Rolling Bearing Diagnosis Based on CNN-LSTM and Various Condition Dataset

Osamu Yoshimatsu[1], Yoshihiro Satou[2], and Kenichi Shibasaki[3]

[1,2,3] *Core Technology R&D Center, NSK Ltd., Fujisawa, Kanagawa, 251-8501, Japan*

Yoshimatsu-o@nsk.com
satou-yos@nsk.com
shibasaki@nsk.com

ABSTRACT

Flaking is typical failure mode in rolling bearings. Therefore, flaking diagnosis plays a critical role in condition monitoring of general rotating machinery. In recent years, there has been an increasing interest in deep learning technique for bearing flaking diagnosis, because it can learn the flaking induced vibration features with no information of bearing specifications nor that of rotating speed. However, most of the studies have only focused on laboratory data using one test rig as well as a small dataset under the limited operating condition. Accordingly, no discussion has been found on the generalization performance of the diagnostic model, i.e., availability for actual rotating machinery, in which vibration feature is affected by various operating conditions and unknown disturbance. In this study, more than 21,000 time-series waveforms of normal and bearing flaking induced machine vibration were prepared from three types of test rig and three bearing types under various operating condition. And deep learning such as Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) models were applied to recognize flaking bearing vibration. The applied models trained with various condition data showed higher accuracy of various condition test data diagnosis than other models trained using single condition data. Furthermore, the applied diagnostic models also showed less accuracy degradation for test data in which additional artificial noise was imposed, than the models trained with single condition data.

1. INTRODUCTION

Condition monitoring of rotating machinery in many industrial fields is important for reducing operational costs and avoiding sudden accidents. In particular, rolling bearings are often used under severe operating conditions, thus bearings have a substantial need for condition monitoring. The most typical damage is flaking. Because of advantage of vibration measurement, e.g. higher SN ratio, compared to other measurement methods, vibrating velocity and acceleration are the widely used for detecting bearing flaking. In the bearing diagnosis using vibration, highly accurate diagnosis is possible if appropriate features are extracted from data, and also threshold is set using information such as bearing specifications and rotational speed (Randall et al., 2011). However, since vibration is affected by transferring path from the source to sensor positions and other components of vibrations, it is difficult even for experts to extract common features from vibration data of many rotating machinery.

In recent years, many studies have been reported that deep learning can extract the features and can achieve high diagnostic accuracy by training using data of rotational machinery without information of bearing specifications nor that of rotating speed. In case of image classification model in deep learning, it is known that common features of training data can be extracted by models trained with sufficient quantity data (Le et al., 2013). Various deep learning methods have been applied for diagnosis of machinery in order to extract features appropriately that can distinguish normal vibration waveform and damaged vibration waveform (Zhao et al. 2016). Convolutional Neural Network (CNN), which is a representative method of deep learning, can extract local features by convolutional kernels and pooling operation. However, one of the features of flaking vibration is periodical phenomena due to repetitive impact when rolling elements of bearing passing through flaking spalls. The amplitude and the interval of these features depend on bearing specification and operating condition. It is hard for CNN to recognize various sequential change of these vibrational features. Therefore, CNN would be insufficient for diagnosis of bearing flaking. In case of a previous study (Zhao et al. 2017), a regression model combining CNN and Long Short-Term Memory (LSTM) has been applied to predict wear condition of tool. LSTM is a kind of deep learning method that enables classification and regression of time-series data. The aim of their model was to recognize local robust features and temporal information, and their model showed better prediction accuracy than the conventional model. It is thought that similar approach may be effective to enhance diagnosis accuracy for the case of flaking induced vibration of rolling bearings.

In this study, deep learning models such as CNN, LSTM and CNN-LSTM(combined model CNN and LSTM) were applied to recognize flaking bearing vibrational features as described above. CNN extracts various types of impact vibration waveform and LSTM recognizes periodical phenomena featuring bearing flaking.

On the other hand, the diagnosis using deep learning in the field application has a various problem. Firstly, diagnosis would be difficult when conditions of training data and those of diagnostic target data are different including the structure of machines to which bearings are mounted. Whereas in many previous studies, the performance of diagnostic models is evaluated using data of a single test rig, thus it has not been verified that the trained diagnostic model was applied to the other test rigs (Zhang et al. 2018; Feng et al. 2018; Mao et al. 2016; Chen et al. 2017).

Secondly, it is quite usual in the field that vibration signal contains not bearing originated component, i.e. noise. Therefore, it is hard to diagnose bearing flaking using a simple exceeding detection method. But almost nothing is discussed on the robustness of the trained bearing diagnosis model for noise-containing data in previous reports.

In this study, in order to make clearer on the above-mentioned problems, training and evaluation of the diagnosis models were carried out using three kinds of datasets in which bearing type and/or test rig were different, in addition to that the artificial noise was imposed to the test data.

2. APPLIED MODELS

2.1. Convolutional Neural Network (CNN)

CNN is one of the deep learning methods originally proposed for image processing (Le Cun et al. 1990; Krizhevsky et al. 2012). In our applied models, CNN consists of multilayers, and each layer has its function, i.e., a convolutional layer or a pooling layer. Operations of the convolutional layers are a summation of multiplications between the vectors of input data and weight coefficients of convolutional kernel. Operations of pooling layers are extractions of features and fixture length of the vectors. Also, time-series data such as acceleration waveforms are processed in 1-D convolutional layers and average pooling layers. Additionally, in order to avoid overfitting of model training, batch normalization (Ioffe et al. 2015) is performed on each outputs of the pooling layers.

2.2. Long Short-Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN), which is a deep learning method that can classify and regress time-series data such as natural languages and voices in consideration of feature changes at each time step. LSTM is an improvement of RNN in order to capture long-term dependencies (Hochreiter et al. 1997; Graves et al. 2005).

2.3. CNN-LSTM

The applied CNN-LSTM consists of three major parts, i.e., the multi-layered CNN, LSTM layer and Fully-Connected (FC) layer. The whole structure of the CNN-LSTM is shown in Figure 1. The subject of bearing flaking diagnosis is defined as a binary classification of normal and flaking vibration, in other words an input of the model is time-series acceleration waveform and output of the model is 2 classes labels. The labels, which are 2-D one-hot vectors, represent normal bearing and fault bearing. Firstly, in CNN layers, the 8192 points waveform are processed to data of 16 points × 256 ch, via 9 sets of convolution layer and average pooling layer. Next, the output of the last CNN layer is divided into 16-time steps, and led to the LSTM layer each time step. In addition to its input, the LSTM also receives the output of the LSTM layer at previous time step as an input. At the last time step, the output of the LSTM layer is input to 2-D FC layer. In the FC layer, outputs present diagnostic result of the input waveform as “normal” or “fault”.

In this study, CNN-LSTM model, which is presumed to be suitable for the diagnosis of bearing flaking, was applied mainly. However, in order to compare the generalization performance, model consisting of 3-CNN layers and 2-FC layers (CNN model), and model consisting of 2-LSTM layers (LSTM model) were applied for the training and the test. For the training and the test of these models, the same datasets described in next chapter were used.

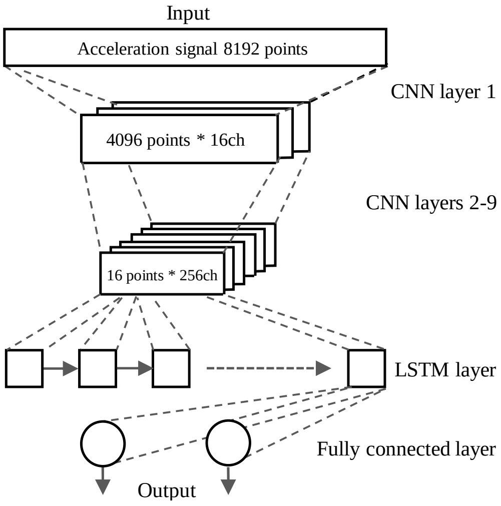

Figure 1. Diagram of the applied CNN-LSTM.

3. TRAINING METHOD

As previously mentioned, three kinds of datasets were prepared both for the training and the test. It is noteworthy that, bearings were damaged artificially, i.e., faults were not actual flaking. Such kind of alternatives is widely adopted from the viewpoint of saving the time of preparation of a number of failed bearings. Further, as the training method applied by the authors, the data of normal (not damaged) bearings are added as training data. This is the simulation of the case, especially in the field, that the trained diagnostic model is applied to new machinery and/or bearings, at which getting failure data due to the damage are quite difficult.

3.1. Description of Datasets

In order to evaluate generalization performance of the applied diagnostic models, three types of datasets (Dataset A, B, C) were prepared. These were acquired with accelerometers, mounted on the test rigs to observe the conditions of the bearings in operation. Table 1 shows details of the acquisition conditions of the datasets. Acceleration waveforms were picked up from each dataset using different test rig. Since three test rigs have their own resonant characteristic, suitable frequency band for diagnosis is different. Also, differences in operating conditions affect amplitude and intervals of impact vibration due to rolling element passing through flaking spall. Therefore, once high diagnostic accuracy is achieved using data of various operating conditions and machinery (test rig), one may expect the model widely available.

Dataset A includes vibration data of ball bearings with artificial defect. These data are published by Case Western Reserve University (Loparo, 2012). In this dataset, acceleration waveforms were acquired using ball bearings with 12 kinds of artificial defects of different sizes under 4 kinds of operating conditions (Smith et al. 2015). Dataset B consists of acceleration waveforms using cylindrical roller bearings with 4 types of artificial defects under 9 different operating conditions. Dataset C consists of acceleration waveforms using spherical roller bearings with 5 types of artificial defects under 24 different operating conditions. The tests for acquisition of dataset B and C were carried out using different test rigs in the authors' company.

Three types of test data and 15 types of training data were arranged using these Dataset A, B and C. Table 2 shows contents of test data and training data. In Table 2, Subscript of An, Bn and Cn mean "normal", i.e., all waveforms are from not failed bearing. The order of the waveforms of each test data and training data was randomly rearranged. Each acceleration waveform was picked up from dataset so that it became sequential 8192 sampled time series data. Each test data consists of 2000 waveforms and each training data consists of 5000 waveforms. There was no duplication between each training data and each test data. The ratio of normal bearing data to damaged bearing data is 1 : 1 in all the

Table 1. Details of the Datasets.

Dataset name	A (CWRU data)		B		C	
Bearing type	Ball bearing		Cylindrical roller bearing		Spherical roller bearing	
Bearing number	6205-2RS JEM (SKF/NTN)		NU2228BMMA (NSK)		230/750CAME4 (NSK)	
Sampling rate [Hz]	48000		48000		800	
Rotational Speed $[\text{min}^{-1}]$	1730 to 1797 (4 types)		1200 to 1750 (3 types)		8 to 20 (4 types)	
Load [types]	4 (Motor horse power)		3 (Radial)		6 (Radial + Axial)	
Types of artificial defect	Place	Size [types]	Place	Size [types]	Place	Size [types]
	None	1	None	1	None	1
	Inner race	4	Inner race	3	Inner race	1
	Outer race	4	Outer race	1	Outer race	4
	Ball	4				

Table 2. List of the combinations of training data and test data

Test No.	Training Data	Test Data
1	A	A
2	B	
3	C	
4	B + C	
5	B + C + An	
6	A	B
7	B	
8	C	
9	A + C	
10	A + C + Bn	
11	A	C
12	B	
13	C	
14	A + B	
15	A + B + Cn	

test data and the training data. Table 3 shows the number of waveforms included in each training data.

In order to avoid the influence of the amplitude of the vibration waveform and focus on the waveform shape, each waveform was normalized as preprocessing so that average value of each waveform becomes 0 and standard deviation becomes 1. Figure 2 shows examples of the normalized waveforms obtained from each dataset. As shown in Figure 2, each example of fault waveform includes periodical phenomena of impact vibration.

Table 3. Number of waveforms in the training data.

Test No.	Data A		Data B		Data C	
	Normal	Fault	Normal	Fault	Normal	Fault
1	2500	2500	0	0	0	0
2	0	0	2500	2500	0	0
3	0	0	0	0	2500	2500
4	0	0	1250	1250	1250	1250
5	834	0	833	1250	833	1250
6	2500	2500	0	0	0	0
7	0	0	2500	2500	0	0
8	0	0	0	0	2500	2500
9	1250	1250	0	0	1250	1250
10	833	1250	834	0	833	1250
11	2500	2500	0	0	0	0
12	0	0	2500	2500	0	0
13	0	0	0	0	2500	2500
14	1250	1250	1250	1250	0	0
15	833	1250	833	1250	834	0

4. EVALUATION PROCEDURE

Training and test were performed for deep learning diagnostic models. Each training was iterated 10 epochs saving trained diagnostic models at each epoch. The epoch was defined as one cycle of training using the training data consist of 5000 waveforms. After that, test data were input to the trained diagnostic models. In each training of the model, the order of training data and initial value of weight coefficient in each layer were set randomly, this influenced the training results. Therefore, each combination of training and test were performed 100 times and the evaluation results are shown later in this paper are the average of these values. As a performance indicator of the trained diagnostic model, F-score (a.k.a. F measure) was calculated in accordance with Table 4 and Eq. (1). When diagnostic accuracy is high, F-score is close to 1. On the other hand, F-score decreases with the increase of diagnostic error. If training of a binary classification fails completely, all outputs of the trained model are the same class. In such a state, F-score decreases to about 0.33 in this case.

In addition, in order to evaluate the robustness of the diagnostic model for noise, which might be important when the model is applied to the field, five levels of noise signals were added to each normalized waveform of test data. The noise signals were generated simulating Gaussian distribution. Numpy, which is the package with Python, was used to generate the noise signals. An average value of each noise signals was 0 and standard deviations σ were 0.1, 0.2, 0.5, 1.0 and 2.0.

The trainings and tests as described above were carried out for CNN-LSTM model, CNN model and LSTM model.

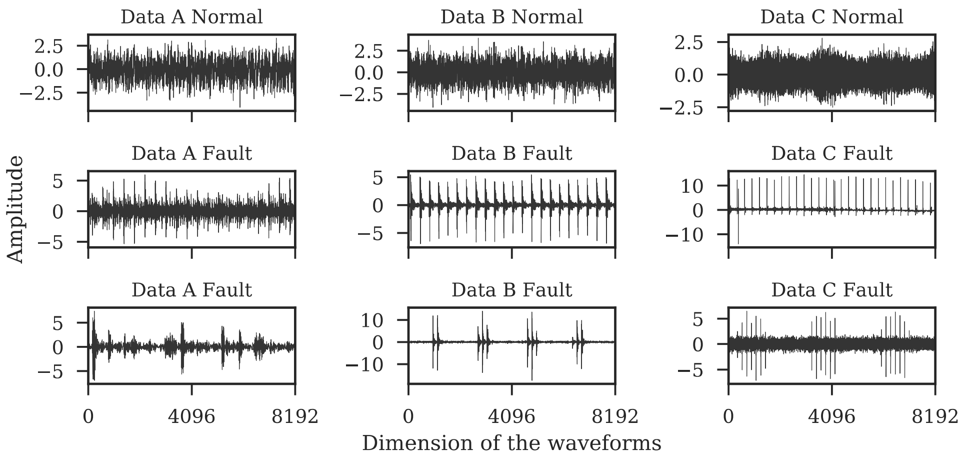

Figure 2. Examples of normalized acceleration waveforms of Data A, B and C.

Table 4. Classification of prediction results. (Confusion matrix).

		Actual class	
		Normal	Fault
Predicted class	Normal	TN (True Negative)	FN (False Negative)
	Fault	FP (False Positive)	TP (True Positive)

$$\text{F-score} = \frac{\frac{TP}{TP+FP}\times\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}} + \frac{\frac{TN}{TN+FP}\times\frac{TN}{TN+FN}}{\frac{TN}{TN+FP}+\frac{TN}{TN+FN}} \qquad (1)$$

5. RESULTS OF EVALUATION

In this section, generalization performance of the diagnostic models is described from evaluation results under influences of various training data. First, the results of the CNN-LSTM model are described, followed by comparison with the results of CNN model and LSTM model.

5.1. Results of CNN-LSTM model

Figure 3 shows test results using single data set for training (Test No.1,2,3,6,7,8,11,12,13). It is quite obvious from Figure 3 that high F-scores are achieved when training data and test data were selected from the same dataset. On the other hand, when the training data and the test data were different, F-scores remarkably deteriorated. This result showed that it is easy to achieve high diagnostic accuracy as long as the training data and the test data are acquired under the same condition. These results also suggested that high accuracy of trained diagnostic model does not mean high generalization performance under the state that training data and test data are acquired in the same condition.

Figures 4, 5 and 6 show the test results when different datasets were used for training and test. Generalized performance of the trained diagnostic models can be improved by adding only normal, i.e., not failed data to training data from the test target data when using different data in training and test.

Figure 4 shows results of tests using test data A in cases that training data does not include fault data of data A (Test No. 2, 3, 4, 5). F-scores were improved when using mixed training data than when using single condition training data.

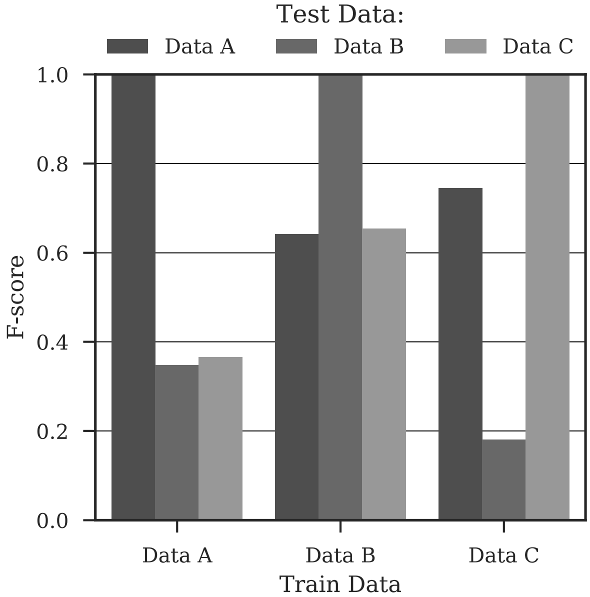

Figure 3. Test results of the model trained with single condition data.

Figure 5 is almost same as Figure 4, but for data B and excluded training data is fault data of data B (Test No. 6, 8, 9, 10). F-score of the diagnostic model using the training data A + C + Bn achieved the maximum score of 0.9, which was the highest in this figure.

Figure 6 is almost same as Figures 4 and 5, but for data C and excluded training data is fault data of data C (Test No. 11, 12, 14, 15). F-score of the diagnostic model using the training data A + B + Cn was a maximum score in this figure. However, F-score of the diagnostic model using the training data A + B were lower than F-score of the diagnostic model using only the training data B.

From these results, it is revealed that generalization performance of diagnostic model tends to be improved by using training data of various conditions data. In particular, it is effective for improvement of generalization performance to add only normal data of a test target to training data.

Figures 7, 8 and 9 show the evaluation results on the influence of noise on diagnostic accuracy. The noise added test data were diagnosed by using each trained diagnostic model at epoch 10.

Figure 7 shows results of tests using test data A (Test No. 1-5) with a noise component. In the case of $\sigma = 2.0$, F-score of the models trained with the single condition data decreased to the level of the failed model, about 0.33. However, the model trained with data B + C + An maintained the F-score over 0.4.

Figure 8 shows results of tests using test data B (Test No. 6-10) with a noise component. In the case of σ = 2.0 added, F-scores decreased to the level of the failed model. However, F-score of the diagnostic model trained with training data A + C + Bn provided a little bit better F-score than the others.

Figure 9 shows results of tests using test data C (Test No. 11-15) with a noise component. In the case of σ = 2.0, F-score of the diagnostic model trained with training data A + B + Cn provided about 0.45, better F-score than the others.

These results showed that the accuracy of the diagnostic model was kept almost same as no noise level for the noise intensity was 0.5 or less. In addition, loss of diagnostic accuracy was less for the noise intensity is 1.0 or more, if the model was trained with various conditions including the same condition data as those of test target.

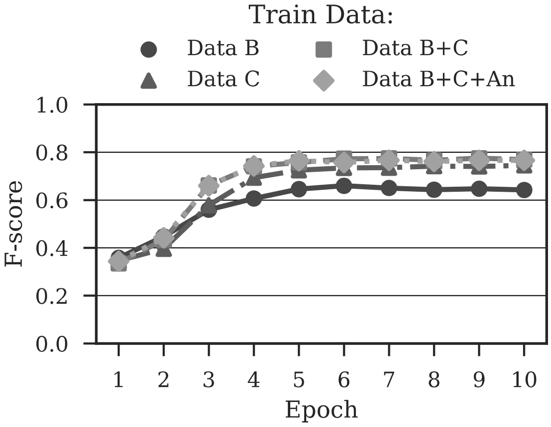

Figure 4. Test results of the CNN-LSTM model for data A trained without fault data of data A.

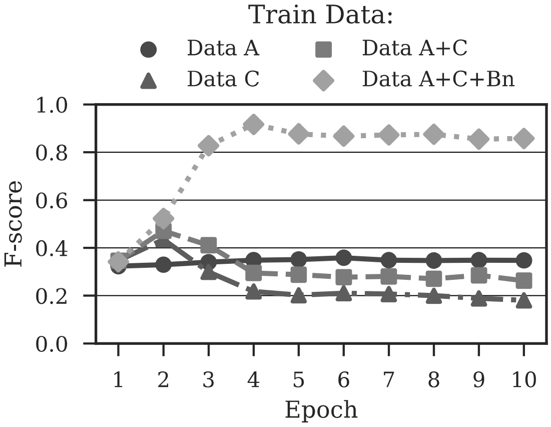

Figure 5. Test results of the CNN-LSTM model for data B trained without fault data of data B.

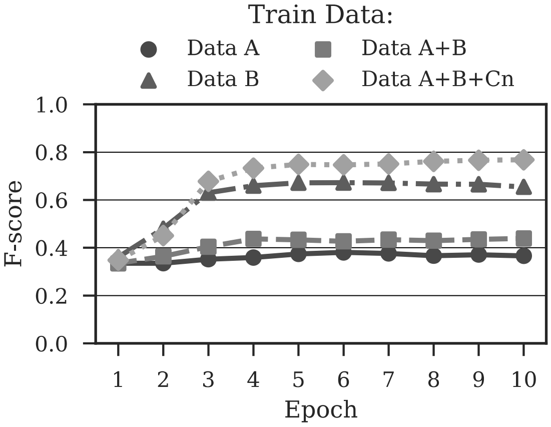

Figure 6. Test results of the CNN-LSTM model for data C trained without fault data of data C.

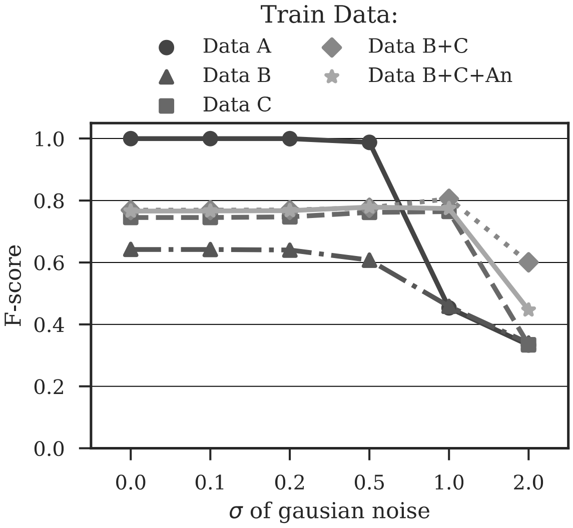

Figure 7. Test results of the CNN-LSTM model for data A with noise component.

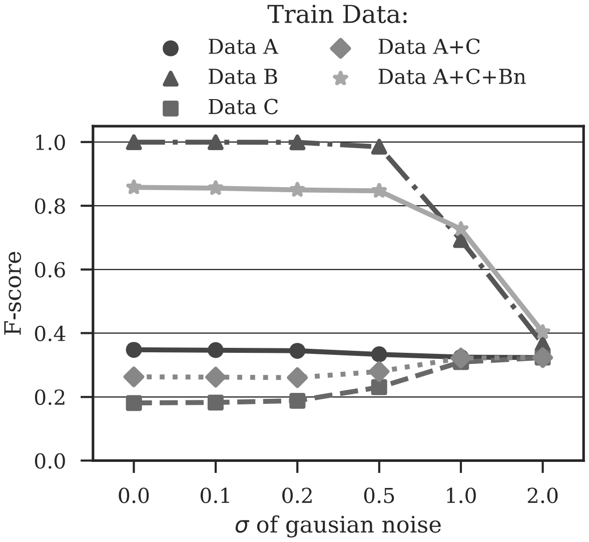

Figure 8. Test results of the CNN-LSTM model for data B with noise component.

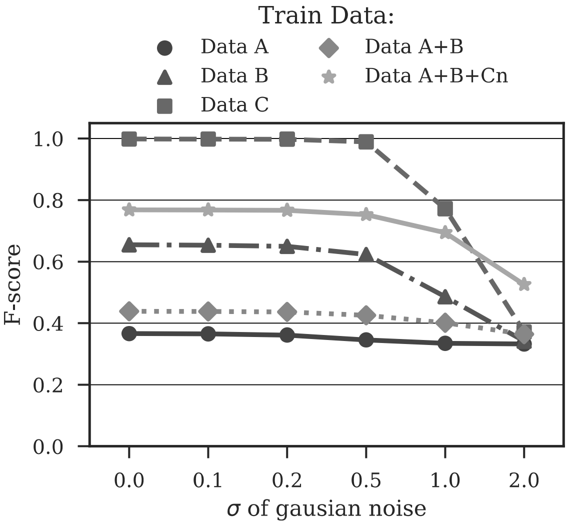

Figure 9. Test results of the CNN-LSTM model for data C with noise component.

5.2. Comparison with other models

Generalization performance was improved in many cases of the CNN model and the LSTM model, as in case of the CNN-LSTM model, by training with mixed data including normal data. Also, even loss of F-score were small for test data noise added. However, the evaluation results for part of test data showed a tendency different from that for the CNN-LSTM model.

Figures 10 and 11 show the part of evaluation results using the CNN model and the LSTM model on the influence of noise on diagnostic accuracy. The noise added test data were diagnosed by using each trained diagnostic model at epoch 10.

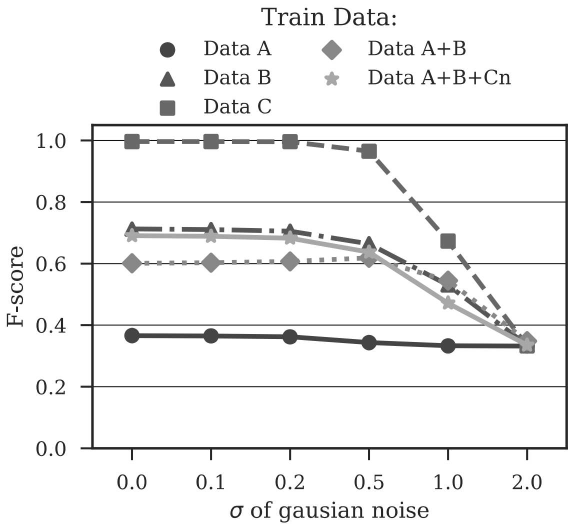

Figure 10. Test results of the CNN model for data C with noise component.

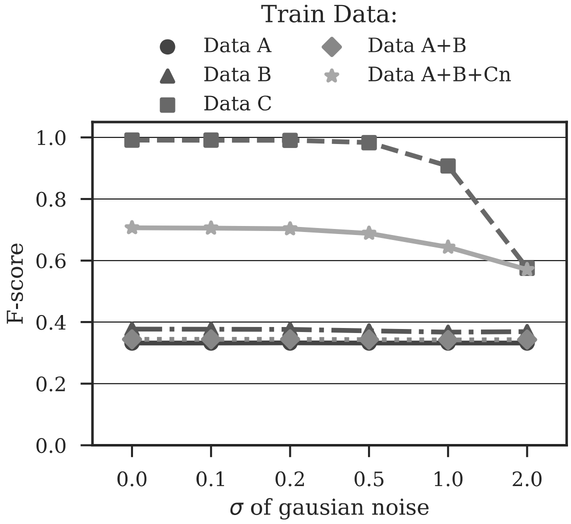

Figure 11. Test results of the LSTM model for data C with noise component.

Figure 10 shows results of test using the CNN model and test data C (Test No.11-15). Unlike in the case of CNN-LSTM model, there were no improvement of generalization performance by training using mixed data. In addition, when the noise σ = 0.5 or more, the loss of F-score due to test data noise were the same level as the evaluation result of the model trained using single condition data.

Figure 11 shows results of test using the LSTM model and test data C (Test No.11-15). Like in the case of CNN-LSTM model, there were improvement of generalization performance by training using mixed data with normal data. Also, the loss of F-score due to test data noise were small when the noise σ = 0.5 or more. However, the F-score in the case of CNN-LSTM model (figure 9) were higher than in the case of LSTM model when noise σ was less than 2.0 and training data were mixed data include normal data. Therefore, it is inferred that more general features were extracted in the case of CNN-LSTM model.

These results show that in the diagnostic model using deep learning, not only the CNN-LSTM model, can be improved of generalization performance by training with mixed data including normal data. In addition, the CNN-LSTM model can obtain generalization performance improvement more stable than the CNN model or the LSTM model.

6. CONCLUSIONS

In this study, the CNN-LSTM diagnostic models have been applied, which are combined with two methods, CNN and LSTM. The applied models are designed to detect the flaking occurred in bearings. The binary classification process is adopted to detect the fault from the vibration waveforms of the accelerometers. The training method is proposed that using data of various test rigs and various bearings with an artificial defect. Generalization performance of the diagnostic models has been investigated with the various combinations of the training datasets acquired from three types of test rigs.

The models trained with single condition dataset showed low diagnostic accuracy for test data which were different from the data for training. If training data consist of the normal / fault data provided from other test rigs and only normal data from the target test rig, it was found that the diagnostic accuracy was improved. It is suggested that damage data of diagnostic target would be unnecessary for condition monitoring in the field. The effect of the noise component included in the normalized test data also has been investigated. When the value of noise σ was 0.5 or less, the diagnostic accuracy of the models did not decrease substantially. When the value of noise σ was 1.0 and 2.0, however, the diagnostic accuracy of all models was lower than in case of no noise, diagnostic accuracy was less likely to decrease of the models trained with proposed method than other models. The CNN-LSTM models gained these advantages more stable than the CNN models and the LSTM models. From these results, CNN-LSTM models could diagnose bearing flaking with high accuracy and robustness by using the proposed training method.

In future work, it would be necessary to verify, whether the essential features of the vibration waveforms due to flaking would be extracted correctly through the applied method, and how it changes as training progress.

REFERENCES

Chen, Z., Deng, S., Chen, X., Li, C., Sanchez, R.V., & Qin, H. (2017). Deep neural networks-based rolling bearing fault diagnosis. *Microelectronics Reliability.* vol. 75, pp. 327-333.

Feng, J., Yaguo, L., Liang, G., Jing, L., & Saibo, X. (2018). A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing.* vol. 272, pp. 619-628.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks.* vol. 18, no. 5-6, pp. 602-610.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation.* vol. 9, no. 8, pp. 1735-1780.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the 2012 advances in neural Information processing systems*. December 3-6, Lake Tahoe, NV. pp. 1097-1105

Laparo, K. A. (2012). Case Western Reserve University Bearing Data Center: http://csegroups.case.edu/bearingdatacenter/home

Le Cun, B. B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard,. L. D., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, November 26–29, Denver, CO.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *IEEE international conference on acoustics, speech and signal processing*, May 26-31, Vancouver, BC. doi:10.1109/ICASSP.2013.6639343

Mao, W., He, J., Li, Y., & Yan, Y. (2016). Bearing fault diagnosis with auto-encoder extreme learning machine: A comparative study. *Proceedings of the Institution of Mechanical Engineers Part C: Journal of Mechanical Engineering Science.* vol. 231, no. 8, pp. 1560-1578.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical systems and signal processing.* vol. 25, no. 2, pp. 485-520.

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical systems and signal processing.* vol. 64, pp. 100-131.

Zhang, W., Li, C., Peng, G., Chen, Y., & Zhang, Z. (2018). A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical systems and signal processing.* vol. 100, pp. 439-453.

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2016). Deep learning and its applications to machine health monitoring: A survey. *arXiv preprint arXiv:1612.07640.*

Zhao, R., Yan, R., Wang, J., & Mao, K. (2017). Learning to monitor machine health with convolutional bi-directional LSTM networks. *Sensors.* vol. 17, no. 2, 273