Deep Residual Networks With Dynamically Weighted Wavelet Coefficients for Fault Diagnosis of Planetary Gearboxes

Minghang Zhao, Myeongsu Kang, *Member, IEEE*, Baoping Tang, and Michael Pecht, *Fellow, IEEE*

***Abstract*—One of the significant tasks in data-driven fault diagnosis methods is to configure a good feature set involving statistical parameters. However, statistical parameters are often incapable of representing the dynamic behavior of planetary gearboxes under variable operating conditions. Although the use of deep learning algorithms to find a good set of features for fault diagnosis has somewhat improved diagnostic performance, the lack of domain knowledge incorporated into deep learning algorithms has limited further improvement. Accordingly, this paper developed a variant of deep residual networks (DRNs), the so-called deep residual networks with dynamically weighted wavelet coefficients (DRN+DWWC) to improve diagnostic performance, which takes a series of sets of wavelet packet coefficients on various frequency bands as an input. Further, the fact that no general consensus has been reached as to which frequency band contains the most intrinsic information about a planetary gearbox's health status calls for "dynamic weighting layers" in the DRN+DWWC and the role of the layers is to dynamically adjust a weight applied to each set of wavelet packet coefficients to find a discriminative set of features that will be further used for planetary gearbox fault diagnosis.**

***Index Terms*—Deep residual learning, fault diagnosis, feature learning, planetary gearbox, wavelet packet transform.**

I. Introduction

PLANETARY gearboxes [1] are superior to general parallel-shaft gearboxes owing to their higher power density and greater efficiency. They are already widely used for transmission in important mechanical systems, including gas turbines, heavy-duty trucks, helicopters, and wind turbines [2]. Due to the harsh operating environment, planetary gearboxes often encounter gear tooth pitting and root cracking [3]. Some of the failures can cause serious damage to the entire mechanical system, which can result in human safety accidents and huge economic losses. The development of an efficient fault diagnosis tool for planetary gearboxes can ensure operational reliability and reduce maintenance costs.

To date, there are two major approaches to fault diagnosis in planetary gearboxes—vibration analysis and data-driven methods. Vibration analysis (e.g., spectral analysis or envelope analysis) detects defect frequencies with the help of signal decomposition techniques, such as wavelet transform and empirical mode decomposition (EMD) [4]. However, it is difficult to observe defect frequencies in the spectrum (or power spectrum) because they are mostly hidden by low-frequency deterministic components (e.g., frequencies of $1\times$ r/min and $2\times$ r/min due to misalignment) and high-frequency noise components. Because planetary gearboxes are mostly connected to the motor, the parallel gearbox, and other rotating parts, the frequency components are more complex. Moreover, varying the rotating speed compounds the problem. Much professional knowledge, skill, and experience are required to detect the defect frequencies. The second fault diagnosis approach is based on data-driven methods. Data-driven methods use feature engineering and machine learning to detect the onset of faults, pinpoint types of faults, and predict remaining useful life [5]. The data-driven methods analyze performance data based on a training dataset. Compared to vibration analysis, these methods can be used in complex systems with multiple and potentially competing failure modes as long as the systems exhibit repeatable behaviors. This study focuses on a data-driven method for fault diagnosis of planetary gearboxes.

In traditional data-driven diagnosis approaches, researchers mostly need to extract a number of statistical parameters (e.g., root-mean-square, energy, kurtosis, and so forth) and feed these statistical parameters into machine learning algorithms, such as k-nearest neighbors, random forest, naive Bayes model, support vector machines (SVMs), and neural networks (NNs). To achieve an accurate diagnostic performance, the distributions of the aforementioned high-dimensional statistical parameters should be separable for each condition (i.e., class) to be considered in diagnosis. However, if the distributions are not separable enough for differentiating conditions, it can be challenging to achieve high diagnostic accuracy. Due to the complex structure,

Manuscript received May 31, 2017; revised September 10, 2017; accepted September 24, 2017. Date of publication October 12, 2017; date of current version January 16, 2018. This work was supported in part by the Technology Innovation Program (or Industrial Strategic Technology Development Program (10076392, Development of Vehicle Self Diagnosis System and Service for Automobile Driving Safety Improvement) funded by the Ministry of Trade, Industry and Energy, Korea) and in part by the National Natural Science Foundation of China under Grant 51775065. The work of M. Zhao was supported by the China Scholarship Council. *(Corresponding author: Myeongsu Kang.)*

M. Zhao and B. Tang are with the State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400044, China (e-mail: minghang.zhao@gmail.com; bptang@cqu.edu.cn).

M. Kang and M. Pecht are with the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: mskang@calce.umd.edu; pecht@calce.umd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2017.2762639

long transmission path, variable operating conditions, and strong background noise, the distributions will be easily overlapped, leading to low diagnostic accuracy.

Although many researchers use their domain knowledge to find a reliable set of statistical parameters by retaining as much discriminative information as possible for fault diagnosis of the planetary gearbox, there is still no guarantee that the statistical parameters can fully represent the dynamic characteristics of planetary gearboxes under variable operating conditions. Therefore, statistical parameter extraction can be an obstacle to achieve higher diagnostic accuracies.

To avoid the above problems caused by the use of statistical parameters, feature learning, which refers to learning useful features from the input observations, has been recently used for machinery fault diagnosis. Classical shallow machine learning algorithms are not able to adequately learn features from the complex, redundant, and highly variable raw vibration signals sampled at thousands or hundreds of thousands of hertz due to their limited nonlinear transformation process. Deep learning, which refers to the representation learning that has multiple layers of nonlinear transformation [6], can be a promising solution to address the limitations of the classical shallow machine learning algorithms.

To be specific, deep learning can enable a hierarchical nonlinear learning of high-level features built on top of low-level features to discriminate different health conditions (e.g., healthy or faulty). Low-level features are the basic details of the health conditions, whereas high-level features are more abstract—that is, high-level features, also known as feature maps, can be obtained by a series of nonlinear transformations through multiple layers. Deep belief networks (DBNs), deep auto-encoders (DAEs), and convolutional NNs (CNNs) are popular deep learning methods used for various fault diagnosis applications in recent years [7]–[14]. DBNs and DAEs can conduct unsupervised pretraining on the weights, which can ease the difficulty of the subsequent supervised training of the deep networks. However, a key problem in DBNs and DAEs is that there are too many weights to train when the inputs are raw vibration signals or their time–frequency representations. In contrast, CNNs and deep residual networks (DRNs) [15] can reduce the number of weights to be optimized using the strategies of local receptive field and weight sharing, which can be further effective for reducing computational burden during the training process. For example, Ince *et al.* [10] used 1-D CNN for motor fault diagnosis, which not only avoided the challenge of manually configuring the statistical feature set, but also achieved high diagnostic accuracy. Likewise, Ding and He [13] used CNNs to automatically learn features from 2-D wavelet packet energy maps to diagnose the spindle bearings under load fluctuations. These methods outperformed conventional shallow machine learning-based fault diagnosis methods.

Although the inclusion of deep learning in fault diagnosis has been effective for learning reliable sets of features, the following issues should be properly addressed to further improve diagnostic performance. First, classical deep learning algorithms using sigmoidal activation functions often encounter the vanishing/exploding gradient problem found in training NNs with gradient-based learning methods and backpropagation. Backpropagation computes gradients by the chain rule, leading to exponential decrease/increase of the gradients with the increase of layers [16]. If, for example, the vanishing gradient problem occurs, weights between layers in deep learning algorithms are not properly optimized. Second, it is generally more difficult for deeper NNs (or NNs with many layers) to update all their trainable parameters to optimal values, which means simply stacking more layers cannot ensure a better performance [15].

Deeper networks are becoming useful for vibration-based fault diagnosis because they are capable of finding a good set of features from complex and highly variable signals. However, as mentioned above, it is challenging to train deeper networks. DRNs are more effective for easing the difficulty of training networks that are substantially deeper than the general CNNs [17], [18] by using identity shortcuts, which help backpropagate errors through multiple layers. For example, a DRN can successfully train networks that have 1001 layers [19]. Accordingly, DRNs have the potential to outperform classical CNNs in learning a good set of features that will be used for fault diagnosis of planetary gearboxes.

The key contribution of this study is the development of a variant of DRNs for vibration-based fault diagnosis, the so-called deep residual networks with dynamically weighted wavelet coefficients (DRN+DWWC). Since a wavelet packet transform is an effective tool for characterizing the transitory features of nonstationary vibration signals, the developed DRN+DWWC uses a series of sets of wavelet packet coefficients corresponding to various frequency bands (i.e., 64 sets of wavelet packet coefficients in this study) as input for fault diagnosis. Additionally, "dynamic weighting layers" are included in the developed DRN+DWWC to automatically adjust a unique weight applied to each set of wavelet packet coefficients for the sake of finding a discriminative set of features for fault diagnosis. This is mainly because of dealing with an issue caused by the fact that no general consensus has been reached as to which frequency band contains the most intrinsic information about the diverse health status of a planetary gearbox. In fact, the inclusion of domain knowledge in deep learning—i.e., dynamic weighting layers in DRNs—improves the ability of DRNs to learn discriminative features (or feature maps) for vibration-based fault diagnosis. The efficacy of the developed DRN+DWWC was verified for planetary gearbox fault diagnosis in this study.

The remainder of this paper is organized as follows. Section II introduces the experimental setup to collect vibration signals from the planetary gearbox under variable operating conditions. Section III then elucidates the developed DRN+DWWC fault diagnosis method. In Section IV, the usefulness of the developed method is verified by comparing with classical and state-of-the-art machine learning algorithms used for fault diagnosis. Finally, Section V gives conclusions.

II. Description of Planetary Gearbox Health States

The developed DRN+DWWC method was used to pinpoint health states (i.e., healthy and various faulty states) in the planetary gearbox. The experiments in this study used a drivetrain

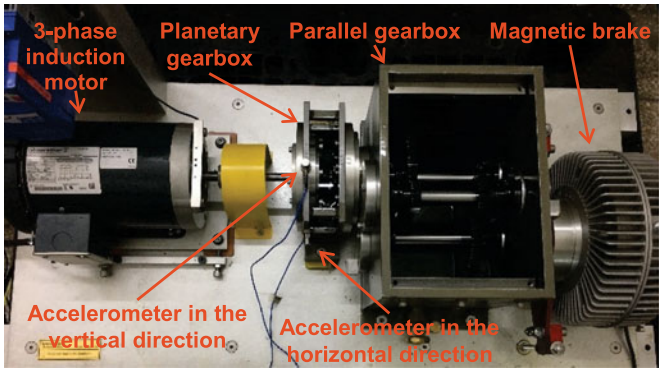

Fig. 1. Drivetrain diagnostics simulator used for experiments.

TABLE I
SUMMARY OF HEALTH CONDITIONS OF THE PLANETARY GEARBOX CONSIDERED IN THIS STUDY

Health condition	Description	Label	Category
Healthy	No seeded fault in the planetary gearbox	H	Class 1
	A seeded fault on a ball in a bearing with healthy gears	SFB + HG	Class 2
	A seeded fault on the inner raceway of a bearing with healthy gears	SFI + HG	Class 3
	A seeded fault on the outer raceway of a bearing with healthy gears	SFO + HG	Class 4
Faulty	Composite seeded faults on a bearing with healthy gears	CSF + HG	Class 5
	A tooth root crack on the planet gear with healthy bearings	TRC + HB	Class 6
	A tooth surface pitting on the planet gear with healthy bearings	TSP + HB	Class 7
	A tooth chipped fault on the planet gear with healthy bearings	TCF + HB	Class 8
	A tooth missing fault on the planet gear with healthy bearings	TMF + HB	Class 9

diagnostics simulator, which mainly consisted of a motor, a two-stage planetary gearbox, a two-stage parallel gearbox, a torque controller, and a magnetic brake, as shown in Fig. 1. Vibration signals sampled at 25.6 kHz were collected via two accelerometers that ensure horizontal and vertical movements were measured in the gearbox.

For the sake of verifying the efficacy of the developed DRN+DWWC fault diagnosis method, nine health states in the planetary gearbox were used for diagnosis, as described in Table I. Additionally, vibration signals collected in the horizontal and vertical directions were used to form dataset 1 and dataset 2, respectively. Likewise, 12 56-s vibration signals were recorded at the variable rotating speed from 20 to 38.7 Hz under three different load conditions, as summarized in Table II. More specifically, each 56-s vibration signal was further divided into 350 0.16-s vibration signals. Hence, the total number of observations for each health state considered in this study is 4200.

TABLE II
SUMMARY OF OBSERVATIONS FOR EACH HEALTH CONDITION

Load condition	Number of 56-s vibration signals for each load condition	Number of observations obtained from a 56-s vibration signal	Total number of observations used for each health condition
0 V	4	350	
4 V	4	350	4200
8 V	4	350	

III. DRN+DWWC

Fig. 2 illustrates an overview of the developed DRN+DWWC fault diagnosis method. As shown in Fig. 2, the DRN+DWWC uses a matrix involving wavelet packet coefficients on various time–frequency bands to learn discriminative features from the matrix. More specifically, the developed DRN+DWWC learns low-level features at shallow layers (i.e., layers close to the input) and mid-/high-level features at deeper layers. Finally, a fully connected output layer is used for classification; the identification of different health conditions in the planetary gearbox is a multiclass classification problem. The primary contribution of this paper is the development of a variant of a DRN for diagnostics.

A. Input of DRN+DWWC

In vibration-based machine fault diagnosis applications, the fact that the properties of the vibration signals vary with time due to variable operating conditions (e.g., rotating speed, loads) usually calls for time–frequency analysis. This is mainly because time–frequency analysis is effective for scrutinizing a 1-D signal in both the time and frequency domains simultaneously. In fact, various time–frequency representations, such as short-time Fourier transform (STFT), EMD, and wavelet transform, have been used for time–frequency analysis. Considering the fixed frequency resolution of STFT and the lack of theoretical foundation of EMD, the developed fault diagnosis method employed wavelet transform [20], which can enable powerful multiresolution time–frequency analysis. More specifically, a discrete wavelet packet transform is preferable in this study because a continuous wavelet transform produces a large amount of redundant information due to overlapping.

Wavelet packet decomposition is a discrete algorithm to analyze nonstationary signals, which decomposes a signal into several frequency bands, and in each frequency band, there are a series of wavelet packet coefficients. Fig. 3 shows the process of building an input matrix for the DRN+DWWC. In this study, the total number of samples in each observation is 4096. Then, each terminal node contains 64 wavelet packet coefficients; $4096/2^{\text{depth}} = 4096/2^6 = 64$, where "depth" refers to the depth of wavelet packet decomposition. Further, the input

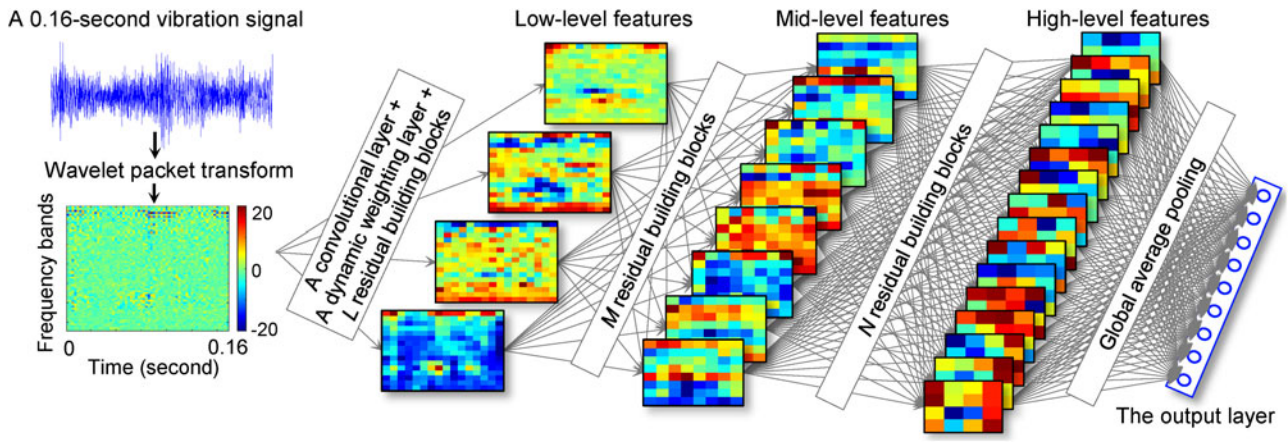

Fig. 2. Developed DRN+DWWC method, where a residual building block mainly involves two convolutional layers, a dynamic weighting layer, and an identity shortcut; *L*, *M*, and *N* are the numbers of residual building blocks used to learn low-, mid-, and high-level features, respectively.

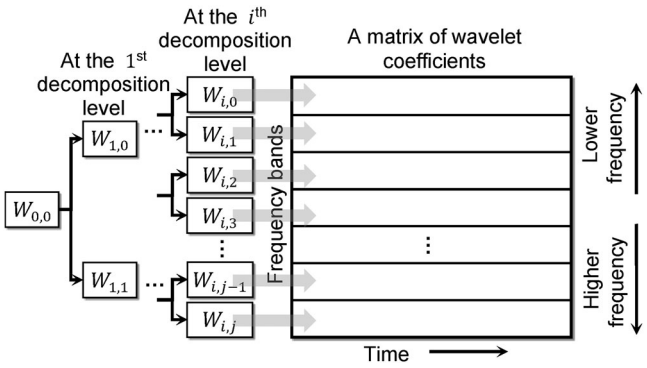

Fig. 3. Matrix involving wavelet coefficients on various time–frequency bands, where $W_{i,j}$ is a set of wavelet coefficients at the *j*th terminal node at the *i*th decomposition level. In this study, $i = 6$ and $j = 2^i - 1$.

matrix is formed by stacking 64 sets of wavelet packet coefficients on various frequency bands, and further used as the input for the DRN+DWWC.

There are several advantages of using a matrix of wavelet packet coefficients as an input of the developed DRN+DWWC for fault diagnosis. First, because there is no general consensus on a frequency band which contains the most intrinsic information about a planetary gearbox's diverse health status, it is worthwhile to learn a good set of features from a series of sets of wavelet packet coefficients corresponding to various frequency bands. Second, using a matrix of wavelet packet coefficients as an input for the developed DRN+DWWC enables learning nonlinear relationships among wavelet packet coefficients on neighboring frequency bands.

B. Feature Learning in DRN+DWWC

CNNs [17] are deep learning methods that use convolutional layers and mostly have high computational complexities. With the development of computational hardware, CNNs have attracted attention due to their excellent performance. Compared with the traditional fully connected deep NNs, CNNs greatly reduce the number of trainable parameters (weights and biases) through the strategies of local receptive field and weight sharing [21]. Local receptive field means that each neuron is only connected to some neurons of the previous layer and the next layer, and weight sharing means the weights of each neuron are also shared with other neurons in the same layer. The two strategies can reduce the number of weights and make it easier to train deeper networks. A DRN is a type of a CNN with identity shortcuts in its architecture. For deeper architectures, the aforementioned trainable parameters are generally not easy to optimize. The identity shortcuts of DRNs can help the backpropagation of gradients, so that the weights and biases can be updated efficiently.

Fig. 4 shows the architecture of the developed DRN+DWWC, which uses the matrix of wavelet packet coefficients as input and consists of standard convolutional layers, dynamic weighting layers, identity shortcuts, a global average pooling layer, a fully connected layer and so forth. More details about the components of the developed DRN+DWWC are given in subsequent sections.

1) Standard Convolutional Layers: The standard convolutional layer uses the strategies of local receptive field and weight sharing. These two strategies are mathematically implemented by means of a convolution operation. In this layer, the input feature map is convolved with one or more convolutional kernels (also called filters). The local receptive field is the same size as the kernel used in the convolutional layer. The convolutional kernel slides on the input map, so that the parameters (weights) of the kernel are shared, mathematically expressed as follows:

$$\boldsymbol{x}_j^l = f\left(\sum_{i \in M_j} \boldsymbol{x}_i^{l-1} * \boldsymbol{k}_{ij}^l + b_j^l\right) \tag{1}$$

where $\boldsymbol{x}_i^{l-1}$ is the *i*th channel of the feature map at the $(l-1)$ th layer, $\boldsymbol{x}_j^l$ is the *j*th channel of the feature map at the *l*th layer, M_j is the selection of channels used for calculating the *l*th output channel, $\boldsymbol{k}$ is the convolutional kernel, b is the bias, and $f(\cdot)$ is the activation function [14].

It is notable that the weights in the convolutional kernels and the biases are optimized by minimizing backpropagation

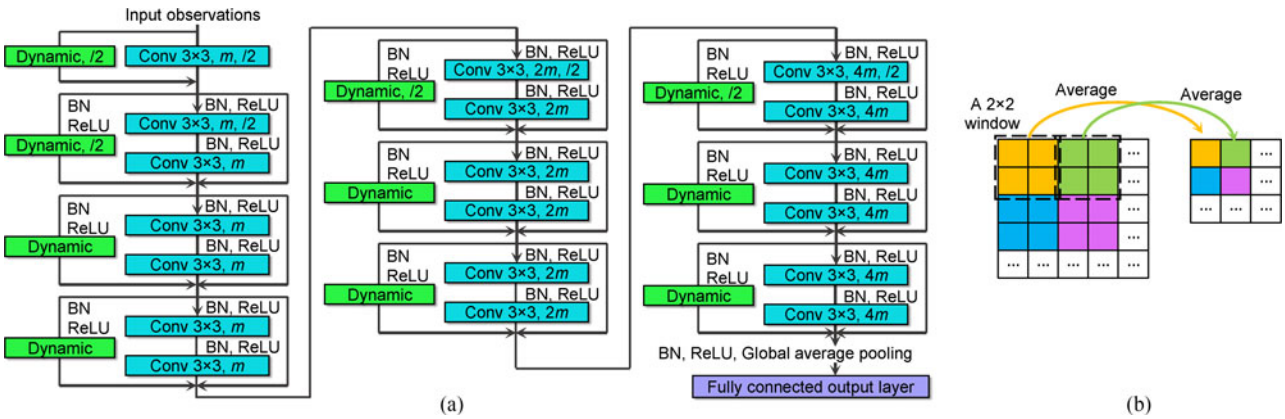

Fig. 4. (a) Architecture of the DRN+DWWC, where "Conv 3 × 3" refers to the convolutional layer with m, $2m$, or $4m$ 3 × 3 kernels, "/2" means reducing the feature map size with a stride of 2, "dynamic" refers to a dynamic weighting layer, "BN" stands for batch normalization, and "ReLU" is a rectified linear unit activation function. (b) 2 × 2 moving window that has a stride of 2, where an averaging operation is carried out for downsampling within the window.

errors during the training process. Nowadays, the 3 × 3 kernel is widely used because it is computationally efficient [22] and large enough for capturing basic local features, including local extremes. Meanwhile, the activation function can be used for achieving nonlinear transformations. In this paper, the rectified linear unit (ReLU) activation function [23] is used, rather than the traditional sigmoid and hyperbolic tangent (tanh) activation functions. The ReLU activation function is expressed by

$$f(x) = \max(x, 0). \tag{2}$$

The ReLU activation function is more effective for avoiding the vanishing gradient problems than the classical activation functions. The major reason is that the absolute values of the derivatives of standard sigmoid and tanh functions are mostly smaller than 1, and the gradients may become very close to 0 when backpropagating the error through multiple layers.

2) Dynamic Weighting Layers: The role of dynamic weighting layers is to apply dynamic weights to the input. More specifically, a single unique weight is applied to the wavelet coefficients on a particular frequency band, which can be expressed by

$$\boldsymbol{y}_i^l = \boldsymbol{x}_i^{l-1} w_i^l \tag{3}$$

where $\boldsymbol{x}_i^{l-1}$ refers to the ith row of the input feature map at the lth dynamic weighting layer, e.g., the ith row of the input feature map at the first dynamic weighting layer contains 64 wavelet coefficients, w_i^l refers to a weight that can be multiplied with each of the wavelet coefficients at the ith row, and $\boldsymbol{y}_i^l$ refers to the ith row of weighted feature map. The process of applying the row-wise weights to the feature map is depicted in Fig. 5, resulting in the weighted feature map.

Note that the weights in the dynamic weighting layer are also optimized during the training process of DRN+DWWC using a gradient descent algorithm. The gradients for the weights at a dynamic weighting layer can be calculated as follows:

$$\frac{\partial E}{\partial w_i} = \sum_{j,k} \frac{\partial E}{\partial y_{ij}^k} \frac{\partial y_{ij}^k}{\partial w_i} = \sum_{j,k} \frac{\partial E}{\partial y_{ij}^k} x_{ij}^k \tag{4}$$

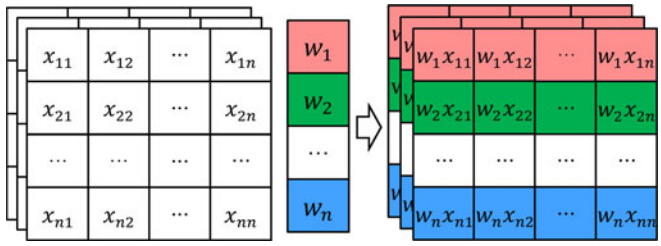

Fig. 5. Weighted feature maps.

where E is the error (i.e., the softmax cross-entropy loss which is introduced subsequently) and $y_{ij}^k = w_i x_{ij}^k$ is an element in the output feature maps of a dynamic weighting layer. Feature maps can be viewed as a 3-D matrix, as depicted in Fig. 5, where i, j, and k are the indexes of row, column, and channel of the feature maps, respectively. Then, the weights can be updated as follows:

$$w_i \leftarrow w_i - \eta \frac{\partial E}{\partial w_i} \tag{5}$$

where η is the learning rate. In addition, to match the feature map size with the convolutional layers, average pooling with a stride of 2 is carried out after the first and three other dynamic weighting layer layers, as illustrated in Fig. 4(b). Likewise, zero padding is used to match the number of channels.

3) Residual Building Block: In a DRN, a residual building block often consists of several convolutional layers, batch normalizations (BNs), ReLU activation functions, and one identity shortcut, as shown in Fig. 6(a). Specifically, BN [24] is a type of normalizing method that can be applied to each batch between the layers in deep NNs. It aims to solve the internal covariate shift problem, i.e., the distribution of feature maps in the layers continuously changes in the training process, which can decrease the training speed. Likewise, a series of building blocks are stacked after the first convolutional layer. The identity shortcuts in these stacked building blocks are useful for optimizing trainable parameters in error backpropagation.

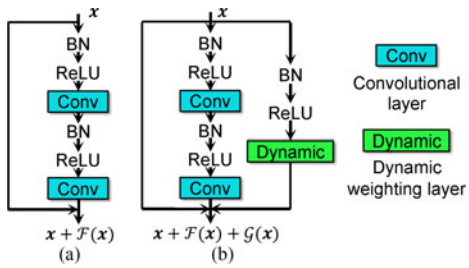

Fig. 6. (a) Residual building block in the DRN and (b) developed residual building block with a dynamic weighting layer.

In the developed DRN+DWWC, a single dynamic weighting layer is employed in a residual building block [see Fig. 6(b)]. Although the inputs passing to building blocks are feature maps obtained by a series of convolutions and nonlinear transformations, the use of dynamic weighting layers can further emphasize different contributions of wavelet packet coefficients on different frequency bands in the process of learning discriminative features that will be used for fault diagnosis of the planetary gearbox. He *et al.* [19] claimed that full preactivation—i.e., BN and ReLU are conducted before the input is propagated to a convolution layer [see Fig. 6(a)]—is helpful for updating trainable model parameters and reducing overfitting. Thus, the underlying concept of full preactivation was employed in the developed DRN+DWWC.

As shown in Fig. 6(a), let $\boldsymbol{x}$ donate the input, and $\boldsymbol{y}$ donate the output, each residual building block in the DRN can be expressed in a general form as follows:

$$\boldsymbol{y} = \boldsymbol{x} + \mathcal{F}\left(\boldsymbol{x}, \boldsymbol{W}_C\right) \tag{6}$$

where $\mathcal{F}(\cdot)$ donates a nonlinear function for the path which contains two BNs, two ReLU activation functions and two convolutional layers, and $\boldsymbol{W}_C$ donates the parameters to be optimized for this path. Likewise, the output of the residual building block in DRN+DWWC can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{x} + \mathcal{F}\left(\boldsymbol{x}, \boldsymbol{W}_C\right) + \mathcal{G}\left(\boldsymbol{x}, \boldsymbol{W}_D\right) \tag{7}$$

where $\mathcal{G}(\cdot)$ donates a nonlinear function for the path which contains a BN, an ReLU activation function, and a dynamic weighting layer, and $\boldsymbol{W}_D$ donates the parameters to be optimized for this path. In DRN+DWWC, the developed residual building blocks are stacked as depicted in Fig. 4(a).

4) Softmax Cross-Entropy Loss Function: The softmax cross-entropy loss function [21] was used in the fully connected output layer of the DRN and adopted in the developed DRN+DWWC method as the objective function to be minimized, rather than the squared error loss function. In classification problems, the probabilities of an observation belonging to all the classes should be in the range of [0, 1] and sum up to 1. Considering this range, the softmax activation function is used, which is expressed by

$$q_j\left(\boldsymbol{x}\right) = \frac{e^{x_j}}{\sum_{i=1}^{N_{\text{class}}} e^{x_i}}, \quad \text{for } j = 1, \ldots, N_{\text{class}} \tag{8}$$

where x_j is the jth input feature of the softmax activation function, $q_j(\boldsymbol{x})$ is the output which can be interpreted as the estimated probability of an observation $\boldsymbol{x}$ belonging to the jth class, and N_{class} is the total number of classes [25]. Then, the cross-entropy loss function is used to measure the error between the outputs and target values (i.e., labels), which is expressed by

$$E\left(p\left(\boldsymbol{x}\right), q\left(\boldsymbol{x}\right)\right) = -\sum_{j=1}^{N_{\text{class}}} p_j\left(\boldsymbol{x}\right) \log\left(q_j\left(\boldsymbol{x}\right)\right) \tag{9}$$

where $p(\boldsymbol{x})$ is the label of the observation $\boldsymbol{x}$, and $p_j(\boldsymbol{x})$ can be interpreted as the real probability of $\boldsymbol{x}$ belonging to the jth class. In general, the cross-entropy loss function has a higher training speed than the squared error function, because, for example, it has larger gradients when the output is close to 0 and the target value is 1 (i.e., the real probability of the observation belonging to the class is 100%), so that the trainable parameters can be updated more efficiently [25].

IV. Experimental Results

The developed DRN+DWWC method was implemented using TensorFlow, which is Google's open source software library for machine learning, and applied for fault diagnosis of the planetary gearbox under variable operating conditions. As mentioned in Section II, the developed method was verified on two different datasets. The developed method was also tested on datasets with a higher level of noise; although each of the 0.16-s vibration signals already contained a certain level of noise, Gaussian noise was artificially embedded into them, yielding a signal-to-noise ratio of 5 dB, for the sake of increasing the level of difficulty in fault diagnosis under the assumption that the vibration signals can contain a higher level of Gaussian noise in real-world fault diagnosis applications. That is, the robustness to noise of the developed DRN+DWWC can also be verified in this study. Likewise, although a comparison between the developed DRN+DWWC method and the classical and state-of-the-art machine learning-based fault diagnosis methods with or without feature learning ability involves unavoidable errors due to the use of different hyperparameters, the goal of the comparison is to show the potential for improved diagnostic performance of the developed method for fault diagnosis of the planetary gearbox rather than to conduct a precise performance comparison.

A. Hyperparameter Setup for DRN+DWWC

There have been many empirical suggestions for the hyperparameters, and this paper sets them according to [15], [16], and [19]. To be specific, the initial learning rate is set to 0.1, divided by 10 at 40 and 80 epochs, and terminated at 100 epochs. The purpose of this schedule is to update the trainable parameters quickly at the beginning, and fine-tune the trainable parameters at the end of the training process.

The mini-batch [21] refers to the group of observations that feed into the networks at the same time, and its size is set to 128. On one hand, if a large training dataset is fed into the deep networks at the same time, it requires a great amount of memory, which is mostly impractical for personal computers. On the other hand, dealing with a batch of observations in each

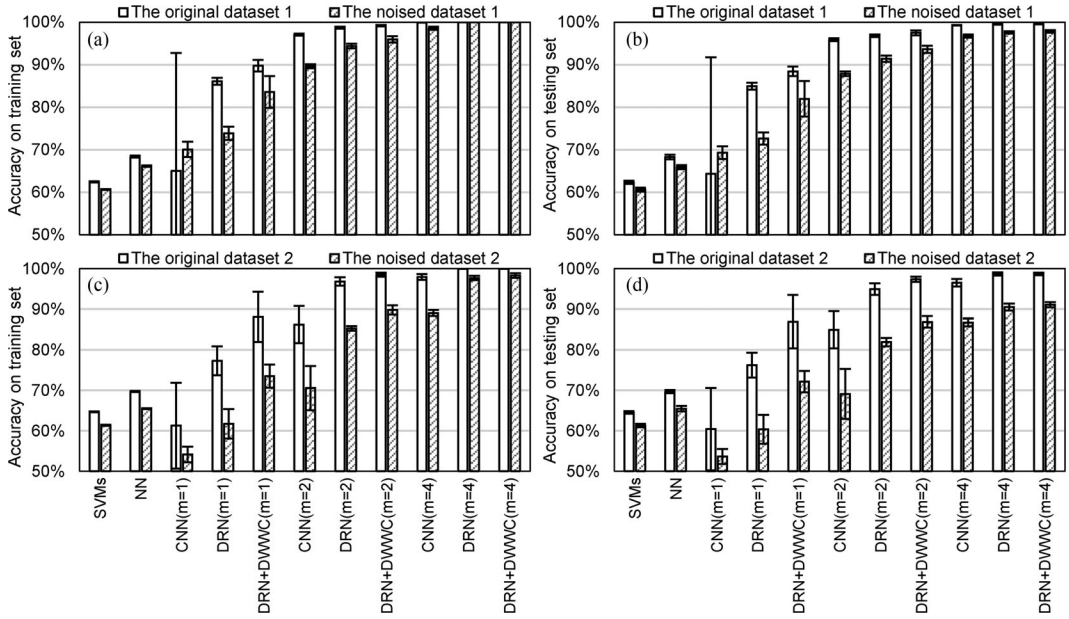

Fig. 7. Diagnostic performance in terms of accuracy. Note that *m* indicates the number of convolutional kernels in Fig. 4.

time can accelerate training by taking advantage of parallel computation [21].

Momentum is a weight-updating strategy that allows the updates of previous iterations to contribute to the update of the current iteration [21]. Its purpose is to accelerate the training process and avoid the local optima. For example, if the weights need to be continuously updated toward the same direction, it will be faster if there is a contribution from the update of the previous iteration; if there are some local optima in this direction before reaching the global optima, there will be a larger chance for the NNs to jump over the local optima with a larger step (the gradient in this iteration plus a contribution of the update from the previous iteration). In general, the contribution of the previous iteration is recommended to be 0.9.

The weights (including the dynamic weights on wavelet coefficients) are initialized according to [16], and the biases are initialized to be zeros. The problem in traditional weight initialization (using zero-mean Gaussian random values with a constant standard deviation) is that the high-level feature maps may have extremely large or small absolute values at the end of deep NNs when using the ReLU activation function [16]. The reason is that the traditional randomly initialized weights have a greater chance to magnify or shrink the input data layer by layer. He *et al.* [16] initialized the weights by trying to keep the variance of input data unchanged when the data go through the networks, so that the values of high-level feature maps can be kept in a reasonable range.

L2 regularization is a strategy that aims to increase the generalization ability (e.g., to ensure relatively high accuracy on the test dataset), because the method that has high accuracy on the training dataset does not essentially have high accuracy on the test dataset. This situation can be caused by the overfitting problem [21]. To be specific, if the weights in the deep NNs have extremely large values, it would be easy for the test data to have large errors after multiplying with these weights, even if the test data have similar values as the training data. Therefore, a penalty, which is known as weight decay, is applied on the weights, so that the weights are preferred to have small absolute values [21]. In the developed DRN+DWWC, the coefficient of weight decay (which defines how strong the penalty is) is set to 0.0001, to remain consistent with the DRN [15].

As shown in Fig. 4(a), m is the number of convolutional kernels in the first convolutional layer. In deep learning methods, a convolutional kernel is actually a trainable feature extractor. When there are more convolutional kernels in the first layer, there are more kinds of basic local features being extracted from the input observation. The more basic local features can be nonlinearly integrated to be much more complex high-level features. After the supervised training process, the high-level features can become discriminative between different classes. In this study, the experiments were conducted with m equal to 1, 2, and 4.

B. Performance Comparisons

The developed DRN+DWWC method was not only compared with traditional fault diagnosis methods employing shallow machine learning algorithms (multiclass SVMs and a NN) with statistical parameters to verify the usefulness of feature learning, but also with the state-of-the-art deep learning-based methods (CNN and DRN) to show the enhanced feature learning ability. The experimental results are shown in Fig. 7, and further discussion is as follows.

1) Usefulness of Feature Learning in Fault Diagnosis: To verify the usefulness of feature learning, classical supervised learning-based diagnosis methods (i.e., multiclass SVMs and a NN) using the statistical parameters in [26] were compared with the diagnosis methods with feature learning ability. In Fig. 7, it is obvious that the performance of the classical diagnosis methods is mostly inferior to the methods involving feature learning in noiseless and noisy environments.

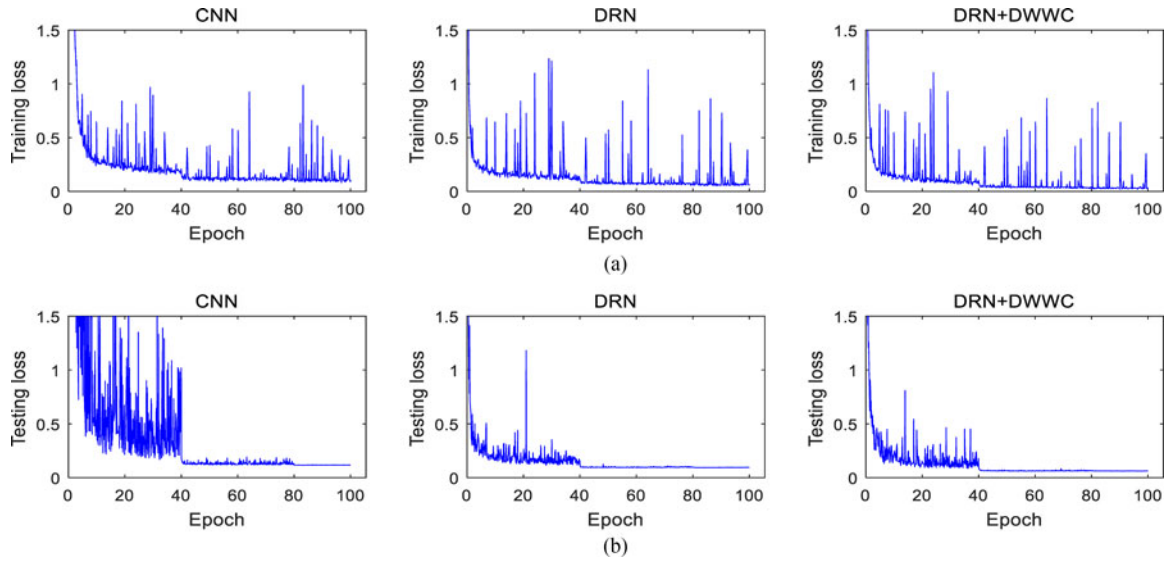

Fig. 8. Example of (a) training losses and (b) testing losses obtained from the CNN, DRN, and DRN+DWWC with $m = 4$ on the original dataset 1.

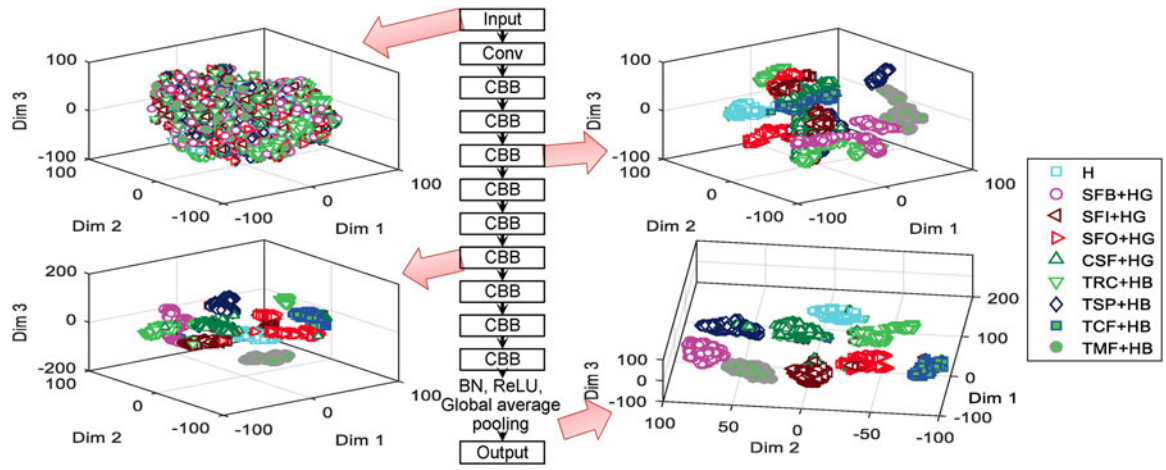

Fig. 9. Three-dimensional representations of high-dimensional feature maps at different layers in the CNN with $m = 4$, where a testing dataset from dataset 1 under the cross-validation scheme was used for the sake of visualizing feature maps in a lower dimensional space. Likewise, the CNN yielded 98.72% accuracy on the testing dataset. "Conv" in the brief architecture refers to a convolutional layer, and "CBB" refers to the convolutional building block, which does not have an identity shortcut when compared with the residual building block.

More specifically, the developed DRN+DWWC method with $m = 4$ (see Fig. 4) at least improved 30.23% and 29.02% of training accuracy and testing accuracy in a noiseless environment, respectively, compared to the conventional fault diagnosis methods using statistical parameters. Additionally, the developed DRN+DWWC method with $m = 4$ was more robust to noise than the conventional fault diagnosis methods by yielding at least 32.82% and 25.62% performance improvements in terms of training accuracy and testing accuracy, respectively, in a noisy environment.

2) Comparison Between the DRN+DWWC Method and the State-of-the-Art Methods With Feature Learning Ability: As shown in Fig. 7, the developed DRN+DWWC method outperformed the other feature learning-based diagnosis methods in terms of training accuracy and testing accuracy. Fig. 8 presents an example of training and testing losses obtained from the CNN, DRN, and DRN+DWWC on the original dataset 1. As shown in Fig. 8, the CNN, DRN, and DRN+DWWC do not encounter overfitting, because their testing losses converged to a certain level after a series of epochs. Consequently, their accuracies are reliable for performance comparisons. One interesting observation in Fig. 8 is that the training losses have more fluctuations than the testing losses. This is mainly because the training losses are calculated from the mini-batches. That is, a small number of observations are randomly selected for a mini-batch and the associated training loss can be significantly influenced by a few misclassifications in the batch. Unlike the training losses, the testing losses are relatively less fluctuated due to the relatively larger number of observations used for testing. Our analysis indicated that the use of dynamic weighting layers to dynamically adjust the significance of wavelet coefficients on different frequency bands improved the ability of learning discriminative features for identifying nine health states in the planetary gearbox.

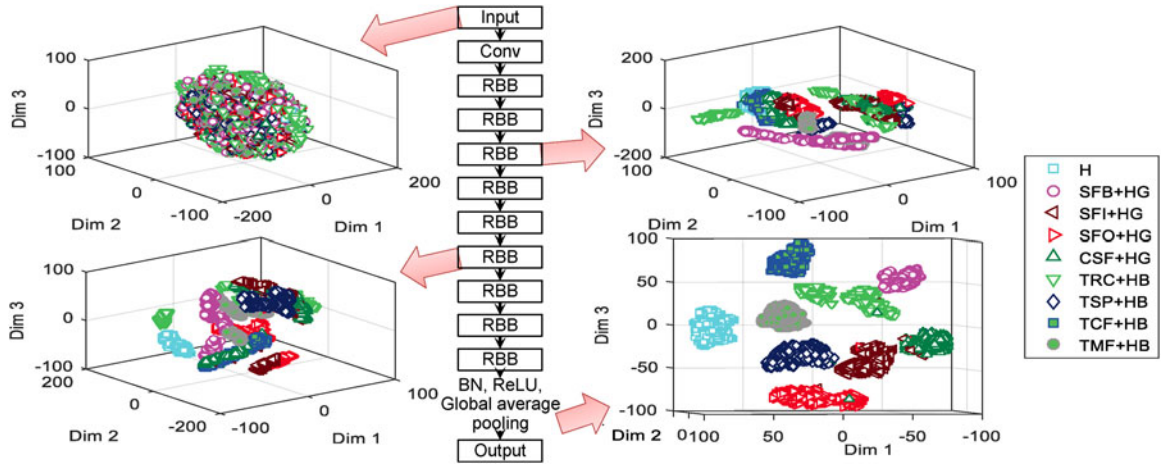

Fig. 10. Three-dimensional representations of high-dimensional feature maps at different layers in the DRN with $m = 4$, where a testing dataset from dataset 1 under the cross-validation scheme was used for the sake of visualizing feature maps in a lower dimensional space. Likewise, the DRN yielded 99.52% accuracy on the testing dataset. "Conv" in the brief architecture refers to a convolutional layer, and "RBB" refers to a residual building block in DRN.

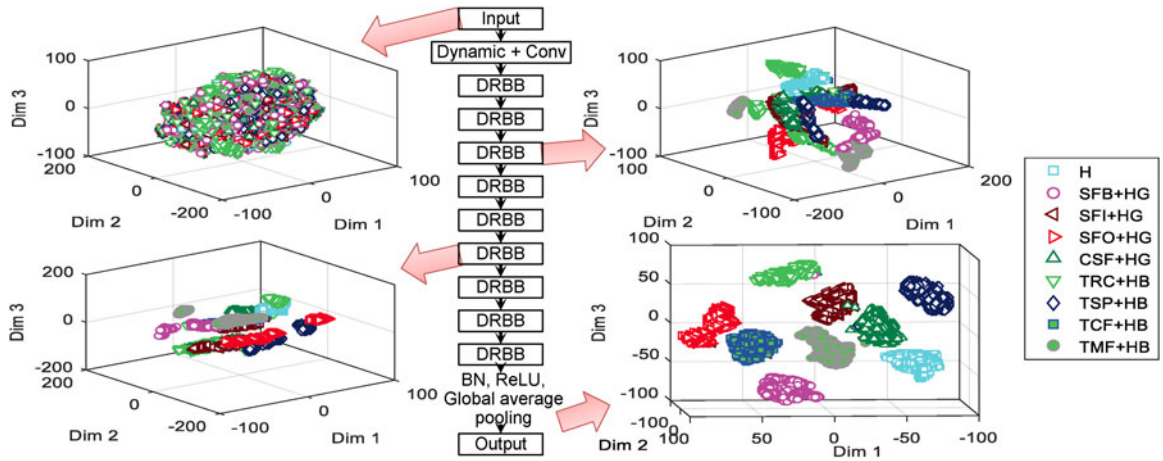

Fig. 11. Three-dimensional representations of high-dimensional feature maps at different layers in the developed DRN+DWWC with $m = 4$, where a testing dataset from dataset 1 under the cross-validation scheme was used for the sake of visualizing feature maps in a lower dimensional space. Likewise, the developed DRN+DWWC yielded 99.60% accuracy on the testing dataset. "Conv" in the brief architecture refers to a convolutional layer, "Dynamic" refers to a dynamic weighting layer, and "DRBB" refers to the developed residual building block.

In addition, a nonlinear dimensionality reduction method, the so-called *t*-distributed stochastic neighbor embedding [27], was employed to provide 3-D representations of high-dimensional feature maps at different layers in the CNN, DRN, and DRN+DWWC, respectively, as shown in Figs. 9–11. Although a comparison between the feature maps in a lower dimensional space involves unavoidable errors due to the loss of information in dimensionality reduction, the goal of the comparison is to explore the effectiveness of each deep learning method in learning a discriminative set of features for fault diagnosis. In Figs. 9–11, it is obvious that diverse health states (or classes) are heavily overlapped at the input layer, whereas they become more separable at deeper layers. Specifically, the developed DRN+DWWC was the most effective for learning a discriminative set of features, yielding 99.60% accuracy (on a testing dataset in Fig. 11). That is, a small number of misclassifications were observable among the health states, as depicted in Fig. 11.

V. Conclusion

Finding a good set of features has been a long-standing issue in the fault diagnosis of planetary gearboxes subject to variable operating conditions. To address this issue, a DRN+DWWC method was developed to learn a set of features that could discriminate diverse health states in the planetary gearbox—one health state and eight faulty states. More specifically, dynamic weighting layers in the developed deep learning architecture were used to optimize weights applied to wavelet coefficients on various frequency bands for the sake of exploring how a series of wavelet coefficients on a particular frequency band contributed to discriminating the gearbox's health states.

The usefulness of feature learning was verified by a comparison between the developed DRN+DWWC method and the classical machine learning-based diagnosis methods (i.e., one-against-one multiclass SVMs and a three-layer NN) employing statistical parameters. The developed method is able to

automatically learn discriminative features from the training data. Experimental results indicated that this method outperformed the SVM- and NN-based fault diagnosis methods by yielding 30.77% and 28.82%, 25.63%, and 23.66% performance improvements in terms of the averages of training accuracies and testing accuracies (under noiseless and noisy environments), respectively. That is, the inclusion of feature learning is significant for fault diagnosis of planetary gearboxes.

Likewise, the developed DRN+DWWC method was superior to the state-of-the-art deep learning-based diagnosis methods with feature learning ability. More specifically, the developed method was more effective for learning discriminative features that could reduce a classifier's burden to discriminate multiple gearbox health states than the state-of-the-art deep learning algorithms. This was mainly due to the inclusion of a series of "dynamic weighting layers" to adjust the importance of wavelet coefficients on different frequency bands during the training process. As a consequence, the developed method resulted in 11.43% and 10.60%, 3.74% and 3.87% performance improvements compared with CNN- and DRN-based methods in terms of the averages of training accuracies and testing accuracies (using different numbers of convolutional kernels in noiseless and noisy environments), respectively.

In this study, the efficacy of the developed DRN+DWWC that learns a good set of features was verified by fault diagnosis of planetary gearboxes. However, this method would be applicable to general data-driven fault diagnosis with minor changes (e.g., other machinery and electronic components or systems).

REFERENCES

[1] E. Gouda, S. Mezani, L. Baghli, and A. Rezzoug, "Comparative study between mechanical and magnetic planetary gears," *IEEE Trans. Magn.*, vol. 47, no. 2, pp. 439–450, Feb. 2011.

[2] Z. Du, X. Chen, H. Zhang, and R. Yan, "Sparse feature identification based on union of redundant dictionary for wind turbine gearbox fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6594–6605, Oct. 2015.

[3] F. Chaari, T. Fakhfakh, and M. Haddar, "Dynamic analysis of a planetary gear failure caused by tooth pitting and cracking," *J. Failure Anal. Prevent.*, vol. 6, no. 2, pp. 73–78, 2006.

[4] Y. Lei, J. Lin, M. Zuo, and Z. He, "Condition monitoring and fault diagnosis of planetary gearboxes: A review," *Measurement*, vol. 48, pp. 292–305, 2014.

[5] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, nos. 1/2, pp. 314–334, 2014.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vol. 72/73, pp. 303–315, 2016.

[8] H. Chen, J. Wang, B. Tang, K. Xiao, and J. Li, "An integrated approach to planetary gearbox fault diagnosis using deep belief networks," *Meas. Sci. Technol.*, vol. 28, no. 2, 2016, Art. no. 025010.

[9] Y. Lei, F. Jia, J. Lin, S. Xing, and S. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[10] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016.

[11] 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, pp. 414(1)–414(15), 2017.

[12] O. Janssens *et al.*, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, 2016.

[13] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 8, pp. 1926–1935, Aug. 2017.

[14] Z. Chen, C. Li, and R. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock Vib.*, vol. 2015, pp. 1–10, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 27–30, 2016, pp. 770–778.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 13–16, 2015, pp. 1026–1034.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NA, USA, Dec. 3–6, 2012, pp. 1097–1105.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, Netherlands, Oct. 8–16, 2016, pp. 630–645.

[20] D. F. Walnut, *An Introduction to Wavelet Analysis*. Boston, MA, USA: Birkhäuser, 2004.

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT, 2016.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NA, USA, Jun. 26–Jul. 1, 2016, pp. 2818–2826.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, Apr. 11–13, 2011, pp. 315–323.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, Jul. 7–9, 2015, pp. 448–456.

[25] P. Zhou and J. Austin, "Learning criteria for training neural network classifiers," *Neural Comput. Appl.*, vol. 7, no. 4, pp. 334–342, 1998.

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

[27] L. J. P. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Minghang Zhao was born in Shandong, China, in June 1991. He received the B.E. degree in mechanical engineering from Chongqing University, Chongqing, China, in 2013, where he is currently working toward the Ph.D. degree under the supervision of Prof. Baoping Tang in the State Key Laboratory of Mechanical Transmission.

He was previously a Visiting Research Scholar with the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD, USA, from 2016 to 2017. His research interests include data-driven fault diagnosis, prognostics and health management of mechanical and electrical systems.

Myeongsu Kang (M'17) received the B.E. and M.S. degrees in computer engineering and information technology and Ph.D. degree in electrical, electronics, and computer engineering from the University of Ulsan, Ulsan, South Korea, in 2008, 2010, and 2015, respectively.

He is currently a Research Scientist with the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD, USA. His current research interests include data-driven anomaly detection, diagnostics, and prognostics of complex systems, such as automotive, railway transportation, and avionics, for which failure would be catastrophic. He has expertise in analytics, machine learning, system modeling, and statistics for prognostics and systems health management.

Baoping Tang received the M.Sc. and Ph.D. degrees in mechanical engineering from Chongqing University, Chongqing, China, in 1996 and 2003, respectively.

He is currently a Professor and Ph.D. Supervisor with the College of Mechanical Engineering, Chongqing University. More than 150 papers has been published in his research career. His main research interests include wireless sensor networks, mechanical and electrical equipment security service and life prediction, and measurement technology and instruments.

Dr. Tang was the recipient of the National Scientific and Technological Progress 2nd Prize of China in 2004 and the National Invention 2nd Prize of China in 2015.

Michael Pecht (S'78–M'83–SM'90–F'92) received the B.S. degree in acoustics, M.S. degrees in electrical engineering and engineering mechanics, and Ph.D. degree in engineering mechanics from the University of Wisconsin at Madison, Madison, WI, USA, in 1976, 1978, 1979, and 1982, respectively.

He is the Founder of the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD, USA, where he is also a Chair Professor. He has been leading a research team in the area of prognostics.

Dr. Pecht is a Professional Engineer and a Fellow of the American Society of Mechanical Engineers. He was the recipient of the IEEE Undergraduate Teaching Award and the International Microelectronics Assembly and Packaging Society William D. Ashman Memorial Achievement Award for his contributions in electronics reliability analysis. He served as the Chief Editor of the IEEE TRANSACTIONS ON RELIABILITY for eight years and an Associate Editor for the IEEE TRANSACTIONS ON COMPONENTS AND PACKAGING TECHNOLOGY.