D CNN, LSTM and GRU, are chosen. Fig. 2 shows the framework architecture of the 1D CNN network, which contains an input layer, two convolutional layers, two pooling layers, two fully connected (FC) layers and a multi-classifier layer. The number of channels and length of input samples are 13 and 100, respectively, which means each input sample consists of 13 vibration parameters and each parameter consists of 100 points. The convolutional layer and pooling layer use one-dimensional Conv1D and MaxPool1D to adapt the input samples. The flatten layer reshapes the network into one-dimensional network. The followed two fully connected layers work just like traditional neural networks. It should be noted that each convolutional layer or each FC layer consists of an

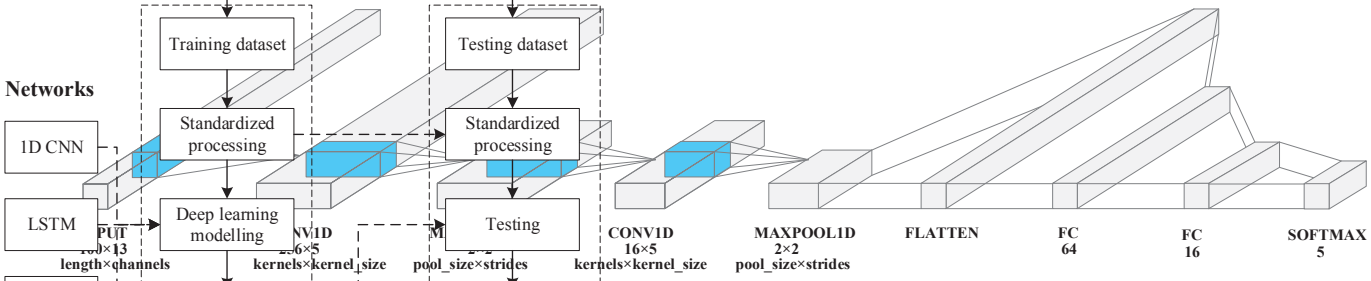


Figure 2. Architecture of 1D CNN network

activation function. The multi-classifier layer uses 'softmax' as the activation function to classify the types of fault diagnosis, including four fault types and one normal type.

The architecture of LSTM (or GRU) network is similar to that of 1D CNN network except that the convolutional layers and pooling layers are replaced with one LSTM (or GRU) layer, or one RNN layer with several LSTM (or GRU) cells, as shown in Fig. 3. Therefore, the LSTM (or GRU) network contains an input layer, a LSTM (or GRU) layer and a multi-classifier layer.

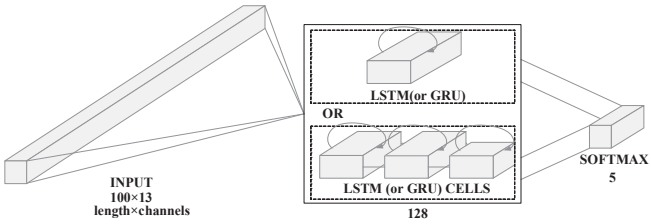


Figure 3. Architecture of LSTM (or GRU) network

C. *Training, validation and testing*

For each deep learning-based model, use standardized training dataset, which consists of a large amount of samples and is split into training part and validation part, to train and validate the model. In order to better validate the reliability of deep learning models, data from several different working conditions is used to train models. And then use the trained model to test the testing dataset. Finally, we get results from three different models for different working conditions.

III. EXPERIMENTAL SETUP

The whole test bench includes drive motor, load motor, test gearbox, load gearbox, etc., as shown in Fig. 4.

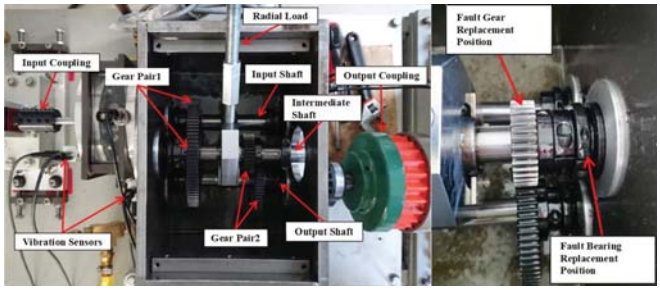
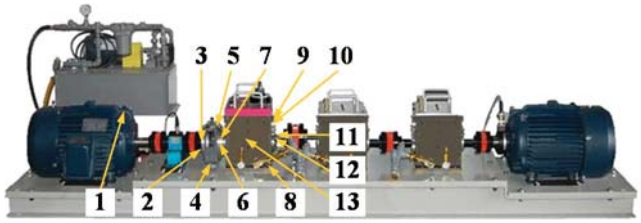


Figure 4. Deployment of test bench

The parallel shaft gearbox is divided into two gear sets, which means the gearbox can be decelerated by two stages, and the deceleration effect is more obvious. There are three shafts in the parallel shaft gearbox test bench, namely the input shaft, the intermediate shaft and the output shaft. The input shaft is connected to the output shaft of the planetary gearbox. The intermediate shaft is the uppermost shaft in Fig. 4. The fault parts replaced are the gear and rolling bearing of the intermediate shaft. The output shaft is connected to the rightmost couplings.



1-motor Z, 2-planet X, 3-planet Z, 4-planet carrier Y, 5-planet carrier Z, 6-parallel middle left Y, 7-parallel middle left Z, 8-parallel box base Z, 9-parallel middle Y, 10-parallel middle Z, 11-parallel input right X, 12-parallel input right Z, and 13-parallel box Y.

Figure 5. Deployment of sensors on the parallel shaft gearbox test bench

The parallel shaft gearbox test bench can be loaded with normal gears (or bearings) or fault gears (or bearings) for life prediction experiments. The load gearbox is used to increase the output speed of the test gearbox output and reduce the torque load.

There are 13 sensors deployed on the test bench, including motor Z, planet X, planet Z, planet carrier Y, planet carrier Z, parallel middle left Y, parallel middle left Z, parallel box base Z, parallel middle Y, parallel middle Z, parallel input right X, parallel input right Z and parallel box Y, as shown in Fig. 5.

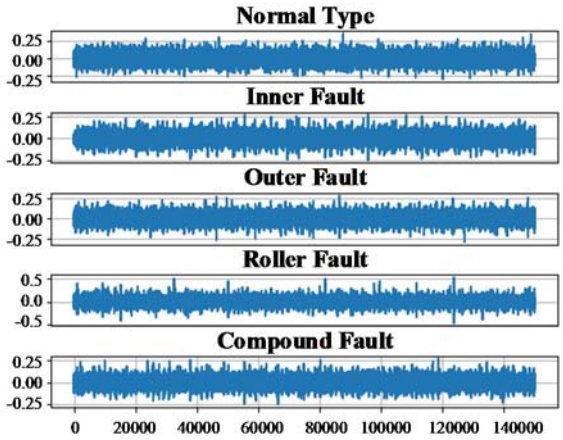
In the experiment, the rotating speed of the motor and the load rate can be adjusted. As designed, 1000 rpm, 2000 rpm and 3000 rpm are chosen as the rotating speed of the motor, and 0%, 3% and 6% of full load are chosen as the load rate. Therefore, there are 9 working conditions in the experiment, which means 9 datasets in different working conditions with 13 vibration parameters are available for analysis. Besides, there are five types of data, including four fault types and one normal type, in the parallel shaft gearbox test bench. The four fault types are inner fault, outer fault, roller fault and compound fault. The summary of dataset, including parameter types and numbers, is listed in Table I.

TABLE I. SUMMARY OF DATASET

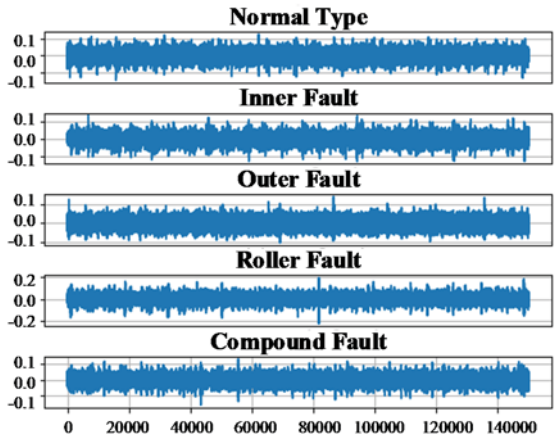
Items	*Vibration parameters*	*Working conditions*		*Types of data state*
		Rotating speed	*Load rate*	
Number	13	3	3	5
		9		
Total number of data series	585			

Table I shows that the total number of data series is 585. For each of the data series, we select 150000 points for training and validation and 20000 points for testing. Therefore, the order of magnitude of the dataset to be analyzed is about 100 million, which is large enough for fault diagnosis analysis.

Original data from planet Z sensor and parallel middle Y sensor are shown in Fig. 6. All the five types of data are plotted in one figure.



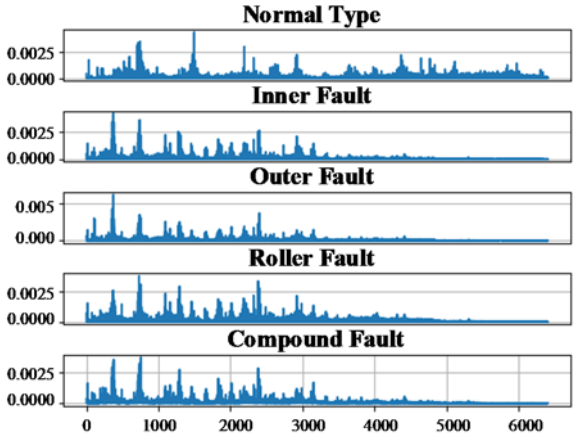
a) Original data from planet Z sensor



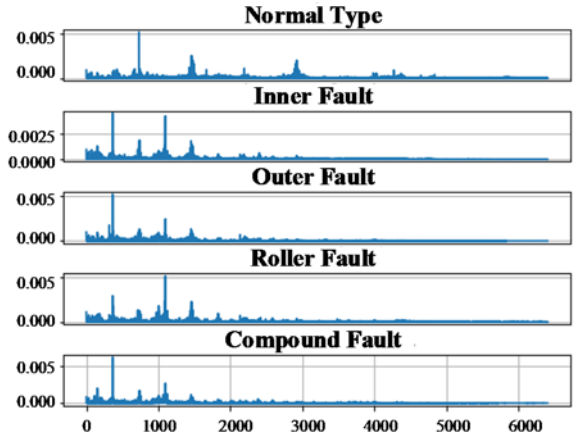
b) Original data from parallel middle Y sensor

Figure 6. Five types of original data from different sensors

Fig. 6 shows that it is not easy to distinguish normal type, inner fault, outer fault and compound fault in the time domain plots. One solution is to perform spectrum analysis and extract features in the frequency domain, as shown in Fig. 7.



a) Spectrums of original data from planet Z sensor



b) Spectrums of original data from parallel middle Y sensor

Figure 7. Spectrums of five types of original data from different sensors

Fig. 7 shows that the differences among the five types of data in frequency domain is more significant than those in time domain. In traditional fault diagnosis methods, amplitude values at specific frequencies will be selected as features in frequency domain. In Fig. 7(b), values at 365 Hz and 1095 Hz may be selected as features, and then the two features can be used as features of machine learning methods, such as SVM, KNN, Decision Tree, Random Forest. In Fig. 7(a), however, it is not easy to find the specific frequencies because the features exist in a frequency region, for example the frequency region around 370Hz and the frequency region around 740 Hz. Therefore, it is still difficult to distinguish all the five types of signal without expert knowledge.

In the paper, a deep learning-based method is used to overcome the mentioned difficulty considering its ability to extract features automatically.

IV. Results and Analysis of Deep Learning Based Methods

For dataset of each working condition, three models, i.e. 1D CNN, LSTM and GRU, are trained respectively. For each model, the hyperparameters are the same. The epoch number is 20 and the batch size is 32. Each sample that is used for training, validation and testing consists a sequence with length of 100. For each time of training, 6750 samples are used as training data and 750 samples are used as validation data. For each time of testing, 1000 samples are used as testing data. We will compare the three models in three aspects, i.e. model complexity, convergence rate and testing accuracy.

A. Comparison of model complexity

The model summaries are listed in Table II for 1D CNN, LSTM and GRU networks. In order to avoid overfitting problem and to improve the training efficiency, a dropout layer is added to 1D CNN network. Each convolutional layer or each FC layer contains an activation function of 'relu'. For LSTM and GRU networks, the LSTM layer and GRU layer use the standard cells with an output dimension of 128 and an activation function of 'tanh'.

Table II. Summary of three models

Models	*Layer (type)*	*Output shape*	*Number of parameters*
1D CNN	conv1d_1 (Conv1D)	(None, 100, 256)	16896
	max_pooling1d_1 (MaxPooling)	(None, 50, 256)	0
	conv1d_2 (Conv1D)	(None, 46, 16)	20496
	max_pooling1d_2 (MaxPooling)	(None, 23, 16)	0
	flatten_1 (Flatten)	(None, 368)	0
	dropout_1 (Dropout)	(None, 368)	0
	dense_1 (Dense)	(None, 64)	23616
	dense_2 (Dense)	(None, 16)	1040
	dense_3 (Dense)	(None, 5)	85
	Total parameters		**62,133**
LSTM	lstm_1 (LSTM)	multiple	72704
	dense_1 (Dense)	multiple	645
	Total parameters		**73,349**
GRU	gru_1 (GRU)	multiple	54528

Models	*Layer (type)*	*Output shape*	*Number of parameters*
	dense_1 (Dense)	multiple	645
	Total parameters		**55,173**

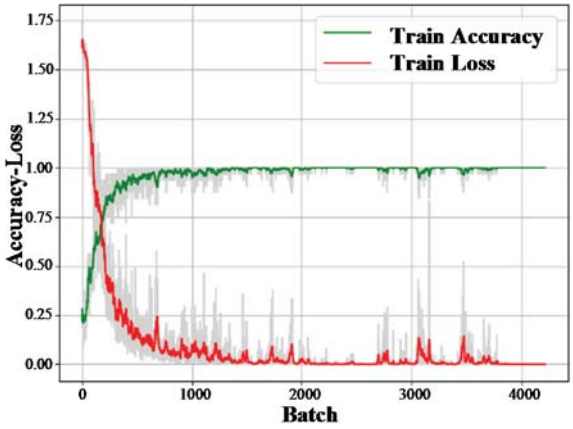
From Table II we can see that the total parameters to be trained of three types of deep learning networks are in the same order of magnitude. In the ascending order of total parameters, they are GRU network, 1D CNN network and LSTM network. However, according to the training results, 1D CNN network is the most efficient (about 12 seconds per epoch), followed by GRU network (about 21 seconds per epoch) and LSTM network (about 32 seconds per epoch).

B. Comparison of convergence rate

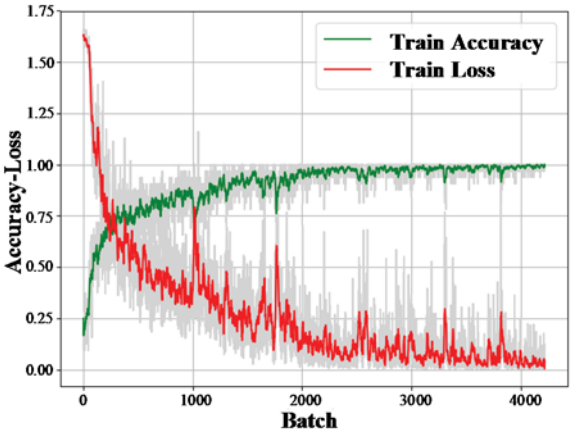
Fig. 8 shows the convergence rates of the three models for the working condition of 2000 rpm rotating speed of the motor and 0% load rate. In this paper, the convergence rate is defined as stable accuracy over stable batch point.

$$convergence\ rate = \frac{stable\ accuracy}{stable\ batch\ point} \quad (1)$$

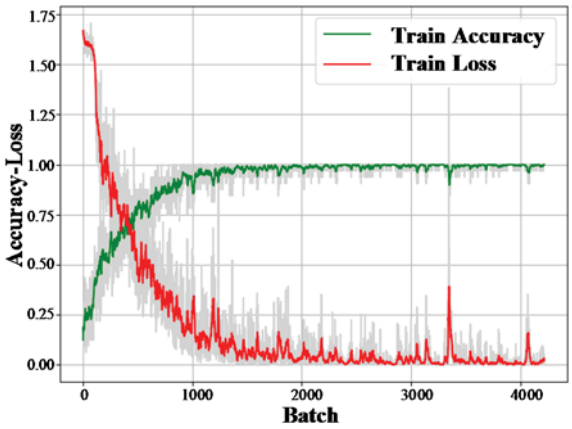
In (1), the stable batch point is the point where the accuracy curve firstly reaches stable accuracy and from where the accuracy stays relatively stable. The convergence rate is a comprehensive representation of the convergence speed and the diagnosis accuracy of the network, representing the efficiency of the network.



a) 1D CNN



b) LSTM



c) GRU

Figure 8. Comparison of convergence rates of three models

As shown in Fig. 8, it takes 750 batches for 1D CNN network to reach at a relative stable accuracy, while 2700 batches for LSTM network and 1450 batches for GRU network. The results show that the convergence rate of 1D CNN network is higher than those of LSTM and GRU networks. Therefore, the convergence rates for 1D CNN, LSTM and GRU are 0.0013, 0.0004 and 0.0007, which means 1D CNN network has the fastest convergence rate and LSTM network has the slowest convergence rate. Besides, form the accuracy-loss curves in Fig. 8, we can see that the 1D CNN network is more stable than LSTM and GRU networks.

C. Comparison of testing accuracy

The testing results of three models for each condition are listed in Table III. The load rate includes 0%, 3% and 6%, and the rotating speed includes 1000 rpm, 2000 rpm and 3000 rpm (therefore nine working conditions in total).

TABLE III. TESTING ACCURACY OF EACH CONDITION

Models		*1D CNN*	*LSTM*	*GRU*
Load rate	*Rotating speed (rpm)*			
0%	1000	1.000	0.984	0.988
	2000	0.998	0.963	0.973
	3000	0.999	0.970	0.976
3%	1000	0.997	0.967	0.985
	2000	0.993	0.960	0.970
	3000	0.999	0.988	0.954
6%	1000	0.998	0.918	0.941
	2000	1.000	0.960	0.973
	3000	0.996	0.951	0.976
Minimum accuracy		0.993	0.918	0.941
Maximum accuracy		1.000	0.988	0.988
Average accuracy		0.998	0.962	0.971

From Table III we can see that, for the nine working conditions, the minimum accuracies of 1D CNN, LSTM and GRU networks are 0.993, 0.918 and 0.941, the maximum accuracies of 1D CNN, LSTM and GRU networks are 1.000, 0.988 and 0.988, and the average accuracies of 1D CNN, LSTM and GRU networks are 0.998, 0.962 and 0.971. The testing results show that 1D CNN network has the highest accuracy in rolling bearing fault diagnosis, followed by GRU, and the worst is LSTM.

V. CONCLUSION

In this paper, rolling bearing fault diagnosis methods based on deep learning are proposed. LSTM, GRU and one-dimensional (1D) CNN are used to build the deep learning network structure respectively. A methodology is proposed for general rolling bearing fault diagnosis process, including data preprocessing, network modeling, training, validation and testing. Test bench data of nine working conditions is used for fault diagnosis and from the results we finally conclude that:

a) Deep learning based end-to-end methods are effective for fault diagnosis in the field of rolling bearing, each average accuracy of the three models is above 0.96.

b) The deep learning model based on 1D CNN network has the best performance in accuracy, efficiency, convergence rate and stability, followed by model based on GRU network and LSTM network.

This research is of great significance for intelligent fault diagnosis, prognosis and health management of marine equipment and systems.

REFERENCES

[1] R. Zhou, W. Bao, N. Li, X. Huang and D. Yu, “Mechanical equipment fault diagnosis based on redundant second generation wavelet packet transform,” Digital Signal Processing, vol. 20, pp. 276-288, January 2010.

[2] T. Boutros and M. Liang, “Mechanical fault detection using fuzzy index fusion,” International Journal of Machine Tools and Manufacture, vol. 47, pp.1702-1714, September 2007.

[3] S. Nandi, H. A. Toliyat and X. Li, “Condition monitoring and fault diagnosis of electrical motors—a review," IEEE Transactions on Energy Conversion, vol. 20, pp. 719-729, December 2005.

[4] J. Lin and L. Qu, “Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis,” Journal of Sound and Vibration, vol. 234, pp. 135-148, June 2000.

[5] C. Nan, F. Khan and M. T. Iqbal, “Real-time fault diagnosis using knowledge-based expert system,” Process Safety and Environmental Protection, vol. 86, pp. 55-71, January 2008.

[6] W. Hao, R. Bie, J. Guo, X. Meng and S. Wang, “Optimized CNN based image recognition through target region selection,” Optik - International Journal for Light and Electron Optics, vol. 156, pp. 772-777, November 2017.

[7] S. Yang and D. Ramanan, “Multi-scale recognition with DAG-CNNs," IEEE International Conference on Computer Vision, pp. 1215-1223, 2015.

[8] C. Liu, Y. Wang, K. Kumar and Y. Gong, “Investigations on speaker adaptation of LSTM RNN models for speech recognition,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[9] G. Alex, N. Jaitly and A. R. Mohamed, “Hybrid speech recognition with deep bidirectional LSTM,” 2013 IEEE workshop on automatic speech recognition and understanding, 2013.

[10] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He and et al., “Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval,” IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 24, pp. 694-707, April 2016.

[11] R. Messina and J. Louradour, “Segmentation-free handwritten Chinese text recognition with LSTM-RNN,” 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 171-175, 2015.

[12] M. He and D. He, “A deep learning based approach for bearing fault diagnosis,” IEEE Transactions on Industry Applications, vol. 53, pp. 3057-3065, June 2017.

[13] S. Zhang, S. Zhang, B. Wang and T. G. Habetler, “Machine learning and deep learning algorithms for bearing fault diagnostics - A comprehensive review,” arXiv:1901.08247, January 2019.

[14] R. Chen, X. Huang, L. Yang and B. Tang, “Rolling bearing fault identification based on convolution neural network and discrete wavelet transform,” Zhendong Gongcheng Xuebao/ Journal of Vibration Engineering, vol. 31, pp. 883-891, October 2018.

[15] D. T. Hoang and H. J. Kang, “Convolutional neural network based bearing fault diagnosis,” International Conference on Intelligent Computing, Cham: Springer, 2017, pp. 105-111.

[16] L. Eren, I. Turker and K. Serkan, “A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier,” Journal of Signal Processing Systems, vol. 91, pp. 179-189, February 2018.

[17] B. Zhang, W. Li, J. Hao, X. Li and M. Zhang, “Adversarial adaptive 1-D convolutional neural networks for bearing fault diagnosis under varying working condition,” arXiv:1805.00778, May 2018.

[18] 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,” IEEE Access, vol. 7, pp. 10278-10293, December 2018.

[19] H. Liu, J. Zhou, Y. Zheng, W. Jiang and Y. Zhang, “Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders,” ISA Transactions, vol. 77, pp. 167-178, April 2018.

[20] T. D. Bruin, K. Verbert and R. Babuska, “Railway track circuit fault diagnosis using recurrent neural networks,” IEEE Transactions on Neural Networks and Learning Systems, vol.28, pp. 523-533, March 2017.

[21] M. Yuan, Y. Wu and L. Lin, “Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network,” 2016 IEEE International Conference on Aircraft Utility Systems (AUS), 2016.

[22] A. Z. Hinchi and M. Tkiouat, “Rolling element bearing remaining useful life estimation based on a convolutional long-short-term memory network,” Procedia Computer Science, vol. 127, pp. 123-132, January 2018.