Iranian Society of Acoustics and Vibration
The 12th International Conference on Acoustics and Vibration ISAV2022
14-15 Dec 2022 Tehran - Iran
Amirkabir University of Technology

The application of audio signals in gear fault diagnosis based on deep learning methods: an end-to-end approach

Hassan Alavi[a,b], Abdolreza Ohadi[a,b]

[a] *Acoustics Research Laboratory, Mechanical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran*

[b] *Vehicle Technology Research Center, Technology Institute of Mechanical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran*

** Corresponding author e-mail: a_r_ohadi@aut.ac.ir*

Abstract

Diagnosing gearbox faults based on audio signal has received less attention in researches, although due to the non-contact nature of the microphone, it makes the diagnosis process more accessible. In this article, based on the methods of deep learning, the diagnosis of crack and uniform wear of the gearbox in three different scenarios of (1) constant fault severity and working conditions, (2) constant fault severity and different working conditions, (3) different fault severity and different working conditions have been investigated. State-of-the-art methods of Convolutional Neural Network (CNN), Deep Residual Neural Network (DRN) and a proposed hybrid network of CNN and Long Short-Term Memory (LSTM), all applied based on end-to-end approach, have been investigated. The results show that the CNN+LSTM has a better performance than other methods, in such a way that in the most difficult case, i.e. different fault severity and different working conditions, it classifies the faults with an accuracy of 88.8%. In addition, the computational cost of training the proposed network is less than other networks.

Keywords: Fault Diagnosis; Gearbox; Deep learning; End-to-end approach.

1. Introduction

Gearbox is a commonly used component in many industrial, transportation, and energy conversion applications. Gear faults may develop during operation due to excessive or improper use. They can be divided into two categories: distributed faults such as tooth wear and localized faults such as tooth root crack. By implementing a reliable diagnosis system, faults can be identified and repair measures can be carried out to prevent high economic costs due to the complete shut-down of the

machine without prior planning. Fault signatures could be extracted from different sources; among them, vibration signals, acoustic emission, audio, temperature, and oil debris are widely discussed.

Although the audio signal is not as reliable as the vibration signal for fault detection, the use of the audio signal makes the diagnosis process more accessible due to the fact that there is no need for the sensor (microphone) to be in direct contact to the machine. Baydar and Ball conducted a comparative study on the ability of vibration and audio signals in diagnosis of localized wear, crack and broken tooth in a gear transmission system[1]. They concluded that the audio signal is more capable in diagnosis of crack but less capable in diagnosis of wear and broken tooth. Hou et al proposed the application of near-field acoustic holography in diagnosis of gear pitting and chipping faults[2]. Vanraj et al proposed the application of Teager-Kaiser energy operator to extract features from audio signal. A kNN classifier has been used to classify different chipping fault severities in gears[3]. Parey and Singh proposed the application of continuous wavelet transform and energy to Shannon entropy as features to diagnose crack and chipping in gears. An adaptive neuro-fuzzy inference system classifies the faults.

In recent years, with the advancement of machine learning theories, and parallel to it, the development of computing ability of computers, deep learning methods have received attention. Deep learning techniques can take the raw or slightly processed signals as input and perform feature extraction and selection as part of the learning process. This approach called “end-to-end” approach in fault diagnosis and is in contrast to traditional methods that require expert’s knowledge to extract features from the signal. The deep learning, as a state-of-the-art topic, has been widely concerned since Hinton et al. proposed a fast learning algorithm to train deep belief networks[4]. Thereafter, other deep learning methods were developed and applied in fault diagnosis of machinery. Among them, Convolutional Neural Networks (CNN)[5-8], Deep Residual Neural Networks (DRNN)[9, 10], and Long Short-Term Memory (LSTM)[11, 12] have a history of use in the fault diagnosis.

In this article, an end-to-end hybrid CNN and LSTM network has been proposed and its ability to classify wear and crack faults in a helical gearbox has been investigated under different scenarios. Furthermore, the accuracy of CNN+LSTM has been compared to conventional CNN and DRNN. The rest of this article is organized as follows. In section 2, the theoretical background on CNN, DRNN, and LSTM has been introduced and detailed architectures of networks has been depicted. In section 3, the experimental test rig, faults and raw signal preparation has been described. In section 4, the results has been presented and discussed. Finally, in section 5, some concluding remarks have been drawn and ideas for future studies has been proposed.

2. Theoretical background

In this section, the theory of conventional CNN and two of its extensions, namely, DRNN and CNN+LSTM has been briefly introduced.

2.1 Convolutional neural networks

A conventional CNN cold be implemented by using a large variety of layers. In the following, the most important and widely used layers has been described.

(1) Convolution layer

The convolution layer convolves a filter or kernel its input matrix. According to the dimension of the input. Depending on the dimensionality of the input, the kernel could be a vector or higher dimensional matrix. In the processing of raw audio signals, the kernel is one-dimensional and has a series of weights and a bias as learnable parameters. Eq. **Error! Reference source not found.** shows how the convolution layer works:

$$u_i^l = f\left(\sum_{m=1}^{M} k_i^l x_{i+m}^{l-1} + b\right) \quad (1)$$

in which x^{l-1} and x^l denote, respectively, the input and output vector of the convolution layer, $f(.)$ is the activation function, $'*'$ represents the convolution operator, k^l is the kernel's weight vector, and b is a scalar bias. Finally, M is the filter (kernel) size.

(2) Activation layer

The most common activation function used in CNNs is Rectified Linear Unit (RELU) function:

$$\sigma(u_i^l) = RELU(u_i^l) = max(0, u_i^l) \quad (2)$$

(3) Batch normalization layer

The application of batch normalization layer is optional, however, it can enhance the rate of convergence of the network via normalizing the mini batch samples in such a way that the mean and standard deviation of its input become close to, respectively, zero and one.

(4) Pooling layer

After each activation layer, usually, a pooling layer reduces the dimensionality by down-sampling. Having pooling layers in the network architecture has two advantages; first, it decreases the likelihood of overfitting in the network via reduction of the overall learnable parameters. Second, it speeds up the network convergence. Among the methods of performing the pooling operation, the Max-Pooling method is the most commonly used in CNN studies. The max-pooling operation can be described as:

$$x_i^l = max(x_{i:i+Q}^{l-1}) \quad (3)$$

where x^{l-1} and x^l are, respectively, the input and output vector of the Max-pooling layer, and Q is the pool size. The pooling layer has no learnable parameters.

(5) Fully-connected layer

A fully connected layer, similar to that used in the multi-layer perceptrons, is used to estimate the probabilities that an input belongs to classes. To transform probabilities to exact decisions, Softmax layer assigns the class label to the most probable class. The loss function of the network, when more than two class labels are exist, is usually defined as the cross-entropy between predicted and expected labels:

$$Loss = -\sum_{i=1}^{N} y_j ln p_j \quad (1)$$

where y_j is expected output, p_j is predicted output, and N is the number of samples.

2.2 Deep residual neural networks

Deep residual neural network is an extension to CNN. There is a problem in the training process of a conventional CNN that the gradients of loss function caused by last layers does not penetrate to first layers in backpropagation algorithm. This phenomenon usually called the "vanishing gradients"[13] and results in difficulties in the deep network training. The presence of a shortcut mapping as shown in Figure 1 could ease the penetration of gradients in backward direction. Ref. [14] gives a clear mathematical description of the residual connections on the performance of deep residual networks. Its worth mentioning that usually the shortcut mapping has been regarded as unity (H(x) = 1), however, it is possible to replace it with other types of layers such as convolution.

2.3 Long short-term memory networks

Another advancement in the field of machine learning is to develop the concept of Recurrent Neural Networks (RNN) that mimics a discrete-time dynamical system by using feedback connections between neurons. A disadvantage of the RNNs is that they suffer from vanishing or suddenly increasing gradients. The concept of LSTM has been developed to overcome this disadvantage by inducing a constant error flow throughout the network[15]. An LSTM unit has three state variables that help the network to reduce the long term dependency to input sequences. These states are called the "forget", the "input", and the "output". The forget state variable eliminates redundancy in sequences. The input state variable processes the incoming sequences, and the output state variable evaluate the dynamic connection between input and current state of the unit. The diagram of an LSTM unit is shown in Figure 2.

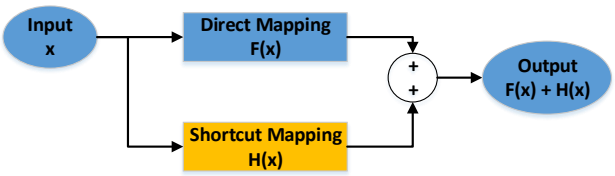


Figure 1 Shortcut connection in residual neural networks

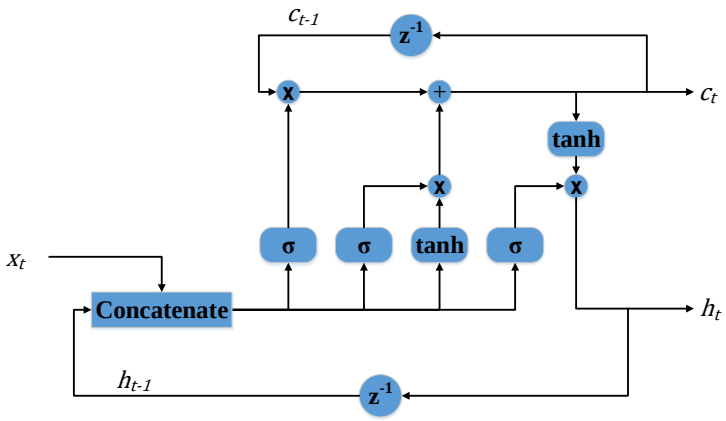


Figure 2 A long short-term memory unit

In the current article, a hybrid CNN+LSTM network has been employed to treat raw audio signals because LSTM layer can enhance features extracted by CNN layers according to the above-mentioned descriptions. To make a combination of LSTM and CNN, sequence folding/unfolding layers is necessary for the sake of compatibility between predefined objects in MATLAB™ environment. Additionally, since the pooling layer is not applicable after LSTM layer to prevent overfitting of the network, a dropout layer has been considered to alleviate overfitting. Detailed block diagrams of the three networks used in this paper has been depicted in Figure 3. The network architectures has been employed to classify wear and crack faults on experimental data sets.

3. Experimental test set-up

In this part, the experimental test rig, induced faults, and data acquisition procedure has been discussed. After that, the preparation of data for feeding to the deep classifiers has been introduced.

3.1 Test rig and faults

Figure 4 shows the experimental test rig and the location of microphone and optical tachometer. The central part of the experimental test rig is a Peugeot/Citroen™ BE3 gearbox. The gearbox is

driven by an induction motor via three parallel V-belts and sheaves. A laser tachometer made by Contrinex™ is located in the vicinity of the gearbox's input shaft, and produces a pulse in each rotation of the gearbox's input shaft. A BWSA Tech™ MA231 microphone measures the sounds produced by the gearbox during operation. An Advantech™ PCI-1712-12bit with sampling frequency of 40KHz digitizes the microphone and the tachometer data.

Figure 5 shows the wear fault at different severities of 5, 10 and 20 micrometre. Furthermore, crack fault at severities of 20%, 50%, and 80% of the tooth root depth has been illustrated in this figure. The tests has been conducted under two torques of 35N.m and 45N.m, and three motor speeds of 30Hz, 35Hz and 40Hz. The audio and tachometer signals has been measured and recorded for 10 seconds and each tests repeated three times.

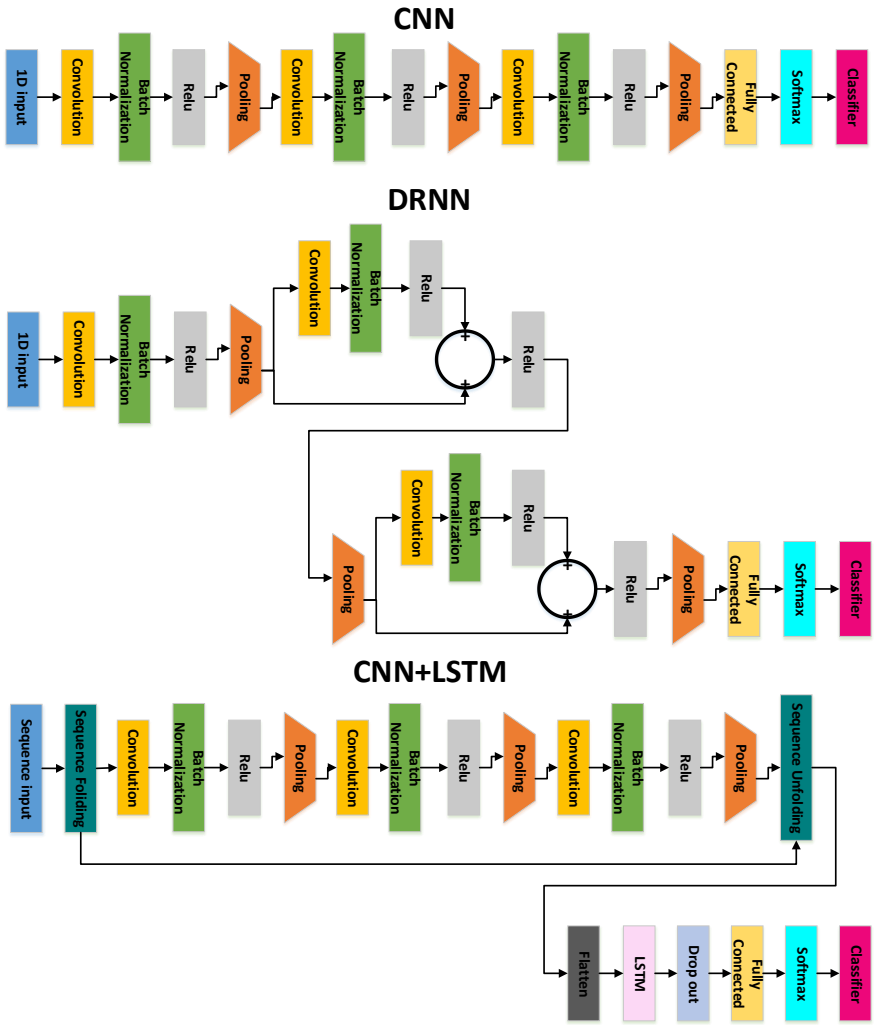
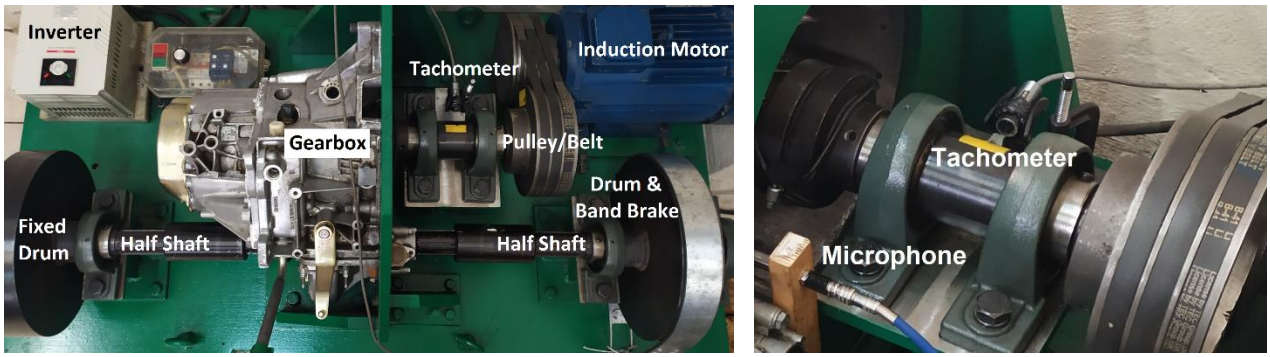


Figure 3 Three deep network architectures of the current study



(a) **(b)**

Figure 4 (a) Experimental test rig and its components **(b)** close-up view of sensors

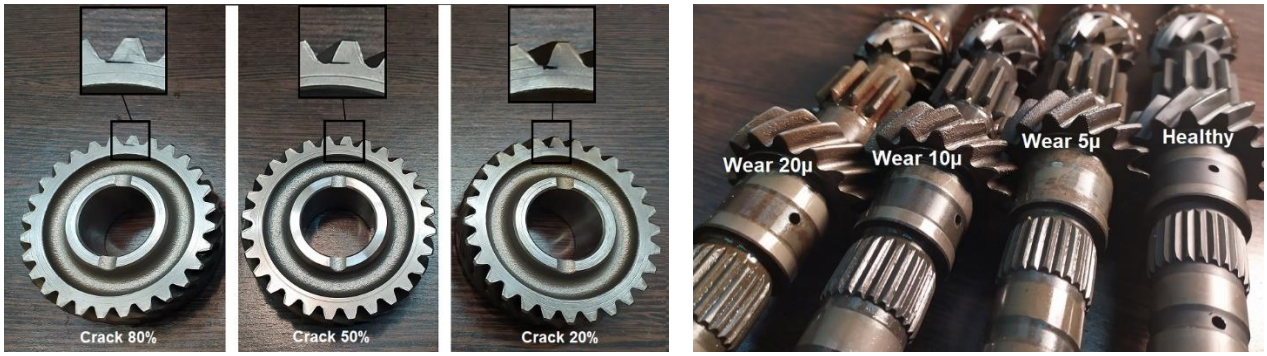


Figure 5 Gear wear and crack faults at different severities

3.2 Raw signal preparation

The recorded signals should be sectioned into equal parts to be fed to the deep learning classifier. In rotating machinery fault diagnosis, each rotation of shaft can be regarded as a complete representation of the condition of the machine in time domain. Therefore, tachometer pulses has been utilized to divide the acquired audio signal into training/validation samples. For this aim, the tachorpm command of MATLAB™ has been used to analyse the tachometer signal and calculate proper sectioning times. Figure 6 shows the result of tachorpm on a sample of tachometer signal. The red "+" signs in the figure depicts the detected pulses. The audio signal between two consecutive detected pulses can be regarded as a sample. To have equal length samples, the audio signal between to pulses of tachometer resampled to a signal with a length of 500 using interp1 command of MATLAB™. The reason behind the resampling is two-fold; first, the signals has been acquired in different speed conditions, and second, even in measured signals at equal motor speeds, the presence of belt and pulley mechanism in transmission of the motor power to the gearbox cause unavoidable slip, and therefore, the length of signals between two adjacent pulses are slightly differ.

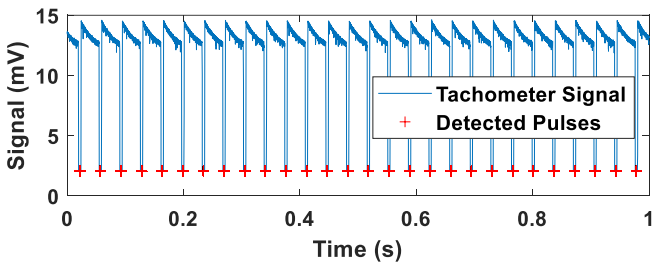


Figure 6 Tachometer signal and detected pulses used for audio signal sectioning

4. Results and discussion

In this section, the performance of the deep learning network architectures presented in section 2 has been evaluated under three different scenarios. First, the fault classification ability of the classifiers under constant fault severity and working conditions has been evaluated. Second, the classification accuracy under constant fault severity and different working conditions has been assessed, and third, the classification accuracy under different fault severity as well as working conditions has been studied.

4.1 Scenario one: Constant fault severity and working condition

The scenario of constant fault severity and working condition can be divided into three sub-scenarios i.e. the mild, the moderate and the severe cases. In this scenario, the classifier should learn to assign three labels to each observation which has been defined in Table 1. In these three cases, the data acquired under the torque of 45N.m and the speed of 40Hz has been used to train the networks. A randomly selected 20% portion of data has been chosen for validation of the network and the remaining 80% has been used for training. The training diagrams of CNN, DRNN, and CNN+LSTM has been shown in Figure 7. For the sake of conciseness, only the training diagrams of mild case has been shown in the figure. It can be seen that the training of DRNN is more challenging since both the training and validation curves fluctuates more than CNN and CNN+LSTM. Furthermore, the DRNN converges more slowly to the trained situation than other two networks.

Table 1 Fault labels in scenarios one and two

Mild		Moderate		Severe	
Label	**Condition**	**Label**	**Condition**	**Label**	**Condition**
H	Healthy	H	Healthy	H	Healthy
C	Crack 20%	C	Crack 50%	C	Crack 80%
W	Wear 5µm	W	Wear 10µm	W	Wear 20µm

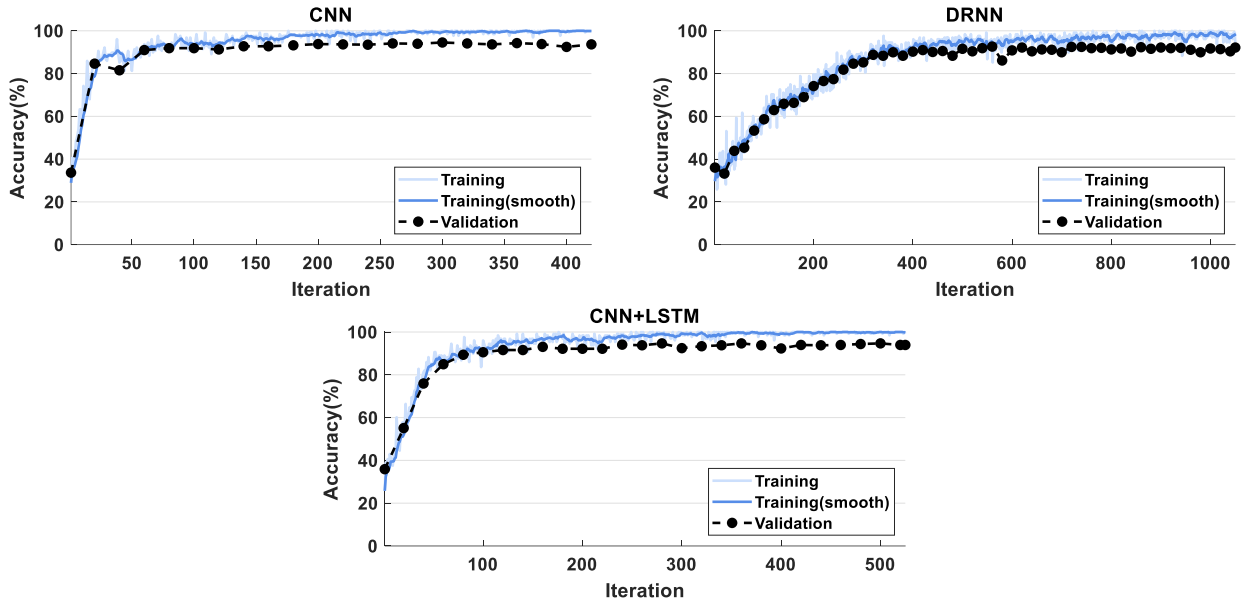


Figure 7 Training chart of CNN,DRNN, and CNN+LSTM networks

The confusion matrices of all sub-scenarios in this sections has been illustrated in Figure 8 and the final validation accuracy and training time of the networks on the three sub-scenarios has been reported in Table 2. It should be noted that the training times evaluated on a system with Intel™ Core i7 equipped with 8GB of RAM and running MATLAB™ 2021b under Windows™ 10 64Bit. It can

be seen that the training time of the CNN+LSTM is slightly lower than CNN, but the training time of the DRNN is dramatically higher than both CNN and CNN+LSTM. The accuracy of CNN+LSTM is slightly higher in mild faults, while the accuracy of CNN at moderate and severe faults is incrementally higher than CNN+LSTM. The accuracy of DRNN is not as good as both CNN and CNN+LSTM. It can be seen from the confusion matrices that more confusions occur between crack and healthy conditions rather than wear and healthy conditions in all fault severities. It means that the diagnosis of crack is more difficult than the diagnosis of wear.

Figure 8 Confusion matrices of CNN, DRNN and CNN+DRNN under mild, moderate and severe faults

Table 2 The validation accuracy of CNN, DRNN and CNN+DRNN under mild, moderate and severe faults

Mild		**Moderate**		**Severe**		
Network	**Accuracy**	**Network**	**Accuracy**	**Network**	**Accuracy**	**Average Training Time (sec)**
CNN	93.8%	CNN	**98.2%**	CNN	**99.4%**	100
DRNN	92.5%	DRNN	96.0%	DRNN	97.8%	482
CNN+LSTM	**94.8%**	CNN+LSTM	97.9%	CNN+LSTM	98.8%	77

4.2 Scenario two: Constant fault severity at different working condition

In this case, the fault labels are similar to Table 1, but the data acquired under all load and torque conditions has been fed to the networks. Since, according to section 4.1, the mild case is more challenging case in fault diagnosis, only the mild case has been considered in this scenario. Again, 20% of total data has been used for validation and 80% has been used for training. The confusion matrices for mild case under different load and torque conditions has been shown in Figure 9. A comparison between accuracy of all networks in mild case under constant and different load/speed has been reported in Table 3. It can be seen that the classification accuracy of CNN+LSTM surpasses the other classifiers as well as it has less drop in accuracy due to different working conditions rather than the other two networks. The confusion matrices depicts that even in different working conditions and mild fault severity, the confusion between wear and crack is minimal and almost all confusions occur between healthy and faulty conditions.

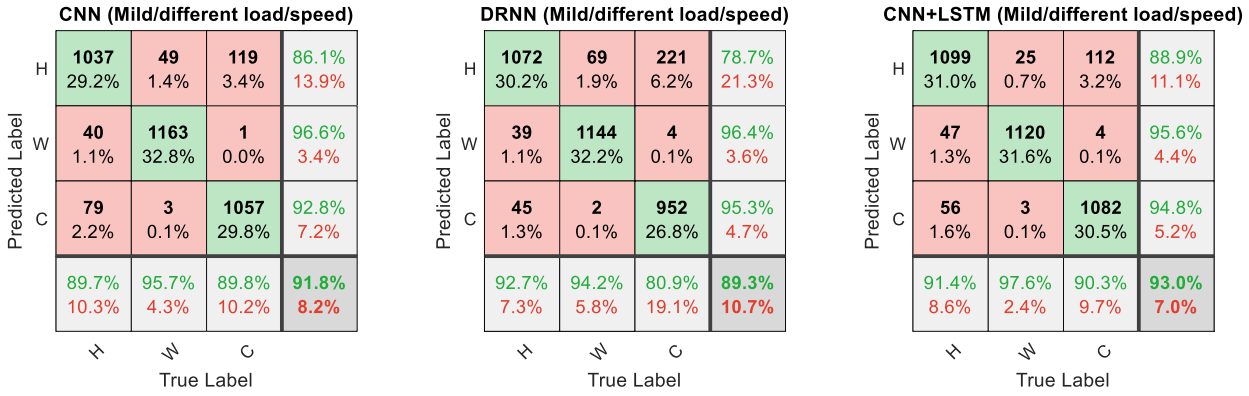


Figure 9 Confusion matrices of CNN, DRNN and CNN+DRNN under mild faults and different working conditions

Table 3 Validation accuracy of CNN, DRNN and CNN+DRNN under mild faults and different working conditions

Network	Accuracy (Mild/Constant load/speed)	Accuracy (Mild/different load/speed)	Percent drop
CNN	93.8%	91.8%	2.1%
DRNN	92.5%	89.3%	3.5%
CNN+LSTM	94.8%	93.0%	**1.9%**

4.3 Scenario 3: Different fault severity at different working conditions

This scenario is the hardest scenario for a classifier. The aim is to discriminate between different type of faults as well as different severities of each fault under different working conditions. In this scenario, the classifier should select a label among a set of seven labels for each observations. The condition labels has been defined in Table 4.

Table 4 Fault labels in the third scenario

Label	Condition	Label	Condition
H	Healthy	W5	Wear 10μm
C20	Crack 20%	W10	Wear 10μm
C50	Crack 50%	W20	Wear 20μm
C80	Crack 80%		

The confusion matrices of the three networks has been portrayed in Figure 10. The final validation accuracy of the networks has been reported in Table 5. It can be seen that in this scenario, the accuracy of all networks is significantly less than those of constant fault severity in scenario 2. Still, the CNN+LSTM has better classification accuracy than CNN, and CNN outperforms DRNN. By investigating the confusion matrices, it can be seen that all networks provide their best performance in detection of severe wear and their weakest performance in detection of mild crack. Furthermore, their overall performance in detection of wear at different severities is superior than their overall performance in detection of different severities of crack. It's worth mentioning that despite the overall superiority of CNN+LSTM in classification accuracy, the misclassification rate of CNN and DRNN in detection of mild crack is less than that of CNN+LSTM.

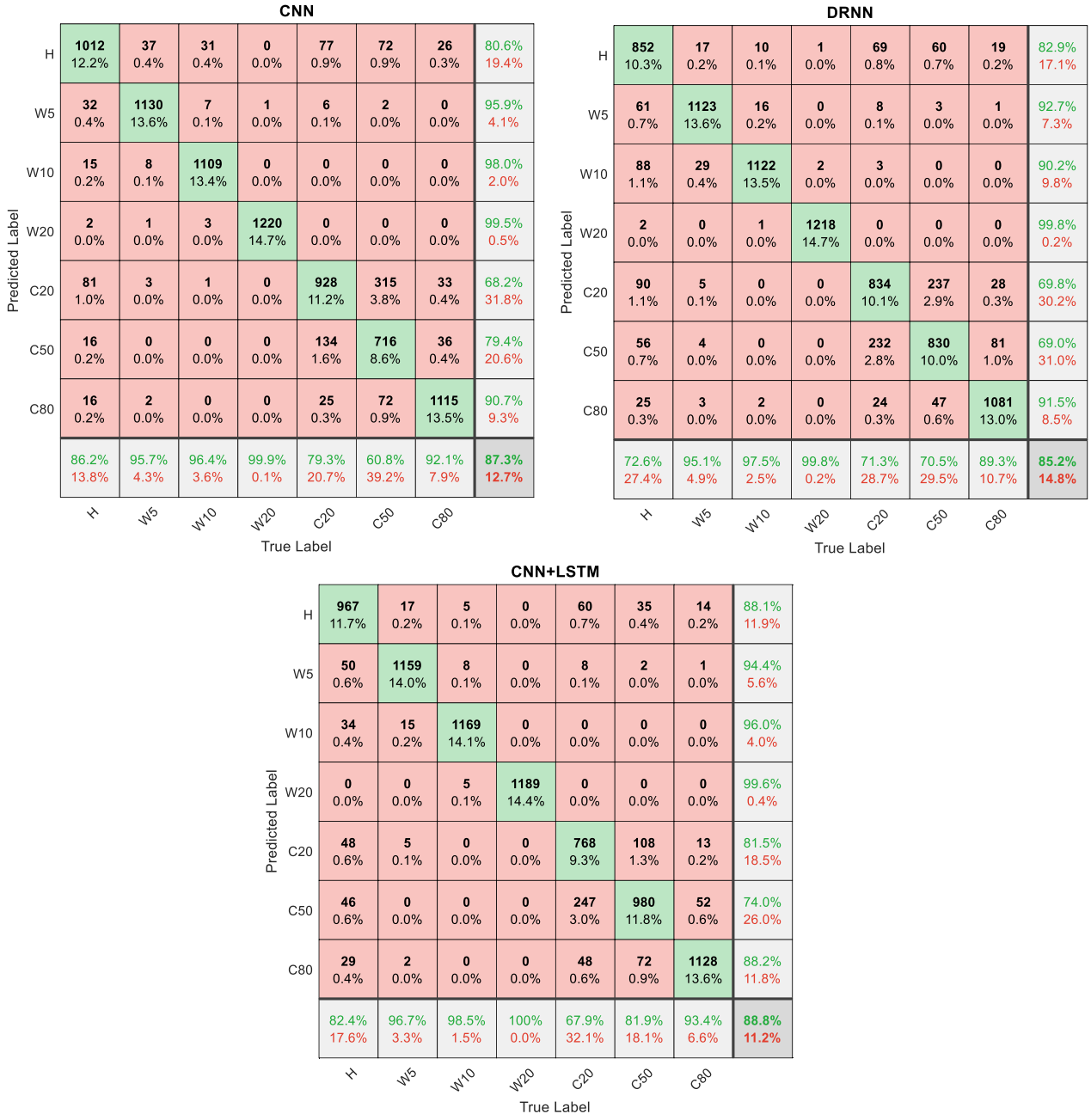


Figure 10 Confusion matrices of CNN, DRNN and CNN+DRNN under third scenario

Table 5 Validation accuracy of CNN, DRNN and CNN+DRNN under third scenario

Network	Accuracy
CNN	87.3%
DRNN	85.2%
CNN+LSTM	**88.8%**

5. Conclusion

In this article, end-to-end approach on diagnosis of wear and crack faults in an automotive gearbox has been considered. In the end-to-end approach, there is no need to experts' knowledge to define and extract features from the signal. Instead, a deep classifier performs both feature extraction and classification of the faults.

An architecture for deep classifier based on a combination of CNN and LSTM has been proposed and investigated under three fault scenarios:

(1) Constant fault severity/constant working condition
(2) Constant fault severity/different working condition
(3) Different fault severity/different working condition

Additionally, a conventional CNN and a DRNN has been used for comparison purposes. The results show that the overall performance of CNN+LSTM regarding all scenarios is better than the other two classifiers. In the first scenario, with mild fault severity as a more challenging case, the classification accuracy of CNN+LSTM is 94.8% while the accuracy of CNN and DRNN are 93.8% and 92.5%, respectively. In the second scenario, the accuracy drop of CNN+LSTM is 1.9%, which is better than that of 2.1% and 3.5% of CNN and DRNN. In the third scenario, the CNN+LSTM outperforms the other two methods by showing an accuracy of 88.8% while the accuracy of CNN and DRNN are 87.3% and 85.2% respectively.

By examining the confusion matrices of the third scenario, it can be seen that the accuracy of mild crack detection in CNN+LSTM is worse than CNN and DRNN while in other fault labels it performs better than CNN and DRNN. It suggests an idea for future works to fuse the decisions of different network architectures to attain higher classification accuracy. Another idea for future works is to fuse different sources of data such as vibration and audio signal or acoustic emission and audio signal.

REFERENCES

[1] N. Baydar and A. Ball, "A comparative study of acoustic and vibration signals in detection of gear failures using Wigner–Ville distribution," *Mechanical systems and signal processing*, vol. 15, no. 6, pp. 1091-1107, 2001.
[2] J. Hou, W. Jiang, and W. Lu, "Application of a near-field acoustic holography-based diagnosis technique in gearbox fault diagnosis," *Journal of Vibration Control*, vol. 19, no. 1, pp. 3-13, 2013.
[3] Vanraj, S. Dhami, and B. Pabla, "Hybrid data fusion approach for fault diagnosis of fixed-axis gearbox," *Structural Health Monitoring*, vol. 17, no. 4, pp. 936-945, 2018.
[4] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527-1554, 2006.
[5] L. Jing, M. Zhao, P. Li, and X. Xu, "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox," *Measurement*, vol. 111, pp. 1-10, 2017.
[6] L. Jing, T. Wang, M. Zhao, and P. Wang, "An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox," *Sensors*, vol. 17, no. 2, p. 414, 2017.
[7] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196-3207, 2018.
[8] S. Kim and J.-H. Choi, "Convolutional neural network for gear fault diagnosis based on signal segmentation approach," *Structural Health Monitoring*, vol. 18, no. 5-6, pp. 1401-1415, 2018.

[9] Z. Minghang, M. Kang, B. Tang, and M. Pecht, "Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes," *IEEE Transactions on Industrial Electronics,* vol. 65, no. 5, pp. 4290-4300, 2017.

[10] Z. Zhao *et al.*, "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Transactions,* vol. 107, pp. 224-255, 2020.

[11] H. Zhao, S. Sun, and B. Jin, "Sequential fault diagnosis based on LSTM neural network," *Ieee Access,* vol. 6, pp. 12929-12939, 2018.

[12] 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)*, 2016: IEEE, pp. 135-140.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[14] Y. Jin, C. Qin, Y. Huang, and C. Liu, "Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network," *Measurement,* vol. 173, p. 108500, 2021.

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