Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks

Rui Zhao, Dongzhe Wang, Ruqiang Yan, *Senior Member, IEEE*, Kezhi Mao, Fei Shen, and Jinjiang Wang

***Abstract*—In modern industries, machine health monitoring systems (MHMS) have been applied wildly with the goal of realizing predictive maintenance including failures tracking, downtime reduction, and assets preservation. In the era of big machinery data, data-driven MHMS have achieved remarkable results in the detection of faults after the occurrence of certain failures (diagnosis) and prediction of the future working conditions and the remaining useful life (prognosis). The numerical representation for raw sensory data is the key stone for various successful MHMS. Conventional methods are the labor-extensive as they usually depend on handcrafted features, which require expert knowledge. Inspired by the success of deep learning methods that redefine representation learning from raw data, we propose local feature-based gated recurrent unit (LFGRU) networks. It is a hybrid approach that combines handcrafted feature design with automatic feature learning for machine health monitoring. First, features from windows of input time series are extracted. Then, an enhanced bidirectional GRU network is designed and applied on the generated sequence of local features to learn the representation. A supervised learning layer is finally trained to predict machine condition. Experiments on three machine health monitoring tasks: tool wear prediction, gearbox fault diagnosis, and incipient bearing fault detection verify the effectiveness and generalization of the proposed LFGRU.**

***Index Terms*—Fault diagnosis, feature engineering, gated recurrent unit (GRU), machine health monitoring (MHM), tool wear prediction.**

I. Introduction

THE development of advanced sensing technologies, wireless communications, and computing systems has generated a huge amount of data for manufacturing systems in recent years [1], [2]. Meanwhile, it motivates the research of data-driven machine health monitoring systems (MHMS) that are capable to detect faults and predict working conditions [2]–[4]. Data-driven MHMS train models based on historical measured data, and make decisions upon the online data collected from sensors of the monitored equipment. It demonstrates that data-driven MHMS can update themselves with online collected data in the real time. As shown in Fig. 1, data-driven MHMS includes two phases: one is training model based on historical sensor signals, and the other one is applying the trained model on online sensor signals to make decisions. However, it is difficult to design a complete set of features to represent machine condition, considering these commonly adopted time, frequency and time-frequency domain analysis provide a broad range of measures. Therefore, feature extraction/selection methods as a kind of information fusion are usually conducted after handcrafted feature design. To learn a more discriminative feature space, feature extraction methods such as principal component analysis [5] and factor analysis [6] have been used to transform original features into a novel informative feature space, while feature selection including fisher ratio [7] and distance measures [8] attempts to select subsets of features. Then, the derived features are fed into machine learning models, e.g., support vector machine and linear regression models, to make predictions of machine condition. For example, Widodo and Yang [9] presented an overview of machine condition monitoring and fault diagnosis based on support vector machine. Prieto *et al.* [10] extracted significant statistical-time features first and then apply a hierarchical neural network for bearing fault classification. Yang *et al.* [11] extracted energy features at various vital frequencies

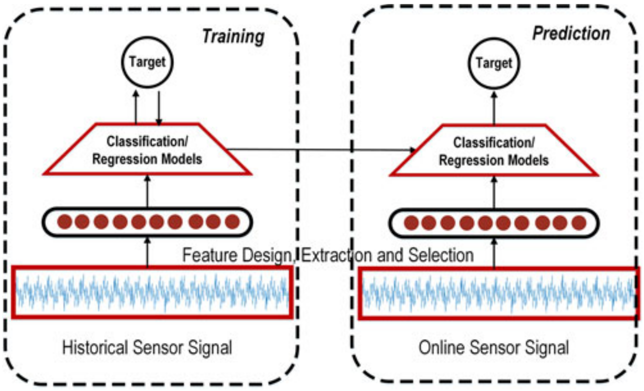
Training
Prediction
Target
Target
Classification/ Regression Models
Classification/ Regression Models
Feature Design, Extraction and Selection
Historical Sensor Signal
Online Sensor Signal

Fig. 1. Illustration of data-driven machine health monitoring system.

Manuscript received January 10, 2017; revised May 10, 2017; accepted June 5, 2017. Date of publication July 31, 2017; date of current version December 8, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 51575102. *(Corresponding author: Rui Zhao.)*

R. Zhao, D. Wang, and K. Mao are with the School of Electrical and Electronic Engineering, Nanyang Technology University, Singapore, 639798 (e-mail: rzhao001@e.ntu.edu.sg; DWANG015@e.ntu.edu.sg; ezkmao@ntu.edu.sg).

R. Yan and F. Shen are with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: ruqiang@seu.edu.cn; 936833562@qq.com).

J. Wang is with the Faculty of Mechanical Engineering, China University of Petroleum, Beijing 102249, China (e-mail: jwang@cup.edu.cn).

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.2017.2733438

and trained a random forest classifier for induction motor fault diagnosis. He *et al.* [12] constructed a k-nearest neighbor (KNN) classifier on time-domain features extracted by empirical mode decomposition for bearing faults detection as one submodule. Regardless of which kind of machine learning models are applied, it is shown that the representation determines the upper-bound performances of machine learning algorithms [13]. However, the above pipeline system may have some potential concerns as follows.

1) *Expert knowledge*: Handcrafted feature design and adaptation of appropriate feature extraction/selection methods all require prior domain knowledge and expert expertise, which may not be met in all scenarios.
2) *Joint optimization*: Considering feature learning and machine learning models that work in a cascaded way, it is impossible to jointly optimize them. One part should be fixed, and the other part can be adjusted.
3) *Temporal information*: The machinery data are usually sampled by sensors and expressed in a sequential form and the sequential information behind sensor data is quite vital to represent machine condition. However, the handcrafted feature design usually extract measures from the whole range of time-series data, which may not capture its intrinsic temporal information.

As a branch of machine learning models, deep learning (DL) provides a powerful solution to these above concerns [14], [15]. DL is featured by its capability of extracting hierarchical representations from input data by building deep neural networks (DNNs) with multiple layers of nonlinear transformations. Intuitively, one layer operation can be regarded as a transformation from input values to output values. The state-of-arts of various application areas including computer vision, automatic speech recognition, natural language processing, audio recognition, and bioinformatics have been achieved by DL techniques. In the past few years, many DL models including stacked autoencoder (SAE) and deep belief network (DBN) have been developed in the field of machine health monitoring [16]. Generally, DL-based MHMS as an end-to-end system has the capability of learning features from raw input directly, which do not require extensive expert knowledge. Since feature learning and target prediction are incorporated into the whole neural network, all model parameters can be trained and optimized jointly. Some researchers focused on the pretraining of DNN. SAE [17] and DBNs [18] are adopted to facilitate the training of DNN and learn discriminative representation for machinery data. Although these pretrained DNN models can directly work on raw sensory time-series data, input dimensionality is easily over 100, even 1000, which will increase model size [19]. The huge number of model parameters may lead to heavy computational cost and overfitting problems. To control the model size, features in a low-dimensional space are usually fed into pretrained DNN models. For example, Jia *et al.* [20] first extracted the frequency spectra features of time-series data, and then fed them into SAE-DNN for rotating machinery fault diagnosis. Guo *et al.* [21] designed multidomain statistical features including time-domain features, frequency-domain features, and time–frequency domain features and built a SAE-DNN upon them for bearing fault diagnosis. Ma *et al.* [22] adopted a DBN-DNN framework for degradation assessment under an accelerated bearing life test, in which the raw input consists of statistical feature, root mean square (rms) fitted by Weibull distribution and the frequency domain features. Chen *et al.* [23] fed a feature vector consisting of load and speed measure, time-domain features and frequency-domain features into DBN-DNN for gearbox fault diagnosis. Therefore, to relief the above concerns, we focus on DL-based machine health monitoring methods here.

In this paper, we present a new framework named local feature-based gated recurrent units (LFGRU) networks as a generalized MHMS. Gated recurrent units networks as a variant of recurrent neural network is able to process memories of sequential data by storing previous inputs in the internal state of networks and map from the entire history of previous inputs to target vectors in principal. In our proposed framework, local features are first extracted from segments or windows of time-series data. Then, an enhanced GRU networks: bidirectional GRU networks with weighted local features averaging has been proposed to learn representation from the sequence of local features. Supervised learning layer is added on the top to map the learned representation to targets. In our framework, DL models are applied to handcrafted features design instead of raw time-series data, considering the model size can be controlled. Different from previous DL-based MHMS that built upon extracted features of the whole range of time-series data, our proposed LFGRU extracts local features from consecutive windows of time-series data, which can keep the order information among windows. What is more, this kind of local-feature extraction are suitable for multisensory data. The features extracted from synchronized segments of multiple sensory data can be concatenated to summarize the information of all sensors at the same time step. Gated recurrent units were proposed to relief the problem of gradient exploding or vanishing in recurrent neural networks (RNN) with a controlled model complexity compared to another RNN variant: long-short term memory network [24]–[26]. In our proposed LFGRU, the bidirectional recurrent structure has been incorporated, which can access the sequential data in two directions including forward and backward ones with two separate hidden layers so that our model can fully explore the context of the input, i.e., the past and future information at each state. The averaging local feature is able to generate representation of sequential data without capturing ordering information, which can provide a supplementary to representation generated by bidirectional GRU. Considering information in the middle of the sequence might be easily lost in the bidirectional GRU, the weighted average of local features is adopted here that assigns large weights to features located in the middle. To verify the effectiveness and generalizability of our proposed LFGRU, three different machine health monitoring tasks including tool wear sensing, gearbox fault diagnosis, and incipient fault diagnosis of rolling element bearings are introduced. Several state-of-the-art models are compared with our proposed model. The main contributions of our work can be summarized as follows.

1) Our proposed framework can be considered as a hybrid approach of handcrafted feature design and automatic

feature learning by the DL model. The local-feature extraction scheme can reduce the model size of the applied enhanced GRU networks and the enhanced GRU networks are able to encode the temporal information on the generated sequence of local features. Therefore, in our proposed framework, not only the model size of GRU can be controlled to prevent overfitting, but also the handcrafted feature design is not required to be at expert level.

2) An enhanced GRU networks has been proposed. Bidirectional recurrent structure is first incorporated to GRU network to capture the future and past context jointly. Next, the center-biased feature averaging operation can provide a direct representation of the input sequence without considering order information, which can be supplementary to the output of bidirectional GRU. The concatenated vector is regarded as the final representation learned by our proposed enhanced GRU networks.
3) This approach is suitable for multisensory scenario, which is illustrated in the following Section III-A and verified in the experimental part.
4) Comprehensive experimental studies including three case studies including tool wear prediction, gearbox fault diagnosis and incipient bearing fault detection are conducted. The effectiveness and generalization capability behind our proposed framework have been verified.

This paper is organized as follows. In Section II, RNN and GRU models are reviewed. Then, our proposed LFGRU is presented in Section III. After that, experimental results are illustrated in Section IV. Finally, concluding remarks are provided in Section V.

II. RNN and GRU Networks

RNN including GRU is related to our work. In this section, a brief introduction to RNN and GRU is given.

As stated in [14], RNN are able to memorize arbitrary-length sequences of input patterns by building connections between units from a directed cycle. The key component in RNN is the transition function in each time step t, which takes the current time information $\mathbf{x}_t$ and the previous hidden output $\mathbf{h}_{t-1}$ and updates the current hidden output as follows:

$$\mathbf{h}_t = \mathbb{H}(\mathbf{x}_t, \mathbf{h}_{t-1}) \tag{1}$$

where $\mathbb{H}$ defines a nonlinear and differentiable transformation function. Due to the recurrent structure, $\mathbf{h}_{t-1}$ in (1) can be regarded as a memory of previous inputs, i.e., RNN can keep the memory of previous inputs in the network's internal state. Therefore, after processing the whole sequence, the hidden output at the last time step, i.e., $\mathbf{h}_T$ can be regarded as a vector encoding the original sequential data. Supervised learning layer is added on top to map the obtained representation $\mathbf{h}_T$ to targets and the model can be trained via backpropagation through time [27]. Different formulations of transition functions derive different RNN models. Vanilla RNN adopts a linear transformation function with a nonlinear activation function as follows:

$$\mathbf{h}_t = \varphi(\mathbf{W}\mathbf{x}_t + \mathbf{H}\mathbf{h}_{t-1} + \mathbf{b}) \tag{2}$$

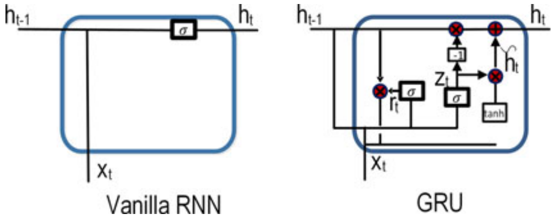

Fig. 2. Illustrations of Vanilla RNN and GRU models.

where $\mathbf{W} \in \mathbb{R}^{d\times k}$ and $\mathbf{H} \in \mathbb{R}^{d\times d}$ represent transformation matrices and $\mathbf{b} \in \mathbb{R}^d$ is the bias vector and φ is the nonlinear activation function such as sigmoid and tanh functions. Due to the vanishing gradient problem during backpropagation for model training, vanilla RNN may not capture long-term dependencies. It means that $\mathbf{h}_t$ may forget the information in the early stage of sequential data. To alleviate this issue, long short-term memory networks were first presented by introducing gates function in the design of transition function [25]. In our paper, we adopt another RNN variant: gated recurrent units (GRUs) that can be regarded as a simpler version of LSTMs [24], [28]. In GRU, two gates including a reset gate r that adjusts the incorporation of new input with the previous memory and an update gate z that controls the preservation of the previous memory are introduced. The transition functions in hidden units of GRU are given as follows:

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{W}^z\mathbf{x}_t + \mathbf{V}^z\mathbf{h}_{t-1} + \mathbf{b}^z) \\ \mathbf{r}_t &= \sigma(\mathbf{W}^r\mathbf{x}_t + \mathbf{V}^r\mathbf{h}_{t-1} + \mathbf{b}^r) \\ \tilde{\mathbf{h}_t} &= \tanh(\mathbf{W}^c\mathbf{x}_t + \mathbf{V}^c(\mathbf{r}_t \odot \mathbf{h}_{t-1})) \\ \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}_t} \end{aligned} \tag{3}$$

where model parameters including all $\mathbf{W} \in \mathbb{R}^{d\times k}$, $\mathbf{V} \in \mathbb{R}^{d\times d}$, and $\mathbf{b} \in \mathbb{R}^d$ are shared by all time steps and learned during model training, $\odot$ denotes the element-wise product, k is a hyperparameter that represents the dimensionality of hidden vectors. For an intuitive illustration, if the update gate is closed, i.e., $\mathbf{z}_t = 0$, the information in the initial time step can be kept no matter how long the sequence is. The illustration of vanilla RNN and GRU has been given in Fig. 2.

In our proposed framework, an enhanced GRU networks has been proposed with the bidirectional structure and weighted feature averaging scheme. First, the bidirectional structure is applied to the capture of future and previous context, which can improve the expressiveness of the GRU model. The adopted weighted feature averaging simply derives the mean vector of representations in each time step with a center-biased weighting scheme. The final representation consists of two parts: one is the output of bidirectional GRU and the other is the weighted mean feature, which are supplementary to each other.

III. LFGRU Networks

In this section, our proposed LFGRU networks will be presented in the scenario of multisensory machine monitoring. As shown in Fig. 3, the enhanced GRU network is applied on the

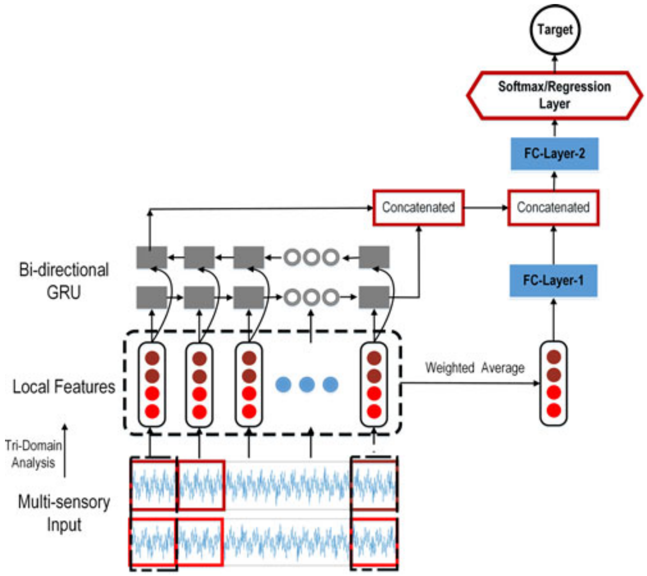

Fig. 3. Architecture of our proposed LFGRU. As an end-to-end system, this framework is able to predict target from multi-sensory input.

sequence of local features extracted from raw sensor inputs to learn representation of machine condition and predict the corresponding target.

We assume each data collected from monitored machine consists of one sensor input, which is time-series data denoted as $\mathbf{x} = [x_1, x_2, \ldots, x_l]$ where l is the length of data sample and training data has a corresponding target value such as fault type or tool wear that are defined in the specific applications.

A. Local-Feature Extraction

Each sensory input is first divided into T local segments and each segment is a window of original signal with a length of $\frac{l}{T}$. For example, the jth local window is a segment starting from time step $(j-1)\frac{l}{T}$ to $(j-1)\frac{l}{T} + \frac{l}{T} - 1$ denoted as $\mathbf{x}_{(j-1)\frac{l}{T}:(j-1)\frac{l}{T}+\frac{l}{T}-1}$. Then, tridomain features including time, frequency, and time-frequency ones are extracted from each local window. The details of these handcraft features design will be elaborated in the following experiments. Therefore, the original sensory input can be transformed to a sequence of local features as

$$\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T] \tag{4}$$

where $\mathbf{c}_j \in \mathbb{R}^m$ consists of m features extracted from the jth local window. It is shown that this operation can be easily extended to multisensory scenario. Local features can be extracted from synchronized windows of multiple sensor signals, respectively, and concatenated together into one feature vector as $\mathbf{c}_i$.

After local-feature extraction, the sequence of feature vectors $\mathbf{c}$ with a length of T can be generated. Compared to the original time series $\mathbf{x}$, the local-feature sequence is much shorter and can convey more discriminative information compared to the noisy original signal. Compared to conventional feature extraction conducted in MHMS that may abandon sequential characteristic, local features are designed and extracted from windows of the original noisy signal, which are arranged in order to form a sequence.

B. Bidirectional GRU With Weighted Feature Averaging

Then, GRU is applied on the generated local feature sequence to learn representation. Here, an enhanced GRU model has been proposed named bidirectional GRU with weighted feature averaging.

1) Bidirectional GRU: The incorporated bidirectionality of recurrent structure can increase the model capacity and flexibility. As shown in Fig. 3, the bidirectional recurrent structure can enable GRU to process the sequence input in two directions including forward and backward ways with two individual hidden layers. Therefore, each hidden layer at one certain time step can capture past (forward direction) and future (backward direction) context jointly. In addition, the bidirectional structure can reinforce the memory of the beginning and the end stages of the raw time-series input. In bidirectional GRU, the complete hidden element representation $\mathbf{h}^T$ at the last time step is the concatenated vector of the outputs of forward and backward processes as follows:

$$\mathbf{h}_T = \overrightarrow{\mathbf{h}}_T \oplus \overleftarrow{\mathbf{h}}_1 \tag{5}$$

where → and ← denote forward and backward processes, respectively, and the corresponding hidden vector is updated as follows:

$$\begin{aligned}\overrightarrow{\mathbf{h}}_t &= \overrightarrow{\mathbb{H}}(\mathbf{x}_t, \overrightarrow{\mathbf{h}}_{t-1})\\ \overleftarrow{\mathbf{h}}_t &= \overleftarrow{\mathbb{H}}(\mathbf{x}_t, \overleftarrow{\mathbf{h}}_{t+1}).\end{aligned} \tag{6}$$

In bidirectional GRU, the function $\mathbb{H}$ is defined by (3).

2) Weighted Feature Averaging: The hidden output of the bidirectional GRU at the last time step $\mathbf{h}_T$ can be regarded as the representation of the raw sensor signal. However, information in the middle range of the sequence might be lost in bidirectional GRU. Considering the beginning and end ranges of sequence contribute a lot to the outputs of backward GRU and forward GRU, respectively. Therefore, weighted feature averaging is introduced to provide another view of the sequence of local features $\mathbf{c}$. The average feature vector $\bar{\mathbf{c}}$ is given as

$$\bar{\mathbf{c}} = \sum_{k=1}^{T} w_k \mathbf{c}_k \tag{7}$$

where k denotes the index for time step. To highlight the impact of the middle local features, weights are designed as follows:

$$w_k = \frac{\exp(q(k))}{\sum_{j=1}^{T} \exp(q(j))} \tag{8}$$

where

$$q(k) = \min(k-1, T-k). \tag{9}$$

To give a clear illustration of the above equations, the weights derived by (8) for a length-10 sequence of local features are illustrated in Fig. 4. Then, the weighted average of local features is fed into a fully-connected dense layer F_1, and the output of

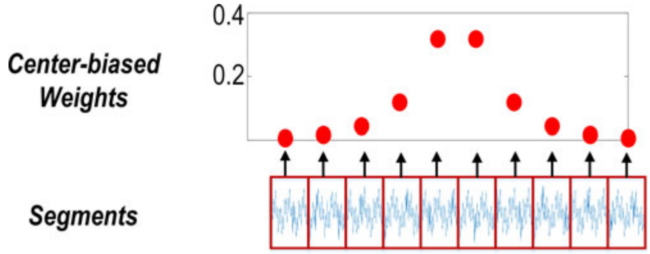

Fig. 4. Visualization of center-biased weights for a sequence of 10 local features.

this dense layer is concatenated to $\mathbf{h}_T$ generated by bidirectional GRU to derive the final representation $\mathbf{u}$

$$\mathbf{u} = \mathbf{h}_T \oplus F_1(\bar{\mathbf{c}}). \tag{10}$$

C. Supervised Learning Layer

At last, the learned final representation $\mathbf{u}$ is passed into another fully-connected dense layer F_2 and supervised learning layer. If the targets are discrete labels such as fault types, the supervised learning layer can be softmax layer, which is defined as

$$P\left(\frac{\tilde{y}=j}{F_2(\mathbf{u})}\right) = \frac{e^{F_2(\mathbf{u})^T \mathbf{w}_j}}{\sum_{k=1}^{K} e^{F_2(\mathbf{u})^T \mathbf{w}_k}} \tag{11}$$

where K is the number of labels and $\mathbf{w}$ denotes parameters of softmax layer. If the targets are continuous values such as remaining useful life (RUL) estimation and tool wear depth, the supervised learning layer can be a liner-regression layer given by

$$\tilde{y} = \mathbf{W} F_2(\mathbf{u}) + b \tag{12}$$

where $\mathbf{W}$ and b denote transformation matrix and bias value in the liner regression layer. The error between predicted values and ground truth values in training data can be calculated and backpropagated to train the parameters in the whole model. Then, the trained model can be applied on the unseen input data to make prediction about machine condition. The whole framework has been illustrated in Fig. 3.

IV. Experiments

To test the performances of our proposed LFGRU MHMS, three real-life case studies including tool wear prediction, gearbox fault diagnosis, and incipient fault diagnosis of rolling element bearings are conducted.

A. Descriptions of Datasets

1) Tool Wear Prediction: This dataset was collected from a high-speed CNC machine operated under dry milling operations [29]. The schematic diagram of experimental platform has been shown in Fig. 5. The operation parameters are as follows: the running speed of the spindle was 10 400 r/min; the feed rate in x-direction was 1555 mm/min; The depth of cut (radial) in y-direction was 0.125 mm; the depth of cut (axial) in z-direction was 0.2 mm. To acquire online data related to this CNC machine's operation condition, a Kistler quartz 3-component platform dynamometer was mounted between the workpiece and the machining table to measure cutting forces, while three Kistler Piezo accelerometers were mounted on the workpiece to measure the machine tool vibration in x-, y-, z-directions, respectively. DAQ NI PCI1200 was adopted to perform in-process measurements including force and vibration in three directions (x, y, z) with a continuous sampling frequency of 50 KHz during the tool wear test. Therefore, the sensory data consists of seven channels: force in three directions, vibration in three directions, and AE-rms. The corresponding flank wear of each individual flute was measured offline using a LEICA MZ12 microscope after finishing each surface, which is considered to be one cut number in the following data analysis, which will be the target value. The task is defined as the prediction of the actual flank wear (offline measurement) from the seven-channel sensory data (online measurement). Finally, three individual cutter records named $c1$, $c4$, and $c6$ were selected as our dataset. Each test contains 315 data samples, while each data sample has a corresponding flank wear. For training/testing splitting, a three-fold setting is adopted that two tests are used as training domain and the rest one is used as testing domain. Therefore, three different testing cases can be created, which are denoted as C1, C4 and C6. For example, the training/testing splitting scenario C1 is referred to the case that $c4$ and $c6$ are adopted as training data and $c1$ is used as testing data.

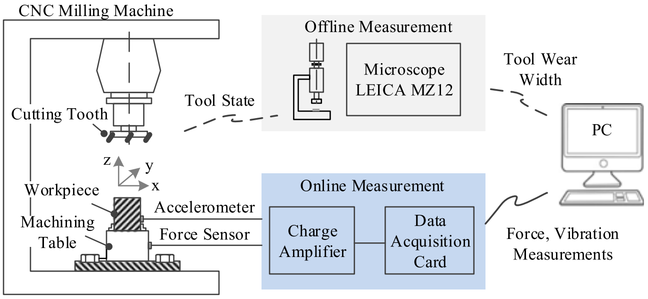

Fig. 5. Schematic of the experimental setup for tool wear prediction [6].

2) Gearbox Fault Diagnosis: This experiment was conducted in the Drivetrain Dynamics Simulator (DDS). The DDS is composed of four units including the motor, the planetary gearbox, the parallel gearbox, and brake as shown in Fig. 6. In the experiment, the faults of gear and bearing were investigated under two different operating conditions where rotating speed and load configuration are set as 20 HZ–0 V and 30 HZ–2 V. These two fault locations have their own fault types as described in Table I, where the real images of these fault types and locations are shown in Fig. 7. Therefore, four different fault diagnosis datasets are created and each of them is a five-class categorization task (four fault conditions and one health condition). The sensing configuration is described as follows: seven vibrating 608A11 sensors whose frequency range, measuring range, and accuracy are 0.5 Hz–10 kHz, ± 50 g, and 100 mV/g were adopted in the surface of DDS test bed. Three of them measured the vibration of planetary gearbox in three directions: x, y, and z, other three of them measured the three direction vibrations of the gear box, and the rest one was used

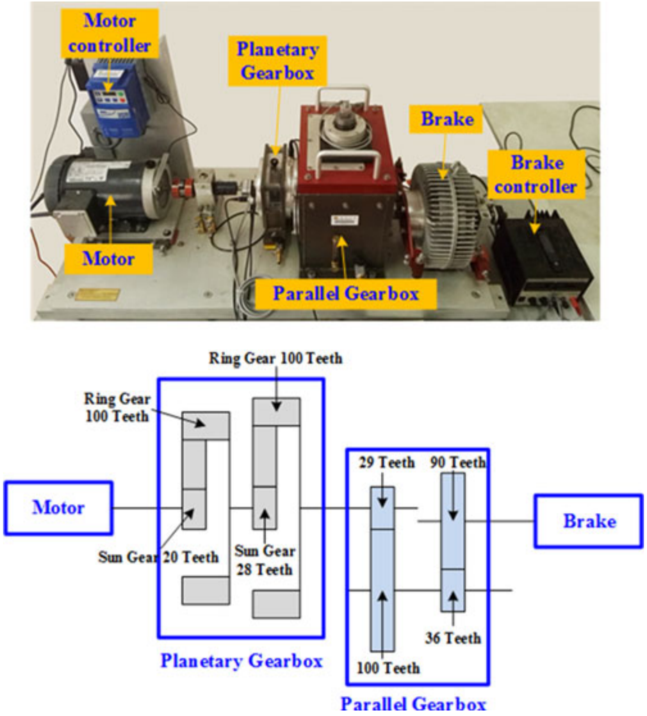

Fig. 6. Schematic of the experimental setup for gearbox fault diagnosis.

TABLE I
FAULT TYPES OF BEARING AND GEAR COMPONENTS

Component	Types	Description
Bearing	Ball	A crack occurs in the ball
	Combo	A crack occurs in both inner and outer ring.
	Inner	A crack occurs in the inner ring.
	Outer	A crack occurs in the outer ring.
Gear	Chipped	A crack occurs in the feet.
	Miss	One of feet is missing.
	Root	A crack occurs in the root of feet.
	Surface	The wear occurs in the surface.

to measure the driving motor. A torque sensor (model: FT293; measuring range: ± 5 V; accuracy: 4 N · m/V) was mounted between the motor and the planetary gearbox to measure the load, and a compact spectra PAD data acquisition instrument (Max 20 channels) was adopted for signal collecting with 1024 Hz sampling frequency and 512 s sampling window.

3) Incipient Fault Diagnosis of Rolling Element Bearings: This testing scenario is introduced to verify the performance of our model in the area of incipient fault diagnosis. Here, we used the experimental data from the bearing data center in the Case Western Reserve University (CWRU)[1] following the procedures adopted in [30] that vibration data under the smallest damage radius 7 m ils are selected. Due to page limits, the description of CWRU experimental platform and data acquisition system are skipped, which can be referred in [30]. Here, we create a four-class classification task. The corresponding classes include health condition, inner fault, ball fault, and outer fault and each class contains 800 data samples. Each data sample is the vibration signal acquired on fan end whose length is fixed to 20 000.

B. Experimental Setup

In our experiments, several methods are compared as follows:

- SVM/SVR: Support vector machine/Regression with rbf kernel;
- MLP: Neural network with two hidden layers;
- KNN: k-nearest neighbors
- SAE-DNN:
  - ◇ RNN: Vanilla RNN;
  - ◇ GRU: Gated recurrent units networks;
  - ◇ BiGRU: Bidirectional gated recurrent units networks;
  - ◇ LFGRU: Our proposed local featured-based gated recurrent units.

Here • denotes the handcrafted feature from the whole range of raw signal and ◇ denotes the local-feature extraction following the steps illustrated in Section III-A.

In tool wear prediction tasks, considering the multisensory input contain seven channels, the dimensionality of the handcrafted feature vector is 70, which includes time-domain, frequency-domain, and time–frequency domain features as illustrated in Table II. Here, the wavelet energy feature is the energy of an eight-level wavelet packet decomposition using db1, which corresponds to the wavelet coefficient with higher energy that is related to the characteristic frequency of the machine. In SVR, we search the best regularization parameter from $[0.001, 0.01, 0.1, 1, 10]$. In KNN, we search the best neighbor from $[3, 5, 7, 9]$. The layer sizes behind MLP and SAE-DNN are unified to be [140, 200] and their top layer is a linear regression layer to predict tool wear depth. For the last four methods, the shape of the input sequence of local features is 20×70 and the sizes of hidden recurrent layer in these four RNN are unified to be the same as 100. In our proposed method, the sizes of FC-Layer 1 F_1 and FC-Layer 2 F_2 are set to be 100 and 400, respectively. One of gradient descent optimization algorithms: RMSprop is adopted to train these four models [31]. The other hyperparameters including learning rates and epoch number are all searched via validation dataset. To quantify the performance of all compared methods, two measures to evaluate regression loss are used including mean absolute error (MAE) and root mean squared error (RMSE). MAE is the average value of the absolute values of the errors. RMSE is the square root of the average of the square of all of the errors. The corresponding equations for the calculations of these two measures over n testing samples are given as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\tilde{y}_i - y_i| \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2} \tag{14}$$

where y_i and $\tilde{y}_i$ are true and predicted tool wear depths.

[1] The dataset has been kindly provided at http://csegroups.case.edu/bearingdatacenter/home.

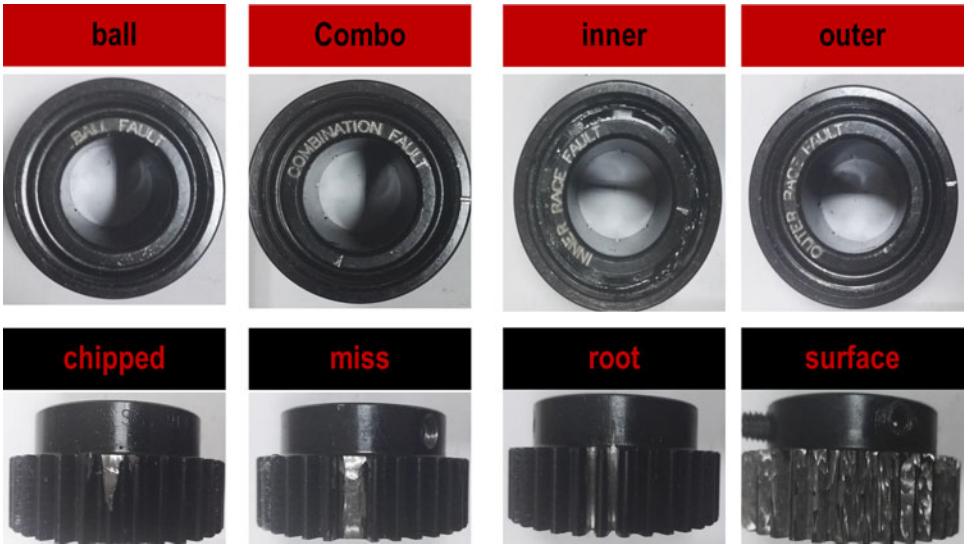

Fig. 7. Fault types Illustrations.

TABLE II
HANDCRAFTED FEATURE SETS FOR GEARBOX FAULT DIAGNOSIS AND TOOL WEAR PREDICTION

Quantity	Equations
RMS	$z_{\text{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} z_i^2}$
Variance	$\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2$
Maximum	$\max(z)$
Skewness	$E[(\frac{z-\mu}{\sigma})^3]$
Kurtosis	$E[(\frac{z-\mu}{\sigma})^4]$
Peak-to-Peak	$\max(z) - \min(z)$
Spectral Skewness	$\sum_{i=1}^{k}(\frac{f_i-\bar{f}}{\sigma})^3 S(f_i)$
Spectral Kurtosis	$\sum_{i=1}^{k}(\frac{f_i-\bar{f}}{\sigma})^4 S(f_i)$
Spectral Power	$\sum_{i=1}^{k}(f_i)^3 S(f_i)$
Wavelet Energy	$\sum_{i=1}^{N} wt_{\phi}^2(i)/N$

TABLE III
(A) MAE AND (B) RMSE ACHIEVED BY COMPARED METHODS IN TOOL WEAR PREDICTION TASKS

	(a) MAE			(b) RMSE		
Algorithms	Cases			Cases		
	C1	C4	C6	C1	C4	C6
SVR	6.9	10.0	31.1	9.6	12.4	34.3
KNN	10.3	12.7	28.7	13.0	15.4	31.6
MLP	11.2	11.6	30.1	13.7	14.1	31.9
SAE-DNN	10.9	9.5	29.5	13.7	11.8	31.2
RNN	8.6	8.1	10.1	10.8	10.7	25.5
GRU	5.9	7.0	11.6	7.9	8.8	12.9
BiGRU	5.5	7.4	9.4	6.8	9.2	11.1
LFGRU	**4.0**	**6.9**	**5.8**	**5.4**	**8.3**	**8.2**

Bold face indicates best performances.

In gearbox fault diagnosis tasks, the dimensionality of the handcrafted feature vector is 27 that the first nine features described in Table II are extracted for three sensors on three directions. In SVM and KNN, we select their best hyperparameters using the same setting discussed above. MLP and SAE-DNN share the same structure of DNN in fine-tuning phase that the layer size is [54, 108] and the top layer is a softmax layer to classify machine conditions. In SAE-DNN, before fine tuning, unsupervised training is performed by a stacked denoising autoencoder. The nonlinear activation function is set to be tanh. For the last four methods, the number of segments is $T = 20$ and, thus, the shape of the input sequence of local features is 20×27 and the sizes of hidden recurrent layer in these four RNN are unified to be the same as 50. In our proposed method, the sizes of FC-Layer 1 F_1 and FC-Layer 2 F_2 are set to be 50 and 200, respectively. The RMSprop is adopted to train these four models and the other hyperparameters including learning rates and epoch number are all searched based on validation dataset. To quantify the performance of all compared methods, the classification accuracy over five-fold training/testing splitting is reported.

In incipient fault diagnosis task, the dimensionality of the handcrafted feature vector is 10 as the one-channel vibration data is used. The layer size in MLP and SAE-DNN is [20, 40]. For the last four methods, the number of segments is $T = 20$ so that the shape of the input sequence of local features is 20×10 and the sizes of hidden recurrent layer in these four RNN are unified to be the same as 20. In our proposed method, the sizes of FC-Layer 1 F_1 and FC-Layer 2 F_2 are set to be 20 and 80, respectively. The other settings for all comparative methods and training/testing splitting scheme are kept the same as reported in the above gearbox fault diagnosis tasks.

C. Experimental Results

In tool wear prediction tasks which is more challenging compared to fault diagnosis tasks, two measures including MAE and RMSE are reported in Table III. It is shown that our proposed LFGRU achieves the lowest regression error among all

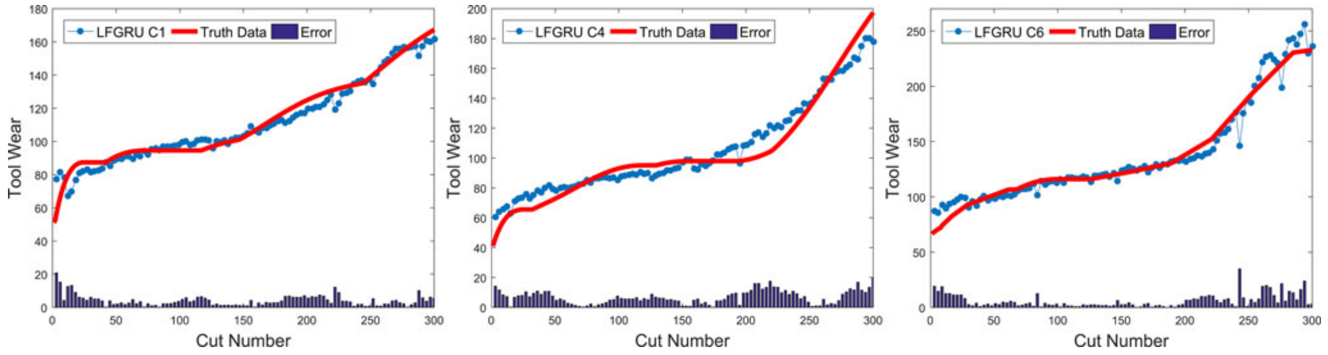

Fig. 8. Regression Analysis of tool wear predicted by LFGRU.

TABLE IV
CLASSIFICATION ACCURACY ACHIEVED BY COMPARED METHODS IN GEARBOX DIAGNOSIS TASKS AND INCIPIENT FAULT DIAGNOSIS

Algorithms	Bearing		Gear		Incipient
	20–0	30–2	20–0	30–2	Faults
SVM	0.833	0.886	0.944	0.901	0.972
KNN	0.808	0.864	0.932	0.892	0.956
MLP	0.865	0.905	0.843	0.906	0.968
SAE-DNN	0.875	0.921	0.927	0.919	0.974
RNN	0.929	0.920	0.923	0.893	0.956
GRU	0.912	0.924	0.938	0.905	0.981
BiGRU	0.930	0.936	0.938	0.907	0.985
LFGRU	**0.932**	**0.940**	**0.948**	**0.958**	**0.996**

Bold face indicates best performances.

compared methods. Compared to the most competitive model: BiGRU, the enhanced GRU networks concatenates the weighted averaging of local features and the output of BiGRU. Therefore, the experimental results have verified the effectiveness of the weighted feature averaging operation in our proposed model and recurrent models based on local features especially three GRU variants outperform the rest models, which have demonstrated that the temporal information encoded by GRU can boost the tool wear prediction. Finally, we found most of compared models perform slightly worse in dataset C6 than that in the rest two datasets. It may be explained by the fact that the distribution of $c6$ is different from the other two datasets $c1$ and $c4$. However, our proposed LFGRU model presents a robust performance over dataset C6. It further verifies the effectiveness of the hybrid combination of local-feature extraction and the enhanced GRU networks. Finally, to qualitatively demonstrate the effectiveness of our proposed model, the predicted tool wears under different datasets are illustrated in Fig 8. The actual tool wear conditions measured offline by a microscope are also displayed, respectively. It is found that the predicted tool wear overall are able to follow the trend of the truth data well.

In gearbox fault diagnosis tasks and incipient fault detection task, the results are shown in Table IV. We compare our proposed method with several state-of-the-art methods. The first observation is that recurrent models including RNN, GRU, BiGRU, and our proposed LFGRU that encode temporal information over the sequence of local features perform better than conventional data-driven models including SVM, KNN and MLP, and the DL model: SAE-DNN. Therefore, the importance of modeling the temporal dependency in fault diagnosis has been verified. Among all recurrent models based on local features, our proposed model outperforms other models. It should be explained by the introduction of bidirectional recurrent structure and center-biased feature averaging scheme. As shown in Table IV, the robust performance achieved by our proposed method in the above two fault diagnosis tasks verify the generalization capability of our method over various fault severity and machinery equipments.

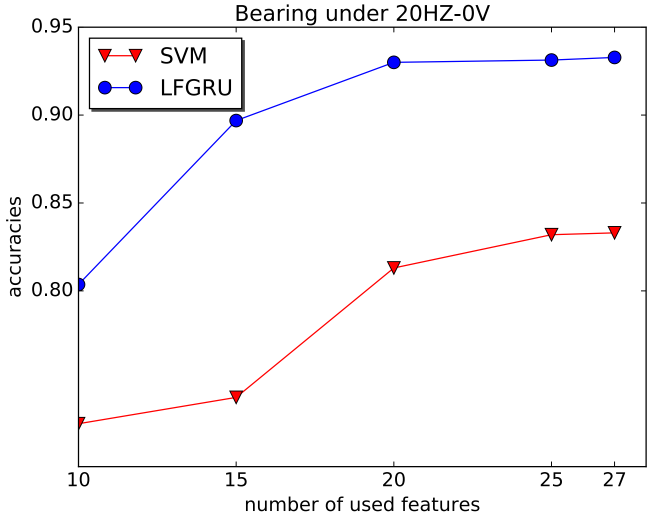

Fig. 9. Performances of SVM and LFGRU under different subsets of input features.

In addition, the robustness of our system on the quality of handcrafted design was investigated. We vary the number of adopted handcrafted features in the range [10, 15, 20, 25, 27] into SVM model and our proposed LFGRU and the classification accuracies on one dataset: bearing (20 HZ–0 V) are shown in Fig. 9. It is clear that even with ten hand-crafted features, the accuracy achieved by our method can be as high as 0.8, which significantly outperforms SVM model. To illustrate the effectiveness of our proposed method intuitively, the learned representation by our model has been visualized in Fig. 10. Here, we only report the results in gear dataset under a rotating speed of 30 HZ and a load of 2 V. t-SNE algorithm [32]

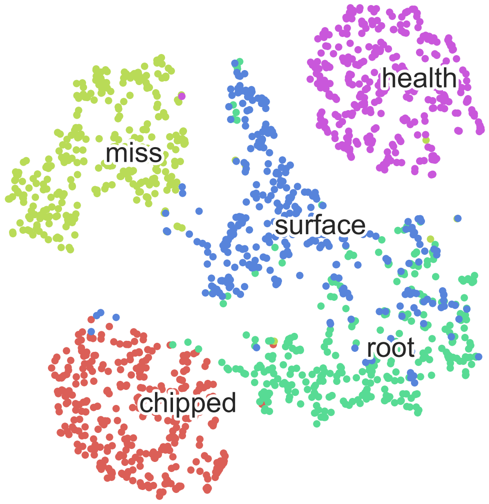

Fig. 10. Two-dimensional (2-D) projection of gear dataset. The representation is learned by LFGRU.

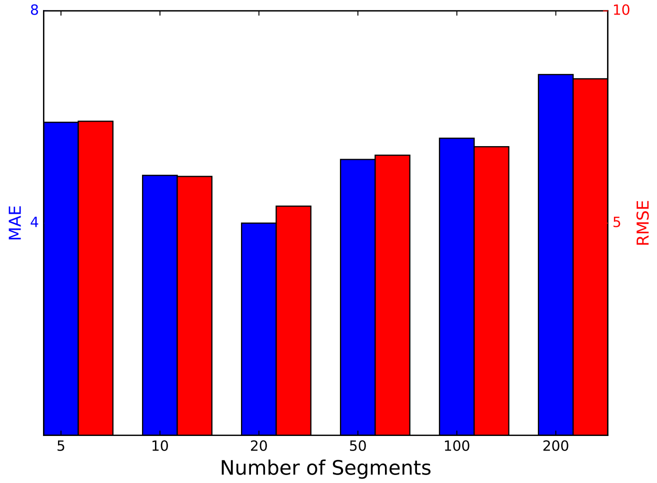

Fig. 11. Performances of our LFGRU under different number of segments.

is adopted to project the high-dimensional representation into a two-dimensional space. The separability demonstrates the capability of our model to learn discriminative and informative representation from mechanical signals.

At last, all of our experiments are conducted using four Nvidia GTX GPUs 1080 on a Linux Server with a 3.60 GHz Intel CPU. The testing time for each sample of our algorithm is only 0.009 s, which is suitable for the real-time monitoring.

D. Sensitivity Analysis on Number of Segments

Number of segments T is a hyperparameter to control the length of the sequential input into our proposed LFGRU model. It is clear that a small T will make the size of local segments large so that the too much sequential information may be lost. Under a large T, the small size of local segment may not be able to derive discriminative local features and the computational burden is increased at the same time. To verify the above statement empirically, we tested our proposed LFGRU under five additional different numbers of segments: $T = [5, 10, 50, 100, 200]$ and our previous setting $T = 20$ over one tool prediction task C1. The other parameters of our proposed LFGRU are kept unchanged as reported before. The RMSE and MAE are compared and shown in Fig. 11. It can be shown that very large and small T both hinder the performance of our proposed LFGRU. Under moderate numbers of segments such as 10 and 20, our proposed LFGRU model is able to achieve optimal performance.

V. Conclusion

A new DL based MHMS, i.e., LFGRU has been proposed. After local-feature extraction, a sequence of local features can be generated, which does not require a high-level expert knowledge. Then, an enhanced GRU networks is adopted to learn representation of the sequence of local features. In three real machine health monitoring tasks, the effectiveness and robustness of our proposed LFGRU model had been verified.

In the future work, we are going to explore the performance of our framework in the prognosis tasks. For example, the estimation of RUL can be casted into a regression problem [33], then our proposed technique may be adopted to predict the target RUL value from raw input signal.

References

[1] D. Lund, C. MacGillivray, V. Turner, and M. Morales, “Worldwide and regional internet of things (IoT) 2014–2020 forecast: A virtuous circle of proven value and demand,” Int. Data Corp., Framingham, MA, USA, Tech. Rep, 2014.

[2] 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.

[3] S. Yin, X. Li, H. Gao, and O. Kaynak, “Data-based techniques focused on modern industry: An overview,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.

[4] Z. Chen, H. Fang, and Y. Chang, “Weighted data-driven fault detection and isolation: A subspace-based approach and algorithms,” *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3290–3298, May 2016.

[5] A. Malhi and R. X. Gao, “PCA-based feature selection scheme for machine defect classification,” *IEEE Trans. Instrum. Meas.*, vol. 53, no. 6, pp. 1517–1525, Dec. 2004.

[6] J. Wang, J. Xie, R. Zhao, L. Zhang, and L. Duan, “Multisensory fusion based virtual tool wear sensing for ubiquitous manufacturing,” *Robot. Comput.-Integrated Manufacturing*, vol. 45, pp. 47–58, 2016.

[7] M. Fuente, G. Garcia, and G. Sainz, “Fault diagnosis in a plant using fisher discriminant analysis,” in *Proc. 16th Mediterranean Conf. Control Autom.*, 2008, pp. 53–58.

[8] Y. Lei, Z. He, Y. Zi, and X. Chen, “New clustering algorithm-based fault diagnosis using compensation distance evaluation technique,” *Mech. Syst. Signal Process.*, vol. 22, no. 2, pp. 419–435, 2008.

[9] A. Widodo and B.-S. Yang, “Support vector machine in machine condition monitoring and fault diagnosis,” *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, 2007.

[10] M. D. Prieto, G. Cirrincione, A. G. Espinosa, J. A. Ortega, and H. Henao, “Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks,” *IEEE Trans. Ind. Electron.*, vol. 60, no. 8, pp. 3398–3407, Aug. 2013.

[11] X. Yang, R. Yan, and R. X. Gao, “Induction motor fault diagnosis based on ensemble classifiers,” in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. Proc.*, May 2016, pp. 1–5.

[12] D. He, R. Li, and J. Zhu, “Plastic bearing fault diagnosis based on a two-step data mining approach,” *IEEE Trans. Ind. Electron.*, vol. 60, no. 8, pp. 3429–3440, Aug. 2013.

[13] 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.

[14] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

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

[16] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring: A survey," Dec. 2016. [Online]. Available: https://arxiv.org/abs/1612.07640

[17] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[18] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural Comput.*, vol. 20, no. 6, pp. 1631–1649, 2008.

[19] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, 2016.

[20] 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, 2016.

[21] L. Guo, H. Gao, H. Huang, X. He, and S. Li, "Multifeatures fusion and nonlinear dimension reduction for intelligent bearing condition monitoring," *Shock Vibration*, vol. 2016, 2016, Art. no. 4632562.

[22] M. Ma, X. Chen, S. Wang, Y. Liu, and W. Li, "Bearing degradation assessment based on weibull distribution and deep belief network," in *Proc. Int. Symp. Flexible Autom.*, 2016, pp. 1–4.

[23] Z. Chen, C. Li, and R.-V. Sánchez, "Multi-layer neural network with deep belief network for gearbox fault diagnosis," *J. Vibroeng.*, vol. 17, no. 5, pp. 2379–2392, 2015.

[24] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop on Deep Learn.*, Dec. 2014.

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

[26] R. Zhao, R. Yan, J. Wang, and K. Mao, "Learning to monitor machine health with convolutional bi-directional LSTM networks," *Sensors*, vol. 17, no. 2, pp. 273–290, 2017.

[27] H. Jaeger, *Tutorial on Training Recurrent Neural Networks, Covering BPPT, RTRL, EKF and the" Echo State Network" Approach*. GMD-Forschungszentrum Informationstechnik, Sankt Augustin, Germany 2002.

[28] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 1724–1734. [Online]. Available: http://www.aclweb.org/anthology/D14-11

[29] "2010 PHM data challenge." 2010. [Online]. Available: https://www.phmsociety.org/competition/phm/10

[30] Z. Wei, J. Gao, X. Zhong, Z. Jiang, and B. Ma, "Incipient fault diagnosis of rolling element bearing based on wavelet packet transform and energy operator," *WSEAS Trans. Syst.*, vol. 10, no. 3, pp. 81–90, 2011.

[31] T. Tieleman and G. Hinton, "Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[32] L. van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[33] G. S. Babu, P. Zhao, and X.-L. Li, "Deep convolutional neural network based regression approach for estimation of remaining useful life," in *Proc. Int. Conf. Database Systems Adv. Appl.*, Springer, 2016, pp. 214–228.

Authors' photographs and biographies not available at the time of publication.