Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox

Guoqian Jiang, *Member, IEEE*, Haibo He, *Fellow, IEEE*, Jun Yan, *Member, IEEE*, and Ping Xie

***Abstract*—This paper proposes a novel intelligent fault diagnosis method to automatically identify different health conditions of wind turbine (WT) gearbox. Unlike traditional approaches, where feature extraction and classification are separately designed and performed, this paper aims to automatically learn effective fault features directly from raw vibration signals while classify the type of faults in a single framework, thus providing an end-to-end learning-based fault diagnosis system for WT gearbox without additional signal processing and diagnostic expertise. Considering the multiscale characteristics inherent in vibration signals of a gearbox, a new multiscale convolutional neural network (MSCNN) architecture is proposed to perform multiscale feature extraction and classification simultaneously. The proposed MSCNN incorporates multiscale learning into the traditional CNN architecture, which has two merits: 1) high-level fault features can be effectively learned by the hierarchical learning structure with multiple pairs of convolutional and pooling layers; and 2) multiscale learning scheme can capture complementary and rich diagnosis information at different scales. This greatly improves the feature learning ability and enables better diagnosis performance. The proposed MSCNN approach is evaluated through experiments on a WT gearbox test rig. Experimental results and comprehensive comparison analysis with respect to the traditional CNN and traditional multiscale feature extractors have demonstrated the superiority of the proposed method.**

***Index Terms*—Convolutional neural network (CNN), classification, deep learning, intelligent fault diagnosis, multiscale feature extraction, wind turbine (WT) gearbox.**

I. Introduction

IN RECENT years, wind energy has experienced a remarkable expansion in response to the increase of worldwide energy demand [1], and accordingly, wind turbines (WTs) have been extensively installed both onshore and offshore. Usually, these turbines are located in remote areas and are exposed to harsh environments, suffering from constantly varying loads and experiencing extreme operational temperature and humidity changes. As a result, they are prone to failures and result in high operation and maintenance (O&M) costs. Gearbox, as a critical component in WTs, often suffers from various failures, such as bearing damage, tooth breakage, and gear crack, and has shown high failure rates and resulted in high maintenance costs [2] due to its long downtimes and complex repair procedures. Statistically, maintenance cost caused by the gearbox is about 13% of the overall WT cost [3]. Therefore, condition monitoring and fault diagnosis of WT gearbox have gained considerable attentions from both academia and industry [4]–[8].

Generally, a fault diagnosis system is defined from a pattern recognition perspective. It consists of three general steps: data acquisition (DAQ), feature extraction, and fault classification, among which the latter two are of significant importance and greatly affect the final diagnosis accuracy. The main objective of feature extraction is to extract representative features, which can characterize the health conditions of the underlying machine and also help the downstream fault recognition task. Up to now, numerous feature extraction methods have been reported and discussed. For example, Zhang *et al.* [5] utilized the correlation coefficient analysis in the time domain and the windowed Fourier transform in the frequency domain to extract fault features and combined mining methods with several data to identify different failure stages of a WT gearbox. Du *et al.* [6] proposed a novel sparse feature identification framework based on the union of redundant dictionary to diagnose multiple faults occurring in a WT gearbox. Yang *et al.* [7] proposed an improved spline-kernelled chirplet transform to extract fault-related frequency features for detecting faults in a WT drive train under the time-varying operational conditions. Wang *et al.* [8] integrated ensemble empirical mode decomposition with independent component analysis to separate fault components for diagnosis of bearing failures in a WT. To sum up, various signal processing methods, including classical spectral analysis, wavelet transform [9], empiri-

Manuscript received May 17, 2017; revised August 27, 2017, November 10, 2017, February 2, 2018, and May 10, 2018; accepted May 20, 2018. Date of publication June 13, 2018; date of current version November 30, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61673336 and the Natural Science Foundation of Hebei Province under Grant F2018203413 and Grant F2016203421. This work was conducted when G. Jiang was a joint Ph.D. student at the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA. *(Corresponding author: Haibo He.)*

G. Jiang is with the School of Electrical Engineering, Yanshan University, Qihuangdao 066004, China (e-mail: jgqysu@gmail.com).

P. Xie is with the School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China (e-mail: pingx@ysu.edu.cn).

H. He is with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881 USA (e-mail: he@ele.uri.edu).

J. Yan is with the Concordia Institute for Information Systems Engineering, Concordia University, Montréal, QC H3G 1M8, Canada (e-mail: jun.yan@concordia.ca).

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

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes a case study for induction motor bearing fault diagnosis that further demonstrates the effectiveness and adaptability of the proposed MSCNN approach. This material is 0.07 MB in size.

Digital Object Identifier 10.1109/TIE.2018.2844805

cal mode decomposition (EMD) [10], manifold learning [11], sparse representation [6], and other variants, have been applied to extract effective fault-related features from vibration signals from different perspectives. Then, these extracted features are fed into some shallow machine learning algorithms including logistic regression and support vector machine (SVM). It has been shown that the extracted features or representations define the upper-bound performances of machine learning algorithms [12]. These traditional shallow learning models can only provide limited diagnosis ability. In summary, traditional intelligent fault diagnosis methods still have some obvious drawbacks as follows.

1) Feature extraction and classification are separately designed and performed, both of which affect the final diagnosis performance. This is a divide and conquer strategy that cannot be optimized simultaneously.
2) All extracted features are naturally hand-crafted, requiring much prior knowledge about signal processing techniques and diagnostic expertise, which is time-consuming and labor-intensive.
3) Most of the existing methods are domain-specific and cannot be updated online and generalize well to new diagnosis domains, thus a general-purpose approach is highly desirable.

Alternatively, to address the above drawbacks, deep learning may provide a promising and effective solution for intelligent fault diagnosis. Deep learning attempts to model high-level representations of data and predict/classify patterns via multiple stacked layers of nonlinear information processing units in a hierarchical architecture [13]. In recent years, deep learning has been emerged as a powerful tool and successfully applied in various domains such as computer vision, speech recognition, natural language processing, and health informatics. Motivated by these successful achievements, deep learning has been recently introduced to the field of fault diagnosis [14]. Different deep models such as stacked denoising autoencoders [15], deep belief network [16], sparse filtering [17], convolutional neural networks (CNNs) [18], and recurrent neural network [19] have demonstrated state-of-the-art diagnosis performance compared to traditional approaches relying on hand-crafted features.

In this paper, our goal is to develop an end-to-end fault diagnosis system based on CNNs, which is motivated by its excellent feature learning ability. The desirable system can automatically learn and discover discriminative features from raw temporal vibration signals and then classify different health conditions of the WT gearbox. However, in practice, WT gearboxes usually work in variable operation conditions, especially under different speeds or loads. Thus, vibration signals measured from the gearbox are characterized with nonlinearity and nonstationarity caused by varying speeds and loads, along with strong environment noises. Therefore, it is quite challenging to extract useful fault features from raw vibration signals for fault diagnosis. On the other hand, it is well-known that WT gearboxes are complex mechanical systems consisting of bearings, gears, shafts, and other mechanical components, and are usually coupled with other subsystems, such as generators. Thus, measured vibration signals using sensors installed on the house of the gearbox will contain multiple intrinsic oscillatory modes due to the interaction and coupling effects among different components and subsystems. Various mechanical rotating and reciprocating frequencies will complicate the measured vibration signals. As a result, vibration signals usually exhibit multiscale characteristics [20] and contain complex patterns at multiple time scales, which has been discussed in literature [20]–[22]. Such inherent multiscale characteristics cannot be captured by the traditional CNN architecture due to the lack of multiscale feature extraction ability. To overcome the limitation, this paper proposes a new multiscale CNN (MSCNN) architecture. This architecture is developed on top of the traditional CNN but incorporates the idea of multiscale learning and provides a more efficient and applicable way to extract fault signatures from multiple scales. As a result, it will enhance the feature learning capability and therefore improve the fault diagnosis performance. The main contributions of this paper are summarized as follows.

1) A new MSCNN architecture is proposed by introducing a multiscale coarse-grained layer into the traditional CNN. This MSCNN architecture can effectively extract the high-level features with a hierarchical learning structure using multiple pairs of convolutional and pooling layers. This also allows MSCNN to capture complementary and rich diagnostic information at different scales of raw vibration signals in parallel. The obtained multiscale feature representations have stronger discriminability and robustness over CNN and allow MSCNN to overcome the limitation of feature extraction in the traditional architecture that handles only a single time scale.
2) A novel end-to-end fault diagnosis system is developed based on the proposed MSCNN for the WT gearbox under different operational conditions. The system can automatically learn discriminative features directly from raw temporal vibration signals and classify different conditions simultaneously. Unlike traditional methods relying on manually defined or extracted features, the proposed design works well at a signal-level that keeps all information from the input. More importantly, it does not require any additional signal processing or expert knowledge, which has great potentials toward a general-purpose framework for intelligent fault diagnosis.
3) The proposed approach is evaluated through experiments on a WT gearbox test rig with a comprehensive performance evaluation. Compared with the traditional CNN and multiscale feature extraction methods, MSCNN achieves significantly better feature learning ability and diagnosis performance. Specifically, MSCNN exhibits superior robustness against noises of a large range than the traditional CNN, a promising result in real-world industrial applications. From the viewpoint of engineering application, the proposed MSCNN provides a new solution to the challenging WT gearbox fault diagnosis problem.

The rest of this paper is organized as follows. Section II briefly reviews the CNNs and their applications for fault diagnosis. Section III details our proposed MSCNN framework for WT gearbox fault diagnosis. Section IV demonstrates the experimental results on a WT gearbox test rig to evaluate the

effectiveness of the proposed framework. Finally, Section V draws the conclusions.

II. Related Works

A. Overview of CNNs

CNNs are a specialized kind of multilayer perceptrons neural network, which is biologically inspired to mimic the behavior of the mammalian visual cortex [23]. CNNs were originally proposed by LeCun *et al.* [24] for handwritten digits classification. Up to now, CNN has been extensively applied in various application domains including automatic speech recognition and natural language processing. It has been shown that CNN can deal with various types of signals, including one-dimensional (1-D) signals and sequences, two-dimensional (2-D) images, and three-dimensional (3-D) videos, due to their powerful ability of automatic feature extraction and classification. Different from the traditional fully connected neural networks, there are three key architectural ideas [25] behind CNNs, i.e., local connectivity, shared weights, and spatial pooling. These properties allow CNNs to optimize fewer parameters, learn more robust translation-invariant features, and achieve better generalization on many recognition tasks [13].

B. CNNs for Fault Diagnosis

In recent years, motivated by the success of CNNs in a variety of classification and recognition tasks, CNNs have also been investigated in the field of fault diagnosis, especially for rotating machineries such as bearings and gearboxes. Different from 2-D image data, sensory signals, such as vibration and current signals collected from the underlying machines, are naturally 1-D time series sampled by sensors with a certain time interval. To apply the 2-D CNN for fault diagnosis, some signal preprocessing or transformation procedures are required before input to the 2-D CNN model. In [26], Chen *et al.* adopted a 2-D CNN for gearbox fault diagnosis, in which the input matrix with a size of 16×16 for CNN was reshaped by a vector containing 256 statistic features including RMS values, standard deviation, skewness, kurtosis, rotating frequency, and applied load. In [27], a 2-D CNN model was utilized to recognize four different conditions of rotating machinery, where the input of CNN was discrete Fourier transform of two-channel vibration signals from two accelerometer sensors. In [28], the vibration signals were first processed using wavelet transform to form spectrogram images of size 32×32 and then these images were fed into a CNN model to learn the invariant representation and recognize the fault status of the rotor system.

In the case of raw sensory signals, 1-D CNN was recently developed on raw current signals for motor bearing fault detection [18], where two major blocks (i.e., feature extraction and classification) of a traditional fault detection approach were integrated together. Also, Abdeljaber *et al.* [29] applied 1-D CNN on normalized vibration signals to perform damage detection and localization of the structural damage in real time. It has been shown that 1-D CNN-based fault diagnosis methods are able to extract optimal fault features automatically from raw signals with proper training, which reduces the dependence on hand-crafted feature extraction. Inspired by these works, in this paper, we focus on exploring multiscale characteristics inherent in raw vibration signals, which are ignored by the traditional CNN, and propose a new MSCNN architecture in Section III.

III. Proposed MSCNN-Based Fault Diagnosis System for WT Gearbox

This paper focuses on the fault diagnosis of the WT gearbox under varying operational conditions. In such cases, for vibration signals measured under different speeds or loads, different conditions of the gearbox will have a considerable internal variability, which further challenges the class separability when simply extracting the fault signatures in a single scale with CNN. We argue that a more thorough and wide understanding of the signals, such as in multiple time or frequency scales, is required to enhance feature extraction and classification performance. In addition, faults occurred in gearbox components will introduce low-frequency impulse components and demodulated components. However, these fault signatures are easily masked by those inherent high-frequency components with dominant amplitudes or those strong background noises. To better extract fault signatures and capture multiscale characteristics of raw vibration signals, this paper proposes a new MSCNN architecture to incorporate multiscale feature learning into the traditional CNN architecture. The overall architecture of the proposed MSCNN is illustrated in Fig. 1. A key advantage of the proposed new architecture is to automatically learn high-level robust and useful fault features at different time scales directly from complex raw vibration signals in a parallel way through the hierarchical learning. It works in an end-to-end learning manner. Generally, it consists of three sequential stages: a multiscale coarse-grained stage, a multiscale feature learning stage, and a classification stage.

A. MSCNN Architecture

1) Multiscale Coarse-Grained Layer: The key idea of the proposed MSCNN is to incorporate multiscale feature learning ability into the traditional CNN architecture. In a recent study [30], the downsampling and smoothing were simultaneously used to implement multiscale operations of a raw time series input. Motivated by the work in [30], but differently, in this paper, we introduce a simple coarse-grained procedure [31] to represent the raw vibration signals at multiple time scales. The coarse-grained process has been successfully used for quantifying the complexity of complex physiologic time series [31] and recently for vibration time series signals [20], [21] measured from mechanical systems.

Given a measured vibration signal $x = \{x_1, x_2, \ldots, x_N\}$, where x_i is the value at time stamp i, and there are N timestamps for each signal. We construct consecutive coarse-grained signals $\{y^{(s)}\}$ by averaging the data points in the original signal x with a nonoverlapping window of increasing length s (also called the scale factor). Fig. 2 illustrates a coarse-grained procedure for scales $s = 2$ and $s = 3$. Each element of a coarse-grained signal is calculated by using the following equation [32]:

$$y_j^{(s)} = \frac{1}{s} \sum_{i=(j-1)s+1}^{js} x_i, 1 \le j \le \frac{N}{s}. \quad (1)$$

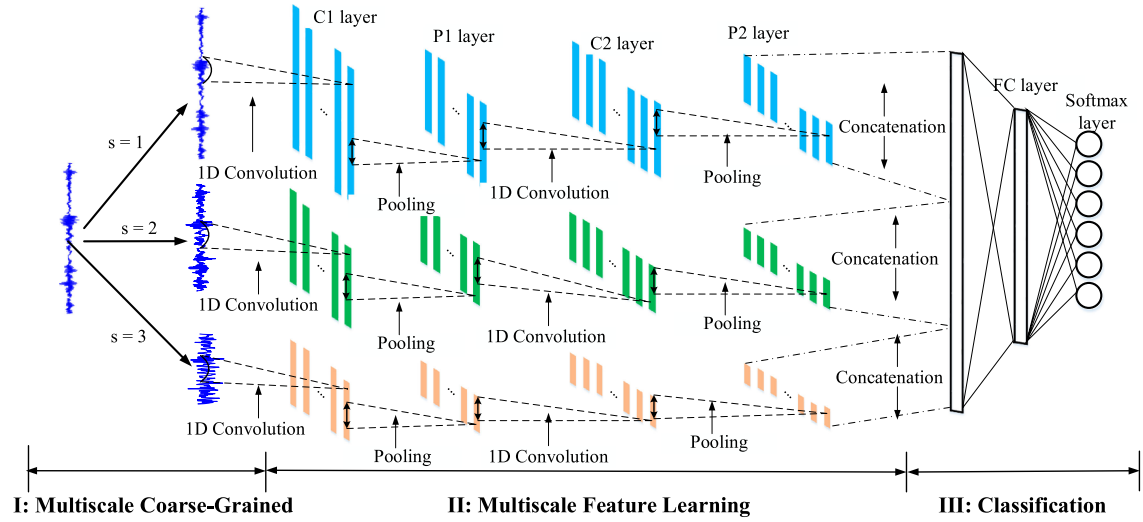


Fig. 1. Proposed MSCNN framework for fault diagnosis using 1-D vibration signals. The *i*th convolutional layer is denoted as C*i* layer, while the *i*th pooling layer is denoted as P*i* layer. The FC layer means fully connected layer. In this illustration, three scales ($s = 1, 2, 3$) are considered.

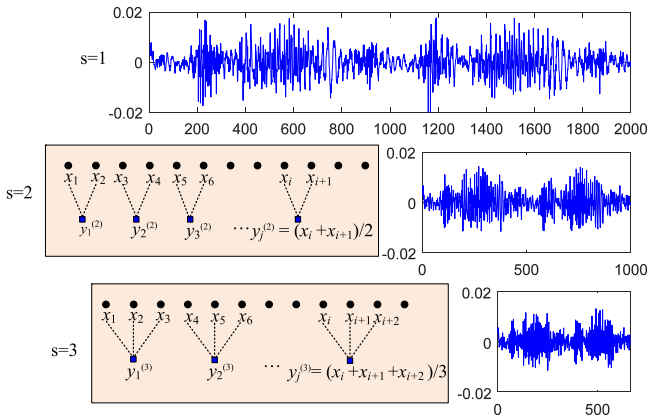


Fig. 2. Illustration of the coarse-grained procedure for $s = 2, 3$.

In fact, the coarse-grained operation naturally performs smoothing and downsampling of the original signals simultaneously. In other words, this operation can also be considered as a simple low-pass filtering process through moving average with a nonoverlapped window, thus those high-frequency perturbations and random noises will be filtered to some extent. Compared to the multiscale transformation used in [30], the coarse-grained operation is simple to implement and can reduce the model complexity and computational cost. As shown in Fig. 2, for different scale factors, we can obtain multiple filtered signals, each of which corresponds to different smoothed representations of the original signals. In the case of fault diagnosis, it provides a more efficient way to characterize vibration signals over a range of scales, and results in multiple coarse-grained signals containing different and complementary fault pattern information. In the following stage, we will learn more useful and robust feature representations from these coarse-grained signals and the original signals with stacked convolutional layers and pooling layers.

2) Multiscale Feature Learning: The goal of this stage is to learn high-level effective fault features from multiple coarse-grained signals from different time scales in a parallel manner through multiple pairs of convolutional layers and pooling layers. Concretely, as shown in Fig. 1, for each coarse-grained signal $\{y^{(s)}\}$, we use two pairs of convolutional layers (C1 and C2) and pooling layers (P1 and P2) are used to learn abstract and robust features. Specifically, the same filter size is used across all coarse-grained signals, and thus shorter signals would produce a larger local receptive field in the original signals. This way, each output feature map captures a different scale of the original signals.

The input of the first convolutional layer (C1) is a coarse-grained signal with length $L = N/s$. We slide the filter with a window size of m over the whole input signal to extract local features. Accordingly, the output z_i of the i-node in the feature map is defined by

$$z_i = \sigma(\mathbf{w}^{\mathrm{T}} y_{i:i+m-1} + b) \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^m$ represents the filter vector, b denotes the bias term, $y_{i:i+m-1}$ is an m-length subsignal of input signal y starting from the ith time step, and $\sigma(\cdot)$ is a nonlinear activation function. In this study, we choose rectified linear units (ReLU) [33] $\sigma(x) = \max(0, x)$ as the nonlinear activation function to prevent the vanishing gradient problem and accelerate the convergence of model.

As defined in (2), the output scalar z_i can be viewed as the activation of the filter on the corresponding subsignal. By sliding the filter from the beginning time step to the ending time step, the feature map of the jth filter can be denoted as

$$\mathbf{z}_j = [z_1, z_2, \ldots, z_{L-m+1}]. \tag{3}$$

Afterwards, the pooling layers (P1) are further applied to the feature maps generated by the convolutional layer (C1), thus enabling to extract the most important and location-invariant features. In this paper, the max-pooling with a pooling length of

p is adopted for calculating the local max value over the input feature map. Then, the kth pooled feature map can be obtained as

$$\mathbf{h}_k = [h_1, h_2, \ldots, h_{\frac{L-m}{p}+1}] \tag{4}$$

$$h_j = \max_{(j-1)p+1 \le i \le jp} \{z_i\}. \tag{5}$$

After the first pair of convolutional layer (C1) and max-pooling layer (P1), we obtain K_1 new feature maps. Then, these feature maps will act as the input of the second pair of convolutional layer (C2) and max-pooling layer (P2), and repeat the same operations in (2)–(5). Supposing that K_2 filters are used and learned in the second pair, the output of the pooling layer (P2) is K_2 new pooled feature maps, in which the kth one can be denoted as $\mathbf{h}'_k$.

Then, for each coarse-grained signal $\{y^{(s)}\}$, we can obtain its corresponding representation $\mathbf{q}^{(s)}$ learned through two alternating convolutional layers and pooling layers, which is the concatenation of all feature maps represented as

$$\mathbf{q}^{(s)} = [\mathbf{h}'_1, \mathbf{h}'_2, \ldots, \mathbf{h}'_{K_2}]. \tag{6}$$

Finally, we simply concatenate the learned representations of each coarse-grained signal to obtain a final multiscale representation of the raw input signal in a long vector, which is denoted as

$$\mathbf{q} = [\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \mathbf{q}^{(3)}]. \tag{7}$$

Obviously, compared to the representation learned from raw vibration signal on only one single scale, the learned multiscale representations may contain complementary and rich fault pattern features at multiple time scales. Thus, they can provide complementary discriminability among different conditions to a significant extent and enable better fault classification ability.

3) Classification: The classification task for the WT gearbox fault diagnosis studied in this paper is a multiclass classification problem. The multiscale representation $\mathbf{q}$ obtained in the feature extraction stage is directly fed to two additional layers. The first is a fully connected layer with ReLU units. In the output layer, we use a softmax function [17] that outputs a conditional probability for each class. Assuming that there are n classes of gearbox health conditions for the input sample x, its corresponding output probability $O_j \in [0, 1]$ for class j is calculated as

$$O_j = \frac{e^{(\theta^{(j)} x)}}{\sum_{j=1}^{n} e^{(\theta^{(j)} x)}}, j = 1, 2, \ldots, n \tag{8}$$

where θ is the model parameter to learn and $\sum_{j=1}^{n} O_j = 1$. It will be automatically optimized based on training samples.

Similar to the traditional CNN, MSCNN is trained using gradient descent through the back propagation (BP) algorithm. The input of MSCNN is the raw temporal vibration signals, and the output is the prediction of the gearbox conditions with different class labels indicating different health conditions. For MSCNN training, we adopt the cross-entropy between the predicted class labels and the true class labels as the loss function. The Adam optimization algorithm [34] is employed to minimize the loss function for its efficient computation and little memory. We also use a dropout technique [35] on fully connected layers in the classification stage of our MSCNN architecture to prevent overfitting. With the dropout technique, individual nodes are randomly removed with a certain probability, typically 0.5, at each training epoch. Also, this technique significantly improves the speed of training.

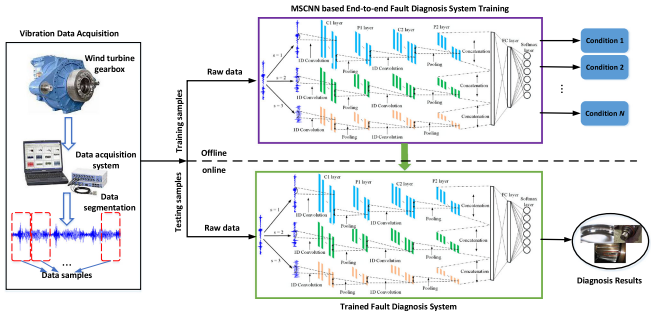


Fig. 3. Flowchart of the developed MSCNN-based fault diagnosis system for WT gearbox.

It is worth noting that our proposed MSCNN architecture is general and flexible, which may have different scales and different depths (i.e., pairs of convolutional layers and pooling layers). Specifically, MSCNN can effectively learn high-level fault features by using the hierarchical learning structure with multiple pairs of convolutional layers and pooling layers and capture complementary and rich diagnosis information at different scales. An MSCNN model with more scales and more layers can learn more robust and abstract fault features from more time scales, and is expected to have improved diagnosis performance. However, there is certain restriction relationship between the length of the input sample, the number of scales, and the depth. As is shown in (1), with the increase of the scale factor s, the length of coarse-grained signal decreases. Therefore, if more scales are considered, the longer input signal will be required to ensure that every coarse-grained signal contains enough information for feature extraction using convolutional layers and pooling layers. On the other hand, the pooling operation will also reduce the length of the extracted feature maps. Therefore, the deeper model also requires the longer input signal. In practical applications, we should make a suitable choice about the number of scales and the depth depending on the length of the input signal, which will be discussed in detail in Section IV.

B. MSCNN-Based Fault Diagnosis

In this section, an end-to-end fault diagnosis system based on the proposed MSCNN architecture is presented. The flowchart is shown in Fig. 3 and the general procedures are summarized as follows.

1) *Step 1:* Collect vibration data from different health conditions of the WT gearbox using the DAQ system. For each condition, we segment the whole signal into several small segments to obtain data samples for model training and testing.
2) *Step 2:* With training samples, we build an end-to-end fault diagnosis system based on the proposed MSCNN with the raw vibration data as input and the correspond-

ing condition labels as output. The whole model is trained offline using the BP algorithm to optimize all parameters, and multiscale feature learning and classification are accomplished simultaneously.

3) *Step 3:* We input the testing samples to the well-trained fault diagnosis system to automatically perform the calculation of multiscale features and directly diagnose the condition of WT gearbox.

Different from traditional fault diagnosis system performed on feature-level relying on those manual specific feature extraction algorithms, our proposed MSCNN-based fault diagnosis system has several important merits as follows.

1) *Multiscale*. MSCNN incorporates the multiscale learning and learns useful fault features at different time scales in a parallel manner, thus obtaining multiscale representations containing more rich and complementary diagnosis information and enabling better fault diagnosis performance.
2) *End-to-end*. MSCNN provides an end-to-end fault diagnosis system which is directly fed with raw input data through a deep network architecture and outputs the final fault classification results. Also, it builds and captures the complex mapping relationship between raw temporal vibration signals and health condition labels.
3) *General-purpose*. More importantly, instead of extracting those hand-crafted features, which are generally specific for certain diagnosis object, MSCNN only depends on the raw temporal signals without any complex time-consuming transformation, domain knowledge, and expert experience, and therefore is a general-purpose approach. Thus, it can be easily extended to deal with fault diagnosis problems of other industrial systems.

IV. Case Study

In this section, the proposed MSCNN framework is evaluated using experiments on a WT gearbox test rig. We start with the description of the test rig and experiments, and then the detailed results are shown and discussed.

A. Test Rig and DAQ

As shown in Fig. 4, an experimental WT gearbox test rig has been set up to illustrate the effectiveness of MSCNN for fault diagnosis. This test rig is designed to simulate the operation of a real WT system. A 3-kW induction motor coupled with a speed reducer with a ratio of 40:1 is operated by the speed controller, and is used to emulate the dynamics of a WT rotor to generate a lower shaft speed. Then, the lower speed at the output shaft of the reducer will increase through a two-stage parallel gearbox with total ratio 1:20 to drive a 3-kW three-phase permanent magnet synchronous generator for power generation. Finally, the generated power was consumed by the load bank.

In this test rig, various common periodic and irregular faults occurred in the gearbox, such as chipped tooth, broken tooth, cracked gear, and bearing outer race damage can be simulated. In this study, eight health conditions of the gearbox, including one normal, three gear faults, three bearing faults, and a shaft

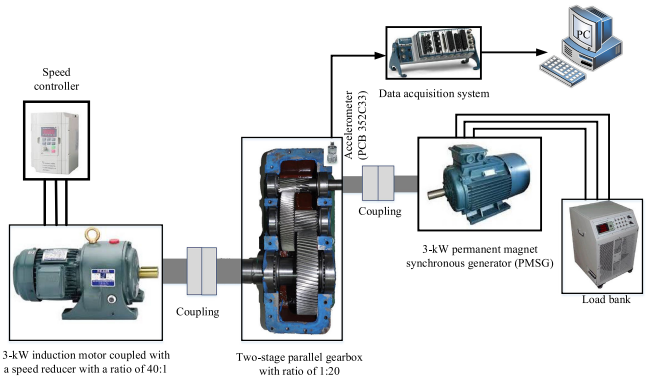


Fig. 4. Schematic of the WT gearbox test rig setup.

TABLE I
Description of Considered Gearbox Health Conditions

Label	Fault Description	Motor Speed (rpm)	Load (kW)
1	Normal	200/300/400/500/600	3/1.5
2	Gear broken tooth	200/300/400/500/600	3/1.5
3	Gear chipped tooth	200/300/400/500/600	3/1.5
4	Gear pitted tooth	200/300/400/500/600	3/1.5
5	Bearing inner race defect	200/300/400/500/600	3/1.5
6	Bearing out race defect	200/300/400/500/600	3/1.5
7	Bearing ball defect	200/300/400/500/600	3/1.5
8	Imbalanced shaft	200/300/400/500/600	3/1.5

fault, are tested. All considered fault conditions occurred on the second stage (i.e., the high-speed end) of the tested gearbox, which is commonly encountered in the large-scale WT gearbox. In each experiment, a damaged component is installed inside the tested gearbox and other components remain normal. Each experiment is conducted under different operation conditions, i.e., five motor driving speeds and two loads. Vibration signals were acquired using a piezoelectric accelerometer (PCB 352C33) mounted on the end cover of the second stage of the gearbox by a DAQ system with a sampling frequency of 10 kHz. The detailed description of the data set and condition labels is summarized in Table I, where eight health conditions are listed. In this data set, the same health condition under different speeds and loads is treated as one class. There are 2600 samples for each health condition under ten operation conditions, and each sample contains 2000 data points. Therefore, the data set totally contains $2600 \times 8 = 20\,800$ samples for eight health conditions. We adopted the 10-fold cross-validation method for performance evaluation, which is commonly used in the literature [29], [36]. We split the original vibration data set into ten equal-sized subsets. Of the ten subsets, a single subset is retained as the testing set for testing the model, and the remaining nine subsets are used as the training set. The cross-validation process is then repeated ten times, with each of the ten subsets used exactly once as the testing set. The average classification results of the testing set over ten folds are reported.

B. Performance Metrics

The fault diagnosis problem studied in this paper is naturally a multiclass classification (eight classes in our case). For per-

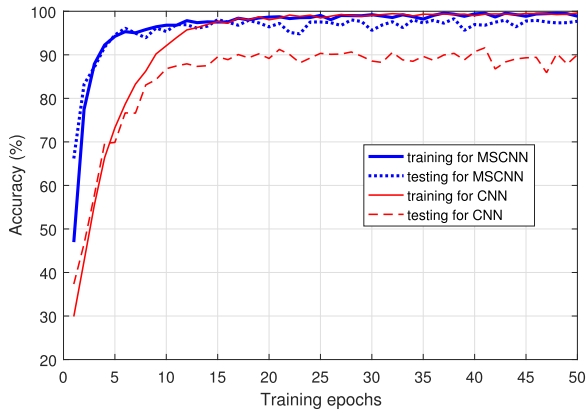


Fig. 5. Training performance curve in terms of overall accuracy.

formance evaluation and comparison, F_1 score [37] is adopted in this study. It is a commonly used comprehensive metric to measure the performance of a classification method, which is defined as

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

where true positive (TP) is correctly classified as positive samples, false positive (FP) is misclassified as positive samples, true negative (TN) is correctly classified as negative samples, and false negative (FN) is misclassified as negative samples. In the following studies, F_1 score for each class (health condition) calculated with the proposed method and the compared methods will be reported.

C. Multiscale Learning Evaluation: MSCNN Versus CNN

We first evaluate the performance of our proposed method by using the collected gearbox data set. In order to extract more rich fault pattern features helpful for classification from raw vibration signal, four scales are considered in our proposed MSCNN framework. The input size of MSCNN is 2000, i.e., the length of a vibration signal sample. The number of filters for two convolutional layers C1 and C2 are 16 and 32, respectively. The same filter length of 100 for all filters is used for capturing enough information from a local region of the input signal. The pooling length for two pooling layers P1 and P2 is set to 2. In the classification stage, the number of neurons in a fully connected hidden layer is chosen as 1024. The output size of MSCNN is 8, corresponding to the number of considered health conditions of the gearbox. Our proposed MSCNN is trained from scratch using the training set described in Section IV-A. As mentioned earlier, MSCNN is optimized by Adam gradient descent optimization algorithm [34] with a mini-batch size of 50 samples. The number of training epochs is 50. The learning rate is initialized to 0.001 with no decay on each update. The dropout rate of 0.5 is used for fully connected layers to avoid the overfitting risk. For comparison, the traditional CNN, i.e., corresponding to the MSCNN with one scale, is considered.

Fig. 5 shows the overall accuracy curves of the training set and the testing set for both MSCNN and CNN methods during the whole training process over 50 epochs for one fold. It can be seen that the training accuracy and the testing accuracy for both methods reach the stable values after 20 epochs; no overfitting is observed. This demonstrated that the training set used here is sufficiently large to train the model with the six-layer network structure. In addition, notice that MSCNN converges to the stable value faster than CNN, which means that training an optimal MSCNN model requires less time in practice. The testing classification results over ten folds with both MSCNN and CNN methods are shown in Fig. 6(a) and (b) using the confusion matrix, respectively. It gives the correctly classified samples and misclassified samples for each health condition. The x-axis and y-axis represent predicted labels and true labels, respectively. Compared with MSCNN and CNN, it is easily found that for each condition, more samples are misclassified when using CNN, leading to a poor diagnosis performance. This implies that fault features in different conditions are masked in noisy raw signals and therefore it is difficult to identify using the traditional CNN performed on only one scale.

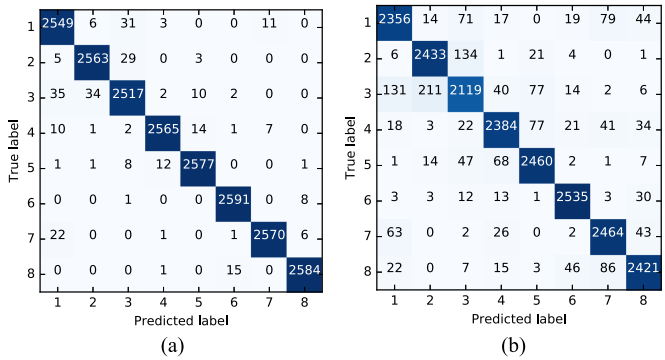
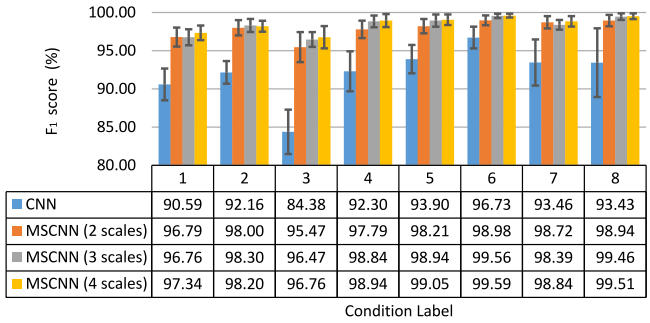


Fig. 6. Confusion matrix results with (a) MSCNN and (b) CNN.

D. Discussions on Effects of Scale and Depth

The proposed MSCNN model involves different number of scales and different depth (i.e., the number of pairs of convolutional layers and pooling layers). Both parameters will impact the diagnosis performance of the proposed MSCNN. Herein, we investigate the effects of scale and depth.

1) Effects of Scale: Vibration signals generally show great variations in multiple observation scales, therefore it is necessary to take multiscale information into consideration. In this study, multiscale coarse-grained signals are used in the MSCNN framework in order to extract fault features at different scales. To quantitatively evaluate the effects of the chosen scales in MSCNN on the classification performance, different scales ranging from two to four are considered. In these experiments, two-layer structure MSCNN (i.e., two convolutional layers and two pooling layers) was adopted. The average results of F_1 score over ten folds for each condition are shown in Fig. 7, where the errorbar represents the standard deviation showing the stability of classification performance. It is easily observed that, for each condition, the proposed MSCNN from two to four scales always outperforms the traditional CNN. In general, as more scales are incorporated, MSCNN can achieve better and more reliable performance. Specifically, a significant increase



	1	2	3	4	5	6	7	8
CNN	90.59	92.16	84.38	92.30	93.90	96.73	93.46	93.43
MSCNN (2 scales)	96.79	98.00	95.47	97.79	98.21	98.98	98.72	98.94
MSCNN (3 scales)	96.76	98.30	96.47	98.84	98.94	99.56	98.39	99.46
MSCNN (4 scales)	97.34	98.20	96.76	98.94	99.05	99.59	98.84	99.51

Fig. 7. Diagnosis performance on the testing set using traditional CNN and the proposed MSCNN with different scales from 2 to 4.

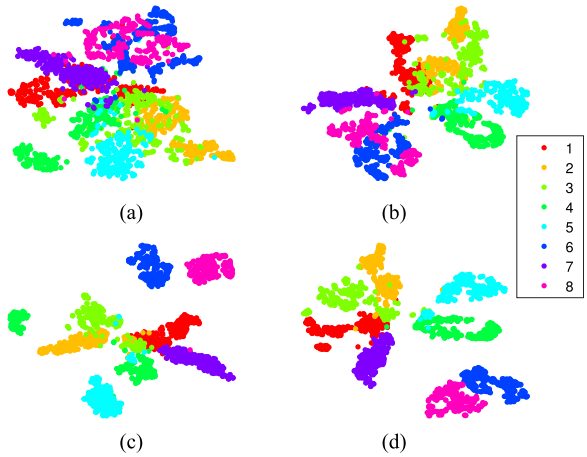


Fig. 8. Feature visualization via t-SNE reduced from the learned multiscale representations for the testing data set with (a) the traditional CNN and the proposed MSCNN with (b) two scales, (c) three scales, and (d) four scales.

from CNN to MSCNN with two scales can be observed for each condition. Furthermore, compared to the traditional CNN, a smaller standard deviation for each condition can be noticed especially for MSCNN with three or four scales, which shows a more reliable performance. This result demonstrates that the proposed MSCNN can learn more discriminative and robust features from different scales of raw vibration signals than the traditional CNN that extracts representative features at only one scale, i.e., raw vibration signal itself. Therefore, the proposed MSCNN can produce an enhanced fault diagnosis performance.

To better understand the benefits of multiscale features incorporation in the proposed MSCNN, we use the t-SNE technique [38] to reduce the dimensions of the learned multiscale representations in two dimensions for visualization. The reduced 2-D feature representations of the testing set with different scales are shown in Fig. 8, where the different colors represent different health conditions described in Table I. From Fig. 8, we can observe that with the increase of scales from 1 (corresponding to the traditional CNN) to 4 for MSCNN, the features within the same health condition present better cluster performance while the features of different health conditions are much more separable, which in turn enables the classifier easier to classify different health conditions. This result shows a strong identifiability and high robustness for different health conditions

TABLE II
COST TIME FOR MSCNN FOR DIFFERENT SCALES AND CNN

Methods	Training Time (s)	Testing Time (ms)
CNN	8.1588	0.1488
MSCNN(2 scales)	11.2809	0.1698
MSCNN(3 scales)	12.4587	0.1899
MSCNN(4 scales)	13.9728	0.2036

of the WT gearbox under ten different speeds and loads. This further proves that incorporating multiscale features can greatly improve feature learning and therefore fault diagnosis ability.

In the present study, all experiments are conducted on a workstation (Intel Core (TM) 3.4 GHz processor with 64 GB of RAM) and an NVIDIA GeForce 1080Ti GPU on a Ubuntu system platform. All considered models are trained from scratch using the training set, and therefore much more training time is required. We evaluate the training time and the testing time for the traditional CNN and the proposed MSCNN with different scales. The results are given in Table II. In this table, the training time of one epoch averaged over 50 epochs and the testing time of one testing sample are calculated, respectively. In terms of training time, the proposed MSCNN consumes much more time than CNN. As more scales are considered, more time is needed. This can be easily explained that due to the incorporation of multiple scales, the MSCNN will introduce more parameters to be trained and therefore require more time. However, as mentioned in Section III-B, all models are trained offline, and then used for online fault diagnosis. Therefore, in fault diagnosis applications, the training time will not directly affect the performance of the fault diagnosis system, while the testing time is the focus when the designed fault diagnosis system is put into use. During the testing phase, the maximum time spent by MSCNN with four scales for one test sample is only 0.2036 ms, which is just slightly larger than that of CNN or MSCNN with fewer scales. Therefore, our MSCNN approach is also suitable for real-time monitoring and diagnosis.

2) Effects of Depth: The depth of the MSCNN can determine the abstraction level of extracted features. Low-level features obtained from the raw signal may greatly suffer from speed or load variations and background noises. The abstraction level of fault features can significantly impact the classification results. To test the effects of the depth on diagnosis performance, MSCNNs with one to three layers were tested, and the results are shown in Fig. 9. It can be first seen that, for both MSCNN and the traditional CNN, the classification performance increases with the depth. This is because MSCNN with more layers can learn and extract more abstract and robust features at higher levels that are helpful in classification. The MSCNN models with two and three layers have similar performance (about 98%) in terms of F_1 score, which is significantly better (with an improvement of 7%) than the MSCNN model with only one layer. We also calculated the training time and the testing time of the CNN and the MSCNN with different layers, and the results are given in Table III. It is obvious that the computation overhead increases with depth for both methods. In order to reduce the computational cost, however, we chose only two layers in the feature extraction. For practical applications, to deal with more

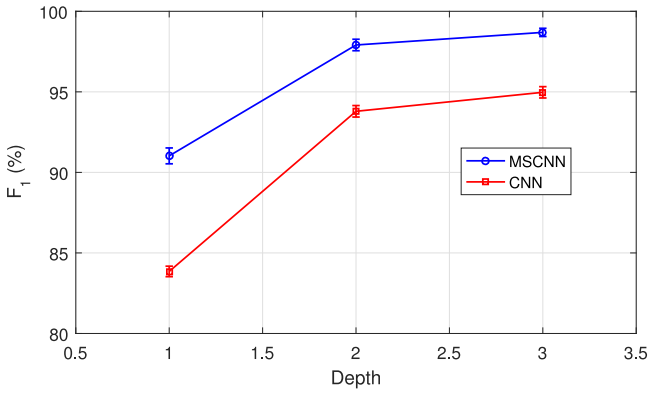


Fig. 9. Effects of depth on diagnosis performance using the proposed MSCNN and the traditional CNN.

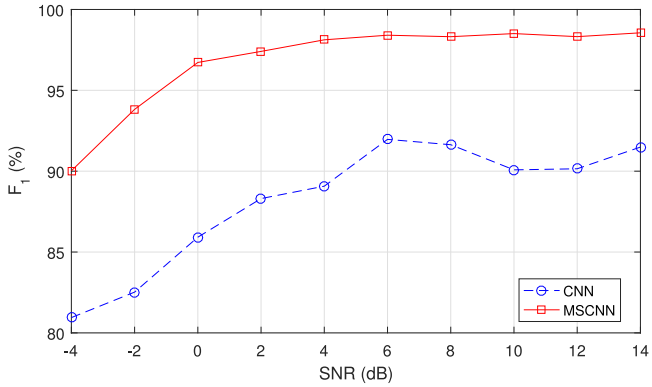


Fig. 10. Diagnosis performance on noisy signals with different SNRs using the proposed MSCNN and the traditional CNN.

TABLE III
COST TIME OF MSCNN AND CNN FOR DIFFERENT DEPTHS

Method	Training time (s)			Testing time (ms)		
	1 layer	2 layers	3 layers	1 layer	2 layers	3 layers
CNN	4.5966	8.1588	11.0719	0.1153	0.1488	0.1676
MSCNN	8.0454	13.9728	18.4642	0.1511	0.2036	0.2299

One layer refers to one pair of the convolutional layer and the pooling layer.

challenging and complex diagnosis tasks, a deeper model may be considered to further improve performance, especially when large amount of data are available.

E. Robustness Against Noise: MSCNC Versus CNN

In real-world applications, the WT gearbox often operates in complicated working environments and the measured vibrations signals are easily contaminated by strong background noises. Therefore, it is necessary to check the robustness of the proposed MSCNN against noise. For this reason, we inject additive Gaussian white noise [39] to the raw vibration signals to construct noisy signals with different signal-to-noise ratios (SNRs). The SNR is defined as follows:

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \tag{10}$$

where P_{signal} and P_{noise} denote the power of the raw signal and the injected noise, respectively.

In this study, we evaluate the proposed MSCNN approach using noisy signals with different SNRs ranging from -4 to 14 dB. The evaluation results for both MSCNN and CNN are shown in Fig. 10, where the average results of F_1 scores for all eight conditions are calculated as the evaluation metric. For MSCNN, four scales are considered. It is obvious that the proposed MSCNN significantly outperforms the traditional CNN, with over 90% testing performance in terms of F_1 score within all considered SNR levels. Specifically, we can find that MSCNN obtains above 95% testing performance even when SNR is 0 dB, where the power of the noise is equal to that of the raw signal, and even increases to 98% at a stable level when SNR is greater than 4 dB. Furthermore, the proposed MSCNN exhibits a more remarkable advantage over the traditional CNN at lower SNR levels from -4 to 2 dB. The largest difference reaches to 11.3% at the SNR level of -2 dB. To this end, the proposed MSCNN presents superior robustness against noise without any additional denoising preprocessing, which means that MSCNN can learn and extract more robust features than the traditional CNN under the noisy environment. The result also shows that the proposed MSCNN is more suitable in practical industrial environments where more complicated environment noises and measurement interferences exist. Our proposed MSCNN method will be further evaluated using the vibration data collected from real WT gearboxes in wind farms in our future work.

F. Comparison With Existing Multiscale CNN Architecture

In a recent study in [30], a similar MCNN was proposed to address the time series classification issue and has shown the state-of-the-art performance on a large number of benchmark data sets. Herein, MCNN is compared with our proposed MSCNN method. In general, MSCNN is different from MCNN in the three following aspects.

1) Convolution locality. MSCNN applies only local convolutions to perform local feature extraction from each coarse-grained subsignals with multiple pairs of convolutional and pooling layers; MCNN applies both local convolutions and full convolutions. Concretely, in MCNN, the local convolution is first used to extract the features for each branch, then the extracted features from different branches are concatenated, and finally full convolution is used to perform feature extraction and fusion.
2) Model structure. In MCNN, there are three branches/subprocesses including identity mapping, smoothing, and downsampling. For the smoothing and downsampling branches, there are multiple subscale signals corresponding to various degrees of smoothness and downsampling rates, respectively. In our proposed MSCNN, each branch corresponds to a coarse-grained subsignal differed only by the scale factors, and the number of branches depends merely on the number of scales considered. In other words, MSCNN focuses on the local feature learning in each coarse-grained signal, while MCNN performs a global feature extraction for each branch.

TABLE IV
COMPARISON (THE AVERAGE VALUE ± THE STANDARD DEVIATION) OVER 50 RANDOM TRIALS BETWEEN MSCNN, MSCNN-II, AND MCNN

Methods	F_1 score (%)	Training time (s)	Testing time (ms)
MCNN	98.81±0.0049	36.7319±0.2198	0.4883±0.0022
MSCNN	98.56±0.0036	14.5389±0.2248	0.1807±0.0024
MSCNN-II	98.03±0.0056	14.5994±0.1758	0.1813±0.0021

3) Multiscale operation. MSCNN uses the multiscale coarse-grained operation, while MSCNN adopts the smoothing and downsampling operation. The coarse-grained operation performs moving average smoothing and downsampling simultaneously, and each coarse-grained signal is calculated using all information from the raw signal. The down-sampling only keeps every kth data points based on the downsampling rate k, thus losing some information contained in raw signals since it will discard some data points; the smoothing uses the moving average with different window sizes. Also, the coarse-grained operation in MSCNN is easier to deploy in hardware implementation and takes less computation overhead compared to the multiscale transformation in MCNN.

Similar to MCNN, we apply both local and global convolutions in our proposed MSCNN model and named it the MSNNN-II model. The comparison results with three different models of MCNN, MSCNN, and MSCNN-II are given in Table IV in terms of F_1 score and computation time (including the average training time per epoch over all training samples and the average testing time per sample) averaged over 50 random trials. The 10-fold cross-validation is performed for each trial. Four scales are considered and two pairs of convolutional layers and pooling layers are used in all three models. Specifically, MCNN has one convolutional layer followed by a pooling layer in the local convolution and full convolution, respectively. It can be found that MCNN is slightly better than MSCNN in terms of F_1 score. Compared with MSCNN and MSCNN-II, MSCNN performs better than MSCNN-II and suggested that the global convolution may not improve the performance of MSCNN. It can be concluded that the key difference between MSNN and MCNN lies in the multiscale operation. Then, we further validated if the performance difference in terms of F_1 score between our proposed MSCNN and MCNN is significant. A two-tailed *t*-test was conducted for validation under the significance level of 0.05 similar to the one in [40]. The *t*-test results demonstrated that given the significance level of 0.05, there is no significant difference ($p = 0.0713 > 0.05$) between MSCNN and MCNN.

Also, as further shown in Table IV, the computation overhead of MSCNN is lower than the MCNN, mainly due to the simpler structure and a smaller number of parameters to learn of MSCNN. It should be noted that the training time and the testing time should be important factors when developing a diagnosis device for online and real-time diagnosis. From Table IV, it is obvious that the training time of MCNN is 2.5 times more than that of MSCNN. It means that MSCNN is more efficient than MCNN especially when much more training data are available and need to be processed in real-world applications. This allows a shorter development cycle of diagnosis systems and faster releases of updates, which hugely lowers the cost of the system development. For online testing, MSCNN also took less time, which means that the diagnosis system can be used efficiently and enable a faster diagnosis in practice, which is helpful to make a timely decision for the operators. Therefore, in practical applications, MSCNN is more computationally efficient at a competitive diagnosis performance. From the comprehensive consideration of F_1 score and training and testing time, it would be more effective to apply the MSCNN method, especially for real-world industrial applications.

G. Comparison With Traditional Multiscale Approaches

In order to demonstrate the merits of our proposed MSCNN, three traditional feature extraction methods, including multiscale entropy (MSE) [20], wavelet package decomposition (WPD) [9], and EMD [10], are compared. MSE is used for calculating the sample entropy over multiple coarse-grained signals from different time scales of a raw vibration signal. WPD and EMD can decompose a complex vibration signal into several different components containing different frequency bands information. These methods have been considered as multiscale feature extraction ones and widely used in vibration-based fault diagnosis for rotary machines [9], [10], [20]. Among three methods, MSE is the simplest multiscale approach only requiring the average operation in the time domain, while EMD and WPD require complex and time-consuming signal transformation or decomposition. For the MSE, entropy features over 20 scales are calculated for each sample. For the WPD, we use Daubechies 4 wavelet to decompose each sample into four levels, thus producing 16 nodes in different frequency bands. Then, energy values of 16 nodes are computed as features. For the EMD, we calculate the energy vector of first six intrinsic mode functions of each sample. All extracted features with different methods are fed to an SVM classifier with a Radial basis kernel function to perform the fault diagnosis task. In the case of multiclass predictions for SVM, the one-versus-one strategy is adopted here. The kernel parameter and penalty factor of SVM are determined by a five-fold cross-validation method.

Table V gives the comparison results between the proposed MSCNN and traditional multiscale feature extraction methods in terms of F_1 score over ten folds, where the classification results for each health condition and the average performance (averaged on all eight conditions) are shown. The best performance is highlighted in bold. It is easily found that the proposed MSCNN method achieves the best overall performance of 98.53%. For each condition, MSCNN obtained the over 97% F_1 score, and more stable performance, which corresponds to a smaller standard deviation. Unfortunately, three traditional feature extraction methods perform worse with below 85% F_1 score. This can be explained that these traditional methods only extract a small number of features at different time scales or different frequency bands, while a lot of information cannot be effectively captured and even will be lost in the feature extraction process. The results also indicate that all extracted energy features and entropy features between different classes may be similar and therefore cannot be distinguished. On the contrary,

TABLE V
COMPARISON RESULTS WITH DIFFERENT METHODS IN TERMS OF F_1 SCORE FOR EACH CONDITION (%)

Methods	Condition Label								Average
	1	2	3	4	5	6	7	8	
MSE-SVM	75.13±4.75	87.41±3.24	81.65±2.95	90.34±1.90	85.36±1.72	89.32±2.54	73.30±3.42	80.53±1.89	82.88
WPD-SVM	80.40±4.31	86.92±2.84	72.50±5.40	96.93±2.99	78.95±3.21	87.18±4.71	93.78±3.47	77.80±3.80	84.31
EMD-SVM	71.01±1.04	70.42±2.32	62.70 ±3.04	72.09±1.97	71.30±1.57	83.26±1.82	82.39±1.14	68.51±2.03	72.71
Proposed MSCNN	**97.34±0.96**	**98.20±0.72**	**96.76±1.46**	**98.94±0.85**	**99.05±0.69**	**99.59±0.26**	**98.84±0.68**	**99.51±0.38**	**98.53**

MSCNN can automatically learn useful fault features at multiple time scales from raw signals without depending on the hand-crafted features, thus largely reducing the demand of prior knowledge and time-consuming signal processing algorithms. In summary, traditional multiscale feature approaches are on feature-level, while MSCNN is performed on signal-level that will keep all information from the input. Therefore, the proposed MSCNN method has the potential to provide a useful alternative as a general-purpose classification procedure for intelligent fault diagnosis due to its end-to-end feature learning ability.

V. CONCLUSION

This paper focused on the multiscale feature learning of complex vibration signals and proposed a new MSCNN architecture for intelligent fault diagnosis of the WT gearbox under different operational conditions. The major contribution of the new architecture was the incorporation of multiscale feature learning by introducing a coarse-grained layer into the traditional CNN. The proposed MSCNN can automatically and effectively learn complementary and rich fault features at different time scales in a parallel way directly from raw vibration signals with multiple layers, which greatly enhances feature learning ability and fault diagnosis performance. Accordingly, an MSCNN-based end-to-end fault diagnosis system was also developed. Different from those traditional fault diagnosis methods greatly relying on the handcrafted features and a shallow classification model, the developed MSCNN-based system can perform automatic feature extraction and classification simultaneously without depending on complex signal processing algorithms and prior knowledge. Experimental results demonstrated that the proposed MSCNN significantly outperformed the traditional CNN in terms of feature learning, robustness against noise, and classification performance. Compared with the traditional multiscale feature extraction methods, our proposed MSCNN presented an end-to-end fault diagnosis capacity and achieved much better performance, which are crucial for WT gearboxes that are playing an important role in reliable operation and emergency response in large-scale wind farms. More importantly, it offers a new general-purpose framework for the field of fault diagnosis and can be easily extended to deal with different machines and industrial systems.

In our future work, we will verify the scalability of our proposed MSCNN on a real large-scale WT gearbox. In addition, in the case of imbalanced data distribution in the fault diagnosis field, we will further investigate the imbalanced learning based multiscale representation learning approach to mitigate the impact from skewed data distribution between normal and faulty data, so that the performance of learning algorithms can be further improved.

REFERENCES

[1] F. Blaabjerg and K. Ma, "Future on power electronics for wind turbine systems," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 1, no. 3, pp. 139–152, Sep. 2013.

[2] W. Qiao and D. Lu, "A survey on wind turbine condition monitoring and fault diagnosis–part i: Components and subsystems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6536–6545, Oct. 2015.

[3] Y. Feng, Y. Qiu, C. J. Crabtree, H. Long, and P. J. Tavner, "Monitoring wind turbine gearboxes," *Wind Energy*, vol. 16, no. 5, pp. 728–740, Jul. 2013.

[4] W. Qiao and D. Lu, "A survey on wind turbine condition monitoring and fault diagnosis–Part II: Signals and signal processing methods," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6546–6557, Oct. 2015.

[5] Z. Zhang, A. Verma, and A. Kusiak, "Fault analysis and condition monitoring of the wind turbine gearbox," *IEEE Trans. Energy Convers.*, vol. 27, no. 2, pp. 526–535, Jun. 2012.

[6] 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.

[7] W. Yang, P. J. Tavner, and W. Tian, "Wind turbine condition monitoring based on an improved spline-kernelled chirplet transform," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6565–6574, Oct. 2015.

[8] J. Wang, R. X. Gao, and R. Yan, "Integration of EEMD and ICA for wind turbine gearbox diagnosis," *Wind Energy*, vol. 17, no. 5, pp. 757–773, May 2014.

[9] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, pp. 1–15, Mar. 2014.

[10] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, no. 1, pp. 108–126, Feb. 2013.

[11] B. Tang, T. Song, F. Li, and L. Deng, "Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine," *Renew. Energy*, vol. 62, pp. 1–9, Feb. 2014.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

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

[14] 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, pp. 303–315, May 2016.

[15] C. Lu, Z.-Y. Wang, W.-L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Process.*, vol. 130, pp. 377–388, Jan. 2017.

[16] M. Gan, C. Wang, and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings," *Mech. Syst. Signal Process.*, vol. 72, pp. 92–104, May. 2016.

[17] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. 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.

[18] 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.

[19] T. de Bruin, K. Verbert, and R. Babuška, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 28, no. 3, pp. 523–533, Mar. 2017.

[20] L. Zhang, G. Xiong, H. Liu, H. Zou, and W. Guo, "Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference," *Expert Syst. with Appl.*, vol. 37, no. 8, pp. 6077–6085, Aug. 2010.

[21] H. Liu and M. Han, "A fault diagnosis method based on local mean decomposition and multi-scale entropy for roller bearings," *Mech. Mach. Theory*, vol. 75, pp. 67–78, May 2014.
[22] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mech. Syst. Signal Process.*, vol. 85, pp. 746–759, Feb. 2017.
[23] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *J. Phys.*, vol. 148, no. 3, pp. 574–591, Oct. 1959.
[24] 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.
[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, Nov. 2016.
[26] Z. Q. Chen, C. Li, and R. V. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock Vibrat.*, vol. 2015, no. 2, pp. 1–10, Apr. 2015.
[27] O. Janssens *et al.*, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vibrat.*, vol. 377, pp. 331–345, Sep. 2016.
[28] J. Wang, J. Zhuang, L. Duan, and W. Cheng, "A multi-scale convolution neural network for featureless fault diagnosis," in *Proc. Int. Symp. Flexible Autom.*, Aug. 2016, pp. 65–70.
[29] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D. J. Inman, "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks," *J. Sound Vibrat.*, vol. 388, pp. 154–170, Feb. 2017.
[30] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," [Online]. Available: https://arxiv.org/abs/1603.06995.
[31] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys. Rev. Lett.*, vol. 89, no. 6, Feb. 2002, Art. no. 068102.
[32] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals." *Phys. Rev. E*, vol. 71, no. 1, Feb. 2005, Art. no. 021906.
[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learning*, Jun. 2010, pp. 807–814.
[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learning Representations*, May 2015.
[35] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting." *J. Mach. Learning Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[36] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.
[37] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[38] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learning Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[39] R. Liu, G. Meng, B. Yang, C. Sun, and X. Chen, "Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1310–1320, Jun. 2017.
[40] H. He and J. A. Starzyk, "A self-organizing learning array system for power quality classification based on wavelet transform," *IEEE Trans. Power Del.*, vol. 21, no. 1, pp. 286–295, Jan. 2006.

Guoqian Jiang (M'18) received the B.S. degree in measurement control technology and instrumentation and the Ph.D. degree in control science and engineering from Yanshan University, Qinhuangdao, China, in 2011 and 2018, respectively.

He was a joint Ph.D. student with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA, from 2015 to 2017. He is currently a Lecturer with the School of Electrical Engineering, Yanshan University. His research interests include advanced signal processing algorithms, intelligent fault diagnostics and prognostics, and deep learning for machine health monitoring.

Haibo He (SM'11–F'18) received the B.S. and M.S. degrees in electrical engineering from Huazhong University of Science and Technology, Huazhong, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from Ohio University, Athens, OH, USA, in 2006.

From 2006 to 2009, he was an Assistant Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology. He is currently the Robert Haas Endowed Chair Professor with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA. He has published one sole-author research book (Wiley), edited one book (Wiley-IEEE) and six conference proceedings (Springer), and authored and coauthored more than 300 peer-reviewed journal and conference papers. His research interests include adaptive dynamic programing, computational intelligence, machine learning and data mining, and various applications.

Dr. He served as the General Chair of the IEEE Symposium Series on Computational Intelligence. He was a recipient of the IEEE International Conference on Communications Best Paper Award (2014), IEEE Computational Intelligence Society Outstanding Early Career Award (2014), National Science Foundation CAREER Award (2011), and Providence Business News "Rising Star Innovator Award" (2011). Currently, he is the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Jun Yan (M'17) received the B.Eng. degree in information and communication engineering from Zhejiang University, Zhejiang Sheng, China, in 2011, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Rhode Island, Kingston, RI, USA, in 2013 and 2017, respectively.

He is currently an Assistant Professor with the Concordia Institute for Information Systems Engineering, Concordia University, Montréal, QC, Canada. His research interests include cyber-physical systems and security, smart and resilient infrastructures, computational intelligence, machine learning, and self-adaptive systems.

Dr. Yan was a recipient of the IEEE International Joint Conference on Neural Networks Best Student Paper Award (2016), the IEEE International Conference on Communications Best Paper Award (2014), the IEEE Communications Society Best Readings Award (2013), and the IEEE TRANSACTIONS ON SMART GRID Best Reviewers Award (2016).

Ping Xie received the M.S. degree in measuring technology and instrumentation and the Ph.D. degree in circuits and systems from Yanshan University, Qinhuangdao, China, in 1996 and 2006, respectively.

Since 2009, she has been a Professor with the School of Electrical Engineering, Yanshan University, Qinhuangdao, China. Her research interests include multimodal information processing and fusion, brain–computer interface, rehabilitation robot biofeedback control, wearable equipment development, and condition monitoring and fault diagnosis of complex systems.